Journal of Biogeography (J. Biogeogr.) (2012) 39, 2225–2239

SPECIAL ISSUE

Parameter and uncertainty estimation for process-oriented population and distribution models: data, statistics and the niche

Glenn Marion[1]*, Greg J. McInerny[2], Jörn Pagel[3,4], Stephen Catterall[1], Alex R. Cook[5], Florian Hartig[6] and Robert B. O'Hara[4]

[1]Biomathematics and Statistics Scotland, Edinburgh, EH9 3JZ, UK, [2]Computational Ecology and Environmental Science Group, Computational Science Laboratory, Microsoft Research, Cambridge, CB3 0FB, UK, [3]Plant Ecology and Conservation Biology, University of Potsdam, 14469, Potsdam, Germany, [4]Biodiversity and Climate Research Centre, D-60325, Frankfurt am Main, Germany, [5]Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore, [6]Department of Ecological Modelling, UFZ – Helmholtz Centre for Environmental Research, 04318, Leipzig, Germany

*Correspondence: Glenn Marion, Biomathematics and Statistics Scotland, James Clerk Maxwell Building, The King's Buildings, Edinburgh EH9 3JZ, UK.
E-mail: glenn@bioss.ac.uk

ABSTRACT

The spatial distribution of a species is determined by dynamic processes such as reproduction, mortality and dispersal. Conventional static species distribution models (SDMs) do not incorporate these processes explicitly. This limits their applicability, particularly for non-equilibrium situations such as invasions or climate change. In this paper we show how dynamic SDMs can be formulated and fitted to data within a Bayesian framework. Our focus is on discrete state-space Markov process models which provide a flexible framework to account for stochasticity in key demographic processes, including dispersal, growth and competition. We show how to construct likelihood functions for such models (both discrete and continuous time versions) and how these can be combined with suitable observation models to conduct Bayesian parameter inference using computational techniques such as Markov chain Monte Carlo. We illustrate the current state-of-the-art with three contrasting examples using both simulated and empirical data. The use of simulated data allows the robustness of the methods to be tested with respect to deficiencies in both data and model. These examples show how mechanistic understanding of the processes that determine distribution and abundance can be combined with different sources of information at a range of spatial and temporal scales. Application of such techniques will enable more reliable inference and projections, e.g. under future climate change scenarios than is possible with purely correlative approaches. Conversely, confronting such process-oriented niche models with abundance and distribution data will test current understanding and may ultimately feedback to improve underlying ecological theory.

Keywords

Bayesian inference, demography, dispersal, dynamic models, dynamic range models, establishment, global change, niche models, species distribution models.

INTRODUCTION

Species' distributions and the processes that determine them are dynamic. However, despite the fact that such processes (which unfold over time) are more naturally incorporated into dynamic models, the vast majority of statistical species distributions models (both spatial and non-spatial) are static in nature. This bias reflects both a paucity of suitable temporal and spatio-temporal data and current difficulties in confronting dynamic species distribution models (SDMs) with such data. The need for species distribution modelling to account for the dynamics of demographic processes has repeatedly been acknowledged (see e.g. Guisan & Thuiller, 2005; Schurr *et al.*, 2007; Thuiller *et al.*, 2008).

Experimental or field-study data from direct observation of processes are typically not sufficient to fully parameterize dynamic SDMs, e.g. if suitable observations are not available for a particular process, or a representative sample of locations, or if models are applied at scales larger than that at which such observations are made. Thus, there is an urgent need to develop appropriate statistical tools for these models that combine a range of sources of information, including experimental and field data, large-scale observational studies and expert knowledge about the ecology of the species. Here

doi:10.1111/j.1365-2699.2012.02772.x

Journal of Biogeography

we present a general framework for statistical inference which may be applied to both static and dynamic/process-based SDMs. The ability to infer parameters and compare competing process-orientated SDMs will lead to better understanding of the mechanisms that determine species distributions and enable more accurate prediction of species range shifts, e.g. under climate change.

There has been extensive application of statistical regression techniques to the analysis of species range data (Guisan & Zimmermann, 2000; Elith & Leathwick, 2009). These techniques are often used to construct ecological niche or SDMs by correlating current observed spatial distribution with climatic and other environmental variables. The resulting models may then be used to predict the potential future distribution of the species, given a particular climate or land use change scenario. Such methods can be applied rapidly to a wide variety of species (Schweiger *et al.*, 2012). They have been used successfully to predict the risk of alien plant invasions (Thuiller *et al.*, 2005), to predict species distributions (Raxworthy *et al.*, 2003) and for simulating past distributional changes (Martínez-Meyer *et al.*, 2004). However, such ecological niche modelling has limitations (Pearson & Dawson, 2003). For example, dispersal is typically ignored (but see Engler & Guisan, 2009; Václavík & Meentemeyer, 2009), and it is generally implicitly assumed that the species' distribution is at equilibrium with the environment (but see Chuine, 2010; Elith *et al.*, 2010; Hulme, 2011). These shortcomings are clearly problematic when the focus is on a species whose range is shifting, e.g. due to climate change (Zurell *et al.*, 2009), or because it is invasive (Hulme, 2009). A key problem is that, even in well-mapped regions, the absence of a species from a given location may have more to do with propagule pressure and constraints on dispersal than with its inherent environmental suitability. These issues can be addressed by correcting for exposure, e.g. using spatio-temporal models that combine habitat suitability/niche components with dispersal (Catterall *et al.*, 2012; Pagel & Schurr, 2012).

Inference for stochastic spatio-temporal models (i.e. dynamic models that are both stochastic and spatial) is an area of ongoing development which forms a bridge between dynamic process models and (non-dynamic) spatial statistical approaches to species distribution modelling (Heikkinen *et al.*, 2006; Latimer *et al.*, 2006). The increase of capabilities in computational statistics in recent years, most notably the development of Markov chain Monte Carlo (MCMC) techniques (Gilks *et al.*, 1996; Gamerman & Lopes, 2006), has enabled statistical parameter estimation in an increasingly wide range of models. This includes continuous time models of infectious disease transmission (O'Neill, 2010) in particular spatially explicit transmission models (Gibson, 1997; Höhle *et al.*, 2005), metapopulation models (e.g. O'Hara *et al.*, 2002; Kuroe *et al.*, 2011) and models for the spread of invasive species (Cook *et al.*, 2007; Catterall *et al.*, 2012) that use environmental covariates to account for critical spatial heterogeneities in habitat suitability (Hastings *et al.*, 2005). Techniques for inference in discrete-time stochastic models (where model time is advanced in a series of fixed-sized steps) have also been applied to population dynamics in both non-spatial (Buckland *et al.*, 2004) and spatial (Wikle, 2003; Hooten *et al.*, 2007) contexts. In this paper the model state-space describes the species distribution at any given point in time (e.g. the abundance at each location modelled) and the state-space history is the full set of distributions at every time point within the period of interest. Thus, using stochastic spatio-temporal models as the basis for inference enables explicit modelling of population responses to environmental change over time. Our focus is on discrete state-spaces (i.e. binary or integer-valued for modelling presence/absence and abundance respectively), and therefore we do not consider stochastic diffusion type models.

A further advantage of stochastic spatio-temporal models is that their dynamic nature makes it relatively easy to incorporate biological understanding of the key processes that define the system under study, e.g. births, deaths, dispersal. In stochastic models, the trajectory (state-space history) varies from one realization to another, even when the parameters and initial conditions are fixed. In contrast, deterministic models constitute a special case for which there is only one possible trajectory for a given set of inputs. The structured randomness of the former is useful insofar as it reflects the variability seen in many biological systems, e.g. due to demographic stochasticity, even in controlled microcosms (e.g. Cook *et al.*, 2007). At larger spatial scales the stochastic nature of biological processes can also play a significant role; for example rare long-range dispersal can be a decisive factor when modelling invasions or species responses to climate change (e.g. Clark *et al.*, 2001). The ability to routinely include such process understanding is likely to enable more reliable extrapolation beyond the range of data used to parameterize such models, potentially critical when considering the impacts of severe climate change.

In the next section we outline a general framework for statistical inference in dynamic SDMs that may include deterministic or stochastic processes. A computationally demanding aspect of such schemes as applied to stochastic models is the need to infer, alongside the model parameters, the unobserved (often referred to as latent or hidden) state-space histories that are consistent with the available data, e.g. time-series of underlying species abundance from presence/absence records (O'Hara *et al.*, 2002). However, we emphasize the benefits of the stochastic formulation as a natural approach to account for spatio-temporal correlations.

We then demonstrate the potential of this approach using three case studies covering a range of typical data that may be available to fit SDMs: spatial presence/absence data at a single point in time; spatio-temporal presence/absence data (distribution maps at two or more points in time); and time series abundance data. These examples also make use of continuous- and discrete-time stochastic processes as well as deterministic differential equations. The

models are tailored to the data available (or assumed availability in the case of simulated data), and variously describe local presence/absence or abundance levels and the processes of local population dynamics, dispersal and establishment. Finally, the relative strengths and weaknesses of each approach are compared and near and long-term goals for methodological developments and future applications are discussed.

