Biol Invasions (2011) 13:2785–2797
DOI 10.1007/s10530-011-9963-4

ORIGINAL PAPER

Use of niche models in invasive species risk assessments

A. Jiménez-Valverde · A. T. Peterson ·
J. Soberón · J. M. Overton · P. Aragón ·
J. M. Lobo

Received: 19 July 2010 / Accepted: 1 February 2011 / Published online: 12 February 2011

Abstract Risk maps summarizing landscape suitability of novel areas for invading species can be valuable tools for preventing species' invasions or controlling their spread, but methods employed for development of such maps remain variable and unstandardized. We discuss several considerations in development of such models, including types of distributional information that should be used, the nature of explanatory variables that should be incorporated, and caveats regarding model testing and evaluation. We highlight that, in the case of invasive species, such distributional predictions should aim to derive the best hypothesis of the potential distribution of the species by using (1) all distributional information available, including information from both the native range and other invaded regions; (2) predictors linked as directly as is feasible to the physiological requirements of the species; and (3) modelling procedures that carefully avoid overfitting to the training data. Finally, model testing and evaluation should focus on well-predicted presences, and less on efficient prediction of absences; a *k*-fold regional cross-validation test is discussed.

Keywords Biological invasions · Model validation · Occurrence data · Potential distribution models

Introduction

Invasive species are colonizer species that have established populations outside their native distributional ranges and that have potential to spread and affect native ecosystems or local human-mediated systems (Lockwood et al. 2007). Species will be able to maintain populations in the long term without immigrational input only in areas meeting their abiotic (physical) needs and harbouring the appropriate biological milieu, but these areas must also be within their dispersal potential (Pulliam 2000; Soberón and Peterson 2005). Under this perspective,

A. Jiménez-Valverde · A. T. Peterson · J. Soberón
Biodiversity Institute, University of Kansas, Lawrence,
KS 66045, USA

Present Address:
A. Jiménez-Valverde
Departamento de Biología Animal, Facultad de Ciencias,
Universidad de Málaga, 29071 Málaga, Spain

J. M. Overton
Landcare Research, Hamilton 2001, New Zealand

P. Aragón
Department of Ecology and Evolution (DEE), Biophore,
University of Lausanne, Lausanne, Switzerland

J. M. Lobo (✉)
Departamento de Biodiversidad y Biología Evolutiva,
Museo Nacional de Ciencias Naturales (CSIC),
c/José Gutiérrez Abascal 2, 28006 Madrid, Spain
e-mail: mcnj117@mncn.csic.es

species' invasions can be conceptualized as the result of an increase in their dispersal capability, generally resulting from human activities (Groves and Di Castri 1996). Human-mediated invasion processes are usually more dynamic, quick, and dramatic than natural ones (Lockwood et al. 2007), and their ecological and economic impacts have made them a central issue in ecology and conservation biology (Bergmans and Blom 2001).

The invasion process consists of a series of steps, including transport, establishment, spread, and impact (Williamson 1996), and different strategies have been developed to address and mitigate invasions at each stage (Shigesada and Kawasaki 1997; Parker et al. 1999; Williamson 1999; Davis et al. 2000; Kolar and Lodge 2002; Cassey et al. 2004; Leung et al. 2004; Mack 2004; Simberloff et al. 2005). As efforts focused on management generally cannot prevent future invasions (Peterson and Vieglais 2001), ecologists agree on the need for preventive steps, given that control or eradication of already-established populations is more difficult and costly (Hobbs and Humphries 1995; Bax et al. 2001). Preventive steps for invasive species would be more effective if the set of environmental conditions within which the species can have a net positive population increase rate could be estimated, since the geographic projection of those conditions provides a preliminary, but reliable, hypothesis of sites suitable for possible invasion (Peterson 2003).

The range of ecological conditions within which a species can maintain such source populations can be defined as its niche (Hutchinson 1978). Unfortunately, these niche conditions have been defined in diverse ways with little standardization and a species may also impact the conditions that allow it to survive. By stressing different features of the niche, and considering (or not) impacts of other species in the definition, very different niche concepts are obtained. Indeed, the term "niche" has evolved into a veritable amalgam of meanings, terminology, and theory (Chase and Leibold 2003; Colwell 1992). For the purpose of this paper, we use the term in its Grinnellian sense (Jackson and Overpeck 2000; Soberón 2007): niches are subsets of a multivariate space of environmental conditions defined in terms of variables that are not consumed or affected (at least in ecologically relevant timespans) by the presence of the species in question within which the species can maintain populations without inmigrational subsidy. The relationships between these environmental subsets and geography are complex, and frequently do not correspond to expected patterns (Angert 2009); as a result, it is necessary to understand the caveats associated with deriving niche estimates from species' occurrences across realized distributions.

The geographic potential of species' invasions can be estimated via so-called ecological niche modelling procedures (see review by Peterson 2003). Niche models are correlational techniques aimed at identifying key suites of environmental conditions within which the species is present, based on occurrence data and environmental data stored in geographic information systems. If historical constraints on dispersal are overcome, additional suitable areas become available to the species. Such procedures have been used to assist preventative actions in tandem with information on propagule pressure to identify corridors and facilitate control of further spread (Peterson 2003; Thuiller et al. 2005). In view of the probable increasing use of these techniques in the future, we examine key conceptual and methodological aspects of niche modelling as applied to invasive species to enhance the discussion within this emerging field of research.

Which niches are discovered using distribution data?

The range of environmental conditions associated with positive intrinsic growth rates is termed the fundamental niche (*FN*), and can probably be estimated only via experimental (mainly physiological) studies (Calosi et al. 2010; Kearney 2006; Kearney et al. 2009). A species may be able to survive under environmental conditions existing at a particular time within the area that is accessible to it (*M* in the BAM framework of Soberón and Peterson 2005; Varela et al. 2009); thus, a species can only potentially inhabit a portion of its *FN* depending on the available environmental space (the so-called potential niche or *PN*; see Jackson and Overpeck 2000). That is, the *FN* might be approximated as the total environmental space encompassed by different expressions of the *PN*s through time. Additionally, within any one *PN* the species may effectively only occupy a subset (G_O in Fig. 1) of the whole

