Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 82–94 doi: 10.1111/j.2041-210x.2012.00264.x

Downscaling of species distribution models: a hierarchical approach

Petr Keil[1,2,*], Jonathan Belmaker[1,3], Adam M. Wilson[1], Philip Unitt[4] and Walter Jetz[1]

[1]*Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT, 06520, USA;* [2]*Center for Theoretical Study, Charles University and the Academy of Sciences of the Czech Republic, Jilska 1, 110 00 Praha 1, Czech Republic;* [3]*Department of Zoology, George s. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, 69978, Israel; and* [4]*Department of Birds and Mammals, San Diego Natural History Museum, PO Box 121390, San Diego, CA, 92112-1390, USA*

Summary

1. Reliable methods to downscale species distributions from coarse to fine grain (equivalent to resolution or support) hold great potential benefit for ecology and conservation. Existing methods have been based on partially unrealistic assumptions and yield mixed results.

2. Here, we introduce a novel and simple approach for downscaling species distribution models based on a hierarchical Bayesian modelling (HBM) framework. Our approach treats putative (unknown) fine-grain presences/absences as latent variables, which are modelled as a function of observed fine-grain environmental variables and constrained by observed coarse-grain presences/absences using logistic regression. The aim is to produce downscaled fine-grain probabilities of species occurrence that (1) closely resemble the probabilities produced by a logistic model parameterized with the observed fine-grain data (the 'reference model') and (2) are improvements over conventional downscaling methods. We additionally test how fine-grain occupancy based on power-law scale-area relationships modifies the downscaling results. We test our approach on 127 bird species from the San Diego breeding atlas data surveyed at 5 km grain.

3. The HBM approach provides unbiased fine-grain probabilities of occurrence whilst the conventional methods (direct approach, point sampling) consistently over-predict occurrence probabilities. Incorporation of the downscaled occupancy further improves reliability of the models, but only in cases when the fine-grain occupancy is estimated accurately.

4. Summing predictions across grid cells and species, HBMs provide better estimates of fine-grain species richness than conventional methods. They also provide better estimates of fine-grain occupancy (prevalence).

5. The presented HBM-based downscaling approach offers improved predictions of fine-grain presence and absence compared with existing methods. The combination of the Bayesian approach with key macroecological relationships (specifically, the scale-area relationship) offers a promising general basis for downscaling distributions that may be extended, for example, using generalized linear or additive models. These approaches enable integrative predictions of spatial biodiversity patterns at fine grains.

Key-words: AUC, birds, California, fractal, logit, niche modelling, realized niche, SDM, spatial scale, San Diego

Introduction

Thanks to increasingly fine-grain climatic and remotely sensed data available for species distribution modelling (SDM), the grain (i.e. the area of a single grid cell in a gridded dataset) of the species occurrence data has become the key limiting factor for many questions in ecology and conservation (Jetz, McPherson & Guralnick 2012). Either implicitly (e.g. point observations with high rate of false negatives and/or large spatial uncertainty) or explicitly (e.g. biodiversity atlas data at defined grain) species distribution knowledge is typically much coarser than the grain of many biological processes, the grain at which management decisions are made and the grain of available environmental data.

There is thus great potential value and opportunity in the *downscaling* of species distributions, that is, the combination of coarse-grain species occurrences with fine-grain environmental data to predict species' distributions at a fine grain, in a single modelling framework. With increasing availability of survey data and fine-grain environmental information, downscaling methods have been of growing interest (Lloyd & Palmer 1998; Araújo *et al.* 2005; McPherson, Jetz & Rogers 2006; Trivedi *et al.* 2008; Niamir *et al.* 2011; Bombi & D'Amen 2012). However, many of the proposed downscaling methods have provided results of mixed quality (Araújo *et al.*

*Correspondence author. E-mail: pkeil@seznam.cz

2005; McPherson, Jetz & Rogers 2006 and references therein), and no methods exist that explicitly account for the hierarchical spatial structure in ecological data.

For example, the *direct approach* (Araújo *et al*. 2005; McPherson, Jetz & Rogers 2006; Bombi & D'Amen 2012) applies model parameters estimated at coarse grains to predict fine-grained species occurrences using fine-grained environmental variables. It relies on the strong assumption that fine-grain species distributions show the same environmental associations as distributions at the coarse grain – an assumption that usually does not hold (e.g. Menke *et al*. 2009). The *iterative approach* (McPherson, Jetz & Rogers 2006) is similar to the direct approach, but instead of applying the coarse-grain model directly to fine grains, it first downscales species distributions to some intermediate grains and then to the fine grains (the algorithm is quite complex; see McPherson, Jetz & Rogers 2006 for details). However, the biologically unrealistic assumption of the direct approach remains. In the third approach, termed *point sampling* (Lloyd & Palmer 1998; McPherson, Jetz & Rogers 2006; Niamir *et al*. 2011), fine-grain cells are sampled (by various algorithms) from the occupied coarse-grain cells. The sampled fine-grain cells are then treated as fine-grain presences and linked to fine-grain environment. This approach requires the different, but also usually untenable assumption that all areas within the occupied coarse-grain cells host conditions suitable for the modelled species. Finally, in the *clustering approach* (McPherson, Jetz & Rogers 2006), the fine-grain cells in the occupied coarse-grain cells are subject to cluster analysis based on environment within the cells. Clusters of fine-grain cells that are present in all of the occupied coarse-grain cells are then expected to contain environments suitable for the species. The approach assumes that each occupied grid cell must have some favourable habitat and that this habitat is more homogeneous from one occupied cell to the next than unfavourable habitat (McPherson, Jetz & Rogers 2006). As a limitation, the method ignores absence information. Several other methods have been proposed, such as the range clipping method based on prior *expert knowledge* of a species' habitat preferences (Jetz, Wilcove & Dobson 2007; Niamir *et al*. 2011; Rondinini *et al*. 2011) or physiological climatic tolerances (Kearney & Porter 2009). Niamir *et al*. (2011) suggested a *hybrid approach* in which point sampling was combined with expert knowledge on habitat preferences. Here, we focus on situations where only species presences/absences and environmental data are available, and thus do not consider models that rely on additional expert knowledge.

Hierarchical Bayesian models (hereafter HBM) (Gelfand *et al*. 2005; Latimer *et al*. 2006; MacKenzie *et al*. 2006; Chakraborty *et al*. 2010; Wilson *et al*. 2011) offer a framework for integrating processes acting at different grains whilst avoiding the potentially untenable assumptions listed above. In spatial ecology, HBMs have been used to combine point species occurrence data (Latimer *et al*. 2006) or abundance data (Chakraborty *et al*. 2010) and gridded environmental conditions. MacKenzie *et al*. (2006) provide basic applications of HBMs in the field of SDM, including the use of species' occurrences at the nonsurveyed sites as latent variables (although not in the multi-scale context). McInerny & Purves (2011) used HBM to retrieve fine-grain species' environmental niches from coarse-grain environmental data with unknown (latent) fine-grain variation. Niamir *et al*. (2011) used the term Bayesian to describe their downscaling method but in their case it refers to the way of combining prior expert knowledge on habitat requirements with fine-grain environment (their MaxEnt SDM is not Bayesian). To our knowledge, a broadly applicable hierarchical Bayesian approach to downscaling species distribution models from coarse to fine grains has not been tested.

Although it is still unclear why, species distributions can be to some degree self-similar (fractal) across grains (Virkkala 1993; Condit *et al*. 2000) or predictably self-dissimilar (Lennon *et al*. 2007; Storch *et al*. 2008; Azaele, Cornell & Kunin 2012). Hence, some of the fine-grain properties of species distributions can be predicted independent of environmental variables using only coarse-grain occurrences. One such property is species' *occupancy*, which is the number of occupied grid cells at a given grain and which is convertible to *prevalence* (proportion of occupied grid cells). It has been shown that the area of occupied grid cells (convertible to occupancy) scales more or less predictably with grain. This scaling relationship is known as the occupancy-area relationship (He & Condit 2007), *scale-area relationship* (Kunin 1998; this is the term that we hereafter use), range area relationship (Harte *et al*. 2005), area–area curve (IUCN Standards & Petitions Subcommittee 2011) or scaling pattern of occupancy (Hui *et al*. 2009). Kunin (1998) used a simple power-law (fractal) scale-area relationship to estimate fine-grain occupancy using only coarse-grain occupancy. The exercise was then refined using other models that do not assume strict fractality (see Azaele, Cornell & Kunin 2012 for review). Although there is no general consensus on which of the models is the best, it is becoming clear that estimating fine-grain species' occupancy from coarse data is a promising prospect (Azaele, Cornell & Kunin 2012). Interestingly, none of the current SDM-downscaling techniques incorporate it.

