# 21

# A Modeling Framework to Estimate and Project Species Distributions in Space and Time

*Niels Raes*[1] *and Jesús Aguirre-Gutiérrez*[1,2,3]

[1] *Naturalis Biodiversity Center, Leiden, Netherlands*
[2] *Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Netherlands*
[3] *Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK*

**Abstract**

Over the past decade, species distribution models (SDMs) have become an indispensable item in the ecologist's toolbox. SDMs, also known as ecological niche models, bioclimatic models or habitat suitability models, characterize the multivariate ecological space delimiting species' distributions and project this subset of ecological space back on to geography, resulting in a map of habitat suitability. Although SDMs build on correlations, they offer an important capacity to elucidate the altitudinal zoning on mountains, to forecast the effects of climate change on the risk of plant invasions onto mountains and to predict the potential of mountains as climate refugia, among many other ecological applications. There are more than 600 million digitized and georeferenced collection records currently available through the Global Biodiversity Information Facility (GBIF) portal, which, together with large amounts of spatial data on past, present and future climatic conditions, quantitative soil conditions and high-resolution topographic data, will further increase the popularity and applications of SDMs. In this chapter, we describe the principles and data requirements of SDMs, and provide an introduction to estimating species ranges from collection records and projecting these estimates in space and time.

**Keywords:** *ecological niche model, species distribution model, ensemble model, climate change, non-analog conditions, range shift*

## 21.1 Species Niches and Their Reciprocal Spatial Distributions

A species' niche consists of three components, and largely builds on Hutchinson's quantification of a species' niche as *n*-dimensional space that reflects suitable values of *n* biologically important and independent variables (e.g., temperature and precipitation) (Hutchinson 1957; Colwell & Rangel 2009; Blonder et al. 2014). Hutchinson's key innovation was the separation of the physical distribution of a species characterized by its geographical coordinates from the local values of *n* environmental conditions at a given time. The definition of a species' niche by *n* environmental attributes allows reciprocal projections between a species' niche and its present, past and future geographical distributions. Importantly, Hutchinson's niche expresses the effects of species interactions, but also the constraints of dispersal limitation (Colwell & Rangel 2009), known as a species' realized niche. Building on Hutchinson's niche concept, Soberón & Peterson (2005) developed the biotic, abiotic, movement (BAM) framework. The first and most important component of a species' niche is represented by the abiotic conditions within which a species population can establish and maintain itself, given its intrinsic physiological limits (Hutchinson 1957; Boulangeat et al. 2012). The second component consists of dispersal or movement limitations, which may prevent species from reaching sites with suitable abiotic conditions (e.g., a mountain range with suitable abiotic conditions separated by a vast lowland region, ocean, ocean strait or large river) (Bateman et al. 2013; Vasudev et al. 2015). The third component represents biotic interactions such as specific plant–pollinator interactions, the presence of pathogens or mutualistic relationships between plants and fungi or soil microbes. A species is present where all three niche components overlap, in what is known as its "realized niche" (Soberón & Nakamura 2009). It is from the realized niche that species presence records are collected, subsequently to be used in species

*Mountains, Climate and Biodiversity*, First Edition. Edited by Carina Hoorn, Allison Perrigo and Alexandre Antonelli.

Companion website: www.wiley.com\go\hoorn\mountains,climateandbiodiversity

distribution models (SDMs). The extent to which the three niche components overlap is often unknown, and this is a caveat of SDMs that should be taken in consideration. Advances are being made towards including biotic interactions (Boulangeat et al. 2012; Giannini et al. 2013; Thuiller et al. 2015) and dispersal limitations (Engler et al. 2012; Miller & Holloway 2015) in SDMs. The majority of SDM studies, however, estimate the spatial distribution of suitable abiotic niche conditions based on a species' realized niche. Despite this caveat, abiotic conditions, both at present and historically, govern at least the broadest outlines of the distribution of species and biomes (Thomas 2010; Boucher-Lalonde et al. 2016; Lee-Yaw et al. 2016). SDMs have successfully been used to forecast the effects of climate change on species' distributions (Thuiller et al. 2011), to identify historical refugia (Waltari et al. 2007) and map past distribution ranges (Raes et al. 2014), to predict the potential geographical ranges of invasive species (Broennimann et al. 2007) and to overcome (at least partly) the Wallacean shortfall (Hortal et al. 2015) or lack of knowledge on geographical distributions of species (Vollering et al. 2016), among many other applications (Araújo & Peterson 2012).

## 21.2 Species Presence Data

Without data on species occurrences from which to infer niche dimensions, it would be impossible to develop an SDM. These records are obtained from survey data and digitized herbarium and natural history museum specimens, which represent verifiable presences. The largest data portal with collection records is arguably the Global Biodiversity Information Facility data portal (www.gbif.org). For South America, the speciesLink data portal (www.splink.cria.org.br) is an additional source. Absence records are far more difficult to obtain, as "the absence of presence does not equal the presence of absence." Some SDM algorithms use presence-only data, while others require pseudo-absences or a background sample as replacement for true absence records (see Table 21.1).

Given that most species are rare, the number of records that are available to model the distributions of many species is limited. This poses a potential problem, as no relationship between species occurrence and abiotic conditions can be inferred based on only a few records. Various authors have used the subjective number of five spatially unique records as the absolute minimum requirement

