PERSPECTIVE

Climate change, genetic markers and species distribution modelling

Nicholas J. Gotelli* and John Stanton-Geddes

Department of Biology, University of Vermont, Burlington, VT 05405, USA

ABSTRACT

Ecologists and biogeographers are currently expending great effort forecasting shifts in species geographical ranges that may result from climate change. However, these efforts are problematic because they have mostly relied on presence-only data that ignore within-species genetic diversity. Technological advances in high-throughput sequencing have now made it cost-effective to survey the genetic structure of populations sampled throughout the range of a species. These data can be used to delineate two or more genetic clusters within the species range, and to identify admixtures of individuals within genetic clusters that reflect different patterns of ancestry. Species distribution models (SDMs) applied to the presence and absence of genetic clusters should provide more realistic forecasts of geographical range shifts that take account of genetic variability. High-throughput sequencing and spatially explicit models may be used to further refine these projections.

*Correspondence: Nicholas J. Gotelli,
Department of Biology, University of Vermont,
Burlington, VT 05405, USA.
E-mail: ngotelli@uvm.edu

INTRODUCTION

A major ecological research programme in this new millennium is to understand the consequences of global climate change for populations and communities (Parmesan, 2006; Burrows *et al.*, 2011). A more specific focus within that programme is to build species distribution models (SDMs; Elith & Leathwick, 2009) that predict the probability of occurrence of a species in contemporary climatic and geographical space (Peterson *et al.*, 2011). The output of a SDM can then be combined with projections of future (or past) climatic conditions to forecast (or hindcast) shifts in the occurrence and geographical range of a species (Williams *et al.*, 2013; Burrows *et al.*, 2014). These projected range maps represent the elemental building blocks for understanding how potentially novel assemblages of species will interact in future climates (Blois *et al.*, 2013) and what kind of ecosystem functions (Cramer *et al.*, 2001) and services (Olesen & Bindi, 2002) they may provide.

This *Perspective* will be divided into two parts. In the first part, we briefly review current practices in species distribution modelling (Elith & Leathwick, 2009), and a recent controversy over two competing algorithms – MaxEnt (maximum entropy) and MaxLike (maximum likelihood) – that are used to estimate the probability of species occurrence. Our contention is that neither of these algorithms is optimal because they both are used mostly with presence-only data, which makes the estimation process complicated and the output uncertain. More generally, standard SDM methods applied to presence-only species occurrence records cannot easily accommodate the possibility of local adaptation (Thuiller *et al.*, 2013) and evolutionary change within different parts of a species geographical range (Atkins & Travis, 2010).

In the second part of this essay, we suggest ways that standard genetic methods (Pritchard *et al.*, 2000) and recent technological advances in high-throughput sequencing (Mardis, 2011) can be combined with traditional field surveys to build new SDMs that incorporate genetic variation within a species geographical range. We also highlight a few recent studies that are moving in this direction. Our main argument is that existing methods that are based on presence-only occurrence records are statistically fragile, and they do not allow us to explore the hypothesis that current genetic variation can predict future responses to climate change. Both problems can be addressed by building forecasting models that are based on genetic data.

SPECIES DISTRIBUTION MODELS

Species distribution modelling has its roots in wildlife studies of animal–habitat associations (Scott *et al.*, 2002) that were themselves based on a conceptual framework of the Grinnelian niche as a spatial model of animal abundance (James, 1971). The SDM framework emerged over a period of several decades, in concert with the development of GIS (Goodchild, 2003) and the availability of georeferenced environmental data layers (Beaumont *et al.*, 2005), an expanding collection of multivariate (McGarigal *et al.*, 2000) and geostatistical (Diggle & Ribeiro, 2007) modelling tools and the ever-increasing speed and storage capacity of personal and mainframe computers, combined with the recent success of R (R Core Team, 2013) and other open-source software platforms (Neteler *et al.*, 2012). SDMs have received greater attention in the past decade as a collective sense of societal and scientific urgency over climate change has emerged (Hansen *et al.*, 2013).

A plethora of statistical and mechanistic models has been developed, and there is a steady stream of publications describing new methods (Pagel & Schurr, 2012; Miller, 2014) and improvements to existing ones (Radosavljevic & Anderson, 2014). But the field is currently dominated by MaxEnt (Phillips *et al.*, 2006), a machine-learning algorithm based on principles of maximum entropy (Elith *et al.*, 2011). The algorithm has been extensively tested and benchmarked (Hernandez *et al.*, 2006; Phillips & Dudík, 2008), and works with presence-only species occurrence records. Phillips *et al.* (2006) carefully laid out the theoretical framework, assumptions and methods of MaxEnt. Although not all SDMs are based on MaxEnt, and not all MaxEnt analyses are of SDMs, the Phillips *et al.* (2006) publication has received over 2600 citations, with over 500 in the year 2014 alone (Web of Science, accessed 25 May 2015), and it is the most widely used tool by ecologists and biogeographers for species distribution modelling.