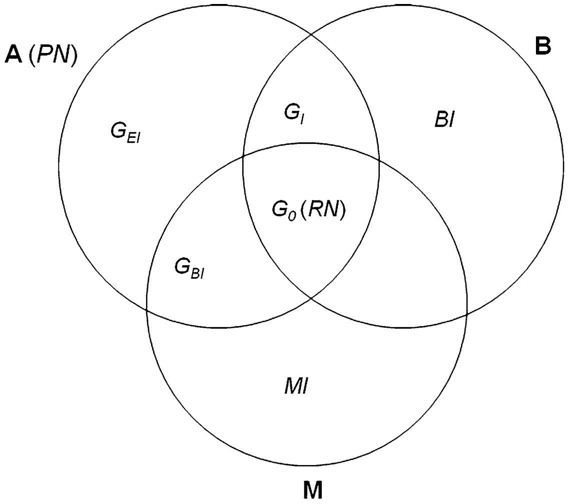

Fig. 1 Diagram showing factors affecting distributions of species (following Soberón and Peterson 2005; Soberón 2007). *A*, is the geographic area in which the environment is suitable at a given time, and where the intrinsic growth rate of the species would be positive (Potential Distribution Area). This area corresponds to the regions meeting the physiological requirements of the species (Potential Niche or *PN* sensu Jackson and Overpeck 2000), so $PN = FN \cap E$, where *FN* is the fundamental niche of the species and *E* represents the existing environmental conditions in the region. *B* is the geographic area where biotic interactions are favourable for species' existence and *M* is the geographic area that is accessible to the species. Thus, G_0 is the actual occupied area (its environments constitute the realized niche or *RN*) and is equivalent to $A \cap B \cap M$. G_I is the geographic area with favourable biotic and environmental conditions that remains un-colonized (an invasible area) or $A \cap B$ outside of M. G_{BI} is the geographic area available and presenting favourable environmental conditions but inappropriate biotic conditions (a probably invasible area if the species were to have a plastic response with regards to biotic interactions or if the required associated species also invade the region). G_{EI} is the presently inaccessible geographic area with favourable environmental conditions but unsuitable biotic interactions (also potentially invasible depending on the species' capacity to overcoming existing dispersal limitations and the biotic response of the species). Finally MI and BI are areas environmentally unsuitable but accessible or biotically appropriate, respectively. Although potential distributions should preferably correspond to the area $G_0 + G_I + G_{BI}$, our ignorance of the role played by biotic interactions, as well as the plastic response of species toward them should guide the precautionary principle of also including G_{EI} areas. Thus, without information about the factors that compose *B*, it is likely that model extrapolations will include undefined portions of *A*

environmentally favourable conditions, which corresponds to the realized niche (*RN*) determined jointly by environmental conditions (*A*), biotic factors (*B*), and dispersal limitations (*M*; Svenning and Skov 2004; Soberón 2007).

As a consequence of the partial reciprocity between niche space and geographic distributions (the so-called Hutchinsonian duality; see Colwell and Rangel 2009), the area currently occupied by a species (the geographic projection of *RN*, or G_O in Fig. 1) can only provide partial environmental information on the full set of environmental conditions under which a species can survive and reproduce (*A* and its subsets G_{BI}, G_I.and G_{EI}; see Fig. 1). Unfortunately, a continuum of disequilibrium at different spatial scales is the norm in observed realized distributions of species (Peterson 2003). Thus, the geographic projection of the niche conditions derived from the occupied area will be biased not only by the role played by restrictive forces such as biotic interactions and dispersal limitations on *PN*, but also because present-day environmental conditions may frequently represent only a portion of the full conditions inhabitable by the species (i.e., $PN < FN$). This fact hinders full estimation of the *FN* of species and thus also of their potential distributions (Colwell and Rangel 2009; Rödder et al. 2009; Soberón and Nakamura 2009; Godsoe 2010), which constraints dramatically both the types of data and the techniques that can be used to built useful risk maps.

Considering that correlative niche modelling techniques focus mostly on the *RN*, and that the *RN* is in fact an unknown subset of the *PN*, which in turn may represent only partially the *FN*, how can we maximize the capacity to unveil species' niches from simple occurrence data? Evidence available regarding similarity of environmental conditions obtained from native and invaded areas suggests that a relatively reliable picture of the *RN* can be obtained in some cases from analyses of distributional data (Richardson and Thuiller 2007). On the other hand, some studies have noted discrepancies in derived niche conditions among regions that suggest that only partial niche representations can be obtained from spatial or temporal data slices (Mau-Crimmins et al. 2006; Broennimann et al. 2007; Varela et al. 2009; Medley 2010), or that the phenotypic plasticity of the species is greater than what may be appreciated from realized distributions (Orr and Smith 1998).

The empirical evidence about invasions is complex (Lambdon et al. 2008), so discrepancies between niches observed on native and invaded distributional areas may result from a variety of factors like acclimation and niche evolution (Jiménez-Valverde and Lobo 2011), or release from competitors and predators that may also change the *RN* (Keane and Crawley 2002). However, other factors can probably affect these discrepancies more easily: (1) if the species is not in distributional equilibrium in a region, the niche models based on distributional data from that region may not be able to capture the species' entire environmental potential; and (2) model extrapolations depend critically on methodological questions, such that niche models must be fit appropriately and not defined overly narrowly, which may explain apparent niche differences as methodological artefacts only (see example described in Peterson and Nakazawa 2008).

Models for invasive species have to date generally been trained on native distributional areas (e.g., Peterson 2003; Ficetola et al. 2007), which have the advantage of having a higher probability of being in distributional equilibrium. However, occurrence data from invaded regions, which may offer additional insights of novel environments and biotic contexts, have been used as well (e.g., Rouget et al. 2004; Muñoz and Real 2006; Ganeshaiah et al. 2003). Occurrence information from invaded ranges has been used principally when data from native regions are unavailable or costly to obtain. Indeed, in some cases (e.g., Kluza et al. 2007), the native region may be unknown, and models based on invaded regions can be used to predict the geographic origin of the introductions (Steiner et al. 2008). Although we are well aware of the perils of training niche models in regions undergoing invasion and thereby not in distributional equilibrium (Peterson 2005a), many invasive species seem to have reached some level of distributional equilibrium in their new ranges, or at least encounter local environmental limits that can be informative for estimating niches. Indeed, some native ranges (e.g., species endemic to small islands) may not suffice to characterize the entirety of the species' niche.

