Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions

Robert P. Anderson[1,2,3,4]

[1]Department of Biology, City College, [2]The Graduate Center, and [3]CREST Institute, The City University of New York, New York, New York. [4]Division of Vertebrate Zoology (Mammalogy), American Museum of Natural History, New York, New York

Address for correspondence: Robert P. Anderson, 526 Marshak Science Building, City College of the City University of New York, 160 Convent Avenue, New York, NY 10031. anderson@sci.ccny.cuny.edu

Recent advances allow harnessing enormous stores of biological and environmental data to model species niches and geographic distributions. Natural history museums hold specimens that represent the only information available for most species. Ecological niche models (sometimes termed *species distribution models*) combine such information with digital environmental data (especially climatic) to offer key insights for conservation biology, management of invasive species, zoonotic human diseases, and other pressing environmental problems. Five major pitfalls seriously hinder such research, especially for cross-space or cross-time uses: (1) incorrect taxonomic identifications; (2) lacking or inadequate databasing and georeferences; (3) effects of sampling bias across geography; (4) violation of assumptions related to selection of the study region; and (5) problems regarding model evaluation to identify optimal model complexity. Large-scale initiatives regarding data availability and quality, technological development, and capacity building should allow high-quality modeling on a scale commensurate with the enormous potential of and need for these techniques.

Keywords: biodiversity; climate change; ecological niche modeling; georeference; natural history museum; species distribution

Introduction

The relevance of museum data and ecological niche modeling to society

The vast majority of species on Earth remain known only from specimens housed in the research collections of natural history museums and herbaria (hereafter, "museums"), creating the need to infer from such information in order to estimate and characterize the planet's biodiversity.[1–6] The range, or geographic distribution, of a species constitutes one fundamental dimension of biodiversity—for many species, the only one feasible for study. Developed largely over the past two decades, computer techniques often termed *ecological niche modeling* (or *species distributional modeling*; see later) can harness information from museum specimens to model a species' environmental requirements and identify geographic areas suitable for it.[7] To do so, they use occurrence records of the species in conjunction with digital environmental data (especially regarding climate), typically interfacing with computer mapping software termed geographic information systems (GIS). Despite the tremendous potential of such techniques and the explosion in their use over the past decade,[8] several factors have conspired to impede realization of their full potential. Based on my fieldwork and research in taxonomy, systematics, and ecological niche modeling, I offer this perspective regarding the enormous promise for, and substantial pitfalls involved in, transforming occurrence data from museums into niche models and their associated distributional predictions.

Here, I consider only modeling based on what are termed *presence-only* occurrence data, the principal data source available for studying species distributions (Table 1). Presence-only data, such as what can be gleaned directly from museum specimens, constitute records of places where a species has been documented—without any information regarding the species' abundance (abundance data) or indicating places where it does not occur (presence–absence data). For the most part, those latter data types exist only for very well-studied taxonomic groups in temperate areas of the world, especially Europe and North America. Hence, for most species,

Table 1. Biodiversity information (definitions here and in other tables largely follow those of a recent theoretical and methodological treatment[7])

Term	Definition
Natural history museum	A museum with research collections and scientists studying biodiversity, often focused on animals. Frequently also contains extensive public exhibits.
Herbarium	A scientific institution with research collections and scientists studying biodiversity, focused on plants. Frequently also contains botanical gardens of living plants.
Taxonomy	The science of documenting and describing biological diversity, typically at the level of the species ("alpha taxonomy," including naming new species). *Revisionary taxonomy* refers to efforts to study a particular group (often a genus) in order to determine what species exist, describe any new to science, and characterize all of them (typically based on morphology, although increasingly leveraging molecular data as well).
Systematics	Often considered to include taxonomy, the science of characterizing biological diversity, including the evolutionary relationships among species and classifying them accordingly (typically via the Linnaean system, usually based on the results of phylogenetic analyses reconstructing evolutionary relationships, most commonly using molecular data).
GIS	Computer software that allows visualization of and calculations regarding spatial data.
Presence-only data	Data sets containing records of where a species has been observed to be present, but lacking any information regarding sites where it is absent.
Presence-background data	Data sets containing records of where a species has been observed to be present, as well as information regarding environmental variation across the study area (the "background") and whether or not sampling has occurred there (and if so, whether or not records of the species exist from those regions).
Presence-pseudoabsence data	Data sets containing records of where a species has been observed to be present, as well as sites where it has not been observed (but note that the species may actually inhabit these latter sites, which can lack records of it due to nonexistent or inadequate sampling).
Presence-absence data	Data sets containing records of where a species has been observed to be present, as well as sites where it is absent, or assumed to be, despite sampling efforts (but note that the species may actually inhabit these latter sites, if sampling is present but inadequate).
Abundance data	Data sets containing information regarding the abundance of a species at various sites.

and especially for poorly studied tropical regions of high biodiversity, presence-only data sets constitute the only information available.[4] Lacking absence data, most niche-modeling algorithms using presence-only occurrence data compare the environmental conditions of the sites that the species is known to inhabit with those in a sample of the "background" available in the study region.[9,a] Although other sources also can provide presence-only occurrence data, I focus on data from museums, which hold the only information available for the vast majority of species. I do not attempt a full summary of the field or its uses[8,10] or a theoretical synthesis (a void largely filled by a recent book,[7] whose terminology I follow). Rather, I aim to present this research area and its great utility to nonspecialists, point out the most overreaching and consistent issues that limit its use, and advocate for an updated vision regarding the mutualistic, but still vastly underdeveloped, relationship between museums and ecological niche modeling.[11]

Niche models based on presence-only occurrence data hold enormous utility in basic and applied biodiversity science, yet even the tremendous increase in the use of such techniques over the past decade pales in comparison with their staggering untapped potential (Table 2). In addition to exciting academic uses in biogeographic, ecological, and evolutionary studies, niche models hold great applied relevance—for example, in conservation biology, the mitigation of invasive species, and public health concerns related to zoonotic human diseases. The relevance of climatic change and other anthropogenic environmental alterations cuts across each of these areas. Largely because of society's need to forecast the future effects of our actions, it is the ability of niche models to predict after climatic change that has attracted the most attention to the field.[12,13] The power of these models to predict suitable conditions in other places and time periods derives from their niche-based nature. Instead of directly delineating in geography the places where a species occurs (its range, or occupied distributional area), they aim to characterize the environmental conditions suitable for the species (its niche; or more precisely, its ex-

isting fundamental Grinnellian niche). Subject to certain critical assumptions,[7] such a niche model then allows identification of the geographic areas that fulfill those requirements (its abiotically suitable area, termed the *potential distribution* by many workers) in the region or time period of interest.[b]