CURRENT CONTROVERSY OVER MAXENT AND MAXLIKE

Royle *et al.* (2012) published an important criticism of the MaxEnt algorithm: it does not estimate the statistic of interest, which is the probability of occurrence of a species, given a set of environmental conditions or habitat variables. Instead, MaxEnt estimates an index of relative habitat suitability, based on presence-only data. Phillips *et al.* (2006) were careful to highlight this distinction, but unfortunately, the scaled index that MaxEnt creates ('logistic output') has almost always been interpreted by the user community as the probability of occurrence (Yackulic *et al.*, 2013). Although suitability and probability of occurrence may be related to one another, the suitability index tends to over-estimate the true occurrence probability outside of a simulated geographical range, but under-estimates it within the range (Royle *et al.*, 2012). This bias may reflect the fact that

MaxEnt default settings for logistic output are based on an overall prevalence $\tau = 0.5$. It is troubling that when the MaxEnt and MaxLike algorithms were applied to the same empirical data sets, their predictions were very different from one another (Fitzpatrick *et al.*, 2013). In these empirical comparisons, cross-validation analysis with random testing and training partitions of presence-only data favoured the MaxLike algorithm. However, MaxLike is not without its own problems. Hastie & Fithian (2013) argued that parameter estimates from the underlying parametric log-linear model of MaxLike can be severely biased if the true functional form of the data is not log-linear. Fitzpatrick *et al.* (2013) found that there can be large uncertainty in estimates of occurrence probabilities, especially for locations that are geographically distant from the sampled area. However, such uncertainty may accurately reflect the limitations of the data, whereas the uncertainty in MaxEnt estimates seems suspiciously low.

Merow & Silander (2014) showed that the underlying models of MaxEnt and MaxLike are similar to one another, and that the slope coefficients from both are comparable to those generated by a generalized linear model (GLM) applied to presence/absence data. Both logistic regression and MaxEnt are formally equivalent to a model of an inhomogeneous Poisson process (IPP; Fithian & Hastie, 2013). However, the logistic output from MaxEnt (at least in its current configuration) should not be used as an estimate of occurrence probability, and some care must be taken to modify the default settings of MaxEnt to generate slope values that are comparable to those of MaxLike (Merow *et al.*, 2013).

Although the jury is still out on these methods, we currently favour the MaxLike model for presence-only data because, for climate change forecasting, we need to estimate the absolute probability of occurrence, not the relative habitat suitability. Two grid cells might have the same habitat suitability index, but if the probability of occurrence is 0.9900, then both grid cells will almost certainly be occupied, whereas if the probability of occurrence is 0.0099, both grid cells will almost certainly be empty. It is these absolute probabilities of occurrence that determine the forecasted shift in species ecological ranges under different climate change scenarios. With presence-only data derived from geo-referenced specimen records, the sampling area for estimating occurrence probabilities is undefined, and this is a problem for both MaxEnt and MaxLike. Operationally, the area often will be associated with the grid size for sampling of environmental data in the study.

The criticisms that Hastie & Fithian (2013) raised against MaxLike seem (to us) comparable to those that arise when fitting a simple regression model to a small data set for which the relationship between *x* and *y* may not be strictly linear. Because MaxEnt with its current default settings is not estimating the probability of species occurrence (Merow & Silander, 2014), we believe its output is not appropriate for estimating range expansions and contractions under scenarios of climate change.

MaxEnt, MaxLike and other complex, non-intuitive alternatives are necessary because ecologists continue to work mostly with presence-only data (Newbold, 2010). If the data set includes both presences and absences, then even a simple logistic regression model (Pearce & Ferrier, 2000) is on firmer statistical ground than a MaxEnt or MaxLike model fit to presence-only data. With presence/absence data randomly sampled across the landscape, logistic regression in the framework of GLMs can be readily fit (McCullagh & Nelder, 1989) to estimate the probability of occurrence. Moreover, well-established geostatistical models (Diggle & Ribeiro, 2007) can be used to incorporate endogenous spatial autocorrelation, which is not explicitly accounted for in the mainstream SDM framework. Analyses of spatial autocorrelation are important because these kinds of models can potentially outperform traditional climate-based models of species occurrence and abundance (Bahn & McGill, 2007).

Looking beyond the current statistical controversy over how to analyse presence-only data, the basic modelling framework of SDMs in general has been criticized on a number of grounds, including: assuming that detection probabilities equal 1.0 (Royle & Dorazio, 2008), failing to incorporate biotic interactions (both positive and negative) that can limit range distributions (Araújo & Luoto, 2007), unrealistically assuming that populations have reached an equilibrium distribution with respect to climate (Svenning & Sandel, 2013), ignoring heterogeneity in population and genetic structure in different parts of a species geographical range (Hampe & Petit, 2005) and assuming a lack of dispersal constraints (Pagel & Schurr, 2012). Incorporating molecular data addresses only the fourth of these assumptions, but it represents an important step forward for modelling the effects of climate change on species geographical ranges (Reusch & Wood, 2007).

