Journal of Applied Ecology 2005
42, 720–730

FORUM

The continuing challenges of testing species distribution models

I. P. VAUGHAN and S. J. ORMEROD
Catchment Research Group, Cardiff School of Biosciences, Cardiff University, Main Building, Cardiff CF10 3TL, UK

Summary

1. Species distribution models could bring manifold benefits across ecology, but require careful testing to prove their reliability and guide users. Shortcomings in testing are often evident, failing to reflect recent methodological developments and changes in the way models are applied. We considered some of the fundamental issues.
2. Generalizability is a basic requirement for predictive models, describing their capacity to produce accurate predictions with new data, i.e. in real applications beyond model training. Tests of generalizability should be as rigorous as possible: ideally using a large number of independent test sites (≥ 200–300) that represent anticipated applications. Bootstrapping identifies the role of overfitting of the training data in limiting a model's generalizability.
3. Predictions from most distribution models are continuous variables. Their accuracy may be described by discrimination and calibration components. Discriminatory ability describes how well a model separates occupied from unoccupied sites. It is independent of species prevalence and is readily comparable between models. Rank correlation coefficients, such as the concordance index, are effective measures.
4. Calibration describes the numerical accuracy of predictions (e.g. whether 40% of sites with predicted probabilities of 0·40 are occupied) but is frequently overlooked in model testing. Poor calibration could mislead any conservation efforts utilizing models to estimate the 'value' of different sites for a given species. Effective assessments can be made using smoothed calibration plots.
5. The effects of species prevalence on nominal presence–absence predictions are well known. The currently preferred accuracy measure, Cohen's κ, has weaknesses. We argue that mutual information measures, based in information theory, may be more appropriate.
6. *Synthesis and applications*. Model evaluation must be informative and should ideally: (i) define generalizability in detail; (ii) separate the discrimination and calibration components of accuracy and test both; (iii) adopt assessment techniques that permit more valid intermodel comparisons; (iv) avoid nominal presence–absence evaluation where possible and consider information-theoretic measures; and (v) utilize the full range of techniques to help diagnose the causes of prediction problems. Few modellers in applied ecology and conservation biology satisfy these needs, making it difficult for others to evaluate models and identify potential misuses. The problems are real, and if uncorrected will damage conservation efforts through the inaccurate assessment of distribution and habitat preferences of important organisms.

Key-words: bootstrapping, calibration, discrimination, generalizability, overfitting, presence–absence data, transportability

Journal of Applied Ecology (2005) **42**, 720–730
doi: 10.1111/j.1365-2664.2005.01052.x

Correspondence: Dr Ian Vaughan, Catchment Research Group, Cardiff School of Biosciences, Cardiff University, Main Building, Cardiff CF10 3TL, UK (e-mail vaughanip@cardiff.ac.uk).

Introduction

Species distribution models are proposed with increasing frequency throughout ecology and conservation biology. Some are focused on improved understanding of species' habitat requirements, whereas many are intended as applied predictive tools (Rushton, Ormerod & Kerby 2004). The potential benefits are manifold (Manel, Williams & Ormerod 2001, table 1), and as a consequence distribution modelling is frequent in applied ecological journals. In the *Journal of Applied Ecology*, around 10% of papers over the 5 years 1999–2003 incorporated distribution modelling. Methodological developments are a frequent theme amongst the *Journal of Applied Ecology's* papers (Buckland & Elston 1993; Manel, Williams & Ormerod 2001; Suárez-Seoane, Osborne & Alonso 2002; McPherson, Jetz & Rogers 2004), culminating in a recent special profile (Volume 41, issue 2; Rushton, Ormerod & Kerby 2004; Frair *et al*. 2004; Gibson *et al*. 2004; Jeganathan *et al*. 2004; Johnson, Seip & Boyce 2004; Cabeza *et al*. 2004; Engler, Guisan & Rechsteiner 2004). Across ecology more generally, the testing of models developed from presence–absence data has been a recurrent focus (Fielding & Bell 1997; Pearce & Ferrier 2000; Manel, Williams & Ormerod 2001; Boyce *et al*. 2002) and this forum represents a further contribution.

Estimates of predictive performance may be used during model development and to assess final model quality (Hastie, Tibshirani & Friedman 2001). The former is illustrated by measures such as Akaike's information criterion, which facilitate model selection by estimating the relative error rates for alternative models built using the same data set (Hastie, Tibshirani & Friedman 2001). Our focus is on the latter: estimating the predictive ability expected from a model in subsequent predictive applications. Such testing involves comparing a sample of a model's predictions against the observed species' distributions. Conceptually this is very simple and represents an effective and universal basis for testing applicable to all types of distribution model. It can provide meaningful estimates of model accuracy and describe how it varies according to the conditions under which the model is applied, providing essential guidance to both the developers and users of a model. Analyses of test performance may help to diagnose problems (Miller, Hui & Tierney 1991). Moreover, testing provides an opportunity to convince ecologists and end users of a model's value by demonstrating its efficacy (Pearce, Ferrier & Scotts 2001).

Whilst conceptually simple, model testing incorporates a range of technical issues regarding which intensive study has been carried out over the last 20 years. Much of this has occurred in medical statistics, and is spread across a wide range of sources, making overall syntheses in ecology important. Even in the short time since some of the seminal ecological model evaluation papers were published (Fielding & Bell 1997; Pearce & Ferrier 2000), important developments have occurred. For example, a recent discussion in medicine has clarified key concepts in testing the wider applicability of distribution models (Justice, Covinsky & Berlin 1999), whilst the recent paper of McPherson, Jetz & Rogers (2004) necessitates a reconsideration of the use of the κ statistic in ecology, an issue that was widely thought to be resolved. In addition, there have been subtle shifts in the way species distribution models are used. Predictions of the probability of species occurrence are increasingly favoured over nominal presence–absence predictions, and this has implications for model evaluation. In this forum we provide an up-to-date discussion of both general concepts and technical issues in the evaluation of species distribution models. Our aims were to raise awareness of important issues that may have made limited inroads into ecology, and provide recommendations regarding what appears to be current best practice.