In this paper, we introduce a HBM approach to downscale species distribution models combining coarse-grain species presences/absences and fine-grain environmental data. It avoids loss of information caused by averaging of fine-grain environment (as in the direct approach; McPherson, Jetz & Rogers 2006) and arbitrary re-sampling of the coarse-grain presences/absences (as in the point sampling approach; McPherson, Jetz & Rogers 2006). It is based on a simple HBM model as follows: (1) the unobserved or 'latent' fine-grain probabilities of species occurrences are modelled as a function of the observed fine-grain environmental conditions (the data). (2) In each coarse-grain grid cell, the fine-grain occurrence probabilities are combined and the resulting coarse-grain occurrence probability is linked to the observed coarse-grain presences/absences (the data). (3) The scale-area relationship (*sensu* Kunin 1998) is incorporated by fitting the scale-area relationship at coarse grains and then extrapolating it to predict occupancy at finer grains. We use fine-grained survey data for 127 bird species to compare the success of this new

method to alternative approaches and to explore its broader applications.

Materials and methods

THE DATA

We tested our downscaling methods using data on 127 breeding bird species surveyed in San Diego County, California, USA (Fig. 1). Specifically, we used data from the San Diego County Bird Atlas (Unitt 2005), collected over a period of 5 years, from March 1997 to February 2002. The presence–absence data are organized in a grid system of 479 cells that are approximately 5×5 km^2 each (Figs 1 and 2a). The dataset is especially suitable for our purposes as the rate of false absences should be negligible (high sampling effort in small region, with extensive expert effort dedicating to minimizing false absence rate), and it covers a heterogeneous landscape with steep environmental gradients (Unitt 2005). We used only breeding season species distributions data. The resulting 5×5 km^2 data (Figs 1 and 2a), which we consider *fine grain*, were used to generate *coarse-grain* data of 15×15 km^2 (Figs 1 and 2b) and *super coarse-grain* data of 30×30 km^2 (Fig. 2c) by setting all coarse-grain grid cells with at least one occurrence in a nested fine-grain grid cell to constitute presences. We excluded some of the fine-grain cells along the margins so that the fine-grain data were fully nested within the coarse-grain grid. We excluded all species that we considered to be accidental observations (based on W.J.'s expert knowledge) and all 'waterbirds' in the broadest sense (e.g. ducks, geese, terns, waders), that is, those that predominantly feed in or around water during the breeding season, because their key habitat requirement is not well captured and not straightforwardly correlated with our main environmental variables. We analysed only species that were present in at least 3 (of 46) coarse-grain grid cells and with a maximum of 44 (of 46) occupied coarse-grain cells. This resulted in 414 fine-grain (Fig. 2a), 46 coarse-grain (Fig. 2b) and 11 super coarse-grain (Fig. 2c) grid cells. We also excluded $\approx$ 20 species for which the nonhierarchical models were impossible to fit realistically (see the model fitting section for details). This resulted in 127 bird species in total (of 389 bird species that breed in San Diego county).

For each of the fine-grain 5×5 km^2 grid cells, we used five environmental variables to predict species distributions. These were as follows: normalized difference vegetation index (NDVI), mean summer temperature, mean annual precipitation, mean elevation and square-root transformed urbanized area (see Menke *et al*. 2009 for detailed description of these variables). Colinearity between predictors can bias parameter estimates and may cause convergence problems in Markov Chain Monte Carlo (MCMC) algorithms (Clark & Gelfand 2006). Hence, we subjected the five variables to principal components analysis (PCA) and extracted the first two PCA axes (centred, standardized, explained 58·9% and 28% of variability, respectively), which then served as our environmental predictors of species distributions. Hereafter, we call them envi1 (correlated with NDVI, Pearson's $r = -0{\cdot}83$; summer temperature, $r = 0{\cdot}88$ and precipitation, $r = -0{\cdot}98$) and envi2 (correlated with urbanized area, $r = -0{\cdot}88$ and elevation, $r = 0{\cdot}62$).

THE MODELS

The models we present below have two levels. The first represents the relationship between fine-grain probability of occurrence and fine-grain environmental conditions – this is the SDM part of the model. We use logistic regression as the SDM part, but it can be readily replaced by other functions. The second level models how the relationship between occurrence probabilities and environment behaves at different grains, and it is the main focus of this study.

Our first model (Model 1) does not involve any downscaling part and was parameterized using fine-grain presence/absence data and the fine-grain environmental variables. Because this model uses the true fine-grain data on species' occurrences, we consider it the 'best-case' scenario and use it as a reference model to judge the performance of the remaining models that use coarse occurrence data. We then present the two conventional models (Models 2 and 3), which are similar to those used in previous downscaling studies (McPherson, Jetz & Rogers 2006). Finally, we introduce three novel HBMs (Models 4–6).

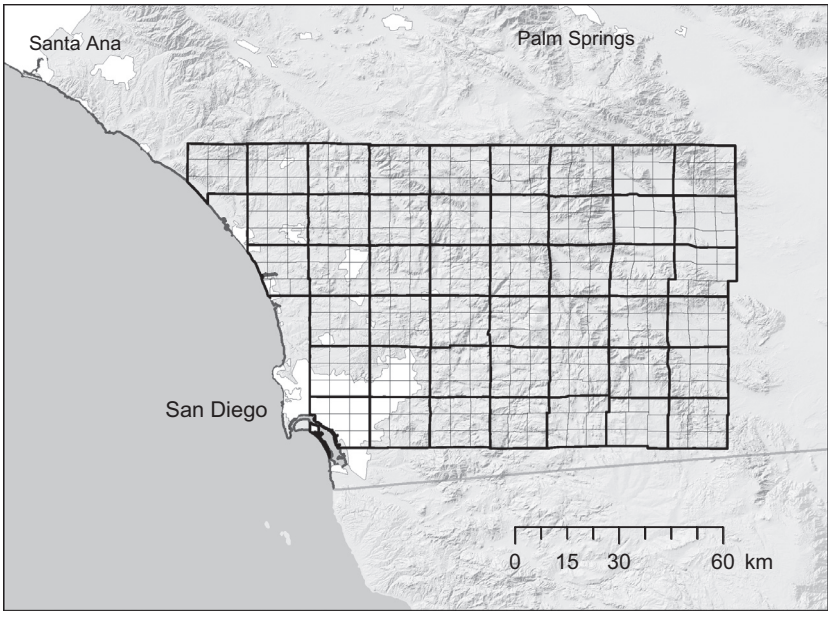

Fig. 1. Geographical extent of our study. Displayed are shaded relief, major urban areas (white), the finest-grain grid (5×5 km) and the coarse-grain grid (15×15 km). The grid system is based on the historic Public Land Survey System (US Department of the Interior 1973).

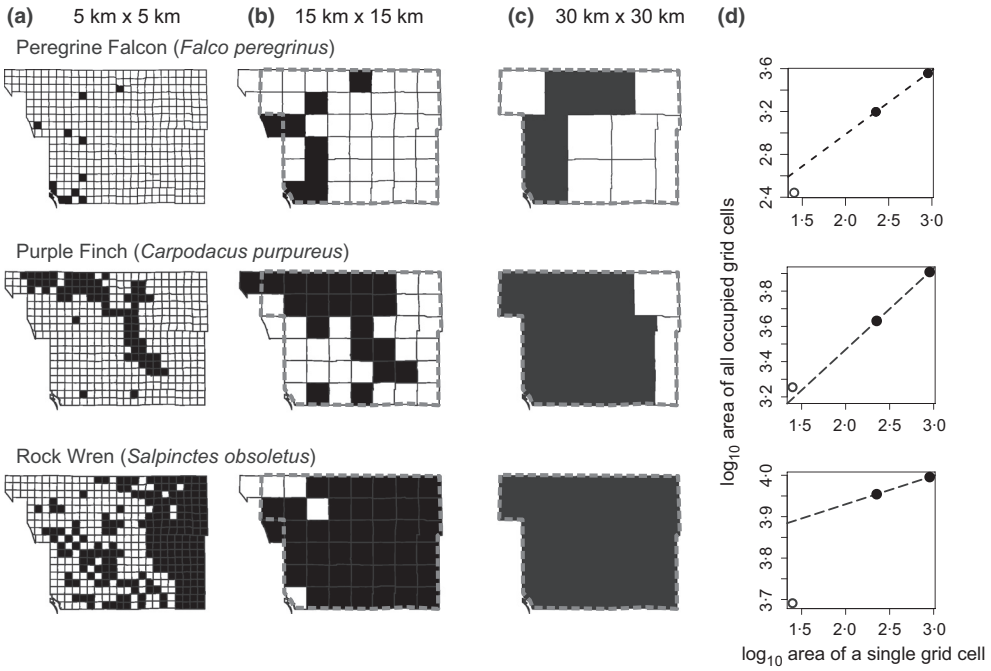

Fig. 2. Example of distribution data for three species (differing in their occupancy – from rare to widespread) at three grains. The dashed line in (b) and (c) delineates areas that were used to downscale the proportion of fine-grain occupancy using the power-law scale-area model. Panel (d) shows how the power-law (dashed line) scale-area model was fitted to the 15 × 15 km data and 30 × 30 km (filled circles) and extrapolated to the 5 × 5 km fine grain for Model 5. Empty circle is the true occupancy at 5 × 5 km.