Here, we propose sampling methods and genetic assays for realistically representing the spatial genetic structure (Hampe *et al.*, 2010) of a species across its geographical range. Next, we describe how genetic clusters can be identified based on structured demes and SNP variants within the geographical range of a species. When the data are organized this way, the clusters will usually generate presence/absence occurrence vectors because most clusters will not be represented at all sampling locations. These presence/absence vectors for each cluster can then be fit with a variety of stacked SDMs to forecast range expansions and contractions that take account of genetic variability within and among populations.

RANGE SHIFTS AND POPULATION GENETIC STRUCTURE

Standard SDMs that use presence (or presence/absence) data combined with geo-referenced measures of biotic and abiotic variables assume that a species has a single environmental niche and that all individuals respond similarly to changes in environmental conditions. At the population genetic level,

this is equivalent to the assumption of a single panmictic population encompassing the entire geographical range of the species, with little or no genetic variation that may respond to changing environmental conditions. But there is abundant evidence that, across their geographical ranges, many species are organized into differentiated genetic lineages that may be adapted to local conditions (Linhart & Grant, 1996; Leimu & Fischer, 2008; Hereford, 2009). Moreover, there is strong evidence that populations differ in their adaptive potential to respond to environmental change (Pujol & Pannell, 2008; Shaw & Etterson, 2012). Therefore, to realistically model shifts in geographical ranges, we need to incorporate this intra-specific variation. We first describe a generic framework for the sampling and analysis of species occurrences that are organized into genetic clusters. We then provide some specific proposals for the kinds of genetic data that can be collected to implement this framework.

INDIVIDUALS, CLUSTERS AND SITES

The methods we propose can be carried out by sampling of random individuals across the geographical range of a species. Because SDMs are inherently a regression-based approach, there is, in theory, no advantage to replicate sampling at a single location, particularly as there is an inevitable trade-off with sampling among locations (Gotelli & Ellison, 2012). However, purely random sampling of a single individual at different locations may require too much travel in the field and be too costly.

Moreover, collecting multiple individuals at one location is important for more thorough sampling of genetic diversity and detection of rare genetic clusters. Even if geo-referenced specimens can be sampled from different locations, environmental data may only be available at coarser spatial grains, so that neighbouring samples must be binned and assigned the same set of locations and environmental variables. Preliminary genetic data can be used to define an optimal strategy based on the allocation of N individuals into K clusters (the optimal_sampling() function in the R package GSTUDIO; Dyer, 2013).

Finally, although purely random sampling provides the strongest basis for general inference, it may not necessarily yield the most useful information for accurate forecasting. Stratified sampling that emphasizes extreme environments and locations near the edge of a species geographical range may provide better resolution than random sampling for forecasting the expansion and contraction of geographical ranges with climate change.

GENETIC ANALYSES

Once the sampled individuals are in hand, their genetic identity can be assayed with a variety of molecular markers, including microsatellites (Broman *et al.*, 1998), amplified fragment-length polymorphisms (AFLPs; Bonin & Manel, 2007) and single nucleotide polymorphisms (SNPs; Vignal

et al., 2002). For ecologists, AFLPs may be the easiest because they require no species-specific information and allow for screening many loci at a low cost (Gugerli et al., 2008). However, with current high-throughput sequencing technologies (e.g. Illumina), it is becoming equally cost-effective to collect SNP data using restriction-associated digestion (RAD) sequencing (Peterson et al., 2012) or genotyping-by-sequencing (GBS; Elshire et al., 2011).

We personally favour the use of SNPs because they provide additional fine-scale resolution of population genetic structure, but for the purposes of identifying major genetic clusters, any of these methods can be used for the analyses we propose here. However, we appreciate that the majority of SNPs are selectively neutral with respect to geography (Coop & Ralph, 2012), so that non-random spatial patterns of SNPs may reflect dispersal and demographic history, rather than adaptive variation. Even for traits that are under selection, fixed differences in selected traits may not remain after neutral demographic history is taken into account (e.g. Coop et al., 2009). So there is certainly a risk that correlates of population genetic structure with contemporary environmental variables will be confounded with the effects of historical demographic processes (Dyer et al., 2010). On the other hand, teasing apart the role of neutral and selective forces in structuring populations will surely give us better forecasting tools than current methods that ignore genetic data and rely only on species occurrence records.

The data structure for these analyses consists of a matrix in which each row is an individual, and each column is an allelic variant. Different columns might correspond to single nucleotides for SNPs, DNA fragment-length polymorphisms for AFLPs and simple sequence repeats for microsatellites. The entries in this data matrix indicate the genetic variant that was scored for each individual at each allelic variant. Programs such as STRUCTURE (Pritchard et al., 2000), INSTRUCT (Gao et al., 2011) and GENELAND (Guillot et al., 2012) use these data to assign individuals to different genetic clusters, although care must be taken in the use of these programs (see Gilbert et al., 2012). Importantly, all of these programs assign individuals with admixtures to multiple clusters. The admixtures reflect the different ancestry and gene clusters represented by a single individual. In contrast, a graph-theoretic approach such as POPGRAPH (Dyer & Nason, 2004) evaluates the connectedness of populations based on their genetic covariances. Groups of connected populations can then be used as the basis for ecological modelling, as in Sork et al. (2010).