Generalizability, test data and the role of resampling methods

Generalizability is a model's capacity to predict a species' distribution as accurately with new data as it does with its own training data (Justice, Covinsky & Berlin 1999). Equally, if quantifying species' habitat preferences is the primary aim of a model (cf. prediction), generalizability describes the broader applicability of the correlations modelled in the training data. It is a vital property for all predictive species distribution models, as their basic purpose is to be applied to new data. Whilst testing cannot provide absolute proof that a model will generalize to a subsequent application, it should be possible to obtain good evidence of its likely efficacy. The strength of such evidence relies directly upon the test data that are collected.

TEST DATA

Independent data are the only rigorous test of a model's generalizability (Chatfield 1995). Surprisingly little work has been carried out to identify the basic requirements for an independent test set (Altman & Royston 2000), but two issues can be highlighted. The first is to maximize the data's representativeness of the applications anticipated for a model. Such applications may differ from the training data in a range of ways (Table 1) and, because the effects of such differences cannot be predicted *a priori*, representative test data are required. For models quantifying habitat preferences, these applications are the range of conditions over which there is a desire to draw inference from the model. The second requirement from a test set is a sufficient sample size to provide precise estimates of performance.

Representativeness of test data

The ideal way to measure a model's generalizability is to collect test data from each of its proposed applications, thereby simulating its use. For example, a sample

Table 1. Potential differences between training data and the subsequent applications for a distribution model. Partly adapted from Justice, Covinsky & Berlin (1999)

Temporal separation	Repeat use of a model at subsequent intervals (e.g. different breeding seasons; Boyce *et al*. 2002) or the prediction of historical or future distributions (e.g. under alternative climate change scenarios; Berry *et al*. 2002)
Geographical separation	Extrapolation to new regions (e.g. different islands; Fielding & Haworth 1995) rather than interpolation between training sites (e.g. Suárez-Seoane, Osborne & Alonso 2002)
Different regions of environmental space	Often related to differences in geographical space. Different ranges of individual environmental gradients or more complex suites of changes (e.g. separate streams with different fluvial characteristics; D'Angelo *et al*. 1995)
Deployment factors	Miscellaneous factors that may affect the performance of a model, such as different users (e.g. the RIVPACS system; Wright 1995) or different data sources/collection methods (e.g. survey data collected for other purposes; Pearce, Ferrier & Scotts 2001)

of test sites could be collected from every island to which a model might be applied (Marsden & Fielding 1999). Successful prediction would provide a high degree of confidence in the model's subsequent applicability and/or the generality of its conclusions. To make testing more informative it may be valuable to sample some areas densely, ideally blocks of contiguous sites, to identify any 'biotic' errors (Fielding & Bell 1997). Processes such as territoriality may affect species' occurrences at neighbouring sites, despite equally suitable physical habitat, leading to characteristic spatial error patterns (Fielding & Bell 1997). Relatively dense sampling would be required to characterize such phenomena and to provide appropriate guidance to the model's users.

Unfortunately, models are often required for situations where they cannot be tested directly, presenting a major challenge. Common examples are the prediction of future distribution patterns, such as species' responses to climate change, and predictions in very remote areas, for which resources are insufficient to allow test data collection. In such situations it will never be possible to attain the same level of confidence as sampling the actual applications. Nevertheless, if a model is shown to generalize successfully to test data collected under a wide range of conditions, greater confidence may be obtained in its generalizability to situations that cannot be sampled (Justice, Covinsky & Berlin 1999). Even testing that seems to be a poor surrogate for subsequent model applications can often identify weaknesses in a model that are later confirmed when representative data are obtained (Charlson *et al*. 1987).

On this basis, we recommend a two-stage approach to testing generalizability when actual applications cannot be sampled. The first stage is to test the model as widely as possible, incorporating situations that resemble future applications as closely as possible. The second stage is to state clearly the conditions under which the model was tested, compared with its anticipated applications. This should indicate the rigour of testing and make limitations to generalizability testing explicit.

Resampling methods (e.g. bootstrapping; Verbyla & Litvaitis 1989) may be the only feasible option for model testing where resources would otherwise have to be diverted from collecting training data, or when very few independent data could be collected such that accuracy estimates would be unreliable (see below). Resampling can also be useful when models are built primarily to investigate species–environment correlations (cf. prediction) if a data set covers the full range of conditions over which inference is of interest (e.g. samples taken across the UK; Gates & Donald 2000). Resampling's weakness is that it only examines model performance under identical conditions to the training data (cf. Table 1).

Independent test data should focus upon aspects of generalizability that can be tested. For example, if a model is intended to predict future distributions in a different region from the training data set, test data collected from that region would allow geographical generalizability to be tested, even though combined spatiotemporal effects could not be tested. Alternatively, an analogue of the application could be tested, such as using historical data to demonstrate temporal generalizability, albeit not to the required time period, on a model required to predict future distributions. Cheddadi, Guiot & Jolly (2001) illustrate this principle, using data from the pollen record to test a model relating Mediterranean vegetation to climate, prior to making predictions of future distributions.

Size of the test set

The size of a test set is critical in ensuring that reliable estimates of model accuracy are obtained and differences in performance between models can be identified. Large sample sizes are often required to obtain precise estimates of the accuracy statistics used with species distribution models. For example, differences of 0·05 in the area under the receiver operating characteristic (ROC) curve (AUC) statistic may reflect major differences in the predictive ability of two models, yet several hundred sites may be

required to identify this statistically (Cumming 2000; Steyerberg *et al*. 2003). Similar sample sizes may be required to separate κ statistics (Donner 1998). For both, the precision of estimates increases with the overall sample size, as species prevalence approaches 50% and as the accuracy of the model increases (Hanley & McNeil 1982; Donner 1998; McPherson, Jetz & Rogers 2004). These relationships make sample size guidelines difficult to formulate, but suggest that 200–300 sites or more are desirable for a test set. Harrell (2001) suggests that a test set should contain at least 100 sites of the less common event (present/absent). Test set size also relates to the generalizability desired from the model. Greater numbers of test sites are likely to be necessary where wider generalizability is needed from a model. This makes daunting demands on field data collection, yet may be essential to provide robust assessments of predictive performance.

