Modeling niches and mapping distributions: progress and promise of ecological niche models for primate research

Kenneth L. Chiou[1] and Mary E. Blair[2]

[1]Department of Anthropology, Washington University, St. Louis, MO 63130, USA
[2]Center for Biodiversity and Conservation, American Museum of Natural History, New York, NY 10024, USA

December 2015

1   Introduction

In this chapter, we briefly review the burgeoning field of ecological niche modeling and explore its relevance to studies of primate ecology, evolution, and conservation. Recent years have witnessed an explosion of interest in ecological niche models, spurred on by the increasing availability of occurrence data and spatially explicit environmental data as well as GIS tools and technologies appropriate for processing the increasingly high-resolution and multidimensional data typical of this field.

Ecological niche models (ENMs), sometimes known as species distribution models (SDMs), relate observed points of occurrence to spatially explicit environmental variables thought to be relevant in shaping the ecological tolerances of a given taxon or population (Franklin, 2009; Peterson et al., 2011). ENMs are often used to predict species distributions—such as in biogeography, conservation management (including climate change assessment), or pathogen modeling—or they can focus on the models themselves as representations of species ecological niches. When strong associations are found between occurrences and environmental predictors, ENMs can lend important insights into the ecological tolerances or habitat preferences of species or populations.

While newcomers can surely appreciate the exciting applications made possible by ENMs, the deluge of alternative methods and implementations can be overwhelming.

Some of these methods emphasize distinct theoretical issues concerning the nature of ecological niches. Others emphasize statistical or computational issues. In light of the abundance of considerations, many newcomers simply run the most available software, usually without modifying default settings. In doing so, they may overlook more suitable methods or fail to correct for biases or other shortcomings in their datasets.

In this chapter, we review previous applications of ENMs in primate studies and highlight areas of primatology for which ENMs can lend valuable insights. We also review some of the factors to consider when selecting and running models, and attempt to illustrate this with a brief case study. Our goal is to minimize the difficulties that many encounter while learning to develop, employ, and interpret ENMs and to encourage others to explore this promising tool for applications in primatology.

2   Ecological Niche Models in Primatology

Despite recent expansion of ENM approaches and their relevance to our understanding of primate ecology, evolution, and conservation, applications to primate populations are still relatively rare in comparison to other taxonomic groups. Here we briefly review some of the ways in which ENMs are being used in studies of primate populations.

2.1   Habitat suitability models

The most common application of ENMs to studies of primate populations has been through the approach of habitat suitability modeling (reviewed in Rode et al., 2013), which focuses on likelihood of occurrence rather than the fitness of species (Hirzel & Le Lay, 2008). The outputs of habitat suitability models are often incorporated into gap analyses—a method for identifying spatial "gaps" in existing protected area management plans by determining whether particular environments are poorly represented in existing areas (Scott et al., 1987)—that can be used for for species-, site-, country-, or regional-level conservation planning (Morales-Jimenez et al., 2005; Rode et al., 2013). Also, for rare and elusive species for which complete distribution data are unavailable, habitat suitability models can be useful for directing future survey efforts (Rode et al., 2013).

In addition to gap analyses, studies have inferred suitable habitat for regional and site-level conservation planning (Peck et al., 2011; Vidal-Garca & Serio-Silva, 2011; Junker et al., 2012; Campos & Jack, 2013; Hickey et al., 2013; Pintea et al., 2014). Coudrat & Nekaris (2013), for instance, used Maxent to model suitable

habitat for four macaque species (*Macaca arctoides*, *M. assamensis*, *M. leonina*, *M. mulatta*) in Nakai-Nam Theun National Protected Area, central-eastern Laos and found important differences in characteristics of suitable habitats across species, with implications for conservation management. Hickey et al. (2013) used biotic and abiotic data (including fragmentation) associated with bonobo (*Pan paniscus*) nests to predict suitable conditions across the species' range.

For rare and elusive species, ENMs can be used to help locate areas that might provide suitable habitat for a species but have not yet been surveyed. Boubli & Lima (2009), for example, modeled the suitable habitat of brown-backed bearded sakis (*Chiropotes israelita*) and black uakaris (*Cacajao melanocephalus*, *C. hosomi*, and *C. ayresi*) in remote regions of western Amazonas, Brazil using Maxent to guide future survey expeditions into these remote areas. Also, Thorn et al. (2009) and Voskamp et al. (2014) modeled suitable habitat in Maxent for Asian slow lorises in Borneo, Java, and Sumatra, and compared the model to protected area and land-use information to define areas that should be high priority survey sites, sites for reintroduction, or sites for possible protected area extensions. Similar approaches may be useful for identifying priority areas for archaeological or paleontological surveys by predicting site distributions using ENMs (Beeton et al., 2013), which can help complement other GIS approaches (e.g., Conroy et al., 2012).

2.2 Modeling the potential effects of future climate change

Although threats such as hunting and habitat disturbance may be more immediate concerns for primate population persistence as compared to stressors related to climate change, assessing likely future exposure to climate change across a species' range can nevertheless play an important role in conservation prioritization and planning for the long-term survival of species (Blair et al., 2012). Models and empirical analyses of range shifts in other taxonomic groups suggest that species show individualistic responses to changing climates (Peterson et al., 2011). Future climate projection data from the Intergovernmental Panel on Climate Change (IPCC) are available for at least three emissions scenarios (A1, A2, B2), seven general circulation models (GCMs), and seven time periods (`http://www.ccafs-climate.org/`). ENMs can be calibrated in the current climate and then projected into the future using IPCC scenarios to identify the parts of a species' range that are expected to be most exposed to changes in temperature and precipitation.

A growing number of studies project primate ENMs into the future using IPCC scenarios. Most studies focus on projecting shifts in climatically suitable areas for a species or group of species, allowing an estimation of potential future habitat loss

as well as refuge areas that might be prioritized in conservation efforts. Examples include studies of Francois' langurs and Tonkin snub-nosed monkeys (Vu et al., 2010, 2011), Sichuan snub-nosed monkeys (Luo et al., 2014), lion tamarins (Meyer et al., 2014), and Bornean orang-utans (Struebig et al., 2015). Another application to primate populations is the use of ENMs to project climatic suitability for vegetation species providing key habitat, rather than to project climatic suitability for the individual species themselves. Wong et al. (2013), for example, used Maxent to model and project the distribution of key vegetation types of importance to the Yunnan snub-nosed monkey under IPCC scenarios to infer and plan for the potential effects of climate change on this species. Common criticism of these applications of ENMs include failure to capture uncertainty across model algorithms and scenarios and a limited ability for validation of future projections. One approach to improve validation is to use a two-step modelling process, forecasting from the recent historical past (when such data are available) and validating the forecast with present occurrence data, then forecasting to the future; Chatterjee et al. (2012) modeled gibbon distributions in China across three time intervals using fossil and historical data, current data, and IPCC scenarios, providing additional validation for their projections across time. Further, ensemble forecasting has been presented as a method that better captures the uncertainty of modeling into the future by calibrating ensembles, or sets of ENMs, using various algorithms projected to a suite of future climate scenarios and exploring the resulting range of uncertainties (Arajo & New, 2007). This approach has been theoretically and empirically shown to outperform forecasts by individual models in predictive ability. Brown & Yoder (2015) modeled ENMs for 57 lemur species under future climate scenarios using an ensemble approach. They found that 60% the lemur species modeled will experience considerable range reductions entirely due to future climate change. When taking into account the uncertainties, climate change projections can be useful to determine general trends in terms of areas and species likely to be the most affected by habitat loss due to climate change. In order to most accurately estimate local extinction risks due to climate change, however, it may be important to couple ENMs with models of dispersal, demography, or interactions among species (Peterson et al., 2011; Sterling et al., 2013).

2.3 Biogeography, niche evolution, and ecological speciation

ENMs can provide considerable insight into the biogeographic history of organisms including species' environmental requirements, potential barriers that limit dispersal, and the degree to which ecological niches change over evolutionary timescales (Wiens, 2004; Raxworthy et al., 2007; MacColl, 2011; Bett et al., 2012; Blair et al., 2013b).

Bett et al. (2012) constructed ENMs for grey-, red-, and black-shanked doucs (*Pygathrix cinerea*, *P. nemaeus*, and *P. nigripes*) to investigate the north-to-south distribution turnover common across several primate groups in Vietnam, Laos, and Cambodia. The ENMs successfully predicted the north-to-south gradient based only on climatic variables, and the authors used the differences between the most important variables in each model to discuss whether a zoogeographic barrier or current or historical climatic shifts are most likely to explain the biogeographic history of this group. A similar approach was applied to explain the parapatric distribution of brown and black-and-gold howler monkeys (Holzmann et al., 2015). Paleoclimate layers are increasingly available and may further inform questions of primate biogeographic histories by facilitating the modeling of past species distributions, calibrated with fossil data.

Blair et al. (2013b) produced ENMs for eight species of *Eulemur* to test for ecological niche conservatism or divergence between sister species pairs, defined as species that are the only descendants of a common ancestral species. Ecological niche divergence refers to the tendency for related species to differ more ecologically than expected by random drift (or simple Brownian motion descent with modification (Losos, 2008), while ecological niche conservatism refers to the tendency for related species to differ less ecologically than expected by random drift. Blair and colleagues tested for ecological niche divergence or conservatism using the null model developed by Warren et al. (2008), in which observed ENM overlap was compared to a null distribution of overlap values generated from random points within the geographic range of the species pair. For three sister-pair comparisons, the tests supported the null model that niches are no more divergent than the available background region. Combined with the presence of a riverine barrier between these pairs, the null results are consistent with an allopatric speciation model. For the sister pair *E. flavifrons-E. macaco*, however, Blair and colleagues found support for significant niche divergence (Figure 1). Consistent with their parapatric distribution on an ecotone and the lack of obvious geographic barriers, these findings support a parapatric model of speciation and support overall the idea that multiple speciation processes are at work among these closely related lemurs.

ENMs can also be used to predict the geography of hybrid zones. Indeed, the overlap between the above ENMs (Blair et al., 2013b) coincides with known hybrid zones between *Eulemur* species, including the known hybrid zone between *E. macaco* and *E. flavifrons* (Figure 1). Innovative work is being done in other taxa to predict hybrid zones and other species interactions in recognition of the fact that biotic interactions may be as prominent as climate in determining species distributions. A recent study, for instance, attempted to disentangle the roles of biotic interactions
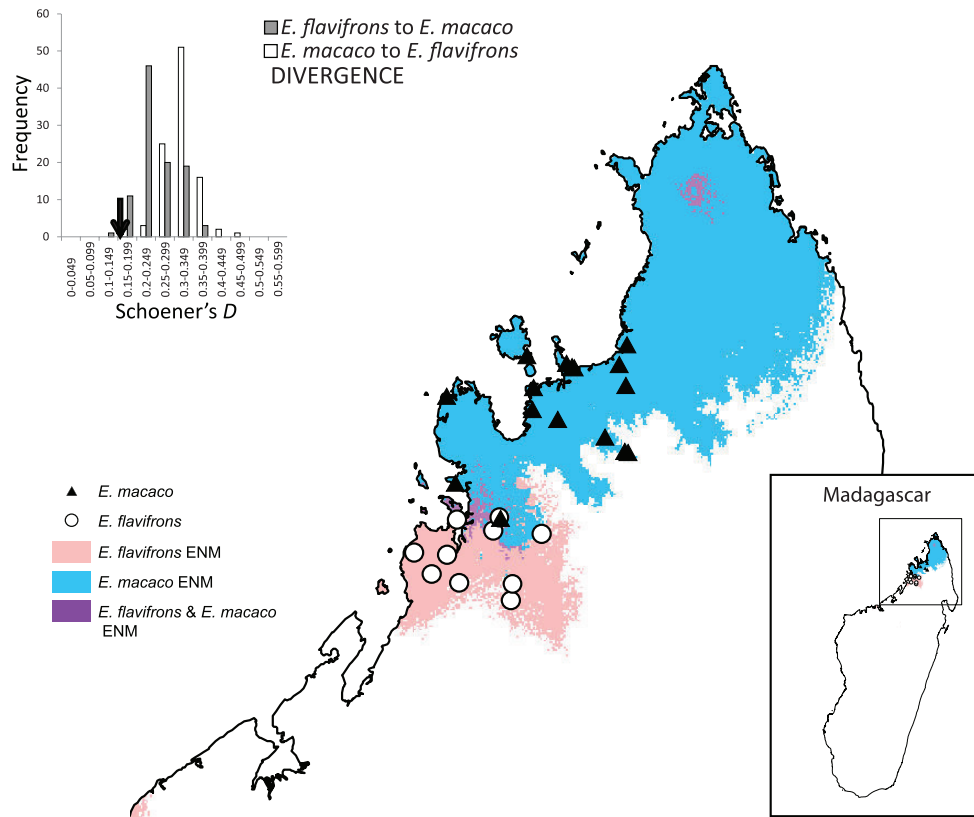
5

Figure 1: ENMs, occurrence records, and histograms showing the results of the null distribution test with 100 replicates for *Eulemur flavifrons-Eulemur macaco*. The black arrow on the histograms shows where the actual overlap between the two ENMs falls compared to the null distribution of pseudo-replicated niche overlap values. Adapted from Blair et al. (2013b).