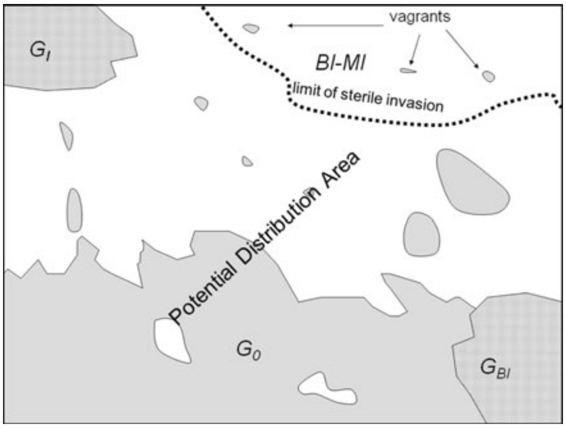

Fig. 2 Schematic representation of a species' distribution (modified from Gorodkov 1986a, b; Gaston 2003). The *gray area* represents the localities at which a species has been found (mostly G_0). Stippled areas represent favourable regions that have not been possible to colonize owing to the presence of a negative biological interaction (G_{BI}) or to the effects of dispersal constraints (G_I). Within the suitable area, absence localities can also result from demographic processes (Pulliam 1988, 2000), while some sink populations may be maintained and vagrants found in unsuitable localities (*BI-MI*) thanks to immigration from source localities. In the case of invaders we are interested in predicting as favourable the whole potential distribution area (the geographic representation of the *PN* over an area much broader than just the native distribution)

Additionally, environmentally suitable sites within the native distributional area may remain unoccupied because of biotic interactions, metapopulation dynamics, dispersal limitations, or historical constraints (Ricklefs and Schluter 1993; Hanski 1998; Pulliam 1988, 2000; Soberón and Peterson 2005; Soberón 2007; see Fig. 2). These restrictive forces may, at least in theory, differ or even be lacking in invaded areas, allowing the species to inhabit a *RN* that supercedes the native-range manifestation, although few credible demonstrations of this idea have been presented (Peterson and Nakazawa 2008). Facilitative interactions, ecological and physiological plasticity, and microevolutionary adaptation may also lead non-native species to occupy novel environmental conditions in invaded areas (Orr and Smith 1998; Rodríguez-Trelles and Rodríguez 1998; Gomulkiewicz et al. 1999; Davis and Shaw 2001; Thomas et al. 2001; Bruno et al. 2003; Maron et al. 2004). Thus, as suggested by a few authors (Welk 2004; Mau-Crimmins et al. 2006; Broennimann and Guisan 2008; Beaumont et al. 2009), an ideal approach to forecasting species' invasions would take into consideration all available information coming from native and invaded regions, as well as those provided by different time slices, since it may enhance characterization of the species' fundamental niche. Still, it must be emphasized that, owing to the impossibility of estimating the *FN* accurately, and although incorporation of all available occurrence information may be advantageous, underestimation of species' potential distributions when models are projected onto novel conditions may be the rule rather than the exception.

Occurrence data and modelling techniques

In general, data documenting presences of species are much more readily available than data documenting absences (Soberón and Peterson 2005), thanks mainly to the difficulty of discriminating genuine absence from inadequate survey effort (MacKenzie et al. 2002; Royle et al. 2005; Anderson 2003; Lobo 2008a; Gu and Swihart 2004). Considerations that may place 'presence' records outside of ecologically suitable conditions include incorrect identifications, typographic or other data-entry errors, and poor taxonomic arrangements, all factors that should not be particularly common. In contrast, myriad factors lead to no data documenting the presence of a species at a particular site, when the species could in fact be present, beginning most simply with the fact that no species' distribution has been sampled exhaustively. Finally, given dispersal limitations (*M* in the BAM framework), large suitable areas will exist that will lack populations of the species: indeed, these areas are the basis for the occurrence of species' invasions.

In general, quality assessment of occurrence data is a *sine qua non* before any modelling exercise is started; spatial bias and mis-georeferencing are common problems (Jiménez-Valverde et al. 2010). On one hand, if the presence localities are not drawn from exhaustive environmental and spatial sampling efforts, biases with respect to environmental dimensions will certainly exist (see, e.g., Hirzel et al. 2001; Zaniewski et al. 2002; Engler et al. 2004; Kadmon et al. 2004; Hortal and Lobo 2005; Reese et al. 2005), which may have profound consequences for estimation of species' niches (Lobo et al. 2007; Hortal et al. 2008), because missing data might correspond to important parts of environmental response surfaces. On the other hand, some presence records may fall far from "optimal" conditions in environmental space for the species and can be troublesome (Dennis and Hardy 1999): some methods simply ignore them (e.g., Busby 1991). However, discriminating sink populations or vagrants from established, reproductive populations that persist under marginal or suboptimal conditions can be difficult, particularly when occurrence information is drawn from large, heterogeneous, historical data sets (Graham et al. 2004). But those environmentally peripheral, but stable, populations are precisely the records that are most informative about the limits of tolerance for the species. Hence, we argue that removing environmentally extreme occurrence data may not be a desirable automatic procedure and should be critically assessed: such information may rather be indispensable in describing the full environmental potential of the species, and thus may help approximate the full diversity of areas and environments at risk of invasion, particularly if some of the restrictive forces acting on the species on its native range are relaxed or absent in the invaded region. Of course, the reliability of these extreme records has to be assessed to avoid incorporating the effects of georeferencing errors and mis-identifications, which may be common in large-scale biodiversity data sets (Jiménez-Valverde et al. 2010).

Numerous techniques have been explored for estimating niches and forecasting potential distributions of species from incomplete occurrence data (Guisan and Zimmermann 2000; Elith et al. 2006). These techniques generally aim to estimate a mathematical function linking potential predictor variables to available occurrence information for the species. They may be divided, for practical purposes, into those techniques requiring presence-absence occurrence data as input, including different forms of pseudoabsences (environmentally unsuitable absences selected at random from areas where the species has not been recorded; e.g., Engler et al. 2004; Lobo et al. 2006; López-Darias et al. 2008) or background data (absences selected at random from the entire study area; e.g., Elith et al. 2006), and those that only work with presence data (Fig. 3). Methods can also be more or less arbitrarily divided by their algorithmic approaches. Three extreme cases are statistical methods (e.g., generalized linear models, generalized additive models, classification and regression trees, etc.; see Guisan and Zimmermann 2000 for an overview); machine-learning methods (e.g., Maxent, Phillips et al. 2006; GARP, Stockwell and Peters 1999; artificial neural networks, Olden et al. 2008) and geometrical methods based on set theory and profiles of distributions of presence data in environmental dimensions (BIOCLIM, Busby 1991; Mahalanobis distance, Farber and Kadmon 2003; Fig. 3).