Regardless of which kind of genetic data are used or what kind of statistical algorithm is used to create the clusters, such analyses assume that: (1) individuals and populations are sampled randomly within the geographical range of the species; and (2) the genetic differences among the clusters potentially reflect adaptive variation to different abiotic or biotic conditions locally present within the geographical range of the species (Hancock et al., 2011).

PRESENCE/ABSENCE DATA AND THE SAMPLING OF EMPTY SITES

Sampling individuals from multiple geo-referenced populations within the genetic range of a species generates a standard 'presence-only' data set that can be combined with abiotic layers for a standard analysis of presence-only data. However, once the genetic clusters have been delineated, each group can be analysed as a set of presence/absence data because it is unlikely that any of the groups will be represented at every site. We emphasize that such an analysis does not impute 'pseudo-absences' to locations where no sample information is available (the typical situation with presence-only occurrence data). Instead, these are true sample zeros, which represent sites where a particular genetic cluster has not been detected with standardized sampling effort (although it might have been with more intensive sampling).

This may seem like a trivial observation, but it means that genetic sampling of populations across the geographical range of a species will yield presence/absence data for genetically delineated groups. It is not necessary to use the complicated (and controversial) algorithms to analyse such data. Instead, the probability of occurrence of each group can be estimated in a straightforward way with a GLM for presence/absence data. Thus, genetic sampling provides an effective operational solution to the problem of obtaining presence/absence data to analyse the probability of occurrence. Clusters designated by genetic analysis may be spatially aggregated in different subregions of the geographical range, although this is not a requirement for analysis, and should not be used as a method for designating the clusters. Unless the clusters delineated by genetic analysis are perfectly allopatric and do not co-occur, at least some of the sampled sites will support more than a single cluster.

In addition to the presence/absence data generated by the delineation of groups within the geographical range, augmenting the sampling beyond the known or suspected geographical margins of the species ranges is still very important, for two reasons. First, if abiotic conditions beyond the range margin do not differ from conditions within the contemporary geographical range, dispersal barriers or biotic interactions may be preventing further range expansion (Gaston, 2009; Marsico & Hellmann, 2009). If populations can be experimentally transplanted beyond the geographical range boundary, much stronger conclusions can be drawn about the proximate causes of range boundaries (Stanton-Geddes et al., 2012; Hargreaves et al., 2014). Second, sampling beyond the existing range boundaries is important because range shifts are already under way with climate change, and we need accurate baseline data to establish the trajectory of ongoing range expansions and contractions.

Nonetheless, it is still a difficult question to decide how much sampling should be done beyond the range boundaries

of a species, and thus how much additional 'absence' data should be included in the statistical analysis. On the one hand, including lots of absence data may tend to lead to over-fitting of climatic variables and simply demonstrate that temperate species are unlikely to occur in tropical areas (Bahn & McGill, 2007). Data sets with too many absence values from extreme climate regions may also obscure effects of metapopulation dynamics (Hanski & Ovaskainen, 2000), limited dispersal (Fenster et al., 2003), philopatry (Weatherhead & Forbes, 1992) and other endogenous processes that can be expressed as spatial autocorrelation. As Bahn & McGill (2007) have shown, models of spatial variation in abundance of North American birds that contain only spatial autocorrelation can often outperform species distribution models that include standard climate variables and assume that samples are statistically independent of one another.

On the other hand, many climate models do forecast strong changes in abiotic variables over the next century, leading to the possibility of no-analogue climates and communities (Urban et al., 2012), for which there is already abundant evidence from the fossil record (Williams et al., 2013). Thus, some effort should be made to include more spatially varied sampling to encompass some of the potential conditions that arise from these forecasts. Whatever decision is made about sampling, forecasting models should certainly explore the consequences of including versus excluding absence data or training data that are collected beyond the current range boundaries of a species (Radosavljevic & Anderson, 2014).

MODELLING SPECIES RANGE LIMITS WITH GENETICALLY DISTINCT CLUSTERS

Although a variety of methods can be used to estimate the probability of species occurrence with simple presence/absence data, we favour a GLM with a binomial distribution and a logit link function (McCullagh & Nelder, 1989). This model properly captures the variance structure of presence/absence data, allows for model selection with the Akaike information criterion (AIC) for a variety of linear or polynomial predictor variables, and allows for modelling of spatial autocorrelation [see Zuur et al. (2009) for details and extensions]. At the level of individual genetic markers, Joost et al. (2007) use similar regression approaches for associations of allelic frequencies at marker loci and environmental variables.

This kind of model could be applied separately to each genetic cluster, raising the interesting possibility that different clusters may be best fit with different sets of predictor variables. The prediction for each of the m clusters is P_{kj}, the probability that cluster k is present in location j. To forecast range shifts under climate change, the set of prediction sites can be expanded beyond the original set of sites where the data were collected.

