Species' Distribution Modeling for Conservation Educators and Practitioners

Author(s): Richard G. Pearson

This article is featured in *Lessons in Conservation,* the official journal of the Network of Conservation Educators and Practitioners (NCEP). NCEP is a collaborative project of the American Museum of Natural History's Center for Biodiversity and Conservation (CBC) and a number of institutions and individuals around the world. *Lessons in Conservation* is designed to introduce NCEP teaching and learning resources (or "modules") to a broad audience. NCEP modules are designed for undergraduate and professional level education. These modules—and many more on a variety of conservation topics—are available for free download at our website, ncep.amnh.org.

To learn more about NCEP, visit our website: ncep.amnh.org.

Illustrations obtained from the American Museum of Natural History's library: images.library.amnh.org/digital/

Species' Distribution Modeling for Conservation Educators and Practitioners

Richard G. Pearson*

* The American Museum of Natural History, New York, NY, U.S.A., email pearson@amnh.org

K. Frey

Table of Contents

Species' Distribution Modeling for Conservation Educators and Practitioners

Richard G. Pearson

Abstract

Models that predict distributions of species by combining known occurrence records with digital layers of environmental variables have much potential for application in conservation. Through using this module, teachers will enable students: to develop species' distribution models; to apply the models across a series of analyses; and to interpret predictions accurately. Part A of the synthesis introduces the modeling approach, outlines key concepts and terminology, and describes questions that may be addressed using the approach. A theoretical framework that is fundamental to ensuring that students understand the uses and limitations of the models is then described. Part B of the synthesis details the main steps in building and testing a distribution model. Types of data that can be used are described and some potential sources of species' occurrence records and environmental layers are listed. The variety of alternative algorithms for developing distribution models is discussed, and software programs available to implement the models are listed. Techniques for assessing the predictive ability of a distribution model are then discussed, and commonly used statistical tests are described. Part C of the synthesis describes three case studies that illustrate applications of the models: 1) Predicting distributions of known and unknown species in Madagascar; 2) Predicting global invasions by plants of South African origin; and 3) Modeling the potential impacts of climate change on species' distributions in Britain and Ireland.

This synthesis document is part of an NCEP (Network of Conservation Educators and Practitioners; http://ncep.amnh.org/) module that also includes a presentation and a practical exercise:
- An introduction to species' distribution modeling: theory and practice (presentation by Richard Pearson)

- Species' distribution modeling using Maxent (practical by Steven Phillips)

This module is targeted at a level suitable for teaching graduate students and conservation professionals.

Learning objectives

Through use of this synthesis, teachers will enable students to:

1. Understand the theoretical underpinnings of species' distribution models
2. Run a distribution model using appropriate data and methods
3. Test the predictive performance of a distribution model
4. Apply distribution models to address a range of conservation questions

Introduction and Theory

Introduction

Predicting species' distributions has become an important component of conservation planning in recent years, and a wide variety of modeling techniques have been developed for this purpose (Guisan and Thuiller, 2005). These models commonly utilize associations between environmental variables and known species' occurrence records to identify environmental conditions within which populations can be maintained. The spatial distribution of environments that are suitable for the species can then be estimated across a study region. This approach has proven valuable for generating biogeographical information that can be applied across a broad range of fields, including conservation biology, ecology, and

evolutionary biology. The focus of this synthesis is on conservation-oriented applications, but the methods and theory discussed are also applicable in other fields (see Table 1 for a list of some uses of species' distribution models in conservation biology and other disciplines).

This synthesis aims to provide an overview of the theory and practice of species' distribution modeling. Through use of the synthesis, teachers will enable students to understand the theoretical basis of distribution models, to run models using a variety of approaches, to test the predictive ability of models, and to apply the models to address a range of questions. Part A of the synthesis introduces the modeling approach and describes the usefulness of the models in addressing conservation questions. Part B details the main steps in building a distribution model, including selecting and obtaining suitable data, choosing a modeling algorithm, and statistically assessing predictive performance. Part C of the synthesis provides three case studies that demonstrate uses of species' distribution models.

What is a species' distribution model?
The most common strategy for estimating the actual or potential geographic distribution of a species is to characterize the environmental conditions that are suitable for the species, and to then identify where suitable environments are distributed in space. For example, if we are interested in modeling the distribution of a plant that is known to thrive in wet clay soils, then simply identifying locations with clay soils and high precipitation can generate an estimate of the species' distribution. There are a number of reasons why the species may not actually occupy all suitable sites (e.g. geographic barriers that limit dispersal, competition from other species), which we will discuss later in this synthesis. However, this is the fundamental strategy common to most distribution models.

The environmental conditions that are suitable for a species may be characterized using either a *mechanistic* or a *correlative* approach. Mechanistic models aim to incorporate physiologically limiting mechanisms in a species' tolerance to environmental conditions. For example, Chuine and Beaubien (2001)

modeled distributions of North American tree species by estimating responses to environmental variables (including mean daily temperature, daily precipitation, and night length) using mechanistic models of factors including frost injury, phenology, and reproductive success. Such mechanistic models require detailed understanding of the physiological response of species to environmental factors and are therefore difficult to develop for all but the most well understood species.

Correlative models aim to estimate the environmental conditions that are suitable for a species by associating known species' occurrence records with suites of environmental variables that can reasonably be expected to affect the species' physiology and probability of persistence. The central premise of this approach is that the observed distribution of a species provides useful information as to the environmental requirements of that species. For example, we may assume that our plant species of interest favors wet clay soils because it has been observed growing in these soils. The limitations of this approach are discussed later in the synthesis, but it has been demonstrated that this method can yield valuable biogeographical information (e.g., Raxworthy et al., 2003; Bourg et al., 2005). Since spatially explicit occurrence records are available for a large number of species, the vast majority of species' distribution models are correlative. The correlative approach to distribution modeling is the focus of this synthesis.

The principal steps required to build and validate a correlative species' distribution model are outlined in Figure 1. Two types of model input data are needed: 1) known species' occurrence records; and 2) a suite of environmental variables. 'Raw' environmental variables, such as daily precipitation records collected from weather stations, are often processed to generate model inputs that are thought to have a direct physiological role in limiting the ability of the species to survive. For example, Pearson et al. (2002) used a suite of seven climate variables and five soil variables to generate five model input variables, including maximum annual temperature, minimum temperature over a 20-year period, and soil moisture availability. At this stage, care should be taken to ensure that data are checked for errors. For example, simply plotting the spe-
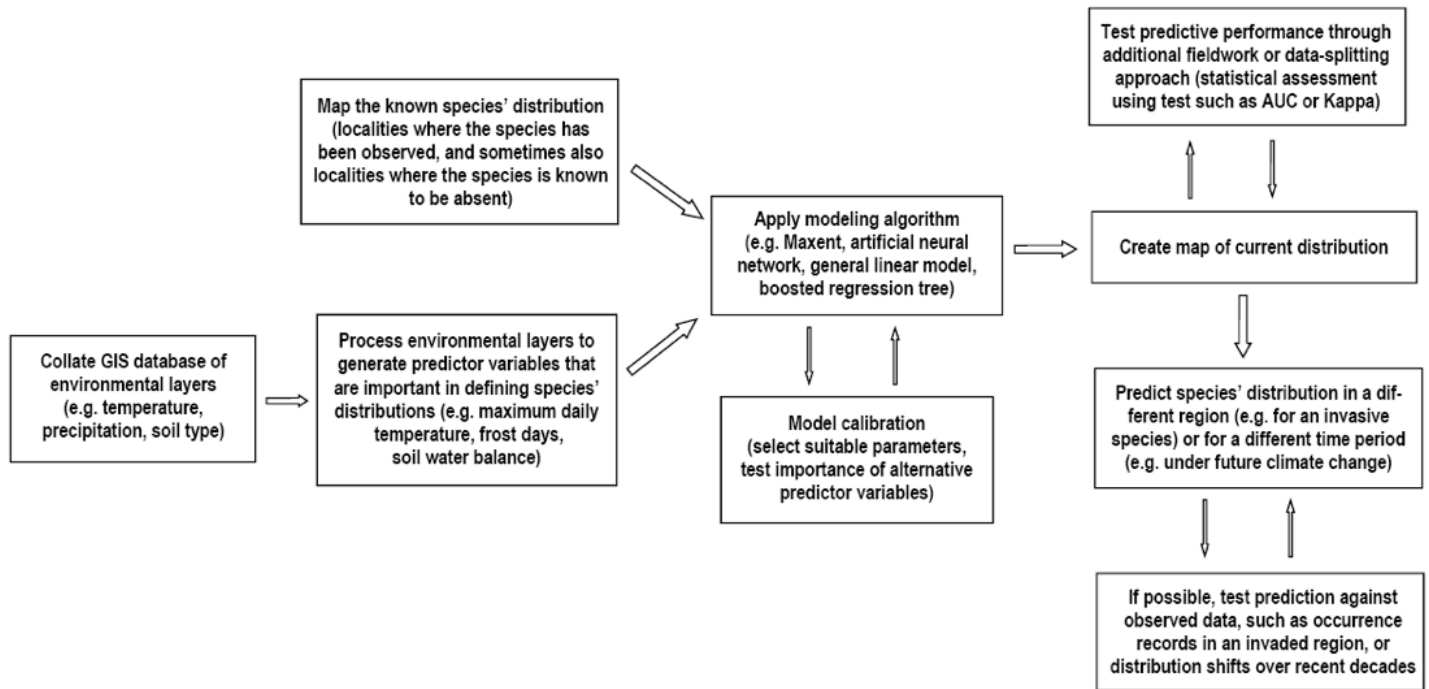
Figure 1. Flow diagram detailing the main steps required for building and validating a correlative species distribution model.

cies' occurrence records in a GIS can help identify records that are distant from other occupied sites and should therefore be checked for accuracy. Data types and sources are discussed in detail in the Section 3: 'Data types and sources'.

The species occurrence records and environmental variables are entered into an algorithm that aims to identify environmental conditions that are associated with species occurrence. If just one or two environmental variables were used, then this task would be relatively straightforward. For example, we may readily discover that our plant species has only been recorded at localities where mean monthly precipitation is above 60mm and soil clay content is above 40%. In practice, we usually seek algorithms that are able to integrate more than two environmental variables, since species are in reality likely to respond to multiple factors. Algorithms that can incorporate interactions among variables are also preferable (Elith et al., 2006). For example, a more accurate description of our plant's requirements may be that it can occur at localities with mean monthly precipitation between 60mm and 70mm if soil clay content is above 60%, and in wetter areas

(>70mm) if clay content is as low as 40%.

A number of modeling algorithms that have been applied to this task are reviewed in Section 4. Depending on the method used, various decisions and tests will need to be made at this stage to ensure the algorithm gives optimal results. For example, a suitable 'regularization' parameter will need to be selected if applying the Maxent method (see Phillips et al., 2006 and Box 3), or the degrees of freedom must be selected if running a generalized additive model (see Guisan et al., 2002). The relative importance of alternative environmental predictor variables may also be assessed at this stage so as to select which variables are used in the final model.

Having run the modeling algorithm, a map can be drawn showing the predicted species' distribution. The ability of the model to predict the known species' distribution should be tested at this stage. A set of species occurrence records that have not previously been used in the modeling should be used as independent test data. The ability of the model to predict the independent data is assessed using a suitable test

statistic. Different approaches to generating test datasets and alternative statistical tests are discussed in the section: 'Assessing predictive performance'. Since a number of modeling algorithms predict a continuous distribution of environmental suitability (i.e. a prediction between 0 and 1, as opposed to a binary prediction of 'suitable' or 'unsuitable'), it is sometimes useful to convert model output into a prediction of suitable (1) or unsuitable (0). This is a necessary step before applying many test statistics; thus, methods for setting a threshold probability, above which the species is predicted as present, are also outlined in the section: 'Assessing predictive performance'.

Once these steps have been completed, and if model validation is successful, the model can be used to predict species' occurrence in areas where the distribution is unknown. Thus, a set of environmental variables for the area of interest is input into the model and the suitability of conditions at a given locality is predicted. In many cases the model is used to 'fill the gaps' around known occurrences (e.g., Anderson et al., 2002a; Ferrier et al., 2002). In other cases, the model may be used to predict species' distributions in new regions (e.g. to study invasion potential, for review see Peterson, 2003) or for a different time period (e.g. to estimate the potential impacts of future climate change, for review see Pearson and Dawson, 2003). Three examples of the use of predictions from species distribution models are presented in Part C. Ideally, model predictions in different regions or for different time periods should be tested against observed data; for example, Thuiller et al. (2005; see Case Study 2) tested predictions of invasion potential using occurrence records from the invaded distribution, while Araújo et al. (2005a) tested predictions of distribution shifts under climate change using observed records from different decades.

This modeling approach has been variously termed 'species distribution', 'ecological niche', 'environmental niche', 'habitat suitability' and 'bioclimate envelope' modeling. Use of the term 'species distribution modeling' is widespread but it should be noted that the term is somewhat misleading since it is actually the distribution of suitable environments that is being modeled, rather than the species' distribution *per se*. Re-

gardless of the name used, the basic modeling process is essentially the same (see Part B) and the theoretical underpinnings of the models are similar. It is essential that these theoretical underpinnings are properly understood in order to interpret model outputs accurately. The following section describes this theoretical framework.

Section 2. Theoretical Framework

This section outlines some of the fundamental concepts that are crucial for understanding how species' distribution models work, what types of questions they are suitable for addressing, and how model output should be interpreted.

Geographical versus environmental space
We are used to thinking about the occurrence of species in geographical space; that is, the species' distribution as plotted on a map. To understand species' distribution models it is important to also think about species occurring in *environmental space*, which is a conceptual space defined by the environmental variables to which the species responds. The concept of environmental space has its foundations in ecological niche theory. The term 'niche' has a long and varied history of use in ecology (Chase and Leibold, 2003), but the definition proposed by Hutchinson (1957) is most useful in the current context. Hutchinson defined the *fundamental niche* of a species as the set of environmental conditions within which a species can survive and persist. The fundamental niche may be thought of as an 'n-dimensional hypervolume', every point in which corresponds to a state of the environment that would permit the species to exist indefinitely (Hutchinson, 1957). It is the axes of this n-dimensional hypervolume that define environmental space.