DIAGNOSING CAUSES OF LIMITED GENERALIZABILITY

Where limits to generalizability are identified during model testing, it is important to analyse these further, both to diagnose the causes and facilitate improvements to the model, and to guide users about when and where it is safe to apply the model. Two basic factors may be implicated in poor generalizability: overfitting and a failure to transport to the differing conditions experienced in new data (Justice, Covinsky & Berlin 1999).

Overfitting

Overfitting is a model development issue, occurring when idiosyncrasies in the training set are modelled in addition to the underlying species–environment relationships (Harrell, Lee & Mark 1996). This results in a misleadingly good fit to the data. In statistical terms, the modelled relationships will not accurately represent those in the population from which the training data were sampled. The potential for overfitting increases when more flexible modelling methods are used (e.g. generalized additive models compared with their generalized linear equivalents), when a greater degree of variable selection is employed (e.g. all subsets regression) or when fewer training data are available (Harrell, Lee & Mark 1996).

Independent test data, where conditions differ from training (Table 1), measure overall generalizability rather than addressing overfitting specifically. Resampling methods, in contrast, derive estimates of model performance from the training data, measuring performance under the training conditions. As a consequence, resampling methods solely address overfitting. Frequently, overfitting is a major cause of limited generalizability, enabling resampling to identify problems prior to the expense of collecting specific test data (Charlson *et al*. 1987; Harrell, Lee & Mark 1996). Equally, good resampling performance may provide the necessary justification to pursue model development further, including the expense of test data collection (Verbyla & Litvaitis 1989).

A review of different resampling methods is provided by Verbyla & Litvaitis (1989). In general, jack-knifing and bootstrapping are the most useful methods, especially when the training set is relatively small (Efron & Gong 1983; Efron 1983). Bootstrapping has the advantage over jack-knifing of producing large 'training' and 'test' data sets at every iteration, making it easier to calculate many model accuracy statistics. Several bootstrapping methods are available, ranging in complexity. Work in medicine using models analogous to those in ecology suggests that one of the simplest approaches (Table 2) performs as well as more complicated ones (Steyerberg *et al*. 2001).

Transportability of the model

Transportability relates to the consistency of the underlying species–environment relationships from the conditions under which a model was trained to those under which the model will be applied. It therefore considers whether a model can maintain its accuracy when the prevailing conditions change, compared with overfitting, which only considers whether the model provides an unbiased fit under the training conditions (Justice, Covinsky & Berlin 1999). In terms of quantifying species–environment relationships, transportability describes

Table 2. A simple bootstrapping method, compatible with a wide range of accuracy measures, for estimating the predictive performance of distribution models. In contrast to other bootstrap methods, which make direct estimates of model accuracy, overfitting in the model development process is estimated and then used to correct the biased accuracy estimate made with the training data. After Efron (1983); Harrell, Lee & Mark (1996); Steyerberg *et al*. (2001)

1.	Estimate accuracy statistic in the training data
2.	Generate a bootstrap of equal size to the training set by sampling training data with replacement
3.	Fit the model in the bootstrap using the same methods as employed to fit it in the original training data; this includes the same variable selection strategy, where applicable
4.	Estimate the accuracy statistic within the bootstrap resample. This simulates an accuracy estimate made with the training data
5.	Using the same model as in step 4, predict the species distribution in the original training set and estimate the accuracy statistic. This simulates the use of independent test data
6.	Overfitting = (training data estimate in step 4) – (test data estimate in step 5)
7.	Repeat steps 2–6 for 100–200 bootstraps. Average the values calculated in step 6 to provide the overall estimate of overfitting
8.	Subtract overfitting estimate from the training data estimate in step 1 to provide an optimism-corrected value

the range of conditions over which conclusions are valid. Failure to transport, and consequently to generalize, to a new application could, for example, result from the presence of novel factors influencing distribution (e.g. different predators or competitors) or ecotypic variation between regions of interest (Oostermeijer & van Swaay 1998).

Unlike overfitting, problems with transportability cannot be diagnosed directly. Instead they are derived by comparing the accuracy of a model when applied to both its training and test data, and when tested by resampling (Justice, Covinsky & Berlin 1999). If the performance of a model decreases from its training data to independent test data by a greater amount than can be accounted for by overfitting, transportability problems are implied.

As a further stage in analysing generalizability, test data should be used to produce a detailed summary regarding the conditions under which it is 'safe' to apply a model and what degree of accuracy might be expected. These aims can be furthered by the methods used to assess model fit (Miller, Hui & Tierney 1991). Nicholls (1989) used regression residuals in the training data to identify environmental conditions under which a predictive model performed poorly. For probabilistic models, such methods may be generalized to a test set by fitting a logistic regression model to relate predictions to the observed presence–absence (see measuring calibration below) and then calculating regression diagnostics (Miller, Hui & Tierney 1991). Alternatively, the accuracy of individual predictions may be quantified by the squared difference between each prediction and its observation (Brier 1950). Groups of sites for which the accuracy is low, or specific conditions under which accuracy is degraded, may help to diagnose transportability problems.