**Table 21.1** SDM algorithms. The most widely used are indicated with bold text.

| SDM | Description | "Absence" data | References |
|---|---|---|---|
| ANNs | Artificial neural networks | Pseudo-absence | Hilbert & Ostendorf (2001) |
| BIOCLIM | Bioclimatic envelope – rectilinear | Presence only | Busby (1991) |
| BRTs | Boosted regression trees | Pseudo-absence | Elith et al. (2008) |
| CART | Classification and regression trees | Pseudo-absence | Breiman et al. (1984); De'ath & Fabricius (2000) |
| DOMAIN | Proximity to presences in multidimensional predictor space measured by the Gower metric | Presence only | Carpenter et al. (1993) |
| ENFA | Ecological niche factor analysis | Background sample | Hirzel et al. (2002) |
| GARP | Genetic algorithm for rule set prediction | Pseudo-absence | Stockwell & Peters (1999) |
| **GAMs** | **Generalized additive models** | **Pseudo-absence** | **Hastie & Tibshirani (1986); Yee & Mitchell (1991)** |
| GBMs | Generalized boosted models | Pseudo-absence | Ridgeway (1999) |
| GDM | Generalized dissimilarity modeling | Pseudo-absence | Ferrier et al. (2007) |
| **GLMs** | **Generalized linear models** | **Pseudo-absence** | **McCullagh & Nelder (1989); Venables & Ripley (2002)** |
| HABITAT | Bioclimatic envelope – convex hull | Presence only | Walker & Cocks (1991) |
| Mahalanobis distance | Multidimensional distance to the mean value for each predictor across presence localities | Presence only | Rotenberry et al. (2006); Calenge et al. (2008) |
| MARS | Multivariate adaptive regression splines | Pseudo-absence | Elith & Leathwick (2007) |
| **Maxent** | **Maximum entropy** | **Background sample** | **Phillips et al. (2006)** |
| MDA | Mixture discriminant analysis | Pseudo-absence | Hastie et al. (1994) |
| RFs | Random forests | Pseudo-absence | Breiman (2001) |
| SVMs | Support vector machines | Presence-only | Guo et al. (2005) |

(Pearson et al. 2007; Raes et al. 2014). A recent study has shown that the minimum required number of presence records depends on the prevalence, or proportional presence area, relative to the study region (van Proosdij et al. 2016). Prevalence values should range between 0.1 and 0.9 in order to obtain reliable results. Taking these considerations into account, and using virtual species distributions with a stringent accuracy test, the results of van Proosdij et al. (2016) indicate that at least 10 spatially unique presence records are required to calibrate an SDM. Additionally, these authors provide a methodology to arrive at an accurate estimate of the minimum required number of presence records for a given study region.

Although the region under study might cover only part of a species' range, it is important to include *all* available presence records for that species to calibrate an SDM, in order to avoid modeling partial or truncated niches (Raes 2012; Hannemann et al. 2016). Partial niche models tend to underestimate the probability of occurrence at the edges of "niche space" covered by the artificially delimited study region, and to overestimate it at the centre (Raes 2012). Furthermore, caution should be taken when modeling the distribution of invasive species. Lack of biotic interactions (e.g., pathogens, predators) or niche shifts in the invaded range can lead to the inclusion of presence records with abiotic conditions that do not exist in the native range, and hence potentially result in overprediction of the native range (Broennimann et al. 2007).

Another issue of concern is taxonomic synonyms. Institutes that contribute data to the global data portals may not always use the same taxonomy, or they may file records under synonymous names. Synonyms can be resolved using the Taxonomic Name Resolution Service (TNRS) (Boyle et al. 2013), while the Encyclopedia of Life (www.eol.org) provides synonyms for a wide taxonomic range of organisms. Moreover, specimens can also be stored under false taxonomic names as a result of misidentification (Goodwin et al. 2015), or as a result of belonging to as of yet undescribed taxa ("the Linnean shortfall") (Hortal et al. 2015).

Finally, geographical coordinates should be checked against specimen locality descriptions. Too often, latitudinal and longitudinal coordinates are reversed or centroid country coordinates are linked to specimens, among other potential sources of errors (Maldonado et al. 2015; Töpel et al. 2017).

## 21.3 Abiotic Spatial Data

### 21.3.1 Bioclimatic Variables

SDM algorithms are powerful tools that identify correlations between species presence records and abiotic – or, in fact, any spatially explicit – variables. Therefore, in order to obtain meaningful SDM results, it is important to select abiotic variables that relate to the ecological niche of the species. For the terrestrial realm, abiotic climatic conditions, such as temperature and precipitation, account for the majority of the spatial variation in the probability-of-occurrence estimation of a species (Boucher-Lalonde et al. 2012, 2014; Lee-Yaw et al. 2016). The widely used Bioclim data set consists of 19 bioclimatic variables derived from monthly minimum and maximum temperatures and monthly precipitation data (Hijmans et al. 2005). Bioclimatic variables represent biological limits such as "minimum temperature of the coldest month" or "precipitation of the driest quarter." Bioclimatic data sets are available at different spatial resolutions, ranging between 0.5 degree (~3000 $km^2$ at the equator) and 30 arc-seconds (~1 $km^2$), and can be downloaded from www.worldclim.org (Hijmans et al. 2005), www.climond.org (Kriticos et al. 2012), www.ccafs-climate.org and www.ecoclimate.org (Lima-Ribeiro et al. 2015).

### 21.3.2 Altitude and Derived Variables

In addition to bioclimatic variables, altitude can be used as an abiotic variable. Altitude seems relevant when modeling species distributions in montane regions. However, it is very often highly negatively correlated with the annual mean temperature, as temperature decreases with increasing altitude (Körner 2007). If the goal of an SDM is to predict the impact of future climate change on species distributions, or to project the model on to past climatic conditions, it is strongly advised not to use altitude as a variable: altitude is static, whereas global climate models (GCMs) predict increasing future temperatures, resulting in upslope range shifts of species.

Related to altitude is "topographic heterogeneity." The NASA Shuttle Radar Topographic Mission (SRTM) has delivered a digital elevation model at 3 arc-seconds, or 90 m spatial resolution, at the equator. When 90 m SRTM data are aggregated to resolutions between 1 $km^2$ and 5 arc-minutes (~9.3 × 9.3 km), the standard deviation (SD) around the mean is a measure of topographic heterogeneity. Altitudinal plains are represented by raster cells with low topographic heterogeneity values, and rugged mountainous terrains have high topographic heterogeneity values.

Additionally, slope and aspect can be derived from altitudinal data. Aspect describes the direction in which a slope faces, and relates to the degree of solar exposure. It should be noted that various other variables can be derived from altitudinal measurements and that the inclusion of topographic heterogeneity, slope and aspect variables in SDMs may be recommended instead of the inclusion of altitude.

### 21.3.3 Quantitative Soil Property Variables

A third category of abiotic variables is made up of quantitative soil variables such as pH, water holding capacity and organic carbon content. These abiotic variables have recently become available through various portals, such as the Harmonized World Soil Database (HWSD) (FAO/IIASA/ISRIC/ISSCAS/JRC 2012), SoilGrids1km (Hengl et al. 2014) and the European Soil DataBase (ESDB). The quantitative soil information is derived from interpolated US Food and Agricultural Organization (FAO) soil profile data, and is also available as categorical variables.