Visualizing a species' distribution in both geographical and environmental space helps us to define some basic concepts that are crucial for species' distribution modeling (Fig. 2). Notice that the observed localities constitute all that is known about the species' *actual distribution*; the species is likely to occur in other areas in which it has not yet been detected (e.g., Fig. 2, area *A*). If the actual distribution is plotted in envi-
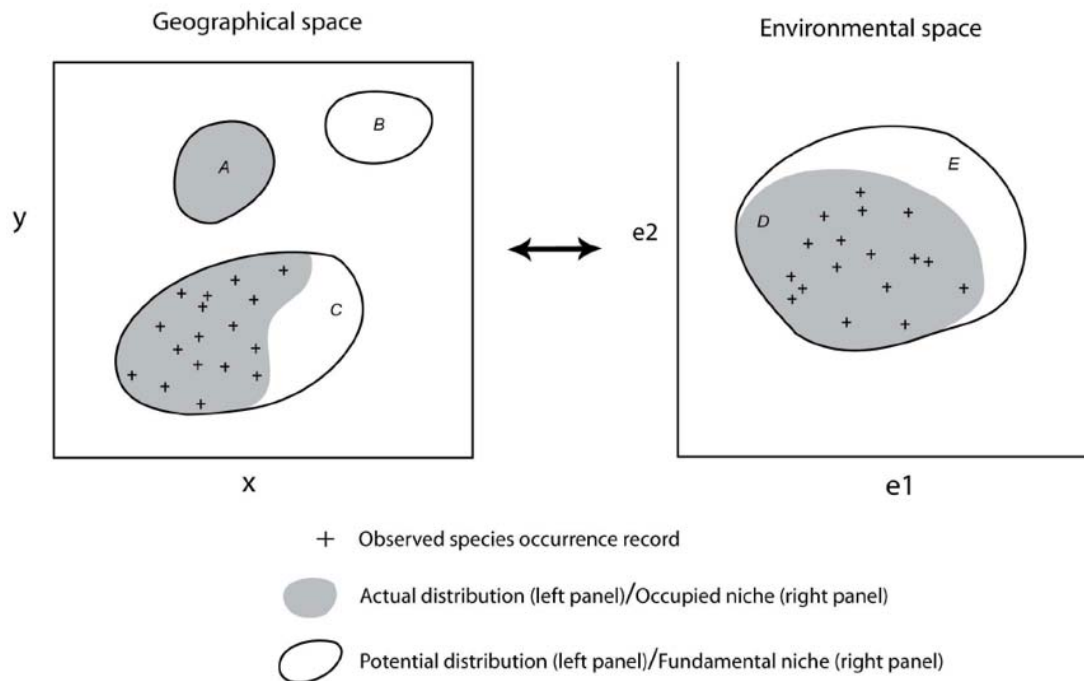
Figure 2. Illustration of the relationship between a hypothetical species' distribution in geographical space and environmental space. Geographical space refers to spatial location as commonly referenced using x and y coordinates. Environmental space refers to Hutchinson's n-dimensional niche, illustrated here for simplicity in only two dimensions (defined by two environmental factors, e_1 and e_2). Crosses represent observed species occurrence records. Grey shading in geographical space represents the species' actual distribution (i.e. those areas that are truly occupied by the species). Notice that some areas of actual distribution may be unknown (e.g. area A is occupied but the species has not been detected there). The grey area in environmental space represents that part of the niche that is occupied by the species: the occupied niche. Again, notice that the observed occurrence records may not identify the full extent of the occupied niche (e.g. the shaded area immediately around label D does not include any known localities). The solid line in environmental space depicts the species' fundamental niche, which represents the full range of abiotic conditions within which the species is viable. In geographical space, the solid lines depict areas with abiotic conditions that fall within the fundamental niche; this is the species' potential distribution. Some regions of the potential distribution may not be inhabited by the species due to biotic interactions or dispersal limitations. For example, area B is environmentally suitable for the species, but is not part of the actual distribution, perhaps because the species has been unable to disperse across unsuitable environments to reach this area. Similarly, the non-shaded area around label C is within the species' potential distribution, but is not inhabited, perhaps due to competition from another species. Thus, the non-shaded area around label E identifies those parts of the fundamental niche that are unoccupied, for example due to biotic interactions or geographical constraints on species dispersal.

ronmental space then we identify that part of environmental space that is occupied by the species, which we can define as the *occupied niche*.

The distinction between the occupied niche and the fundamental niche is similar, but not identical, to Hutchinson's (1957) distinction between the realized niche and the fundamental niche. With reference to the case of two species utilizing a common resource, Hutchinson described the realized niche as comprising that portion of the fundamental niche from which a species is not excluded due to biotic competition. The definition of the occupied niche used in this synthesis broadens this concept to include geographical and historical constraints resulting from a species' limited ability to reach or re-occupy all suitable areas, along with biotic interactions of all forms (competition, predation, symbiosis, and parasitism). Thus, the occupied niche reflects all constraints imposed on the actual distribution, including spatial constraints due to

limited dispersal ability, and multiple interactions with other organisms.

If the environmental conditions encapsulated within the fundamental niche are plotted in geographical space then we have the *potential distribution*. Notice that some regions of the potential distribution may not be inhabited by the species (Fig. 2, areas B and C), either because the species is: excluded from the area by biotic interactions (e.g., presence of a competitor or absence of a food source); the species has not dispersed into the area (e.g., there is a geographic barrier to dispersal, such as a mountain range, or there has been insufficient time for dispersal); or the species has been extirpated from the area (e.g. due to human modification of the landscape).

Before we go on to discuss how these concepts are used in distribution modeling, it is important to appreciate that the environmental variables used in a distribution model are unlikely to define all possible dimensions of environmental space. Hutchinson (1957) originally proposed that all variables, "both physical and biological" are required to define the fundamental niche. However, the variables available for modeling are likely to represent only a subset of possible environmental factors that influence the distribution of the species. Variables used in modeling most commonly describe the physical environment (e.g. temperature, precipitation, soil type), though aspects of the biological environment are sometimes incorporated (e.g. Araújo and Luoto 2007, Heikkinen et al. 2007). However, the distinction between biotic and abiotic variables is often problematic; for example, land cover type is likely to incorporate both abiotic (e.g. urban) and biotic (e.g. deciduous forest) classes.

Another important factor that we must be aware of is source-sink dynamics, which may cause a species to be observed in unsuitable environments. 'Source-sink' refers to the situation whereby an area (the 'sink') may not provide the necessary environmental conditions to support a viable popu-

lation, yet may be frequently visited by individuals that have dispersed from a nearby area that does support a viable population (the 'source'). In this situation, species occurrence may be recorded in sink areas that do not represent suitable habitat, meaning that the species is present outside its fundamental niche (Pulliam, 2000). We can logically expect this situation to occur most frequently in species with high dispersal ability, such as birds. In such cases, it is useful to only utilize records for modeling that are known to be from breeding distributions, rather than migrating individuals. Because correlative species distribution models utilize observed species occurrence records to identify suitable habitat, inclusion of occurrence localities from sink populations is problematic. However, it is often assumed that observations from source areas will be much more frequent than observations from sink areas, so this source of potential error is commonly overlooked.

One more thing to be aware of before we move on is that some studies explicitly aim to only investigate one part of the fundamental niche, by using a limited set of predictor variables. For example, it is common when investigating the potential impacts of future climate change to focus only on how climate variables impact species' distributions. A species' niche defined only in terms of climate variables may be termed the

Humpback whales have been recently sighted in unlikely waters around Hong Kong and New York Harbor (Source: Chad King)

climatic niche (Pearson and Dawson, 2003), which represents the climatic conditions that are suitable for species existence. An approximation of the climatic niche may then be mapped in geographical space, giving what is commonly termed the *bioclimate envelope* (Huntley et al., 1995; Pearson and Dawson, 2003).

Estimating niches and distributions

Let us now consider the extent to which species' distribution models can be used to estimate the niche and distribution of a species. We will assume in this section that the model algorithm is excellent at defining the relationship between observed occurrence localities and environmental variables; this will enable us to focus on understanding the ecological assumptions underlying distribution models. The ability of different modeling algorithms to identify the relationship between occurrence localities and environmental variables is discussed in the section: 'Modeling Algorithms'.

Let us first ask what the aim of the modeling is: what element of a species' distribution are we trying to estimate? There are many potential uses of the approach (Table 1) and these require modeling either the actual distribution or the potential distribution. For example, if a model is being used with the purpose of selecting sites that should be given high conservation priority, then modeling the actual distribution will be the aim (since there would be less priority given to conserving sites where the environment is suitable for the species, but the species is not present). In contrast, if the purpose is to identify sites that may be suitable for the reintroduction of an endangered species, then modeling the potential distribution is an appropriate aim. We will now consider the degree to which alternative aims are achievable using the species' distribution modeling approach.

Correlative species' distribution models rely on observed occurrence records for providing information on the niche and distribution of a species. Two key factors are important when considering the degree to which observed species occurrence records can be used to estimate the niche and distribution of a species:

Table 1. Some published uses of species' distribution models in conservation biology (based in part on Guisan and Thuiller, 2005).

Type of use	Example reference(s)
Guiding field surveys to find populations of known species	Bourg et al., 2005; Guisan et al., 2006
Guiding field surveys to accelerate the discovery of unknown species	Raxworthy et al., 2003
Projecting potential impacts of climate change	Iverson and Prasad, 1998; Berry et al., 2002; Hannah et al., 2005; for review see Pearson and Dawson, 2003
Predicting species' invasion	Higgins et al., 1999; Thuiller et al., 2005; for review see Peterson, 2003
Exploring speciation mechanisms	Kozak and Wiens, 2006; Graham et al., 2004b
Supporting conservation prioritization and reserve selection	Araújo and Williams, 2000; Ferrier et al., 2002; Leathwick et al., 2005
Species delimitation	Raxworthy et al., 2007
Assessing the impacts of land cover change on species' distributions	Pearson et al., 2004
Testing ecological theory	Graham et al., 2006; Anderson et al., 2002b
Comparing paleodistributions and phylogeography	Hugall et al., 2002
Guiding reintroduction of endangered species	Pearce and Lindenmayer, 1998
Assessing disease risk	Peterson et al., 2006, 2007

1) *The degree to which the species is at 'equilibrium' with current environmental conditions.*

A species is said to be at equilibrium with the physical environment if it occurs in all suitable areas, while being absent from all unsuitable areas. The degree of equilibrium depends both on biotic interactions (for example, competitive exclusion from an area) and dispersal ability (organisms with higher dispersal ability are expected to be closer to equilibrium than organisms with lower dispersal ability) (Araújo and Pearson, 2005). When using the concept of 'equilibrium,' we should remember that

species distributions change over time, so the term should not be used to imply stasis. However, the concept is useful for us here to help understand that some species are more likely than others to occupy areas that are abiotically suitable.

2) *The extent to which observed occurrence records provide a sample of the environmental space occupied by the species.*
    In cases where very few occurrence records are available, perhaps due to limited survey effort (Anderson and Martinez-Meyer, 2004) or low probability of detection (Pearson et al., 2007), the available records are unlikely to provide a sufficient sample to enable the full range of environmental conditions occupied by the species to be identified. In other cases, surveys may provide extensive occurrence records that provide an accurate picture as to the environments inhabited by a species in a particular region (for example, breeding bird distributions in the United Kingdom and Ireland are well known) (Gibbons et al., 1993). It should be noted that there is not necessarily a direct relationship between sampling in geographical space and in environmental space. It is quite possible that poor sampling in geographical space could still result in good sampling in environmental space.

Each of these factors should be carefully considered to ensure appropriate use of a species' distribution model (see Box 1). In reality, species are unlikely to be at equilibrium (as illustrated by area B in Fig. 2, which is environmentally suitable but is not part of the actual distribution) and occurrence records will not completely reflect the range of environments occupied by the species (illustrated by that part of the occupied niche that has not been sampled around label D in Fig. 2). Fig. 3 illustrates how a species' distribution model may be fit under these circumstances. Notice that the model is calibrated (i.e. built) in environmental space and then projected into geographical space. In environmental space, the model identifies neither the occupied niche nor the fundamental niche; instead, the model fits only to that portion of the niche that is represented by the observed records. Similarly, the model identifies only some parts of the actual and potential distribu-

tions when projected back into geographical space. Therefore, it should not be expected that species' distribution models are able to predict the full extent of either the actual distribution or the potential distribution.

This observation may be regarded as a failure of the modeling approach (Woodward and Beerling, 1997; Lawton, 2000; Hampe, 2004). However, we can identify three types of model predictions that yield important biogeographical information: species' distribution models may identify 1) the area around the observed occurrence records that is expected to be occupied (Fig. 3, area 1); 2) a part of the actual distribution that is currently unknown (Fig. 3, area 2); and/or 3) part of the potential distribution that is not occupied (Fig. 3, area 3). Prediction types 2 and 3 can prove very useful in a range of applications, as we will see in the following section.

Uses of species' distribution models
Consider modeled area 2 in Fig. 3, which identifies part of the actual distribution for which no occurrence records have been collected. Although the model does not predict the full extent of the actual distribution, additional sampling in the area identified may yield new occurrence records. A number of studies have demonstrated the utility of species' distribution modeling for guiding field surveys toward regions where there is an increased probability of finding new populations of a known species (Fleishman et al., 2002; Bourg et al., 2005; Guisan et al., 2006; also see Case Study 1). Accelerating the discovery of new populations in this way may prove extremely useful for conservation planning, especially in poorly known and highly threatened landscapes.