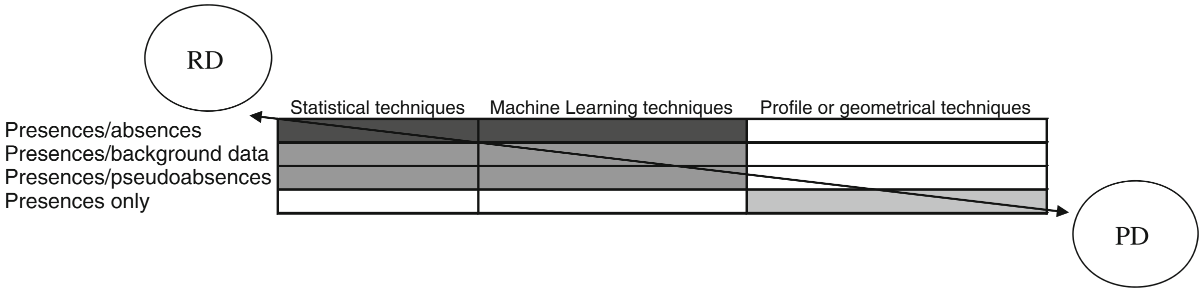

Fig. 3 Diagram of types of distributional data for the species of interest and complexity of the modelling technique necessary to obtain estimations of the realized (RD; *black*) or the potential (PD; *light grey*) distribution of species

These distinctions are to some degree naïve, as types of data suitable for different methods overlap broadly, but they help to highlight what we believe is a relevant polarity: techniques that play down the importance of absence information may be better suited to estimation of the ecological and distributional potential of the species, whereas methods incorporating absence information more directly may be more suitable for estimating actual distributions of species (Jiménez-Valverde et al. 2008; Sutherst and Bourne 2009; Fig. 3). Contingent factors discussed above (e.g., biotic interactions, historical events, disturbance, metapopulation dynamics, dispersal limitation) can cause absences of species from areas that are environmentally suitable in the native region (contingent absences; see Fig. 2). Such absences should ideally not be considered in modelling if the goal is to estimate the complete area at risk for invasion. In any case, the only absence information that could be included in estimating the distributional potential of invaders is that obtained from localities with environmental conditions that negate unequivocally the presence of the species, and that ostensibly come from within the dispersal potential of the species (Engler et al. 2004; Lobo et al. 2006, 2010). In the BAM terminology of Soberón and Peterson (2005), the only relevant areas would be within *M* but outside of *A* (Fig. 1). Inclusion of such absences, especially if they are located near to the boundary of suitable environmental conditions, can facilitate correct classification of the absence zone by providing explicit contrasts for the modelling algorithm (Thuiller et al. 2004; Austin 2007; Lobo et al. 2010). Absences that are environmentally very distinct from the domain defined by presences are safest, but are not particularly informative (Lobo et al. 2010). However, discerning which sites fall into each category of absence is challenging (Lobo et al. 2010). Indeed, even if contingent absences are not used explicitly in model development, this drawback exists, because even presence-only methods are not able to include among the presence data the favourable areas submitted to these restrictive forces. This shortcoming is especially important in describing the niche of the species when the regions in which these contingent absences appear have particular environmental conditions not present in the observed area of distribution.

Environmental variables for modelling

To estimate the niche of a species, one would ideally know beforehand the specific environmental factors that limit its distribution and determine its fitness, as well as the shapes of its response curves and tolerance limits. However, such detailed knowledge is lacking for the vast majority of species, and may be generated only by detailed physiological and ecological experiments (Kearney and Porter 2004; Kearney 2006; Kilroy et al. 2007). As a consequence, most users resort not to the most appropriate environmental variables, but rather to those that are available in digital formats across broad geographic extents (e.g., Hijmans et al. 2005). Users also generally rely on automatic parameterization and variable-selection procedures to train correlative niche models, with the expectation that the variables selected by the algorithm will serve as effective surrogates of the genuinely causal ones.

Therefore, some thought should be put into the process, to assure that the resulting models are robust.

Indirect variables (i.e., those not having direct effects on the physiology of the species) should be avoided, because associations between these variables and true causal variables can frequently be location-specific (Austin 2007). As an example, we can consider the case of elevation, which is often closely correlated with species' occurrences, and can easily be incorporated into niche models by automatic variable-selection procedures, but that has a physiological 'meaning' that varies dramatically (3,000 m elevation in the Tropics holds cloud forest, whereas the same elevation near the poles would be covered with snow and ice). Rödder et al. (2009) showed that models based on variables related directly to the ecology of the species had higher discrimination capacity that models based on a random subset of variables when predicting invaded ranges. A promising exploratory procedure to identify major environmental requirements of species consists of using Ecological-Niche Factor Analysis or analogous procedures (ENFA; see Hirzel et al. 2002; Basille et al. 2008; Calenge and Basille 2008; Calenge et al. 2008). By examining the response of the species to environmental variables across the study area, ENFA calculates both marginality (the ecological distance between the species' optimum and the mean conditions within the study area) and specialization (the ratio of the ecological variance in mean conditions to that associated with the focal species) of species against potential predictors, identifying those most likely to limit the distribution. Variables thus identified can then be prioritized for inclusion in ecological niche models to estimate the potential of colonization, but inclusion of variables that are closely related to contingent effects such as dispersal barriers and biotic interactions should be avoided (Lobo 2008b).

Given that niches can be complex, but sample sizes of occurrence data are often small, overfitting must be avoided carefully (Peterson and Nakazawa 2008). When overfitting, models may reconstruct training data sets well, but not predict independent testing data accurately; such models should not be used to transfer geographic projections onto novel regions. Modellers thus should avoid overparameterization, and use environmental spaces that are not overly dimensional (Peterson 2005b; Peterson and Nakazawa 2008). Environmental variables are frequently highly correlated, which may have complex effects in automatic variable-selection procedures (Feinstein 1996). Hence, preliminary multi-collinearity analyses, factor analyses, or other steps to reduce dimensionality, may help in selecting appropriate predictors (e.g., Jiménez-Valverde et al. 2007, 2011). These steps not only simplify the already-complex procedures of model development, but also assist in avoiding overfitting. An additional consideration in order to avoid overfitting is not to consider too much complex relationships between the occurrence of the species and the predictors (Jiménez-Valverde et al. 2008) as this would only increase the chances of accounting for contingent and local factors that would hamper the transferability capacity of the models.

