Global Change Biology (2005) **11,** 1504–1513, doi: 10.1111/j.1365-2486.2005.001000.x

Validation of species–climate impact models under climate change

MIGUEL B. ARAÚJO*†, RICHARD G. PEARSON*‡, WILFRIED THUILLER§ and MARKUS ERHARD¶

*Biodiversity Research Group, School of Geography and Environment, University of Oxford, Mansfield Road, Oxford OX1 3TD, UK, †Biogeography and Conservation Laboratory, Natural History Museum, Cromwell Road, London SW7 5BD, UK, ‡Macroecology and Conservation Unit, University of Évora, Estrada dos Leões, 7000-730 Évora, Portugal, §Climate Change Research Group, Kirstenbosch Research Centre, South African National Biodiversity Institute, Private Bag x7, Claremont 7735, Cape Town, South Africa, ¶Institute for Meteorology and Climate Research, Forschungszentrum Karlsruhe, Postfach 3640, 76021 Karlsruhe, Germany

Abstract

Increasing concern over the implications of climate change for biodiversity has led to the use of species–climate envelope models to project species extinction risk under climate-change scenarios. However, recent studies have demonstrated significant variability in model predictions and there remains a pressing need to validate models and to reduce uncertainties. Model validation is problematic as predictions are made for events that have not yet occurred. Resubstituition and data partitioning of present-day data sets are, therefore, commonly used to test the predictive performance of models. However, these approaches suffer from the problems of spatial and temporal autocorrelation in the calibration and validation sets. Using observed distribution shifts among 116 British breeding-bird species over the past $\sim$20 years, we are able to provide a first independent validation of four envelope modelling techniques under climate change. Results showed good to fair predictive performance on independent validation, although rules used to assess model performance are difficult to interpret in a decision-planning context. We also showed that measures of performance on nonindependent data provided optimistic estimates of models' predictive ability on independent data. Artificial neural networks and generalized additive models provided generally more accurate predictions of species range shifts than generalized linear models or classification tree analysis. Data for independent model validation and replication of this study are rare and we argue that perfect validation may not in fact be conceptually possible. We also note that usefulness of models is contingent on both the questions being asked and the techniques used. Implementations of species–climate envelope models for testing hypotheses and predicting future events may prove wrong, while being potentially useful if put into appropriate context.

Keywords: bioclimatic-envelope models, breeding birds, Britain, climate change, model accuracy, uncertainty, validation

Received 3 November 2004; revised version received 24 January 2005; accepted 8 March 2005

Correspondence: Miguel B. Araújo, Departamento de Biodiversidad y Biologia Evolutiva, Museo Nacional de Ciencias Naturales, CSIC, C/Jose Gutierrez Abascal, 2, 28006 Madrid, Spain, tel. + 34 91411328, fax + 34 915645078, e-mail: maraujo@mncn.csic.es

Introduction

Attempts to predict climate-change impacts on biodiversity have often relied on the species–climate 'envelope' modelling approach (also known as ecological niche models), whereby present day distributions of species are combined with environmental variables to project distributions of species under future climates (for review, see Pearson & Dawson, 2003). In spite of the

inherent limitations of correlative models (for review, see Guisan & Zimmermann, 2000), projections arising from species–climate envelope models have been used to support estimates of species' extinction risk under climate change for a variety of taxa and parts of the world (e.g. Bakkenes *et al*., 2002; Erasmus *et al*., 2002; Midgley *et al*., 2002; Peterson *et al*., 2002; Thomas *et al*., 2004a). The impact of these estimates within political and public debate is potentially high, yet there is great deal of scope for misrepresenting the science behind such studies (Ladle *et al*., 2004). Recent studies have reported that projections arising from species–climate models may be highly sensitive to the assumptions, algorithms and parameterizations of different methods (e.g. Thuiller, 2004; Thuiller *et al*.; 2004a, Pearson *et al*., 2005). These studies have raised a number of methodological issues that lead to a degree of uncertainty which has been underestimated, or simply overlooked, in previous assessments of climate impacts on biodiversity. We argue that when results of a particular analysis contribute to the discussion of the weight of evidence required to support important societal decisions, the demand that models' predictive accuracy be assessed is eminently reasonable.

Nevertheless, validation (also referred to as evaluation) of species–climate envelope models under climate changes remains poorly explored. The reason is that events being predicted have either been poorly documented or have yet not occurred. Consequently, assessments of accuracy are usually limited to a process of 'resubstituition', in which the data used to calibrate (or train) models are also used to validate (test) them (Fig. 1a; for review, see Table 1). A problem with the resubstituition approach is that models may overfit to the calibration data, leaving users unable to judge whether high accuracy on nonindependent data reflect good predictive accuracy on independent data sets. Some authors also caution against possible bias in estimates of model-prediction errors as the models are optimized to deal with the 'noise' in the data and might consequently lose generality outside the original data (for discussion, see Olden & Jackson, 2000; Olden *et al*., 2002). To address these problems, a growing number of studies have used data partitioning methods for the allocation of cases to calibration and validation data sets. The most familiar technique is one-time data-splitting, whereby data are split into calibration and validation samples by random process (Fig. 1b, Table 1). There are alternative techniques including grouped cross-validation (also known as *k* fold partitioning, hold out, or external method), bootstrapping, and jackknifing (also known as leave-one-out) (for discussion, see Harrell, 2001), but they all share the assumption that randomly selected samples from original data