INFERENCE IN DYNAMIC SDMS

In this section we set out a generic framework for Bayesian inference in dynamic SDMs (see Fig. 1 for an overview).

Likelihood

Statistical inference via both maximum-likelihood and Bayesian statistics is based on one central quantity, the likelihood. The likelihood is defined as the probability of making observation D (the data) given a model structure and associated parameter values and boundary and initial conditions (see Hartig *et al.*, 2012). For state-space Markov process models it is usual to split the model definition into a process or latent model, and an observation model. The latent component describes the underlying biological assumptions and represents in some sense the true unobserved (i.e. hidden or latent) state of the system of interest. The observation model describes the probability of making the observations D for any underlying state of the system described by the latent model.

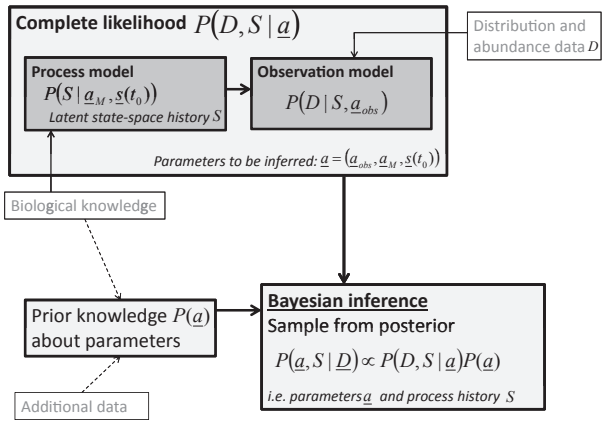

Figure 1 Overview of inference in dynamic stochastic models. Dark grey boxes denote model components, inference elements are light grey and biological data and information represented by white boxes. Understanding of the biological processes that determine species distributions enables specification of the stochastic process model, which defines the likelihood $P(S|\underline{a}_M, \underline{s}(t_0))$ of any complete state-space history S (i.e. the species distributions for all times of interest, which are unobserved in practice). The observation model $P(D|S, \underline{a}_{\text{obs}})$ describes the probability of observing the data D given any underlying history S. The product of these is the complete likelihood for the parameters $\underline{a} = (\underline{a}_{\text{obs}}, \underline{a}_M, \underline{s}(t_0))$, which include the parameters of the noise and process models ($\underline{a}_{\text{obs}}$ and $\underline{a}_M$ respectively), as well as the initial state $\underline{s}(t_0)$ (e.g. the initial species distribution). The parameter prior $P(\underline{a})$ encodes knowledge, sometimes obtained directly from alternative data sets, that constrains the range of parameter values. Bayesian inference is based on the posterior distribution of the unknowns ($\underline{a}$, S) given the data, and is proportional to the product of the prior and complete likelihood. In practice, techniques such as MCMC are used to draw samples from the posterior, and, for example, parameter statistics $P(\underline{a}|D)$ can be obtained by marginalizing over possible histories S (see text for details).

Process likelihood

In this paper we consider SDMs on a set of geo-referenced locations L, where variables $s_i(t)$ describe the presence/absence (binary) or abundance (e.g. integer) at each site $i \in L$ and time t. The vector of all such variables $\underline{s}(t)$ (throughout, an underscore will denote a vector) is known as the model *state-space*. The complete realization of the entire history of the process in the time interval from an initial time t_0 to some final time T is denoted by $S = \{\underline{s}(t) : t \in [t_0, T]\}$ and is the set of state-spaces representing the underlying state of the system at all times during the interval of interest $[t_0, T]$. This model should be a close representation of the actual dynamics of the species and the methods presented generalize to more complex models that account for additional factors such as different age classes or even multiple species at each location.

For any stochastic model, one can derive the process (or latent component) of the likelihood $P(S|\underline{a}_M, \underline{s}(t_0))$, i.e. the probability of the entire history of the process S given the model parameters $\underline{a}_M$ (e.g. demographic rates) and the initial state of the system $\underline{s}(t_0)$. The formulation of the likelihood follows naturally from the definition of the model. However, in some cases calculation of the process likelihood can be computationally expensive and one has to resort to stochastic simulation where a number of possible paths through the state space are simulated and used to approximate the likelihood (see Hartig *et al.*, 2011 for an overview).

Often, one can formulate dynamic models such that they are Markovian, i.e. the next change in state only depends on the current state and not on any previous states. This allows the process likelihood to be factorized into separate probabilities for the change from one time point to the next:

$$P(S|\underline{a}_M, \underline{s}(t_0)) = \prod_{k=1}^{N} P(\underline{s}(t_k)|\underline{a}_M, \underline{s}(t_{k-1})). \qquad (1)$$

The individual terms in the above product are the transition probabilities between the states of the system at time t_{k-1} and t_k. If the model is a discrete time model, it is defined from the start in terms of these transition probabilities between consecutive times t_{k-1} and t_k. Below we explore specific exemplar models of the dynamic processes that determine species distributions, and the corresponding calculation of the likelihood in both discrete- and continuous-time stochastic processes. In addition equation (1) can be

applied to static models (e.g. see below), where S is simply the state of the system at one point in time for which the species is assumed to be in equilibrium (for stochastic models this is typically taken to be the quasi-equilibrium when the true equilibrium is extinction).

Observation model

For static models and both deterministic and stochastic dynamic models the actual states are never perfectly known from the data. Like all observations, monitoring data are imprecise and subject to sampling variations and measurement errors (Kéry & Schmid, 2006). Whilst the states modelled by the process model are not independent (the dynamic model itself implies the relationship), the observation errors are typically assumed to be independent. Observation errors are described by means of an observation model, $P(D|S, \underline{a}_{\text{obs}})$, which is parameterized in terms of some additional parameters $\underline{a}_{\text{obs}}$. It describes the probability of obtaining the data D, given the underlying state of the system during the same period S. While observation models can be developed for various types of species distribution data (Royle & Dorazio, 2008), the precise specification of the observation model also depends on the process model considered and in turn the observation model may influence the form of the process model used.

The complete likelihood

Combining the observation model and the process likelihood gives

$$P(D, S|\underline{a}) = P(D|S, \underline{a}_{\text{obs}})P(S|\underline{a}_M, \underline{s}(t_0)), \tag{2}$$

which is known as the complete likelihood or complete-data likelihood. In the above equation the initial conditions, the parameters from the underlying model $\underline{a}_M$, and those in the observation model $\underline{a}_{\text{obs}}$, are combined into a single vector of parameters to be estimated, $\underline{a}$. If the underlying model is deterministic then there is only one state-space trajectory corresponding to the deterministic solution which starts at $\underline{s}(t_0)$. In other words, the entire history S can be determined without uncertainty from $\underline{a}_M$ and $\underline{s}(t_0)$ so that the likelihood simplifies to

$$P(D|\underline{a}) = P\big(D|S^{\text{det}}(\underline{a}_M, \underline{s}(t_0)), \underline{a}_{\text{obs}}\big). \tag{3}$$

Bayesian inference

Here, we describe the Bayesian approach to inferring the parameters $\underline{a}$ and the history S given the data D and discuss its implementation using the computational method MCMC.

Priors

In addition to the likelihood, Bayesian analysis requires the definition of the *prior* $P(\underline{a})$ describing any external sources of information on parameters or initial conditions additional to the primary data set, D. This allows existing knowledge on process parameters to be incorporated into the analysis. In practice, priors are typically chosen to be independent across the components of $\underline{a}$ and reflect knowledge about the range of values different parameters can take, or to reflect ignorance by giving support to a wide range of parameter values. In some cases certain parameters or initial conditions can be dropped from the vector $\underline{a}$ if they are assumed to be known with great precision, although if there is actually uncertainty in the value of the parameter, doing so will lead to conclusions that are unrealistically precise. Kass & Wasserman (1996) provide a review of more formal approaches to the selection of Bayesian priors, but see also Hartig *et al.* (2012).

The posterior

Applying Bayes' theorem, which simply tells us how different conditional probabilities are related, we obtain the so-called *posterior distribution* of the unknowns, namely the model parameters and the set of underlying states of the system throughout the period of interest, conditional on what we know, i.e. the data, our assumptions regarding the model structure, and any sources of information manifested in the prior:

$$P(\underline{a}, S|\underline{D}) = \frac{P(D, S|\underline{a})P(\underline{a})}{P(D)}. \tag{4}$$

All Bayesian inference (see e.g. Lee, 2004) is based on the posterior distribution, which combines the prior information with that obtained from the data via the likelihood. Two important points to note are that unknown parameter values and the unobserved states are both considered to be subject to inference and that the greater the information content of the data the less influence the prior assumptions have. As we saw above, when the underlying model is deterministic the likelihood simplifies to $P(D|\underline{a})$, and the latent variables (i.e. the state-space history S) also 'drop out' of the posterior, which simplifies to:

$$P(\underline{a}|\underline{D}) = \frac{P(D|\underline{a})P(\underline{a})}{P(D)}. \tag{5}$$

For most models of interest, including deterministic ones, modern computational tools such as MCMC are required to explore the posterior (Box 1).