Evaluating model predictions

Niche models must be evaluated rigorously prior to any use in forecasting or risk mapping (Peterson 2005b). Generally, in invasive species applications, testing is or should be carried out first within the training region, to assure significant predictive ability before any transfer of model parameters to other regions. Ideally, such initial tests are based on spatially stratified subsets of the training area, to present serious predictive challenges to the models (Hortal et al. 2007). Simple random subsetting or cross-validation does not provide a particularly difficult challenge for model testing, given that spatial autocorrelation reduces independence of training and testing data sets.

Because of the potential distribution emphasis in niche modelling, it is necessary to weight Type I and Type II statistical errors differently (Peterson et al. 2008). In particular, maximization of sensitivity (i.e., presences correctly predicted as presences) is desired (Anderson et al. 2003). Absences, as we have discussed above, can be caused by several contingent factors, so specificity (i.e., absences correctly predicted as absences) should be of less concern if the focus of the forecasting effort is on identifying localities that are suitable environmentally for the species in question. Put another way, overprediction (false positives), commonly cited as a problem with niche model results (Fielding and Haworth 1995; Araújo and Williams 2000; Stockwell and Peterson 2002; Brotons et al. 2004; Stockman et al. 2006) should be understood as a necessary part of the niche modelling process, when it is aimed at estimating the potential distribution area. Indeed, this 'error' is not a shortcoming when the goal is to identify areas susceptible to colonization by invasive species (Peterson 2006); rather, the fact that species rarely inhabit the entire spatial footprint of their ecologically suitable potential distributions should be acknowledged (Peterson 2003), and models trained, interpreted, and utilized accordingly. This is in fact a complex issue for the case of invasive species, because the pool of false positives may frequently include both suitable areas not yet reached by the species (not-so-false positives) and true errors inherent in the data or the modelling technique (real false positives). A partial assessment of more reliable error might be detecting those false positives that fall outside the physiological tolerances of the species, which can be approached by performing spatial projections of physiological information coming from empirical studies (e.g., Aragón et al. 2010).

If the invasion has already occurred, or if the species has already invaded some other region beside the region of interest, presence data may also be available from invaded regions, permitting tests of model predictive success there as well. Here again, the species may frequently not be in distributional equilibrium, so focus should also be on optimizing sensitivity, examining carefully the error rate of presences, since this measure can inform about our capacity to estimate the niche of species from distributional data. Screening distributions of omission errors (false absences) in the invaded region for spatially non-random concentrations can help greatly in detecting possible confounding variables that should be added to or subtracted from models (Miller et al. 2004). These ideas echo recent recommendations based on theoretical considerations that different kind of errors (omission vs. commission) should be accorded different weight in tests of distributional models depending on the question under consideration (Soberón and Peterson 2005; Jiménez-Valverde et al. 2008; Lobo et al. 2008; Peterson et al. 2008).

If enough data are available, we suggest a multistep approach for evaluation of niche models for invasive species based on training and validating across multiple regions (Fitzpatrick et al. 2007). For instance, consider a species native to region X that is already present in regions Y and Z as an invader, and that is of interest regarding invasive potential in region W. Initial validation would be carried out within region X, but then would pass on to a *k*-fold regional cross-validation—here, models trained in X and Y would be tested in Z, models trained in X and Z would be tested in Y, and models trained in Y and Z would be tested in X. In this way, it is possible to evaluate quantitatively the possibility and magnitude of effects of competitive release or other factors possibly broadening the species' ecological potential as an invader. Finally, all three current distributional areas can be used to train an overall model to be projected onto region W of interest as a forecast.

The regionalized *k*-fold cross-validation recommended here is a variation on the customary validation or cross-validation approach, but one in which data subsets are chosen spatially rather than randomly. In a normal *k*-fold cross-validation, data are split randomly into *k* subsets, the model trained based on $k - 1$ subsets, and predictions tested on the remaining subset; the process is repeated until each subset has been used for testing. In this case, we subset the data regionally, using each major distributional area as a subset. This approach mimics the forecasting situation almost exactly: one has knowledge of the species' ecological 'behaviour' in one or several regions, but would like to forecast its behaviour in yet another—the *k*-fold partitioning manipulation directly tests the ability of the models to achieve this goal. At the end of the process, one has a quantitative measure of ability to predict the species' invasive potential in *k* regions, and can thereby evaluate one's likely success in understanding its invasion of a $(k + 1)$th region.

Evaluation of model predictions cannot be reduced to a simple discrimination power analysis. Among the many sources of uncertainty that affect niche modelling (Elith et al. 2002; Rocchini et al. 2010), two are of special relevance to invasion risk assessment. First, extrapolation beyond environmental limits in the training data is a dangerous and potentially unreliable practice; thus, it is advisable to highlight areas where the niche models are being applied to novel combinations of environmental variables, such that their predictions may have a relatively high degree of uncertainty (Fitzpatrick and Hargrove 2009; Elith et al. 2010; Jiménez-Valverde et al. 2011). Second, when correlational modelling methods are used, the correlation structure among environmental variables must be similar across regions (Jiménez-Valverde et al. 2009). The constancy of the correlation structure can be assessed using, for example, Mantel tests between correlation matrices (e.g., Jiménez-Valverde et al. 2011); further ideas for mapping the information associated to this source of uncertainty can be found in Elith et al. (2010).