6

and climate in determining the location of a moving hybrid zone between the breeding ranges of two parapatric passerines in Europe, and found that interspecific interactions, not climate, accounted for the present location of the contact zone (Engler et al., 2013).

Researchers have also used ENMs to identify cryptic species lineages for which processes of morphological differentiation, lineage sorting, and the formation of reproductive barriers may be incomplete but ecological attributes have differentiated. This approach has been most commonly applied to reptiles, amphibians, or birds (e.g., Raxworthy et al., 2007), but has also been applied to primate populations. Kumara et al. (2009), for example, mapped the distribution of various subspecies of slender lorises in peninsular India using ENMs. Their results indicated the presence of a previously undescribed, geographically disjunct, and ecologically unique subspecies occupying a distinct and intermediate climate region running along the eastern fringe of the southern Western Ghats. The study also described morphological differences, and recommended an urgent need for detailed exploration and conservation action. Other recent studies have documented how both environmental variation and rivers shape patterns of genetic differentiation between chimpanzee subspecies (Mitchell et al., 2015; Sesink Clee et al., 2015).

3   Which Ecological Niche Model?

Selecting and implementing a particular type of ENM for any given research question can be an intimidating task requiring rumination over issues such as sample size, sample bias, spatial scale, environmental correlations, model complexity, desired statistical metrics, and accessibility of algorithms and software. The process of model selection can be particularly agonizing due to the abundance of methods to choose from (Table 1 and Figure 2) and the knowledge that discrepancies among model performances can sometimes be very large, particularly for the projection of species distributions into independent or unknown circumstances (Thuiller, 2003; Pearson et al., 2006; Peterson et al., 2007). Our goal in this section is not to cover all of these methods, but rather to highlight how they differ and to indicate important issues to consider when choosing among them. We hope that this review will encourage readers to explore available methods, justify their choices, and to apply them effectively. Because our focus is on the practical implementation of ENMs, we use the terms "methods" and "models" broadly to refer to modeling approaches that have been published and are generally available as software. This encompasses not only their algorithms but also the factors that influence other key elements of the modeling process such as characteristics of occurrence data, choice of environmental variables,

selection of decision thresholds, and prevention of overfitting.

The majority of studies discussed in this chapter thus far have used correlative ENMs, which use associations between species occurrence records and environmental variables to characterize the environments within which species can exist or are likely to exist. The advantages and disadvantages of these correlative ENMs as compared to mechanistic or process-based ENMs (Kearney & Porter, 2009) have been widely discussed in the literature (e.g., Pearson & Dawson, 2003; Dawson et al., 2011). Criticisms of correlative ENMs have commonly focused on their inability to yield information about the underlying mechanisms that limit species distributions. An important advantage of correlative ENMs, however, is that detailed knowledge of the functional traits of organisms is not required. Correlative methods therefore have the potential to be applied rapidly to a large number of species, as well as to rare or poorly understood species including many primates. Lehmann et al. (2010) used a mechanistic (time budget) model to link climate variables to the behavior and biogeography of great apes, for which detailed information is available. Another recent study, however, compared a mechanistic time-budget model to a correlative Maxent model for vervets and reported that the two approaches produced remarkably similar predictions for vervet distribution despite the conceptual and methodological differences between these two modeling approaches (Willems & Hill, 2009). The remainder of this chapter focuses on correlative approaches, which we anticipate to be more readily applicable to primate datasets in the near term.

3.1   Accessibility

One of the most important considerations for selecting ENMs is more practical than methodological. While many methods have been published, not all are equally accessible. Some software implementations are not made available. Others may be disseminated as software extensions or libraries, or as standalone programs, but nevertheless remain inaccessible due to cost (e.g., software is proprietary), system requirements (e.g., software is platform-specific), or difficulty of use. For some projects, certain methods may be undesirable due to the computational time involved.

The most accessible methods are released as software that is freely available and cross-platform, with detailed documentation on model algorithms, settings, and parameters. The inclusion of tutorials or sample datasets can be an effective complement to other documentation (e.g., Phillips, 2009). The availability of a graphical user interface (GUI) can also increase accessibility by reducing the learning curve for operating the tool. Maxent and GARP, for instance, both have standalone GUI versions (Scachetti-Pereira, 2002; Phillips et al., 2006) while ModEco (Guo & Liu, 2010)

Table 1: Partial list of ecological niche modeling methods

Method	Description	Statistical approach	Data[1]	Software[2]	References
ANN	artificial neural networks	machine learning	PA*/Ab	R (nnet), SNNS, openModeller, ModEco	Manel et al. 1999 Pearson et al. 2002
BIOCLIM	bioclimatic envelope model	distance	PO	R (dismo, biomod2), openModeller, DIVA-GIS, ModEco	Nix 1986 Busby 1991
BN	Bayesian networks	machine learning	PA*	ModEco, Elvira	Friedman et al. 1997 Aguilera et al. 2010
BRT	boosted regression trees	machine learning	PA*	R (gbm, dismo)	Leathwick et al. 2006b Elith et al. 2008
BRUTO	generalized additive model with adaptive back-fitting	regression	PA*/Ab	R (mda)	Leathwick et al. 2006a
CART	classification and regression trees (also known as decision trees)	machine learning	PA*/Ab	R (rpart), ModEco	De'ath & Fabricius 2000 Rouget et al. 2001
DOMAIN	continuous point-to-point similarity metric (Gower metric)	distance	PO	R (dismo), openModeller, DIVA-GIS, ModEco	Carpenter et al. 1993
ENFA	ecological niche factor analysis	distance	PB	Biomapper, openModeller	Hirzel et al. 2002
GAM	generalized additive models	regression	PA*/Ab	R (mgcv, gam)	Guisan et al. 2002
GARP	genetic algorithm for rule set production	machine learning	PP	openModeller, DesktopGarp	Stockwell & Peters 1999
GDM	generalized dissimilarity models (including community and single-species implementations, see Elith et al., 2006)	regression	PA*/Ab	Unreleased	Ferrier et al. 2002 Elith et al. 2006
GLM	generalized linear models	regression	PA*/Ab	R (stats), ModEco	Guisan et al. 2002
HABITAT	HABITAT envelope procedure	distance	PO	Unreleased	Walker & Cocks 1991
LIVES	limiting variable and environmental suitability	distance	PO	Unreleased	Li & Hilbert 2008
MARS	multivariate adaptive regression splines (including community and interaction implementations, see Elith et al., 2006)	regression	PA*/Ab	R (mda)	Friedman 1991 Moisen & Frescino 2002
MAXENT	maximum entropy	machine learning	PB	Maxent, R (dismo via maxent.jar), openModeller, ModEco	Phillips et al. 2006 Phillips & Dudk 2008
MD	Mahalanobis distance	distance	PO	R (dismo)	Farber & Kadmon 2003 Rotenberry et al. 2006
PPM	Point process models	regression	PO	R (spatstat, ppmlasso), Maxent	Renner et al. 2015
RF	random forests	machine learning	PA*	R (randomForest), openModeller	Cutler et al. 2007
RS	rough set	machine learning	PA*	ModEco	Pawlak 1991 Guo & Liu 2010
SVM	support vector machines	machine learning	PO (one-class SVM); PA*/Ab (rest)	R (e1071, kernlab), openModeller, ModEco	Guo et al. 2005 Drake et al. 2006

[1] PO = presence-only; PA = presence/absence; PP = presence/pseudo-absence; PB = presence/background; Ab = abundance data; * note that presence/absence methods can be applied to presence-only datasets through the generation of pseudo-absences.

[2] All software in this list is freely available. For methods implemented in R, relevant packages are listed in parentheses.
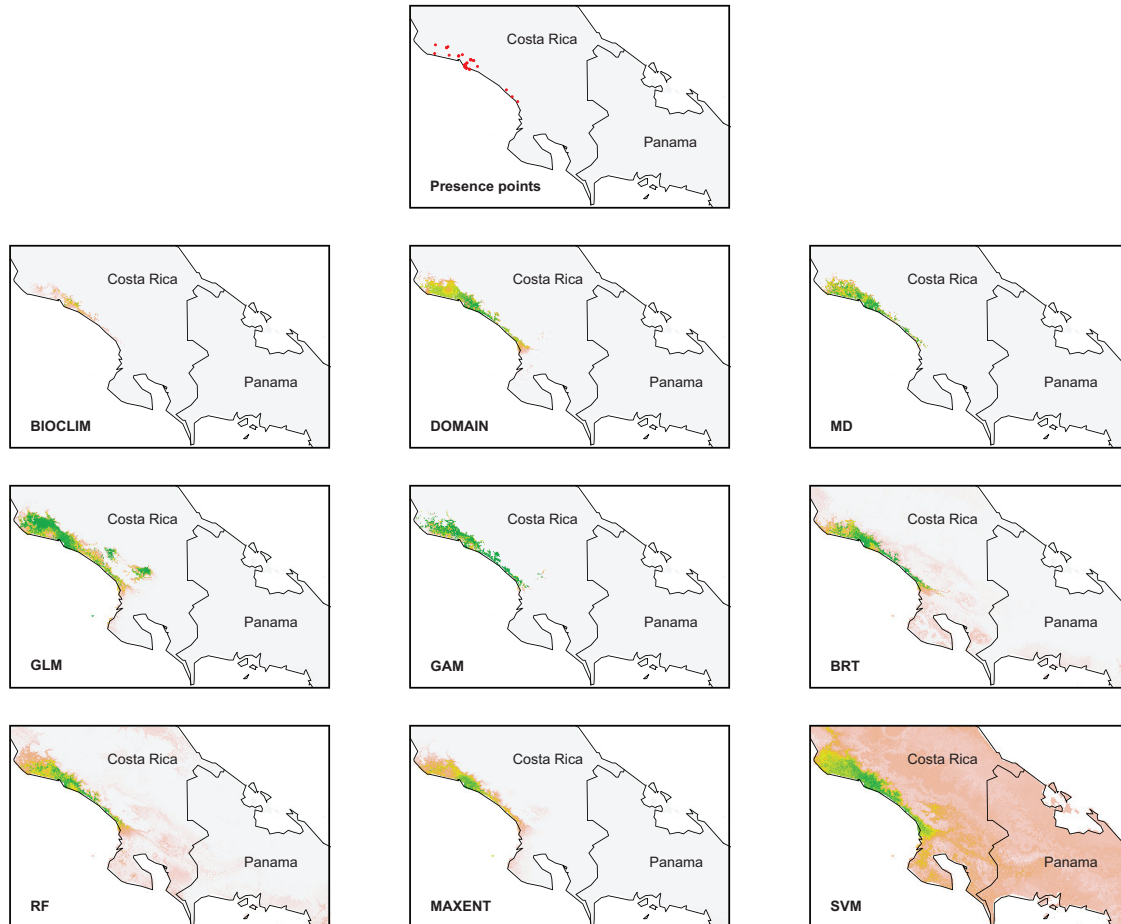
Figure 2: A comparison of model outputs using nine of the ENM algorithms listed in Table 1. Models were trained using presence points for *Saimiri oerstedii citrinellus*, shown in red, that were compiled from several sources. Modeling generally followed procedures described in Box 1. The values mapped represent the model outputs, but their interpretations depend on the method used. For envelope methods, they represent distance values. For other methods, they represent suitability scores, but not necessarily occurrence probabilities. Green pixels represent higher values, but scales are not shown because values are relative.

is a relatively recent GUI application employing a wide range of algorithms. Despite these advantages, however, an active community of users and developers continue to support non-GUI systems. The most prominent of these systems is R (R Core Team, 2013), which is both a programming language and a command-line statistical program. R's use of the command line can be intimidating for novices, but we find that the command line provides greater control over analysis and fosters a greater understanding and appreciation of model settings and parameters. This helps avoid the blunder common to novice GUI users of hitting "go" without changing or even checking the default settings.