EXAMPLE APPLICATIONS

In what follows, we describe three different formulations of SDMs and show how the framework described above can be used to infer parameters from both real and simulated data sets. The examples chosen illustrate a range of data types and the use of contrasting classes of model, e.g. deterministic or stochastic, discrete- or continuous-time.

BOX 1: SAMPLING FROM THE POSTERIOR: COMPUTATIONAL ISSUES

Markov chain Monte Carlo (MCMC) techniques, allow samples of the parameters $\underline{a}$ and state-space histories S to be drawn directly from the posterior (which is proportional to the likelihood and the prior, $P(\underline{a}, S|\underline{D}) = P(D, S|\underline{a})P(\underline{a})/P(D)$ – see main text for details) without having to calculate the normalization constant $P(D)$, which can be extremely computationally demanding for many models of interest. The procedure generates successive values of $\underline{a}$ and S using a Markov chain designed so that its equilibrium distribution is the posterior. Typically, initial samples must be discarded until the Markov chain has reached this dynamic equilibrium, so that, only after a 'burn-in' phase are the generated values genuine samples from the posterior. Estimates of the posterior improve as the number of samples generated from the Markov chain increases, and usually many thousands are required. Thinning of samples is sometimes applied to reduce computational requirements, but this reduces the precision of estimates obtained from the posterior (Link & Eaton, 2012). Although any suitable MCMC algorithm will have the desired equilibrium distribution, getting the algorithm to work efficiently can sometimes take considerable skill. Unfortunately, it is not possible in general to rigorously determine when the burn-in phase is complete, but in practice this is not usually a major drawback and, for example, Cowles & Carlin (1996) provide a set of heuristics to assess convergence. Here, we focus on the standard MCMC methodology (see e.g. Andrieu *et al.*, 2003) but there are also a range of alternative approaches, for example rejection sampling and particle filters (e.g. Arulampalam *et al.*, 2002), which are becoming more widely used.

Once samples have been generated from the posterior $P(\underline{a}, S|\underline{D})$, e.g. as described above, subsequent analyses are usually very simple because they involve little more than calculating summary statistics of transformations of the MCMC output. For example, the marginal distribution of parameters $P(\underline{a}|D)$ may be estimated by the histogram of the sampled parameter values. The marginal distribution of any single component of $\underline{a}$, or the joint distribution of two or more, may be obtained in a similar fashion. From such distributions it is then possible to obtain various measures such as means and credible intervals. More complex properties of the model can similarly be estimated by drawing repeated samples from the posterior and then calculating the property of interest by running the model (or submodel) for each sample. Statistics of interest include maps of underlying habitat suitability or niche distribution, or projections of species' future distributions. Inference of the state-space histories also allows for probabilistic reconstruction of unobserved past states, e.g. historical distributions (see Fig. 3). Future projections are said to be sampled from the *predictive posterior distribution*, whereas estimates of model parameters and inferred system states, or simple functions of them, are derived directly from the posterior distribution.

Presence/absence data from a single point in time

In many applications, presence/absence data are available only at a single point in time. We illustrate the treatment of such data using a simulated data study based on the (static) equilibrium solution of a deterministic process-model.

A deterministic equilibrium model for species abundance

Consider a simple single-species demographic model where all intra-specific effects are accounted for by spatial variation in the carrying capacity k_i and the dynamics of local abundance $s_i(t)$ in each grid cell $i \in L$ is described by deterministic logistic growth. The spatially heterogeneous carrying capacity is a function of local environmental conditions

$$k_i = k_{\max} \prod_c \exp\left(-\left[\frac{x_{c,i} - \mu_c}{\tau_c} \right]^2 \right). \quad (6)$$

In this formulation the niche is a Gaussian function with species' optimum μ_c and tolerance τ_c for the environmental variable c, which has value $x_{c,i}$ in cell i. When all such environmental conditions are optimal the carrying capacity is $k_{\max}$. Thus, the maximum potential carrying capacity is reduced by the extent to which the local environment does not match the ideal conditions for the focal species. If this model is run to an equilibrium, $s_i(t)$ will converge to the carrying capacity, k_i. Hence, if we assume this model is reasonable and the species is in equilibrium, we can model the observed distribution in terms of the effect that environmental covariates have on the carrying capacity, k_i. For illustration we assume that just two environmental variables, $x_{1,i}$ and $x_{2,i}$, affect the carrying capacity of the species and thus the model is parameterized by five niche parameters, i.e. $\underline{a}_M = (k_{\max}, \mu_1, \mu_2, \tau_1, \tau_2)$.

Observation model

To link this abundance model to presence/absence data we assume a simple form for the probability of detection

$$\psi_i(t) = 1 - \exp(-d_i s_i(t)), \quad (7)$$

which increases with local abundance $s_i(t)$ and the detection parameters d_i which can be adjusted, e.g. to reflect cryptic-ness of taxa, or heterogeneous observer effort. The data are the observed presences–absences, i.e. for site i we use $\pi_i^t = 1(0)$ to represent an observed presence (absence) assumed to be observed at some set of sites, Θ, when the system is in equilibrium. The likelihood of this set of observational data D given the true states S, i.e. the modelled abundances $s_i(t)$ at each site is

$$P(D|S, \underline{a}_{\text{obs}}) = \prod_{\Theta} \text{Bernoulli}\left(\pi_i^t | \psi_i(t)\right). \quad (8)$$

Note that the Bernoulli distribution is simply $\psi_i(t)$ where a presence is observed and $1-\psi_i(t)$ otherwise and the observation model parameters are $\underline{a}_{\text{obs}} = (d_1, d_2, \ldots, d_{|L|})$ where $|L|$ represents the number of sites modelled.

The likelihood and posterior

The deterministic nature of the model means that the likelihood simplifies to equation (3) with $S^{\text{det}}(\underline{a}_M, \underline{s}(t_0))$ determined by assuming the system has reached equilibrium at each site as given by equation (6). The posterior is then given by equation (5). We note that this model conveniently reduces to the standard statistical form of a generalized linear model (McCullagh & Nelder, 1989) with a complementary log–log link function that transforms ψ_i into a quadratic function of the parameters $\underline{a} = (k_{\max}, \mu_1, \mu_2, \tau_1, \tau_2, d_1, d_2, \ldots, d_L)$. A more complex function (e.g. generalized additive models, GAMs; see e.g. Guisan *et al.*, 2002) could be used to represent the (unknown) response of the carrying capacity to the environment.

Simulated data and inference

In order to simulate data, the model is assumed to be in equilibrium with abundances in site i given by equation (6), $s_i(t) = k_i$, and observed presences sampled from each site with probability ψ_i in equation (7) (see also software provided in Appendix S1 of the Supporting information). This generates presence/absence data at a single point in time, which we use to infer the niche parameters. In the figures presented here we also assume known spatially homogeneous detection rates $d_i = 0.03$. Weakly informative, uniform and independent priors for the five remaining parameters to be inferred are chosen to be $k_{\max} \sim U[0.0001, 800]$; $\mu_1, \mu_2 \sim U[-100, 200]$; $\tau_1, \tau_2 \sim U[1, 200]$. Samples are drawn from the corresponding posterior using a freely available software package FILZBACH, that implements Metropolis Hastings MCMC within a user friendly set of *C* libraries (see Appendix S1; further examples found in Purves, 2009; McInerny & Purves, 2011; Williams & Purves, 2011). This is not a computationally demanding analysis and can be carried out on a standard PC. The posterior mean and credible intervals (i.e. point and interval estimates) are calculated using every 100th parameter set from 200,000 iterations following a generous 200,000 sample burn-in. The results show that the presence/absence data allow reliable inference of the underlying abundances (Fig. 2).