Selection of accuracy statistics: probabilistic predictions

Many accuracy statistics of varying utility have been applied to species distribution models and much has been published about their desirable properties (Forbes 1995; Fielding & Bell 1997; Fielding 2002). Two properties in particular are worth highlighting. The first is the ability to describe model accuracy in terms of the observed predictive performance. A statistic measuring the probability of making the correct prediction at a site, for example, is more useful than many goodness-of-fit measures (e.g. R^2). The latter statistics describe overall prediction–observation agreement yet are difficult to equate to the observed predictive performance. The second, and perhaps most important, property for accuracy measures is generality: the ability to compare accuracy meaningfully between the same model in different applications or between models developed for different species or with different training and test data. Such comparisons require accuracy statistics either largely independent of, or possibly corrected for, potentially confounding properties of the particular data used to test a model, such as species prevalence (Miller, Hui & Tierney 1991; Manel, Williams & Ormerod 2001; Fielding 2002). Minimizing the reliance upon test set properties represents a substantial challenge and most accuracy measures used to evaluate species distribution models are deficient in this respect. Flawed measures may result in attempts to compare models for different species, or from different data sets, being confounded by statistical artefacts (McPherson, Jetz & Rogers 2004).

Many distribution modelling methods produce probabilistic, or other types of quantitative, predictions, at least as their initial outputs prior to dichotomization either by software or the user (Fielding 2002). Such predictions are often considered as estimates of habitat suitability or quality (Buckland & Elston 1993). They convey more information of conservation value than nominal presence–absence predictions, indicating the suitability of individual sites, rather than dividing all sites crudely into two opposing categories that are likely to disguise a wide range of habitat variation. Critically for model testing, it is straightforward to describe the accuracy of quantitative predictions in a way that can be interpreted generally (cf. nominal presence–absence predictions): by distinguishing two components of accuracy, discriminatory ability and calibration, the role of a species' prevalence can be separated from the underlying predictive ability of a model. The measurement of these two components will be considered in turn.

DISCRIMINATORY ABILITY

Discriminatory ability is the capacity of a model to distinguish occupied from unoccupied sites (Harrell, Lee & Mark 1996), manifested ultimately as the ability to place sites in rank order of probability of occupancy or suitability for a species. Discrimination is therefore the fundamental component of prediction accuracy, equating to 'ecological skill' in detecting differences between sites that correlate with species' distributions. It is readily comparable between studies, being an intrinsic property of a model that is independent of species' prevalences and other characteristics of particular data sets.

Non-parametric correlation coefficients are effective measures of discriminatory ability for probabilistic models (Miller, Hui & Tierney 1991). By comparing the rank orders of predictions with observed presence–absence, they focus solely upon discrimination. Parametric correlation coefficients, in contrast, incorporate the size of the discrepancies between individual predictions and observed presence–absence (Miller, Hui & Tierney 1991). This makes them heavily reliant upon the particular data set used to test a model. For example, if the habitats in one data set are polarized between highly suitable and highly unsuitable, and in another are all of low to moderate suitability, a model that ranks sites equally well in both cases could have a very different average prediction–observation discrepancy. Consequently, whilst a rank correlation coefficient would

be the same in both cases, a parametric one could be very different, erroneously suggesting different discriminatory ability and confounding intermodel comparisons.

A range of rank correlation coefficients could be used to assess discriminatory ability, including Somers' D_{XY}, Goodman and Kruskal's γ and Kendall's τ (Harrell, Lee & Mark 1996; SAS Institute 1999). The concordance index (*c*-index; Harrell *et al*. 1982) is probably the most useful for distribution modelling because it is equivalent to the Wilcoxon statistic and the non-parametric AUC statistic used increasingly in ecology (Hanley & McNeil 1982). Models developed for their quantitative predictions can thus be readily compared with those whose predictions will be dichotomized, and so for which ROC methods may be useful in selecting a classification threshold. Concordance is an intuitive measure, indicating the probability that a model will place two sites (one occupied, the other unoccupied) selected at random in the correct rank order of likelihood of occupancy (Harrell *et al*. 1982). This means that chance/random performance has a clear definition, 0·5, and perfect discrimination is unity.

The major weakness of non-parametric correlation coefficients is the poor efficiency of relying upon rank orders, necessitating large test sample sizes. A particular problem arises when the prevalence is very low (« 50%), as with the rarer species often emphasized in conservation biology. Only paired occupied–unoccupied sites are used to calculate *c*, with the result that large increases in the size of a test set may be required to increase the effective sample size for calculating the statistic. The strength of this effect is demonstrated by the large impact that small changes in prevalence have near the extremes for a fixed sample size (McPherson, Jetz & Rogers 2004).

CALIBRATION

In contrast to discriminatory ability, calibration is concerned directly with species prevalence, both for a complete application (or test set) and for subsets of sites. It describes the numerical accuracy of the predictions (Harrell, Lee & Mark 1996): whether, for example, sites given predicted probabilities of 0·60 have a 60% chance of being occupied and whether this is twice as likely as for sites given labels of 0·30. If the predictions are not scaled between zero and one, some analogous property could be considered: whether the habitat suitability scores are proportional to prevalence, for example.

The importance of testing calibration varies according to a model's intended uses. Often, a model may only be required to rank sites according to their relative suitability/probability of being occupied: treating its predictions as being ordinal, rather than quantitative. A typical example is the selection of the 'best' *n*% of sites within an area for wildlife reserve location or targeting management activities (Pearce, Ferrier & Scotts 2001). In such instances, a test of discrimination may be deemed sufficient. If a model is to be used in such a way, however, the lack of calibration testing should be stated clearly, and we strongly recommend that predictions be displayed as ranked categories to reflect the ordinal nature of model testing (discrimination). If values between zero and one are given, even with a clear statement of their use solely for ranking sites, there is a risk that inexperienced users of the predictions could attempt to interpret them quantitatively.