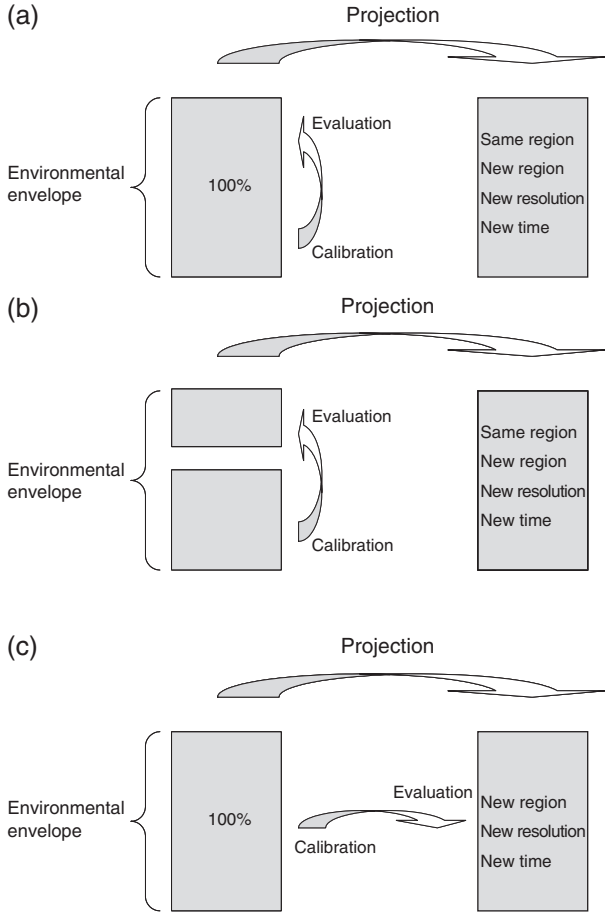

Fig. 1 Species-climate envelope modelling framework under three calibration and validation strategies: (a) resubstituition; (b) data splitting; and (c) independent validation.

constitute independent observations, hence suitable for model validation. Although these validation strategies have generally been accepted to provide more robust measures of predictive success than resubstituition (e.g. Fielding & Bell, 1997), they may not avoid two of the most important pitfalls of correlative models. The first is that of spatial autocorrelation in the distribution of species and environmental variables (e.g. Hampe, 2004). This is a problem because modelling techniques assume that modelled events are independent, which is not true in the case of spatially autocorrelated data. This problem is not overridden by resampling the original data randomly, nor is it by carrying additional field sampling for testing models within the modelled region, because any of these validation strategies would use test data that is spatially autocorrelated with data to calibrate models. The second is that of temporal correlation in biological and environmental phenomena. This is another form of autocorrelation in the data, and implies that observations in time series are

Table 1 Four approaches used to validate species–climate envelope models under climate change

Reference	Resubstituition	Bootstrap	Data-splitting	Independent validation
Araújo *et al*. (2004)			1	
Bakkenes *et al*. (2002)	1			
Beaumont & Hughes (2002)*				
Berry *et al*. (2002)			1	
Burns *et al*. (2003)	1			
Erasmus *et al*. (2002)	1			
Guisan & Theurillat (2000)			1	
Huntley (1995)	1			
Huntley *et al*. (1995)	1			
Huntley *et al*. (2004)	1			
Iverson & Prasad (1998)			1	
Iverson *et al*. (1999)	1			
Martinez-Meyer *et al*. (2004)				1
Midgley *et al*. (2002)	1			
Midgley *et al*. (2003)	1			
Miles *et al*. (2004)	1			
Pearson *et al*. (2002)			1	
Pearson *et al*. (2005)			1	
Peterson (2003b)		1		
Peterson *et al*. (2002)		1		
Peterson *et al*. (2001)	1			
Saetersdal *et al*. (1998)*				
Skov & Svenning (2004)	1			
Sykes *et al*. (1996)	1			
Teixeira & Arntzen (2002)	1			
Thuiller (2003)			1	
Thuiller (2004)			1	
Thuiller *et al*. (2004a)			1	
Thuiller *et al*. (2004b)			1	

Few studies (*) have not attempted to validate the predictive accuracy of their models.

nonrandom because of lack of independence between data points that are adjacent in time. Consequently, projections of observed current distributions closer in time are likely to be more similar than projections made further apart. The interplay of spatial and temporal autocorrelation make it conceptually difficult to discard the possibility that models' goodness-of-fit to the data represent an over-optimistic estimate of their predictive ability outside the initial spatial and temporal conditions defining the training set (e.g. Beutel *et al*., 1999). Thus, the number of degrees of freedom is overestimated, causing unrealistically small estimates of the standard errors of the model outputs. In addition, as temporal autocorrelation can introduce slow changes (i.e. low-frequency variability) in the time series, it can affect the estimate of the degree of estimated changes.

It may be argued that the predictive accuracy of species–climate envelope models can only be fully tested by means of validation studies using direct comparison of model predictions with independent empirical observations (Fig. 1c). Attempts to perform such tests are relatively rare. A limited number of studies have attempted independent validation using known distributions in different regions (Beerling *et al*., 1995; Fielding & Haworth, 1995; Peterson, 2003a), data at different resolutions (Pearson *et al*., 2004; Araújo *et al*., 2005a), field observations in previously unsampled regions where species' occurrences are predicted (Raxworthy *et al*., 2003), fossil records of mammal distributions under Pleistocene climates (Martinez-Meyer *et al*., 2004), and visual comparison between simulated and observed range changes for butterflies in the UK over the 20th century (Hill *et al*., 1999). However, statistical validation using independent data describing range shifts under recent climate change has not previously been undertaken.