A growing number of ENM methods are supported in R through packages such as *dismo* and *biomod2*. Many of these methods, especially those included in the same packages, share many details of their implementation and can therefore be readily run together with other methods, which is very helpful for comparing or combining models (Thuiller et al., 2009). When prepared as scripts, analyses can be generalized and reused such that a series of procedures can be reapplied to evolving or independent datasets, thereby reducing considerably the time and effort needed to set up, adjust, and rerun analyses. Because of its power, flexibility, and popularity in the ENM community, we strongly encourage readers to consider using R for running ENMs. Hijmans & Elith (2013) and Georges & Thuiller (2013) provide excellent guidance for getting started.

3.2   Treatment of presence, absence, and abundance

All ENM methods require *a priori* information on species distributions as part of their approaches. The methods can be classified according to two criteria: (1) whether the input data are treated as continuous measures of abundance or as binary measures of presence/absence and (2) how absence data are provided and interpreted (particularly for presence/absence methods).

Abundance-based methods require prior data not only on *whether* species occur, but also on the number (or relative number) of individuals present. Estimates of relative abundance are virtually impossible to obtain without intensive sampling and seldom cover extensive geographic space due to practical considerations. For primates and other animals, abundance comparisons are further confounded by factors such as range shifts, group size/structure, activity patterns, and detection ability. Abundance methods are best-suited for sampling designs for which abundance information is recorded and the detection probability can be addressed. For primates and other animals, these methods may therefore be effectively employed together with field methods such as camera-trap surveys, capture-recapture sampling, or sight-

resight of animals with individually identifiable features. Abundance methods are also well-suited for modeling plant distributions from plot data (Potts & Elith, 2006) and may therefore have useful applications for assessing primate habitats (see also Rode et al., 2013).

Presence/absence approaches treat their input data as binary (i.e., species are either recorded or not recorded) and are therefore more appropriate for occurrence data lacking information on abundance. An inherent challenge with presence/absence approaches, however, relates to detection probability and the relative reliability of presences compared to absences. Notwithstanding any errors in identification or georeferencing, geographic presences are in general more reliable than absences because of the possibility that "absent" species are in fact present but undetected (e.g., Kry et al., 2010). Several analytical approaches attempt to address this uncertainty by eliminating the need to provide information on absences. Because most data on primate distributions are likely to be geographically biased with relatively small samples, approaches not requiring absence information may be of great utility. While these approaches are sometimes labeled as "presence-only" methods in the literature, this label can be misleading when applied to certain methods that nevertheless incorporate indirect measures of absence into their models (e.g., "background" points in Maxent). In our discussion, we reserve the label "presence-only" for methods not involving any form of absence. Presence-only data, in contrast, refer to data for which absence and abundance information is not available.

The Mahalanobis distance (Farber & Kadmon, 2003) is one illustrative example of a true presence-only method. In this method, environmental values across the study region are positioned in environmental space and a mean vector representing optimal conditions is computed for presence records only. The covariance of the presence sample is also used to compute an ellipsoidal distribution surrounding the data. The suitability of all cells (the Mahalanobis distance) is calculated according to their proximity to this vector in environmental space, scaled by the width of the ellipsoid in the test cell's direction. The Mahalanobis distance thus provides an intuitive scale-invariant means of comparing an unknown sample to a known sample while taking into account covariance in the dataset.

Presence/absence approaches, in contrast, incorporate both presence and absence into their models. These methods include traditional statistical models such as generalized linear models (GLM) and generalized additive models (GAM), both of which use linear regression to fit a main presence-absence pattern. When absence data are not available, presence/absence methods can nevertheless be used with presence-only datasets by randomly generating "pseudo-absences" in place of true absences. The pseudo-absences can be randomly derived from the entire study area (i.e., the back-

12

ground) or from a subset of it, such as all non-presence cells. The software GARP, for instance, randomly generates pseudo-absences from the total study area in lieu of user-defined absences (Stockwell & Peters, 1999). This strategy is similar to presence/background methods (described below) but differs in that the pseudo-absences representing the background are still treated as absences in the building of the model.

Presence/background approaches use presence records together with data from the entire study area without treating any parts of the background as absences. Maxent, for instance, uses sample points and background samples to generate a distribution that maximizes the relative entropy between the probability densities estimated from each (Elith et al., 2011). Ecological niche factor analysis (ENFA), by contrast, uses factor analysis to compare the distribution of ecological values associated with presence records to the distribution of values comprising the full background in multidimensional space (Hirzel et al., 2002). Because the background distribution is based on the defined study area, presence/background methods can be considerably affected by the choice of study extent (Anderson & Raza, 2010).

3.3   Model approach, evaluation, and performance

Every ENM method will generate a prediction when properly run. Understanding how good a model is and how suitable it is for a particular purpose is considerably more difficult (e.g., Lozier et al., 2009). The manner in which a method is implemented dictates how its results should be interpreted. Choosing a model based on its statistical or ecological suitability is an important but difficult task that is covered in more detail elsewhere (Franklin, 2009; Peterson et al., 2011). When comparing models, however, we find it useful to ask whether models make sense geometrically (does the shape of the modeled relationships make sense?), ecologically (is the model sensible given the ecological relationships being modeled?), and spatially (are the model's predictions plausible when mapped out in geographic space?).

The simplest ENM approaches use a distance-based "envelope" to define the boundaries of suitable habitat in multidimensional environmental space. These include the BIOCLIM model (Nix, 1986; Busby, 1991), which defines the envelope as the space delimited by the minimum and maximum values for all presences, the DOMAIN model (Carpenter et al., 1993), which uses a multivariate point-to-point similarity coefficient known as the Gower metric to assign classification values to sites based on the proximity of the most similar record in environmental space, and the Mahalanobis distance (Farber & Kadmon, 2003), which is the standardized difference between the environmental values for any point and the mean values for the same variables across all presence points. Distance methods do not estimate a response

function and cannot tease apart the relative importance of environmental predictors. They also assume that organisms are found in optimal habitats, are well-sampled in environmental space, and that their habitat variables are not dynamic. Despite these assumptions and limitations, distance methods continue to be widely used (e.g., Booth et al., 2014).

A number of ENM approaches use statistical methods such as regression to estimate fitted response functions in their models. These include generalized linear models (GLM), generalized additive models (GAM), and multivariate adaptive regression splines (MARS). GLM is an extension of basic linear regression that uses link functions to fit linear predictors to a flexible range of distributions (e.g., Gaussian, Poisson, binomial, negative binomial, gamma) in the response variable. Through transformations of predictor variables, GLM can be made to accommodate nonlinear relationships between predictors and responses. GAM is a nonparametric (or semi-parametric) extension of GLM that uses smooth functions to fit predictors to complex, nonlinear, and nonmonotonic responses. MARS is similar to GAM in that it makes fewer assumptions about the form of the response function. Unlike GAM, however, MARS fits the response in a stepwise, adaptive manner, resulting in a series of connected linear segments rather than smooth curves in GAM (Friedman, 1991). Unlike distance approaches, regression models can accommodate categorical predictors and can address individually the contributions of environmental variables to habitat suitability due to additivity in the models.

A diverse group of ENM approaches are derived from the field of machine learning and are hence referred to as machine learning methods. There is considerable overlap, however, between the fields of machine learning and inferential statistics and the classification of these ENMs is not without ambiguity (Franklin, 2009). In general, machine learning approaches differ in that they learn the mapping function (or classification rules) inductively from the training data while statistical approaches estimate parameters from the data but require that distributions be set by the user (Breiman, 2001). While machine learning models can require a shift in thinking for researchers accustomed to statistical paradigms, some have well-formed statistical properties that have been dissected for ecological applications (e.g., Elith et al., 2008, 2011). Machine learning methods include decision trees, genetic algorithms, and maximum entropy. We describe these methods below because they are commonly used and have accessible software implementations.

Decision trees, or classification and regression trees (CART), refer to two related algorithms and their outputs: classification trees (CT) and regression trees (RT). Both CT and RT are assembled as recursive binary splits that classify observations based on threshold values of single predictors (De'ath & Fabricius, 2000). In other

14

words, the model can be thought of as a tree-shaped series of rules which at each branching point (node) uses a conditional statement (e.g., annual rainfall < 15 cm) to divide the response into two classes, each of which is relatively homogeneous. Decision trees partition the predictor space using a series of rules to identify areas having the most homogeneous responses. At each node, the responses then take as their values either the majority class (for CT) or the average value (for RT) of the training data. Decision trees are particularly effective at modeling complex (nonlinear and nonadditive) relationships between predictors and responses, including categorical predictors, which are difficult to parameterize using linear models when they have many categories. Decision trees are also very robust at handling missing values and outliers (De'ath & Fabricius, 2000). They do have some drawbacks, however, including poor categorization of linear or smooth species responses, poor categorization of rare classes, especially with limited observations, and potential instability to changing inputs (Franklin, 2009). Ensemble techniques that combine decision trees with algorithms for boosting, such as boosted regression trees (Elith et al., 2008), and bagging, such as random forests (Cutler et al., 2007), have been shown to improve performance by compensating for some of these deficiencies.

Genetic algorithms have been extensively used for niche modeling through the use of the genetic algorithm for rule-set production (GARP) software (Stockwell & Peters, 1999). Like decision trees, genetic algorithm models are expressed in terms of conditional decision rules. Unlike decision trees, however, the rules are generated as part of a population of rules that are "evolved" iteratively through modifications (i.e., "mutation") and evaluation based on ability to predict known cases (i.e., "natural selection"). Through successive iterations, an optimal set of rules emerge. In GARP, the population of rules includes several different kinds of models (e.g., envelope, logistic regression) from which the algorithm chooses the optimal set (Stockwell & Peters, 1999). GARP and other genetic algorithms are stochastic algorithms that produce different results when applied to the same data. A more robust model can be produced by running the algorithm multiple times and summarizing across runs (Anderson et al., 2003).