Having estimates of sampling effort are of course crucial and the example shown here can be easily extended to allow for spatially heterogeneous sampling effort. Readers can implement this whole analysis by running the code supplied in Appendix S1, and may then adjust the code to implement the case of heterogeneous sampling for themselves (also see McInerny & Purves, 2011). Without taking into account sampling effort, we would underestimate the maximum response of a species at its optimum, overestimate its tolerance for that variable and introduce large uncertainty into

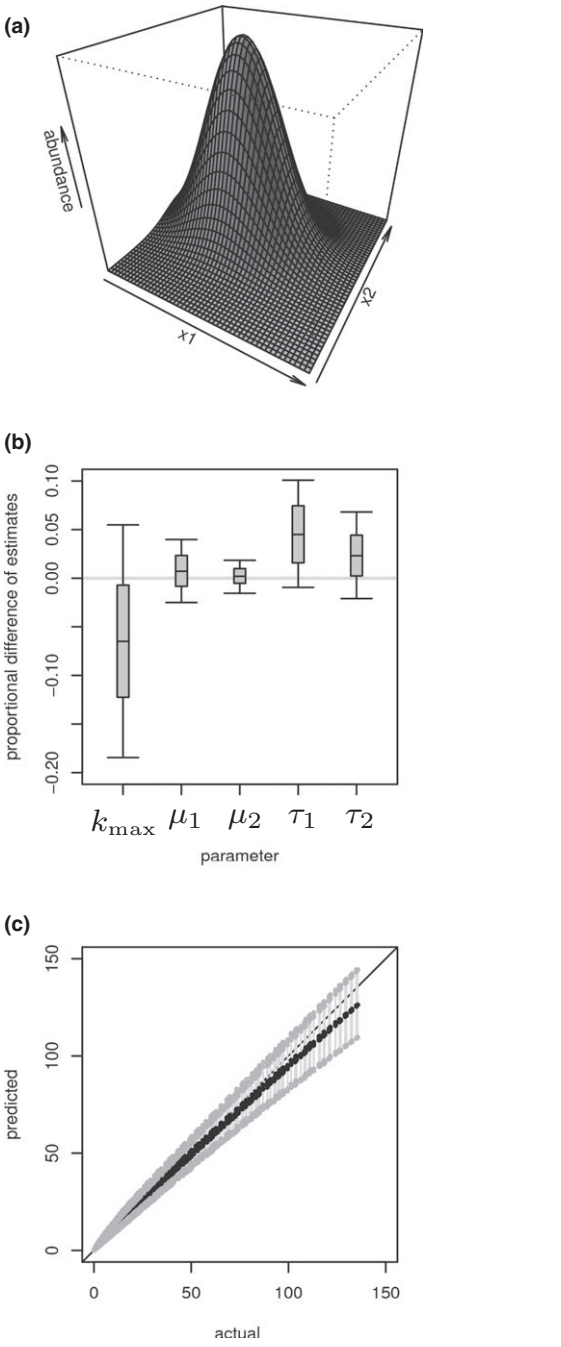

Figure 2 (a) The abundance model (6) viewed as a surface across two environmental variables, $x1$ and $x2$. Simulated data are generated from this underlying abundance distribution by applying equation (7) to randomly assign presence/absence to each location and Bayesian inference is then applied to this data (see text for details). (b) The reliability of this parameter inference is shown by 95% (error bars) and 67% (grey box) credibility intervals, and the posterior mean, of the proportional difference in parameter estimates relative to the actual value, i.e. [(predicted − actual)/actual]. The plot shows these quantities for each of the inferred parameters $k_{\max}$, μ_1, μ_2, τ_1, τ_2 and in all cases the 95% credibility interval includes the actual parameter value. (c) Plotting actual versus predicted values shows that robust inference of abundance is possible in this scenario even from presence/absence data.

the parameter estimates (e.g. Frost & Thompson, 2000; McInerny & Purves, 2011). Detection probabilities (i.e. d_i) may be derived from a variety of sources – survey records, analysis of survey data or inferred from species discovery, site accessibility and, of course, expert knowledge.

Presence/absence data at two or more points in time

The previous application is very similar to current methods for estimating SDMs: it is (in essence) a regression of the presence against static covariates. In the next example we show how temporal information can be incorporated into dynamic SDMs. The inclusion of temporal information is desirable because it offers information that is complementary to spatial distribution data with respect to various aspects of a species' dynamics, and the reaction of a species to climate change and other environmental change will depend on such dynamics (Schurr *et al.*, 2012). Thus this second example, shows how a spatio-temporal model can be fitted to spatio-temporal data using the methods developed by Catterall *et al.* (2012), and their previously unpublished application to data on the spread of the invasive plant species *Rhododendron ponticum* L. in Great Britain.

Species distribution data

We use distribution maps for *R. ponticum* (see Fig. 3a) from the *New Atlas of the British and Irish Flora* (Preston *et al.*, 2002), which indicate that the occurrence pattern is not static over time and therefore that range dynamics should be accounted for. This non-stationarity makes application of standard SDM approaches problematic and, for example, the in an analysis of riparian invasive species Collingham *et al.* (2000) omit data from regions where species 'probably only occupy a fraction of all suitable squares'. Here, we use all available data and take the distribution given by the 1970 map as our initial data and then assume additional presences observed in the 2000 map were colonized during the interval (1970, 2000). Note that although there is a map for 1987, it is generally considered to be of lower quality (Preston *et al.*, 2002) and so is not used for fitting purposes. We also assume that there are no errors in the observed 1970 and 2000 distributions, which means that the observation model assigns zero probability to any state-space history that does not coincide exactly with the data, enabling computationally efficient likelihood calculations to be used (for details see Catterall *et al.*, 2012).

Modelling the spread of invasive vascular plants

We model $s_i(t)$, the presence $(s_i(t) = 1)$ or absence $(s_i(t) = 0)$ of the focal species at each of the sites i at time t. Each location is further characterized in terms of several environmental covariates $\underline{x}_i$. The model is a continuous time Markov process in which the colonization of each site j by propagules from already colonized sites occurs at rate

$$c_j(\underline{s}(t)) = f\left(\underline{x}_j, \underline{\beta}\right) \times \left(1 - s_j(t)\right) \times \sum_{i:d_{ij} \le d_{\max}} \frac{\phi(d_{ij}, \lambda)}{m(\lambda)} s_i(t). \quad (9)$$

The right-hand third term includes a summation over all possible donor sites i within a distance $d_{\max} = 150$ km, and represents the connectivity of site j expressed in terms of the dispersal kernel $\phi(d, \lambda) = d^{-2\lambda}$, and the normalization factor $m(\lambda) = \sum_{k:d_{jk} \le d_{\max}} \phi(d_{jk}, \lambda)$, where d_{jk} is the distance between sites j and k. The middle term simply ensures that only uncolonized sites can be colonized, whilst the left-most term represents landscape heterogeneity in terms of the relative suitability for colonization,

$$f\left(\underline{x}_i, \underline{\beta}\right) = \exp(\alpha A_i) \exp\left(\tau(T_i - \overline{T})\right) \sum_{k=0}^{9} b_k H_{i,k}. \quad (10)$$

As detailed in Catterall *et al.* (2012), the following environmental covariates for each of 2838 standard Ordnance Survey hectads (10 km × 10 km grid squares) in Great Britain are used: the proportion (by area) covered by each of 10 land-cover types $\underline{H}_i$; the mean annual temperature over the period 1920–1999 T_i (with $\overline{T}$ the average temperature across all hectads); and the mean elevation in each hectad A_i. Thus, $\underline{\beta} = (\alpha, \tau, b_0, b_1, \ldots, b_9)$, where α controls the effect of elevation, τ that of temperature, and $b_k \in [0, \infty)$ quantifies the suitability of habitat class k. Note, we assume $b_0 = 0$ as habitat type zero is sea. Thus, we wish to estimate both the dispersal and the niche parameters $\underline{a}_M = (\lambda, \underline{\beta})$.

Inference of colonizability, historical distributions and future spread

The inference framework described above is applied to this model using the *R. ponticum* and covariate data described earlier. The Markovian nature of the model leads to exponentially distributed times between colonization events (Cox & Miller, 1965). The upshot of this is that for any given set of ordered colonization times $\{t_k : k = 1, \ldots, N\}$, where site $j(k)$ is colonized at time t_k, the components of the factorization shown in equation (1) are easily calculated, up to a constant of proportionality, to be

$$P(\underline{s}(t_k)|\underline{a}_M, \underline{s}(t_{k-1})) \propto c_j(s(t_{k-1}))\ \exp(-C(\underline{a}, s(t_{k-1}))[t_k - t_{k-1}]), \quad (11)$$

where $C(\underline{a}, \underline{s}(t)) = \sum_{j=1}^{L} c_j(\underline{s}(t))$ is the total colonization rate across all sites. In this case the missing state space information which is inferred alongside the parameter values, are the colonization times of sites observed to be invaded by *R. ponticum*. This is achieved in an MCMC context by proposing changes to the parameters and colonization times and accepting/rejecting them using a Metropolis–Hastings algorithm (see Catterall *et al.*, 2012).