Consider now predicted area 3 in Fig. 3. Here, the model identifies an area of potential distribution that is environmentally similar to where the species is known to occur, but which is not inhabited. The full extent of the potential distribution is not predicted, but the model can be useful for identifying sites that may be suitable for reintroduction of a species (Pearce and Lindenmayer, 1998) or sites where a species is most likely to become invasive (if it overcomes dispersal barriers and if biotic competition does not prevent establishment) (Peterson,
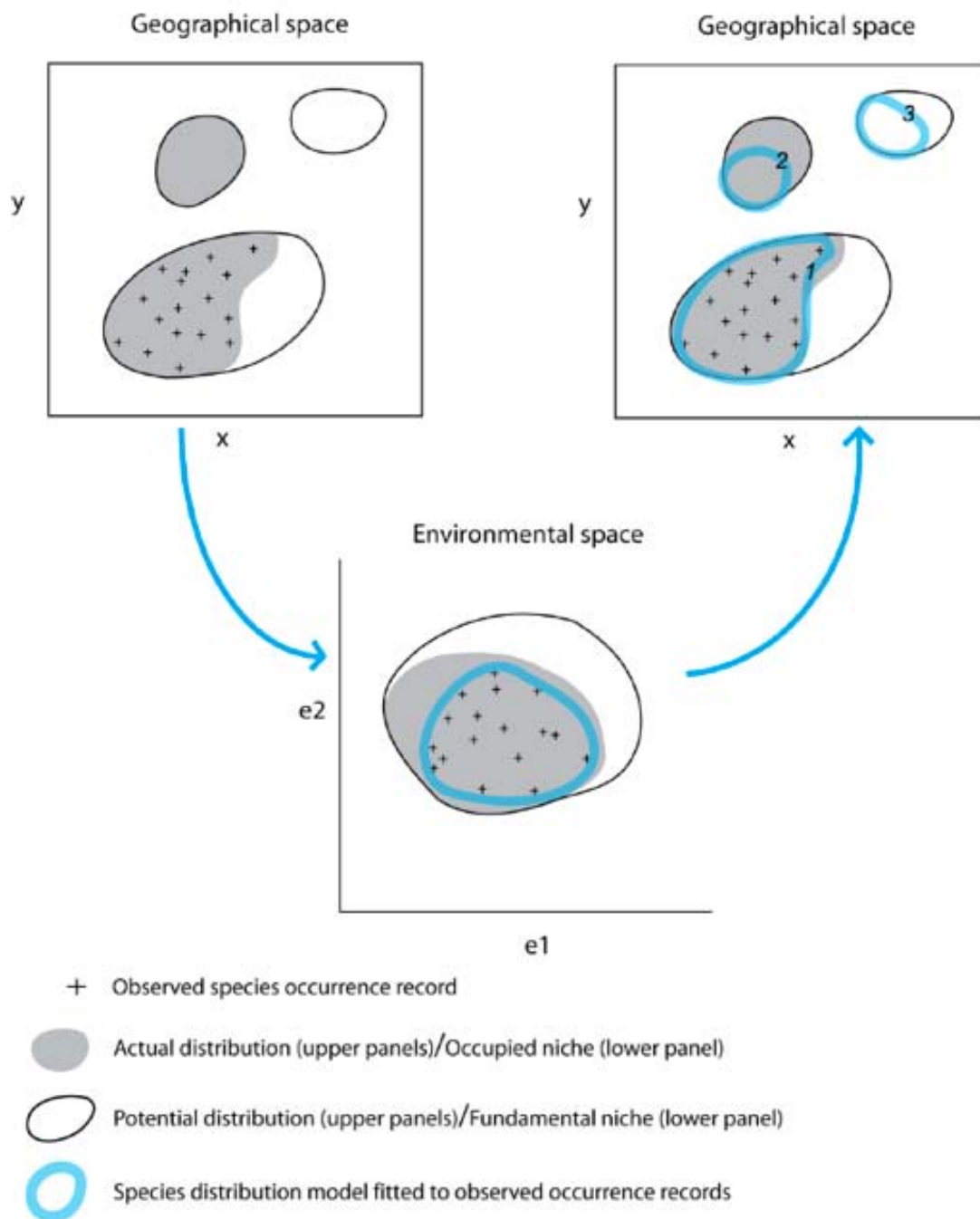
Figure 3. Diagram illustrating how a hypothetical species' distribution model may be fitted to observed species occurrence records (using the same hypothetical case as in Fig. 2). A modelling technique (e.g. GARP, Maxent) is used to characterize the species' niche in environmental space by relating observed occurrence localities to a suite of environmental variables. Notice that, in environmental space, the model may not identify either the species' occupied niche or fundamental niche; rather, the model identifies only that part of the niche defined by the observed records. When projected back into geographical space, the model will identify parts of the actual distribution and potential distribution. For example, the model projection labeled *1* identifies the known distributional area. Projected area *2* identifies part of the actual distribution that is currently unknown; however, a portion of the actual distribution is not predicted because the observed occurrence records do not identify the full extent of the occupied niche (i.e. there is incomplete sampling; see area *D* in Fig. *2*). Similarly, modeled area *3* identifies an area of potential distribution that is not inhabited (the full extent of the potential distribution is not identified because the observed occurrence records do not identify the full extent of the fundamental niche due to, for example, incomplete sampling, biotic interactions, or constraints on species dispersal; see areas *D* and *E* in Fig. 2).

2003). Model predictions of this type also have the potential to accelerate the discovery of previously unknown species that are closely related to the modeled species and that occupy similar environmental space but different geographical space (Raxworthy et al., 2003; see Case Study 1).

Model predictions as illustrated in Fig. 3 therefore have the potential to yield useful information, even though species are not expected to inhabit all suitable locations and sampling may be poor. Additional uses of species' distribution modeling include identifying potential areas for disease outbreaks (Peterson et al., 2006), examining niche evolution (Peterson et al., 1999; Kozak and Wiens, 2006) and informing taxonomy (Raxworthy et al., 2007). However, some potential applications require an estimation of the actual distribution of a species. For example, if a model is being used with the purpose of selecting priority sites for conservation, then an estimate of the actual species' distribution is desired since it would be inefficient to conserve sites where the species is not present (Loiselle et al., 2003). In such cases, it should be remembered that modeled distributions represent environmentally suitable regions but do not necessarily correspond closely with the actual distribution. Additional processing of model output may be required to improve predictions of the actual distribution. For example, predicted areas that are isolated from observed occurrence records by a dispersal barrier may be removed (Peterson et al., 2002) and the influence of competing species may be incorporated (Anderson et al., 2002b).

It is useful to note that mechanistic distribution models (e.g., Chuine and Beaubien, 2001) are subject to the same basic caveat as correlative approaches: the models aim to identify areas with suitable environmental conditions, but do not inform us which areas are actually occupied. Mechanistic models are ideally suited to identifying a species' fundamental niche, and hence its potential distribution. This is because mechanistic approaches model physiological limitations in a species' environmental tolerance, without relying on known occurrence records to define suitable environments. However, the detailed understanding of species' physiology that is required to build mechanistic models prohibits their use in many instances.

The discussion in this section should help clarify the theoretical basis of the species' distribution modeling approach. It is crucial that any application of these models has a sound theoretical basis and that model outputs are interpreted in the context of this framework (see Box 1). It should now be apparent why the terminology used to describe these models is so varied throughout the literature. The terms 'ecological niche model', 'environmental niche model', 'bioclimate envelope model' and 'environmental suitability model' usually refer to attempts to estimate the potential distribution of a species. Use of the term 'species distribution model' implies that the aim is to simulate the actual distribution of the species. Nevertheless, each of these terms refers to the same basic approach, which can be summarized as follows: 1) the study area is modeled as a raster map composed of grid cells at a specified resolution; 2) the dependent variable is the known species' distribution; 3) a suite of environmental variables are collated to characterize each cell; and 4) a function of the environmental variables is generated so as to classify the degree to which each cell is suitable for the species (Hirzel et al., 2002). Part B of this synthesis details the principal steps required to build a distribution model, including selecting and obtaining suitable data, choosing a modeling algorithm, and statistically assessing predictive performance.

Developing a Species' Distribution Model

Data Types and Sources

Correlative species' distribution models require two types of data input: *biological data*, describing the known species' distribution, and *environmental data*, describing the landscape in which the species is found. This section discusses the types of data that are suitable for distribution modeling, and reviews some possible sources of data (see Table 2).

Data used for distribution modeling are usually stored in a Geographical Information System (GIS; see Box 2). The data may be stored either as point localities (termed 'point vector' data; e.g. sites where a species has been observed, or locations of weather stations), as polygons defining an area (termed

---

Box 1. Caution! On the use and misuse of models

Garbage in, garbage out: This old adage is as relevant to distribution modeling as it is to other fields. Put simply, a model is only as good as the data it contains. Thus, if the occurrence records used to build a correlative species' distribution model do not provide useful information as to the environmental requirements of the species, then the model cannot provide useful output. If you put garbage into the model, you will get garbage out.

Model extrapolation: 'Extrapolation' refers to the use of a model to make predictions for areas with environmental values that are beyond the range of the data used to calibrate (i.e. develop) the model. For example, suppose a distribution model was calibrated using occurrence records that spanned a temperature range of 10–20°C. If the model is used to predict the species' distribution in a different region (or perhaps under a future climate scenario) where the temperature reaches 25°C, then the model is extrapolating. In this case, because the model has no prior information regarding the probability of the species' occurrence at 25°C, the prediction may be extremely uncertain (see Pearson et al., 2006). Model extrapolation should be treated with a great deal of caution.

The lure of complicated technology: Many approaches to modeling species' distributions utilize complex computational technology (e.g. machine learning tools such as artificial neural networks and genetic algorithms) along with huge GIS databases of digital environmental layers. In some cases, these approaches can yield highly successful predictions. However, there is a risk that model users will be swayed by the apparent complexity of the technology: "it is so complicated, it must be correct!" Always remember that a model can only be useful if the theoretical underpinnings on which it is based are sound.

For additional discussion of the limitations of ecological models, see the NCEP module: *Applications of Remote Sensing to Ecological Modeling.*

'polygon vector' data; e.g. areas with different soil types) or as a grid of cells (termed 'raster' data; e.g. land cover types derived from remote sensing [see NCEP module: *Remote Sensing for Conservation Biology*]). For use in a distribution model, it is usual to reformat all environmental data to a raster grid. For example, temperature records from weather stations may be interpolated to give continuous data over a grid (Hijmans et al., 2005). Formatting all data to the same raster grid ensures that environmental data are available for every cell in which biological data have been recorded. These cells, containing both biological data and environmental data, are used to build the species' distribution model. After the model is constructed, its fit to test occurrence records is evaluated (see section: 'Assessing Predictive Performance') and, if the fit is judged to be acceptable, the occurrence of species in cells for which only environmental data are available can be predicted. Note that in some applications the model may be used to predict the species' distribution in a region from which data were not used to build the model (e.g. to predict the spread of an inva-

sive species) or under a future climate scenario. In these cases, environmental data for the new region or climate scenario must also be collated.

A consideration when collating data is the spatial scale at which the model will operate. Spatial scale has two components: *extent* and *resolution* (see NCEP module *Applications of remote sensing to ecological modeling*). Extent refers to the size of the region over which the model is run (e.g. New York state or the whole of North America) whilst resolution refers to the size of grid cells (e.g. 1 km^2 or 10 km^2). Note that it is common for datasets with large extent to have coarse resolution (e.g. data for North America at 10 km^2) and datasets with small extent to have fine resolution (e.g. New York state at 1 km^2). Spatial scale can play an important role in the application of a species' distribution model. In particular, ideally the data resolution should be relevant to the species under consideration: the appropriate data resolution for studying ants is likely to be very different from that for studying elephants.

Biological data

Data describing the known distribution of a species may be obtained in a variety of ways:

1) *Personal collection*: occurrence localities can be obtained during field surveys by an individual or small group of researchers. For example, Fleishman et al. (2001) built models using butterfly occurrence records collected by the researchers during surveys in Nevada, USA.

2) *Large surveys*: distribution information may be available from surveys undertaken by a large number of people. For example, Araújo et al. (2005a) built distribution models using data from *The New Atlas of Breeding Birds in Britain and Ireland: 1988-1991* (Gibbons et al., 1993), which represents the sampling effort of hundreds of volunteers.

3) *Museum collections:* occurrence localities can be obtained from collections in natural history museums. For example, Raxworthy et al. (2003) utilized occurrence records of chameleons in Madagascar that are held in museum collections.

4) *Online resources:* distributional data from a variety of sources are increasing being made available over the internet (see Table 2). For example, the Global Biodiversity Information Facility (www.gbif.org) is collating available datasets from a diversity of sources and making the infor-

mation available online via a searchable web portal.

Species distribution data may be either *presence-only* (i.e. records of localities where the species has been observed) or *presence/absence* (i.e. records of presence and absence of the species at sampled localities). Different modeling approaches have been developed to deal with each of these cases (see section on: 'Modeling Algorithms'). In some instances, the inclusion of absence records has been shown to improve model performance (Brotons et al., 2004). However, absence records are often not available and may be unreliable in some cases. In particular, absences may be recorded when the species was not detected even though the environment was suitable. These cases are often referred to as 'false absences' because the model will interpret the record as denoting unsuitable environmental conditions, even though this is not the case. False absences can occur when a species could not be detected although it was present, or when the species was absent but the environment was in fact suitable (e.g. due to dispersal limitation or metapopulation dynamics). Inclusion of false absence records may seriously bias analyses, so absence data should be used with care (Hirzel et al., 2002).