Concluding remarks

We do not recommend rejecting taxonomically *reliable* occurrence data based on environmental novelty, as very important information regarding the full environmental domain that the species can inhabit may be discarded. Predictions of potential distributions of invasive species should be considered as risk maps, in which *overprediction* is a desirable property that reflects the very nature of invasive species. Predicting actual distributions require true-absence information (Jiménez-Valverde et al. 2008; Václavík and Meentemeyer 2009; Lobo et al. 2010), but we suggest avoiding the use of absence data in estimating potential distributions as those required in invader risk maps; because they can misrepresent actually suitable areas, erroneously transfer specific contingent factors, transfer drawbacks from one region to another, and bias model predictions. If possible, independent variables used in modelling should be linked as directly as possible to the physiology (or other mechanistic processes) of the target species. Although recent tendencies in niche modelling have been towards more complex modelling approaches (Elith et al. 2006), relatively simple techniques can be most adequate to estimate the geographic representation of the fundamental niche (see Sutherst and Bourne 2009). Finally, focus should be on generating a geographic representation of the potential distribution of the invader, which requires special evaluation and testing procedures: in absence of physiological data, independent data documenting presences correctly predicted as presences should be more powerful to estimate the accuracy of such models than other information. Despite our emphasis on the importance of maximizing the information content of the occurrence data and the preference of overprediction as opposed to underprediction when dealing with invasive species, the costs and losses of both types of errors must be assessed carefully. For example, if the costs of preventive actions are much higher than losses caused by the species' invasion, then overprediction may be not as valuable as in the opposite situation.

Niche models can be extremely useful in exploring a number of biodiversity phenomena, but each application has different requirements in terms of input information, output content, and interpretation of results (Peterson 2006; Jiménez-Valverde et al. 2008). We have discussed important considerations for forecasting potential distributions of invasive species in novel regions. As we have seen, uncertainty is a key feature in the process. The fact that distributions inform only partially about species' niches simply because distributions are the spatial projections of many different contingent forces, difficulty in selection of appropriate predictors, filtering of occurrence data, and variation in the results depending on the modelling technique, makes necessary the collaboration between specialists on different disciplines.

Acknowledgments This paper was supported by the Spanish MEC project CGL2004-0439/BOS, a Fundación BBVA Project, and the European Distributed Institute of Taxonomy (EDIT) project. AJ-V was supported by a MEC (Ministerio de Educación y Ciencia, Spain) postdoctoral fellowship (Ref.: EX-2007-0381) and the MEC Juan de la Cierva Program. ATP and JS were supported by a grant from Microsoft Research.

References

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

Anderson RP, Lew D, Peterson AT (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol Model 162:211–232

Angert AL (2009) The niche, limits to species' distributions, and spatiotemporal variation in demography across the elevation ranges of two monkeyflowers. Proc Natl Acad Sci USA 106:19693–19698

Aragón P, Baselga A, Lobo JM (2010) Global estimation of invasion risk zones for the western corn rootworm *Diabrotica virgifera virgifera*: integrating distribution models and physiological thresholds to assess climatic favourability. J Appl Ecol 47:1026–1035

Araújo MB, Williams PH (2000) Selecting areas for species persistence using occurrence data. Biol Conserv 96:331–345

Austin M (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol Model 200:1–19

Basille M, Calenge C, Marboutin E et al (2008) Assessing habitat selection using multivariate statistics: some refinements of the ecological-niche factors analysis. Ecol Model 211:233–240

Bax N, Carlton JT, Mathews-Amos A et al (2001) The control of biological invasions in the world's oceans. Conserv Biol 15:1234–1246

Beaumont LJ, Gallagher RV, Thuiller W et al (2009) Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. Divers Distrib 15:409–420

Bergmans W, Blom E (2001) Invasive plants and animals. Is there a way out? The Netherlands Committee for IUCN, Amsterdam

Broennimann O, Guisan A (2008) Predicting current and future biological invasions: both native and invaded ranges matter. Biol Lett 4:585–589

Broennimann O, Treier UA, Müller-Schärer H et al (2007) Evidence of climatic niche shift during biological invasion. Ecol Lett 10:701–709

Brotons L, Thuiller W, Araújo MB et al (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27:437–448

Bruno JF, Stachowicz JJ, Bertness MD (2003) Inclusion of facilitation into ecological theory. Trends Ecol Evol 18:119–125

Busby JR (1991) BIOCLIM: a bioclimate analysis and prediction system. In: Margules CR, Austin MP (eds) Nature conservation: cost effective biological surveys and data analysis. CSIRO, Melbourne, pp 64–68

Calenge C, Basille M (2008) A general framework for the statistical exploration of the ecological niche. J Theor Biol 252:674–685, 543

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

Calosi P, Bilton DT, Spicer JI et al (2010) What determines a species' geographical range? Thermal biology and latitudinal range size relationships in European diving beetles (Coleoptera, Dytiscidae). J Anim Ecol 79:194–204

Cassey P, Blackburn TM, Sol D et al (2004) Global patterns of introduction effort and establishment success in birds. Proc R Soc Lond B Biol 271:S405–S408

Chase JM, Leibold M (2003) Ecological niches: linking classical and contemporary approaches. University of Chicago Press, Chicago

Colwell RK (1992) Niche: a bifurcation in the conceptual lineage of the term. In: Keller EF, Lloyd EA (eds) Keywords in evolutionary biology. Harvard University Press, Cambridge, pp 241–248

Colwell RK, Rangel TF (2009) Hutchinson's duality: the once and future niche. Proc Natl Acad Sci USA 106:19651–19658

Davis MB, Shaw RG (2001) Range shifts and adaptive responses to Quaternary climate change. Science 292:673–679

Davis MA, Grime JP, Thompson K (2000) Fluctuating resources in plant communities: a general theory of invasibility. J Ecol 88:528–534

Dennis RLH, Hardy PB (1999) Targeting squares for survey: predicting species richness and incidence of species for a butterfly atlas. Global Ecol Biogeogr 8:443–454

Elith J, Burgman MA, Regan HM (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. Ecol Model 157:313–329

Elith J, Graham CH, Anderson RP et al (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151

Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. Methods Ecol Evol 1:330–342

Engler R, Guisan A, Rechsteiner L (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J Appl Ecol 41:263–274

Farber O, Kadmon R (2003) Assessment of alternative approaches for bioclimatic modelling with special emphasis on the Mahalanobis distance. Ecol Model 160:115–130

Feinstein AR (1996) Multivariable analysis: an introduction. Yale University Press, New Haven

Ficetola GF, Thuiller W, Miaud C (2007) Prediction and validation of the potential global distribution of a problematic alien invasive species, the American bullfrog. Divers Distrib 13:476–485

Fielding AH, Haworth PF (1995) Testing the generality of bird-habitat models. Conserv Biol 9:1466–1481

Fitzpatrick MC, Hargrove WW (2009) The projection of species distribution models and the problem of non-analog climate. Biodivers Conserv 18:22–2261

Fitzpatrick MC, Weltzin JF, Sanders N et al (2007) The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? Global Ecol Biogeogr 16:24–33