As models projecting species' distributional shifts under future climate change are unlikely to be validated in most circumstances because of data limitations, it is important to improve understanding

of the underlying characteristics of data and methods that contribute uncertainty to predictions. Because most model evaluations assess accuracy to the calibration, or nonindependent validation data (also referred to as verification), it is important to investigate the degree to which these measures correlate with proper validations on independent data sets. These questions can be addressed only when independent data adequate for model validation are available and this is a rare circumstance for climate-change impact assessments. We make a first attempt to address these problems using British-breeding bird distributional records in two periods between the 1960s and the 1990s. We assume these are independent events, although we acknowledge that some degree of nonindependence may arise given that data were recorded in the same region and in two periods of time only 20-years apart. However, they do constitute a rare record of observed range shifts, and one of the few examples of species range-shift data that allows direct comparison between observations in each recording period, without the need to correct for sampling bias. Furthermore, they also have the advantage of including species reported to shift northward in apparent response to recent regional climate changes (Thomas & Lennon, 1999). The unprecedented quality of these data allows researchers to explore issues of bioclimate envelope model validation that have not yet been addressed in the literature. In particular, we ask: (1) how well do models perform on an independent validation dataset? (2) does validation using nonindependent distribution data provide a good surrogate for accuracy on independent data? (3) do particular modelling techniques perform consistently better than others?

Data and methods

Species data

We used distributional records in Britain for 116 native breeding-bird species recorded during the periods 1968–1972 (t_1) and 1988–1991 (t_2) (Sharrock, 1976, Gibbons *et al*., 1993). Volunteer recorders achieved 100% cover of the British 2831 10 km squares, with the total number of nonduplicate 10 km squares receiving records for the second period being within 1% of the 217 615 10 km squares records received for 1968–1972. This has allowed researchers to make comparisons between occupancy of squares in each recording period, without the need to correct for sampling bias (e.g. Thomas & Lennon, 1999; Thomas *et al*., 2004b). Our analyses of bird distributions did not include marine, waterfowl, and aquatic shorebirds. Species with less than 20 records in the first recording period were also excluded from analysis to avoid problems related to modelling data with excessively small sample sizes (e.g. Stockwell & Peterson, 2002). The minimum number of records for a species in this period was 25, the median number was 1560, and the maximum was 2405.

Climate data

A set of aggregated climate parameters were derived from an updated version of the CRU (Climate Research Unit at the University of East Anglia, UK) monthly climate data (New *et al*., 2000). The updated data set provides monthly values for the years 1901–2000 at $10' \times 10'$ spatial resolution (Mitchell *et al*., 2004). Average monthly temperature, precipitation and cloud cover of 1416 grid cells covering the area of the UK (7°30′ E–1°40′ W and 50°N–61°N) were used to calculate mean values of six different climate parameters in two different time slices (1967–1972, 1987–1991). Variables include mean annual temperature within time slices (°C), mean temperature of the coldest month (°C), mean temperature of the warmest month (°C), mean annual summed precipitation (mm), and mean sum of precipitation between July–September (mm), and growing season, defined as the temperature sum of all consecutive days with mean temperature greater than 5 °C. The six variables were selected on the basis that they are known to impose constraints upon species distributions as a result of widely shared physiological limitations (Crick, 2004).

Species–climate modelling

Breeding bird species distribution records in Britain were modelled using SPLUS-based BIOMOD (Thuiller, 2003). Modelling procedures included (1) generalized linear models (GLM) with linear, quadratic and polynomial terms (second and third order). A stepwise procedure using the AIC criterion was used to select the most significant variables (Akaike, 1974); (2) generalized additive models (GAM) with cubic-smooth splines. The degree of smoothness was bounded to four for each variable. As for GLM, a stepwise procedure was used to select the most parsimonious model; (3) classification tree analysis (CTA) using a 10-fold cross-validation to select the best trade-off between the number of leaves of the tree and the explained deviance; and (4) feed-forward artificial neural networks (ANN) with seven hidden units in a single layer and with weight decay equal to 0.03. Because of the heuristic nature of ANN models were run 10 times and the mean prediction was used. This procedure of averaging predictions over the collection of networks

is often preferred to using the solution giving the lowest error (Ripley, 1996).

Two runs were made with each modelling technique. In the first run, models were calibrated on a 70% random sample of the original time t_1 data and predictive accuracy was evaluated on the remaining 30% of the data (Fig. 1b). The size of the calibration set was determined by application of a commonly used heuristic for identifying the ratio of training and cross-validation sets in presence and absence models: $[1 + (p-1)^{1/2}]^{-1}$, where p is the number of predictor (here climate) variables (Fielding & Bell, 1997). In the second run, models were calibrated using 100% of the original time t_1 data and evaluated on the original time t_2 data (Fig. 1c). In each run, we tested agreement between observed and projected distributions by calculating Cohen's κ statistic of similarity (κ) and the area under curve (AUC) of the receiver operating characteristic (ROC) approach (Fielding & Bell, 1997). We used the κ approach after maximising the statistic over a range of thresholds above which model outputs are considered to represent species' presence. We calculated AUC using the nonparametric method based on the derivation of the Wilcoxon statistic (Fielding & Bell, 1997). Values of AUC range from ≤ 0.5 for models with no predictive ability, to 1.0 for models giving perfect predictions. κ values range from 0.0 (no predictive ability) to 1.0 (perfect predictive ability). There are a number of rules-of-thumb available to help interpreting measures of agreement between observed and projected events. For example, when using the κ statistic approach, Landis & Koch (1977) suggest the following ranges of agreement: excellent $K > 0.75$; good $0.40 > K < 0.75$; and poor $K < 0.40$. When using the ROC procedure, Swets (1988) recommends interpreting range values as: excellent $AUC > 0.90$; good $0.80 > AUC < 0.90$; fair $0.70 > AUC < 0.80$; poor $0.60 > AUC < 0.70$; fail $0.50 > AUC < 0.60$.