Maximum entropy is a machine learning model that has been widely applied to ENM research through the Maxent algorithm and software (Phillips et al., 2006; Phillips & Dudk, 2008). It is based on the principle that, given limited information about a set of features, the probability distribution that best models that information is the distribution of maximum entropy (i.e., that is most spread out, or closest to uniform), subject to the constraint that the expected value of each feature match its empirical mean. Applied to ENMs, environmental variables are the features comprising the space upon which the probability distribution is defined and the

values found at presence locations supply the constraints. The maximum entropy distribution is effective at modeling incomplete information because it agrees with everything that is known without assuming anything that is not known (Phillips et al., 2006).

Point process models (PPM) have recently emerged as highly promising tools for presence-only modeling (Renner et al., 2015). By demonstrating close connections with both Maxent and traditional GLM (Renner & Warton, 2013), PPM can claim the performance benefits of Maxent while clarifying important points of statistical interpretation and implementation. Because the target of interest in point process models is intensity, however, point process models may be most appropriate for modeling abundance based on suitably sampled presence data.

Aside from algorithms such as those described in this chapter, a number of methods are available for improving model accuracy, minimizing overfitting, and evaluating model performance. These methods are in general not restricted to particular models, and we cover some of them in the following section on "Implementing ecological niche models." We mention them here, however, because modeling approaches and (particularly) software vary in the availability and implementation of these additional methods and this can be an important consideration when choosing an ENM method.

Finally, another factor to consider is how well methods perform in comparison to one another. A number of studies have systematically compared ENM methods (e.g., Loiselle et al., 2003; Segurado & Arajo, 2004; Elith et al., 2006; Pearson et al., 2006; Wisz et al., 2008; Elith & Graham, 2009) and have consistently demonstrated differences in performance among methods. These studies differ in their focus on different applications, for instance projections into future climate scenarios (Arajo et al., 2005; Pearson et al., 2006), projections into unsampled areas (Peterson et al., 2007), modeling of presence-only data (Elith et al., 2006; Tsoar et al., 2007), or modeling with small sample sizes (Pearson et al., 2007).

Elith et al. (2006) conducted one of the most comprehensive comparative ENM studies to date in terms of the number of models compared (16), global coverage (6 regions across the globe), and number of species (226). For each species, they trained models based on presence-only occurrence records (such as those typical from unplanned surveys or museum collections) and evaluated models based on presence and absence records collected from planned surveys of the same regions. They found that machine learning models such as Maxent and boosted regression trees performed relatively well due to their ability to predict species distributions. A number of publications, however, caution against simplistic interpretations of these comparisons (e.g., Peterson et al., 2011), as different evaluation strategies emphasize different

goals and consequently make different assumptions. Elith and colleagues, for instance, evaluated models based on their predictions of species occupied areas rather than all abiotically suitable areas, the latter of which is more expansive since abiotically suitable areas may nevertheless be unsuitable due to biotic factors. Evaluation based on species occupied areas assumes that species are in equilibrium with their environment and is therefore problematic for applications such as projections into future climate scenarios (Arajo & New, 2007), modeling niche evolution (Warren et al., 2008), or predicting range shifts in non-native species (Peterson, 2003).

Box 1: Case Study: Testing for niche divergence among subspecies of the Central American squirrel monkey

Squirrel monkeys (genus *Saimiri*) occur throughout the Amazon, and also in a disjunct, highly restricted area along the Pacific coast of Central America. A recent biogeographical analysis of the genus suggests that the Central American species, *S. oerstedii*, is a northern ancestral remnant of a population expansion event from the *S. ustus* A group approximately 1.35 and 1 Ma (Lynch Alfaro et al., 2015). The two subspecies of *S. oerstedii* (*S. oerstedii oerstedii* and *S. oerstedii citrinellus*) are distinguished by pelage differences (Hershkovitz, 1984; Rylands & Mittermeier, 2009) and have disjunct geographic distributions on either side of the large Trraba River. Recent genetic studies have lended support for the reciprocal monophyly of these subspecies (Blair et al., 2013a; Lynch Alfaro et al., 2015). The split between the subspecies has been dated to between 0.25 to 0.15 Ma (Lynch Alfaro et al., 2015) or 0.16 to 0.11 Ma (Chiou et al., 2011), consistent with the hypothesis that the subspecies separated during a sea-level rise of ca. 100 m in the Middle-Upper Pleistocene (Nores, 1999; Ford, 2006). *S. oerstedii* are restricted to lowland settings (below 500 m asl) and are therefore constrained to the Central and Southern Pacific coasts by the Central and Talamanca Cordilleras. High water levels in the Pleistocene would thus have resulted in their isolation (Ford, 2006), which is now maintained by the Trraba River.

In this case study, we use ENMs to test hypotheses about ecological niche divergence between the two subspecies of *S. oerstedii*. The timing of genetic divergence coincides with a potential isolation event due to sea-level rise in the Middle-Upper Pleistocene, suggesting an allopatric divergence process for these populations. An allopatric process is essentially a spatial process, which may be facilitated by ecological niche conservatism, where failure to adapt to new envi-

ronments maintains separation between populations (Wiens, 2004). Allopatric divergence process would be most supported by a finding of niche conservatism, or a failure to reject a null model (Losos, 2008; Blair et al., 2013b). By contrast, a finding of niche divergence might suggest a process of ecological divergence, if distributed along an ecotone, or post-divergence ecological specialization.

We sourced occurrence data for *S. o. citrinellus* from prior fieldwork (Blair & Melnick, 2012a, 2012b) and for *S. o. oerstedii* from published (Rodrguez-Vargas, 2003; Solano Rojas, 2007) and unpublished sources (A. Mora & G. Gutirrez-Espeleta, unpubl. data). Presence points were assessed visually for obvious errors. We used 19 bioclimatic variables in the WorldClim dataset (Hijmans et al., 2005) to characterize the environmental background of our study region. We set a window encompassing the study area, then ran a principal components analysis (Peterson et al., 2007) to create a final environmental dataset of 8 independent variables encompassing most (>99%) of the bioclimatic variation. To minimize spatial autocorrelation, we filtered presence records by sampling for environmentally equidistant points (i.e., points that were most distant from other records in environmental space) (Oliveira et al., 2014). The final dataset comprised 23 presence records for *S. o. citrinellus* and 206 presence records for *S. o. oerstedii*. For methods requiring absence information, we generated pseudoabsences randomly from the background for each species with a ratio of 9 pseudoabsences per presence.

Modeling was complicated by differences in spatial bias in the occurrence datasets. Points for *S. o. citrinellus* were few, but reasonably well-distributed and representative of the restricted range of the subspecies. Points for *S. o. oerstedii* were more numerous, but exhibited much more uneven coverage, even following filtering.

For both species, we ran five algorithms to model the ecological niche (Table 2), calculating predictions across the study region for each model. To evaluate the predictions, we cross-validated the models via 5-fold partitioning using two measures of model performance: area under the receiver operating characteristic curve (AUC) and the maximum (i.e., threshold-independent) true skill statistic (Allouche et al., 2006).

As expected, model performance measures were generally higher for *S. o. citrinellus* (Table 2), reflecting the more even sampling design. Because of the biases in our occurrence dataset, we created a more robust model for each species by combining all five models into an ensemble forecast of subspecies distributions, using the minimum training presence as a classification threshold (Figure 3).

18

Table 2: Model evaluation results for two subspecies using five algorithms.

| Subspecies	Method	AUC	TSS (maximum)
S. o. citrinellus	MD	0.9861668	0.9251109
S. o. citrinellus	ANN	0.9779671	0.8973745
S. o. citrinellus	SVM	0.9781707	0.9562147
S. o. citrinellus	BRT	0.9817845	0.8986909
S. o. citrinellus	MAXENT	0.9863144	0.9213667
S. o. oerstedii	MD	0.9777778	0.9625000
S. o. oerstedii	ANN	0.9983673	0.9959184
S. o. oerstedii	SVM	0.9942857	0.9755102
S. o. oerstedii	BRT	0.9979167	0.9875000
S. o. oerstedii	MAXENT	0.9975510	0.9877551
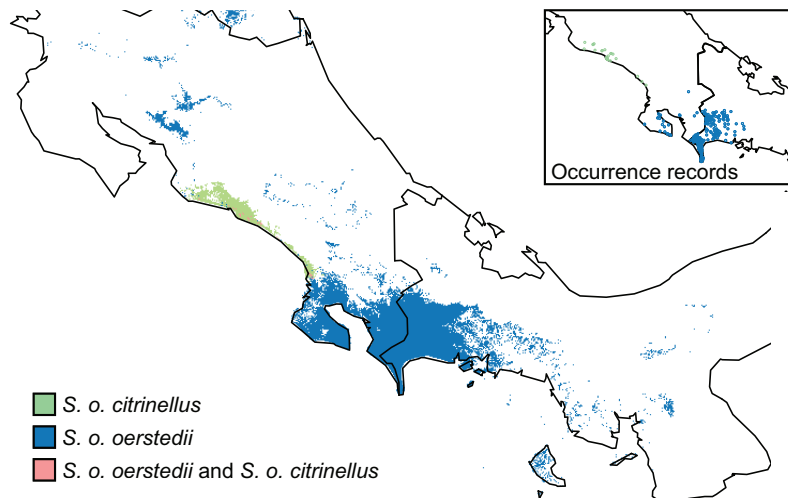
Figure 3: Projected distributions for *Saimiri oerstedii oerstedii* and *Saimiri citrinellus citrinellus* using an ensemble approach.

To test for niche overlap, we calculated similarity statistics (Warren et al., 2008) and conducted background randomization tests (Warren et al., 2010) to assess whether inferred niches were more or less different than expected by chance based on differences in the environmental backgrounds in which they occur. Background areas for each subspecies were defined based on areas falling within 10 km of a known occurrence. Randomization tests in both directions using 100 replicates with Maxent models revealed significant niche divergence for the D similarity statistic ($p < 0.01$, $p < 0.01$), but not the I similarity statistic ($p = 0.33$, $p = 0.33$).

Based on these results, we find support for significant niche divergence under one statistic, but cannot reject the null hypothesis under the other statistic. Given a lack of a clear ecological gradient between the distributions of two subspecies, our results are most consistent with an allopatric divergence scenario. A cautious interpretation is needed, however, given known biases in the sampling of each subspecies. Ultimately, these models illustrate the limitations of inferences based on incomplete or biased sampling. A reasonable conclusion, therefore, is that even the best models may fail when there are unaddressed shortcomings in the data.

4   Ecological Niche Models in Practice

Whereas the previous section focused on the diversity of ENM approaches, in this section we describe the practice of designing and running ENMs. Our review below is largely model-independent but in some cases we concentrate our discussion on Maxent because of its popularity, high-performance, and suitability for presence-only data.

Before beginning the modeling process, the model objectives should be clearly expressed. What is being modeled? Is the focus on modeling niches or distributions? Is the focus on the fundamental niche or the realized niche? The fundamental niche (or potential niche) refers to the set of all environmental states that permit a species to survive while the realized niche refers to the set of all environmental states that permit a species to survive in the presence of competitors or other biotic or movement factors. Whereas the fundamental niche represents the range of theoretical possibilities, the realized niche represents the range that is observable in nature due to real-world circumstances, which may be spatially or temporally specific. While the realized niche is often understood as a subset of the fundamental niche, this is not necessarily the case, for instance in sink habitats (low quality patches where species occupancy is maintained by an influx of individuals from high quality patches) or in cases where there is facilitation among species (see Levi et al., 2013 for a primate example). In general, the fundamental niche is more relevant for assessing potential distributions whereas the realized niche is more relevant for assessing occupied distributions. We recommend that readers consult Sobern & Peterson (2005) and Peterson et al. (2011) for a more thorough discussion.

4.1   Sample data

Sample data provide information relevant for characterizing areas of environmental space as being part of, or not part of, the ecological niche. They virtually always take the form of geographic presence/absence/abundance locations that provide training information for the models through their spatially associated environmental values, which we discuss in the following section.