---

[b]The niche represents a central, unifying concept in ecology, albeit a complex one with varying perspectives and definitions presented over the past decades.[14–16] To place the ecological niche models (Table 2) discussed here in context, I follow the rearrangement of niche concepts explained at length in a recent theoretical and methodological treatment.[7] Although the division of niche concepts into complementary Grinnellian and Eltonian perspectives (that arguably constitute two ends of a continuum) espoused therein requires simplifying assumptions, this separation represents a highly useful distinction defensible in many circumstances. Under this framework, niche models using presence-only occurrence records (and typically a background sample of the study region) are based on environmental variables not affected by the presence of the species (termed *scenopoetic variables*). Generally most relevant and measured at coarse grains, such variables characterize the species' Grinnellian niche, delimiting density-independent factors that permit positive population growth rates for the species. Conveniently, this simplified perspective of the niche can be modeled as static sets of numbers. Such models differ from those of what has been termed the *Eltonian niche*[7] (detailed in a comprehensive treatment[14]), which considers variables modified by the presence of the species. Often relevant and measured at a fine spatial grain and small, local extent, these variables appear as density-dependent terms in population growth equations, requiring substantially more complicated mathematical formulations. Consideration of the Grinnellian niche via static sets of environmental conditions allows for definition of subsets of it that correspond to important biological situations, inspired by the long-standing concepts of fundamental and realized components of the niche (parallel to that distinction relevant for Eltonian niches[14]). For example, the full scenopoetic Grinnellian fundamental niche of a species likely contains various smaller subsets that represent conditions that can be termed the: existing fundamental niche, biotically reduced niche, invadable niche, and occupied niche. Researchers should consider such realities affecting the occurrence data for a species when planning a study using ecological niche modeling and when interpreting the resulting predictions. For example, dispersal limitations, biotic interactions, and even the limited set of environmental conditions existing on Earth today may cause a species to inhabit less than its full fundamental Grinnellian niche, violating assumptions of modeling.[7,17–19]

---

[a]Alternatively, researchers sometimes use a "pseudoabsence" sample taken from all pixels lacking a record of the species.

Table 2. Ecological niche modeling and species distribution modeling

Ecological niche modeling	Modeling of the [existing fundamental] niche of a species in environmental space with the intent of estimating the areas in geographic space holding suitable conditions for it [= the abiotically suitable area; termed the potential distribution by many workers], whether or not the species truly occupies those areas (Figs. 1 and 2). In practice (and subject to critical and clearly stated assumptions[7]), this is typically carried out by modeling the [abiotically] suitable conditions [= existing fundamental niche] for the species using: (1) records of its presence; and (2) [scenopoetic] environmental variables not affected by the presence of the species; these data sets are assembled via selection of a study region matching the assumptions of niche modeling; see Pitfall 4. The niche model is then applied to geography to identify the [abiotically] suitable area for the species. When desired, the prediction of the [abiotically] suitable area can be processed subsequently to estimate other distributional areas for the species—for example, and often of special interest, its range [= occupied distributional area]—by considering information regarding dispersal and biotic interactions. Niche models hold predictive ability across space and time.
Species distribution modeling	Modeling with the intent of estimating the range [= occupied distributional area] of a species in geographic space directly, without first passing through an estimate of its [existing fundamental] niche in environmental space or the corresponding [abiotically] suitable area in geographic space. In practice, this is typically carried out by modeling the range [= occupied distributional area] of the species using: (1) records of the presence of a species, and sometimes also information regarding sites where it is absent; and (2) various kinds of variables, including [scenopoetic] environmental variables not affected by the presence of the species, and/or spatial variables (e.g., latitude and longitude); these data sets are assembled via selection of a study region matching the assumptions of species distribution modeling and likely violating some of those of niche modeling (see Pitfall 4), resulting in models that include the effects of correlations of the two data sets with dispersal limitation and biotic interactions. The species distribution model is then applied to geography to identify the range [= occupied distributional area] of the species. The prediction of the range [= occupied distributional area] cannot be processed subsequently to estimate the [abiotically] suitable area for the species. Species distribution models are not designed to hold predictive ability across space or time.

Note: Much confusion exists in the literature regarding the terms *ecological niche modeling* and *species distribution modeling*. Although some workers do not recognize a difference between the two (generally using the term species distribution modeling for applications that are clearly niche-based), others maintain a conceptual distinction, requiring two different terms for clarity.[7] Following the latter perspective, I offer the definitions below to provide unambiguous meaning for my usage of "ecological niche modeling" of Grinnellian niches (footnote *b*). Following these definitions, many (but not all) recent studies conducting what the respective authors termed "species distribution modeling" would be recategorized as having carried out "ecological niche modeling," at least in intent. Nonspecialists may derive more clarity from reading without inclusion of the terms supplied in brackets for the specialist.