In recent years, predictions have been used increasingly in a quantitative manner. The clearest example of this trend is the increasing use of spatial maps to show the probability of species occurrence/habitat suitability generated by the proposed model (e.g. Franco, Brito & Almeida 2000; Suárez-Seoane, Osborne & Alonso 2002; Johnson, Seip & Boyce 2004; Venier *et al*. 2004). Apart from their visual impact, such maps illustrate the information that quantitative predictions of occurrence can convey. Aside from using predictions as absolute estimates of probability of occurrence, sites can be compared quantitatively, suggesting how much more likely to be occupied, or suitable for a species, one site is than another. This should increase the value of models, allowing sites of similar suitability/probability of occupancy to be identified and management strategies to be formulated according to the suitability of different sites. Probabilities also provide estimates of confidence in predicted occurrences/absences. It is rare, however, to see evidence that the calibration of predictions has been tested (Carroll, Zielinski & Noss 1999). It seems remarkable that after the repeated emphasis upon careful model testing, predictions are often used quantitatively with apparently no attempt to check their numerical accuracy first. Part of the problem may stem from a limited recognition of the distinction between discrimination and calibration, as good discrimination (e.g. high AUC) need not imply good calibration. Unfortunately, without prior tests of calibration, model predictions could mislead conservation efforts. It is virtually impossible to estimate what damage may already have been done to conservation through the use of poorly calibrated predictions that were considered accurate following tests that examined only discrimination.

The calibration plot is the basic tool for assessing the calibration of probabilistic predictions and serves to illustrate the two major types of calibration error. At its simplest, a plot of average predicted probability (*x*-axis) vs. species prevalence (*y*-axis) can be drawn for discrete groups of sites across the probability scale (Fig. 1). Groups may be formed at regular probability intervals (e.g. deciles) or for fixed group sizes across the probability range. Problems with calibration are evident as deviations in the agreement between predictions and observations from the 45° diagonal (Fig. 1).

The first type of calibration error is a consistent under- or overestimation of a species' prevalence (Fig. 1). Such overall bias typically results when a species' prevalence differs from the training data (Pearce & Ferrier 2000). If predictions are to be subsequently dichotomized, such bias will inflate either the false positive or false negative rate. Fortunately, comparing the mean

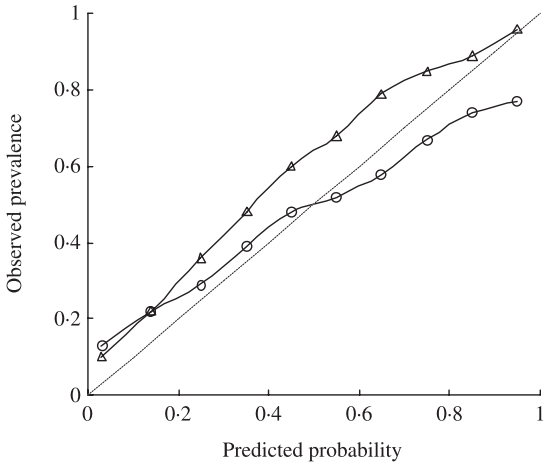

Fig. 1. Example of a calibration plot, comparing the predicted and observed species prevalence within decile probability intervals. The 45° diagonal (fine dashed line) indicates perfect calibration. The other two lines represent calibration problems: overfitting (circular symbols), characterized by a slope < 45°, and an overall underestimation of a species' prevalence across the habitats in the test set (triangular symbols).

predicted probability to the observed prevalence is an easy test of overall calibration. Similarly, overall calibration is easily adjusted to match the species' prevalence (Poses *et al*. 1986).

The second type of calibration problem is more complicated and may involve probabilities being over- or underestimated at some sites, well calibrated at others or a combination of the different types over the probability scale. These problems may still exist when the overall calibration is accurate. They often result from problems in the development of the model but fortunately the pattern of calibration problems may be characteristic. For example, at least in a regression context, overfitting is indicated by a slope shallower than 45°, with predictions at the lower end of the probability scale tending to underestimate prevalence and those at the upper end of the scale overestimating prevalence (Fig. 1). This pattern develops because the coefficients of overfitted models tend to be biased towards more extreme values (Miller, Hui & Tierney 1991). Calibration testing can therefore perform an important diagnostic role.

Harrell, Lee & Mark (1996) suggested the use of scatterplot smoothers, such as LOWESS, on the raw predictions–observations as an improved approach to producing calibration plots. Such plots are very versatile, because by altering the degree of smoothing used they can reveal either the overall calibration pattern or localized problems within the probability range. Overall, though, calibration plots are largely subjective, both in requiring an arbitrary choice of group size or smoothing neighbourhood, and in judging what constitutes an important calibration problem.

Cox (1958) proposed an objective method for assessing calibration, which includes a basis for testing the significance of calibration problems. Pearce & Ferrier (2000) describe the method in detail. It relates the probabilistic predictions (logit-transformed) made by a model to the observed species presence–absence using logistic regression. The intercept in the resulting regression equation addresses the overall calibration, whilst the slope coefficient considers the pattern of calibration across the probability scale. A perfectly calibrated model has an intercept of zero and a slope of one, and deviations can be tested using likelihood ratio tests (Miller, Hui & Tierney 1991; Pearce & Ferrier 2000). The slope is the more useful coefficient, as values between zero and one are indicative of overfitting (Steyerberg *et al*. 2000). Minor overfitting in regression models may be corrected by multiplying their coefficients by the calibration slope coefficient to 'shrink' them back towards their true values (Steyerberg *et al*. 2000).

The use of logistic regression to assess calibration has some weaknesses. With relatively small sets of test data, the estimation of regression coefficients may be unreliable, affecting both the identification and correction of calibration problems. This again argues for a large test set. Estimation of the slope and intercept coefficients is interrelated, making their interpretation difficult (Seillier-Moiseiwitsch 1996). For example, if the slope is different from one, the intercept describes calibration at a probability of 0·50 rather than overall (Miller, Hui & Tierney 1991). Finally, the Cox approach cannot describe non-linear calibration problems (Harrell 2001).

For most applications, calibration plots are more useful, combining simplicity, ease of interpretation, flexibility and diagnostic power, such as the ability to characterize non-linear calibration problems and identify characteristic patterns (e.g. overfitting). The exception is where the focus is on assessing the impact of overfitting upon calibration and potentially attempting simple corrections (shrinkage), for which the Cox approach is more useful.