SDM algorithms that use regression modeling transform categorical data into presence/absence dummy variables, with one dummy variable for each category in the data layer (Franklin 2009). Thus, for regression models, the use of categorical variables may be unwanted – especially when many environmental variables are used and few species presence records are available, which may result in overfitted SDMs. Decision-tree algorithms may be a better option when handling categorical variables.

### 21.3.4 Land-Cover Data and Satellite Imagery

Land-cover data and satellite imagery with global coverage can be useful, but should be used with caution. Most land-cover data are interpretations of satellite images and/or aerial photos. The earliest satellite images, from Landsat 1, were taken in 1973, but many species collection records predate that year; a specimen collected in 1960 can easily be associated with agricultural land based on satellite imagery postdating 1973. Furthermore, while land cover, the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI) data may be useful for modeling animal distributions, we advise against the use of these sources for plant distributions, as this can be classified as circular reasoning. Furthermore, when the intention is to predict future (or past) species distributions under different climate change scenarios, it should be kept in mind that no future land-cover, NDVI or EVI data are readily available, although advances are being made (Martinuzzi et al. 2015). Land-cover data are useful, however, for correcting the predicted distributions of species for remaining natural vegetation cover by removing all areas classified as "urban" and "agricultural land" from the predicted distribution range.

### 21.3.5 Selecting Uncorrelated Abiotic Variables

Most SDM algorithms require uncorrelated predictor variables, in order to avoid problems with collinearity (Dormann et al. 2013). Once ecologically relevant predictors are identified, these can be tested for correlations with a Pearson's r-correlation test, or with a Spearman's rank correlation test in the case of non-normally distributed variables. As a rule of thumb, Pearson's $|r| > 0.7$ or Spearman's $|rho| > 0.7$ is an appropriate indicator for when collinearity begins to severely distort model estimations and subsequent predictions (Dormann et al. 2013).

Another measure of variable correlation or collinearity is the Variance Inflation Factor (VIF). A VIF value of >10 is often used to indicate high collinearity (O'brien 2007). From sets of correlated variables, the one with the highest ecological relevance should be kept to develop the SDM. Once all correlated predictors are removed, the correlation table should not have values above 0.7, or VIF values should not exceed 10.

### 21.3.6 Future and Past Bioclimatic Data

When the aim is to predict the impacts of future climate change on the distributions of species, data from global climate/circulation models (GCMs) are required. The latest report from the Intergovernmental Panel on Climate Change (IPCC) uses four different scenarios for global development, known as representative concentration pathways (RCPs), which lead to increased global average temperatures of between 2 and 4 °C (IPCC 2013). At local scales, the predicted increase in temperature can be much higher or lower, however. No fewer than 61 different GCMs, developed by 20 different institutes, have contributed to the latest IPCC-AR5 report (IPCC 2013). Details of the different GCMs can be found in the Climate Model Intercomparison Project – phase 5 (CMIP5) portal (Taylor et al. 2012). Given the complexity of GCMs, the spatial resolutions of the data are coarse, typically ranging between 1.0 and 2.75°.

To predict the future distributions of species, data from GCMs need to be downscaled to the desired spatial resolution. Two different methods are widely used: the Delta method (GCM portal) and the bias-corrected method (www.worldclim.org). The Delta method calculates the difference (anomaly) between predicted future values and recorded present values at the coarse spatial resolution of the GCM. These anomalies are then interpolated to the desired high spatial resolution used for modeling. Finally, the interpolated values are added to the present high-resolution values in order to maintain the high-resolution climate differences related to, for example, topographic differences. The bias-corrected method calculates the difference between the predicted future GCM values and the predicted present GCM values at the coarse resolution of the GCM. Not all GCMs correctly predict present values as derived from weather stations. The anomalies between predicted present and predicted future values are then interpolated to the

desired spatial resolution and added to the present data, which are interpolated from weather stations. This procedure corrects for biases in GCM predictions concerning present climatic conditions.

It might be equally interesting to assess past distributions of species; for example, to identify glacial refugia (Waltari et al. 2007) or to predict the vegetation types that covered exposed sea beds during glacial periods (Raes et al. 2014). Varela et al. (2015) provide a detailed summary of the available paleoclimatic data.

## 21.4 Species Distribution Models

The applications of SDMs are twofold. They can be used (i) to predict habitat suitability for areas where species collection records are lacking (Wallacean shortfall) and (ii) to describe a species' ecology based on its occurrence records and the abiotic conditions at those localities. SDMs combine data from species presence/absence records – taxonomically synonymized and georeferenced – with a selection of ecologically relevant and uncorrelated predictor variables (see Figure 21.1).

Over the past 2 decades, many different algorithms have been developed, compared and scrutinized in comparative tests (Elith et al. 2006; Aguirre-Gutiérrez et al. 2013; Qiao et al. 2015). Three main classes of modeling algorithms can be distinguished based on their requirements with respect to absence records (Table 21.1). The first class requires presence records only. The second requires absences, or pseudo-absences if true absences are lacking. Pseudo-absences are randomly drawn absences from the study area, taken from any locality where no presence was recorded. The third class do not require any absence data, but use a background sample defined as randomly drawn localities from the entire study area, including presence localities. Depending on the SDM algorithm, the "distributions" are either assumed to be parametric (normal, binomial, Poisson distribution) or are more relaxed in their assumptions (semi-parametric or nonparametric). We do not intend to be exhaustive here, nor to provide detailed descriptions of the different SDM algorithms. For that purpose, we refer to the textbooks of Franklin (2009), Peterson et al. (2011) and Guisan et al. (2017) and the references in Table 21.1. Presently, Maxent, GLM and GAM are the most widely used algorithms (Merow et al. 2013; Qiao et al. 2015).

Several of the SDM algorithms listed in Table 21.1 are implemented in applications (software) with a graphical user interface (GUI), notably Maxent. Most of them can be operated directly through R (R Development Core Team 2014). Several R-libraries have been developed especially for species distribution modeling, including "dismo" (Hijmans et al. 2015), "biomod2" (Thuiller et al. 2014) and "SSDM" (Schmitt et al. 2016). The R-vignette (manual) "Species Distribution Modeling with R" (Hijmans & Elith 2016) is highly recommended and covers the entire modeling process for various algorithms using the R framework.