There are a number of additional potential sources of bias and error that should be carefully considered when collating spe-

---

Box 2. The Role of Geographic Information Systems (GIS)

GIS is a vital tool in species' distribution modeling. The large datasets of biological and environmental data that are used in distribution modeling are ideally suited to being stored, viewed, and formatted in a GIS. For example, useful GIS operations include changing geographic reference systems (it is essential that all data are referenced to a common coordinate system, so occurrence records can be matched with the environmental conditions at the site), reformatting spatial resolution, and interpolating point locality data to a grid. GIS is also crucial for visualizing model results and carrying out additional processing of model output, such as removing predicted areas that are isolated from observed species records by a dispersal barrier (Peterson et al., 2002). However, the distribution modeling itself is usually undertaken outside the GIS framework. With few exceptions (e.g. Ferrier et al., 2002), the distribution model does not 'see' geographical coordinates; instead, the model operates in environmental space (see section: 'Theoretical Framework'). Some GIS platforms now incorporate distribution modeling tools (e.g. DIVA GIS: www.diva-gis.org/, IDRISI: http://www.clarklabs.org/), or have add-in scripts that enable distribution models to be run (e.g. BIOCLIM script for ArcView: http://arcscripts.esri.com/details.asp?dbid=13745), but running the model within a GIS is not necessary.

Table 2. Some example sources of biological and environmental data for use in species' distribution modeling

Type of data	Source
Species' distributions	
– Data for a wide range of organisms in many regions of the world	Global Biodiversity Information Facility (GBIF): www.gbif.org
– Data for a range of organisms, mostly rare or endangered, and primarily in North America	NatureServe: www.NatureServe.org
Climate	
– Interpolated climate surfaces for the globe at 1km resolution	WorldClim:http://www.worldclim.org/
– Scenarios of future climate change for the globe	Intergovernmental Panel on Climate Change (IPCC): http://www.ipcc–data.org/
– Reconstructed palaeoclimates	NOAA: http://www.ncdc.noaa.gov/paleo/paleo.html
Topography	
– Elevation and related variables for the globe at 1km resolution	USGS: http://edc.usgs.gov/products/elevation/gtopo30/hydro/index.html
Remote sensing (satellite)	
– Various land cover datasets	Global Landcover Facility: http://glcf.umiacs.umd.edu/data/
– Various atmospheric and land products from the MODIS instrument	NASA:http://modis.gsfc.nasa.gov/data/
Soils	
– Global soil types	UNEP: http://www.grid.unep.ch/data/data.php?category=lithosphere
Marine	
– Various datasets describing the world's oceans	NOAA: www.nodc.noaa.gov

cies' distribution data. Errors may arise through the incorrect identification of species, or inaccurate spatial referencing of samples. Biases can also be introduced because collectors tend to sample in easily accessible locations, such as along roads and rivers and near towns or biological stations (Graham et al., 2004a). In some cases, biased sampling in geographical space may lead to non–representative sampling of the available environmental conditions, although this is not necessarily the case. When utilizing records from museum collections, it should be remembered that these data were not generally collected with the purpose of determining the distributional limits of a species; rather, sampling for museum collections tends to be biased toward rare and previously unknown species.

Environmental data

A wide range of environmental input variables have been employed in species' distribution modeling. Most common are variables relating to climate (e.g. temperature, precipitation), topography (e.g., elevation, aspect), soil type and land cover type (see Table 2). Variables tend to describe primarily the abiotic environment, although there is potential to include biotic interactions within the modeling. For example, Heikkinen et al. (2007) used the distribution of woodpecker species to predict owl distributions in Finland since woodpeckers excavate cavities in trees that provide nesting sites for owls.

As noted in the previous section: 'Introduction' variables are often processed to generate new variables that are thought to

have a direct physiological or behavioral role in determining species' distributions. In general, it is advisable to avoid predictor variables that have an indirect influence on species' distributions, since indirect associations may cause erroneous predictions when models are used to predict the species' distribution in new regions or under alternative climate scenarios (Guisan and Thuiller, 2005). For example, species do not respond directly to elevation, but rather to changes in temperature and air pressure that are affected by elevation. Thus, a species characterized as living at high elevation in a low latitude region may, in fact, be associated with lower elevations in areas with higher latitude, since the regional climate is cooler.

Environmental variables may comprise either *continuous data* (data that can take any value within a certain range, such as temperature or precipitation) or *categorical data* (data that are split into discrete categories, such as land cover type or soil type). Categorical data cannot be used with a number of common modeling algorithms (see section: 'Modeling Algorithms'). In these cases, it may be possible to generate a continuous variable from the categorical data. For example, Pearson et al. (2002) estimated soil water holding capacity from categorical soil data, and used these values within a water balance model to generate continuous predictions of soil moisture surplus and deficit.

Modern technologies, including remote sensing (see NCEP module *Applications of remote sensing to ecological modeling*), the internet, and GIS have greatly facilitated the collection and dissemination of environmental datasets (see Table 2). In addition, global climate models have been used to generate scenarios of future climates and to simulate climatic conditions since the end of the last glacial period (see Table 2). Predicted future climate scenarios can be used to estimate the potential impacts of climate change on biodiversity (e.g. Thomas et al. 2004; see case study 3), whilst simulations of past climates can be used to test the predictive ability of models (e.g. Martinez-Meyer et al., 2004). Given the vast amounts of data that are available, it is especially important to remain critical as to which variables are suitable for inclusion in the model. Some studies have demonstrated good predictive ability using only

three variables (e.g., Huntley et al., 1995), whilst other studies have applied methodologies that can incorporate many more variables (e.g., Phillips et al., [2006] utilized 14 environmental variables, although some of these variables are likely to have been rejected by the algorithm applied because they did not provide useful information beyond that which was included in other variables).

Modeling Algorithms

A number of alternative modeling algorithms have been applied to classify the probability of species' presence (and absence) as a function of a set of environmental variables. The task is to identify potentially complex non-linear relationships in multi-dimensional environmental space. In the section: 'Theoretical Frameworks', we assumed that the modeling algorithm is excellent, enabling us to identify three expected types of model prediction, illustrated by modeled areas *1, 2* and *3* in Fig. 3. If we now admit some degree of error in the algorithm's ability to fit the observed records, then a fourth type of prediction will occur: the model will predict as suitable areas that are part of neither the actual nor the potential distribution. The most useful algorithms will limit these erroneous 'type 4' predictions.

Table 3 lists some commonly used approaches for species' distribution modeling. Some methods that have been applied are statistical (e.g. generalized linear models [GLMs] and generalized additive models [GAMs]), whilst other approaches are based on machine-learning techniques (e.g., maximum entropy [Maxent] and artificial neural networks [ANNs]). Published studies have often applied one or more of these algorithms and have given the resulting model a name or acronym (e.g., 'Maxent' refers to an implementation of the maximum entropy method, while 'BIOMOD' is the acronym given to a model that implements a number of methods, including GLMs and GAMs). Often these models have been implemented in user-friendly software that is free and easy to obtain (Table 3).

There are some important differences between among model

algorithms that should be carefully considered when selecting which method(s) to apply. One key factor is whether the algorithm requires data on observed species absence (see 'Data Types and Sources'). Some algorithms operate by contrasting sites where the species has been detected with sites where the species has been recorded as absent (e.g., GLMs, GAMs, ANNs). However, reliable absence data often are not available

(see: 'Data Types and Sources'), so other methods have been applied that do not require absence data. We can distinguish three types of presence-only methods:

1) Methods that rely solely on presence records (e.g. BIO-CLIM, DOMAIN). These methods are truly 'presence-only' since the prediction is made without any reference

Table 3. Some published methods for species' distribution modeling			
Method(s)[1]	**Model/soft-ware name[2]**	**Species data type**	**Key reference/URL**
Gower Metric	DOMAIN★	presence-only	Carpenter et al., 1993 http://www.cifor.cgiar.org/docs/_ref/research_tools/domain/ http://diva-gis.org
Ecological Niche Factor Analysis (ENFA)	BIOMAPPER★	presence and background	Hirzel et al., 2002 http://www2.unil.ch/biomapper/
Maximum Entropy	MAXENT★	presence and background	Phillips et al., 2006 http://www.cs.princeton.edu/~schapire/maxent/
Genetic algorithm (GA)	GARP[3]★	pseudo-absence[4]	Stockwell and Peters, 1999 http://www.lifemapper.org/desktopgarp/
Artificial Neural Network (ANN)	SPECIES	presence and absence (or pseudo-absence)	Pearson et al., 2002
Regression: generalized linear model (GLM), generalized additive model (GAM), boosted regression trees (BRT), multivariate adaptive regression splines (MARS)	Implemented in R[5]	presence and absence (or pseudo-absence)	Lehman et al., 2002 Elith et al., 2006 Leathwick et al., 2006 Elith et al., 2007
Multiple methods	BIOMOD	presence and absence (or pseudo-absence)	Thuiller, 2003
Multiple methods	OpenModeller	depends on method implemented	http://openmodeller.sourceforge.net/

[1] 'Method' refers to a statistical or machine-learning technique. [2]'Model/software name' refers to a name (or acronym) given to a published model that implements the method(s) stated. Software to implement the method for species' distribution modeling is readily available at no cost for those models marked with an asterisk (★); other models are available at the discretion of the author(s). [3]The genetic algorithm for rule-set prediction (GARP) includes within its processing multiple methods, including GLM. [4]Note that Pseudo-absence here refers to the sampling approach implemented in the GARP software; in principle, any presence-absence method can be implemented using pseudo absences. [5]R is a freely available (at no cost) software environment for statistical computing and graphics (http://www.r-project.org/). Based in part on Elith et al. (2006) and Guisan and Thuiller (2005).

to other samples from the study area.

2) Methods that use 'background' environmental data for the entire study area (e.g. Maxent, ENFA). These methods focus on how the environment where the species is known to occur relates to the environment across the rest of the study area (the 'background'). An important point is that the occurrence localities are also included as part of the background.

3) Methods that sample 'pseudo-absences' from the study area. In principle, any presence/absence algorithm can be implemented using pseudo-absences. The aim here is to assess differences between the occurrence localities and a set of localities chosen from the study area that are used in place of real absence data. The set of 'pseudo-absences' may be selected randomly (e.g., Stockwell and Peters, 1999) or according to a set of weighting criteria (e.g., Engler et al., 2004; Zaniewski et al., 2002). An important difference between the pseudo-absence approach and the background approach is that pseudo-absence models do not include occurrence localities within the set of pseudo-absences.

Another key difference among model algorithms is their ability to incorporate categorical environmental variables (see the section: Data Types and Sources'). Methods also differ in the form of their output, which is most commonly a continuous prediction (e.g. a probability value ranging from 0 to 1)

but may be a binary prediction (with '0' as a prediction of unsuitable environmental conditions or species absence, and '1' a prediction of highly suitable environmental conditions or species presence). To generate a binary prediction from a model that gives continuous output, it is necessary to set a threshold value above which the prediction is classified as 'highly suitable' or 'present' (see the section: 'Assessing Predictive Performance').

A further consideration when selecting a modeling algorithm is whether it is important to determine the relative influence of different input variables on the model's fit or predictive capacity. Some models may have excellent predictive power but do not enable us to easily understand how the algorithm is operating; such models are often termed 'black box' since the model takes input and produces output but the internal workings are somewhat opaque. For example, artificial neural networks have shown good predictive ability (e.g., Pearson et al., 2002; Thuiller, 2003; Segurado and Araújo, 2004), but identifying the relative contribution of each input variable to the prediction is difficult (sensitivity analysis may be used, but this requires additional analyses). In contrast, a GLM builds a regression equation from which the relative contributions of different variables are immediately apparent (Guisan et al., 2002).

It is not possible within the scope of this synthesis to describe

---

Box 3. Maximum Entropy (Maxent) Modeling of Species Distributions (based on Phillips et al., 2006)

Maxent is a general-purpose method for characterizing probability distributions from incomplete information. In estimating the probability distribution defining a species' distribution across a study area, Maxent formalizes the principle that the estimated distribution must agree with everything that is known (or inferred from the environmental conditions where the species has been observed) but should avoid making any assumptions that are not supported by the data. The approach is thus to find the probability distribution of maximum entropy (the distribution that is most spread-out, or closest to uniform) subject to constraints imposed by the information available regarding the observed distribution of the species and environmental conditions across the study area.

The Maxent method does not require absence data for the species being modeled; instead it uses background environmental data for the entire study area. The method can utilize both continuous and categorical variables

---

Box 3. Maximum Entropy (Maxent) Modeling of Species Distributions (Continued)

and the output is a continuous prediction (either a raw probability or, more commonly, a cumulative probability ranging from 0 to 100 that indicates relative suitability). Maxent has been shown to perform well in comparison with alternative methods (Elith et al., 2006; Phillips et al., 2006; Pearson et al., 2007). One drawback of the Maxent approach is that it uses an exponential model that can predict high suitability for environmental conditions that are outside the range present in the study area (i.e. extrapolation, see Box 1). To alleviate this problem, when predicting for variable values that are outside the range found in the study area, these values are reset (or 'clamped') to match the upper or lower values found in the study area.

   For a concise mathematic definition of Maxent and for more detailed discussion of its application to species distribution modeling see Phillips et al. (2004, 2006). These authors have developed software with a user-friendly interface to implement the Maxent method for modeling species distributions (for free download see web link in Table 3). The software also calculates a number of alternative thresholds (see Section 5), computes model validation statistics (see section: 'Assessing Predictive Performance'), and enables the user to run a jackknife procedure to determine which environmental variables contribute most to the model prediction (see the practical exercise by Steven Phillips that accompanies this synthesis).

---

the theory, advantages and disadvantages of a large number of modeling algorithms; the reader is referred to the literature cited in Table 3. However, see Box 3 and the practical exercise by Steven Phillips that accompanies this Synthesis for a more detailed description of one method, Maxent.

The model algorithm is in some ways the 'core' of the distribution model, but it should be remembered that the algorithm is just one part of the broader modeling process; other factors, including selection of environmental variables (see section: 'Data Types and Sources') and application of a decision threshold (see section: 'Assessing Predictive Performance'), are key elements of the modeling process that affect model results and may be varied regardless of the model algorithm being used. Nevertheless, studies comparing different modeling algorithms have demonstrated substantial differences between predictions from alternative methods. The importance of selecting an appropriate algorithm is discussed below.