When necessary in a particular study (which is often the case), this prediction of suitable areas then can be processed to take into account dispersal barriers and the distributions of key biotic interactors (e.g., competitors, parasites, or mutualists) to estimate the species' occupied distributional area (or range, sometimes termed the realized distribution;[19,20] Figs. 1 and 2). In contrast to niche models,
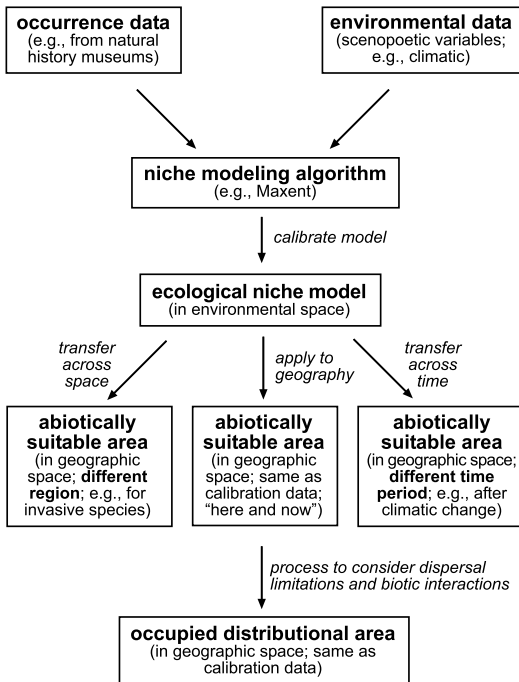
Figure 1. Diagram showing various steps involved in ecological niche modeling. Note the following sequential steps: data acquisition; calibration of the ecological niche model (in environmental space); application to geographic space (including possible transferral to other geographic regions or time periods); and processing to consider dispersal limitations and biotic interactions.[7]

approaches that estimate the species' distribution directly do not aim to predict in other places or time periods.

Most scientists implicitly or explicitly acknowledge that all models of nature remain imperfect— but that some (like, hopefully, ecological niche models) constitute quite useful approximations. Whereas the consequences of errors, or of mere inadequacies, of niche models for an academic (e.g., biogeographic) study may not incur large negative consequences for society, erroneous models for applied uses can lead to grave repercussions—for example, an endangered species poorly managed, resources for combating an invasive species misallocated, or public health policies gone awry. A chilling example comes from the related zoonotic diseases lymphatic filariasis and loa loa in West Africa; effective prophylaxis protects against filariasis, but its administration to a person infected with loa loa can be fatal.[21,22] Hence, niche models, applied to geog-

raphy and then processed to take into account dispersal barriers and relevant biotic interactors, must identify areas where filiariasis is transmitted but loa loa is not. As concluded recently[7] (p. 236), "such situations, in which human lives depend on being 'right,' should give pause to the prospective modeler." Because of these applications of such high importance to society, modelers and those associated with museums have the justification for making ecological niche models match the reality of nature as best as possible.

Five major pitfalls: data and methodology

So why does the full utility of ecological niche modeling remain underrealized? The answer lies in both the incomplete availability of the necessary occurrence data and the nascent nature of the field, with relatively few researchers well trained conceptually and methodologically. Over a decade ago, the field of biodiversity informatics reached a consensus regarding the plan of action for making high-quality museum data available over the internet and developed the requisite technology (e.g., via the pioneering Mammal Networked Information System, MaNIS[23]). However, despite impressive progress, implementation of that vision remains vastly incomplete.[3,24] In addition, researchers often implement the techniques poorly, usually because of lack of knowledge (and a paucity of clarity and consensus in the literature). I highlight five particular pitfalls that limit the effective use of ecological niche modeling, expanding upon each in turn and noting their interrelatedness: (1) incorrect taxonomic identifications; (2) lacking or inadequate databasing and georeferences (coordinates of latitude and longitude); (3) effects of sampling bias across geography; (4) selection of the study region; and (5) model evaluation to identify optimal model complexity (Table 3). The first three concern data quality and availability, where I draw attention to the increased necessity for (and value of) high-quality data, in light of recent progress in ecological niche modeling. In contrast, the latter two represent conceptual and methodological issues in niche modeling. Whereas other works expand on various such topics in much greater detail,[7,25] I highlight these two because of their cross-cutting nature and the strong associated repercussions when transferring a model to another place or time.[13] Other important areas, especially
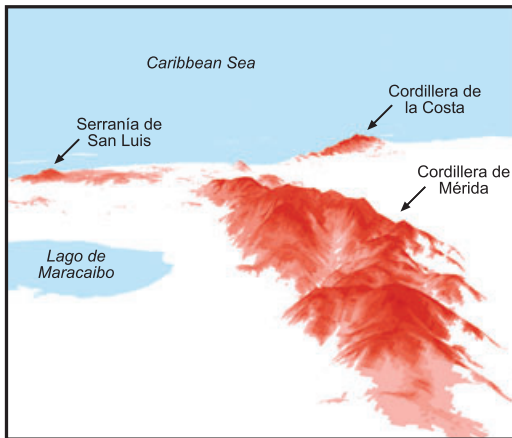
Figure 2. Example of the prediction of a species' abiotically suitable area generated by an ecological niche model. Areas shown in red indicate those estimated as suitable for the rodent *Nephelomys caracolus* in northern Venezuela based on climatic variables; the prediction appears draped over a three-dimensional representation of elevation, with increasingly dark tones indicating progressively stronger predictions for the species.[17] In addition to documented areas inhabited by *N. caracolus* in the Cordillera de la Costa, the model indicates abiotically suitable areas in the Cordillera de Mérida (which holds records of the congener *N. meridensis*) and the Serranía de San Luis (which lacks published records of either species).

improved environmental data such as remotely sensed information from satellites, offer great promise for improving the predictions of niche models,[26,27] but I present these five, which I consider fundamental to progress in the field. They apply to all niche-modeling algorithms using presence-only data along with a background (or pseudoabsence) sample of the environments available in the study region. In addition, some are germane for techniques that use presence-only occurrence data without even taking a background or pseudoabsence sample, as well as for modeling based on presence–absence and abundance data. Those concerning data availability also, of course, hold relevance for myriad other uses of museum data, far beyond those associated with ecological niche modeling.[28] Fortunately, despite the gravity of these five obstacles, all are surmountable.

Pitfall 1: Incorrect taxonomic identifications
Taxonomic knowledge—even of relatively well-studied groups such as birds, mammals, and butterflies—remains inadequate. Termed the Linnaean shortfall, this problem is especially marked in

tropical regions of the greatest biodiversity.[2,29] As a result, the extremely variable quality of species identifications for museum specimens plagues their use in niche modeling.[5] For example, in my research on the spiny pocket mice (*Heteromys*) of Colombia, I found that circa one-third of the specimens lacked correct identifications, with misidentified specimens sometimes (falsely) "documenting" the species hundreds of kilometers from where it actually occurs.[30] Furthermore, whereas only two species were known from South America prior to that date, my collaborators and I subsequently described three species of South American *Heteromys* new to science based on existing (previously misidentified) specimens.[31–33] Reading of the literature and conversations with other taxonomists indicate that such underestimated and mischaracterized diversity represents the norm for small nonvolant mammals throughout the tropics. These anecdotes exemplify the fact that many studies using identifications from museum databases in reality model entities that do not represent the intended species. In a nutshell, taxonomic studies (often termed *revisionary taxonomy*) are necessary to determine what the real species are and identify each specimen correctly. Furthermore, in museum databases, each identified specimen requires associated fields indicating the source and trustworthiness of the identification. The justifications for rectifying taxonomic inadequacies of museum data sources, and the strategies for doing so, intertwine with those for the second pitfall, so I consider them together later.