The estimates of the habitat suitability parameters thus obtained indicate that this invader strongly favours broadleaf forest and avoids both urban and arable habitats. Calculation of the mean posterior suitability for each hectad shows the high colonizability of southern and western areas of Britain,

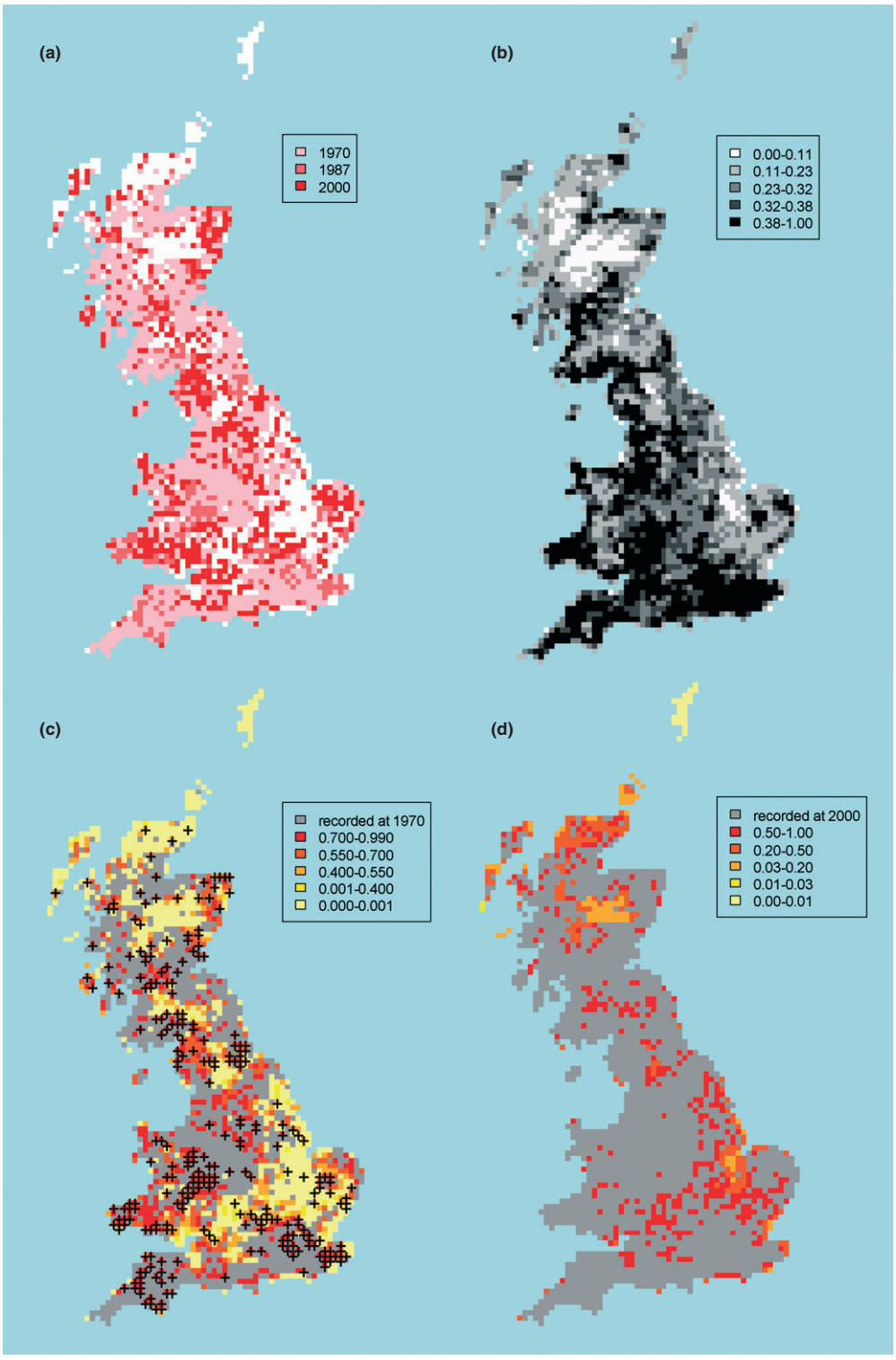

Figure 3 (a) Records of *Rhododendron ponticum* across hectads (10 km ×10 km squares) in Great Britain corresponding to recording periods (1970, 1987 and 2000) used in Preston *et al.* (2002). The records for 1970 show all hectads in which *R. ponticum* was observed up to 1970 whilst the locations marked as 1987 and 2000 show all new records since 1970 and 1987, respectively. (b) Suitability of each hectad for colonization by *R. ponticum* under the spatio-temporal colonization–dispersal model based on estimated habitat suitability and the response to mean annual temperature (°C) averaged over the period 1920–1999 and elevation. The value calculated for each hectad is the mean of the posterior distribution of suitability. (c) Historical prediction. Probability of colonization with *R. ponticum* by 1987, with small crosses indicating populations which were recorded in the 1987 distribution map but not the 1970 distribution map. (d) Long-term predictions. Probability of colonization with *R. ponticum* by 2030, based on 10,000 simulations and accounting for both variability in the modelled process of invasive spread and uncertainty in the parameters estimated from the observed data.

which reflects the fact that most of Britain's arable land is in the east of the country (Fig. 3b). Historical prediction of the *R. ponticum* distribution in 1987 provides evidence for the validity of the modelling, as there is a high level of agreement between the predicted and observed colonizations (Fig. 3c). Prediction of the distribution of *R. ponticum* in 2030 suggests that most of Britain is at risk of invasion, with the exception of some Scottish upland areas and a few outlying islands (Fig. 3d). The inclusion of land-use and temperature covariates means that the model described here could be used to produce similar risk maps under future land use and climatic scenarios.

Catterall *et al.* (2012) apply the methods used here to British floristic atlas data for *Heracleum mantegazzianum* (giant hogweed) assessed at a 10 km × 10 km resolution in 1970 and 2000. Use of simulated data to validate the inference algorithm suggests that such data are sufficient to infer model parameters and that increasing the number of time points does not significantly improve reliability. Using data at 1970 and 2000 they are able to predict future species distributions accurately under simulated scenarios and also, from real data, the results of a limited field survey conducted in 2004.

Combining data types collected at contrasting spatio-temporal scales

The previous examples have dealt with single data types, but this example demonstrates how different sorts of data can be combined. It follows Pagel & Schurr (2012) who present a dynamic range model (DRM) framework to estimate discrete-time stochastic models of spatial population dynamics in variable environments from presence/absence maps and local abundance time-series at a more restricted set of locations.

A stochastic model of local abundance and dispersal

The model describes the dynamics of abundance $s_i(t_k)$ at each site i at discrete times t_k by combining a stochastic model of local population dynamics with a dispersal kernel for long-distance dispersal (LDD). Dispersal precedes reproduction, so at time step t_k the post-dispersal population distribution is

$$\tilde{s}_i(t_k) = \sum_j P_{j\to i}(f_{\mathrm{LDD}}, \alpha)\, s_j(t_{k-1}), \tag{12}$$

where the probability of an individual moving from the cell of origin j to cell i, $P_{j\to i}(f_{\mathrm{LDD}}, \alpha)$ assumes that a fraction $(1 - f_{\mathrm{LDD}})$ does not disperse beyond cell borders, with the remaining proportion f_{LDD} dispersing a distance determined by a negative exponential distribution with mean distance α. Subsequent reproduction is described by the Ricker model (Ricker, 1954) for density-dependent growth. The deterministic Ricker model gives the expected mean population size at time t_k as $\tilde{s}_i(t_{k-1})\exp(r_i(t_{k-1}) - h\tilde{s}_i(t_{k-1}))$, where $r_i(t_k)$ is the intrinsic growth rate and h the competition intensity. The carrying capacity is thereby implicitly defined as $r_i(t_{k-1})/h$. The competition intensity h is considered to be a constant species property, whereas the intrinsic growth rate $r_i(t_k)$ varies across space and time in dependence on environmental conditions. It is assumed that the local environment variables described by the vector $\underline{x}_i$ affect the growth rate $r_i(t_k)$ linearly as described by the niche parameter vector $\underline{\beta}$, but there is additional unexplained variation σ^2_r, so that the growth rate is described by a normal regression model:

$$r_i(t_k) \sim N\left(\underline{\beta}^T \underline{x}_i(t_k), \sigma_r^2\right). \tag{13}$$

The resulting spatio-temporal variation in the intrinsic growth rate is the main driver of range dynamics and represents Hutchinson's realized niche (when and where the growth rate is positive, i.e. $r_i(t) > 0$).

The deterministic Ricker model outlined above is furthermore extended by a Poisson error term describing demographic stochasticity and a log-normal error term with variance σ_p^2 that accounts for possible uncertainty in the specification of population dynamics. Based on the resulting stochastic population dynamics models, the transition probabilities between abundance states $s_i(t_k)$ at discrete times take the form of a Poisson lognormal distribution:

$$P(\underline{s}(t_k)|\underline{a}_M, \underline{s}(t_{k-1})) = \prod_{i=1}^{L} \mathrm{PoissonLN}\left(s_i(t_k)\middle|\eta_i(t_{k-1}), \sqrt{\sigma_p^2 + \sigma_r^2}\right), \tag{14}$$

where the product is over all cells $i \in L$. The scale parameter $\sqrt{\sigma_p^2 + \sigma_r^2}$ combines uncertainties of both the niche model and the population dynamics model. The location parameters $\eta_i(t_{k-1})$ are the mean local population sizes on a log scale, which result from the combination of the Ricker model with the niche model and are given by

$$\eta_i(t_{k-1}) = \ln(\tilde{s}_i(t_{k-1})) + \underline{\beta}^T \underline{x}_i(t_{k-1}) - h\tilde{s}_i(t_{k-1}). \tag{15}$$

Thus, the full parameter vector $\underline{a}_M = \left(f_{\mathrm{LDD}}, \alpha, \underline{\beta}, h, \sqrt{\sigma_p^2 + \sigma_r^2}\right)$ comprises dispersal parameters $(f_{\mathrm{LDD}}, \alpha)$, niche parameters $\underline{\beta}$, the competition strength h, and the combined model uncertainty $\sqrt{\sigma_p^2 + \sigma_r^2}$. Recalling that the complete state-space history in the period of interest is $S = \{\underline{s}(t) : t \in [t_0, T]\}$ then the above equation provides the factorization required by equation (1) for $P(S|\underline{a}_M, \underline{s}(t_0))$ and therefore forms the basis of inference in this model.