Ganeshaiah KN, Barve N, Nath K et al (2003) Predicting the potential geographical distribution of the sugarcane woolly aphid using GARP and DIVA-GIS. Curr Sci India 85:1526–1528

Gaston KJ (2003) The structure and dynamics of geographic ranges. Oxford University Press, Oxford

Godsoe W (2010) I can't define the niche but I know it when I see it: a formal link between statistical theory and the ecological niche. Oikos 119:53–60

Gomulkiewicz R, Holt RD, Barfield M (1999) The effects of density dependence and immigration on local adaptation and niche evolution in a black-hole sink environment. Theor Popul Biol 55:283–296

Gorodkov KB (1986a) Three-dimensional climatic model of potential range and some of its characteristics I. Entomol Rev 65:1–18

Gorodkov KB (1986b) Three-dimensional climatic model of potential range and some of its characteristics II. Entomol Rev 65:19–35

Graham CH, Ferrier S, Huettman F et al (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol Evol 19:497–503

Groves RH, Di Castri F (1996) Biogeography of mediterranean invasions. Cambridge University Press, Cambridge, p 495

Gu W, Swihart RK (2004) Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. Biol Conserv 116:195–203

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecol Model 135:147–186

Hanski I (1998) Metapopulation dynamics. Nature 396:41–49

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25:1965–1978

Hirzel AH, Helfer V, Metral F (2001) Assessing habitat-suitability models with a virtual species. Ecol Model 145: 111–121

Hirzel AH, Hausser J, Chessel D et al (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83:2027–2036

Hobbs RJ, Humphries SE (1995) An integrated approach to the ecology and management of plant invasions. Conserv Biol 9:761–770

Hortal J, Lobo JM (2005) An ED-based protocol for optimal sampling of biodiversity. Biodivers Conserv 14: 2013–2947

Hortal J, Lobo JM, Jiménez-Valverde A (2007) Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. Conserv Biol 21:853–863

Hortal J, Jiménez-Valverde A, Gómez J et al (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117:847–858

Hutchinson GE (1978) An introduction to population ecology. Yale University Press, New Haven

Jackson ST, Overpeck JT (2000) Responses of plant populations and communities to environmental changes of the late Quaternary. Paleobiology 26:194–220

Jiménez-Valverde A, Lobo JM (2011) Tolerance limits, animal. In: Simberloff D, Rejmánek M (eds) Encyclopedia of biological invasions. University of California Press, CA, pp 661–663

Jiménez-Valverde A, Ortuño V, Lobo JM (2007) Exploring the distribution of *Sterocorax* Ortuño, 1990 (Coleoptera, Carabidae) species in the Iberian Peninsula. J Biogeogr 34:1426–1438

Jiménez-Valverde A, Lobo JM, Hortal J (2008) Not as good as they seem: the importance of concepts in species distribution modelling. Divers Distrib 14:885–890

Jiménez-Valverde A, Nakazawa Y, Lira-Noriega A, Peterson AT (2009) Environmental correlation structure and ecological niche model projections. Biodivers Inf 6:28–35

Jiménez-Valverde A, Lira-Noriega A, Peterson AT, Soberón J (2010) Marshalling existing biodiversity data to evaluate biodiversity status and trends in planning exercises. Ecol Res 25:947–957

Jiménez-Valverde A, Decae AE, Arnedo MA (2011) Environmental suitability of new reported localities of the funnelweb spider *Macrothele calpeiana*: an assessment using potential distribution modelling with presence-only techniques. J Biogeogr (in press)

Kadmon R, Oren F, Avinoam D (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecol Appl 14:401–413

Keane RM, Crawley MJ (2002) Exotic plant invasions and the enemy release hypothesis. Trends Ecol Evol 16:199–204

Kearney M (2006) Habitat, environment and niche: what are we modelling? Oikos 115:186–191

Kearney M, Porter WP (2004) Mapping the fundamental niche: physiology, climate and the distribution of a nocturnal lizard. Ecology 85:3119–3131

Kearney M, Porter WP, Williams C et al (2009) Integrating biophysical models and evolutionary theory to predict climatic impacts on species' ranges: the dengue mosquito *Aedes aegypti* in Australia. Funct Ecol 23:528–538

Kilroy C, Snelder TH, Floerl O et al (2007) A rapid technique for assessing the suitability of areas for invasive species to New Zealand's rivers. Divers Distrib 14:262–272

Kluza DA, Vieglais DA, Andreasen JK et al (2007) Sudden oak death: geographic risk estimates and predictions of origins. Plant Pathol 56:580–587

Kolar CS, Lodge DM (2002) Ecological predictions and risk assessment for alien fishes in North America. Science 298:1233–1236

Lambdon PW, Lloret F, Hulme PE (2008) Do alien plants on Mediterranean islands tend to invade different niches from native species? Biol Invasions 10:703–716

Leung B, Drake JM, Lodge DM (2004) Predicting invasions: propagule pressure and the gravity of Allee effects. Ecology 85:1651–1660

Lobo JM (2008a) Database records as a surrogate for sampling effort provide higher species richness estimations. Biodivers Conserv 17:873–881

Lobo JM (2008b) More complex distribution models or more representative data? Biodivers Informatics 5:14–19

Lobo JM, Verdú JR, Numa C (2006) Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). Divers Distrib 12:179–188

Lobo JM, Baselga A, Hortal J et al (2007) How does the knowledge on the spatial distribution of species increase? Divers Distrib 13:772–780

Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Global Ecol Biogeogr 17:145–151

Lobo JM, Jiménez-Valverde A, Hortal J (2010) The uncertain nature of absences and their importance in species distribution modelling. Ecography 33:103–114

Lockwood JL, Hoopes MF, Marchetti MP (2007) Invasion ecology. Blackwell, Oxford

López-Darias M, Lobo JM, Gouat P (2008) Predicting potential distributions of invasive species: the exotic Barbary ground squirrel in the Canarian archipelago and the west Mediterranean region. Biol Invasions 10:1027–1040

Mack RN (2004) Global plant dispersal, naturalization, and invasion: pathways, modes and circumstances. In: Ruiz GM, Carlton JT (eds) Invasive species: vectors and management strategies. Island Press, Washington DC, pp 3–30

MacKenzie DI, Nichols JD, Lachman GB et al (2002) Estimating site occupancy when detection probabilities are less than one. Ecology 83:2248–2255