Pitfall 2: Lacking or inadequate databasing and georeferences
Despite remarkable progress, the overwhelming majority of museum holdings linger undigitized, often with nonexistent or inadequate georeferences.[5,24,c] Overall, vertebrates and plants enjoy much higher rates of databasing, but even with vertebrates, information on the holdings of many critical, vast collections often remains in paper format. Furthermore, even when species are well characterized

---

[c]Although documentation of biodiversity via continued fieldwork and collection of museum specimens constitutes a critical endeavor, especially in tropical regions and for areas experiencing high rates of habitat loss,[34] I focus here on the need for making existing biodiversity information available for ecological niche modeling.

Table 3. Terms related to major pitfalls hindering ecological niche modeling

Term	Definition
Georeference	Coordinates of latitude and longitude (or another system) indicating the position of a point in space.
GPS (global positioning system)	A device that detects information from satellites to determine the current position of the user on Earth. Other features often include the ability to calculate the distances and bearings of other sites relative to the current position.
Sampling bias	Variation in the probability that a site has been sampled by biologists. Generally, such bias corresponds to accessibility in geographic space and often also leads to sampling bias in environmental space.
Transferability	The application of a model (calibrated in one region) to another place in geography and/or to another time period.
Model complexity	The level of detail of a model, here often in terms of how many variables are included, and the weights assigned to them (but also can include how complex the functions are that model a species' response to particular variables).

taxonomically, identifications are correct, and online museum databases exist, ecological niche modeling requires a high-quality georeference for each specimen (typically, geographic coordinates of latitude and longitude). This bottleneck in data availability hinders modeling at least as much as the previously mentioned inadequacies. In addition to developing technology to serve museum data on the internet, the biodiversity informatics community also established suggested standards for georeferencing.[23] First of all, georeferences should be given to the greatest precision possible. Second, a radius of possible error must be associated with each georeference, so that the user can filter records, discarding all that are not of sufficient quality for the study at hand.

I provide another anecdote from *Heteromys*, for one particularly important, yet difficult to locate, locality in Colombia: "Serranía del Darién, Alto de Barrigonal." The only record for the *H. desmarestianus* species group in South America corresponds to this site. The Serranía del Darién, a mountain range that is geologically Central American, runs along the border between Colombia and Panama, but I could not find the precise locality Alto de Barrigonal on any map or in any gazetteer. Later, I noticed that the staff of the museum housing the specimen had georeferenced it—to degrees, minutes, and seconds! Because they could not find the specific locality either, coordinates had been assigned for the middle of the mountain range. However, without an

associated possible error radius, this (otherwise reasonable) georeference was useless for most studies or, worse, perhaps even hugely misinformative. Fortunately, upon consultation, the original collector provided specific information that allowed location of the site with a fairly small radius of error.[30]

Improving the quality of georeferences holds great promise for increasing the utility of niche models.[35] Despite substantial progress with automated georeferencing,[36] determining coordinates by consulting detailed maps, field notes in museum archives, and the original collectors allows for substantially more-accurate coordinates.[33,37] Such efforts often are necessary to arrive at coordinates that are both sufficiently precise and accurate to match the approximate resolution of currently available climatic data interpolated from weather stations.[38] Other data sources, especially remotely sensed data regarding vegetation and land cover derived from satellites, may require even finer georeferences, such as those from GPS readings. Fortunately, many specimens (of various species) often come from the same locality, so not every specimen needs to be georeferenced individually.

Of course, inadequacies of taxonomic knowledge, databasing, and georeferencing (Pitfalls 1 and 2) were recognized in the past, with pioneering and visionary documents produced by the overlapping communities of taxonomic/systematic biology and biodiversity informatics, outlining the steps necessary to rectify them while taking into

Table 4. Issues related to calibrating ecological niche models

Term	Definition
Spatial bias	Bias (see sampling bias; Table 3) in geographic space.
Environmental bias	Bias (see sampling bias; Table 3) in environmental space (e.g., for the environmental variables used in ecological niche modeling).
Noise	Random variation due to the effects of sampling, without any systematic bias.
Signal	The true information concerning an entity being studied. Here, from data representing a species' niche accurately.
Calibration	The step or steps involved in forming a model, here one that estimates a species' niche based on occurrence data and environmental variables.
Evaluation	Here, the use of data not used in model calibration in order to determine model performance and significance, ideally using evaluation data fully independent of those employed in model calibration.
Sampling effort	Generally, the strength or intensity of sampling by biologists, but also may include consideration of the suite of techniques employed.
Target group	Those species that can be observed or collected with the same sampling techniques as the focal species of interest.

account necessary safeguards regarding the online dissemination of sensitive information (e.g., for endangered species vulnerable due to the pet trade[3,39]). Indeed, some museums and taxonomic communities have achieved spectacular success in databasing, and sometimes also georeferencing. However, progress for other museums and groups of organisms lags behind, especially with regard to the level and quality of georeferencing.[24] Furthermore, almost all groups of organisms still suffer from the Linnaean shortfall (with the lack of trained and employed systematists hindering progress), especially for tropical regions and for hyper-diverse invertebrates.[2]

The necessary philosophies and technologies exist for all of these challenges, awaiting implementation once advocacy leads to the necessary funding. The excellent progress in databasing and georeferencing should continue, and advancement in taxonomic studies must accelerate. Immense investment in specimen-based taxonomic studies and in pervasive digitization and georeferencing of museum holdings should lead to high-quality occurrence data available over the internet for a wide array of plants and animals.[d] In my opinion, such

funding can be justified—by making the value of the resulting data *to society as a whole* understood broadly by policy makers and governmental officials, especially those making budgetary decisions. The literature from the past few years makes it increasingly clear that those museums and taxonomic groups that achieve high-quality data available online will get their data used (showing the relevance of museum collections to society), whereas those that do not, will not. Furthermore, the better the data, the more realistic the resulting models—providing additional incentive for improvement in data quality for museum-based scientists concerned with the misuse of inadequate data if served online. The critical applications of museum data via ecological niche modeling represent an opportunity for museums to contribute information and solutions to key societal issues, as well as a compelling justification for investment in the taxonomic studies and concomitant discovery of biodiversity that have been and remain a core mission of museums.