Differences between methods and selection of 'best' models
Given the variety of possible modeling methods (Table 3), it is important to consider the degree to which different methods yield different results. Furthermore, if model predictions differ substantially, how should we choose which method to apply? This is an active area of research, and unfortunately there are

no simple answers.

A number of studies have demonstrated that different modeling approaches have the potential to yield substantially different predictions (e.g., Loiselle et al., 2003; Thuiller, 2003; Brotons et al., 2004; Segurado and Araújo, 2004; Thuiller et al., 2004; Elith et al., 2006; Pearson et al., 2006). Pearson et al. (2006) found especially large differences among predictions of changes in range size under future climate change scenarios based on nine alternative modeling methods. Predicted changes in range size differed in both magnitude and direction (e.g. from 92% range reduction to 322% range increase for a single species). In another study, Loiselle et al. (2003) demonstrated markedly different results when alternative distribution models were used alongside a reserve selection algorithm for identifying priority sites for conservation.

The most comprehensive model comparison to date was provided by Elith et al. (2006). The authors compared sixteen modeling methods using 226 species across six regions of the world. All of the models included in the study were implemented using presence-only data for calibration (some methods required the use of pseudo-absence data), but model performance was assessed using data on both presence and absence. These analyses found differences between predictions

---

from alternative methods, but also found that some methods consistently outperformed others. In general, models classified as 'best' were those that were able to identify complex relationships that existed in the data, including interactions among environmental variables.

Several additional factors that lead to differences among predictions from alternative algorithms have been identified. These include (1) whether the model uses presence-absence or presence-only data (Brotons et al., 2004; Pearson et al., 2006), (2) if the model does not use absence data, whether the model uses solely presence records, 'background' data, or 'pseudo-absences' (Elith et al., 2006), (3) whether the algorithm is parametric or non-parametric (Segurado and Araújo, 2004), and (4) how the model 'extrapolates' beyond the range of data used for its calibration (Pearson et al., 2006; and see Box 1).

In view of these differences among models, selection of an appropriate algorithm is both difficult and crucial. Identifying models that are generically 'best' is problematic since the approach used to assess predictive performance depends on the aim of the modeling. For example, Elith et al. (2006) assessed the ability of models to simulate *actual* distributions by using statistical tests that reward models for correctly classifying both presences and absences (see Section 5). In contrast, Pearson et al. (2007) assessed predictive performance based only on the model's ability to predict observed presences, arguing that the purpose of the modeling was to identify *potential* distributions (in which case use of absence data in assessing performance is invalid since a site classified as absent still may be environmentally suitable). The relative merits of aiming to predict actual versus potential distributions were discussed in Section 2. We will also return to the question of how to identify 'best' models when describing various statistical approaches for assessing predictive performance in Section 5. However, the important point is that it is not straightforward to identify which methods are best, and it is therefore not possible to recommend use of one method over another.

In practice, model selection will be influenced by factors in-

cluding whether observed absence data are available, whether data on some of the environmental variables are categorical, and whether it is important to evaluate the influence of different variables on the model prediction. We recommend that modeling efforts apply and examine predictions from a range of methods in order to quantify uncertainty arising from the choice of method and to identify when different models are in agreement. Perhaps most importantly, it is vital that the assumptions and behavior of any model are properly understood (e.g. how does the model deal with environmental conditions that are beyond the range of the data used to calibrate the model?) so that model output can be accurately interpreted.

Assessing Predictive Performance

Assessing the accuracy of a model's predictions is commonly termed 'validation' or 'evaluation', and is a vital step in model development. Application of the model will have little merit if we have not assessed the accuracy of its predictions. Validation thus enables us to determine the suitability of a model for a specific application and to compare different modeling methods (Pearce and Ferrier, 2000). This section discusses different approaches for assessing predictive performance, including strategies for obtaining data against which the predictions can be compared, methods for selecting thresholds of occurrence, and various test statistics. As in previous sections, there is no single approach that can be recommended for use in all modeling exercises; rather, the choice of validation strategy will be influenced by the aim of the modeling effort, the types of data available, and the modeling method used.

Strategies for obtaining test data
In order to test predictive performance it is necessary to have data against which the model predictions can be compared. We can refer to these as *test data* (sometimes called *evaluation data*) to distinguish them from the *calibration data* (sometimes called *training data*) that are used to build the model. It is fairly common for studies to assess predictive performance by simply testing the ability of the model to predict the calibration data (i.e. calibration and test datasets are identical). However,

this approach makes it difficult to identify models that have over-fit the calibration data (meaning the model is able to accurately classify the calibration data, but the model performs poorly when predicting test data), making it impossible for users to judge how well the model may perform when making predictions (Araújo et al., 2005a). It is therefore preferable to use test data that are different from the calibration data.

Ideally, test data would be collected independently from the data used to calibrate the model. For example, Fleishman et al. (2002) modeled the occurrence of butterfly species in Nevada, USA, using species inventory data collected during the period 1996-1999, and then tested the models using data collected from new sites during 2000-2001 (see Case Study 1 for a comparable study). Other researchers have undertaken validation using independent data from different regions (e.g. Beerling et al., 1995; Peterson, 2003), data at different spatial resolution (e.g., Pearson et al., 2004; Araújo et al., 2005b), data from different time periods (e.g., Araújo et al., 2005a), and data from surveys conducted by other researchers (Elith et al., 2006).

However, in practice it may not be possible to obtain independent test data and it is, therefore, common to partition the available data into calibration and test datasets. Several strategies are available for partitioning data, the simplest being a one-time split in which the available data are assigned to calibration and test datasets either randomly (e.g., Pearson et al., 2002) or by dividing the data spatially (e.g., Peterson and Shaw, 2003). The relative proportions of data included in each data set are somewhat arbitrary, and dependent on the total number of locality points available (though using 70% for calibration and 30% for testing is common, following guidelines provided by Huberty (1994)). An alternative to a one-time split is 'bootstrapping', whereby the data are split multiple times. Bootstrapping methods sample the original set of data randomly with replacement (i.e. the same occurrence record could be included in the test data more than once). Multiple models are thus built, and in each case predictive performance is assessed against the corresponding test data. Validation statistics can then be reported as the mean and range from the set

of bootstrap samples (e.g., Verbyla and Litaitis, 1989; Buckland and Elston, 1993). An approach similar to bootstrapping, but sampling without replacement (i.e. the same occurrence record cannot be included in the test data more than once), can also be applied and may be termed 'randomization' (Fielding and Bell, 1997).

Another useful data partitioning method is k-fold partitioning. Here, data are split into k parts of roughly equal size ($k > 2$) and each part is used as a test set with the other k-1 sets used for model calibration. Thus, if we select $k = 5$ then five models will be calibrated and each model tested against the excluded test data. Validation statistics are then reported as the mean and range from the set of k tests (Fielding and Bell, 1997). An extreme form of k-fold partitioning, with k equal to the number of occurrence localities, is recommended for use with very low sample sizes (e.g., < 20; Pearson et al., 2007). This method is termed 'jackknifing' or 'leave-one-out' since each occurrence locality is excluded from model calibration during one partition.

The following sub-sections describe validation statistics that can be calculated after test data have been obtained using one of the above approaches.

The presence/absence confusion matrix
If a model is used to predict a set of test data, predictive performance can be summarized in a *confusion matrix*. Note that binary model predictions (i.e. predictions of suitable and unsuitable, rather than probabilities; see section 4) are required in order to complete the confusion matrix. Later subsections describe methods for converting continuous model outputs into binary predictions (see *Selecting thresholds of occurrence*) and for assessing predictive performance using continuous predictions (see *Threshold-independent assessment*). However, in order to understand these later sections it is important to first look at the confusion matrix.

The confusion matrix is rather more straightforward than its name suggests, and is alternatively termed an 'error matrix' or a 'contingency table'. The confusion matrix records the fre-
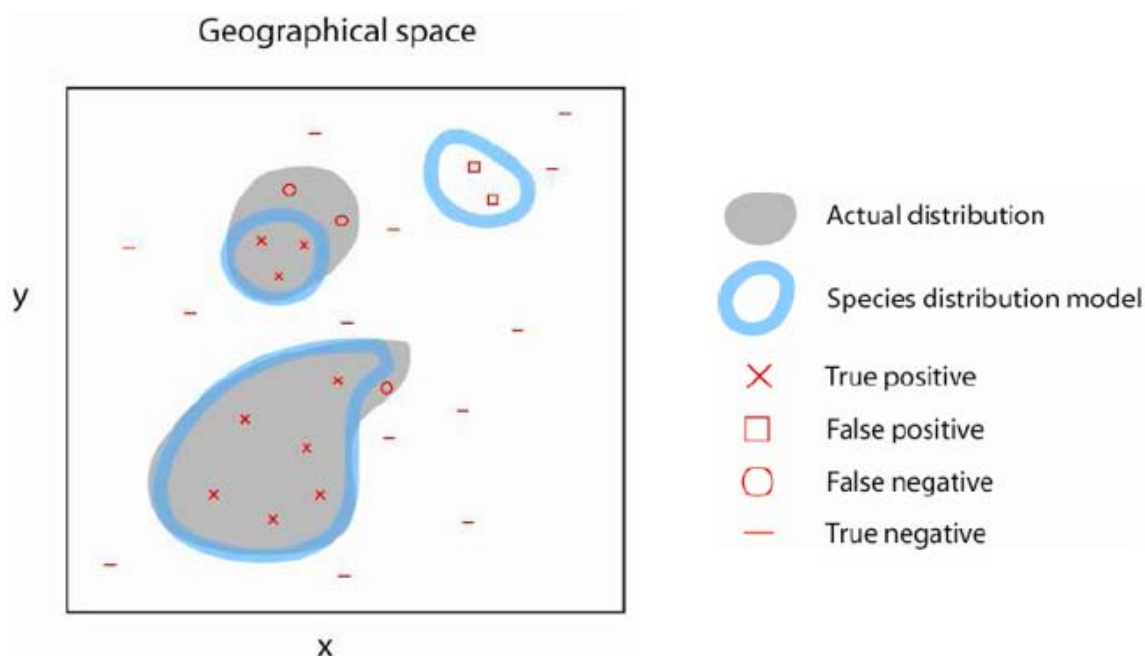
Geographical space

Figure 4. Diagram illustrating the four types of outcomes that are possible when assessing the predictive performance of a species distribution model: true positive, false positive, false negative and true negative. The diagram uses the same hypothetical actual and modeled distributions as in Figure 3. Each instance of a symbol (x, □, o, -) on the map depicts a site that has been surveyed and presence or absence of the species recorded (it is assumed here that if a site falls within the actual distribution then the species will be detected). These survey records constitute the test data. Frequencies of each type of outcome are commonly entered into a confusion matrix (see main text).

quencies of each of the four possible types of prediction from analysis of test data: (a) true positive (the model predicts that the species is present and test data confirms this to be true), (b) false positive (the model predicts presence but test data show absence), (c) false negative (the model predicts absence but test data show presence), (d) true negative (the model predicts and the test data show absence). Frequencies are commonly recorded in a confusion matrix with the following form:

	recorded present	recorded absent
predicted present	*a (true positive)*	*b (false positive)*
predicted absent	*c (false negative)*	*d (true negative)*

Each element of the confusion matrix can be visualized in geographical space as illustrated in Figure 4. In the example depicted, 27 test localities have been sampled and presence or absence of the species recorded at each site. The use of

test data comprising only presence localities (i.e. species occurrence records) is discussed below, but for completion of the confusion matrix we require both presence and absence records. Thus, the hypothetical case shown in Figure 4 would yield the following confusion matrix:

	recorded present	recorded absent
predicted present	*9*	*2*
predicted absent	*3*	*13*

The frequencies in the confusion matrix form the basis for a variety of different statistical tests that can be used to assess model performance. Most of the commonly used tests are described below. Terminology related to model performance often varies from study to study and is sometimes not intuitive. In particular, false negative predictions are commonly termed errors of 'omission', whilst false positive predictions

are termed errors of 'commission'.

Test statistics derived from the confusion matrix
A simple measure of predictive performance that can be derived from the confusion matrix is the proportion (or percentage) of test localities that are correctly predicted, calculated as

$$(a+d)/(a+b+c+d)$$

This measure may be termed 'accuracy' or 'correct classification rate'. The concept of accuracy is simple and logical, but it is possible to obtain high accuracy using a poor model when a species' prevalence (the proportion of sampled sites in which the species is recorded present) is relatively high or low. For example, if prevalence is 5% then 95% of test localities can be correctly classified simply by predicting all sites as 'absent'. To circumvent this problem, Cohen (1960) introduced a measure of accuracy that is adjusted to account for chance agreement between predicted and observed values. The statistic, Kappa (k), is similar to accuracy but the proportion of correct predictions expected by chance is taken into account (for full derivation see Monserud and Leemans, 1992). Kappa is calculated as:

$$\frac{[(a+d)-(((a+c)(a+b)+(b+d)(c+d))/n)]}{[n-(((a+c)(a+b)+(b+d)(c+d)/n]}$$

Accuracy and Kappa statistics use all values in the confusion matrix and therefore require both presence and absence data. However, absence data are often unavailable (e.g. when using specimens from museum collections) and are inappropriate for use when the aim is to estimate the *potential distribution* (since the environment may be suitable even though the species is absent).