Selection of accuracy statistics: nominal presence–absence predictions

Nominal presence–absence predictions are made in a range of ecological situations: where a dedicated classifier is used to predict presence–absence (e.g. a classification tree), if a particular application demands presence–absence predictions (e.g. predicted species lists for sites) or where a simple approach to the integration of the costs associated with different prediction errors is required (Guisan & Zimmermann 2000). Attempts to test such predictions are intrinsically problematic especially if, as in most instances, the predictions result from dichotomizing a quantitative output. Once dichotomized, it is impossible to separate discrimination from calibration, making model evaluation dependent upon the particular test data and confounding inter-model comparisons. The selection of a particular threshold means that accuracy is tested under only one scenario out of the full range between zero and one (Pearce & Ferrier 2000) and the less informative nature

of threshold-dependent testing limits the capacity to describe generalizability and identify the causes of prediction errors. We therefore recommend that probabilistic predictions be used whenever possible and, if nominal predictions need to be derived from quantitative ones, that both threshold independent and dependent measures for a model are given, to provide both a general assessment of performance and one specific to the particular application (and threshold).

The inability to separate discrimination from calibration means that accuracy measures for nominal presence–absence predictions have to compensate for the influence of species' prevalences and the associated problem of 'chance' agreement (Manel, Williams & Ormerod 2001). Without such compensation, accuracy estimates can be severely misleading. High overall prediction success (correct classification rate), for example, can be achieved by a model with little or no predictive ability if the prevalence is extreme, e.g. always predicting that a rare species will be absent (Fielding & Bell 1997). This problem appeared to have been overcome with the adoption of Cohen's κ statistic, which measures the chance-corrected agreement between predictions and observations (Fielding & Bell 1997). In a recent paper, however, McPherson, Jetz & Rogers (2004) questioned the value of κ for distribution modelling. Two concerns with κ are, first, the interpretation of chance-corrected agreement, and secondly the way in which 'chance' is defined.

Agreement beyond chance is a combined property of a model's accuracy and a species' prevalence in the test data, making it difficult to compare between studies where prevalence differs. Such agreement does not describe how often a model's predictions are correct and, indeed, may mislead potential users. At the extremes of prevalence the potential for chance agreement is large, making it virtually impossible even for an accurate model to demonstrate much agreement *beyond* chance. This is evident in the unimodal relationship observed between prevalence and Cohen's κ (e.g. McPherson, Jetz & Rogers 2004). One consequence is that widely used standardized scales for the interpretation of κ (e.g. Landis & Koch 1977) have little meaning. A second consequence is that accurate and valuable models for rare species, recorded at low prevalence, may achieve only moderate κ values and be disregarded. The only way to make meaningful intermodel comparisons based upon κ would be to standardize species' prevalences across studies (McPherson, Jetz & Rogers 2004). Where κ is used, a second statistic, such as overall prediction success, is required to indicate the overall level of accuracy.

The definition of 'chance' also causes problems. The model of chance selected can strongly influence the value of κ and therefore the assessment made of a model (Brennan & Prediger 1981). Different models of chance define different formulations of κ, and there has been debate in the social and medical sciences concerning which is most useful (Brennan & Prediger 1981; Feinstein & Cicchetti 1990). This debate has surfaced in ecology with a consideration of a form of κ sometimes called the τ coefficient, in place of Cohen's κ (Fielding & Bell 1997; Couto 2003). An alternative, and perhaps better, approach is to simulate 'chance' agreement directly using randomization testing (Olden, Jackson & Peres-Neto 2002).

Given these potentially serious limitations, measures other than κ may be preferable. Paired sensitivity and specificity measures are often used to summarize medical test performance, are independent of prevalence and can be calculated with and without correction for chance agreement (Brenner & Gefeller 1994). The use of two measures to describe performance may further complicate intermodel comparisons, however. If both sensitivity and specificity vary between models it can be difficult to say which models are better (Glas *et al.* 2003), especially as the relative importance of sensitivity and specificity may vary between applications. Another measure widely used in medicine is the odds ratio. It has seen little application in ecology, largely because of the problems associated with its calculation when the confusion matrix contains zero values (Manel, Williams & Ormerod 2001). This can be overcome with a simple continuity correction, adding 0·5 to each of the cells in the matrix (Forbes 1995). The odds ratio provides a definition of chance performance (unity), rather than attempting to correct for it. Unfortunately, at extremes of prevalence, commonly encountered with rare species, small changes in accuracy can have large effects upon the odds ratio, exaggerating agreement and making it difficult to interpret (Kraemer 2004).

A more productive line of research may be to investigate information-theoretic methods which, by providing an alternative paradigm for model evaluation, may allow more general model comparisons. Rather than attempting to describe classification accuracy directly, information-theoretic measures aim to quantify the amount of information that a set of predictions provides about their matched observations (Finn 1993; Forbes 1995). This information can also be considered as the information that the observations provide about the predictions, and so is denoted mutual information (Forbes 1995). If a model's predictions are totally unrelated to the observed presence–absence, mutual information is zero. Mutual information becomes maximal when knowledge of predictions allows perfect classification (Forbes 1995). This need not be perfect classification of predictions, however, because a model that always predicts the reverse of what is observed can be used to infer perfect classification. This is the non-monotonic behaviour of mutual information measures upon which information-theoretic criteria have been criticized (Fielding & Bell 1997). It is analogous to an AUC < 0·5, indicating that a model has predictive ability but that it is in the opposite direction from that expected, i.e. predicting absence in place of presence. A cursory examination of the confusion matrix immediately reveals whether this is happening, allowing the mutual information statistic to be qualified (Forbes 1995).