Next, we wish to predict the probability that the species as a whole occurs in a particular location. Following the method of Pearman et al. (2010), this joint probability is:

$$P(occurrence)_j = 1 - \prod_{k=1}^{m}(1 - P_{kj})$$

which is the probability that at least one of the m clusters is present at site j. As in other SDM analyses, these probabilities can be used in conjunction with forecasts based on future climate scenarios (Collins et al., 2006), or hindcasts based on palaeoclimate models (Williams et al., 2013).

A related type of diversity index would be the expected number of groups in each location:

$$E(number\ of\ groups)_j = \sum_{k=1}^{m} P_{kj}$$

Both of these calculations assume that the occurrence probabilities for each cluster are independent (or at least uncorrelated). However, with natural selection, this might not be the case, because the prevalence of one group or genotype may occur at the expense of other groups (Waters et al., 2013). Moreover, genetic mixing of local demes means that the clusters themselves may be changing dynamically (Keller & Taylor, 2010). However, these complications may be difficult to model with 'snapshot' data of current genetic structure.

Assignment to clusters with STRUCTURE allows for the possibility of admixture: some individuals may be assigned probabilistically to more than one cluster, which reflects interbreeding between two previously separated populations. To use the full information present in admixture assignments, we propose using the non-zero admixture values for each individual as weights for the SDM created for that cluster. The data vector is still a binary sequence of presences or absences, but the non-zero admixtures serve as weights. Individuals with no admixture would be given a full weight of 1, as would individuals that are not represented in a cluster (Table 1).

EXAMPLES AND EXTENSIONS

To our knowledge, no one has carried out precisely the protocol that we have advocated, although individual steps have been featured in several other studies. For example, the premise of using SDMs at a lower taxonomic level than the species has already generated promising results. Gonzalez et al. (2011) modelled the distribution of 15 subspecies of Peromyscus mice in the south-eastern USA, as did D'Amen et al. (2013) for phylogeographical lineages of 10 African mammal species. In both studies, the composite model provided a better fit to the data (Gonzalez et al., 2011) than species-level models, which tended to over-predict areas of suitable habitat for African mammals (D'Amen et al., 2013). However, phylogeographical delineation of lineages may reflect genetic similarities from the deep past. These divisions may not necessarily give good results for forecasting future range shifts. For forecasting purposes, we favour analyses of SNPs, AFLPs or microsatellites, which will better reflect contemporary genetic structure of populations.

Table 1 Hypothetical data from a STRUCTURE analysis. Each row represents a different individual, with genetic clustering resulting in three clusters (A,B,C). Columns 2–4 give the proportion ancestry (Anc) for each of the three clusters, which must sum to 1.0 for each individual. Columns 5–7 give the binary presence/absence vector (Vec) that would be used in the generalized linear model (GLM). The last three columns give the weights (W) that would be used with the GLM to indicate admixture proportions and also absences, which receive full weights of 1.0. Other columns (not shown) would give georeferenced coordinates for each individual, and environmental variables or covariates measured for the location of each individual.

Individual	Anc (A)	Anc (B)	Anc (C)	Vec (A)	Vec (B)	Vec (C)	W (A)	W (B)	W (C)
1	1.0	0.0	0.0	1	0	0	1.0	1.0	1.0
2	0.0	0.5	0.5	0	1	1	1.0	0.5	0.5
3	0.2	0.1	0.7	1	1	1	0.2	0.1	0.7
300	0.2	0.0	0.8	1	0	1	0.2	1.0	0.8

Admixture analyses of molecular markers have revealed strong genetic and spatial structuring of populations that are likely to respond to climate change. For example, in a study of California valley oak, Sork et al. (2010) assayed six chloroplast microsatellite markers, and used POPGRAPH to delineate a network of populations connected by gene flow. They then selected four spatially and genetically distinct subgroups and used MaxEnt to forecast future distributions under different scenarios of climate change. Using a different approach, Jay et al. (2012) assayed between 57 and 187 AFLPs per species of 20 European alpine plants. They created Bayesian ancestry distribution models (ADMs) to estimate genetic ancestry and clusters that reflect correlations of genetic and environmental variables. The inferred contact zone between warm and cold clusters was expected to move 92 km with 2 °C warming and 188 km with 4 °C warming. One strength of the ADM approach is its focus on ancestry, thus allowing for the possibility of changes in the genetic structure of populations in the face of climate change.

SUMMARY

We have proposed a method of data collection (molecular data from individuals collected throughout a species geographical range) and analysis (SDMs of presence/absence data for different genetic clusters) to provide more realistic forecasts of species ranges under climate change. We think these methods are an improvement on the current paradigm, which has until recently emphasized the analysis of presence-only data with no genetic information. Our method incorporates genetic variation at a geographical scale, and uses a GLM logistic regression analysis of presence/absence data for the genetic clusters, which is an improvement over statistically fragile methods for presence-only data that do not incorporate genetic information.