When only presence records are used, the proportion of observed occurrences correctly predicted can be calculated as:

$$a/(a+c)$$

This measure is sometimes termed 'sensitivity' or 'true positive fraction'. Alternatively, we may calculate:

$$c/(a+c)$$

which is often termed 'omission rate' or 'false positive fraction'. Note that these two measures – sensitivity and omission rate – sum to 1. Thus, high sensitivity means low omission, and low sensitivity means high omission. Although sensitivity and omission rate avoid the use of absence records, a serious disadvantage of these tests is that it is possible to achieve very high sensitivity (and low omission) simply by predicting that the species is present at an excessively large proportion of the study area. In short, it is possible to cheat: if the model predicts the entire study area to be suitable, then sensitivity will equal 1, and omission rate will be zero. To avoid this problem, it is necessary to test the statistical significance of a sensitivity or omission rate score.

To test for statistical significance, we ask whether the accuracy of our predictions is greater than would be expected by chance. Imagine, for example, that we are blindfolded and asked to throw darts at a map of our study area. The sites identified by our random throws are then used as species' occurrence localities for model testing. The probability of landing darts in the area predicted by the model to be suitable for the species is equal to the proportion of the study area that is predicted as suitable. Thus, if the model predicts that the species will be present in 40% of the study area, then our probability of successfully landing a dart in the area predicted as suitable is 0.4.

We can apply the same logic to assess the statistical significance of a sensitivity (or omission rate) score. In this case, we use an exact one-tailed binomial test (or for larger sample sizes a chi-square test; for description of binomial and chi-square tests see Zar, 1996) to calculate the probability of obtaining a sensitivity result by chance alone (Anderson et al., 2002a). For example, suppose that the model in Figure 4 predicts that 30% of the study area is suitable for the species. The probability of success by chance alone for each test locality is therefore

0.3. We can calculate from the confusion matrix that the sensitivity $= 9/(9+3) = 0.75$, and we can use an exact one-tailed binomial test to calculate that the probability of making nine or more successful predictions of presence by chance alone is 0.0017. We may therefore conclude that our result is statistically significant ($p < 0.01$).

A similar assessment of predictive performance can be conducted when only very few occurrence localities are available and test data have been generated using a jackknifing approach (see subsection *Strategies for obtaining test data*). In this case the number of successful predictions from a set of jackknife trials can be calculated (e.g., 9 successes in 12 jackknife trials) and a *p*-value can be calculated using the method presented in Pearson et al. (2007).

In practice, binomial and chi-square tests can be performed in most standard statistical packages, whilst the jackknife *p*-value can be calculated using software provided as Supplementary Material to the Pearson et al. (2007) paper. Various test statistics and thresholds (including those discussed in the remainder of this section) are sometimes calculated automatically by software designed for species distribution modeling (see Table 3), and test statistics may also be calculated using more general applications such as DIVA-GIS (for free download see http://www.diva-gis.org/).

Another statistic that can be derived from the confusion matrix is the proportion of observed absences that are correctly predicted, calculated as:

$$d/(b+d)$$

This statistic is commonly termed 'specificity' or 'true negative fraction'. Specificity is rarely used as a test statistic on its own, since specificity focuses solely on observed absence records. However, specificity is an important measure used in setting decision thresholds and in ROC analysis, which are described in the following two subsections.

Selecting thresholds of occurrence

Binary predictions of 'present' or 'absent' are necessary to test model performance using statistics derived from the confusion matrix. It is therefore often useful to convert continuous model output into binary predictions by setting a threshold probability value above which the species is predicted to be present. Although an alternative test statistic that does not require a threshold is available (Area Under the ROC curve; see the following subsection), this approach is not suitable in many circumstances (notably when absence records are not available, Boyce et al., 2002, but see Phillips et al., 2006). Furthermore, it is essential to learn the techniques described in this subsection to understand how the AUC test operates.

A number of different methods have been employed for selecting thresholds of occurrence (Table 4). Perhaps the simplest approach is to use an arbitrary value, but this method is subjective and lacks ecological reasoning (Liu et al., 2005). Other methods use criteria that are based on the data used to calibrate the model. One approach is to use the lowest predicted value of environmental suitability, or probability of presence, across the set of sites at which a species has been detected. This method assumes that species presence is restricted to locations equally or more suitable than those at which the species has been observed. The approach therefore identifies the minimum area in which the species occurs whilst ensuring that no localities at which the species has been observed are omitted (i.e. omission rate $= 0$, and sensitivity $= 1$). An alternative approach is to set the threshold to allow a certain amount of omission (e.g. 5%), which is analogous to setting a fixed sensitivity (e.g. 0.95). This method is less sensitive than the lowest predicted value method to 'outliers' (i.e. locations in which the species is detected despite a low predicted probability of occurrence or suitability), but errors of omission are imposed (i.e. some observed localities will be omitted from the prediction).

Many methods for setting thresholds can be implemented by calculating statistics derived from the confusion matrix across the range of possible thresholds. For example, sensitivity and specificity may be calculated at thresholds increasing in incre-

Table 4. Some published methods for setting thresholds of occurrence

Method	Definition	Species data type[1]	Reference(s)
Fixed value	An arbitrary fixed value (e.g. probability = 0.5)	presence-only	Manel et al., 1999 Robertson et al., 2001
Lowest predicted value	The lowest predicted value corresponding with an observed occurrence record	presence-only	Pearson et al., 2006 Phillips et al., 2006
Fixed sensitivity	The threshold at which an arbitrary fixed sensitivity is reached (e.g. 0.95, meaning that 95% of observed localities will be included in the prediction)	presence-only	Pearson et al., 2004
Sensitivity-specificity equality	The threshold at which sensitivity and specificity are equal	presence and absence	Pearson et al., 2004
Sensitivity-specificity sum maximization	The sum of sensitivity and specificity is maximized	presence and absence	Manel et al., 2001
Maximize Kappa	The threshold at which Cohen's Kappa statistic is maximized	presence and absence	Huntley et al., 1995 Elith et al., 2006
Average probability/suitability	The mean value across model output	presence-only	Cramer, 2003
Equal prevalence	Species' prevalence (the proportion of presences relative to the number of sites) is maintained the same in the prediction as in the calibration data.	presence and absence	Cramer, 2003

[1] Species occurrence records required to set the threshold. Based in part on Liu et al. (2005).

ments of 0.01 from 0 to 1 (i.e. 0, 0.01, 0.02, 0.03…0.99, 1). As the threshold increases, the proportion of the study area predicted to be suitable for the species, or in which the species is predicted to be 'present,' will decrease. Consequently, the proportion of observed presences that are correctly predicted decreases (i.e. decreasing sensitivity) and the proportion of observed absences that are correctly predicted increases (i.e. increasing specificity) (Figure 5A). From these data we can select the threshold at which sensitivity and specificity are equal (labeled *a* in Figure 5A) or at which their sum is maximized. Similarly, it is common to calculate Kappa across the range of possible thresholds and to select the threshold at which the statistic is maximized (Figure 5B).

Choice of an appropriate decision threshold is dependent on the type of data that are available and the question that is being addressed. Some methods require presence and absence records, while others require presence-only records (Table 4). When using both presence and absence records, the general approach is to balance the number of observed presences and absences that are correctly predicted; in effect, to maximize agreement between observed and predicted distributions. Thus, we must be willing to increase the omission rate (i.e. decrease sensitivity) in order to increase the proportion of observed absences that are correctly predicted (i.e. increase specificity). Liu et al. (2005) tested twelve methods for setting thresholds using presence and absence data for two European plant species. Based on four assessments of predictive perfor-
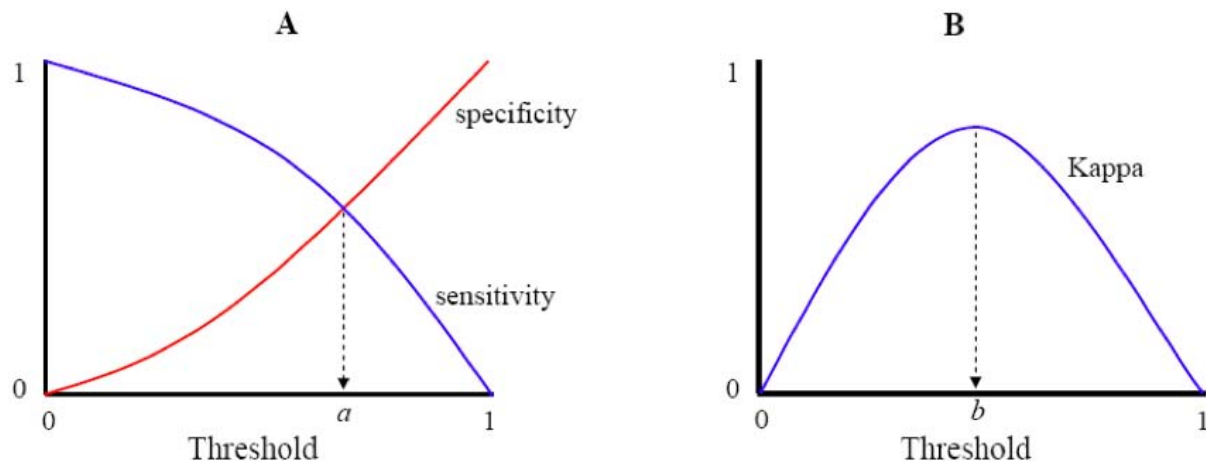
Figure 5. Plots showing changes in test statistics as the threshold of occurrence is adjusted. **A** Decrease in sensitivity and increase in specificity as the threshold is increased. The threshold labeled *a* corresponds to the specificity-sensitivity equality threshold. **B** Changes in the Kappa statistic as the threshold is adjusted. The threshold labeled *b* corresponds to that threshold at which Kappa is maximized.

mance (sensitivity, specificity, accuracy and Kappa), they concluded that the best methods for setting thresholds included maximizing the sum of sensitivity and specificity, using the average probability/suitability score, and setting equal prevalence between the calibration data and the prediction (Table 4). Maximizing Kappa did not perform well, and use of an arbitrary fixed value performed worst.

Methods that use only presence records for setting a threshold are required for cases in which absence data are unavailable. Presence-only methods can also be justified on the grounds that they avoid false absences (Section 3): it may be argued that we should be primarily concerned with maximizing the number of observed presences that are correctly predicted, rather than minimizing the number of absences that are incorrectly predicted as presences (since some absences may be recorded in apparently suitable environments; Pearson et al., 2006). For example, Pearson et al. (2007) used a dataset comprising very few presence-only records for geckos in Madagascar. Because confidence was high that the localities and species identification were correct, and because these species are not highly mobile and are therefore unlikely to be found in unsuitable habitat (i.e. sink habitat; see Section 2), omission of any occurrence record was considered a clear model error.

Therefore, the minimum predicted value corresponding to an observed presence was selected as a threshold to ensure zero omission. Distribution models thus predicted that many regions of the study area were suitable although no presences had been detected there. This approach suited the aim of the study, which was to prioritize regions for future surveys by estimating the potential distribution (see Case Study 1).

As a final illustration of the importance of selecting an appropriate decision threshold, we can return to an example raised in Section 2. If the purpose of modeling is to identify areas within which disturbance may impact a species negatively (e.g. as part of an environmental impact assessment), then the threshold may be set low to identify a larger area of potentially suitable habitat. In contrast, if the model was intended to identify potential introduction or reintroduction sites for an endangered species or species of recreational value, then it would be appropriate to choose a relatively high threshold. Choosing a high threshold reduces the risk of choosing unsuitable sites by identifying those areas with highest suitability (Pearce and Ferrier, 2000).

Threshold-independent assessment
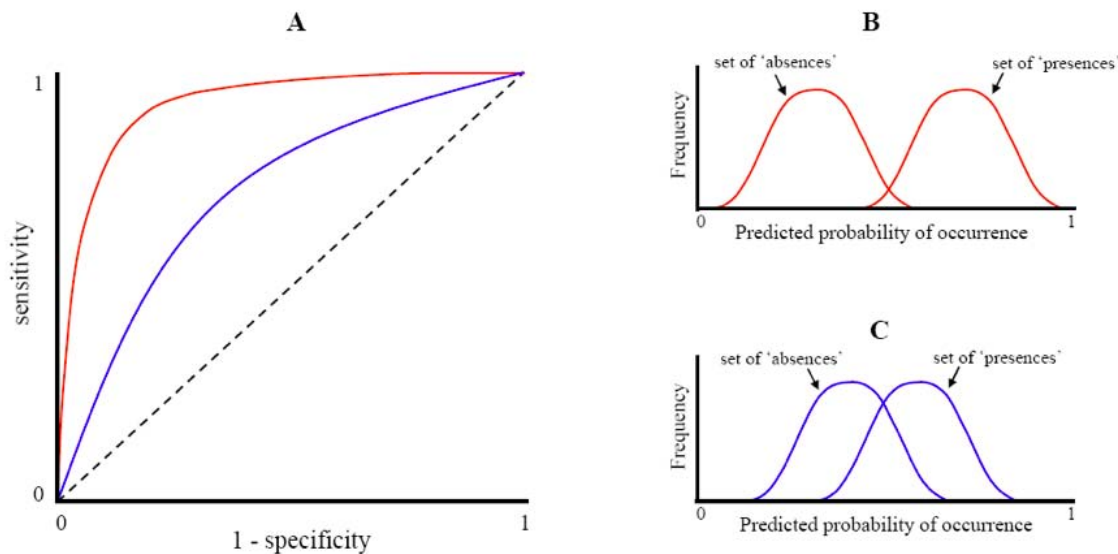When model output is continuous, assessment of predictive

Figure 6. Example Receiver Operating Characteristic (ROC) Curves and illustrative frequency distributions. **A** ROC curves formed by plotting sensitivity against '1 − specificity'. Two ROC curves are shown, the upper curve (red) signifying superior predictive ability. The dashed 1:1 line signifies random predictive ability, whereby there is no ability to distinguish occupied and unoccupied sites. **B** and **C** show example frequency distributions of probabilities predicted by a model for observed 'presences' and 'absences'. The results shown in **B** reveal good ability to distinguish presence from absence, whilst results in **C** show more overlap between the frequency distributions thus revealing poorer classification ability. The case shown in **B** would produce an ROC curve similar to the upper (red) curve in **A**. The case shown in **C** would give an ROC curve more like the lower (blue) curve in **A**.

performance using statistics derived from the confusion matrix will be sensitive to the method used to select a threshold for creating a binary prediction. Furthermore, if predictions are binary, the assessment of performance does not take into account all of the information provided by the model (Fielding and Bell, 1997). Therefore, it is often useful to derive a test statistic that provides a single measure of predictive performance across the full range of possible thresholds. This can be achieved using a statistic known as AUC: the Area Under the Receiver Operating Characteristic Curve.