Pitfall 3: Effects of sampling bias across geography

The very nature of the collecting expeditions that have led to the unique and irreplaceable museum holdings that constitute our primary documentation of biodiversity creates a major problem for ecological niche modeling (Table 4)—one that

---

[d]I leave to others better capable the calculations regarding the monetary amount of funding necessary.

researchers must, and can, address in order to produce realistic models.[1,5] Sampling effort varies tremendously across the globe, for example, with much higher effort in populated areas, temperate regions, and areas near roads and rivers that serve as access points.[40,41] This spatial bias affects the model-calibration process greatly (and also leads to problems for evaluating model quality, see later). The most serious problems occur when biases in the environments sampled accompany the spatial biases. The aim, of course, is to model the requirements of the species (its niche), not any bias in collection effort (in space and/or environment). When sampling effort across geography can be quantified, it should be integrated into the model-calibration process itself. Although the information necessary for direct quantification of sampling effort (e.g., via consultation of detailed field notes) does not exist for most museums, one suitable proxy for it does.

Because museum specimens themselves represent the product of sampling, they collectively can serve as a surrogate of sampling effort. This requires selection of a "target group" of species that are sampled with the same techniques as the focal species of interest (e.g., small nonvolant mammals, such as rodents, marsupials, and shrews in the Neotropics[42]). Such information—which indicates which areas, and hence environments, have been more thoroughly sampled—can then be integrated into model calibration, in effect correcting for sampling bias.[43,e] The beauty lays in the fact that digitization and georeferencing of all records of such a group (rectification of Pitfall 2) allows researchers to circumvent the serious problems posed by Pitfall 3, further leveraging such efforts![f] Fortunately, for this use, the specimen identifications only need to be correct for the focal species being modeled, with specimens of other species simply identified as members of the target group. Hence, this strategy can be implemented before taxonomists fully rectify the Linnaean shortfall (Pitfall 1).

Pitfall 4: Selection of the study region

Determining the relevant study region for model calibration represents a topic of great importance—especially when transferring a niche model to another place or time. For years, this puzzling question did not seem a primary determinant of model quality. However, with the advent of advanced techniques capable of fitting very complex models, it came to the forefront. Recent research has clarified that environmental data from regions that may hold suitable conditions but in which the species is absent for other reasons should not be included in background samples (which are intended to represent the environments available to the species).[17,44] Specifically, a species may be absent from such areas due to dispersal barriers or because biotic interactions lead to a negative population growth rate for the focal species (for example, because of the presence of a competitor or the lack of a key mutualist[18]). Environmental information from such areas provides false negative signal that thwarts efforts to model

---

[e]Lacking data regarding sampling effort (quantified directly, or approximated via indices calculated from data regarding the target group), modelers face a quandary: to reduce bias (geographic and possibly environmental) but not signal (correct information regarding the species' niche). To ameliorate problems associated with biased sampling effort, researchers can filter localities spatially, for example, removing nearby localities (e.g., maintaining the largest set of localities possible, subject to the constraint that all lie at least x km from each other[17]). Determination of the appropriate distance x remains arbitrary at present and likely depends on the level of sampling bias and the heterogeneity of the environment (which itself likely varies across the study region).[7] If the chosen x is too small, not enough bias will be removed; conversely, if it is too large, the species' true niche signal will be diluted. Hence, although spatial filtering represents a helpful approach when no other solution is available, inclusion of information regarding sampling effort in the model-calibration process (e.g., via information on the target group) constitutes a far-superior tactic.[42,43]

[f]Information regarding the target group provides an additional valuable benefit for ecological niche models: the possibility of testing for artifactual absences. The projection of a niche model onto geography indicates the abiotically suitable areas for the species, but dispersal limitations[19] or biotic interactions[20] may limit it from occupying all abiotically suitable areas. In order to identify such cases, researchers must consider whether sampling effort has been sufficient to demonstrate a species' absence from a region that holds abiotically suitable conditions but lacks records of the species. Such tests for artifactual absences can be accomplished using indices of sampling effort provided by data regarding the target group.[42]

Table 5. Challenges related to transferring an ecological niche model to another place or time period and to model evaluation

Term	Definition
Interpolation	Here, prediction between known values of an environmental variable.
Extrapolation	Here, prediction into environmental values beyond the range (in environmental space) of the geographic area on which the model was calibrated (common when a model is applied to cross-time or cross-space situations).
Nonanalog environments	Environmental conditions (often climatic) in one place or time period that do not exist in another place or time (e.g., that used in model calibration). In ecological niche modeling, non analog environments require extrapolation in environmental space to make a prediction, which generally should be interpreted with great caution.
Truncated response curves	Curves of the species' response to a particular variable that do not include the full domain of an environmental variable. Such a situation can lead to nonanalog environments when transferring a model to another place or time period.
Performance	Characterization of how well or poorly a model predicts independent data, but not necessarily including statistical assessment of model significance.
Significance	Determination via statistical tests whether or not predictions of evaluation data differ from a random null hypothesis with a particular level of probabilistic confidence. Often based on some measure of model performance, tests of model significance typically assess whether the model predicts evaluation data better than random expectations (one-tailed hypothesis).
Overfitting	The situation when model complexity is excessive and a model shows close fit to calibration data but is less able to predict independent evaluation data. Note that overfitting can be to noise and/or to sampling bias.

the species' niche and the corresponding abiotically suitable areas in geography.[g]