Absence and abundance data generally take the form of areas, usually grid cells, rather than points. Unlike presence data, absence and abundance areas require considerable survey effort in order to minimize false negatives, i.e., calling species absent when they are in fact present. Rigorous survey data are also beneficial for presence-only data—for instance by minimizing spatial bias or by increasing coverage of areas where species are rare but present—but are frequently not available.

Presence data are often provided as georeferenced points in space, but not always. Presence data can also exist as areas, such as in the case of home range data or distributional data with well-defined boundaries (e.g., islands or national parks). If presence areas are smaller than the resolution of the environmental data, they can be treated as points. If they are larger, however, the environmental values of pixels from across each presence area may be required in order to characterize the occurrence in terms of their mean, range, variance, etc. (Franklin, 2009). Presence data can also be expressed as implicit, but non-georeferenced, points of varying precision (e.g., "in Kaohsiung, Taiwan" or "1.2 km southwest of Fibwe Hide, Kasanka National Park"). These occurrences can be interpreted as areas conservatively encompassing the occurrence point. If these areas are small, like presence areas they can be treated as point data. If they are large, however, additional judgments must be made. In many cases, the data may be deemed unusable.

There are many sources for occurrence data but the most reliable are from field studies. Georeferenced occurrences are often obtained from the literature or from colleagues working in the field. Natural history collections are another major source of data on species distributions and are particularly invaluable as a source of information on historical distributions. These records vary in their accuracy, both geographic and taxonomic. These issues, however, are sometimes documented through metadata and in many cases the primary source material (i.e., "vouchered" specimen") has been retained and can be reexamined. Occurrence data may also be obtained from secondary sources such as scanned distribution maps, although this is seldom advisable due to numerous issues including mapmaking precision and georeferencing errors.

The Internet is a major source of occurrence data, both through scientific and community sources. The Global Biodiversity Information Facility (GBIF) is the

most prominent repository and includes occurrence data from many natural history collections. Other sources are newer and less established, but have the potential to be of great utility for primate ENM studies. GPS collars are being increasingly deployed and their data are beginning to be deposited in repositories such as Movebank (`http://movebank.org`). These data are especially promising for high-resolution local studies. Camera trap surveys are a promising source of occurrence data for rare or elusive species, as are newer methods such as metagenomic DNA surveys (e.g., Calvignac-Spencer et al., 2013). Repositories for these kinds of data do not yet exist to our knowledge, but could become invaluable. Aside from scientific sources, occurrence data are also available through photo-sharing websites such as Flickr (`http://flickr.com`) or citizen science efforts such as Project Noah (`http://projectnoah.org`). We caution, however, that many of the photos on these websites are of captive animals, are not georeferenced, or are georeferenced using imprecise methods that are not necessarily indicated in the metadata (e.g., smartphone location services with wide errors or visual geotagging applied *post hoc*). These websites also contain numerous errors in taxonomic assignment, although these can sometimes be corrected based on the photos themselves. These issues are important and may ultimately turn researchers away from incorporating these data into their projects. We note, however, that the same issues are prevalent in scientific repositories such as GBIF (Yesson et al., 2007; Newbold, 2010) and we remain altogether enthusiastic about the promise of citizen science approaches to ENM research (see also Hochachka et al., 2012).

Creating a good occurrence dataset ultimately requires a balance between the desire for an adequate sample size, the avoidance of positional or taxonomic errors, and the minimization of sample bias. While large samples may seem desirable, some studies (e.g., Pearson et al., 2007) have shown that acceptable models can be built from small samples, especially using certain methods (Stockwell & Peterson, 2002; Elith et al., 2006; Wisz et al., 2008). A recent study (Beck et al., 2013) compared GBIF data to independent compilation efforts for European hawkmoths and found that GBIF contributed less information for niche modeling despite containing many more distribution records. It is therefore better to have small well-distributed samples with low bias/error than to have large samples for which bias/error is prevalent. At least fifty presence observations per species (Kadmon et al., 2003) seems to work well as a rule of thumb, with lower targets for restricted, rare, or elusive species (Pearson et al., 2007). For presence-absence methods, previous studies have suggested targeting a ratio of one absence point for every presence point (e.g., McPherson et al., 2004). This suggested ratio, however, depends on species prevalence and should be higher for rare species (Franklin, 2009). This ratio also does not apply to presence-
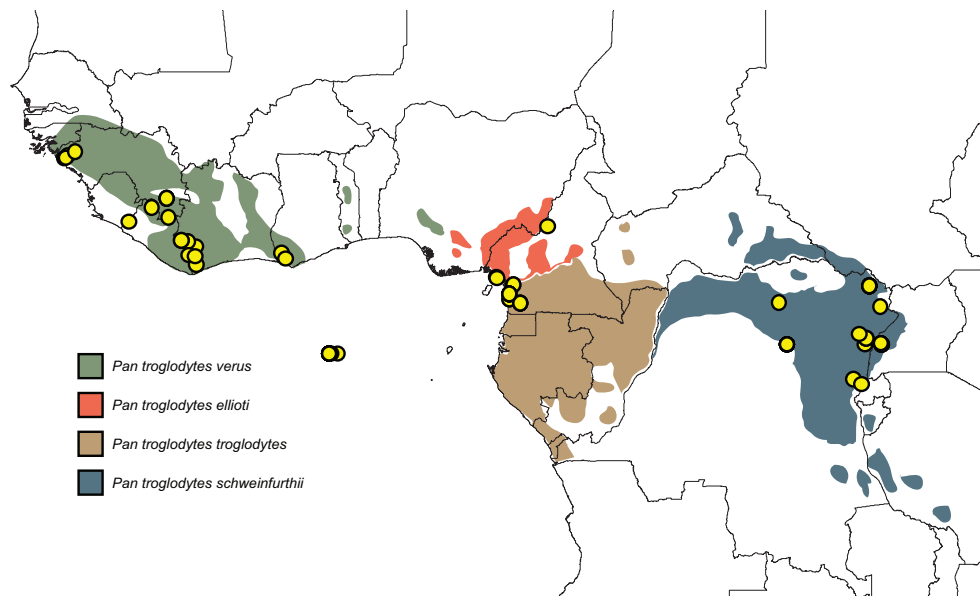
Figure 4: Spatial bias and positional errors in sample locales are evident in occurrence data for common chimpanzees (*Pan troglodytes* spp.) downloaded from the Global Biodiversity Information Facility (`http://gbif.org`; accessed December 15, 2013). In total, 116 records had spatial coordinates. These sample points (yellow circles) are shown superimposed over a chimpanzee distribution map adapted from Bjork et al. (2011). Not shown are six points located on other continents. Notice that some points occur in the middle of the Atlantic Ocean, indicating that errors are present in the dataset (there are no marine chimpanzees). Not all errors, unfortunately, can be so easily identified.

background methods involving randomly generated pseudoabsences, as selection of these pseudoabsences follows other considerations (VanDerWal et al., 2009).

It is less important for the sample size to be large than for it to be representative of the niche distribution. In reality, most ENM datasets are biased in both geographic and environmental space (Figure 3). In a recent literature review, for instance, Yackulic et al. (2013) found that 87% of datasets used in Maxent studies were subject to sample selection bias and that few studies attempted to correct for it. Spatial bias can result from many factors, such as from concentrating survey effort around areas of convenience such as roads or campsites or from skewed representation of areas and taxa that have historically received greater attention. A few methods exist for reducing these biases or for taking them into account. When measures of sampling effort are available, this information can be incorporated into models. In

Maxent, this is accomplished using a "bias file" that specifies the relative sampling effort across cells. When information on sampling effort is not available, the density of occurrence records can be used as a proxy for relative sampling effort (Kramer-Schadt et al., 2013). For datasets with uneven sample density, spatial bias can also be reduced through spatial filtering, i.e., by randomly filtering records from areas with high record density (e.g., Boria et al., 2014; Oliveira et al., 2014). Kramer-Schadt et al. (2013) compared the bias file and spatial filtering approaches and found that both improved model predictions and were therefore preferable to not correcting for biases at all. Of the two methods, they found spatial filtering to be more effective.

4.2   Environmental data

Environmental data are comprised of the features that might influence occupancy across time and space. Stated another way, environmental data constitute the mathematical space upon which ENMs are defined. In the model training stage, they provide the values at sample locations that are used to develop the model, as well as for the background sample locations required for some approaches. In the model prediction stage, they provide the values used to calculate suitability scores across the test region. The environmental data for the training region and test region are often equivalent, but they can differ, such as when models are projected into different regions or time periods.

Environmental variables are ultimately important for ENMs because of the expectation that they are related to aspects of niches and distributions. One goal of variable selection is therefore to identify which variables best provide this information. This is affected to a great extent by the spatial scale of the study. Climatic variables, for instance, are widely acknowledged to be important determinants of thermal, moisture, and light regimes that influence range limits at coarse scales, while factors such as land cover, habitat structure, vegetation phenology, and competitor or predator density are likely to be more relevant at medium and fine scales.

The decision regarding which variables to include in analysis should consider whether the model focuses on the potential or realized niche/distribution. If the focus is on the potential distribution, variables that remain relatively static over large scales and timeframes may be more relevant. Conversely, if the focus is on the occupied distribution, variables that are more dynamic and local may become important.

In the past, one of the biggest barriers for ENM research was the scarcity of suitable environmental datasets. Now digital environmental datasets are quite common and their availability has helped spur the explosion of interest in ENMs in

recent years. The most influential datasets have succeeded due to their high, often global coverage, the standardization of methodology across broad areas, and their ability to provide meaningful information for niche models. This includes the World-clim dataset (Hijmans et al., 2005), which provides high resolution (30 arcseconds, or about 1 km) temperature and precipitation surfaces for land areas covering the entire Earth. These surfaces have been used to derive a set of nineteen bioclimatic variables known as Bioclim variables because they were first developed for early bioclimatic envelope modeling studies using the BIOCLIM method and software (Booth et al., 2014). These nineteen variables are intended to represent annual trends, seasonality, and extremes in temperature and precipitation conditions.

While the availability of these datasets is exciting, it is important that users be mindful of possible quality issues and of the pitfalls of using the data incorrectly (Barry & Elith, 2006). Values from the Worldclim dataset, for example, were interpolated from data collected by weather stations across the world. These weather stations are not distributed evenly across space and in many cases the tropical regions inhabited by most primates have the poorest coverage (Figure 4). Areas with low coverage are not only prone to missing potentially important information on aspects of microclimate, but they are also disproportionately impacted by errors in weather station data. It is therefore advisable to dedicate at least some effort towards understanding how much trust to place in these data for applications. In some cases, data products from downscaled regional climate models might be available or more appropriate for a given area of interest, particularly for areas with considerable topographic complexity.

Aside from climatic variables, variables derived from satellite or aerial imagery are widely used for ENM research. This includes digital elevation models, land cover classifications, and vegetation indices. Vegetation is likely to be of particular importance for primate distributions as both a source of food and as a critical habitat component. Vegetation and other habitat variables can be classified from multispectral or hyperspectral imagery or from data collected in the field (Rode et al., 2013). For these data, however, it can be challenging to describe a meaningful relationship between vegetation classifications and habitat suitability, especially at greater geographic scales (e.g., Chapman et al., 2002). Distributions of other species are also sometimes used as predictors (e.g., Heikkinen et al., 2007). While they suffer from similar scale issues, such predictors may nevertheless provide meaningful information on biotic interactions such as with food resources, competitors, predators, or facilitators.