### 21.4.1 Measures of SDM Accuracy and the Null-Model Test

Testing the accuracy of SDMs is challenging because independent test data are generally lacking. As a solution, presence records are often partitioned into a training and a testing data set. Either single partitions, multiple random partitions or *k*-fold partitions are used (e.g., 75% for training and 25% for testing) to develop SDMs and assess their predictive power on the test data. When the number of presence records is small, a jackknife (or "leave-one-out") procedure can be used (Pearson et al. 2007). For each run, one record is left out of the training data set and is used to measure the predictive accuracy. Both *k*-fold partitioning and the jackknife procedure result in a distribution of accuracy values that can be interpreted as the sensitivity of the SDM to different partitions of the data.

Most measures of SDM accuracy depend on a binary confusion matrix (Fielding & Bell 1997). A confusion matrix is a $2 \times 2$ contingency table that captures (i) the number of presences correctly predicted as present ("sensitivity"), (ii) the number of absences falsely predicted as present ("false positives" or "commission error"), (iii) the number of presences falsely predicted as absent ("false negatives" or "omission error") and (iv) the number of absences correctly predicted as absent ("specificity"). To calculate the different fractions of the confusion matrix, the continuous SDM output should first be converted into a discrete presence/absence prediction based on a threshold value. For an overview of different threshold rules, we refer to the work of Liu et al. (2013). We advocate the use of the "10 percentile training presence threshold." This is a conservative threshold that excludes 10% of the presence records with the lowest probability of occurrence from the predicted presence range and does not rely on absences (which are replaced by pseudo-absences). This threshold accounts for taxonomic misidentifications and georeferencing errors.

From the available threshold-dependent measures of SDM accuracy, Cohen's kappa statistic and true skill statistic (TSS) are widely used (Allouche et al. 2006). Arguably the most widespread, and one of the few threshold independent measures of SDM accuracy, is the area under the curve (AUC) of the receiver operating characteristic (ROC) plot (Hanley & McNeil 1982). The major advantage of the AUC value, in addition to its threshold independence, is that it is relatively insensitive

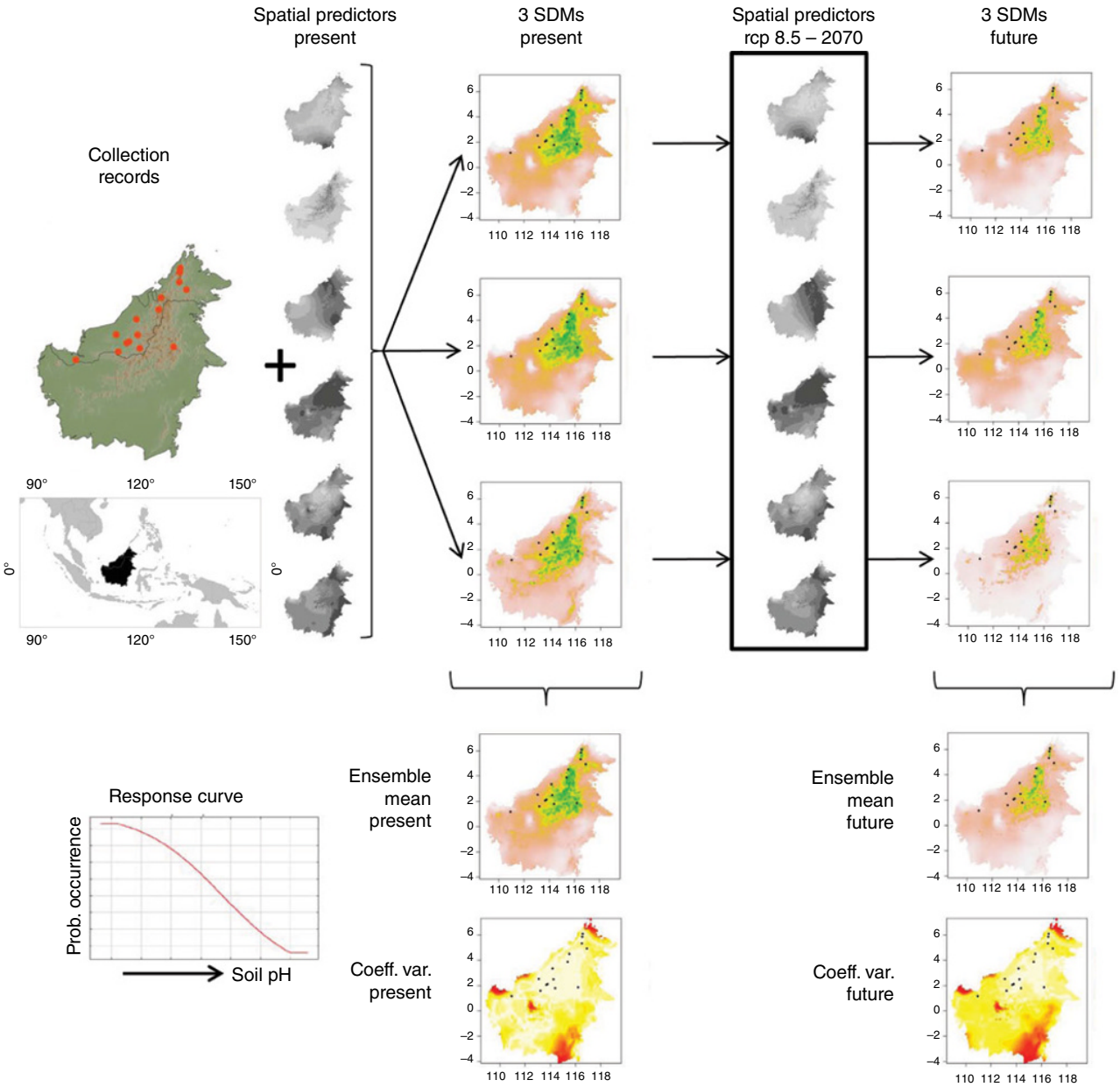


**Figure 21.1** Species distribution model (SDM) workflow for *Vaccinium phillyreoides* occurring on Borneo. Collection records (dots) and uncorrelated spatial predictors of present conditions are used to create three different SDMs using different algorithms (Table 21.1); white indicates high probability of occurrence. An ensemble (mean) of the three SDMs shows where the models agree and mapping of the coefficient of variation identifies areas where predictions are least consistent (dark gray). The SDMs are then projected to future climatic conditions (here, scenario RCP8.5, the most pessimistic climate change scenario, where greenhouse gas emissions continue increasing after the year 2100), resulting in three individual future projections. These are assembled in an ensemble mean forecast. The lower left corner shows a response curve of probability of occurrence decreasing with increasing Soil pH. See also Plate 34 in color plate section.

to prevalence (McPherson et al. 2004). The AUC value is a measure of the area under the curve of sensitivity (proportion of correctly predicted presences) plotted against 1-specificity (proportion of correctly predicted absences) for the range of all possible thresholds, and hence is threshold-independent. The AUC value is interpreted as the chance that a randomly drawn presence record has a higher predicted probability of occurrence value than a randomly drawn absence. For SDMs developed with presence and absence data, AUC values >0.7 are generally accepted as useful models, and an AUC value of 1 indicates perfect model fit (Swets et al. 2000).