The normalized mutual information (NMI) has seen some application in ecology (Manel, Williams & Ormerod 2001; Wright & Fielding 2002). It is the difference between the overall information contained in the confusion matrix and that in the predictions, divided by the information contained in the observed presence–absence, all taken from one (Forbes 1995). The formula based on the confusion matrix is given in table 4 of Manel, Williams & Ormerod (2001), except that the quantity defined should be subtracted from one. The NMI is scaled so that it ranges from zero (no predictive ability) to one (complete information). Its derivation means that it does not correct for chance agreement or provide a clear definition of it (Forbes 1995). However, data sets with extreme prevalence contain little information and, critically, models that purely predict the most common class provide virtually no information about the classification problem; low NMI results even with a high level of prediction–observation agreement. Empirical demonstrations of the NMI indicate that it does not vary systematically with prevalence (Manel, Williams & Ormerod 2001).

Finn (1993) and Couto (2003) describe a measure closely related to the NMI: the average mutual information (AMI). Mutual information may prove to be very useful in ecology, summarizing the most fundamental concern about predictions – how much they reveal about the actual distribution – whilst circumventing the problems associated with chance correction. More work to assess its potential would be valuable.

Conclusions

Testing is a vital stage in developing predictive distribution models (Rushton, Ormerod & Kerby 2004). Yet, because adequate testing is still scarce and errors seldom diagnosed, the true value of species distribution modelling in ecology cannot yet be appraised. Many published examples could have serious limitations. Ideally, testing should have three aims, to: (i) provide an overall assessment of a model's predictive performance, including its generalizability, that can be used to assess its overall potential and allow comparison with other models; (ii) provide clear guidance for the use of a model and/or its predictions to the planners and practitioners dependent on its outputs; (iii) perform a diagnostic function, identifying weaknesses in predictive performance, possible causes and identifying priorities for future model development.

To optimize the testing of species distribution models, we make the following recommendations.

1. Wherever possible, use quantitative (e.g. probabilistic) predictions in preference to nominal presence–absence.

2. Perform the most rigorous possible tests of generalizability. The 'gold standard' is to test a model with samples of data from its actual applications, ideally including 200 or more sites to ensure precise accuracy estimates. Where this is not possible, aim to demonstrate the widest possible generalizability, encompassing the closest possible analogues to anticipated applications.

3. Utilize resampling methods, such as bootstrapping, to diagnose the causes of limited generalizability. Bootstrapping can also provide a convenient basis for making unbiased estimates of model performance prior to the expense of collecting test data.

4. Separate the discrimination and calibration components of accuracy to maximize the generality of model testing. The *c*-index/non-parametric AUC is an excellent statistic for measuring discriminatory ability.

5. Calibration should be tested whenever quantitative predictions will be interpreted directly. Calibration plots are a simple and effective method. Cox's regression method can be useful if overfitting and shrinkage are particular concerns.

6. Cohen's κ for testing nominal presence–absence predictions has limitations. Information-theoretic measures such as the NMI may be more profitable.

Acknowledgements

This work on model evaluation was funded by the Environment Agency. We would like to thank Dr Rob Freckleton and three anonymous referees, all of whose comments allowed us to make useful improvements to the manuscript.

References

Altman, D.G. & Royston, P. (2000) What do we mean by validating a prognostic model? *Statistics in Medicine*, **19**, 453–473.

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

Boyce, M.S., Vernier, P.R., Nielsen, S.E. & Schmiegelow, F.K.A. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.

Brennan, R.L. & Prediger, D.J. (1981) Coefficient kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement*, **41**, 687–699.

Brenner, H. & Gefeller, O. (1994) Chance-corrected measures of the validity of a binary diagnostic test. *Journal of Clinical Epidemiology*, **47**, 627–633.

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Buckland, S.T. & Elston, D.A. (1993) Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology*, **30**, 478–495.

Cabeza, M., Araújo, M.B., Wilson, R.J., Thomas, C.D., Cowley, M.J.R. & Moilanen, A. (2004) Combining probabilities of occurrence with spatial reserve design. *Journal of Applied Ecology*, **41**, 252–262.

Carroll, C., Zielinski, W.J. & Noss, R.F. (1999) Using presence–absence data to build and test spatial habitat models for the fisher in the Klamath region, USA. *Conservation Biology*, **13**, 1344–1359.

Charlson, M.E., Ales, K.L., Simon, R. & MacKenzie, C.R. (1987) Why predictive indexes perform less well in validation studies: is it magic or methods? *Archives of Internal Medicine*, **147**, 2155–2161.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A*, **158**, 419–466.

Cheddadi, R., Guiot, J. & Jolly, D. (2001) The Mediterranean vegetation: what if the atmospheric CO_2 increased? *Landscape Ecology*, **16**, 667–675.

Couto, P. (2003) Assessing the accuracy of spatial simulation models. *Ecological Modelling*, **167**, 181–198.

Cox, D.R. (1958) Two further applications of a model for binary regression. *Biometrika*, **45**, 562–565.

Cumming, G.S. (2000) Using between-model comparisons to fine-tune linear model of species ranges. *Journal of Biogeography*, **27**, 441–455.

D'Angelo, D.J., Howard, L.M., Meyer, J.L., Gregory, S.V. & Ashkenas, L.R. (1995) Ecological uses for genetic algorithms: predicting fish distribution in complex physical habitats. *Canadian Journal of Fisheries and Aquatic Science*, **52**, 1893–1908.

Donner, A. (1998) Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine*, **17**, 1157–1168.

Efron, B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, **78**, 316–331.

Efron, B. & Gong, G. (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, **37**, 36–48.

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. *Journal of Applied Ecology*, **41**, 263–274.

Feinstein, A.R. & Cicchetti, D.V. (1990) High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, **43**, 543–549.

Fielding, A.H. (2002) What are the appropriate characteristics of an accuracy measure? *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall & F.B. Samson), pp. 271–280. Island Press, Washington, DC.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Fielding, A.H. & Haworth, P.F. (1995) Testing the generality of bird–habitat models. *Conservation Biology*, **9**, 1466–1481.

Finn, J.T. (1993) Use of the average mutual information index in evaluating classification error and consistency. *International Journal of Geographical Information Systems*, **7**, 349–366.