Results

How well do models perform on an independent validation dataset?

Our results demonstrate that models' predictive accuracy on independent validation were good around median values with AUC assessment (i.e. $0.80 < AUC < 0.90$ except for CTA and GLM), but only fair near the lower quartile distribution of accuracy values (i.e. $0.70 < AUC < 0.80$, Table 2). With κ assessment, models also provided good agreement around median values (i.e. $0.40 < \kappa < 0.75$ except for GLM), while lower quartile distribution values of accuracy were classified as poor (i.e. $\kappa < 0.40$). In both cases, upper quartile accuracy values were below 'excellent' threshold values (i.e. $AUC < 0.90$ and $\kappa < 0.75$).

Does validation on nonindependent distribution data provide a good surrogate for accuracy on independent data?

As most assessments of model accuracy use nonindependent data, it is useful to estimate the degree to which predictive accuracy measured with nonindependent t_1 distribution data provides a good surrogate for accuracy on t_2 independent data. Our results show that model accuracy evaluated on nonindependent 30% subset of t_1 data was always higher than accuracy on

Table 2 Predictive accuracy of different modelling techniques (ANN, CTA, GAM and GLM), calibrated with 70% data from time t_1 and verified against remaining 30% data of time t_1 (Fig. 1b), or calibrated with 100% of time t_1 data and validated against 100% time t_2 (Fig. 1c)

	Calibration 70% t_1	Validation 30% t_1	Δ_b	Calibration 100% t_1	Validation 100% t_2	Δ_c
British breeding birds						
κ						
ANN	0.59 (0.48, 0.70)	0.59 (0.43, 0.69)	*0*	0.60 (0.47, 0.69)	0.46 (0.26, 0.56)	*−0.14*
CTA	0.57 (0.47, 0.67)	0.53 (0.38, 0.62)	*−0.04*	0.57 (0.45, 0.66)	0.40 (0.25, 0.53)	*−0.17*
GAM	0.53 (0.41, 0.66)	0.58 (0.40, 0.67)	*0.05*	0.53 (0.42, 0.66)	0.43 (0.29, 0.54)	*−0.10*
GLM	0.53 (0.42, 0.66)	0.57 (0.41, 0.67)	*0.04*	0.54 (0.42, 0.66)	0.37 (0.22, 0.50)	*−0.17*
AUC						
ANN	0.92 (0.87, 0.94)	0.90 (0.85, 0.93)	*−0.02*	0.92 (0.87, 0.94)	0.84 (0.78, 0.88)	*−0.08*
CTA	0.88 (0.82, 0.91)	0.86 (0.78, 0.89)	*−0.02*	0.87 (0.81, 0.91)	0.77 (0.70, 0.83)	*−0.10*
GAM	0.91 (0.85, 0.94)	0.90 (0.85, 0.93)	*−0.01*	0.91 (0.85, 0.94)	0.82 (0.75, 0.89)	*−0.09*
GLM	0.91 (0.85, 0.93)	0.90 (0.85, 0.93)	*−0.01*	0.91 (0.86, 0.93)	0.78 (0.68, 0.85)	*−0.13*

Values correspond to median (lower quartile, upper quartile) accuracy measures (κ and ROC) obtained for selected British breeding birds ($n = 116$); Δ values correspond to the difference between median accuracy measured on the 30% randomly chosen t_1 data or 100% time t_2 validation sets and median accuracy measured on calibration sets.

independent t_2 data (Table 2), supporting concerns that models' predictive accuracy measured on nonindependent data are likely to provide a generally overoptimistic assessment of model performance on independent data. Drops in accuracy from calibration to validation sets (i.e. accuracy in validation minus accuracy on calibration sets) were always lower for models validated with nonindependent t_1 validation data (Fig. 1b) than for models validated with independent t_2 validation data (Fig. 1c, Table 2). For example when models were validated on a 30% random sample of the original t_1 data, the maximum median drop of accuracy was −0.04 (CTA κ) with the best result being an increase in accuracy of 0.05 (GAM κ). When models were validated with an independent data set for t_2, then the maximum median drop of accuracy was −0.17 (CTA and GLM κ) whereas the best result was a drop of −0.08 (ANN AUC).

Generally, the two methods yielding the highest accuracies on independent time t_2 data (ANN and GAM with both κ and AUC) also had the lowest decreases in predictive accuracy from calibration in time t_1 to validation in time t_2 (Table 2). Because results in Table 2 are described by only three descriptors of the frequency distribution of accuracy values (median, lower quartile and upper quartile), we explored predictability of accuracies using the whole set of results for each individual species (Figs 2 and 3). As expected, greater correlations are observed between model accuracies on the calibration and nonindependent t_1 validation sets (Fig. 2) than for accuracies on calibration and independent t_2 validation (Fig. 3). GLM and GAM had the highest correlation between accuracies in calibration and nonindependent validation sets (Fig. 2), whereas model accuracies on calibration and independent validation were more stable with GAM and ANN (Fig. 3).