At least two other forecasting challenges remain. The first problem is dispersal. Even if there is a reservoir of genetic variation that can allow for persistence, individuals with those genotypes have to be able to reach those regions of suitable climate (Marsico & Hellmann, 2009). The second problem is time. Current forecasting scenarios are static, and examine future climates as simple snapshots, without considering the velocity of climate change (Loarie et al., 2009), and

the dynamic conditions that species will face along the way. Perhaps spatially explicit simulation models (Gotelli et al., 2009) that forecast climate change and limited-distance dispersal over short time steps may be the best approach. Looking to the future, the combination of high-density molecular data (Santure et al., 2013) and phenotype data for ecologically important traits (Etterson, 2004) may allow for improved forecasts that take into account the potential for evolutionary change (Hancock et al., 2011; Fitzpatrick & Keller, 2015).

REFERENCES

Araújo, M.B. & Luoto, M. (2007) The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography*, **16**, 743–753.

Atkins, K.E. & Travis, J.M.J. (2010) Local adaptation and the evolution of species' ranges under climate change. *Journal of Theoretical Biology*, **266**, 449–457.

Bahn, V. & McGill, B.J. (2007) Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, **16**, 733–742.

Beaumont, L.J., Hughes, L. & Poulsen, M. (2005) Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. *Ecological Modelling*, **186**, 250–269.

Blois, J.L., Zarnetske, P.L., Fitzpatrick, M.C. & Finnegan, S. (2013) Climate change and the past, present & future of biotic interactions. *Science*, **341**, 499–504.

Bonin, A.E.D. & Manel, S. (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Molecular Ecology*, **16**, 3737–3758.

Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. & Weber, J.L. (1998) Comprehensive human genetic maps:

individual and sex-specific variation in recombination. *American Journal of Human Genetics*, **63**, 861–869.

Burrows, M.T., Schoeman, D.S., Buckley, L.B., Moore, P., Poloczanska, E.S., Brander, K.M., Brown, C., Bruno, J.F., Duarte, C.M., Halpern, B.S., Holding, J., Kappel, C.V., Kiessling, W., O'Connor, M.I., Pandolfi, J.M., Parmesan, C., Schwing, F.B., Sydeman, W.J. & Richardson, A.J. (2011) The pace of shifting climate in marine and terrestrial ecosystems. *Science*, **334**, 652–655.

Burrows, M.T., Schoeman, D.S., Richardson, A.J. *et al.* (2014) Geographical limits to species-range shifts are suggested by climate velocity. *Nature*, **507**, 492.

Collins, W.D., Bitz, C.M., Blackmon, M.L., Bonan, G.B., Bretherton, C.S., Carton, J.A., Chang, P., Doney, S.C., Hack, J.J., Henderson, T.B., Kiehl, J.T., Large, W.G., McKenna, D.S., Santer, B.D. & Smith, R.D. (2006) The community climate system model version 3 (CCSM3). *Journal of Climate*, **19**, 2122–2143.

Coop, G. & Ralph, P. (2012) Patterns of neutral diversity under general models of selective sweeps. *Genetics*, **192**, 205–240.

Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W. & Pritchard, J.K. (2009) The role of geography in human adaptation. *PLoS Genetics*, **5**, e1000500.

Cramer, W., Bondeau, A., Woodward, F.I., Prentice, I.C., Betts, R.A., Brovkin, V., Cox, P.M., Fisher, V., Foley, J.A., Friend, A.D., Kucharik, C., Lomas, M.R., Ramankutty, N., Sitch, S., Smith, B., White, A. & Young-Molling, C. (2001) Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models. *Global Change Biology*, **7**, 357–373.

D'Amen, M., Zimmermann, N.E. & Pearman, P.B. (2013) Conservation of phylogeographic lineages under climate change. *Global Ecology and Biogeography*, **22**, 93–104.

Diggle, P.J. & Ribeiro, P.J. (2007) *Model-based geostatistics*. Springer-Verlag, New York.

Dyer, R.J. (2013) *gstudio: analyses and functions related to the spatial analysis of genetic marker data*. Available at: https://github.com/dyerlab/ (last accessed 18 June 2015).

Dyer, R.J. & Nason, J.D. (2004) Population graphs: the graph theoretic shape of genetic structure. *Molecular Ecology*, **13**, 1713–1727.

Dyer, R.J., Nason, J.D. & Garrick, R.C. (2010) Landscape modelling of gene flow: improved power using conditional genetic distance derived from the topology of population networks. *Molecular Ecology*, **19**, 3746–3759.

Elith, J. & Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. & Yates, C.J. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. & Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379–e19379.

Etterson, J.R. (2004) Evolutionary potential of *Chamaecrista fasciculata* in relation to climate change. I. Clinal patterns of selection along an environmental gradient in the great plains. *Evolution*, **58**, 1446–1458.

Fenster, C.B., Vekemans, X. & Hardy, O.J. (2003) Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (Leguminosae). *Evolution*, **57**, 995–1007.

Fithian, W. & Hastie, T. (2013) Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, **7**, 1917–1939.

Fitzpatrick, M.C. & Keller, S.R. (2015) Ecological genomics meets community-level modeling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, **18**, 1–16.

Fitzpatrick, M.C., Gotelli, N.J. & Ellison, A.M. (2013) MaxEnt versus MaxLike: empirical comparisons with ant species distributions. *Ecosphere*, **4**, Article 55.