The AUC test is derived from the Receiver Operating Characteristic (ROC) Curve. The ROC curve is defined by plotting sensitivity against '1 − specificity' across the range of possible thresholds (Figure 6A). Sensitivity and specificity are used because these two measures take into account all four elements of the confusion matrix (true and false presences and absences). It is conventional to subtract specificity from 1 (i.e. 1 − specificity) so that both sensitivity and specificity vary

in the same direction when the decision threshold is adjusted (Pearce and Ferrier, 2000). The ROC curve thus describes the relationship between the proportion of observed presences correctly predicted (sensitivity) and the proportion of observed absences incorrectly predicted (1 − specificity). Therefore, a model that predicts perfectly will generate an ROC curve that follows the left axis and top of the plot, whilst a model with predictions that are no better than random (i.e. is unable to classify accurately sites at which the species is present and absent) will generate a ROC curve that follows the 1:1 line (Figure 6A).

In order to summarize predictive performance across the full range of thresholds we can measure the area under the ROC curve (the AUC), expressed as a proportion of the total area of the square defined by the axes (Swets, 1988). The AUC thus ranges from 0.5 for models that are no better than random to 1.0 for models with perfect predictive ability. We can think of AUC in terms of the frequency distributions of probabilities

predicted for locations at which we have empirical data on presence and absence (Figure 6, B and C). A high AUC score reflects that the model can discriminate accurately between locations at which the species is present or absent. In fact, AUC can be interpreted as the probability that a model will correctly distinguish between a presence record and an absence record if each record is selected randomly from the set of presences and absences. Thus, an AUC value of 0.8 means the probability is 0.8 that a record selected at random from the set of presences will have a predicted value greater than a record selected at random from the set of absences (Fielding and Bell, 1997; Pearce and Ferrier, 2000).

AUC is a test that uses both presence and absence records. However, Phillips et al. (2006) have demonstrated how the test can be applied using randomly selected 'pseudo-absence' records in lieu of observed absences. In this case, AUC tests whether the model classifies presence more accurately than a random prediction, rather than whether the model is able to accurately distinguish presence from absence.

A number of different methods can be used to compute the AUC (Pearce and Ferrier, 2000). Some distribution modeling programs automatically calculate the AUC (e.g. Maxent; Table 3). The statistic can also be calculated using numerous other free software packages (e.g. R, for free download see http://www.r-project.org/).

Choosing a suitable test statistic
The choice of test statistic depends largely on how the model will be applied. If the aim is to predict the actual distribution (Section 2) then use of a test that incorporates both presence and absence records may be preferable (e.g. accuracy or Kappa). A good model will successfully predict both presences and absences with equal frequency. However, when using this approach it is important to realize that predictions that an unoccupied area is environmentally suitable (type 3 predictions, Figure 3) are considered model errors. Because type 3 predictions are theoretically expected (Section 2), models may be judged poor although they make biologically sound predictions.

If the aim of the modeling is to estimate the potential distribution (Section 2) then presence-only assessment of model performance using sensitivity and statistical significance is likely to be preferable. In this case we cannot test if the model is *correct* (since we do not know the true potential distribution), but rather we test if the model is *useful*. Our criteria for usefulness are that the model successfully predicts presence in a high proportion of test localities (i.e., known occurrences) whilst not predicting that an excessively large proportion of the study area is suitable. Thus, a model that successfully predicts whether the species is present at all test localities whilst classifying most of the study area as suitable may be correct (the environment may truly be suitable throughout most of the study area); however, the model is not useful because it is not more informative than a random prediction.

Subjective guidelines can be used to decide what values of a test statistic correspond to 'good' model performance. For example, Landis and Koch (1977) suggested that Kappa scores >0.75 represent an 'excellent' model, whilst Swets (1988) classified any AUC score >0.9 as 'very good'. However, the only true test of the model is whether it is useful for a given application. There are numerous potential applications of these methods (Table 1) and the final part of this synthesis describes three representative case studies.

Case Studies

Case Study 1: Predicting Distributions of Known and Unknown Species in Madagascar (based on Raxworthy et al., 2003)

Our knowledge of the identity and distribution of species on Earth is remarkably poor, with many species yet to be described and catalogued. This problem has two key elements, which may be termed the 'Linnean' and 'Wallacean' shortfalls (Whittaker et al., 2005). The Linnean shortfall refers to our lack of knowledge of how many, and what kind, of species exist. The term is a reference to Carl Linnaeus, who laid the foundations of modern taxonomy and the 18th century. The

Linnean shortfall concerns our highly incomplete knowledge of the diversity of life that exists on Earth.

The Wallacean shortfall refers to our inadequate knowledge of the distributions of species. This term is a reference to Alfred Russel Wallace who, as well as contributing to the early development of evolutionary theory, was an expert on the geographical distribution of species (he is sometimes referred to as 'the father of biogeography'). The Wallacean shortfall thus refers to our poor knowledge of the biogeography of most species. Species distribution modeling offers a powerful tool to address both the Linnean and Wallacean shortfalls, as demonstrated in a study by Raxworthy et al. (2003).

Raxworthy et al. (2003) modeled the distributions of 11 species of chameleon that are endemic to the island of Madagascar. They used species occurrence records from recent surveys and from older specimens deposited in collections of natural history museums. No observed absence records were available for building the models. Environmental variables were derived from remote sensing data, from a digital elevation model, and from weather station data that had been interpolated to a grid (i.e. converted from point vector to raster data). In all, 25 GIS layers were used in the modeling, including environmental variables describing temperature, precipitation, land cover, and elevation. All analyses were undertaken at a resolution of 1 km^2. The modeling algorithm used was GARP (see Table 3), which generated an output ranging from $0 - 10$ at increments of 1. Two alternative thresholds of occurrence were used: threshold = 1 (termed by the authors "any model predicts"), and threshold = 10 (termed "all models predict").

Predictive performance of the models was first evaluated by splitting the available data into two parts, 50% for calibrating the model and 50% for testing the model. The authors calculated the number of test localities at which the species was correctly predicted to be present and tested the statistical significance of the results using a chi-square test. Performance of the 11 models was generally good, with overall prediction success as high as 83%. Predictions usually were better than random. A second evaluation tested model performance using independent test data from herpetological surveys undertaken at 11 sites after the models had been built. In this case, model evaluation was based on both presence and absence records, since surveys were sufficiently thorough that detection probability was high. The success of these predictions was more than 70% and levels of statistical significance were uniformly high.

Raxworthy et al. (2003) thus demonstrated the potential for species' distribution models to be used to guide new field surveys toward areas in which the probability of species presence was high. This approach takes advantage of the type of model prediction illustrated by area *2* in Figure 3: the model identifies an area that is environmentally similar to where the species has already been found, but for which no occurrence data are available. The models can thus help to address the Wallacean shortfall, by improving our knowledge of the distributions of known species.

Raxworthy et al. (2003) also demonstrated that the models can help to address the Linnean shortfall by guiding field surveys toward areas where species new to science are most likely to be discovered. In this case, the approach makes use of the type of model prediction illustrated by area *3* in Figure 3: areas are identified that are unoccupied by the species being modeled, but where closely-related species that occupy similar environmental space are most likely to be found. By surveying sites identified by the distribution models for known species, Raxworthy et al. (2003) discovered seven new species, considerably greater than the number that would usually be expected on the basis of similar survey effort across a less-targeted area.

Case Study 2: Species' Distribution Modeling as a Tool for Predicting Invasions of Non-Native Plants (based on Thuiller et al., 2005)

Invasive species are increasingly a global concern, with invasions altering ecosystem functioning, threatening native biodiversity, and negatively impacting agriculture, forestry, and human health. Species' distribution modeling can be used to

identify areas that are most likely to be colonized by a known invader. The general approach is to model the distribution of a species using occurrence records from its native range, then project the model into new regions to assess susceptibility to invasion. The approach makes use of the type of model prediction illustrated by area *3* in Figure 3: areas are identified that are part of the potential, but not actual, distribution. Thuiller et al. (2005) used this method to identify parts of the world that are potentially susceptible to invasion by plant species native to South Africa.

Thuiller and colleagues developed distribution models for 96 South African plant taxa that are invasive in other regions of the globe. Species distribution data were extracted from large databases of occurrence records that have been collated for South Africa. Since the surveys incorporated within these databases were fairly comprehensive, the authors argued that absences within the databases are reliable and the modeling was thus undertaken using presence and absence data. Four climate-related variables that are known to affect plant physiology and growth were developed and used as input to the models. These included two measures of temperature (growing degree days and temperature of the coldest month) along with indexes of humidity and plant productivity. All the analyses were undertaken at a resolution of 25x25 km, which was considered sufficiently fine to identify environmental differences between regions at a global scale. Generalized Additive Models (implemented using the Splus-based BIOMOD application; Table 3) were used to build the distribution models.

The first step in the study was to model each species' distribution based on its native range in South Africa. Models were calibrated using 70% of the available records for each species, with the remaining 30% retained for model testing. The AUC validation test was applied using presence and absence test data. Because AUC assesses predictive performance based on continuous model output, it was not necessary to set a threshold of occurrence. Validation statistics for the test data were generally very good, with a median AUC score across all 96 species of 0.94 (minimum = 0.68, maximum = 1.0).

The second step in the study was to project the calibrated models worldwide. The accuracy of the global predictions was assessed for three example species using presence–only records from their non-native distributions (for example, in Europe, Australia, and New Zealand). Absence records were not available for these tests, so model performance was assessed using chi-square tests (see section 5). For each of the species, predictions of potentially suitable areas outside South Africa showed considerable agreement with observed records of invasions (chi-square test, $P<0.05$).

In a further step, Thuiller and colleagues summed the probability surfaces for all 96 taxa to produce a global map for risk of invasion by species of South African origin. Parts of the world most susceptible to invasion included six biodiversity 'hotspots', including the Mediterranean Basin, California Floristic Province, and southwest Australia. This study demonstrates that species distribution modeling can be a valuable tool for identifying sites prone to invasion. Such sites may be prioritized for monitoring and quarantine measures can be put in place to help avoid the establishment of invasive species.

Case Study 3: Modeling the Potential Impacts of Climate Change on Species' Distributions in Britain and Ireland (based on Berry et al., 2002)

Climate change has the potential to significantly impact the distribution of species. Species' distribution models have been used in a number of studies that aim to predict the likely redistribution of species under projected climate change over the coming century. The general approach is to calibrate the models based on current distributions of species and then predict future distributions of those species across landscapes for which the environmental input variables have been perturbed to reflect expected changes.

Berry et al. (2002) modeled 54 species that were chosen to represent a range of habitats common in Britain and Ireland. Species distribution data were obtained for the whole of Europe (i.e. European extent) from range maps for European

plants, amphibians, butterflies, and mammals. Since survey effort across Europe is generally very high, areas where a species had not been recorded were considered reliable measures of absence, and both presence and absence data were therefore used in the modeling. Five environmental variables describing temperature, precipitation, and soil type were used. These variables were generated using both current climate data and predictions of future climate from a General Circulation Model (GCM). GCMs are complex simulations that predict future climates using scenarios of greenhouse gas emissions. The environmental variables were calculated at a coarse resolution (~50x50 km) for Europe, and also at a finer resolution (10x10 km) for Britain and Ireland. The modeling algorithm was an Artificial Neural Network (Table 3), which gave predictions of relative suitability ranging from 0 to 1. These continuous predictions were converted into binary predictions of presence and absence by applying a threshold of occurrence that maximized the Kappa statistic.

An important part of Berry et al.'s (2002) method was that calibration of the distribution model was carried out at the European scale (large extent and coarse resolution) and then the model was used to predict distributions in Britain and Ireland (smaller extent and finer resolution). This approach ensured that when distributions were predicted under future climate scenarios in Britain and Ireland, the model was not required to extrapolate beyond the range of data for which it was calibrated (since future climates in Britain and Ireland are expected to be similar to conditions currently experienced elsewhere in Europe).

Evaluation of the models was undertaken at the European scale by comparing model predictions against a test dataset, which comprised one third of the available data that had been randomly selected and not used in model calibration. Since both presence and absence records were available, the Kappa statistic was applied. Results from these tests revealed generally good model performance, with 27 out of 54 species achieving Kappa scores >0.75, and only 7 of 54 species with Kappa scores <0.6.

The models calibrated for Europe were then used to predict distributions in Britain and Ireland under current and projected future climates. Berry et al. (2002) emphasized that they did not predict actual distributions, but rather 'bioclimate envelopes', or suitable climate space. It is important to remember that actual future distributions will be determined by many factors that are not taken into account in the modeling, including the ability of species to colonize areas that become suitable. Nevertheless, the distribution models enabled each species to be placed in one of three categories: those expected to lose suitable climate space, those expected to gain suitable climate space, and those showing little change. Species' distribution modeling thus enables preliminary assessments of the possible impacts of climate change to be made, providing information that may be valuable in developing conservation policies to address the threat.

Acknowledgements

Terms of Use

Literature Cited

Anderson, R.P., M. Gomez-Laverde, and A.T. Peterson. 2002a. Geographical distributions of spiny pocket mice in South America: Insights from predictive models. Global Ecology and Biogeography 11, 131-141.

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

Anderson, R.P., D. Lew, and A.T. Peterson. 2003. Evaluating predictive models of species' distributions: Criteria for selecting optimal models. Ecological Modelling 162: 211-232.