All of our models are variants of logistic regression for the binary response variables (Guisan, Edwards & Hastie 2002). The main reason for using logistic regression is its simplicity, transparency and its generally strong performance and wide-spread use for presence–absence data (Elith *et al.* 2006). The model formula is:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \times \text{envi1} + \gamma \times \text{envi2} \qquad \text{eqn 1}$$

where p is the vector of probabilities of presence of a modelled species at sites described by vectors of environmental conditions envi1 and envi2. α, β and γ are regression coefficients. For illustration, we prefer the α, β and γ notation to the equivalent β_0, β_1 and β_2 (or 'betas') used in some literature. Note that eqn 1 does not explicitly refer to any particular grain yet.

Model 1 – the fine-grain reference model

To estimate the accuracy of the downscaling techniques (Models 2–6), we first need a reference model that is parameterized using both fine-grain environmental data and fine-grain data on presences/absences, that is, a reference fine-grain model (hereafter Model 1). The closer the parameters of Models 2–6 are to those of Model 1 the better. Model 1 is described as follows: Let p_{ij} be the probability of a species' occurrence at fine-grain grid cell j within a coarse-grain grid cell i. p_{ij} is related to fine-grain environmental variables envi1_{ij} and envi2_{ij}:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha + \beta \times \text{envi1}_{ij} + \gamma \times \text{envi2}_{ij}. \qquad \text{eqn 2}$$

We fitted eqn 2 using maximum likelihood and estimated 95% confidence intervals for the model parameters.

Model 2 – direct approach

Model 2 represents the conventionally used 'direct approach' (Araújo *et al.* 2005; McPherson, Jetz & Rogers 2006; Bombi & D'Amen 2012). It assumes that species distributions at fine grains are driven by the same processes as at coarse grains. First, we fitted the following formula:

$$\log\left(\frac{P_i}{1-P_i}\right) = \alpha + \beta \times \text{envi1}_i + \gamma \times \text{envi2}_i \qquad \text{eqn 3}$$

where P_i is the probability of a species' occurrence in a coarse-grain cell i, and envi1_i and envi2_i are mean values of environmental conditions of the fine-grain grid cells that lay within coarse-grain cell i. We fitted eqn 3 to the coarse-grain presence/absence data using maximum likelihood. Estimated values of α, β and γ from eqn 3 were then used *directly* in eqn 2 to predict the fine-grain probabilities p_{ij}.

Model 3 – point sampling

The point sampling approach (McPherson, Jetz & Rogers 2006; Bombi & D'Amen 2012) avoids coarsening the fine-grain environmental variables (as carried out in Model 2) by randomly choosing, within each coarse-grain cell, a fixed number (one in our case) of fine-grain cells. Point sampling assumes that all fine-grain cells within the occupied coarse-grain cell host conditions equally suitable for the modelled species. These fine-grain cells were assigned species presence/absence values according to values in the coarse-grain cell to which they belong. The resulting sub-sampled fine-grain data were then used to fit eqn 2. The sampling procedure was repeated 100 times for each species, and we estimated values of α, β and γ as averages of the 100 outcomes. We also ran the model with 2–8 sampled fine-grain cells, and the results were nearly identical.

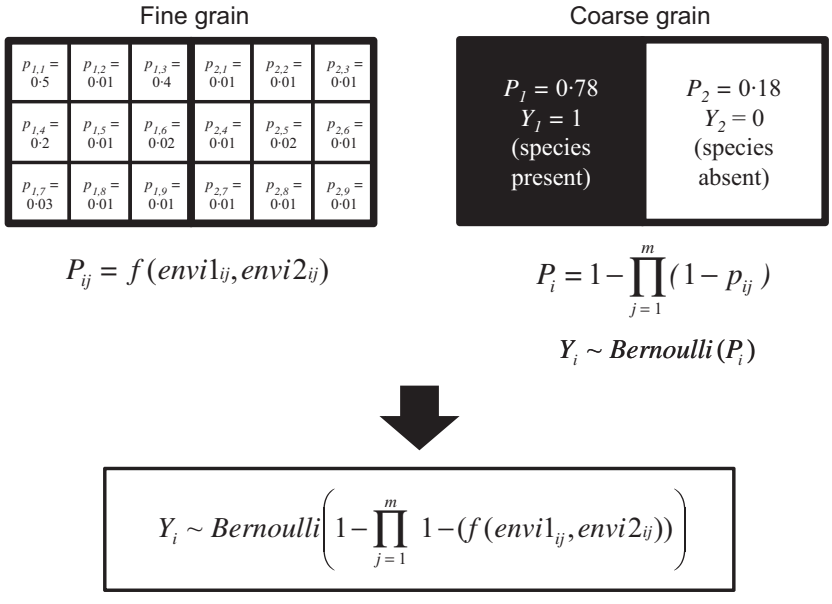

Fig. 3. Rationale of the hierarchical Bayesian modelling approach used in this study. Fine-grain probabilities p_{ij} are linked to fine-grain environment ($envi1_{ij}$ and $envi2_{ij}$) by function $f()$, which is the logistic function (eqn 2) in this particular study. However, $f()$ can be any function, and hence the approach is flexible. Coarse-grain probabilities P_i are then calculated from the fine-grain probabilities p_{ij} using eqn 5. The observed coarse-grain occurrences (Y_i) are a result of a random draw from Bernoulli distribution with probability P_i. Equation in the box is the integrated solution of the relationship between coarse-grain occurrence data, fine-grain probabilities of occurrence and fine-grain environment. The ultimate goal is to find the posterior distribution of the parameters in $f()$, which is carried out by MCMC sampling.

Model 4 – HBM with environment only

This is the simplest of our HBMs (also summarized in Fig. 3). The model requires only coarse-grain species presences/absences and fine-grain environmental variables (similarly to Models 1–3). The model is described as follows: Let p_{ij} be the probability of a species' (unobserved, or 'latent') occurrence at fine-grain grid cell j within a coarse-grain grid cell i. p_{ij} is related to environmental variables $envi1_{ij}$ and $envi2_{ij}$ (in the same way as in eqn 2):

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha + \beta \times \text{envi1}_{ij} + \gamma \times \text{envi2}_{ij}$$

where parameters α, β and γ are random variables. The next level in the model links the fine-grain probabilities to the coarse-grain observed occurrence data. Let Y_i be the observed presence/absence (1/0) of the modelled species at coarse-grain grid cell i. Y_i is thus an outcome of a Bernoulli trial:

$$Y_i \sim \text{Bernoulli}(P_i) \qquad \text{eqn 4}$$

where P_i is the probability that at least one of the m fine-grain grid cells will be occupied by the species. It equals one minus the union of probabilities that the species will be absent at each of the m fine-grain grid cells. Because we are treating the fine-grain absences as independent events, the union probability of fine-grain absences equals to the product of individual probabilities. Hence, we get:

$$P_i = 1 - \prod_{j=1}^{m}(1 - p_{ij}) \qquad \text{eqn 5}$$

where $1 - p_{ij}$ is the probability of absence of the species in the fine-grain cell j within a coarse-grain cell i. See MacKenzie *et al.* (2006) and Marcer *et al.* (2012) for other examples of use of eqn 5 in the field of SDM. The code for Model 4 in the BUGS language is provided in Appendix S1.