Maron JL, Vilà M, Bommarco R et al (2004) Rapid evolution of an invasive plant. Ecol Monogr 74:261–280

Mau-Crimmins TM, Schussman HR, Geiger EL (2006) Can the invaded range of a species be predicted sufficiently using only native-range data? Lehmann lovegrass (*Eragrostis lehmanniana*) in the southwestern United States. Ecol Model 193:736–746

Medley KA (2010) Niche shifts during the global invasion of the Asian tiger mosquito, *Aedes albopictus* Skuse (Culicidae), revealed by reciprocal distribution models. Global Ecol Biogeogr 19:122–133497

Miller JR, Turner MG, Smithwick EAH et al (2004) Spatial extrapolation: the science of predicting ecological patterns and processes. Bioscience 54:310–320

Muñoz AR, Real R (2006) Assessing the potential range expansion of the exotic monk parakeet in Spain. Divers Distrib 12:656–665

Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. Q Rev Biol 83:171–193

Orr MR, Smith TB (1998) Ecology and speciation. Trends Ecol Evol 13:502–506

Parker IM, Simberloff D, Lonsdale WM et al (1999) Impact: toward a framework for understanding the ecological effects of invaders. Biol Invasions 1:3–19

Peterson AT (2003) Predicting the geography of species invasion via ecological niche modelling. Q Rev Biol 78:419–433

Peterson AT (2005a) Predicting potential geographic distributions of invading species. Curr Sci India 89:9

Peterson AT (2005b) Kansas Gap analysis: the importance of validating distributional models before using them. Southwest Nat 50:230–236

Peterson AT (2006) Uses and requirements of ecological niche models and related distributional models. Biodivers Inf 3:59–72

Peterson AT, Nakazawa Y (2008) Environmental data sets matter in ecological niche modelling: an example with *Solenopsis invicta* and *Solenopsis richteri*. Global Ecol Biogeogr 17:135–144

Peterson AT, Vieglais DA (2001) Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. Bioscience 51:363–371

Peterson AT, Papeş M, Soberón J (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecol Model 213:63–72

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modelling of species geographic distributions. Ecol Model 190:231–259

Pulliam HR (1988) Sources, sinks and population regulation. Am Nat 132:652–661

Pulliam HR (2000) On the relationship between niche and distribution. Ecol Lett 3:349–361

Reese GC, Wilson KR, Hoeting JA et al (2005) Factors affecting species distribution predictions: a simulation model experiment. Ecol Appl 15:554–564

Richardson DM, Thuiller W (2007) Home away from home objective mapping of high-risk source areas for plant introductions. Divers Distrib 13:299–312

Ricklefs RE, Schluter D (1993) Species diversity in ecological communities: historical and geographical perspectives. University of Chicago Press, Chicago

Rocchini D, Hortal J, Lengyel S, Lobo JM, Jiménez-Valverde A, Ricotta C, Bacaro G, Chiarucci A (2010) Uncertainty in species distribution mapping and the need for maps of ignorance. Progr Phys Geogr (in press)

Rödder D, Schmidtlein S, Veith M, Lötters S (2009) Alien invasive slider turtle in unpredicted habitat: a matter of niche shift or of predictors studied? PLoS ONE 4:e7843

Rodríguez-Trelles F, Rodríguez MA (1998) Rapid microevolution and loss of chromosomal diversity in *Drosophila* in response to climate warming. Evol Ecol 12:829–838

Rouget M, Richardson DM, Nel JL et al (2004) Mapping the potential ranges of major plant invaders in South Africa, Lesotho and Swaziland using climatic suitability. Divers Distrib 10:475–484

Royle JA, Nichols JD, Kéry M (2005) Modelling occurrence and abundance of species when detection is imperfect. Oikos 110:353–359

Shigesada N, Kawasaki K (1997) Biological invasions: theory and practice. Oxford University Press, Oxford

Simberloff D, Parker IM, Windle PN (2005) Introduced species policy, management, and future research needs. Front Ecol Environ 3:12–20

Soberón J (2007) Grinnellian and Eltonian niches and geographic distributions of species. Ecol Lett 10:1115–1123

Soberón J, Nakamura M (2009) Niches and distributional areas: concepts, methods, and assumptions. Proc Natl Acad Sci USA 106:19644–19650

Soberón J, Peterson AT (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. Biodivers Inf 2:1–10

Steiner FM, Schlick-Steiner BC, VanDerWal J et al (2008) Combined modelling of distribution and niche in invasion biology: a case study of two invasive *Tetramorium* ant species. Divers Distrib 14:538–545

Stockman AK, Beamer DA, Bond JE (2006) An evaluation of a GARP model as an approach to predicting the spatial distribution of non-vagile invertebrate species. Divers Distrib 12:81–89

Stockwell DRB, Peters D (1999) The GARP modeling system: problems and solutions to automated spatial prediction. Int J Geogr Inf Sci 13:143–158

Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Model 148:1–13

Sutherst RW, Bourne AS (2009) Modelling non-equilibrium distributions of invasive species: a tale of two modeling paradigms. Biol Invasions 11:1231–1237

Svenning J-C, Skov F (2004) Limited filling of the potential range in European tree species. Ecol Lett 7:565–573

Thomas CD, Bodsworth EJ, Wilson RJ et al (2001) Ecological and evolutionary processes at expanding range margins. Nature 411:577–581

Thuiller W, Brotons L, Araújo MB et al (2004) Effects of restricting environmental range of data to project current and future species distributions. Ecography 27: 165–172

Thuiller W, Richardson DM, Pysek P et al (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. Glob Change Biol 11: 2234–2250

Václavík T, Meentemeyer RK (2009) Invasive species distribution modeling (iSDM): are absence data and dispersal constraints needed to predict actual distributions? Ecol Model 220:3248–3258

Varela S, Rodríguez J, Lobo JM (2009) Is current climatic equilibrium a guarantee for the transferability of distribution model predictions? A case study of the spotted hyena. J Biogeogr 36:1645–1655

Welk E (2004) Constraints in range predictions of invasive plant species due to non-equilibrium distribution patterns:

purple loosestrife (*Lythrum salicaria*) in North America. Ecol Model 179:551–567

Williamson M (1996) Biological invasions. Chapman & Hall, London

Williamson M (1999) Invasions. Ecography 22:5–12

Zaniewski AE, Lehmann A, Overton JM (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecol Model 157:261–280