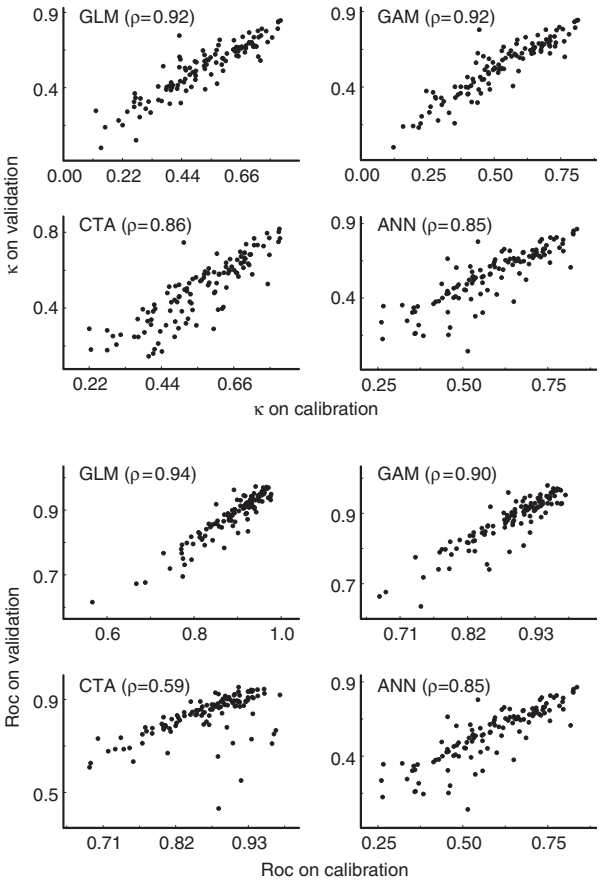

Fig. 2 Nonindependent validation. Relationship between accuracy measured on 70% subsample of original distribution data in time t_1, against accuracy as measured on the remaining 30%. Models were calibrated on a 70% random sample of the original distribution data in time t_1. Each dot represents the model accuracy obtained for individual British breeding bird species.

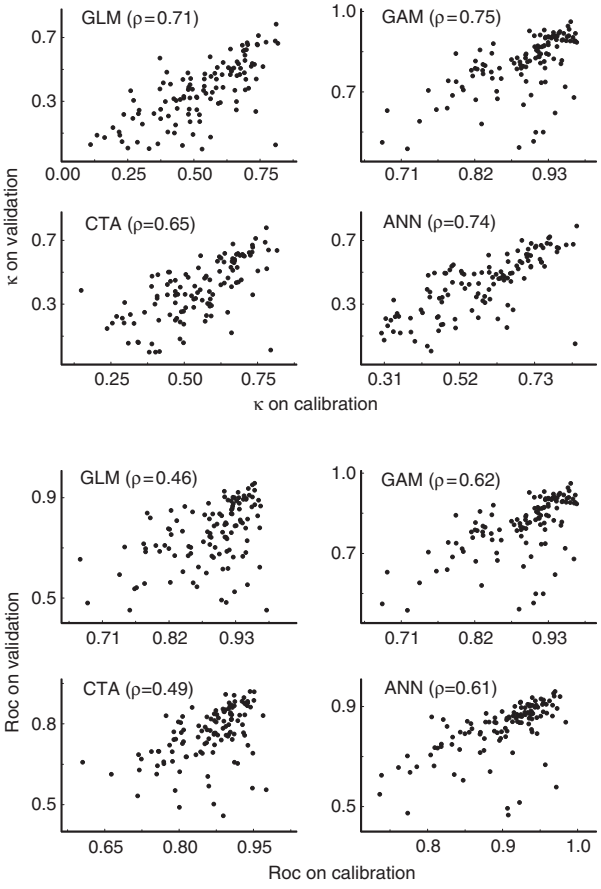

Fig. 3 Independent validation. Relationship between accuracy measured on 100% of original distribution data in time t_1, against accuracy as measured on the 100% of data in time t_2. Models were calibrated on 100% of the original distribution data in time t_1. Each dot represents the model accuracy obtained for individual British breeding bird species.

Do some modelling techniques perform consistently better than others?

ANN provided projections yielding generally higher accuracies on calibration and validation sets than any other methods (Table 2). GAM performed second best with good results in validation data sets, but not always as good in calibration sets. GLM achieved similar accuracy as GAM in calibration data but lost predictive accuracy when modelled distributions were tested against validation sets. CTA seem to overfit in some cases (especially with κ method), as models yielded large accuracies on calibration sets but showed some of the largest median drops in predictive accuracy in validation sets (Table 2).

Discussion

This study reports a first attempt to extensively investigate accuracy of species–climate envelope models, using observed range shifts of 116 British breeding-bird species under recent climate change. Accuracy was assessed using simple rules-of-thumb for interpreting measures of agreement between observed and projected events (Landis & Koch, 1977; Swets, 1988). Results showed good to fair predictive performance on an independent dataset, which is encouraging for applications of the models. However, the arbitrary rules used to assess model performance are difficult to interpret as they are difficult to translate into clear guidelines as to what are acceptable levels of model uncertainty from a user's perspective. Furthermore, they do not allow for a distinction between models' ability to predict absences and presences. It may be argued that the ability to predict absences may not be of exceptional significance, especially for species with a limited number of presences (J. Elith, personal communication). These rules do, however, provide a relative basis to make comparisons and assess changes in models performance. For example, it was shown that models validated with nonindependent t_1 distribution data provide overoptimistic assessments of predictive accuracy when compared with their ability to predict independent t_2 data. We anticipate that models' predictive ability might further decrease as the time-period considered for projections increases. This is because the effect of inflated performance arising from modelling spatially and temporally autocorrelated data should decrease as observed and modelled events become increasingly independent from each other. We also found that the models with highest accuracies on nonindependent data tended to have smaller reductions in accuracy when confronted with independent data. This result supports cautious use of measurements of accuracy on nonindependent data as a surrogate for accuracy on independent data (but see Elith & Burgman, 2002; Araújo *et al*., 2005a). Finally, we showed that, with our data, ANN and GAM provided generally more accurate projections of species range shifts under climate change than GLM and CTA. This pattern of performance across modelling techniques is consistent with previous assessments of performance of species–climate envelope models with nonindependent data (for reviews see Olden & Jackson, 2002; Segurado & Araújo, 2004), and suggests that modelling techniques capable of summarising complex nonlinear relationships are more likely to provide useful projections of species responses to climate change. CTA is one such technique, although a tendency for overfitting was recorded in our study. This is unsurprising as CTA is bound to overfit in three directions: searching for best predictors, for best splits, and searching multiple times (Harrell, 2001). The high performance of complex nonlinear techniques suggests that relatively unexplored methodologies such as multivariate adaptive regression splines , adaptive logistic regression (boosting) and generalized multiplicative models (for review see Hastie *et al*., 2001) might deserve future testing.