Model 5 – HBM with downscaled occupancy

This model builds upon Model 4 (it uses eqns 2, 4 and 5) to link fine-grain probabilities of occurrence to fine-grain environment. However, it also incorporates an estimate of fine-grain occupancy from the power-law model of Kunin (1998) as a further constraint. We thus obtain two kinds of occupancy estimates. The first is λ_{PL}, which is the fine-grain occupancy estimated by the power-law model (see the next section for details). The second is λ_B, which is fine-grain occupancy estimated using the summed p_{ij}:

$$\lambda_B = \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \qquad \text{eqn 6}$$

where n is the number of all coarse-grain grid cells in the studied region and m is the number of fine-grain grid cells within one coarse-grain grid cell ($n = 46$ and $m = 9$ in our dataset; see below). To apply the constraint on model occupancy, we use λ_{PL} as data whilst λ_B is estimated during the model fitting process. We link the separate provenances by assuming that λ_{PL} (data) is an outcome drawn from Poisson distribution with mean λ_B:

$$\lambda_{PL} \sim \text{Poisson}(\lambda_B). \qquad \text{eqn 7}$$

We used Poisson distribution here as it is the simplest way to describe mean and variation in the number of occupied fine-grain grid cells – it constrains λ_B to be reasonably close to λ_{PL} but does not require it to be identical. The code for this model in the BUGS language is provided in Appendix S2.

Model 6 – HBM with true occupancy

This model is nearly identical to Model 5. The only difference is that, instead of using of the occupancy estimated by the power-law model

(λ_{PL}), we use the true occupancy observed at the fine grain, which we call λ_{TRUE} and hence:

$$\lambda_{TRUE} \sim \text{Poisson}(\lambda_B). \quad \text{eqn 8}$$

This model explores the ideal, but rare, situation in which we have a perfect estimate of the fine-grain occupancy. This model is used here to assess the relative performances of Models 4 and 5. The implementation of the model in the BUGS language is provided in the Appendix S2.

ESTIMATION OF FINE-GRAIN OCCUPANCY BY THE POWER-LAW SCALE-AREA RELATIONSHIP

Here, we describe the way we calculated λ_{PL} (eqn 7) using a simple power-law scale-area relationship. We assume that the relationship between the logarithm of area of occupied grid cells (convertible to occupancy) and the logarithm of grain (i.e. area of a single grid cell) can be described by a linear function (Fig. 2d). This is equivalent to the power-law model of self-similar (fractal) spatial distribution of occupied grid cells (Kunin 1998). In the fractal model, the area of occupied grid cells decreases at a constant rate towards finer grains. Hence, we can parameterize the model at coarse grains and then use it to predict occupancy at fine grain. For this first demonstration, we chose the power-law model for its simplicity and transparency, but acknowledge that alternative, more complex scale-area models are available that may ultimately offer stronger fits (see e.g. Azaele, Cornell & Kunin 2012).

To fit the power-law model, we used the coarse-grain (Fig. 2b) and super coarse-grain data (Fig. 2c) and then extrapolated the log–log linear regression down to the fine grain to predict the fine-grain occupancy (Fig. 2d). Additionally, we explored the Poisson (Wright 1991) and negative binomial scale-area models (He & Gaston 2000). These models were fitted using stochastic global optimization of squared log errors (Azaele, Cornell & Kunin 2012). However, as there were only two data points in each species to fit the models, we had difficulties to find a stable set of parameters. Therefore, we only report results based on the power-law model.

MODEL FITTING

We used maximum likelihood approach to find parameters of Models 1–3 and their 95% confidence intervals (function glm() in R, binomial family, logit link function; R Development Core Team 2009). Some of the nonhierarchical models (Models 2 and 3) were impossible to fit realistically for ≈20 species; they provided parameter estimates that differed by >2 orders of magnitude from the reference Model 1 (because of the complete separation problem; see Albert & Anderson 1984). We excluded these species from the analysis and do not show the results here.

To fit Models 4–6, we used OpenBugs 3.2.1 (Lunn *et al.* 2009) to estimate medians of the posterior distributions of α, β and γ and their 95% credible intervals. As an uninformative prior distribution for α, β and γ, we used normal distribution with zero mean and variance of 100. We used three chains and 20 000 iterations from which 10 000 were discarded as burn-in. By visual inspection in 30 species with different values of occupancy, we estimated that the Markov Chains converged after about 1000–3000 iterations. This quick convergence is likely due to (1) the simplicity of the model, (2) the lack of colinearity between predictors and (3) the relatively small data set. Roughly, running the MCMC procedure for Model 6 for a single species required about 3 min using a 2 GHz Intel® Centrino® CoreTM Duo CPU. However, we note that for larger datasets and more complex models, this time can increase substantially. To run the procedure over all of the 127 bird species, we used an R script and the R2OpenBUGS package. The implementation of the models in the BUGS language is provided in Appendices S1 and S2.

MODEL EVALUATION

The goal of our model evaluation was to identify how well the downscaling modelling approach (Models 2–6) is able to approximate the success (or failure) of reference Model 1. Doing the comparison to the fine-grain reference model rather than to the empirical fine-grain data offers clearer and more detailed differences between the models, especially in species with weaker performance of the fine-grained models (i.e. weak association between species' fine-grain probability of occurrence and the two environmental variables). Our goal was thus to identify the modelling approach (Models 2–6) that gives the highest concordance with the reference Model 1.

We measured the discrimination capacity of each of the models by AUC (area under receiver operating characteristic curve; Liu, White & Newell 2011), and we measured the reliability (or goodness-of-fit) of the models using R^2 (using formula in Ash & Shwartz 1999 and Liu, White & Newell 2011). Note that because we are comparing binary with continuous data, the expected R^2 will be lower than when comparing two continuous variables, and it will be correlated with prevalence (occupancy) of a species (Ash & Shwartz 1999). AUC and R^2 were calculated using the predicted fine-grain probabilities of occurrence and the actual fine-grain presences or absences (Liu, White & Newell 2011). We calculated $\Delta\alpha$, $\Delta\beta$, $\Delta\gamma$, ΔAUC and ΔR^2 by subtracting the estimated parameter value for Models 2–6 from the value of Model 1. The closer the Δ values to 0 the better was the concordance of Models 2–6 to the fine-grain reference Model 1. We were also interested in how species occupancy can influence the downscaling accuracy. Hence, we plotted all of the diagnostic Δ measures mentioned above against the species' coarse-grain occupancy (and we incorporated the span of the confidence or credible intervals into the plots).

We performed the model evaluation separately for two sets of species. The first set contained all of the 127 species, regardless of the fit of the fine-grain reference Model 1. The second group consisted of 37 species with AUC > 0·85 (an arbitrary criterion) for Model 1. These were the species for which the environmental variables and the logistic model provided good discrimination capacity.

To evaluate the predictions across all of the species, we calculated species richness at each grid cell as predicted by each of the six models. The expected value of species richness in a given cell is equal to the sum of predicted probabilities of occurrence of individual species (Storch, Šizling & Gaston 2003; Šizling & Storch 2003; MacKenzie *et al.* 2006). As the same logic can be applied to the calculation of predicted occupancy, we also compared values of occupancy predicted by each of the models.

Results

MODEL PERFORMANCE

Values of α, β, γ, AUC and R^2 of the six models are provided in Table 1 and Fig. S1 – these show how well the models predict the empirical fine-grain occurrences, and they show that all three HBMs (Models 4–6) outperform the conventional

Table 1. Median values of R^2, AUC, α, β and γ for the six downscaling models assessed and applied to distributions of birds in San Diego

	All 127 species					37 species with Model 1 AUC > 0·85				
	R^2	AUC	α	β	γ	R^2	AUC	α	β	γ
Model 1 – reference fine-grain model	0·078	0·757	−2·127	−0·05	−0·213	0·374	0·904	−2·814	−0·146	0·024
Model 2 – direct approach	−1·139	0·753	0·176	−0·096	−0·178	−0·615	0·883	−1·353	−0·686	0·573
Model 3 – point sampling	−1·016	0·755	0·179	−0·117	−0·152	−0·556	0·884	−0·95	−0·477	0·356
Model 4 – HBM, only environment	0·016	0·755	−2·71	−0·055	−0·261	0·307	0·884	−4·311	−0·69	0·158
Model 5 – HBM, downscaled occupancy	−0·0013	0·757	−2·065	−0·084	−0·308	0·27	0·882	−4·244	−0·797	−0·069
Model 6 – HBM, true occupancy	0·042	0·755	−2·443	−0·094	−0·282	0·307	0·882	−4·13	−0·792	0·116

The medians were calculated across species. Note that the R^2 values of Models 2–3 can be lower than 0 – a problem that arises when a regression model is fitted to different data than those used to parameterize the model. R^2 (Ash & Shwartz 1999) and AUC (Liu, White & Newell 2011) were calculated using the fine-grain probabilities predicted by the models and true fine-grain presences/absences.
HBM, hierarchical Bayesian modelling.