After determining what variables are available for niche modeling, it is important to ask how many and which variables to include. In some cases, it may be necessary
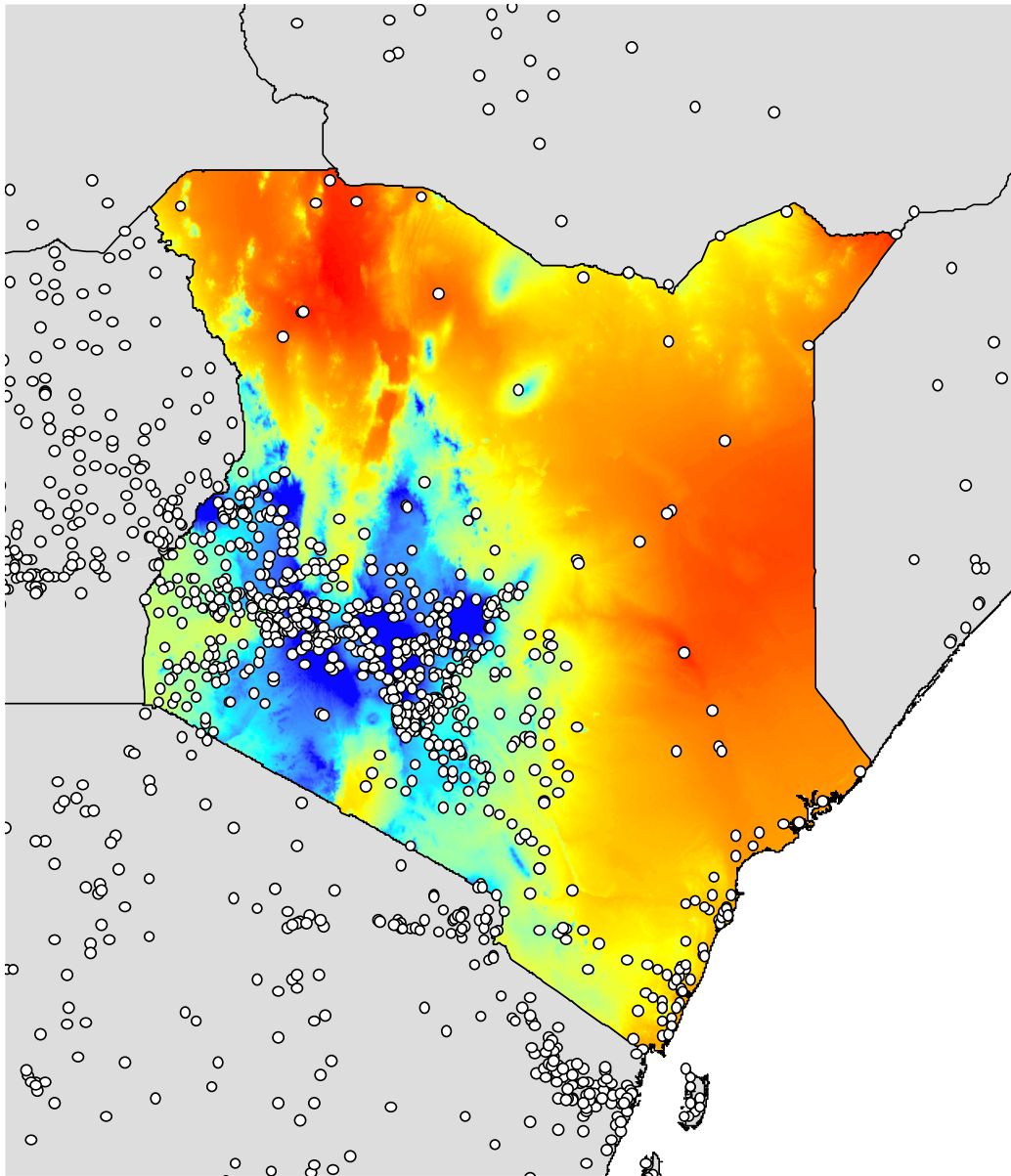
25

Figure 5: An example of low or uneven spatial coverage of weather stations contributing data to interpolated climate datasets. Here, annual mean temperature (BIO1) from the Worldclim dataset (Hijmans et al., 2005) is visualized in Kenya at 30 arcsecond (~1km) resolution along with the locations of weather stations (white circles) contributing raw data to the interpolation. Temperatures are displayed on a red-blue (hot-cold) color scale, with a 2.5 standard deviation stretch function applied in order to increase contrast.

to eliminate variables that do not correspond temporally with other variables or with the sample data. Variables may also prove to be unusable due to spatial issues, such as with incompatible projections, resolutions, or extents, although methods exist for resolving these issues.

Many of the primate ENM studies reviewed in this chapter ran their models using all nineteen Bioclim variables in the Worldclim dataset. This "all but the kitchen sink" approach is understandable given that information on key factors limiting primate species is rarely known. It leads, however, to high degrees of collinearity among variables, which can have adverse effects on statistically based methods (Dormann et al., 2013). Furthermore, inclusion of many variables tends to result in overfit models or a loss of degrees of freedom (Peterson & Nakazawa, 2008). It is generally advisable to select among variables when they correlate. The decision regarding which variables to include, however, can have important impacts on the model. Braunisch et al. (2013) demonstrated this to be the case particularly for projections into future climate change scenarios, where correlative relationships among climate parameters do not necessarily persist. They found that, even when collinearity among variables was maintained across time periods, selecting different variables resulted in different predictions.

When information is available regarding life history and physiological tolerances, the best approach is to use that information to choose an appropriate set of variables (e.g., Blair et al., 2013b). In the absence of *a priori* information, there are some existing methods that have been used to reduce dimensionality in ENMs and to explore relationships among variables (reviewed in Dormann et al., 2013). One approach is to use correlation analysis to identify a set of least-correlated variables, although this approach does not necessarily correct for collinearity (Dormann et al., 2013). Another approach is to use principal components analysis or similar ordination techniques to transform the variables onto new orthogonal axes, then to use the most important dimensions as modeling inputs (e.g., Peterson et al., 2007).

While reducing the number of variables is generally desirable, this can also have adverse effects. Using too few variables can cause the model to miss important factors and to estimate overly broad distributional areas. For some methods, particularly machine-learning methods, collinearity may not be as important of an issue (Guisan & Thuiller, 2005; Elith et al., 2011). One effective strategy might be to run the ENM using many variables with collinear relationships, then to let the model determine which variables contribute the most information relevant to the model. Maxent, for instance, iteratively changes coefficients of environmental features and assigns the increases in the model gain (fit) to the variables the features depend on. These cumulative gains are then used to assess percentage contribution to the models.

27

Additionally, the jackknife test in Maxent measures variable contributions to the models when they are the only variable included in the model and when they are the only variable excluded from the model. This helps shed light on how much relevant information each variable provides as well as how much relevant information each variable *uniquely* provides (i.e., that is not provided by other variables).

Because Maxent and other presence-background (or presence-pseudoabsence) approaches train their models by incorporating a sample of cells taken from the study region, they potentially require additional consideration regarding the extent of the training region used. Anderson & Raza (2010) demonstrated that study regions that were too wide resulted in predictions that were overfit to conditions found near occupied localities. This occurs because wide regions are more likely to include areas that are suitable for species, but that are not occupied due to movement limitations or biotic interactions. In Maxent, "clamping" (i.e., treating variables outside the training range as if they were at the limit of the training range) can be used to help overcome this problem.

Spatial autocorrelation is another important potential issue affecting environmental datasets (Keitt et al., 2002). Spatial autocorrelation results from sample points in space having values that are more similar than expected by chance. When it is positive, points that are nearer exhibit more similar values than those that are far apart. The use of autocorrelated variables may place undue emphasis on environmental factors that may not be important influences on niches and distributions. Several methods have been proposed to correct for this problem (reviewed in Dormann et al., 2007). One method that has been widely adopted corrects for spatial correlation by calculating a measure of autocorrelation, the autocovariate, as a new term and including it as an additional variable in the environmental dataset (Dormann et al., 2007). The procedure of correcting for spatial autocorrelation may be more important for regression approaches than for machine-learning methods. Vclavk et al. (2011) found that while accounting for spatial autocorrelation using multiple methods and geographic scales improved the performance of both a Maxent and GLM model, it reduced spatial autocorrelation in the residuals of only the GLM model.

4.3 Evaluating model performance

Once the model inputs have been determined and the model has been run, several measures exist for assessing performance and significance. Performance is generally assessed based on the ability of the model to predict occurrences in geographical or, less commonly, environmental space. High-performing models should be expected

to minimize commission (false positive) and omission (false negative) errors. Of these two error categories, omission errors are more worrisome because they indicate either model failure or erroneous training data, with a few exceptions (e.g., sink populations). Commission errors are more difficult to interpret due to uncertainty whether negative data reflect true absences or an absence of knowledge.

Evaluation datasets can take the form of fully independent occurrence data, but we anticipate that these will rarely be available for primate research. A more practical approach to evaluation and one of the most widely used is known as k-fold cross-validation, or cross-partitioning. In this evaluation scheme, the occurrences are divided into k random subsets of even size. In successive stages, new models are run using each subset as evaluation data and the remaining subsets as training data. In this manner, a distribution of k estimates of commission/omission error can be estimated.

Estimating commission/omission requires the use of a discrimination threshold (cutoff value) to determine which cells should be judged as suitable or unsuitable habitats for species. Thresholds are also necessary for the mapping of predicted distributions into geographic space. A number of thresholds have been proposed and employed in ENM research (Liu et al., 2005). One intuitive threshold that is well-suited for presence-only data is the "minimum training presence" (MTP). The MTP sets as a threshold the lowest prediction value corresponding with a known presence. A predicted distribution based on the MTP as a threshold can be interpreted as encompassing the areas that are at least as suitable as the least suitable location the species is known to inhabit. It is therefore a conservative measure that identifies the minimum possible area while maintaining zero omission error (Pearson et al., 2007). The MTP, however, can perform poorly if questionable presence observations exist in the dataset or if presence observations are not otherwise representative of suitable habitats. Still more conservative thresholds such as the 10% minimum training presence and the equal training sensitivity and specificity are useful in these cases.

One goal of evaluation, however, might be to assess model performance independent of thresholds. The receiver operating characteristic (ROC) curve is commonly used for this purpose. ROC plots the lack of omission error (true positive rate) on the y-axis against the commission error (false positive rate) on the x-axis. In contrast to threshold-based evaluation measures, ROC calculates numerous omission and commission errors across a range of prediction strengths. The area under the curve (AUC) of the ROC plot represents an overall measure of the performance of the model across a range of thresholds. Its characteristics make it a popular choice as a performance metric (Peterson et al., 2011). Care should be taken to use appropriate settings (e.g., Muscarella et al., 2014), however, as AUC can be inflated in

29

presence-only models.

Evaluation methods and metrics make possible critical assessments of factors influencing model performance and strategies for improving performance. These include topics already discussed such as choosing effective sample sizes, minimizing sample selection bias, correcting for collinearity among environmental variables, and correcting for spatial autocorrelation. Performance can also sometimes be improved by modifying model parameters usually treated as defaults. In Maxent, for example, default parameters were set using average values judged to be optimal using extensive empirical "tuning" (Phillips & Dudk, 2008). Species-specific tuning of parameters shows promise as another method for increasing performance in Maxent, particularly for applications where generality and transferability are desirable (Radosavljevic & Anderson, 2014).

5 Conclusion

A great deal of questions of interest to primatologists can be explored using ENMs, from incorporating climate change into conservation plans to predicting the geographies of hybridization and mixed species associations. Paired with coalescent models or genetic data including next-generation genetic sequencing, ENMs can be used for even more robust inferences about biogeography and population histories. ENMs hold great promise for primatologists, and we hope this chapter will encourage expanded use of the approach amongst our colleagues.

Acknowledgments

References

Aguilera, P. A., Fernndez, A., Reche, F., & Rum, R. (2010). Hybrid Bayesian network classifiers: application to species distribution models. *Environmental Modelling and Software*, *25*, 1630–1639.