Critically, these smaller study regions necessary for model calibration in the "here and now" highlight a key limitation of applying niche models to other places or time periods: the problem of extrapolating a model (in environmental space) to make predictions for environmental conditions that do not exist in the calibration study region (e.g., nonanalog climates; Table 5[13,45]). For example, a species may survive and reproduce very well at the warmest temperatures present in the study region, but how can we estimate what its response would be to even warmer temperatures on another continent or under future climatic change? This problem, often termed "truncated response curves," has no easy solution, with the required laboratory or greenhouse physiological experiments laborious and seldom feasible.[46] Hence, in these cases, the researcher must make some assumption regarding the species' response in order to derive a prediction from the model.[17] In practice, some algorithms flag affected pixels of a cross-space or cross-time prediction, indicating where extrapolation in environmental space occurred and how strong an effect it had on the prediction—and therefore, where caution should be taken when interpreting predictions.[47] Although few current studies take into account these paramount principles of study-region selection and extrapolation in environmental space, they now exist clearly in the literature, and I predict that their consideration will become standard. They also

---

[g]In most cases, it will be difficult to identify such regions operationally in great detail (in order to exclude them from the study region used for model calibration). However, fairly reasonable study regions can be approximated (for example, by taking into account the distributions of major vegetation types), matching the relevant assumptions of modeling much more closely than the excessively expansive study regions often used presently.[17]

embody key principles for model evaluation,[7] as shown later.

Pitfall 5: Model evaluation to identify optimal model complexity

Researchers should demonstrate good performance (and statistical significance) for niche models before interpreting and using them for academic or applied uses, especially for those models that require transferral to another place or time (Table 5). Many factors can lead to poorly performing models, but the need to achieve (or at least approximate) optimal model complexity cuts across all algorithms and uses and is especially germane for cross-space or cross-time studies.[7] Unfortunately, evaluations that allow valid quantification of model performance and identification of optimal model complexity often evade researchers.[h] Having only presence-only data and a background sample (i.e., no information regarding sites where the species is absent)—plus the very objective of niche modeling: identifying the abiotically suitable areas for the species, rather than its occupied distributional area—seriously limits the options for model evaluation. However, several relevant strategies exist.

The overreaching principle of model evaluation should be to determine whether the model predicts independent data well and—for many applications—whether it has the ability to predict across space and/or time. A good model describes the species' requirements sufficiently (and better than a random prediction) but does not overfit to the peculiarities—be they bias or noise—of the calibration data. Hence, with techniques capable of producing complex models, controlling model complexity—avoiding overfitting—becomes vital.[6,7] This is especially important given the pervasive biases of biodiversity data and the small sample sizes available for many species, the latter of which

leads to concomitant problems regarding noise. Algorithmic settings that influence model complexity (e.g., the level of regularization in Maxent[50]) can affect model output greatly, creating the need to vary model parameters and select those that lead to the highest performance in evaluations made with independent data.[51–53]

Meaningful evaluations with these goals assess a model's prediction using localities that are independent of those used to calibrate it.[i] For example, researchers should not use evaluation localities that lie close to the calibration ones. That nonindependence inflates estimates of performance and significance because such sites represent the same environmental information without constituting truly independent data points.[49,j] In an even broader sense, transferring a model to another place or time period (e.g., after climatic change) requires knowledge that the model performs well under some kind of transferral. Occurrence data from other time periods generally do not exist for studies of the effects of climatic change. In such cases, the only option for evaluating transferral with independent data is to transfer across space, for example, via spatially structured subsampling of available occurrence records into

---

[h]Confusion regarding evaluations of model quality permeates much of the literature, primarily because researchers frequently use measures appropriate for presence–absence evaluation data but that misrepresent reality when applied to presence only evaluation data (along with a background or pseudoabsence sample). Perhaps most importantly, false positive rates (= commission error rates) suffer from overinflation, often very substantially.[48] Myriad critical technical points regarding model evaluation appear elsewhere.[7,25,49]

[i]As intimated above (Pitfall 4), theory indicates that the study region appropriate for model calibration also constitutes that appropriate for model evaluation.[7] Violation of this principle pervades the literature, leading to a double whammy: poor models (because of an improper calibration region) that appear to be very good (because of the same, inappropriate, region used for evaluation). Typically, this occurs with extremely large study regions that include extensive areas of abiotically suitable conditions from which the species is absent due to dispersal limitations associated with contingent events of its evolutionary history[19] (although the limitations also could derive from biotic interactions that lead to a negative population growth rate for the focal species[20]). The resulting false negative signal from such regions leads to models that vastly underestimate the abiotically suitable area for the species; concomitantly, the evaluation erroneously suggests that such a prediction is correct, because no occurrence record for the species exists from those regions (which lie on the other side of the dispersal barrier). [j]Spatial filtering of occurrence data (mentioned above for Pitfall 3) likely ameliorates this problem, but the degree of filtering remains arbitrary.

---

calibration and evaluation datasets.[6,7,54,k] Although few recent studies employ evaluation strategies effective in identifying optimal model complexity and the ability to predict across space and/or time, I predict that they will become pervasive.

Agenda: Making museum data and ecological niche modeling ready to address critical environmental issues of the 21st century

Despite these largely cautionary reflections, tremendous opportunity for progress exists, both with regard to data availability and the implementation of ecological niche modeling. Concerning data availability, museum holdings only will be maximally useful to society when high-quality information on those specimens is available across regions and taxonomic groups, ready to be accessed *when the particular problem for a species of interest presents itself*. For example, studies of an emerging zoonotic disease or recently detected invasive species call for quick action. In such a situation, society needs to be able to count on the data being ready, rather than requiring a special (and often prohibitively slow) effort to clarify the taxonomy and digitize and georeference the records of the species of interest (Pitfalls 1 and 2). Furthermore, correcting for collection biases (a critical issue in niche modeling; Pitfall 3) will be possible when data for the whole target group exist for all the museums providing records of the focal species itself.