A major drawback of all measures of SDM accuracy is that they rely on true absences, which are lacking in most

cases and are replaced by pseudo-absences or a background sample. Under such conditions, part of the pseudo-absences or background samples are randomly drawn from the species presence area proportional to the species' true presence distribution (prevalence). The maximum AUC value under these conditions is not 1, but $1 - a/2$, where $a$ stands in for the species' true prevalence, which is typically not known (Phillips et al. 2006; Raes & ter Steege 2007). For example, for a species with a prevalence of 0.4, the maximum AUC value is 0.8 (1 – 0.4/2). Therefore, measures of SDM accuracy that rely on standard threshold values (e.g., AUC > 0.7) that are calculated using pseudo-absences or a background sample instead of true absences are flawed. It should be noted, however, that when the aim is to compare the performance of different SDM algorithms on the same input data, the SDM with the highest AUC value is the most accurate.

Recognition of this caveat led Raes and ter Steege (2007) to develop a null-model that tests whether SDM accuracy values significantly deviate from random expectation. The procedure is straightforward, and uses a random sample of pseudo-presence records from the study area with the same number of records as was used for the real SDM. This is replicated 999 times. These 999 random sets of pseudo-presence records are modeled in a similar way as the real species, using the same SDM algorithm and abiotic spatial data. The 999 measures of SDM accuracy, plus the one measure of accuracy for the real species, are subsequently ranked from high to low. If the real species' measure of SDM accuracy ranks among the top 5%, then the chance that a random set of presence points can produce an equally good model is less than 5%; hence, significantly better than random expectation. The test can be further improved by drawing the pseudo-presence records from a target group background sample (Phillips et al. 2009). The target group background sample represents all presence records in the study area from species of the same group (e.g., same genus) to which the species being modeled belongs. This procedure also corrects for collection biases (Phillips et al. 2009).

### 21.4.2 SDM Complexity

Many of the SDM algorithms listed in Table 21.1 can fit very complex relationships between species presence records and spatial predictors. Complex SDMs often have very high model accuracy values but limited predictive power, as a result of model overparameterization or overfitting (Merow et al. 2014). Model overparameterization refers to the inclusion of too many predictor variables relative to the number of presences (and absences), or the inclusion of predictors that do not relate to the ecology of the study species. An overfit SDM is fitted to noise in the presence data, and fails to capture the species' response to environmental gradients. Ecological niche theory suggests that species' response curves are (at least for fundamental niches) often unimodal (Dolédec et al. 2000; Austin 2005, 2007), and hence quadratic responses to environmental gradients may be most appropriate (Figure 21.2) (Merow et al. 2013). When only part of a unimodal response is captured by the study area, a linear response might be sufficient. Threshold responses are appealing when physiological tolerance limits exist, such as a freezing intolerance resulting in predicted presence for areas where the temperature in the coldest month is above 0 °C. Any other modeling rules should only be included based on ecologically motivated reasoning.

### 21.4.3 Ensemble Models

Different SDM algorithms (Table 21.1), given their statistical assumptions and ways to handle absence data, result in different outputs from the same input data (Figure 21.1). The variation in output can occur not only when comparing different SDM algorithms, but also when comparing models from a single algorithm across multiple cross-validation runs and with different model parameterization. The between and within modeling variability in SDM outputs has been widely documented (Araújo & New 2007; Elith & Graham 2009), and has led to the development of ensemble models (EMs). EMs are rooted in the

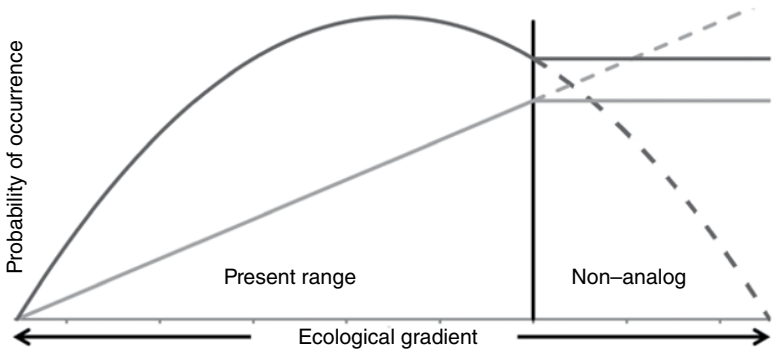


**Figure 21.2** A linear (light gray) and a unimodal or quadratic response curve (dark gray), covering the present and future/past non-analog range of values (abiotic conditions that are not present in the model training data) of an ecological gradient. Dotted lines represent extrapolation to non-analog future/past conditions. Horizontal lines represent the clamped values (future probability values are set constant at the value of the present range edge). The vertical line represents no extrapolation and the edge of the present range of values.

general idea described by Bates and Granger (1969) that, "Given that unknown conditions cannot be exactly predicted, then an ensemble of predictions may render a smaller error than any single prediction." Ensemble modeling can be summarized as a technique that captures the uncertainty in model predictions generated by different SDM algorithms, multiple cross-validation runs or different model parameterizations using a single SDM algorithm. Moreover, EMs may also render more consistent predictions when projecting models from different SDM algorithms to future climate scenarios.

A growing number of studies have compared the outputs of single SDMs against an EM. Aguirre-Gutiérrez et al. (2013) showed that EMs were among the best performing models, consistent across spatial scales, for different prevalence classes (i.e., widely to narrowly distributed species) and for rare to common species. Buisson et al. (2010) showed, using EMs of fish distributions in France, that most of the variation in future climate projections could be attributed to different SDM algorithms, followed by differences in GCMs. Similar conclusions were drawn by Diniz-Filho et al. (2009) for projections of bird SDMs onto future climate projections in South America. Next to differences in the outputs of different SDM algorithms and projections to different global climate change scenarios, differences in ensemble rules also yield different EM outputs.