Gao, H., Bryc, K. & Bustamante, C. (2011) On identifying the optimal number of population clusters via the deviance information criterion. *PLoS ONE*, **6**, e21014.

Gaston, K.J. (2009) Geographic range limits: achieving synthesis. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 1395–1406.

Gilbert, K.J., Andrew, R.L., Bock, D.G., Franklin, M.T., Kane, N.C., Moore, J.-S., Moyers, B.T., Renaut, S., Rennison, D.J., Veen, T. & Vines, T.H. (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. *Molecular Ecology*, **21**, 4925–4930.

Gonzalez, S.C., Soto-Centeno, J.A. & Reed, D.L. (2011) Population distribution models: species distributions are better modeled using biologically relevant data partitions. *BMC Ecology*, **11**, 20.

Goodchild, M.E. (2003) Geographic information science and systems for environmental management. *Annual Review of Environment and Resources*, **28**, 493–519.

Gotelli, N.J. & Ellison, A.M. (2012) *A primer of ecological statistics*, 2nd edn. Sinauer Associates, Sunderland, MA.

Gotelli, N.J., Anderson, M.J., Arita, H.T. *et al.* (2009) Patterns and causes of species richness: a general simulation model for macroecology. *Ecology Letters*, **12**, 873–886.

Gugerli, F., Englisch, T., Niklfeld, H., Tribsch, A., Mirek, Z., Ronikier, M., Zimmermann, N.E., Holderegger, R. & Taberlet, P. (2008) Relationships among levels of biodiversity and the relevance of intraspecific diversity in conservation – a project synopsis. *Perspectives in Plant Ecology, Evolution and Systematics*, **10**, 259–281.

Guillot, G., Renaud, S., Ledevin, R., Michaux, J. & Claude, J. (2012) A unifying model for the analysis of phenotypic, genetic and geographic data. *Systematic Biology*, **61**, 897–911.

Hampe, A. & Petit, R.J. (2005) Conserving biodiversity under climate change: the rear edge matters. *Ecology Letters*, **8**, 461–467.

Hampe, A., El Masri, L. & Petit, R.J. (2010) Origin of spatial genetic structure in an expanding oak population. *Molecular Ecology*, **19**, 459–471.

Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C., Roux, F. & Bergelson, J. (2011) Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, **334**, 83–86.

Hansen, J., Kharecha, P., Sato, M., Masson-Delmotte, V., Ackerman, F., Beerling, D.J., Hearty, P.J., Hoegh-Guldberg, O., Hsu, S.L., Parmesan, C., Rockstrom, J., Rohling, E.J., Sachs, J., Smith, P., Steffen, K., Van Susteren, L., von Schuckmann, K. & Zachos, J.C. (2013) Assessing "dangerous climate change": required reduction of carbon emissions to protect young people, future generations and nature. *PLoS ONE*, **8**, e81648.

Hanski, I. & Ovaskainen, O. (2000) The metapopulation capacity of a fragmented landscape. *Nature*, **404**, 755–758.

Hargreaves, A.L., Samis, K.E. & Eckert, C.G. (2014) Are species' range limits simply niche limits writ large? A review of transplant experiments beyond the range. *The American Naturalist*, **183**, 157–173.

Hastie, T. & Fithian, W. (2013) Inference from presence-only data; the ongoing controversy. *Ecography*, **36**, 864–867.

Hereford, J. (2009) A quantitative survey of local adaptation and fitness trade-offs. *The American Naturalist*, **173**, 579–588.

Hernandez, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

James, F.C. (1971) Ordinations of habitat relationships among breeding birds. *The Wilson Bulletin*, **83**, 215–236.

Jay, F., Manel, S., Alvarez, N., Durand, E.Y., Thuiller, W., Holderegger, R., Taberlet, P. & François, O. (2012) Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, **21**, 2354–2368.

Joost, S., Bonin, A., Bruford, M.W., Despres, L., Connord, C., Erhardt, G. & Taberlet, P. (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955–3969.

Keller, S.R. & Taylor, D.R. (2010) Genomic admixture increases fitness during a biological invasion. *Journal of Evolutionary Biology*, **23**, 1720–1731.

Leimu, R. & Fischer, M. (2008) A meta-analysis of local adaptation in plants. *PLoS ONE*, **3**, e4010.

Linhart, Y.B. & Grant, M.C. (1996) Evolutionary significance of local genetic differentiation in plants. *Annual Review of Ecology and Systematics*, **27**, 237–277.

Loarie, S.R., Duffy, P.B., Hamilton, H., Asner, G.P., Field, C.B. & Ackerly, D.D. (2009) The velocity of climate change. *Nature*, **462**, 1052–1055.

Mardis, E.R. (2011) A decade's perspective on DNA sequencing technology. *Nature*, **470**, 198–203.

Marsico, T.D. & Hellmann, J.J. (2009) Dispersal limitation inferred from an experimental translocation of *Lomatium* (Apiaceae) species outside their geographic ranges. *Oikos*, **118**, 1783–1792.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*, 2nd edn. Chapman & Hall, London.

McGarigal, K., Cushman, S. & Stafford, S. (2000) *Multivariate statistics for wildlife and ecology research*. Springer-Verlag, New York.