Observation model

As before, in addition to the probabilistic process model, we define an observation model that links latent states (abundances) to monitoring data. The extension here is to consider two different types of data: presence/absence data and abundance count data. For the presence/absence part of the data, the model is similar to the first example. The data are presence/absence records π_i^t for site i at time t. It is assumed that there is a constant detection probability p_{pa} and a presence is recorded

if at least one individual is found, i.e. with probability $\Pr(\pi_i^t = 1) = \psi_i(t) = 1 - (1 - p_{pa})^{s_i(t)}$. The contribution to the likelihood from each data point is Bernoulli with parameter $\psi_i(t)$ [see equation (8)]. The model for the observed abundance n_i^t at location i and time t is similar, in that a constant probability, now p_{ab}, of sampling each individual is assumed. But the total number of individuals is counted, so this is assumed to follow a binomial distribution, $n_i^t \sim Bin(s_i(t), p_{\mathrm{ab}})$. The total likelihood of the data given the population states is then the product of the likelihoods from the two sources of data:

$$P(D|S, \underline{a}_{\mathrm{obs}}) = \prod_{\Theta} \mathrm{Binomial}\left(n_i^t | s_i(t), p_{\mathrm{ab}}\right) \times \prod_{\Omega} \mathrm{Bernoulli}\left(\pi_i^t | \psi_\tau(t)\right), \qquad (16)$$

where Θ and Ω are subsets of sites and times for which presence/absence or abundance data are available, respectively. Hence, it is not assumed that both data exist for all sites at all times. For instance, if presence/absence maps are combined with local abundance time-series, the presence/absence data set Θ typically comprises most sites at few different times, whereas the abundance data set Ω comprises few sites for a number of consecutive times. Such data are sufficient to infer model parameters because the model assumes that: (1) probability of observing a species presence is related to local abundance, and (2) population dynamics respond consistently to environmental variables between sites.

Bayesian inference

As described above the product of the observer model and the process likelihood forms the complete likelihood $P(D, S|\underline{a}) = P(D|S, \underline{a}_{\mathrm{obs}})P(S|\underline{a}_M, \underline{s}(t_0))$ for the observer and process model parameters $\underline{a} = (\underline{a}_M, \underline{a}_{\mathrm{obs}})$. When combined with the priors [see equation (4)] this yields the posterior distribution of the unknowns (in this case the parameter values and complete population dynamics – changes in population at each location and time step) given the knowns, i.e. the abundance time-series and presence/absence data. Details of how samples from this joint posterior distribution $P(\underline{a}, S|D)$ can be generated by an MCMC algorithm are described in detail in Appendix S1 of Pagel & Schurr (2012).

Model assessment in a virtual study

The model outlined above was tested by Pagel & Schurr (2012) in a range of virtual case studies. To this end, range dynamics of a set of hypothetical species were simulated in an artificial landscape (40 cells × 20 cells) exposed to environmental change. From the simulated abundance distributions, presence/absence and count data were then sampled probabilistically (using a 'Virtual Ecologist' approach; Zurell *et al.*, 2010) to generate two presence/absence maps in a 10-year time interval and 30 local abundance time series of annual count data for the same 10 years. Based on these data, the posterior $P(\underline{a}, S|D)$ was sampled using MCMC for two independent chains (using every 50th sample from the second half of 100,000 iterations from each Markov chain). Note that this analysis used moderately informative priors for dispersal rates and detection probabilities, and non-informative priors for all other parameters. Forecasts of range dynamics under ongoing environmental change were then generated by stochastic forward simulations of the model using samples from the joint posterior of parameters and population states $\underline{s}(t = t_{\max})$ for the last year of the observation interval (Fig. 4a). The assessment of model predictions demonstrates that the obtained posterior (i.e. the estimation of the niche, dispersal capabilities and local population dynamics) enabled accurate predictions of future ranges even if a species suffers from severe dispersal limitation (Fig. 4b), whereas a static SDM would fail to predict the resulting mismatch between potential and realized range (see Pagel & Schurr, 2012). Because model states of the DRM describe abundances, the model provides not only predictions of future ranges but also of global and local abundance dynamics (Fig. 4c). The prediction of future model states accumulates not only uncertainty in parameter estimates but also uncertainty in estimated model states $\underline{s}(t = t_{\max})$, model uncertainty and inherent stochasticity of the population dynamics. Hence, predictive uncertainty will typically be large compared to range predictions and we suppose that useful abundance forecasting will therefore require both models that describe the actual population dynamics well and data that enable an accurate estimation of local abundances at the end of the observation period.

DISCUSSION

The response of species' distributions to climatic and other environmental change is governed by mechanisms that unfold over time, so that SDMs should be both dynamic and process-orientated. Ignoring such dynamics may lead to biased inference of niche parameters, e.g. absence of a species from a given location may be explained more by dispersal than by the location's inherent suitability. In this paper we have outlined a statistical framework for the fitting of dynamic process-orientated models to data.

This enables inference for models that reflect the structure of systems under study which in turn facilitates the use of multiple sources of data; information on different aspects of the system can be used in model fitting by including suitable components in the process model. For example, the availability of presence/absence maps may suggest modelling the process in terms of presence/absence, whereas abundance time-series data, even at a limited number of locations, can be used if the model is extended to model local population dynamics. It may also be possible to make use of phylogenetic data by incorporating measures of relatedness between populations into models. Moreover, information on model parameters obtained from direct measurement or from analysis of alternative data sets can be incorporated within the Bayesian

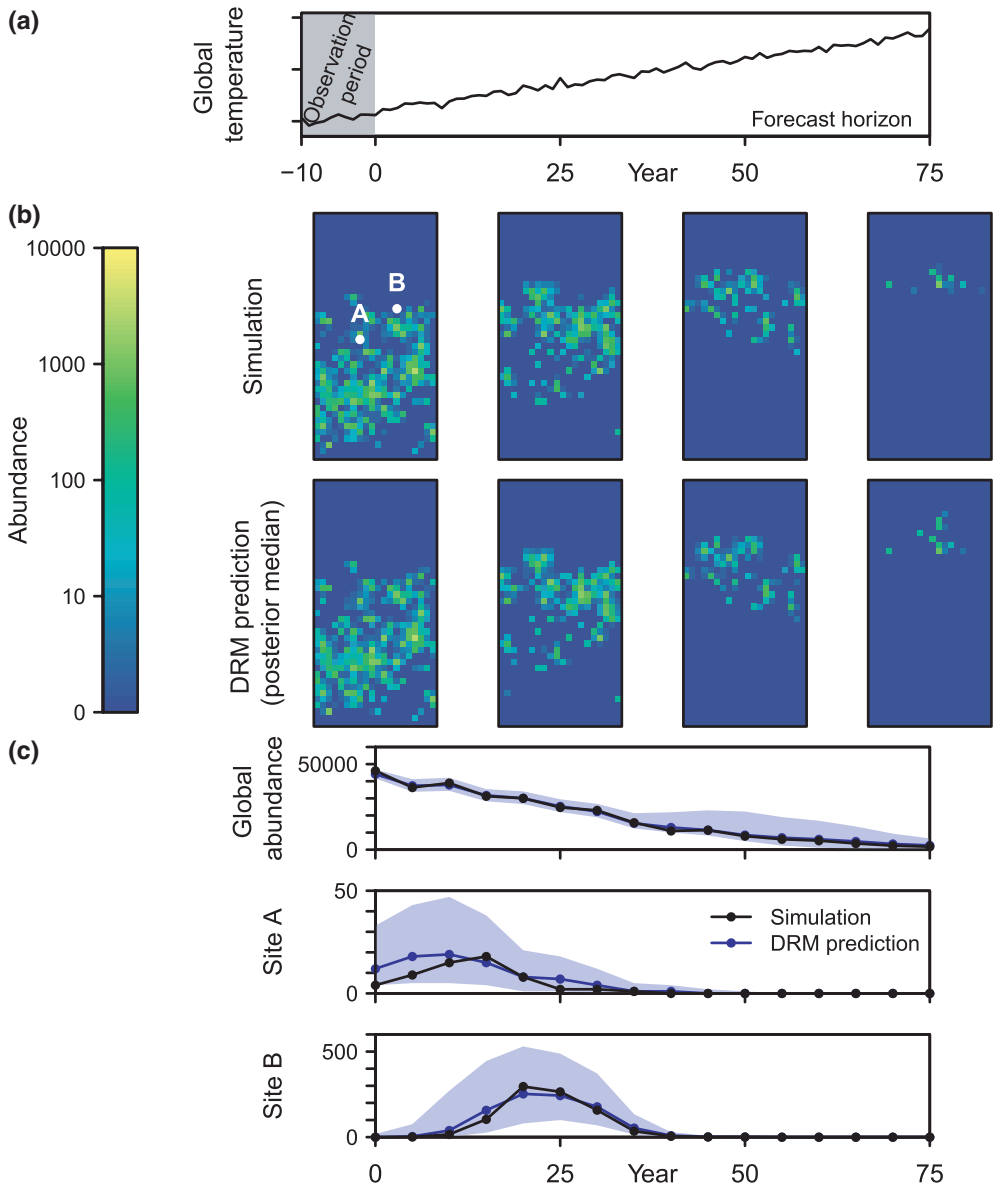

Figure 4 Range forecasts in a virtual case study to assess the dynamic range model (DRM) framework (the results shown here are derived from simulations described in Pagel & Schurr, 2012). (a) Virtual data are sampled in a 10-year observation period and the thereof estimated model is used for predictions within another period of 75 years (forecast horizon) that exhibits continuous increase in global mean temperature. (b) Simulated 'true' spatio-temporal abundance dynamics compared to model predictions of the estimated DRM. (c) Assessment of predictions of global abundance and local abundance for two sites one of which (A) is initially in the core range whilst the other (B) is initially in the periphery. Note that in the long term both locations fall outside the projected range. For DRM predictions, lines show the median and shaded areas the 95% credibility interval of predictive posteriors.