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

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

Anderson, R. P., Lew, D., & Peterson, A. T. (2003). Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling*, *162*, 211–232.

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

Arajo, M. B., Pearson, R. G., Thuiller, W., & Erhard, M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology*, *11*, 1504–1513.

Barry, S., & Elith, J. (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology*, *43*, 413–423.

Beck, J., Ballesteros-Mejia, L., Nagel, P., & Kitching, I. J. (2013). Online solutions and the "Wallacean shortfall": what does GBIF contribute to our knowledge of species' ranges?. *Diversity and Distributions*, *19*, 1043–1050.

Beeton, T. A., Glantz, M. M., Trainer, A. K., Temirbekov, S. S., & Reich, R. M. (2013). The fundamental hominin niche in late Pleistocene Central Asia: a preliminary refugium model. *Journal of Biogeography*, *41*, 95–110.

Bett, N. N., Blair, M. E., & Sterling, E. J. (2012). Ecological niche conservatism in doucs (genus *Pygathrix*). *International Journal of Primatology*, *33*, 972–988.

Bjork, A., Liu, W., Wertheim, J. O., Hahn, B. H., & Worobey, M. (2011). Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. *Molecular Biology and Evolution*, *28*, 615–623.

Blair, M. E., & Melnick, D. J. (2012a). Genetic evidence for dispersal by both sexes in the Central American squirrel monkey, *Saimiri oerstedii citrinellus*. *American Journal of Primatology*, *74*, 37–47.

Blair, M. E., & Melnick, D. J. (2012b). Scale-dependent effects of a heterogeneous landscape on genetic differentiation in the central american squirrel monkey (*Saimiri oerstedii*). *PLoS One*, *7*, e43027.

Blair, M. E., Gutirrez-Espeleta, G. A., & Melnick, D. J. (2013a). Subspecies of the Central American squirrel monkey (*Saimiri oerstedii*) as units for conservation. *International Journal of Primatology*, *34*, 86–98.

Blair, M. E., Rose, R. A., Ersts, P. J., Sanderson, E. W., Redford, K. H., Didier, K., Sterling, E. J., & Pearson, R. G. (2012). Incorporating climate change into conservation planning: identifying priority areas across a species' range. *Frontiers of Biogeography*, *4*, 157–167.

Blair, M. E., Sterling, E. J., Dusch, M., Raxworthy, C. J., & Pearson, R. G. (2013b). Ecological divergence and speciation between lemur (*Eulemur*) sister species in Madagascar. *Journal of Evolutionary Biology*, *26*, 1790–1801.

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, *20*, 1–9.

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77.

Boubli, J. P., & Lima, M. G. de. (2009). Modeling the geographical distribution and fundamental niches of *Cacajao* spp. and *Chiropotes israelita* in northwestern Amazonia via a maximum entropy algorithm. *International Journal of Primatology*, *30*, 217–228.

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

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science*, *16*, 199–215.

Brown, J. L., & Yoder, A. D. (2015). Shifting ranges and conservation challenges for lemurs in the face of climate change. *Ecology and Evolution*, *5*, 1131–1142.

Busby, J. R. (1991). BIOCLIM - a bioclimate analysis and prediction system. *Plant Protection Quarterly*, *6*, 8–9.

Calvignac-Spencer, S., Merkel, K., Kutzner, N., Khl, H., Boesch, C., Kappeler, P. M., Metzger, S., Schubert, G., & Leendertz, F. H. (2013). Carrion fly-derived DNA as a tool for comprehensive and cost-effective assessment of mammalian biodiversity. *Molecular Ecology*, *22*, 915–924.

Campos, F. A., & Jack, K. M. (2013). A potential distribution model and conservation plan for the critically endangered Ecuadorian capuchin, *Cebus albifrons aequatorialis*. *International Journal of Primatology*, *34*, 899–916.

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.

Chapman, C. A., Chapman, L. J., & Gillespie, T. R. (2002). Scale issues in the study of primate foraging: red colobus of Kibale National Park. *American Journal of Physical Anthropology*, *117*, 349–363.

Chatterjee, H. J., Tse, J. S. Y., & Turvey, S. T. (2012). Using ecological niche modelling to predict spatial and temporal distribution patterns in Chinese gibbons: lessons from the present and the past. *Folia Primatologica*, *83*, 85–99.

Chiou, K. L., Pozzi, L., Lynch Alfaro, J. W., & Di Fiore, A. (2011). Pleistocene diversification of living squirrel monkeys (*Saimiri* spp.) inferred from complete mitochondrial genome sequences. *Molecular Phylogenetics and Evolution*, *59*, 736–745.

Conroy, G. C., Emerson, C. W., Anemone, R. L., & Townsend, K. E. B. (2012). Let your fingers do the walking: a simple spectral signature model for "remote" fossil prospecting. *Journal of Human Evolution*, *63*, 79–84.

Coudrat, C., & Nekaris, K. A.-I. (2013). Modelling niche differentiation of co-existing, elusive and morphologically similar species: a case study of four macaque species in Nakai-Nam Theun National Protected Area, Laos. *Animals*, *3*, 45–62.

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

Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C., & Mace, G. M. (2011). Beyond predictions: biodiversity conservation in a changing climate. *Science*, *332*, 53–58.

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

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carr, G., Marquz, J. R. G., Gruber, B., Lafourcade, B., Leito, P. J., Mnkemller, T., McClean, C., Osborne, P. E., Reineking, B., Schrder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*, 027–046.

Dormann, C. F., McPherson, J. M., Arajo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Khn, I., Ohlemller, R., Peres-Neto, P. R., Reineking, B., Schrder, B., Schurr, F. M., & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, *30*, 609–628.

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

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., Graham, C. H., Anderson, R. P., Dudk, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Sobern, J., Williams, S., Wisz, M. S., & Zimmermann, N. E. (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., Phillips, S. J., Hastie, T., Dudk, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*, 43–57.

Engler, J. O., Rdder, D., Elle, O., Hochkirch, A., & Secondi, J. (2013). Species distribution models contribute to determine the effect of climate and interspecific interactions in moving hybrid zones. *Journal of Evolutionary Biology*, *26*, 2487–2496.

Farber, O., & Kadmon, R. (2003). Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, *160*, 115–130.

Ferrier, S., Drielsma, M., Manion, G., & Watson, G. (2002). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and Conservation*, *11*, 2309–2338.

Ford, S. M. (2006). The biogeographic history of Mesoamerican primates., in: Estrada, A., Garber, P.A., Pavelka, M., Luecke, L. (Eds.), New Perspectives in the Study of Mesoamerican Primates (pp. 81–114). New York: Springer.

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

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*, 1–67.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*, 131–163.

Georges, D., & Thuiller, W. (2013, July). An example of species distribution modeling with biomod2. R CRAN Project. Tutorial supplied with the 'biomod2' package; R CRAN Project. Tutorial supplied with the 'biomod2' package.

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

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

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

34

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.

Heikkinen, R. K., Luoto, M., Virkkala, R., Pearson, R. G., & Krber, J.-H. (2007). Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography*, *16*, 754–763.

Hershkovitz, P. (1984). Taxonomy of squirrel monkeys genus *Saimiri* (Cebidae, Platyrrhini): a preliminary report with description of a hitherto unnamed form. *American Journal of Primatology*, *7*, 155–210.

Hickey, J. R., Nackoney, J., Nibbelink, N. P., Blake, S., Bonyenge, A., Coxe, S., Dupain, J., Emetshu, M., Furuichi, T., Grossmann, F., Guislain, P., Hart, J., Hashimoto, C., Ikembelo, B., Ilambu, O., Inogwabini, B.-I., Liengola, I., Lokasola, A. L., Lushimba, A., Maisels, F., Masselink, J., Mbenzo, V., Mulavwa, N. M., Naky, P., Ndunda, N. M., Nkumu, P., Omasombo, V., Reinartz, G. E., Rose, R., Sakamaki, T., Strindberg, S., Takemoto, H., Vosper, A., & Khl, H. S. (2013). Human proximity and habitat fragmentation are key drivers of the rangewide bonobo distribution. *Biodiversity and Conservation*, *22*, 3085–3104.

Hijmans, R. J., & Elith, J. (2013, November). Species distribution modeling with R. R CRAN Project. Tutorial supplied with the 'dismo' package; R CRAN Project. Tutorial supplied with the 'dismo' package.

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.

Hirzel, A. H., & Le Lay, G. (2008). Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, *45*, 1372–1381.

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

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W.-K., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution*, *27*, 130–137.

Holzmann, I., Agostini, I., DeMatteo, K., Areta, J. I., Merino, M. L., & Di Bitetti, M. S. (2015). Using species distribution modeling to assess factors that determine the distribution of two parapatric howlers (*Alouatta* spp.) in South America. *International Journal of Primatology*, *36*, 18–32.

Junker, J., Blake, S., Boesch, C., Campbell, G., Toit, L. du, Duvall, C., Ekobo, A., Etoga, G., Galat-Luong, A., Gamys, J., Ganas-Swaray, J., Gatti, S., Ghiurghi, A., Granier, N., Hart, J., Head, J., Herbinger, I., Hicks, T. C., Huijbregts, B., Imong, I. S., Kuempel, N., Lahm, S., Lindsell, J., Maisels, F., McLennan, M., Martinez, L., Morgan, B., Morgan, D., Mulindahabi, F., Mundry, R., N'Goran, K. P., Normand, E., Ntongho, A., Okon, D. T., Petre, C.-A., Plumptre, A., Rainey, H., Regnaut, S., Sanz, C., Stokes, E., Tondossama, A., Tranquilli, S., Sunderland-Groves, J., Walsh, P., Warren, Y., Williamson, E. A., & Kuehl, H. S. (2012). Recent decline in suitable environmental conditions for African great apes. *Diversity and Distributions*, *18*, 1077–1091.

Kadmon, R., Farber, O., & Danin, A. (2003). A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, *13*, 853–867.

Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, *12*, 334–350.

Keitt, T. H., Bjornstad, O. N., Dixon, P. M., & Citron-Pousty, S. (2002). Accounting for spatial pattern when modeling organism-environment interactions. *Ecography*, *25*, 616–625.

Kry, M., Gardner, B., & Monnerat, C. (2010). Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, *37*, 1851–1862.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schrder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A. K., Augeri, D. M., Cheyne, S. M., Hearn, A. J., Ross, J., Macdonald, D. W., Mathai, J., Eaton, J., Marshall, A. J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J. W., Breitenmoser-Wuersten, C., Belant, J. L., Hofer, H., & Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*, 1366–1379.

Kumara, H. N., Irfan-Ullah, M., & Kumar, S. (2009). Mapping potential distribution of slender loris subspecies in peninsular India. *Endangered Species Research*, *7*, 29–38.

Leathwick, J. R., Elith, J., & Hastie, T. (2006a). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling*, *199*, 188–196.

Leathwick, J. R., Elith, J., Francis, M. P., Hastie, T., & Taylor, P. (2006b). Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, *321*, 267–281.

Lehmann, J., Korstjens, A. H., & Dunbar, R. I. M. (2010). Apes in a changing world - the effects of global warming on the behaviour and distribution of African apes. *Journal of Biogeography*, *37*, 2217–2231.

Levi, T., Silvius, K. M., Oliveira, L. F. B., Cummings, A. R., & Fragoso, J. M. V. (2013). Competition and facilitation in the capuchin-squirrel monkey relationship. *Biotropica*, *45*, 636–643.

Li, J., & Hilbert, D. W. (2008). LIVES: a new habitat modelling technique for predicting the distribution of species' occurrences using presence-only data based on limiting factor theory. *Biodiversity and Conservation*, *17*, 3079–3095.

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.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K. G., & Williams, P. H. (2003). Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology*, *17*, 1591–1600.

Losos, J. B. (2008). Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology Letters*, *11*, 995–1003.