In my view, museums (or better yet, consortia of museums) can make effective arguments for major funding to realize these societal needs. The key lies in leveraging unique museum holdings to justify investment in taxonomic studies, identification of specimens, databasing, and georeferencing of localities. Particular subgroups more likely to be of economic or medical relevance (e.g., rodents or flies) and those already in a relatively advanced state of taxonomic knowledge surely can sell these ideas more easily. However, in principle, the utility of in-

vestments in making museum data useful to society should ring clear across taxonomic groups and geographic regions. For the short term, I suggest continued large-scale, across-the-board databasing and georeferencing initiatives, with a variety of high-priority taxonomic projects realized in parallel. Such progress will allow niche modeling for the most-critical focal species, taking advantage of digitized and georeferenced records of the full target group in order to correct for collection biases (even if the taxonomic information of some members of the target group remains incomplete). In the medium term, further taxonomic efforts should fill in the gaps in high-quality identifications as rapidly as possible. Although progress rectifying the Linnaean shortfall surely will proceed unevenly with respect to taxonomic group and geographic region, effective advocacy should lead to ever-broadening coverage over time.

Parallel to such data-availability efforts, society needs advances in ecological niche modeling, with regard both to technology and—especially—to human capacity. For the former, the field requires software that achieves an appropriate balance between automation and supervision. Currently, many software packages implementing niche-modeling algorithms provide default settings that allow efficient processing of data automatically, yielding niche models for many species simultaneously. However, such ease of use also promotes ease of misuse. Biologists with knowledge of the species and geographic area must select the relevant study region for model calibration carefully (Pitfall 4). Similarly, well-trained modelers must supervise the evaluation process, because default settings do not necessarily lead to models with optimal levels of model complexity (Pitfall 5). At present, it remains time consuming to implement the steps of model calibration and evaluation necessary to produce and document highly performing models, even for skilled users programming batch files via code. This situation calls for software that automates repetitive aspects of the process, while allowing (and forcing) the user to provide input when critical biological and conceptual decisions need to be made. Ideally, such software will be general with respect to the actual modeling algorithm or algorithms used to create the model—that is, umbrella tools that can be employed in concert with any particular modeling algorithm desired.

---

[k]As is appropriate for tests of transferability, these evaluations require the same caveats and assumptions related to truncated response curves and nonanalog environments mentioned earlier,[45] as well as the assumption of no differences in inherited niche physiology across geography or time.[7]

Complementarily, the field must produce a much larger number of scientists capable of building and applying high-quality niche models, as well as a broad community able to appraise their quality and utility. Rather similar to the trend in job searches in museums and universities in the 1980s and 1990s, which often targeted taxonomists and evolutionary biologists with molecular skills (e.g., protein electrophoresis and later DNA sequencing), I predict that the next decade will see a wave of openings specifically desiring scientists fluent in ecological niche modeling and associated spatial analyses.[11] To train this generation of researchers and expand the knowledge base of those educated before them requires both clear literature and extensive educational opportunities. Although this research area remains in extremely rapid development, the literature published over the past few years seems to be leading towards a synthesis of ideas and a maturation of methodologies.[7] Obviously, editors, reviewers, and the pool of scientists worldwide all play vital roles in helping achieve these goals via productive debate, insistence upon rigor, and a fostering of creativity. Furthermore, continued proliferation of workshops and graduate courses will be necessary to train the scientists needed by society. This training must include both the theoretical and the methodological aspects of niche modeling. In my experience, such venues bring together fascinating mixes of people excited to apply emerging technology, consider new ideas, and answer biogeographic questions regarding the systems they study.

Indeed, I have found that ecological niche modeling and biodiversity informatics attract scientists with exceptionally diverse backgrounds and interests. That, together with their intelligence and creativity, epitomizes the interdisciplinary nature of the field and has made it a fascinating area of research over the past decade. With a meeting of societal needs, stores of critical data, blossoming technologies, and capable minds, the next years provide the opportunity to realize the potential that museums and ecological niche modeling offer society. Given the enormous relevance of biodiversity and the environment to humankind over the next century—and the irreversibility of many of the biological changes that may occur—a significant component of history will depend on our success.

Acknowledgments

Conflicts of interest

The author declares no conflicts of interest.

References

1. Funk, V.A. & K.S. Richardson. 2002. Systematic data in biodiversity studies: use it or lose it. *Syst. Biol.* **51:** 303–316.
2. Wilson, E.O. 2003. The encyclopedia of life. *Trends Ecol. Evol.* **18:** 77–80.
3. Soberón, J. & A.T. Peterson. 2004. Biodiversity informatics: managing and applying primary biodiversity data. *Phil. Trans. R. Soc. Lond. B.* **359:** 689–698.

4. Boakes, E.H., P.J.K. McGowan, R.A. Fuller, *et al.* 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLOS Biology.* **8:** 1–11.

5. Newbold, T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progr. Phys. Geogr.* **34:** 3–22.

6. Jiménez-Valverde, A., A.T. Peterson, J. Soberón, *et al.* 2011. Use of niche models in invasive species risk assessments. *Biol. Invasions.* **13:** 2785–2797.

7. Peterson, A.T., J. Soberón, R.G. Pearson, *et al.* 2011. *Ecological Niches and Geographic Distributions.* Monographs in Population Biology 49, Princeton University Press. Princeton, NJ.

8. Zimmermann, N.E., T.C. Edwards, Jr., C.H. Graham, *et al.* 2010. New trends in species distribution modelling. *Ecography* **33:** 985–989.

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

10. Kozak, K.H., C.H. Graham & J.J. Wiens. 2008. Integrating GIS-based environmental data into evolutionary biology. *Trends Ecol. Evol.* **23:** 141–148.

11. Swenson, N.G. 2008. The past and future influence of geographic information systems on hybrid zone, phylogeographic and speciation research. *J. Evol. Biol.* **21:** 421–434.

12. Thomas, C.D., A. Cameron, R.E. Green, *et al.* 2004. Extinction risk from climate change. *Nature* **427:** 145–148.

13. Elith, J. & J.R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* **440:** 677–697.

14. Chase, J.M. & M.A. Leibold. 2003. *Ecological Niches: Linking Classical and Contemporary Approaches.* University of Chicago Press. Chicago.

15. Holt, R.D. 2009. Bringing the Hutchinsonian niche into the 21st Century: ecological and evolutionary perspectives. *Proc. Nat. Acad. Sci. U.S.A.* **106:** 19659–19665.

16. Soberón, J.M. 2010. Niche and area of distribution modeling: a population ecology perspective. *Ecography* **33:** 159–167.