Forbes, A.D. (1995) Classification-algorithm evaluation: five performance measures based on confusion matrices. *Journal of Clinical Monitoring*, **11**, 189–206.

Franco, A.M.A., Brito, J.C. & Almeida, J. (2000) Modelling habitat selection of common cranes *Grus grus* wintering in Portugal using multiple logistic regression. *Ibis*, **142**, 351–358.

Frair, J.L., Nielsen, S.E., Merrill, E.H., Lele, S.R., Boyce, M.S., Munro, R.H.M., Stenhouse, G.B. & Beyer, H.L. (2004) Removing GPS collar bias in habitat selection studies. *Journal of Applied Ecology*, **41**, 201–212.

Gates, S. & Donald, P.F. (2000) Local extinction of British farmland birds and the prediction of further loss. *Journal of Applied Ecology*, **37**, 806–820.

Gibson, L.A., Wilson, B.A., Cahill, D.M. & Hill, J. (2004) Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. *Journal of Applied Ecology*, **41**, 213–223.

Glas, A.S., Lijmer, J.G., Prins, M.H., Bonsel, G.J. & Bossuyt, P.M.M. (2003) The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology*, **56**, 1129–1135.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

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

Harrell, F.E. (2001) *Regression Modeling Strategies*. Springer-Verlag, New York, NY.

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L. & Rosati, R.A. (1982) Evaluating the yield of medical tests. *Journal of the American Medical Association*, **247**, 2543–2546.

Harrell, F.E., Lee, K.L. & Mark, D.B. (1996) Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning*. Springer-Verlag, New York, NY.

Jeganathan, P., Green, R.E., Norris, K., Vogiatzakis, I.N., Bartsch, A., Wotton, S.R., Bowden, C.G.R., Griffiths, G.H., Pain, D. & Rahmani, A.R. (2004) Modelling habitat selection and distribution of the critically endangered Jerdon's courser *Rhinoptilus bitorquatus* in scrub jungle: an application of a new tracking method. *Journal of Applied Ecology*, **41**, 224–237.

Johnson, C.J., Seip, D.R. & Boyce, M.S. (2004) A quantitative approach to conservation planning: using resource selection functions to map the distribution of mountain caribou at multiple spatial scales. *Journal of Applied Ecology*, **41**, 238–251.

Justice, A.C., Covinsky, K.E. & Berlin, J.A. (1999) Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, **130**, 515–524.

Kraemer, H.C. (2004) Reconsidering the odds ratio as a measure of 2×2 association in a population. *Statistics in Medicine*, **23**, 257–270.

Landis, J.R. & Koch, G.C. (1977) The measurement of observer agreement of categorical data. *Biometrics*, **33**, 159–174.

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

Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921–931.

Marsden, S. & Fielding, A. (1999) Habitat associations of parrots on the Wallacean islands of Buru, Seram and Sumba. *Journal of Biogeography*, **26**, 439–446.

Miller, M.E., Hui, S.L. & Tierney, W.M. (1991) Validation techniques for logistic regression models. *Statistics in Medicine*, **10**, 1213–1226.

Nicholls, A.O. (1989) How to make biological surveys go further with generalised linear models. *Biological Conservation*, **50**, 51–75.

Olden, J.D., Jackson, D.A. & Peres-Neto, P.R. (2002) Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329–336.

Oostermeijer, J.G.B. & van Swaay, C.A.M. (1998) The relationship between butterflies and environmental indicator values: a tool for conservation in a changing landscape. *Biological Conservation*, **86**, 271–280.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.

Pearce, J., Ferrier, S. & Scotts, D. (2001) An evaluation of the predictive performance of distributional models for flora and fauna in north-east New South Wales. *Journal of Environmental Management*, **62**, 171–184.

Poses, R.M., Cebul, R.D., Collins, M. & Fager, S.S. (1986) The importance of disease prevalence in transporting clinical prediction rules. *Annals of Internal Medicine*, **105**, 586–591.

Rushton, S.P., Ormerod, S.J. & Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41**, 193–200.

SAS Institute Inc. (1999) *SAS Onlinedoc®, Version 8.* SAS Institute Inc., Cary, NC.

Seillier-Moiseiwitsch, F. (1996) Predictive diagnostics for logistic models. *Statistics in Medicine*, **15**, 2149–2160.

Steyerberg, E.W., Bleeker, S.E., Moll, H.A., Grobbee, D.E. & Moons, K.G.M. (2003) Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology*, **56**, 441–447.

Steyerberg, E.W., Eijkemans, M.J.C., Harrell, F.E. & Habbema, J.D.F. (2000) Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine*, **19**, 1059–1079.

Steyerberg, E.W., Harrell, F.E., Borsboom, G.J.J.M., Eijkemans, M.J.C., Vergouwe, Y. & Habbema, J.D.F. (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, **54**, 774–781.

Suárez-Seoane, S., Osborne, P.E. & Alonso, J.C. (2002) Large-scale habitat selection by agricultural steppe birds in Spain: identifying species–habitat responses using generalized additive models. *Journal of Applied Ecology*, **39**, 755–771.

Venier, L.A., Pearce, J., McKee, J.E., McKenney, D.W. & Niemi, G.J. (2004) Climate and satellite-derived land cover for predicting breeding bird distribution in the Great Lakes Region. *Journal of Biogeography*, **31**, 315–331.

Verbyla, D.L. & Litvaitis, J.A. (1989) Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management*, **13**, 783–787.

Wright, J.F. (1995) Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. *Australian Journal of Ecology*, **20**, 181–197.

Wright, A. & Fielding, A.H. (2002) Modeling wildlife distribution within urbanized environments: an example with the Eurasian badger *Meles meles* L. in Britain. *Predicting Species Occurrences: issues of Accuracy and Scale* (eds J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall & F.B. Samson), pp. 255–262. Island Press, Washington, DC.

Received 17 September 2004; final copy received 28 February 2005
Editor: Rob Freckleton