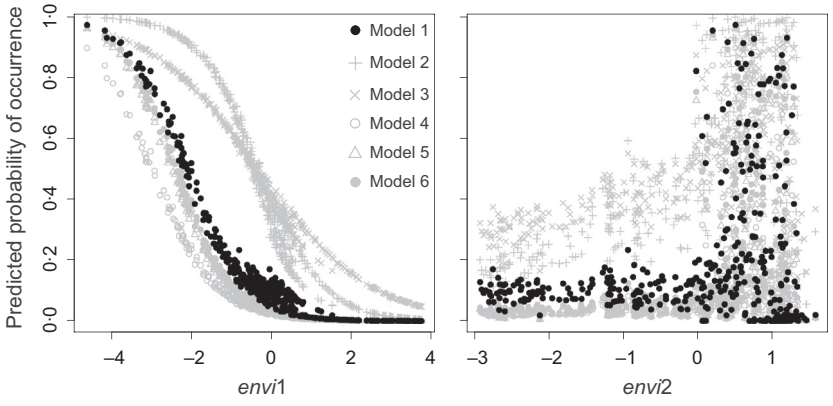

Fig. 4. Example of the six models predicting occurrences of the purple finch (*Carpodacus purpureus*, Gmelin, 1789). Both direct approach (Model 2) and point sampling (Model 3) over-predict the occurrence probabilities when compared with Model 1. All three hierarchical Bayesian modellings give predictions that are in much better concordance with Model 1. Also, note how incorporation of the occupancy (Models 5 and 6) improved the prediction.

Models 2 and 3. The difference between performance of our downscaling models is well illustrated in three species (Figs 4 and 5), which we selected as a representative species of low (Peregrine Falcon, *Falco peregrinus*), medium (Purple Finch, *Carpodacus purpureus*) and high (Rock Wren, *Salpinctes obsoletus*) occupancy. Both of the conventional Models 2–3 (direct method and point sampling) systematically overestimated the fine-grain probabilities of occurrence (Figs 4 and 5), which was mostly caused by unrealistically large values of parameter α (represented by Δα in Fig. 6). They were also less reliable than Model 1 (represented by ΔR^2 in Fig. 6), but produced unbiased values of β, γ and AUC (Fig. 6, Fig. S2). Importantly, the span of 95% confidence intervals of parameter α did not correspond well with the accuracy of the estimated α (Fig. 6).

In contrast to the conventional methods, our HBM models 4–6 produced nearly unbiased estimates of α, β, γ, AUC and R^2 relative to Model 1 (Fig. 6, Fig. S2). Some deviations were apparent in species with extremely high or low occupancy (under/overestimated α, and lower AUC than Model 1) (Fig. 6, Fig. S2). These deviations were less severe than in the conventional Models 2–3 (Fig. 6), and the uncertainty was captured, as is evident in the larger span of the 95% credible intervals for low-occupancy species. Model 6 showed little systematic deviation from Model 1 and generally performed best (Fig. 6, Fig. S2).

The accuracy of some of the downscaling methods (especially Models 2, 3 and 5) was dependent on the occupancy of the species (Fig. 6; Fig. S2). All methods tended to overestimate parameters α in widespread (i.e. high occupancy) species and tended to underestimate α in species with low occupancy (Fig. 6, Fig. S2). Finally, our results hold for both the complete set of 127 species and for the subset of 37 species with AUC > 0·85 of the fine-grain reference Models 1 (Fig. S2).

INCORPORATION OF DOWNSCALED OCCUPANCY

Our efforts to downscale occupancy (λ_{PL}) gave mixed results. Although the downscaled occupancy estimated from the power-law model was highly correlated with the true fine-grain occupancy (see legend of Fig. 7 for R^2), it tended to overpredict the true occupancy (Fig. 7). The incorporation of the downscaled occupancy from the power-law model (λ_{PL}) did

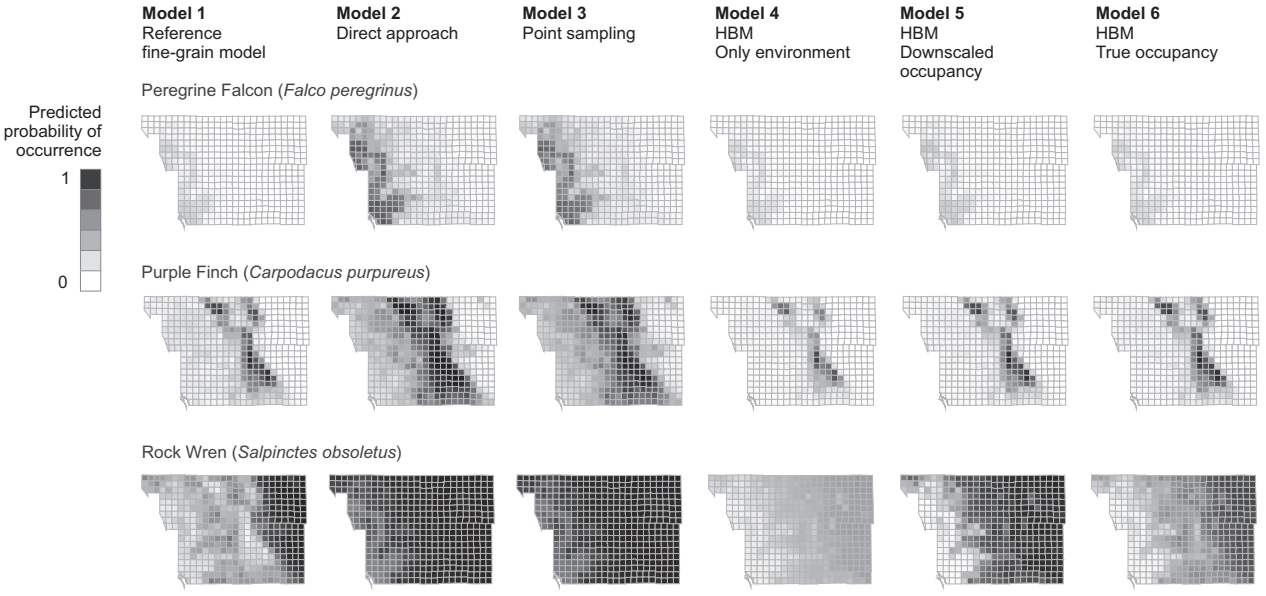

Fig. 5. Example of the six models predicting occurrences of three bird species with different occupancy. Both direct approach and point sampling over-predict compared with Model 1, whilst the three hierarchical Bayesian modellings offer better concordance.

not improve performance of Model 5 over Model 4 (Fig. 6; Fig. S2). Note that Model 4 did not use occupancy estimates at all.

PREDICTED OCCUPANCY AND SPECIES RICHNESS

Model 1 predicted occupancy values almost perfectly (Fig. 8b), which is unsurprising given that logistic regression minimizes the difference between predicted and observed occupancy by definition. Nevertheless, Model 1 had problems with reproducing realistic patterns of species richness (Fig. 8a, Fig. S3). It failed to produce the very high and very low values of richness (Fig. 8a), and it generally smoothed out the species richness maps (Fig. S3). Conventional Models 2 and 3 consistently over-predicted both species richness and occupancy when compared with Model 1 (Fig. 8). On the other hand, all three HBM models produced species richness and occupancy patterns similar to Model 1 (Fig. 8, Fig. S3). Model 5, which incorporated the occupancy estimated from the fractal model (already overestimated; Fig. 7), overestimated both occupancy and species richness, which is expected based on model performance for individual species. Model 4 produced systematically unbiased estimates of both richness and occupancy in comparison with Model 1 (Fig. 8).

Discussion

Overall, the HBMs showed good performance in estimating unbiased, fine-grain species occurrence probabilities. This result is encouraging for a more widespread estimation of species distributions at grains finer than available species occurrence data, with benefits for ecology and conservation. Furthermore, the flexible nature of HBMs opens the possibility of combining species data collected at multiple grains (i.e. point data, coarse gridded data and species lists) to make fine-grain predictions.