Many studies have used good model fits on nonindependent validation data to support results pertaining to the potential impacts of future climate change on biodiversity (see references in Table 1). However, we have demonstrated here that confidence in predictive ability on independent data for a different time-period is reduced, leading to less optimistic estimates of predictive ability. There are many reasons, additionally to the effects of autocorrelation in the data, why good model fits on present-day distribution data (i.e. nonindependent validation data) do not necessarily translate into good predictions of future ranges. Such factors may include the presence of spurious correlations between response (i.e. species) and predictor (i.e. climate) variables, which may translate into poor predictions on independent validation data (e.g. Guisan & Zimmermann, 2000). Problems may also arise when projections of future distributions extrapolate beyond fitted values among predictors (e.g. Thuiller *et al*., 2004b). More fundamentally, models assume immediate responses of species to climate change (e.g. Araújo & Pearson, 2005), when restricted dispersal ability, changes to existing networks of biotic interactions, and possible rapid evolutionary adaptations may prevent such responses from occurring (for reviews see Loehle & LeBlanc, 1996; Pearson & Dawson, 2003). Furthermore, there can be no assurance that models that show good predictive ability for past range shifts will give reliable predictions of future shifts, as climate change over the next century is projected to be

potentially more rapid and of greater magnitude than has been experienced during the last 1 000 years (Houghton *et al*., 2001).

There are clearly limits to the ability of any model to predict the future distribution of species under climate change, and model validation thus becomes a conceptually difficult problem. This is a familiar problem throughout science, as epitomized by Oreskes *et al*. (1994) in the following example: 'If it rains tomorrow, I will stay home and revise this paper. The next day it rains, but you find that I am not home. (...) You conclude that my original statement was false. But in fact it was my intention to stay home and work on my paper. The formulation was a true statement of my intent. Later, you find that I left the house because my mother died, and you realise that my original formulation was not false, but incomplete. It did not allow for the possibility of extenuating circumstances' (p. 641). Here the attempt to validate the preposition was unsuccessful because all influencing factors were not – or could not be – incorporated. Taking a more familiar example, consider predictions of the potential impacts of climate change on the distribution of Red-backed shrike (*Lanius colurio*) in Britain. Araújo *et al*. (2005b) used results from species–climate envelope models to predict that the Red-backed shrike should have expanded its range northwards over the second half of the 20th century. This general projection was coincident for many of the breeding birds studied, as expected (Thomas & Lennon, 1999). However, the Red-backed shrike has not undergone such an expansion, with its distribution in fact having contracted evenly across its range (Araújo *et al*., 2002). It is probable that nonclimatic factors, that are not incorporated within species–climate envelope models, such as habitat change, interactions with other species, or events occurring within the species' wintering grounds in the southernmost parts of Africa, may have been the major driver of distributional shift for this species. Validation of the model prediction for this species was, therefore, unsuccessful, although the chief aim of species–climate-envelope modelling – to characterize a species' suitable climate space (or *potential* range) – remains untested as we have only *realized*, rather than *potential*, ranges against which to validate. These examples illustrate a problem that is common to all models attempting to predict (independent) future events based on calibration of existing ones. As modelled systems are not closed (Oreskes *et al*., 1994), it becomes impossible to account for all potential factors driving changes in the state of modelled events. Errors are thus an inherent property of models.

We conclude by noting that successful validation does not necessarily imply that a model is valid for a particular application (e.g. Oreskes, 1998). Indeed, agreement between modelled and observed events may occasionally occur by chance as errors in one component of the model may be offset by errors in another component. Hence, even when projections from models are entirely consistent with observed independent data, they cannot be formally said to 'prove' the model but to fail to disprove it. Conversely, unsuccessful validation does not always mean that a model is wrong. A model may utilize parameters that are relevant to the underlying processes affecting species distributions but its predictive performance be overridden by governing processes operating at different spatial or temporal scales (Rastetter, 1996). However, as stressed by George Box, 'models are never true, but fortunately it is only necessary that they be useful. For this it is usually needful only that they are not grossly wrong' (Box, 1979, p. 2). The critical questions for species–climate envelope models are thus (1) how can the realism of model assumptions, algorithms and parameters be improved? and (2) which questions make particular model applications useful? While the first question has been the subject of several empirical studies (including this one) and reviews, the second remains largely unexplored (but see Hodges, 1991). Philosophers have argued that models provide useful tools for formulating hypotheses and exploring 'what if' questions, thereby illuminating which aspects of a problem are most in need of further investigation and where more empirical data are needed. However, they caution against the use of such models for hypotheses testing and prediction, arguing that the value of models for policy-making and planning is mainly heuristic (e.g. Shrader-Frechette & McCoy, 1993; Oreskes *et al*., 1994; Oreskes, 1998; Van Horne, 2002). This perspective contrasts with particular applications that used species–climate envelope models for testing hypotheses of species–climate relationships (e.g. Huntley *et al*., 2004), or for making predictions of extinction risk under future climate-change scenarios (e.g. Thomas *et al*., 2004a). Although our results seem to support suggestions that envelope models can be useful for providing a first approximation as to likely general impacts in climate-driven range changes, it may be conceptually inadequate to use these projections as face value for making predictions of future events. It is therefore vital that the models are applied critically and that validation against nonindependent data does not lead to unrealistically optimistic estimates of predictive ability.