Currently, the R-libraries "biomod2" (Thuiller et al. 2014), "BiodiversityR" (Kindt & Coe 2005) and "SSDM" (Schmitt et al. 2016) facilitate a semi-automated construction of EMs. First, individual SDMs using different algorithms are constructed and tested for their predictive power: only SDMs above a predefined threshold may be retained. Second, the ensemble rule should be selected. Among the choices are the mean, median and weighted mean. The latter applies a weight to the different SDMs according to the results of individual model evaluations. In this way, better performing SDMs drive the outcome of the final EM. Although it is tempting to present only the final projected EM, this output should be accompanied by a representation of the measure of uncertainty (Thuiller 2014). The uncertainty measure can be obtained by computing, for example, the coefficient of variation of single SDM predictions used to construct the EM (Figure 21.1). The measure of uncertainty indicates where single SDMs differ in their predictions.

### 21.4.4 The Ecology of Species as Derived from SDMs

Not all selected and uncorrelated abiotic variables contribute equally to an SDM. Often, just two or three variables largely determine the ecology of a species (Aguirre-Gutiérrez et al. 2015). Several methods have been developed to estimate the importance of each environmental variable to the final SDM. The first is a randomization procedure where the values of the variable under investigation are randomly permutated. Subsequently, the Pearson correlation between the predictions of the original SDM and the SDM with one permutated variable is determined. If the correlation is high (i.e., little difference between the two predictions), the permutated variable is considered unimportant for the SDM (Thuiller et al. 2009). This procedure is implemented in the R-library "biomod2" and can be used for all SDM algorithms (Thuiller et al. 2014). Maxent offers a similar permutation methodology, but here the relative drop in AUC value is used to determine the importance of the permutated variable (Phillips et al. 2006). Another method is a jackknife, or leave-one-out, analysis. Each environmental variable is left out of the SDM, and the relative drop in predictive power can be used as a measure of the variable's importance. Alternatively, an SDM can be fitted on a single environmental variable, and the predictive power can be used as a performance indicator for this variable alone.

Once the variables that determine a species' distribution are identified, the responses to these environmental gradients can be analyzed. For that purpose, Elith et al. (2005) developed the "evaluation strip," which allows the user to plot the response curves of an SDM to a single environmental gradient. The evaluation strip consists of generated environmental data, where the range of values of each variable in turn is systematically varied over its range, while all other variables are held constant at their mean (or minimum or maximum). Plotting of the predicted probability of occurrence values on the evaluation strip (response curves) shows how the model responds to increasing values of each variable independently (Figures 21.1 and 21.2). The evaluation strip is included in Maxent and in the R-library "biomod2" (Thuiller et al. 2014). Maxent has one additional type of response curve that is based on the predicted probability of occurrence of an SDM developed with a single environmental predictor, which can be readily plotted.

## 21.5 Projecting SDMs in Time and Space

When the aim of an SDM is to predict the impact of future climate change on species distributions (time), or to assess the invasive potential of a species in a certain area (space), the created SDM is projected to the future abiotic (often climatic) conditions, or to the new area of interest. In both situations, the data sets can include abiotic conditions that are not present in the training

data set, known as "novel" or "non-analog" conditions (Figure 21.2) (Williams & Jackson 2007). Projection of SDMs to novel conditions requires extrapolation of the species' responses along ecological gradients. This is especially troublesome if this involves linear responses (e.g., if the probability of occurrence increases with increasing temperature, in which case extrapolation results in continuous increasing probabilities beyond the present range of temperatures) (Figure 21.2).

Multivariate environmental similarity surfaces (MESS) measure the similarity of any given point to a reference set of points, and allow an analysis of the extent to which future values exceed the present range of values (Elith et al. 2010). Negative values indicate the maximum extension of the present range of values, expressed as a percentage; for example, if present temperatures range from 10–20 °C and the future value at a locality is 25 °C, then the MESS value is $(20 - 25)/(20 - 10) \times 100 = -50$. A future temperature of 5 °C results in the same negative MESS value. Future values within the present range have MESS values between 0 and 100. The most negative MESS value across all environmental variables reported for each locality is plotted, resulting in a MESS map. Predicted SDM probabilities of occurrence in areas with highly negative MESS values should be treated with caution. The MESS analysis is implemented in Maxent and in the R-library "dismo" (Hijmans et al. 2015).

Several SDM algorithms allow regulation of the degree of extrapolation to non-analog future values. The first option is not to extrapolate, effectively enforcing a value for the probability of occurrence of 0 at localities with future values exceeding the present range of values (vertical line in Figure 21.2). A second option is to "clamp" the probability of occurrence values beyond the present range of values. This sets the future probability values constant at the value of the present range edge (horizontal lines in Figure 21.2). The third option is to extrapolate the probability values to non-analog conditions (dashed lines in Figure 21.2).

A final word is dedicated to the dispersal capacity of species to track suitable future abiotic conditions. If a species is a poor disperser, it may not be able to migrate fast enough to keep up with shifting environmental conditions. The R-library "MigClim" (Engler et al. 2012) allows the user to integrate dispersal constraints on future SDM projections. Key to the calibration, however, is to select a realistic dispersal kernel (a probability density function of distance of dispersal), which is difficult because the kernel shape is determined by many variables. Furthermore, it is hard to include rare long-distance dispersal events. Therefore, most future SDM projections use either no dispersal, full dispersal or both.

## 21.6 Conclusion

SDMs are powerful statistical models that relate species presences to climatic and landscape features in order to predict their distributions and the potential impacts of future and past climatic conditions. We would like to stress that ecological knowledge cannot be disregarded and should be taken into account when selecting the predictors used to model species distributions. A model is only as good as the data used to train it. Although SDM outputs often look beautiful, it is of utmost importance to determine whether they reflect reality.

## Acknowledgements

We thank Rafael O. Wüest and Jose A.F. Diniz-Filho for useful comment on the manuscript, and Vitor Freitas for sharing R-scripts.

## References

Aguirre-Gutiérrez, J., Carvalheiro, L.G., Polce, C. et al. (2013) Fit-for-purpose: species distribution model performance depends on evaluation criteria – dutch hoverflies as a case study. *PLoS ONE* **8**, e63708.