17. Anderson, R.P. & A. Raza. 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. *J. Biogeogr.* **37:** 1378–1393.

18. Gaston, K.J. 2003. *The Structure and Dynamics of Geographic Ranges.* Oxford University Press. Oxford.

19. Anderson, R.P., A.T. Peterson & M. Gómez-Laverde. 2002. Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos* **98:** 3–16.

20. Anderson, R.P., M. Gómez-Laverde & A.T. Peterson. 2002. Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecol. Biogeogr.* **11:** 131–141.

21. Thomson, MC.,V. Obsomer, M. Dunne, *et al.* 2000. Satellite mapping of loa loa prevalence in relation to ivermectin use in west and central Africa. *Lancet* **356:** 1077–1078.

22. Gyapong, J.O., D. Kyelem, I. Kleinschmidt, *et al.* 2002. The use of spatial analysis in mapping the distribution of ban-

croftian filariasis in four West African countries. *Annals. Trop. Med. Paras.* **96:** 695–705.

23. Stein, B.R., & J. Wieczorek. 2004. Mammals of the world: MaNIS as an example of data integration in a distributed network environment. *Biodiv. Informatics* **1:** 14–22.

24. Ariño, A.H. 2010. Approaches to estimating the universe of natural history collections data. *Biodiv. Informatics.* **7:** 81–92.

25. Araújo, M.B. & A. Guisan. 2006. Five (or so) challenges for species distribution modelling. *J. Biogeogr.* **33:** 1677–1688.

26. Bradley, B.A. & E. Fleishman. 2008. Can remote sensing of land cover improve species distribution modeling? *J. Biogeogr.* **35:** 1158–1159.

27. Austin, M.P. & K.P. Van Neil. 2010. Improving species distribution models for climate change studies: variable selection and scale. *J. Biogeogr.* **38:** 1–8.

28. Graham, C.H., S. Ferrier, F. Huettman [*sic*], *et al.* 2004. New developments in museum-based informatics and application in biodiversity analysis. *Trends Ecol. Evol.* **19:** 497–503.

29. Whittaker, R.J., M.B. Araújo, P. Jepson, *et al.* 2005. Conservation biogeography: assessment and prospect. *Diversity Distrib.* **11:** 3–23.

30. Anderson, R.P. 1999 [2000]. Preliminary review of the systematics and biogeography of the spiny pocket mice (*Heteromys*) of Colombia. *Rev. Acad. Colomb. Cienc. Exactas Físicas Nat.* **23**(suplemento especial): 613–630.

31. Anderson, R.P. & P. Jarrín-V. 2002. A new species of spiny pocket mouse (Heteromyidae: *Heteromys*) endemic to western Ecuador. *Amer. Mus. Novitates* **3382:** 1–26.

32. Anderson, R.P. 2003. Taxonomy, distribution, and natural history of the genus *Heteromys* (Rodentia: Heteromyidae) in western Venezuela, with the description of a dwarf species from the Península de Paraguaná. *Amer. Mus. Novitates* **3396:** 1–43.

33. Anderson, R.P. & E.E. Gutiérrez. 2009. Taxonomy, distribution, and natural history of the genus *Heteromys* (Rodentia: Heteromyidae) in central and eastern Venezuela, with the description of a new species from the Cordillera de la Costa. *Systematic Mammalogy: Contributions in Honor of Guy G. Musser* (ed. by R.S. Voss and M.D. Carleton). *Bull. Amer. Mus. Nat. Hist.* **331:** 33–93.

34. Raven, P.H. & E.O. Wilson. 1992. A fifty-year plan for biodiversity surveys. *Science* **258:** 1099–1100.

35. Feeley, K.H. & M.R. Silman. 2010. Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. *J. Biogeogr.* **37:** 733–740.

36. Guralnick, R.P., J. Wieczorek, R. Beaman, *et al.* 2006. BioGeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biol.* **4:** 1901–1909.

37. García-Milagros, E. & V.A. Funk. 2010. Improving the use of information from museum specimens: using Google Earth© to georeference Guiana Shield specimens in the US National Herbarium. *Front. Biogeogr.* **2:** 71–77.

38. Hijmans, R.J., S.E. Cameron, J.L. Parra, *et al.* 2005. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25:** 1965–1978.

39. Samper, C. 2004. Taxonomy and environmental policy. *Phil. Trans. R. Soc. Lond. B.* **359:** 721–728.

40. Reddy, S. & L.M. Dávalos. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. *J. Biogeogr.* **30:** 1719–1727.

41. Kadmon, R., O. Farber & A. Danin. 2004. Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecol. Appl.* **14:** 401–413.

42. Anderson, R.P. 2003. Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *J. Biogeogr.* **30:** 591–605.

43. Phillips, S.J., M. Dudík, J. Elith, *et al*. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* **19:** 181–197.

44. Barve, N., V. Barve, A. Jiménez-Valverde, *et al*. 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol. Model.* **222:** 1810–1819.

45. Williams, J.W., S.T. Jackson & J.E. Kutzbach. 2007. Projected distributions of novel and disappearing climates by 2100 AD. *Proc. Nat. Acad. Sci. U.S.A.* **104:** 5738–5742.

46. Williams, J.W. & S.T. Jackson. 2007. Novel climates, no-analog communities, and ecological surprises. *Front. Ecol. Environ.* **5:** 475–482.

47. Elith, J., S.J. Phillips, T. Hastie, *et al*. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity Distrib.* **17:** 43–57.

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

49. Veloz, S.D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *J. Biogeogr.* **36:** 2290–2299.

50. Phillips, S.J., R.P. Anderson & R.E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **190:** 231–259.

51. Elith, J., M. Kearney & S. Phillips. 2010. The art of modelling range-shifting species. *Meth. Ecol. Evol.* **1:** 330–342.

52. Anderson, R.P. & I. Gonzalez, Jr. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. *Ecol. Model.* **222:** 2796–2811.

53. Warren, D.L. & S.N. Seifert. 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecol. Appl.* **21:** 335–342.

54. Araújo, M.B. & C. Rahbek. 2006. How does climate change affect biodiversity? *Science* **313:** 1396–1397.