Acknowledgements

We thank the many volunteer fieldworkers who contributed the atlas records; Jane Elith and Mary Wisz for comments on the

manuscript. Research by M. B. A., R. G. P., and W. T. is supported by the EC Integrated FP6 ALARM (GOCE-CT-2003-506675) project. M. B. A. is a EC FP6 Marie Curie Research Fellow.

References

Akaike H (1974) A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **AU-19**, 716–722.

Araújo MB, Cabeza M, Thuiller W *et al*. (2004) Would climate change drive species out of reserves? An assessment of existing reserve selection methods. *Global Change Biology*, **10**, 1618–1626.

Araújo MB, Pearson RG (2005) Equilibrium of species' distributions with climate. *Ecography* (in press).

Araújo MB, Thuiller W, Williams PH *et al*. (2005a) Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. *Global Ecology and Biogeography*, **14**, 17–30.

Araújo MB, Whittaker RJ, Ladle R *et al*. (2005b) Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, DOI: 10.1111/j.1466-822x.2005.00182.x.

Araújo MB, Williams PH, Fuller RJ (2002) Dynamics of extinction and the selection of nature reserves. *Proceedings of the Royal Society London Series B – Biological Sciences*, **269**, 1971–1980.

Bakkenes M, Alkemade RM, Ihle F *et al*. (2002) Assessing effects of forecasted climate change on the diversity and distribution of European higher plants for 2050. *Global Change Biology*, **8**, 390–407.

Beaumont LJ, Hughes L (2002) Potential changes in the distributions of latitudinally restricted butterfly species in response to climate change. *Global Change Biology*, **8**, 954–971.

Beerling DJ, Huntley B, Bailey JP (1995) Climate and the distribution of *Fallopia japonica*: use of an introduced species to test the predictive capacity of response surface. *Journal of Vegetation Science*, **6**, 269–282.

Berry PM, Dawson TE, Harrison PA *et al*. (2002) Modelling potential impacts of climate change on the bioclimatic envelope of species in Britain and Ireland. *Global Ecology and Biogeography*, **11**, 453–462.

Beutel TS, Beeton RJS, Baxter GS (1999) Building better wildlife – habitat models. *Ecography*, **22**, 219–223.

Box GEP (1979) Some problems of statistics and everyday life. *Journal of the American Statistical Association*, **74**, 1–4.

Burns CE, Johnston KM, Schmitz OJ (2003) Global climate change and mammalian species diversity in U.S. national parks. *PNAS*, **100**, 11474–11477.

Crick HQP (2004) The impact of climate change on birds. *IBIS*, **146**, 48–56.

Elith J, Burgman MA (2002) Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. In: *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds Scott JM, Heglund PJ, Morrison Ml, Raphael MG, Wall WA, Samson FB), pp. 303–314. Island Press, Covelo, CA.

Erasmus BFN, Van Jaarsweld AS, Chown SL *et al*. (2002) Vulnerability of South African animal taxa to climate change. *Global Change Biology*, **8**, 679–693.

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

Fielding AH, Haworth PF (1995) Testing the generality of bird-habitat models. *Conservation Biology*, **9**, 1466–1481.

Gibbons DW, Reid JB, Chapman RA (1993) *The New Atlas of Breeding Birds in Britain and Ireland: 1988–1991*. Poyser, London.

Guisan A, Theurillat JP (2000) Assessing alpine vulnerability to climate change: a modelling perspective. *Integrated Assessment*, **1**, 307–320.

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

Hampe A (2004) Bioclimate envelope models: what they detect and what they hide. *Global Ecology and Biogeography*, **13**, 469–471.

Harrell FE (2001) *Regression Modeling Strategies – with Applications to Linear Models, Logistic Regression and Survival Analysis*. Springer-Verlag, New York. USA.

Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning*. Springer, Berlin.

Hill JK, Thomas CD, Huntley B (1999) Climate and habitat availability determine 20th century changes in a butterfly's range margin. *Proceedings of the Royal Society London Series B – Biological Sciences*, **266**, 1197–1206.

Hodges JS (1991) Six (or so) things you can do with a bad model. *Operations Research*, **39**, 355–365.

Houghton JT, Ding Y, Griggs DJ *et al*. (2001) *Climate Change 2001: the Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge.

Huntley B (1995) Plant species' response to climate change: implications for the conservation of European birds. *IBIS*, **137**, 127–138.

Huntley B, Berry PM, Cramer W *et al*. (1995) Modelling present and potential future ranges of some European higher plants using climate response. *Journal of Biogeography*, **22**, 967–1001.

Huntley B, Green RE, Collingham Y *et al*. (2004) The performance of models relating species geographical distributions to climate is independent of trophic level. *Ecology Letters*, **7**, 417–426.

Iverson LR, Prasad A (1998) Predicting abundance for 80 tree species following climate change in the Eastern United States. *Ecological Monographs*, **68**, 465–485.

Iverson LR, Prasad A, Schwartz MW (1999) Modelling potential future individual tree-species distributions in the Eastern United States under climate change scenario: a case study with *Pinus virginiana*. *Ecological Modelling*, **115**, 77–93.