Anderson, R. P. and E. Martínez-Meyer. 2004. Modeling species' geographic distributions for conservation assessments: an implementation with the spiny pocket mice (Heteromys) of Ecuador. Biological Conservation 116: 167-179.

Araújo, M.B. and M. Luoto. 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography 16: 743-753.

Araújo, M.B., and R.G. Pearson. 2005. Equilibrium of species' distributions with climate. Ecography 28: 693-695.

Araújo, M.B., R.G. Pearson, W. Thuiller, and M. Erhard. 2005a. Validation of species–climate envelope models under climate change. Global Change Biology 11: 1504–1513.

Araújo, M.B., W. Thuiller, P.H. Williams, and I. Reginster. 2005b. Downscaling European species atlas distributions to a finer resolution: Implications for conservation planning. Global Ecology and Biogeography 14: 17-30.

Araújo, M. B., and P.H. Williams. 2000. Selecting areas for species persistence using occurrence data. Biological Conservation 96: 331–345.

Beerling, D.J., B. Huntley, and J.P. Bailey. 1995. Climate and the distribution of *fallopia japonica*: Use of an introduced species to test the predictive capacity of response surfaces. Journal of Vegetation Science 6: 269-282.

Berry, P.M., T.P. Dawson, P.A. Harrison, and R.G. Pearson. 2002. Modelling potential impacts of climate change on the bioclimatic envelope of species in Britain and Ireland. Global Ecology and Biogeography 11: 453-462.

Bourg, N.A., W.J. McShea, and D.E. Gill. 2005. Putting a cart before the search: Successful habitat prediction for a rare forest herb. Ecology 86: 2793-2804.

Boyce, M.S., P.R. Vernier, S.E. Nielsen, and F.K.A. Schmiegelow. 2002. Evaluating resource selection functions. Ecological Modelling 157: 281-300.

Brotons, L., W. Thuiller, M.B. Araújo, and A.H. Hirzel. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27: 437-448.

Buckland, S.T., and D.A. Elston. 1993. Empirical models for the spatial distribution of wildlife. Journal of Applied Ecology 30: 478-95.

Carpenter, G., A.N. Gillson, and J. Winter. 1993. DOMAIN: a flexible modeling procedure for mapping potential distri-

butions of plants and animals. Biodiversity and Conservation 2: 667–680.

Chase, J.M., and M.A. Leibold. 2003. Ecological niches: Linking classical and contemporary approaches University of Chicago Press, Chicago, USA.

Chuine, I., and E.G. Beaubien. 2001. Phenology is a major determinant of temperate tree distributions. Ecology Letters 4, 500-510.

Cramer, J. S. 2003. Logit Models: From Economics and Other Fields. Cambridge University Press.

Elith, J., C. Graham, and the NCEAS species distribution modeling group. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129-151.

Elith, J. and Leathwick, J.R. 2007. Predicting species' distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity and Distributions 13: 165-175.

Engler, R., A. Guisan, and L. Rechsteiner. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41: 263-274.

Ferrier, S., G. Watson, J. Pearce, and Drielsma 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast new south wales. I. Species-level modelling. Biodiversity and Conservation 11: 2275–2307.

Fielding, A.H. and J.F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24: 38-49.

Fleishman, E., R. Mac Nally, and J.P. Fay. 2002. Validation tests of predictive models of butterfly occurrence based on environmental variables. Conservation Biology, 17: 806-817.

Fleishman, E., R. Mac Nally, J.P. Fray, and D.D. Murph. 2001. Modeling and predicting species occurrence using broad-scale environmental variables: An example with butterflies of the great basin. Conservation Biology 15: 1674-1685.

Gibbons, D.W., J.B. Reid, and R.A. Chapman. 1993. The new atlas of breeding birds in Britain and Ireland: 1988–1991 Poyser, London, UK.

Graham, C.H., S. Ferrier, F. Huettman, C. Moritz, and A.T. Peterson. 2004a. New developments in museum-based in-

formatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19: 497–503.

Graham, C.H., S. R. Ron, J.C. Santos, C.J. Schneider, and C. Moritz. 2004b. Integrating phylogenetics and environmental niche models to explore speciation mechanisms in Dendrobatid frogs. Evolution 58: 1781–1793.

Graham, C.H., C. Moritz, and S. E. Williams. 2006 Habitat history improves prediction of biodiversity in rainforest fauna. PNAS 103: 632–636.

Guisan, A., O. Broennimann, R. Engler, M. Vust, N.G. Yoccoz, A. Lehman, and N.E. Zimmermann. 2006. Using niche-based models to improve the sampling of rare species. Conservation Biology 20: 501–511.

Guisan, A., T.C. Edwards Jr., and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: Setting the scene. Ecological Modelling 157: 89-100.

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

Hampe, A. 2004. Bioclimatic models: what they detect and what they hide. Global Ecology and Biogeography 11: 469-471.

Hannah, L., G.F. Midgley, G. Hughes, and B. Bomhard. 2005. The view from the cape: extinction risk, protected areas, and climate change. BioScience 55: 231-242.

Heikinnen, R.K., M. Luoto, R. Virkkala, R.G. Pearson, and J-H. Körber. 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. Global Ecology and Biogeography 16: 754–763.

Higgins, S.I., D.M. Richardson, R.M. Cowling, and T.H. Trinder-Smith. 1999. Predicting the landscape-scale distribution of alien plants and their threat to plant diversity. Conservation Biology 13: 303-313.

Hijmans, R. J., S.E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

Hirzel, A.H., J. Hausser, D. Chessel, and N. Perrin. 2002. Ecological-niche factor analysis: How to compute habitat-suitability map without absence data. Ecology 83: 2027-2036.

Huberty, C.J. 1994. Applied Discriminant Analysis. Wiley Interscience, New York, USA.

Hugall, A., C. Moritz, A. Moussalli and. J. Stanisic. 2002. Reconciling paleodistribution models and comparative phylogeography in the wey tropics rainforest land snail Gnarosophia bellendenkerensis. PNAS 99: 6112-6117.

Huntley, B., P.M. Berry, W. Cramer, and A.P. Mcdonald. 1995. Modelling present and potential future ranges of some European higher plants using climate response surfaces. Journal of Biogeography 22: 967-1001.

Hutchinson, G.E. 1957. Concluding remarks. Cold Spring Harbor Symposium on Quantitative Biology 22: 415-457.

Iverson, L.R., and A.M. Prasad. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. Ecological Monographs 68: 465-485.

Kozak, J.H. and J.J. Wiens. 2006. Does niche conservatism promote speciation? A case study in North American salamanders. Evolution 60: 2604-2621.

Landis, J.R., and G.C. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33: 159-174.

Lawler, J. J., D. White, R.P. Neilson, and A.R. Blaustein. 2006. Predicting climate-induced range shifts: model differences and model reliability. Global Change Biology 12: 1568-1584.

Lawton, J. L. 2000. Concluding remarks: a review of some open questions. Pages 401-424 in M. J. Hutchings, E. John, and A. J. A. Stewart, editors. Ecological Consequences of Heterogeneity. Cambridge University Press.

Leathwick, J.R., D. Rowe, J. Richardson, J. Elith and T. Hastie. 2005. Using multivariate adaptive regression splines to predict the distribution of New Zealand's freshwater diadromous fish. Freshwater Biology 50, 2034-2052.

Leathwick, J.R., J. Elith, and T. Hastie. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling 199: 188-196.

Lehman, A., J.M. Overton, and J.R. Leathwick. 2002. GRASP: generalized regression analysis and spatial prediction. Ecological Modelling 157: 189-207.

Lindenmayer, D. B., H.A. Nix, J.P. McMahon, M.F. Hutchinson, and M.T. Tanton. 1991. The conservation of Leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. Journal of Biogeography 18: 371-383.

Liu, C., P.M. Berry, T.P. Dawson, and R.G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28: 385-393.

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

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

Manel, S., H.C. Williams, and S.J. Ormerod. 2001. Evaluating presences-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38: 921-931.

Martinez-Meyer, E., A.T. Peterson, and W.W. Hargrove. 2004. Ecological niches as stable distributional constraints on mammal species, with implications for pleistocene extinctions and climate change projections for biodiversity. Global Ecology and Biogeography 13: 305-314.

Monserud, R.A., and R. Leemans. 1992. Comparing global vegetation maps with the kappa statistic. Ecological Modelling 62: 275-293.

Pearce, J., and D.B. Lindenmayer. 1998. Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (*lichenostomus melanops cassidix*) in southeastern Australia. Restoration Ecology 6: 238-243.

Pearce, J., and S. Ferrier. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133: 225-245.

Pearson, R.G., and T.P. Dawson. 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., T.P. Dawson, P.M. Berry, and P.A. Harrison. 2002. Species: A spatial evaluation of climate impact on the envelope of species. Ecological Modelling 154: 289-300.

Pearson, R.G., T.P. Dawson, and C. Liu. 2004. Modelling spe-

cies distributions in Britain: A hierarchical integration of climate and land-cover data. Ecography 27: 285-298.

Pearson, R.G., C.J. Raxworthy, M. Nakamura, and A.T. Peterson. 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., W. Thuiller, M.B. Araújo, E. Martinez-Meyer, L. Brotons, C. McClean, L. Miles, P. Segurado, T.P. Dawson, and D. Lees. 2006. Model-based uncertainty in species' range prediction. Journal of Biogeography 33: 1704-1711.

Peterson, A.T. 2003. Predicting the geography of species' invasions via ecological niche modeling. Quarterly Review of Biology 78: 419-433.

Peterson, A.T., and J.J. Shaw. 2003. *Lutzomyia* vectors for cutaneous leishmaniasis in southern brazil: Ecological niche models, predicted geographic distributions, and climate change effects. International Journal of Parasitology 33: 919-931.

Peterson, A.T., B.W. Benz, and M. Pape. 2007. Highly pathogenic H5N1 avian influenza: Entry pathways into North America via bird migration. PLoSONE 2(2): e261. doi:10.1371/journal.pone.0000261.

Peterson, A. T., J. Soberon, and V. Sanchez-Cordero. 1999. Conservatism of Ecological Niches in Evolutionary Time. Science 285: 1265-1267.

Peterson, A. T., R. R. Lash, D. S. Carroll, and K. M. Johnson. 2006. Geographic potential for outbreaks of Marburg hemorrhagic fever. American Journal of Tropical Medicine & Hygiene 75: 9-15.

Peterson, A.T., M.A. Ortega-Huerta, J. Bartley, V. Sanchez-Cordero, J. Soberon, R.W. Buddemeier, and D.R.B. Stockwell. 2002. Future projections for mexican faunas under global climate change scenarios. Nature 416: 626-629.

Phillips, S.J., R.P. Anderson, and R.E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190, 231-259.

Phillips, S.J., M. Dudik, and R.E. Schapire. 2004. A maximum entropy approach to species distribution modeling. In Proceedings of the 21st international conference on machine learning, pages 655-662. AMC Press, New York, USA.

Pulliam, H. R. 2000. On the relationship between niche and distribution. Ecology Letters, 3:349-361.

Raxworthy, C.J., C. Ingram, N. Rabibosa, and R.G. Pearson. 2007. Species delimitation applications for ecological niche modeling: a review and empirical evaluation using Phelsuma day gecko groups from Madagascar. Systematic Biology.

Raxworthy, C.J., E. Martinez-Meyer, N. Horning, R.A. Nussbaum, G.E. Schneider, M.A. Ortega-Huerta, and A.T. Peterson. 2003. Predicting distributions of known and unknown reptile species in Madagascar. Nature 426: 837-841.

Robertson, M. P., N. Caithness, and M.H. Villet. 2001. A PCA-based modeling technique for predicting environmental suitability for organisms from presence records. Diversity and Distributions 7: 15-27.

Segurado, P. and M.B. Araújo. 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography.

Stockwell, D.R.B. and D.P. Peters. 1999. The GARP modelling system: Problems and solutions to automated spatial prediction. International Journal of Geographical Information Systems 13: 143-158.

Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. Science 240: 1285-1293.

Thomas, C.D., A. Cameron, R.E. Green, M. Bakkenes, L.J. Beaumont, Y.C. Collingham, B.F. Erasmus, M.F. De Siqueira, A. Grainger, L. Hannah, L. Hughes, B. Huntley, A.S. Van Jaarsveld, G.F. Midgley, L. Miles, M.A. Ortega-Huerta, A.T. Peterson, O.L. Phillips, and S.E. Williams. 2004. Extinction risk from climate change. Nature 427: 145-148.

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., M.B. Araújo, R.G. Pearson, R.J. Whittaker, L. Brotons, and S. Lavorel. 2004. Uncertainty in predictions of extinction risk. Nature 430: 33 (doi: 10.1038/Nature02716).

Thuiller, W., D.M. Richardson, P. Pysek, G.F. Midgley, G.O. Hughes, and M. Rouget. 2005. Niche-based modeling as a tool for predicting the global risk of alien plant invasions. Global Change Biology 11: 2234-2250.

Verbyla, D.L., and J.A. Litaitis. 1989. Resampling methods for evaluating classification accuracy of wildlife habitat models. Environmental Management 13: 783-787.

Whittaker, R.J., M.B. Araújo, P. Jepson, R.J. Ladle, J.E.M. Watson, and K.J. Willis. 2005. Conservation biogeography: Assessment and prospect. Diversity and Distributions 11: 3-23.

Woodward, F. I., and D. J. Beerling. 1997. The dynamics of vegetation change: health warnings for equilibrium 'dodo' models. Global Ecology and Biogeography 6: 413-418.

Zaniewski, A.E., A. Lehman, and J.M. Overton. 2002. Predicting species spatial distributions using presence–only data: A case study of native New Zealand ferns. Ecological Modelling 157: 261–280.

Zar, J.H. 1996. Biostatistical Analysis 3rd Edition. Prentice Hall, New Jersey, USA.