framework by informing the choice of prior distributions. The influence of this prior information diminishes as the information contained in the data used to fit the model increases and this represents a continuum between forward parameterization and full calibration of process-based models discussed in Dormann *et al.* (2012). However, it should be noted that within grid cell heterogeneity means that inferences drawn at one scale are not directly comparable with those from another scale (see e.g. Collingham *et al.*, 2000), although it may be possible to correct for some of these effects (McInerny & Purves, 2011). Such scale effects will also affect the relationship between model parameters and small-scale demographic measurements (Schurr *et al.*, 2012).

Although the framework presented in this paper is quite general there is still much work to be done in making it more routinely applicable to modelling species distributions. As the complexity of such models increases so does the computational cost of implementing inference and this may require a switch from the explicit likelihood approach adopted in this paper to likelihood-free approaches (Marjoram *et al.*, 2003), approximate Bayesian computation (see e.g. Hartig *et al.*, 2011) or particle MCMC methods (Andrieu

et al., 2010). Further developments are also needed in methods to allow the statistical comparison between models (Johnson & Omland, 2004), particularly when latent states must be inferred (Celeux *et al.*, 2006).

For broad-scale species distribution data, spatial and temporal variations in recording effort are inevitable (Aikio *et al.*, 2010) and lead to heterogeneities in rates of false absences. The framework outlined here accounts for non-detection via definition of the observation model, but often such rates are assumed to be homogeneous in space and time. Although, the results described in Catterall *et al.* (2012) were found to be rather robust to heterogeneities in non-detection, inference can be made more robust by allowing for spatially varying non-detection. Bierman *et al.* (2010) parameterize non-detection rates using expert knowledge on recording activity in a purely spatial SDM, and it would be interesting to extend their approach to the dynamic models described here. We also suggest there may be cases where we could estimate sampling intensities using latent variable methods, similar to McInerny & Purves (2011), by using multiple species.

To improve the ability of dynamic SDMs to represent the ecological niche we see the need for improvements in the following areas: (1) improved mechanistic models for the relationship between the environment and the species' presence (e.g. Higgins *et al.*, 2012) such as those based on plant phenology (Chuine, 2010; Hulme, 2011); (2) the further development of demographic models (Schurr *et al.*, 2012); and (3) attempts to include multispecies interactions (Kissling *et al.*, 2012) and modulation of the environment, e.g. by ecosystem engineers (Linder *et al.*, 2012).

Finally, we emphasize that advances in methodology such as those described here are only part of the solution to understanding the ecological niche and making more reliable predictions of range shifts under environmental change. The successful and widespread application of these methods will also require enhancements in the collation, availability, quality and scope of data on species distributions. Some suitable data are currently available. For example, distribution maps are available at two time points (years 1970 and 2000) for over 1000 vascular plant species in Britain (Preston *et al.*, 2002). These data combine reported sightings with intensive, targeted fieldwork that aims to reduce the rate of false absences to a uniformly low level across Great Britain. Often such atlas data, presented as representing a single time slice, contain detailed information on the times at which individual records were observed, and this can be used to fit dynamic models (Cook *et al.*, 2007; Kadoya & Washitani, 2010) although results from such analyses must be interpreted with care. On regional to continental scales, comprehensive data on variation in abundance remain rare, yet for relatively easily observed taxa citizen science projects can be valuable data sources (Devictor *et al.*, 2010). For instance, long-term data sets, including local abundance estimates, exist for North American birds (Pardieck & Sauer, 2007) and British butterflies (Pollard & Yates, 1993). Another approach to abundance data may be to use the proportion of occupied grid cells at one scale as a measure of abundance at another. For example, tetrad (2 km × 2 km grid square) resolution distribution data are available for vascular plants for a limited number of counties in Great Britain (see e.g. Collingham *et al.*, 2000) and current schemes are collecting tetrad resolution data for all such species right across Great Britain (Walker *et al.*, 2010). In addition to those data discussed above, developments in experimental and observational methods are likely to generate new data and new data types that the framework outlined here is well placed to exploit with suitably tailored models.

ACKNOWLEDGEMENTS

We would like to thank the organizers of the workshop on 'The ecological niche as a window to biodiversity' in Frankfurt (Christine Römermann, Steve Higgins and Bob O'Hara) for inviting us to attend, and to write this review. We would also like to thank the financial supporters of the workshop, the LOEWE initiative for scientific and economic excellence of the German federal state of Hesse. G.M. and S.C. were funded by the Scottish Government's RESAS. F.H. was supported by ERC advanced grant 233066. G.M. wishes to acknowledge Gil Scott-Heron (1949–2011) for his clarity and inspiration. The *Rhododendron ponticum* data were obtained from the National Biodiversity Network Gateway, and were compiled from numerous sources including the Countryside Council for Wales, Bristol Regional Environmental Records Centre, The Scottish Wildlife Trust, and Scottish Borders Biological Records Centre (see http://www.nbn.org.uk for details).

REFERENCES

Aikio, S., Duncan, R.P. & Hulme, P.E. (2010) Herbarium records identify the role of long-distance spread in the spatial distribution of alien plants in New Zealand. *Journal of Biogeography*, **37**, 1740–1751.

Andrieu, C., de Freitas, N., Doucet, A. & Jordan, M.I. (2003) An introduction to MCMC for machine learning. *Machine Learning*, **50**, 5–43.

Andrieu, C., Doucet, A. & Holenstein, R. (2010) Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, **72**, 269–342.

Arulampalam, M.S., Maskell, S., Gordon, N. & Clapp, T. (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, **50**, 174–188.

Bierman, S., Butler, A., Marion, G. & Kühn, I. (2010) Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data. *Ecography*, **33**, 451–460.

Buckland, S.T., Newman, K.B., Thomas, L. & Koesters, N.B. (2004) State-space models for the dynamics of wild animal populations. *Ecological Modelling*, **171**, 157–175.

Catterall, S., Cook, A.R., Marion, G., Butler, A. & Hulme, P. E. (2012) Accounting for uncertainty in colonisation times: a novel approach to modelling the spatio-temporal dynamics of alien invasions using distribution data. *Ecography*, doi:10.1111/j.1600-0587.2011.07190.x.

Celeux, G., Forbes, F., Robert, C.P. & Titterington, D.M. (2006) Deviation information criteria for missing data models. *Bayesian Analysis*, **1**, 651–706.

Chuine, I. (2010) Why does phenology drive species distribution? *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 3149–3160.

Clark, J.S., Lewis, M.A. & Horvath, L. (2001) Invasion by extremes: population spread with variation in dispersal and reproduction. *The American Naturalist*, **157**, 537–554.

Collingham, Y.C., Wadsworth, R.A., Huntely, B. & Hulme, P.E. (2000) Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent. *Journal of Applied Ecology*, **37**(Suppl. 1), 13–27.

Cook, A., Marion, G., Butler, A. & Gibson, G. (2007) Bayesian inference for the spatio-temporal invasion of alien species. *Bulletin of Mathematical Biology*, **69**, 2005–2025.

Cook, A., Otten, W., Marion, G., Gibson, G. & Gilligan, C. (2007) Estimation of multiple transmission rates for epidemics in heterogeneous populations. *Proceedings of the National Academy of Sciences USA*, **104**, 20392–20397.

Cowles, M.K. & Carlin, B.P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, 883–904.

Cox, D.R. & Miller, H.D. (1965) *The theory of stochastic processes*. Chapman Hall, London.

Devictor, V., Whittaker, R.J. & Beltrame, C. (2010) Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, **16**, 354–362.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B. & Singer, A. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

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., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Engler, R. & Guisan, A. (2009) MigClim: predicting plant distribution and dispersal in a changing climate. *Diversity and Distributions*, **15**, 590–601.