Ladle R, Jepson P, Araújo MB *et al*. (2004) Dangers of crying wolf over risks of extinction. *Nature*, **428**, 799.

Landis JR, Koch GC (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Loehle C, LeBlanc D (1996) Model-based assessments of climate change effects on forests: a critical review. *Ecological Modelling*, **90**, 1–31.

Martinez-Meyer E, Townsend Peterson A, Hargrove WW (2004) Ecological niches as stable distributional constraints on mammal species, with implications for Pleistocene extinctions and climate change projections for biodiversity. *Global Ecology and Biogeography*, **13**, 305–314.

Midgley GF, Hannah L, Millar D *et al*. (2002) Assessing the vulnerability of species richness to anthropogenic climate change in a biodiversity hotspot. *Global Ecology and Biogeography*, **11**, 445–451.

Midgley GF, Hannah L, Millar D *et al*. (2003) Developing regional and species-level assessments of climate change impacts on biodiversity: a preliminary study in the Cape Floristic Region. *Biological Conservation*, **112**, 87–97.

Miles LJ, Grainger A, Phillips O (2004) The impact of global climate change on tropical forest biodiversity in Amazonia. *Global Ecology and Biogeography*, **13**, 553–566.

Mitchell TD, Carter TR, Jones PD *et al*. (2004) A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: observed record (1901–2000) and 16 scenarios (2001–2100). Working Paper 55, Tyndall Centre for Climate Change Research Norwich (Vol. in review).

New M, Hulme M, Jones PD (2000) Representing twentieth century spacetime climate variability. Part 2: development of 1901–96 monthly grids of terrestrial surface climate. *Journal of Climate*, **13**, 2217–2238.

Olden JD, Jackson DA (2000) Torturing data for the sake of generality: how valid are our regression models? *Ecoscience*, **7**, 501–510.

Olden JD, Jackson DA (2002) A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology*, **47**, 1976–1995.

Olden JD, Jackson DA, Peres-Neto PR (2002) Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329–336.

Oreskes N (1998) Evaluation (not validation) of quantitative models. *Environmental Health in Perspective*, **106**, 1453–1460.

Oreskes N, Shrader-Frechette KS, Belitz K (1994) Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, **263**, 641–646.

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

Pearson RG, Dawson TE, Berry PM *et al*. (2002) SPECIES: a Spatial Evaluation of Climate Impact on the Envelope of Species. *Ecological Modelling*, **154**, 289–300.

Pearson RG, Dawson TE, Liu C (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**, 285–298.

Pearson RG, Thuiller W, Araújo MB *et al*. (2005) Model-based uncertainty in species' range prediction (in review).

Peterson AT (2003a) Predicting the greography of species' invasions via ecological niche modeling. *The Quaterly Review of Biology*, **78**, 419–433.

Peterson AT (2003b) Projected climate change effects on Rocky Mountain and Great Plain birds: generalities of biodiversity consequences. *Global Change Biology*, **9**, 647–655.

Peterson AT, Ortega-Huerta MA, Bartley J *et al*. (2002) Future projections for Mexican faunas under global climate change scenarios. *Nature*, **416**, 626–629.

Peterson AT, Sánchez-Cordero V, Soberón J *et al*. (2001) Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecological Modelling*, **144**, 21–30.

Rastetter EB (1996) Validating models of ecosystem response to global change. *BioScience*, **46**, 190–198.

Raxworthy CJ, Martinez-Meyer E, Horning N *et al*. (2003) Predicting distributions of known and unknown reptile species in Madagascar. *Nature*, **426**, 837–841.

Ripley BD (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Saetersdal M, Birks HJB, Peglar SM (1998) Predicting changes in Fennoscandian vascular-plant species richness as a result of future climatic change. *Journal of Biogeography*, **25**, 111–122.

Segurado P, Araújo MB (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555–1568.

Sharrock JTR (1976) *The atlas of breeding birds of Britain and Ireland*. Poyser, Berkhamsted.

Shrader-Frechette KS, McCoy ED (1993) *Method in Ecology: Strategies for Conservation*. Cambridge University Press, Cambridge.

Skov F, Svenning J-C (2004) Potential impact of climatic change on the distribution of forest herbs in Europe. *Ecography*, **27**, 366–380.

Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Swets KA (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Sykes MT, Prentice IC, Cramer W (1996) A bioclimatic model for the potential distributions of north European tree species under present and future climate. *Journal of Biogeography*, **23**, 203–233.

Teixeira J, Arntzen JW (2002) Potential impact of climate warming on the distribution of the Golden-striped salamander, *Chioglossa lusitanica*, on the Iberian Peninsula. *Biodiversity and Conservation*, **11**, 2167–2176.

Thomas CD, Cameron A, Green RE *et al*. (2004a) Extinction risk from climate change. *Nature*, **427**, 145–148.

Thomas CD, Lennon JJ (1999) Birds extend their ranges northwards. *Nature*, **399**, 213.

Thomas JA, Telfer MG, Roy DB *et al*. (2004b) Comparative loss of British butterflies, birds and plants and the global extinction crisis. *Science*, **303**, 1879–1881.

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

Thuiller W (2004) Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, **10**, 2220–2227.

Thuiller W, Araújo MB, Pearson RG *et al*. (2004a) Uncertainty in predictions of extinction risk. *Nature*, **430**, doi: 10.1038/nature02716..

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

Van Horne B (2002) Approaches to habitat modeling: the tensions between pattern and process and between specificity and generality. In: *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds Scott JM, Heglund PJ, Morrison ML, Raphael MG, Wall WA, Samson FB), Island Press, Covelo, CA.