Lozier, J. D., Aniello, P., & Hickerson, M. J. (2009). Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. *Journal of Biogeography*, *36*, 1623–1627.

Luo, Z., Zhou, S., Yu, W., Yu, H., Yang, J., Tian, Y., Zhao, M., & Wu, H. (2014). Impacts of climate change on the distribution of Sichuan snub-nosed monkeys (*Rhinopithecus roxellana*) in Shennongjia area, China. *American Journal of Primatology*, *77*, 135–151.

Lynch Alfaro, J. W., Boubli, J. P., Paim, F. P., Ribas, C. C., Nazareth, M., Messias, M., Rhe, F., Mercs, M. P., Sousa e Silva, J. de, Jr, Silva, C. R., Pinho, G. M., Koshkarian, G., Nguyen, M. T. T., Harada, M. L., Rabelo, R. M., Queiroz, H. L., Alfaro, M. E., & Farias, I. (2015). Biogeography of squirrel monkeys (genus *Saimiri*): south-central Amazon origin and pan-Amazonian diversification of a lowland primate. *Molecular Phylogenetics and Evolution*, *82B*, 436–454.

MacColl, A. D. C. (2011). The ecological causes of evolution. *Trends in Ecology & Evolution*, *26*, 514–522.

Manel, S., Dias, J.-M., & Ormerod, S. J. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, *120*, 337–347.

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.

Meyer, A. L. S., Pie, M. R., & Passos, F. C. (2014). Assessing the exposure of lion tamarins (*Leontopithecus* spp.) to future climate change. *American Journal of Primatology*, *76*, 551–562.

Mitchell, M. W., Locatelli, S., Sesink Clee, P. R., Thomassen, H. A., & Gonder, M. (2015). Environmental variation and rivers govern the structure of chimpanzee genetic diversity in a biodiversity hotspot. *BMC Evolutionary Biology*, *15*, 1.

Moisen, G. G., & Frescino, T. S. (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling*, *157*, 209–225.

Morales-Jimenez, A. L., Nekaris, A., Lee, J., & Thompson, S. (2005). Modeling distributions for Colombian spider monkeys (*Ateles* spp.) to find priorities for conservation. *American Journal of Primatology*, *66 (suppl. 1)*, 131.

Muscarella, R., Galante, P. J., Soley-Guardia, M., Boria, R. A., Kass, J. M., Uriarte, M., & Anderson, R. P. (2014). ENMeval: an R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods in Ecology and Evolution*, *5*, 1198–1205.

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, *34*, 3–22.

Nix, H. A. (1986). A biogeographic analysis of Australian elapid snakes., in: Longmore, R. (Ed.), Atlas of Elapid Snakes of Australia (pp. 4–15). Canberra, Australia: Australian Government Publishing Service.

Nores, M. (1999). An alternative hypothesis for the origin of Amazonian bird diversity. *Journal of Biogeography*, *26*, 475–485.

Oliveira, G. de, Rangel, T. F., Lima-Ribeiro, M. S., Terribile, L. C., & Diniz-Filho, J. A. F. (2014). Evaluating, partitioning, and mapping the spatial autocorrelation component in ecological niche modeling: a new approach based on environmentally equidistant records. *Ecography*, *37*, 637–647.

Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning about Data*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful?. *Global Ecology and Biogeography*, *12*, 361–371.

Pearson, R. G., Dawson, T. P., Berry, P. M., & Harrison, P. A. (2002). SPECIES: a spatial evaluation of climate impact on the envelope of species. *Ecological Modelling*, *154*, 289–300.

Pearson, R. G., Raxworthy, C. J., Nakamura, M., & Peterson, A. T. (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, *34*, 102–117.

Pearson, R. G., Thuiller, W., Arajo, M. B., Martnez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T. P., & Lees, D. C. (2006). Model-based uncertainty in species range prediction. *Journal of Biogeography*, *33*, 1704–1711.

Peck, M., Thorn, J., Mariscal, A., Baird, A., Tirira, D., & Kniveton, D. (2011). Focusing conservation efforts for the critically endangered brown-headed spider monkey (*Ateles fusciceps*) using remote sensing, modeling, and playback survey methods. *International Journal of Primatology*, *32*, 134–148.

Peterson, A. T. (2003). Predicting the geography of species' invasions via ecological niche modeling. *The Quarterly Review of Biology*, *78*, 419–433.

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

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

Peterson, A. T., Sobern, J., Pearson, R. G., Anderson, R. P., Martnez-Meyer, E., Nakamura, M., & Arajo, M. B. (2011). *Ecological Niches and Geographic Distributions*. Princeton, NJ: Princeton University Press.

Phillips, S. J. (2009). A brief tutorial on MaxEnt.

Phillips, S. J., & Dudk, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, *31*, 161–175.

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

Pintea, L., Jantz, S., Nackoney, J. R., & Hansen, M. C. (2014). The first high resolution maps of chimpanzee habitat health in Africa., in: IUCN World Parks Congress. Sydney, Australia.

Potts, J. M., & Elith, J. (2006). Comparing species abundance models. *Ecological Modelling*, *199*, 153–163.

R Core Team. (2013). R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org; R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org.

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, *41*, 629–643.

Raxworthy, C. J., Ingram, C. M., Rabibisoa, N., & Pearson, R. G. (2007). Applications of ecological niche modeling for species delimitation: a review and empirical evaluation using day geckos (*Phelsuma*) from Madagascar. *Systematic Biology*, *56*, 907–923.

Renner, I. W., & Warton, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, *69*, 274–281.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., & Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*, 366–379.

Rode, E. J., Stengel, C. J., & Nekaris, K. A.-I. (2013). Habitat assessment and species niche modeling., in: Sterling, E.J., Bynum, N., Blair, M.E. (Eds.), Primate Ecology and Conservation: A Handbook of Techniques (pp. 79–102). Oxford: Oxford University Press.

Rodrguez-Vargas, A. R. (2003). Analysis of the hypothetical population structure of the squirrel monkey (*Saimiri oerstedii*) in Panam., in: Marsh, L.K. (Ed.), Primates in Fragments: Ecology and Conservation (pp. 53–62). New York: Kluwer Academic Publishers.

Rotenberry, J. T., Preston, K. L., & Knick, S. T. (2006). GIS-based niche modeling for mapping species' habitat. *Ecology*, *87*, 1458–1464.

Rouget, M., Richardson, D. M., Lavorel, S., Vayreda, J., Gracia, C., & Milton, S. J. (2001). Determinants of distribution of six *Pinus* species in Catalonia, Spain. *Journal of Vegetation Science*, *12*, 491–502.

Rylands, A. B., & Mittermeier, R. A. (2009). The diversity of the New World primates (Platyrrhini): an annotated taxonomy., in: Garber, P.A., Estrada, A., Bicca-Marques, J.C., Heymann, E.W., Strier, K.B. (Eds.), South American Primates. New York, NY: Springer.

Scachetti-Pereira, R. (2002). DesktopGarp v1.1.6. Available at http://www.nhm.ku.edu/desktopgarp/.

Scott, J. M., Csuti, B., Jacobi, J. D., & Estes, J. E. (1987). Species richness: a geographic approach to protecting future biological diversity. *BioScience*, *37*, 782–788.

Segurado, P., & Arajo, M. B. (2004). An evaluation of methods for modelling species distributions. *Journal of Biogeography*, *31*, 1555–1568.

Sesink Clee, P. R., Abwe, E. E., Ambahe, R. D., Anthony, N. M., Fotso, R., Locatelli, S., Maisels, F., Mitchell, M. W., Morgan, B. J., Pokempner, A. A., & Gonder, M. K. (2015). Chimpanzee population structure in Cameroon and Nigeria is associated with habitat variation that may be lost under climate change. *BMC Evolutionary Biology*, *15*, 2.

Sobern, J., & Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, *2*, 1–10.

Solano Rojas, D. (2007). *Evaluacin del hbitat, el paisaje y la poblacin del mono tit (Cebidae, Platyrrhini:* Saimiri oerstedii oerstedii*) en la Pennsula de Osa, Costa Rica* (Master's thesis). Universidad Nacional de Costa Rica; Universidad Nacional de Costa Rica, Heredia, Costa Rica.

Sterling, E. J., Bynum, N., & Blair, M. E. (2013). Conclusion: the future of studying primates in a changing world., in: Sterling, E.J., Bynum, N., Blair, M.E. (Eds.), Primate Ecology and Conservation: A Handbook of Techniques (pp. 346–350). Oxford: Oxford University Press.

Stockwell, D. R. B., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, *148*, 1–13.

Stockwell, D., & Peters, D. (1999). The GARP modeling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, *13*, 143–158.

Struebig, M. J., Fischer, M., Gaveau, D. L. A., Meijaard, E., Wich, S. A., Gonner, C., Sykes, R., Wilting, A., & Kramer-Schadt, S. (2015). Anticipated climate and landcover changes reveal refuge areas for Borneo's orangutans. *Global Change Biology*, *21*, 2891–2904.

Thorn, J. S., Nijman, V., Smith, D., & Nekaris, K. A.-I. (2009). Ecological niche modelling as a technique for assessing threats and setting conservation priorities for Asian slow lorises (Primates: *Nycticebus*). *Diversity and Distributions*, *15*, 289–298.

Thuiller, W. (2003). BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, *9*, 1353–1362.

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

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, *13*, 397–405.

VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know?. *Ecological Modelling*, *220*, 589–594.

Vclavk, T., Kupfer, J. A., & Meentemeyer, R. K. (2011). Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). *Journal of Biogeography*, *39*, 42–55.

Vidal-Garca, F., & Serio-Silva, J. C. (2011). Potential distribution of Mexican primates: modeling the ecological niche with the maximum entropy algorithm. *Primates*, *52*, 261–270.

Voskamp, A., Rode, E. J., Coudrat, C. N. Z., Wirdateti, Abinawanto, Wilson, R. J., & Nekaris, K. A.-I. (2014). Modelling the habitat use and distribution of the threatened Javan slow loris *Nycticebus javanicus*. *Endangered Species Research*, *23*, 277–286.

Vu, M. V., Thach, H. M., & Pham, V. T. (2010). Using environmental niche model to study the distribution of Tonkin snub-nosed monkey (*Rhinopithecus avunculus*) in Northeastern Vietnam under some climate change scenarios. *Proceedings of the 24th International Conference on Informatics for Environmental Protection, Cologne/Bonn, Germany*, 156–164.

Vu, M. V., Thach, H. M., Le, M. T. T., & Pham, V. T. (2011). Study on the using of environmental niche model Bioclim to estimate the distribution of Francois's Langur (*Trachypithecus francoisi*) in Northern of Vietnam under climate change of IPCC scenario A_2. *VNU Journal of Science: Natural Sciences & Technology*, *27*, 70–76.

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.

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

Warren, D. L., Glor, R. E., & Turelli, M. (2010). ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography*, *33*, 607–611.

Wiens, J. J. (2004). Speciation and ecology revisited: phylogenetic niche conservatism and the origin of species. *Evolution*, *58*, 193–197.

Willems, E. P., & Hill, R. A. (2009). A critical assessment of two species distribution models: a case study of the vervet monkey (*Cercopithecus aethiops*). *Journal of Biogeography*, *36*, 2300–2312.

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

Wong, M. H. G., Li, R., Xu, M., & Long, Y. (2013). An integrative approach to assessing the potential impacts of climate change on the Yunnan snub-nosed monkey. *Biological Conservation*, *158*, 401–409.

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., & Veran, S. (2013). Presence-only modelling using MAXENT: when can we trust the inferences?. *Methods in Ecology and Evolution*, *4*, 236–243.

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A., & Culham, A. (2007). How global is the Global Biodiversity Information Facility?. *PLoS One*, *2*, e1124.