RELATIVE PERFORMANCE OF DOWNSCALING METHODS

Our dataset had limited spatial extent and explored a relatively narrow range of scales. Yet, even this constrained setting was sufficient to unveil striking differences in model performance. The HBM approach yielded downscaled occurrence probabilities that were in better concordance with the fine-grain reference model than the conventional downscaling models. The differences were primarily in the values of parameter α, that is, the 'intercept' of the logistic regression for binary response variables and in the R^2 values (the reliability of the model). If we plot the occurrence probabilities against environment (as in Fig. 4), then any change of α moves the whole sigmoidal curve along the environmental x-axis whilst the overall shape of the curve is preserved. This causes severe distortion of the predicted probabilistic maps (as in Fig. 5) and decrease of the R^2. On the other hand, it does not affect the AUC of the model (the discrimination capacity), which in principle only correlates rank of probabilities in grid cells with presences and absences. This has important implications for SDM. Although the conventional methods predict shifted probabilities of occurrence, the relative rank of these probabilities is similar to the rank produced by our HBMs (as shown by the AUC values). It implies that conventional downscaling methods can produce acceptable binary presences or absences when used with an appropriate probability threshold (Liu *et al.* 2005; Allouche, Tsoar & Kadmon 2006; Bombi & D'Amen 2012). However, in practice, there is no straightforward way to estimate such a threshold without a (fine-grain) evaluation dataset. Conceivably, point records could be joined to the fine-grain grid and then used for thresholding, but point

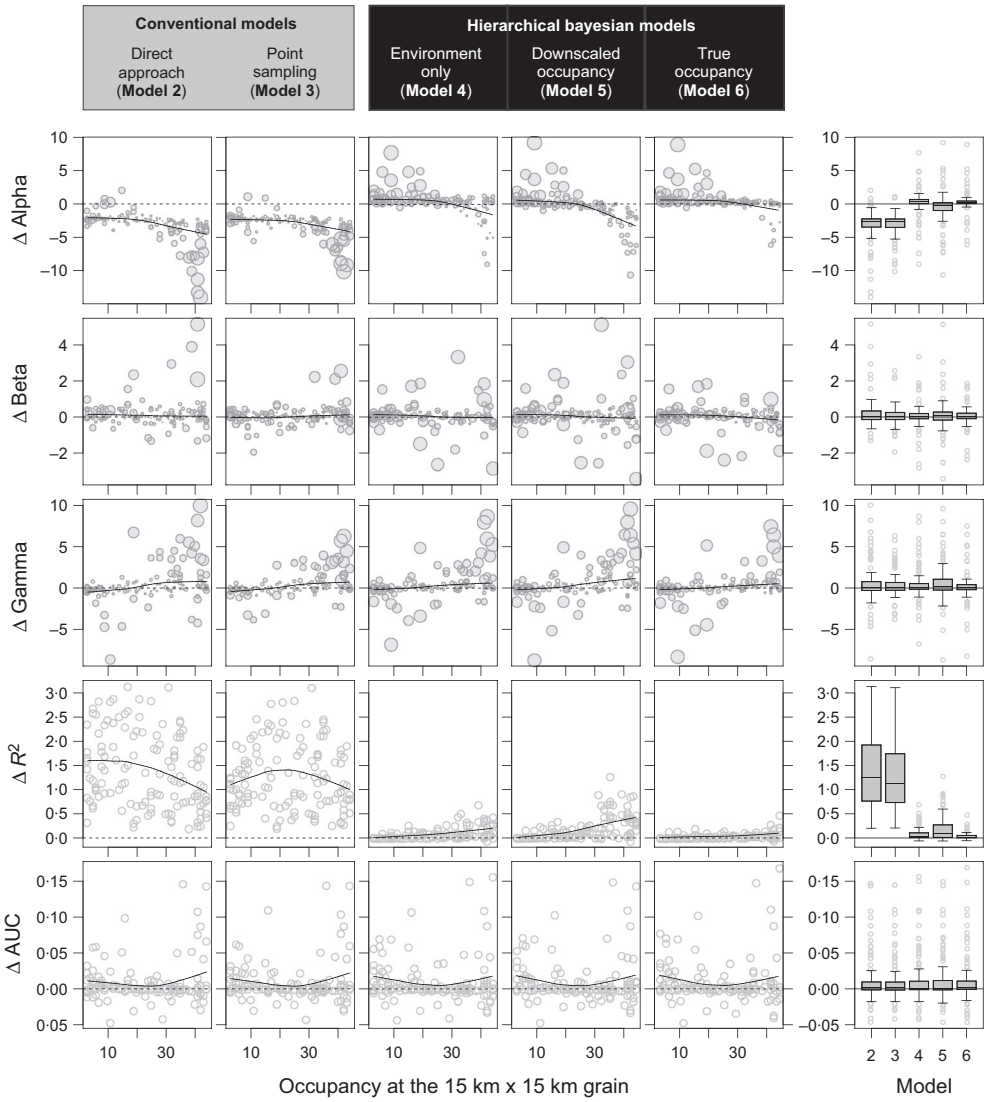

Fig. 6. Relative differences (Δ) in parameter estimation and predictive performance of the five downscaling techniques (Models 2–6) for the 127 species (points) analysed. The Δ values were calculated as the value for the fine-grain reference Model 1 minus the value of a given downscaled model (Models 2–6). The dashed line is the hypothetical perfect match (zero difference between Model 1 and the downscaled model). Solid lines are LOWESS regressions (smoothing span = 2/3, degree = 1), which were weighted by the inverse of the 95% confidence interval (Models 2–3) or 95% credible interval (Models 4–6). Size of the points for $\Delta\alpha$, $\Delta\beta$, $\Delta\gamma$ is proportional to the 95% confidence interval (Models 2–3) or 95% credible interval (Models 4–6) – the larger the point the higher the uncertainty in the parameter estimate. The closer the solid line is to the dashed line the higher the congruence with Model 1. Box-plot panels at the right represent the distribution of the same data across all species regardless of occupancy and have identical y-axes as the corresponding scatter plots (medians, quartiles and extreme values are shown). Occupancy is the number of occupied grid cells at the 15 × 15 km grain.

data are not always available or their tendency for false absences may make their use futile.

One of the advantages of our method set in a logistic regression framework is that it produces actual probabilities of occurrence (with appropriately quantified uncertainty), which allows us to avoid the arbitrary conversion to binary presences/absences in many ecological applications. The raw probabilities can be used to calculate accurate estimates of species richness and their confidence intervals (Storch, Šizling & Gaston 2003), to calculate area occupied by a given species at a given grain (Šizling & Storch 2003; MacKenzie *et al.* 2006), to calculate threshold-free estimates of the most probable range boundaries, to calculate information-rich beta diversity measures such as Bray–Curtis distances or to perform accurate ordination analyses of community composition based on these distances (Legendre & Legendre 1998).

THE ROLE OF OCCUPANCY

Several important issues related to occupancy emerged in our study. First, some parameters of the downscaled logistic function tended to be systematically biased in species with extremely low and high occupancy. At the moment, we do not have a firm explanation for this, and we also could not find one in the current SDM literature, which tends to be focused on evaluation of model predictions not parameters.

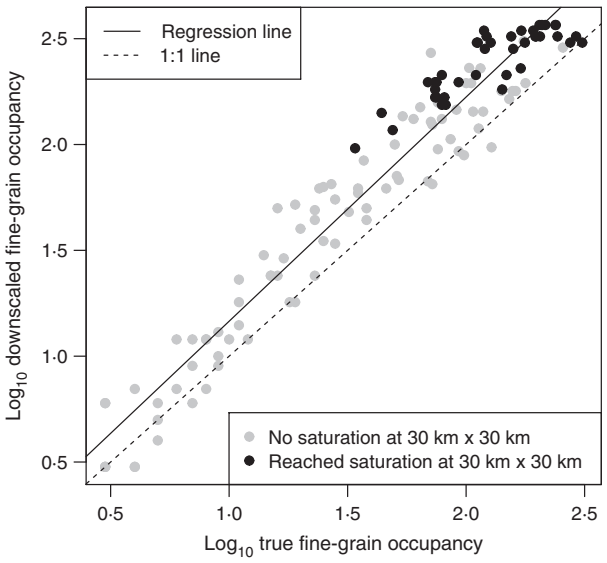

Fig. 7. Ability of the power-law scale-area model (as demonstrated in Fig. 2d) to predict occupancy λ_{PL} (i.e. number of occupied grid cells) at the fine grain for 127 bird species. Note that this figure does not represent performance of the actual downscaling model (Models 1–6). Dashed line is the line of identity ($y = x$). R^2 of the identity line is 0·843. Linear regression (solid line) of downscaled occupancy against true occupancy (in log–log, as plotted) gives R^2 of 0·937, intercept of 0·098 (±0·04 SE), slope of 1·052 (±0·024 SE). Clearly, the power-law model over-predicted occupancy in most of the species, and regardless of whether they reached the saturation scale (100% of occupied grid cells at the coarsest grain).