Aguirre-Gutiérrez, J. & Serna-Chavez, H.M., Villalobos-Arambula, A.R. et al. (2015) Similar but not equivalent: ecological niche comparison across closely-related Mexican white pines. *Diversity and Distributions* **21**, 245–257.

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

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

Araújo, M.B. & Peterson, A.T. (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology* **93**, 1527–1539.

Austin, M.P. (2005) Vegetation and environment: discontinuities and continuities. In: van der Maarel, E. (ed.) *Vegetation Ecology*. Hoboken, NJ: Blackwell Science, pp. 52–84.

Austin, M.P. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling* **200**, 1–19.

Bateman, B.L., Murphy, H.T., Reside, A.E. et al. (2013) Appropriateness of full-, partial- and no-dispersal scenarios in climate change impact modelling. *Diversity and Distributions* **19**, 1224–1234.

Bates, J.M. & Granger, C.W. (1969) The combination of forecasts. *Or* **20**, 451–468.

Blonder, B., Lamanna, C., Violle, C. et al. (2014) The n-dimensional hypervolume. *Global Ecology and Biogeography* **23**, 595–609.

Boucher-Lalonde, V., Morin, A. & Currie, D.J. (2012) How are tree species distributed in climatic space? A simple and general pattern. *Global Ecology and Biogeography* **21**, 1157–1166.

Boucher-Lalonde, V., Morin, A. & Currie, D.J. (2014) A consistent occupancy–climate relationship across birds and mammals of the Americas. *Oikos* **123**, 1029–1036.

Boucher-Lalonde, V., Morin, A. & Currie, D.J. (2016) Can the richness–climate relationship be explained by systematic variations in how individual species' ranges relate to climate? *Global Ecology and Biogeography* **25**, 527–539.

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.

Boyle, B., Hopkins, N., Lu, Z. et al. (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics* **14**, 16.

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

Breiman, L., Friedman, J.H., Olshen, R.A. et al. (1984) *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth.

Broennimann, O., Treier, U.A., Muller-Scharer, H. et al. (2007) Evidence of climatic niche shift during biological invasion. *Ecology Letters* **10**, 701–709.

Buisson, L., Thuiller, W., Casajus, N. et al. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology* **16**, 1145–1157.

Busby, J.R. (1991) BIOCLIM – a bioclimate analysis and prediction system. In: Margules, R. & Austin, M.P. (eds.) *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*. Canberra: CSIRO, pp. 64–68.

Calenge, C., Darmon, G., Basille, M. et al. (2008) The factorial decomposition of the Mahalanobis distances in habitat selection studies. *Ecology* **89**, 555–566.

Carpenter, G., Gillison, A.N. & Winter, J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation* **2**, 667–680.

Colwell, R.K. & Rangel, T.F. (2009) Hutchinson's duality: the once and future niche. *Proceedings of the National Academy of Sciences* **106**, 19 651–19 658.

De'ath, G. & Fabricius, K.E. (2000) Classification and regression trees: a powerful yet simple technique for eological data analysis. *Ecology* **81**, 3178–3192.

Diniz-Filho, J.A.F., Bini, L.M., Rangel, T.F. et al. (2009) Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography* **32**, 897–906.

Dolédec, S., Chessel, D. & Gimaret-Carpentier, C. (2000) Niche separation in community analysis: a new method. *Ecology* **81**, 2914–2927.

Elith, J. & Graham, C.H. (2009) Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography* **32**, 66–77.

Elith, J. & Leathwick, J. (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* **13**, 265–275.

Elith, J., Ferrier, S., Huettmann, F. et al. (2005) The evaluation strip: a new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling* **186**, 280–289.

Elith, J., Graham, C.H., Anderson, R.P. et al. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129–151.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* **77**, 802–813.

Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution* **1**, 330–342.

Dormann, C.F., Elith, J., Bacher, S. et al. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**, 27–46.

Engler, R., Hordijk, W. & Guisan, A. (2012) The MIGCLIM R package – seamless integration of dispersal constraints into projections of species distribution models. *Ecography* **35**, 872–878.

FAO/IIASA/ISRIC/ISSCAS/JRC (2012) Harmonized World Soil Database, version 1.2. Available from: http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/ (last accessed August 30, 2017).

Ferrier, S., Manion, G., Elith, J. et al. (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions* **13**, 252–264.

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.

Franklin, J. (2009) *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge: Cambridge University Press.

Giannini, T.C., Chapman, D.S., Saraiva, A.M. et al. (2013) Improving species distribution models using biotic interactions: a case study of parasites, pollinators and plants. *Ecography* **36**, 649–656.

Goodwin, Z.A., Harris, D.J., Filer, D. et al. (2015) Widespread mistaken identity in tropical plant collections. *Current Biology* **25**, R1066–R1067.

Guisan, A., Thuiller, W. & Zimmermann, N.E. (2017) *Habitat suitability and distribution models; with applications in R*. Cambridge: Cambridge University Press.

Guo, Q., Kelly, M. & Graham, C.H. (2005) Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling* **182**, 75–90.

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

Hannemann, H., Willis, K.J. & Macias-Fauria, M. (2016) The devil is in the detail: unstable response functions in species distribution models challenge bulk ensemble modelling. *Global Ecology and Biogeography* **25**, 26–35.

Hastie, T. & Tibshirani, R. (1986) Generalized additive models. *Statistical Science* **1**, 297–310.

Hastie, T., Tibshirani, R. & Buja, A. (1994) Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association* **89**, 1255–1270.

Hengl, T., de Jesus, J.M., MacMillan, R.A. et al. (2014) SoilGrids1km – global soil information based on automated mapping. *PLoS ONE* **9**, e105992.

Hijmans, R. & Elith, J. (2013). Species distribution modeling with R. *Encyclopedia of Biodiversity* **6**, 10.1016/B978-0-12-384719-5.00318-X.

Hijmans, R.J., Cameron, S.E., Parra, J.L. et al. (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. et al. (2015) Dismo: Species Distribution Modeling. R package version 1.0–12.

Hilbert, D.W. & Ostendorf, B. (2001) The utility of artificial neural networks for modelling the distribution of vegetation in past, present and future climates. *Ecological Modelling* **146**, 311–327.

Hirzel, A.H., Hauser, J., Chessel, D. et al. (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data. *Ecology* **83**, 2027–2036.