Merow, C. & Silander, J.A. (2014) A comparison of Maxlike and Maxent for modelling species distributions. *Methods in Ecology and Evolution*, **5**, 215–225.

Merow, C., Smith, M.J. & Silander, J.A. (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Miller, J.A. (2014) Virtual species distribution models: using simulated data to evaluate aspects of model performance. *Progress in Physical Geography*, **38**, 117–128.

Neteler, M., Bowman, M.H., Landa, M. & Metz, M. (2012) GRASS GIS: a multi-purpose open source GIS. *Environmental Modelling & Software*, **31**, 124–130.

Newbold, T., Reader, T., El-Gabbas, A., Berg, W., Shohdi, W.M., Zalat, S., El Din, S.B. & Gilbert, F. (2010) Testing the accuracy of species distribution models using species records from a new field survey. *Oikos*, **119**, 1326–1334.

Olesen, J.E. & Bindi, M. (2002) Consequences of climate change for European agricultural productivity, land use and policy. *European Journal of Agronomy*, **16**, 239–262.

Pagel, J. & Schurr, F.M. (2012) Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography*, **21**, 293–304.

Parmesan, C. (2006) Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 637–669.

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

Pearman, P.B., D'Amen, M., Graham, C.H., Thuiller, W. & Zimmermann, N.E. (2010) Within-taxon niche structure: niche conservatism, divergence and predicted effects of climate change. *Ecography*, **33**, 990–1003.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, M.N. & Araújo, M.B. (2011) *Ecological niches and geographic distributions*. Princeton University Press, Princeton, NJ.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. & Hoekstra, H.E. (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pujol, B. & Pannell, J.R. (2008) Reduced responses to selection after species range expansion. *Science*, **321**, 96.

R Core Team (2013) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radosavljevic, A. & Anderson, R.P. (2014) Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, **41**, 629–643.

Reusch, T.B.H. & Wood, T.E. (2007) Molecular ecology of global change. *Molecular Ecology*, **16**, 3973–3992.

Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Elsevier, London.

Royle, J.A., Chandler, R.B., Yackulic, C. & Nichols, J.D. (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554.

Santure, A.W., De Cauwer, I., Robinson, M.R., Poissant, J., Sheldon, B.C. & Slate, J. (2013) Genomic dissection of variation in clutch size and egg mass in a wild great tit (*Parus major*) population. *Molecular Ecology*, **22**, 3949–3962.

Scott, J.M., Heglund, P.J. & Morrison, M.I. (2002) *Predicting species occurrences: issues of accuracy and scale*. Island Press, Washington, DC.

Shaw, R.G. & Etterson, J.R. (2012) Rapid climate change and the rate of adaptation: insight from experimental quantitative genetics. *New Phytologist*, **195**, 752–765.

Sork, V.L., Davis, F.W., Westfall, R., Flint, A., Ikegami, M., Wang, H. & Grivet, D. (2010) Gene movement and genetic association with regional climate gradients in California valley oak (*Quercus lobata* Née) in the face of climate change. *Molecular Ecology*, **19**, 3806–3823.

Stanton-Geddes, J., Tiffin, P. & Shaw, R.G. (2012) Role of climate and competitors in limiting fitness across range edges of an annual plant. *Ecology*, **93**, 1604–1613.

Svenning, J.C. & Sandel, B. (2013) Disequilibrium vegetation dynamics under future climate change. *American Journal of Botany*, **100**, 1266–1286.

Thuiller, W., Munkemuller, T., Lavergne, S., Mouillot, D., Mouquet, N., Schiffers, K. & Gravel, D. (2013) A road map for integrating eco-evolutionary processes into biodiversity models. *Ecology Letters*, **16**, 94–105.

Urban, M., Tewksbury, J. & Sheldon, K. (2012) On a collision course: competition and dispersal differences create no-analogue communities and cause extinctions during climate change. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 2072–2080.

Vignal, A., Milan, D., SanCristobal, M. & Eggen, A. (2002) A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, **34**, 275–305.

Waters, J.M., Fraser, C.I. & Hewitt, G.M. (2013) Founder takes all: density-dependent processes structure biodiversity. *Trends in Ecology and Evolution*, **28**, 78–85.

Weatherhead, P.J. & Forbes, M.R.L. (1992) Natal philopatry in passerine birds: genetic or ecological influences. *Behavioral Ecology*, **5**, 426–433.

Williams, J.W., Blois, J.L., Gill, J.L., Gonzalez, L.M., Grimm, E.C., Jackson, S.T., Ordonez, A., Shuman, B. & Veloz, S.D. (2013) Model systems for a no-analog future: species associations and climates during the last deglaciation. *Annals of the New York Academy of Sciences*, **1297**, 29–43.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Grant, E.H.C. & Veran, S. (2013) Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, **4**, 236–243.

Zuur, A., Ieno, E.N., Walker, N., Savaliev, A.A. & Smith, G.M. (2009) *Mixed effects models and extensions in ecology with R*. Springer-Verlag, New York.

Editor: Richard Ladle