We need to raise this as a remaining concern and as something that will require further research. However, the good news is that the biased model parameters of the HBMs (e.g. $\Delta\alpha \neq 0$) were also associated with higher uncertainty (represented by larger circle symbols in Fig. 6), especially in less prevalent species. In contrast, the direct approach and point sampling associated the biased parameters with low uncertainty (indicating high confidence in an incorrect value), which we consider to be a serious flaw.

The second issue is related to the power-law scale-area relationship. We had expected that the power law would perform poorly, especially in species with very high occupancy (here measured at the 15 × 15 km grain), because they are close to reaching their 'saturation scale' of 100% of occupied grid cells (Halley *et al.* 2004; Azaele, Cornell & Kunin 2012). Note that we had already excluded species that occupied more than 44 (of 46) 15 × 15 km grid cells. However, we found that, in accordance with previous work (Kunin 1998; He & Gaston 2000; Azaele, Cornell & Kunin 2012), the power-law fractal model overestimated fine-grain occupancy in all species, regardless of their actual occupancy or whether the occupancy reached saturation at the coarsest grain. One may argue that the inclusion of a primitive and inaccurate scaling relationship of occupancy is still better than ignoring the scaling relationship completely. However, what we found is that, whilst accurate estimates of fine-grain occupancy can improve downscaling performance (as in Model 6), not modelling fine-grain occupancy is superior to use of inaccurate (overestimated) occupancy from the power-law model. Nevertheless, there are new promising methods being developed to downscale occupancy such as the shot noise Cox processes (Azaele, Cornell & Kunin 2012), and we expect that larger datasets may facilitate more accurate estimates of fine-grain occupancy and significantly improve the performance of HBM downscaling (as was the case of our Model 6).

Finally, we can regard our HBMs, and especially Model 4, as a stand-alone method to predict fine-grain occupancy (similarly to the power-law scale-area relationship). Model 4 uses logistic regression, which, by definition, fits the sigmoidal curve so that the sum of predicted probabilities is identical to the number of occupied grid cells. This makes it the first choice model in any environmentally informed attempts to downscale occupancy. Whilst Kunin (1998) set the power-law model as the most user-friendly (although unrealistic) model to downscale occupancy with no environmental data at hand, we suggest that HBMs based on logistic regression can be used as the baseline for developing methods to downscale occupancy that do take environment (i.e. species niches) into account (e.g. sensu Jiménez-Alfaro, Draper & Nogués-Bravo 2012).

PROSPECTS FOR HBMS FOR CROSS-SCALE PREDICTIONS OF BIODIVERSITY

We suggest that the flexible nature of the Bayesian modelling framework brings exciting prospects for adding more complexity to the models. Obvious additional elements to incorporate include spatial autocorrelation, for example, in the form of conditional autoregressive models (CAR; Latimer *et al.* 2006). This would help in addressing the independence between grid cells, which we assumed both at fine grain and coarse grains (e.g. in eqn 5).

The sort of HBMs we present could also be extended to include expert knowledge of species habitat requirements (McPherson, Jetz & Rogers 2006; Jetz, Wilcove & Dobson 2007; Kearney & Porter 2009; Niamir *et al.* 2011; Rondinini *et al.* 2011), more complex associations between environmental variables and species presence (e.g. using unimodal rather than sigmoidal species responses curves; McInerny & Purves 2011) or different types of species occurrence data, such as species lists from unequal-sized survey areas. In the future, an incorporation of species abundances or population dynamics (Pagel & Schurr 2012) and joint modelling of multi-species distributions (Ferrier & Guisan 2006) offers additional potential for improved downscaling predictions. There may also be ways to adjust other existing SDM methods (other than GLM) into the HBM framework. An obvious example would be generalized additive models (GAM; Guisan, Edwards & Hastie 2002), which can be easily extended to incorporate HBM as they are parametric and produce actual probabilities of occurrence. Incorporating our HBM approach for methods such as regression trees or MaxEnt (Elith *et al.* 2006) would be a more difficult challenge.

Having outlined all of these prospects, we also note that our HBM approach can be computationally demanding. With increasing complexity of the model, increasingly fine

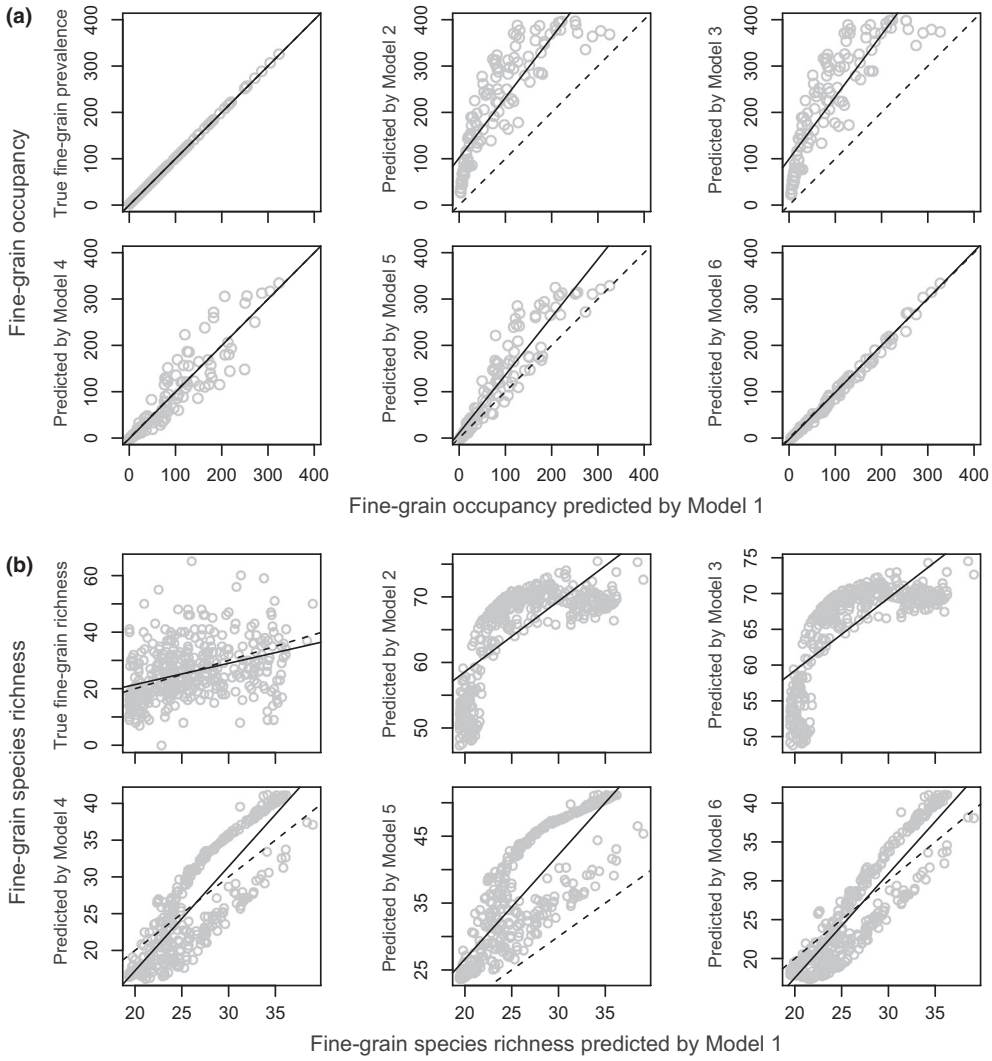

Fig. 8. (a) Fine-grain occupancy (i.e. number of occupied fine-grain grid cells) values predicted by Models 2–6 plotted against occupancy predicted by Model 1. Each point is one species. (b) Species richness predicted by Models 2–6 against species richness predicted by Model 1. Each point is a grid cell (Fig. S3). We calculated occupancy and species richness by summing up the mean probabilities predicted by each model (Storch, Šizling & Gaston 2003; MacKenzie *et al.* 2006). Dashed line is the line of identity; solid line is ordinary least-squares regression prediction. The dashed line is not visible in panel b) in Models 2 and 3 species richness is over-predicting by more than 10 species. In Models 4–6, the predicted lines cluster around two lines, which are caused by different responses of species richness to the two environmental variables. See Fig. S4 for detailed explanation.

grain but large-extent datasets, the computational requirements increase dramatically. This may limit the use of our method primarily to regional rather than continental or global analyses. On the other hand, there is an increasing availability of high-performance cluster computing and associated techniques, including running multiple MCMC chains on separate CPUs, and within-chain parallelization (Chakraborty *et al.* 2010), which may enable the use of HBMs for large spatial extents.