Frost, C. & Thompson, S.G. (2000) Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A*, **163**, 173–189.

Gamerman, D. & Lopes, H.F. (2006) *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman & Hall, London.

Gibson, G.J. (1997) Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *Applied Statistics*, **46**, 215–233.

Gilks, W.R., Richardson, S. & Spiegelhalter, D.J. (eds) (1996) *Markov chain Monte Carlo in practice*. Chapman & Hall, London.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

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

Guisan, A., Edwards, T.C. & Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89–100.

Hartig, F., Calabrese, J.M., Reineking, B., Wiegand, T. & Huth, A. (2011) Statistical inference for stochastic simulation models – theory and application. *Ecology Letters*, **14**, 816–827.

Hartig, F., Dyke, J., Hickler, T., Higgins, S.I., O'Hara, R.B., Scheiter, S. & Huth, A. (2012) Connecting dynamic vegetation models to data – an inverse perspective. *Journal of Biogeography*, **39**, 2240–2252.

Hastings, A., Cuddington, K., Davies, K.F., Dugaw, C.J., Elmendorf, S., Freestone, A., Harrison, S., Holland, M., Lambrinos, J., Malvadkar, U., Melbourne, B.A., Moore, K., Taylor, C. & Thomson, D. (2005) The spatial spread of invasions: new developments in theory and evidence. *Ecology Letters*, **8**, 91–101.

Heikkinen, R.K., Luoto, M., Araújo, M.B., Virkkala, R., Thuiller, W. & Sykes, M.T. (2006) Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography*, **30**, 1–27.

Higgins, S.I., O'Hara, R.B., Bykova, O., Cramer, M.D., Chuine, I., Gerstner, E.-M., Hickler, T., Morin, X., Kearney, M.R., Midgley, G.F. & Scheiter, S. (2012) A physiological analogy of the niche for projecting the potential distribution of plants. *Journal of Biogeography*, **39**, 2132–2145.

Höhle, M., Jørgensen, E. & O'Neill, P.D. (2005) Inference in disease transmission experiments by using stochastic epidemic models. *Applied Statistics*, **54**, 349–366.

Hooten, M.B., Wikle, C.K., Dorazio, R.M. & Royle, J.A. (2007) Hierarchical spatiotemporal matrix models for characterizing invasions. *Biometrics*, **63**, 558–567.

Hulme, P.E. (2009) Relative roles of life-form, land use and climate in recent dynamics of alien plant distributions in the British Isles. *Weed Research*, **49**, 19–28.

Hulme, P.E. (2011) Consistent flowering response to global warming by European plants introduced into North America. *Functional Ecology*, **25**, 1189–1196.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19**, 101–108.

Kadoya, T. & Washitani, I. (2010) Predicting the rate of range expansion of an invasive alien bumblebee (*Bombus*

terrestris) using a stochastic spatio-temporal model. *Biological Conservation*, **143**, 8–1235.

Kass, R.E. & Wasserman, L. (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Kéry, M. & Schmid, H. (2006) Estimating species richness: calibrating a large avian monitoring programme. *Journal of Applied Ecology*, **43**, 101–110.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Svenning, J.-C., Zimmermann, N.E. & O'Hara, R.B. (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial scales. *Journal of Biogeography*, **39**, 2163–2178.

Kuroe, M., Yamaguchi N., Kadoya.T. & Miyashita, T. (2011) Matrix heterogeneity affects population size of the harvest mice: Bayesian estimation of matrix resistance and model validation. *Oikos*, **120**, 271–279.

Latimer, A.M., Wu, S.S., Gelfand, A.E. & Silander, J.A. (2006) Building statistical models to analyze species distributions. *Ecological Applications*, **16**, 33–50.

Lee, P.M. (2004) *Bayesian statistics: an introduction*. Arnold, London.

Linder, H.P., Bykova, O., Dyke, J., Etienne, R.S., Hickler, T., Kühn, I., Marion, G., Ohlemüller, R., Schymanski, S.J. & Singer, A. (2012) Environmental modifiers, environmental modulation and species distribution models. *Journal of Biogeography*, **39**, 2179–2190.

Link, W.A. & Eaton, M.J. (2012) On thinning of chains in MCMC. *Methods in Ecology and Evolution*, **3**, 112–115.

Marjoram, P., Molitor, J., Plagnol, V. & Tavare, S. (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences USA*, **100**, 15324–15328.

Martínez-Meyer, E., Peterson, A.T. & Hargrove, W.W. (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.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*, 2nd edn. Chapman & Hall/CRC Press, Boca Raton, FL.

McInerny, G.J. & Purves, D.W. (2011) Fine scale environmental variation in species distribution models: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2**, 248–257.

O'Hara, R.B., Arjas, E., Toivonen, H. & Hanski, I. (2002) Bayesian analysis of metapopulation data. *Ecology*, **83**, 2408–2415.

O'Neill, P.D. (2010) Introduction and snapshot review: relating infectious disease transmission models to data. *Statistics in Medicine*, **29**, 2069–2077.

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.

Pardieck, K.L. & Sauer, J.R. (2007) The 1999–2003 summary of the North American Breeding Bird Survey. *Bird Populations*, **8**, 28–45.

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

Pollard, E. & Yates, T.J. (1993) *Monitoring butterflies for ecology and conservation*. Chapman and Hall, London.

Preston, C.D., Pearman, D.A. & Dines, T.D. (2002) *New atlas of the British and Irish flora*. Oxford University Press, Oxford.

Purves, D.W. (2009) The demography of range boundaries versus range cores in eastern US tree species. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 1477–1484.

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

Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical modeling and inference in ecology: the analysis of data from populations and communities*. Academic Press, San Diego, CA.

Schurr, F.M., Midgley, G.F., Rebelo, A.G., Reeves, G., Poschlod, P. & Higgins, S.I. (2007) Colonization and persistence ability explain the extent to which plant species fill their potential range. *Global Ecology and Biogeography*, **16**, 449–459.

Schurr, F.M., Pagel, J., Cabral, J.S., Groeneveld, J., Bykova, O., O'Hara, R.B., Hartig, F., Kissling, W.D., Linder, H.P., Midgley, G.F., Schröder, B., Singer, A. & Zimmermann, N. E. (2012) How to understand species' niches and range dynamics: a demographic research agenda for biogeography. *Journal of Biogeography*, **39**, 2146–2162.

Schweiger, O., Heikkinen, R.K., Harpke, A., Hickler, T., Klotz, S., Kudrna, O., Kühn, I., Pöyry, J. & Settele, J. (2012) Increasing range mismatching of interacting species under global change is related to their ecological characteristics. *Global Ecology and Biogeography*, **21**, 88–99.

Thuiller, W., Richardson, D.M., Pyšek, P., Midgley, G.F., Hughes, G.O. & Rouget, M. (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, **11**, 2234–2250.

Thuiller, W., Albert, C.H., Araújo, M.B., Berry, P.M., Cabeza, M., Guisan, G., Hickler, T., Midgley, G.F., Paterson, J., Schurr, F.M., Sykes, M.T. & Zimmermann, N.E. (2008) Predicting global change impacts on plant species distributions: future challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, **9**, 137–152.

Václavík, T. & Meentemeyer, R.K. (2009) Invasive species distribution modelling (iSDM): are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, **220**, 3248–3258.

Walker, K.J., Pearman, D.A., Ellis, R.W., McIntosh, J.W. & Lockton, A. (2010) *Recording the British and Irish flora, 2010–2020*. Botanical Society of the British Isles, London.

Wikle, C.K. (2003) Hierarchical Bayesian methods for predicting the spread of ecological processes. *Ecology*, **84**, 1382–1394.

Williams, R.J. & Purves, D.W. (2011) The probabilistic niche model reveals substantial variation in the niche structure of empirical food webs. *Ecology*, **92**, 1849–1857.

Zurell, D., Jeltsch, F., Dormann, C.F. & Schröder, B. (2009) Static species distribution models in dynamically changing systems: how good can predictions really be? *Ecography*, **32**, 733–744.

Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B. & Grimm, V. (2010) The virtual ecologist approach: simulating data and observers. *Oikos*, **119**, 622–635.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Software and documentation for the first example: presence/absence data from a single point in time.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCH

Glenn Marion is a mathematical biologist with interests in developing specific models and generic mathematical and statistical methods for dynamic process-orientated models with broad applications in ecology and epidemiology.

Author contributions: All authors contributed to the development of the ideas and to the writing. G.M. led the writing.

Editor: Peter Linder

The papers in this Special Issue arose from two workshops entitled 'The ecological niche as a window to biodiversity' held on 26–30 July 2010 and 24–27 January 2011 in Arnoldshain near Frankfurt, Germany. The workshops combined recent advances in our empirical and theoretical understanding of the niche with advances in statistical modelling, with the aim of developing a more mechanistic theory of the niche. Funding for the workshops was provided by the Biodiversity and Climate Research Centre (BiK-F), which is part of the LOEWE programme 'Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz' of Hesse's Ministry of Higher Education, Research and the Arts.