Hortal, J., Bello, F.d., Diniz-Filho, J.A.F. et al. (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* **46**, 523–549.

Hutchinson, G.E. (1957) Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology, pp. 415–427.

IPCC (2013) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Kindt, R. & Coe, R. (2005) *Tree Diversity Analysis. A Manual and Software for Common Statistical Methods for Ecological and Biodiversity Studies*. Nairobi: World Agroforestry Centre (ICRAF).

Körner, C. (2007) The use of "altitude" in ecological research. *Trends in Ecology & Evolution* **22**, 569–574.

Kriticos, D.J., Webber, B.L., Leriche, A. et al. (2012) CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. *Methods in Ecology and Evolution* **3**, 53–64.

Lee-Yaw, J.A., Kharouba, H.M., Bontrager, M. et al. (2016) A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits. *Ecology Letters* **19**, 710–722.

Lima-Ribeiro, M.S., Varela, S., González-Hernández, J. et al. (2015) EcoClimate: a database of climate data from multiple models for past, present, and future for macroecologists and biogeographers. *Biodiversity Informatics* **10**, 1–21.

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.

Maldonado, C., Molina, C.I., Zizka, A. et al. (2015) Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? *Global Ecology and Biogeography* **24**, 973–984.

Martinuzzi, S., Radeloff, V.C., Joppa, L.N. et al. (2015) Scenarios of future land use change around United States' protected areas. *Biological Conservation* **184**, 446–455.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*. Boca Raton, FL: CRC Press.

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

Merow, C., Smith, M.J. & Silander, J.A. (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* **36**, 1058–1069.

Merow, C., Smith, M.J., Edwards, T.C. et al. (2014) What do we gain from simplicity versus complexity in species distribution models? *Ecography* **37**, 1267–1281.

Miller, J.A. & Holloway, P. (2015) Incorporating movement in species distribution models. *Progress in Physical Geography* **39**, 837–849.

O'brien, R.M. (2007) A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity* **41**, 673–690.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. et al. (2007) Predicting species distributions from small

numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography* **34**, 102–117.
Peterson, A.T., Soberón, J., Pearson, R.G. et al. (2011) *Ecological Niches and Geographic Distributions*. Princeton, NJ: Princeton University Press.
Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**, 231–259.
Phillips, S.J., Dudík, M., Elith, J. et al. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* **19**, 181–197.
Qiao, H., Soberón, J. & Peterson, T.A. (2015) No silver bullets in correlative ecological niche modeling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution* **6**, 1126–1136.
R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Raes, N. (2012) Partial versus full species distribution models. *Natureza & Conservação* **10**, 127–138.
Raes, N., Cannon, C.H., Hijmans, R.J. et al. (2014) Historical distribution of Sundaland's Dipterocarp rainforests at Quaternary glacial maxima. *Proceedings of the National Academy of Sciences* **111**, 16790–16795.
Raes, N. & ter Steege, H. (2007) A null-model for significance testing of presence-only species distribution models. *Ecography* **30**, 727–736.
Ridgeway, G. (1999) The state of boosting. *Computing Science and Statistics* **31**, 172–181.
Rotenberry, J.T., Preston, K.L. & Knick, S.T. (2006) GIS-based niche modeling fro mapping species' habitat. *Ecology* **87**, 1458–1464.
Schmitt, S., Pouteau, R., Justeau, D. et al. (2016) SSDM: Stacked Species Distribution Modelling. R package version 0.1.1.
Soberón, J. & Nakamura, M. (2009) Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences* **106**, 19644–19650.
Soberón, J. & Peterson, A.T. (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics* **2**, 1–10.
Stockwell, D. & Peters, D. (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* **13**, 143–158.
Swets, J.A., Dawes, R.M. & Monahan, J. (2000) Better decisions through science. *Scientific American* **283**, 82–87.
Taylor, K.E., Stouffer, R.J. & Meehl, G.A. (2012) An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society* **93**, 485–498.
Thomas, C.D. (2010) Climate, climate change and range boundaries. *Diversity and Distributions* **16**, 488–495.
Thuiller, W. (2014) Editorial commentary on "BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change." *Global Change Biology* **20**, 3591–3592.
Thuiller, W., Lafourcade, B., Engler, R. et al. (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography* **32**, 369–373.
Thuiller, W., Lavergne, S., Roquet, C. et al. (2011) Consequences of climate change on the tree of life in Europe. *Nature* **470**, 531–534.
Thuiller, W., Georges, D. & Engler, R. (2014) Biomod2: ensemble platform for species distribution modeling. R package version 3.1-64.
Thuiller, W., Pollock, L.J., Gueguen, M. et al. (2015) From species distributions to meta-communities. *Ecology Letters* **18**, 1321–1328.
Töpel, M., Zizka, A., Calió, M.F. et al. (2017) SpeciesGeoCoder: Fast Categorization of Species Occurrences for Analyses of Biodiversity, Biogeography, Ecology, and Evolution. *Systematic Biology* **66**, 145–151.
van Proosdij, A.S.J., Sosef, M.S.M., Wieringa, J.J. et al. (2016) Minimum required number of specimen records to develop accurate species distribution models. *Ecography* **39**, 542–552.
Varela, S., Lima-Ribeiro, M.S. & Terribile, L.C. (2015) A short guide to the climatic variables of the Last Glacial Maximum for biogeographers. *PLoS ONE* **10**, e0129037.
Vasudev, D., Fletcher, R.J., Goswami, V.R. et al. (2015) From dispersal constraints to landscape connectivity: lessons from species distribution modeling. *Ecography* **38**, 967–978.
Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S-PLUS*, 4th edn. New York: Springer.
Vollering, J., Schuiteman, A., de Vogel, E. et al. (2016) Phytogeography of New Guinean orchids: patterns of species richness and turnover. *Journal of Biogeography* **43**, 204–214.
Walker, P.A. & Cocks, K.D. (1991) HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters* **1**, 108–118.
Waltari, E., Hijmans, R.J., Peterson, A.T. et al. (2007) Locating pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS ONE* **2**, e563.
Williams, J.W. & Jackson, S.T. (2007) Novel climates, no-analog communities, and ecological surprises. *Frontiers in Ecology and the Environment* **5**, 475–482.
Yee, T.W. & Mitchell, N.D. (1991) Generalized additive models in plant ecology. *Journal of Vegetation Science* **2**, 587–602.