In contrast to conventional methods, the HBM approach provides a full posterior distribution for the estimated probabilities of occurrence in each fine-grain grid cell that incorporates the uncertainty introduced by the downscaling procedure. Although we do not go into such details here, these distributions can be summarized to answer more specific ecological questions whilst fully accounting for model uncertainty. For example, the range size for a species, the probability that two (or more) species co-occur or the expected species lists for any arbitrary region could each be estimated with 95% (or other) credible intervals. Finally, the approach does not require that the fine-grain cells be perfectly nested within the coarser cells, and thus, it is possible to use a HBM to incorporate different types of species occurrence data, such as species lists from any specific regions.

Understanding the spatial distribution of species is critical for understanding and conserving ecological and evolutionary processes, but our knowledge is geographically biased and orders of magnitude coarser than available fine-grain environmental datasets (Jetz, McPherson & Guralnick 2012). In this study, we have illustrated a novel framework for estimating species occurrence probabilities at fine grains by combining

fine-grain environmental data with coarse-grain species occurrence thus improved our fine-grain understanding of species distributions.

Acknowledgements

We are grateful to three anonymous referees for exhaustive comments that improved the manuscript. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no 302868. W.J. acknowledges support for the project by NSF grant DBI 0960550 and DEB 1026764 as well as NASA Biodiversity Grant NNX11AP72G.

References

Albert, A. & Anderson, J.A. (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**, 1–10.

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: occupancy, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Araújo, M.B., Thuiller, W., Williams, P.H. & Reginster, I. (2005) Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. *Global Ecology and Biogeography*, **14**, 17–30.

Ash, A. & Shwartz, M. (1999) R^2: a useful measure of model performance when predicting a dichotomous outcome. *Statistics in Medicine*, **18**, 375–384.

Azaele, S., Cornell, S.J. & Kunin, W.E. (2012) Downscaling species occupancy from coarse spatial scales. *Ecological Applications*, **22**, 1004–1014.

Bombi, P. & D'Amen, M. (2012) Scaling down distribution maps from atlas data: a test of different approaches with virtual species. *Journal of Biogeography*, **39**, 640–651.

Chakraborty, A., Gelfand, A.E., Wilson, A.M., Latimer, A.M. & Silander, J.A. (2010) Modeling large scale species abundance with latent spatial processes. *The Annals of Applied Statistics*, **4**, 1403–1429.

Clark, J.S. & Gelfand, A.E. (2006) *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications*. Oxford University Press, Oxford.

Condit, R., Ashton, P.S., Baker, P., Bunyavejchewin, S., Gunatilleke, S., Gunatilleke, N., Hubbell, S.P., Foster, R.B., Itoh, A., LaFrankie, J.V., Lee, H.S., Losos, E., Manokaran, N., Sukumar, R. & Yamakura, T. (2000) Spatial patterns in the distribution of tropical tree species. *Science*, **288**, 1414–1418.

Elith, J., Graham, C.H., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.C., Townsend Peterson, A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, **43**, 393–404.

Gelfand, A.E., Schmidt, A.M., Wu, S., Silander, J.A., Latimer, A. & Rebelo, A.G. (2005) Modelling species diversity through species level hierarchical modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **54**, 1–20.

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

Halley, J.M., Hartley, S., Kallimanis, A.S., Kunin, W.E., Lennon, J.J. & Sgardelis, S.P. (2004) Uses and abuses of fractal methodology in ecology. *Ecology Letters*, **7**, 254–271.

Harte, J., Colinsk, E., Ostling, A., Green, J.L. & Smith, A.B. (2005) A theory of spatial structure in ecological communities at multiple spatial scales. *Ecological Monographs*, **75**, 179–197.

He, F. & Condit, R. (2007) The distribution of species: occupancy, scale, and rarity. *Scaling Biodiversity* (eds D. Storch, P.A. Marquet & J.H. Brown), pp. 32–50. Cambridge University Press, Cambridge.

He, F. & Gaston, K.J. (2000) Estimating species abundance from occurrence. *American Naturalist*, **165**, 553–559.

Hui, C., McGeoch, M.A., Reyers, B., le Roux, P.C., Greve, M. & Chown, S.L. (2009) Extrapolating population size from the occupancy-abundance relationship and the scaling pattern of occupancy. *Ecological Applications*, **19**, 2038–2048.

IUCN Standards and Petitions Subcommittee (2011) Guidelines for Using the IUCN Red List Categories and Criteria. Version 9.0. Prepared by the Standards and Petitions Subcommittee.

Jetz, W., McPherson, J.M. & Guralnick, R.P. (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology & Evolution*, **27**, 151–159.

Jetz, W., Wilcove, D.S. & Dobson, A.P. (2007) Projected impacts of climate and land-use change on the global diversity of birds. *PLoS Biology*, **5**, 1211–1219.

Jiménez-Alfaro, B., Draper, D. & Nogués-Bravo, D. (2012) Modeling the potential area of occupancy at fine resolution may reduce uncertainty in species range estimates. *Biological Conservation*, **124**, 190–196.

Kearney, M. & Porter, W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, **12**, 334–350.

Kunin, W.E. (1998) Extrapolating species abundance across spatial scales. *Science*, **281**, 1513–1515.

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

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*. Elsevier, Amsterdam.

Lennon, J.J., Kunin, W.E., Hartley, S. & Gaston, K.J. (2007) Species distribution patterns, diversity scaling and testing for fractals in southern African birds. *Scaling Biodiversity* (eds D. Storch, P.A. Marquet & J.H. Brown), pp. 51–76. Cambridge University Press, Cambridge.

Liu, C., White, M. & Newell, G. (2011) Measuring and comparing the accuracy of species distribution models with presence–absence data. *Ecography*, **34**, 232–243.

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**, 385–393.

Lloyd, P. & Palmer, A.R. (1998) Abiotic factors as predictors of distribution in Southern African bulbuls. *The Auk*, **115**, 404–411.

Lunn, D., Spiegelhalter, D., Thomas, A. & Best, N. (2009) The BUGS project: evolution, critique, and future directions. *Statistics in Medicine*, **28**, 3049–3067.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy Estimation and Modelling*. Academic Press, Burlington, MA, USA.

Marcer, A., Pino, J., Pons, X. & Brotons, L. (2012) Modelling invasive alien species distributions from digital biodiversity atlases. Model upscaling as a means of reconciling data at different scales. *Diversity and Distributions*, in press.

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

McPherson, J.M., Jetz, W. & Rogers, D.J. (2006) Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions - possibilities and limitations. *Ecological Modelling*, **192**, 499–522.

Menke, S.B., Holway, D.A., Fisher, R.N. & Jetz, W. (2009) Characterizing and predicting species distributions across environments and scales: Argentine ant occurrences in the eye of the beholder. *Global Ecology and Biogeography*, **18**, 50–63.

Niamir, A., Skidmore, A.K., Toxopeus, A.G., Muñoz, A.R. & Real, R. (2011) Finessing atlas data for species distribution modelling. *Diversity and Distributions*, **17**, 1173–1185.

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.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rondinini, C., Di Marco, M., Chiozza, F., Santulli, G., Baisero, D., Visconti, P., *et al.* (2011) Global habitat suitability models of terrestrial mammals. *Philosophical Transactions of the Royal Society. B, Biological Sciences*, **366**, 2633–2641.

Šizling, A.L. & Storch, D. (2003) Power-law species-area relationships and self-similar species distributions within finite areas. *Ecology Letters*, **7**, 60–68.

Storch, D., Šizling, A.L. & Gaston, K.J. (2003) Geometry of the species–area relationship in central European birds: testing the mechanism. *Journal of Animal Ecology*, **72**, 509–519.

Storch, D., Šizling, A.L., Reif, J., Polechová, J., Šizlingová, E. & Gaston, K.J. (2008) The quest for a null model for macroecological patterns: geometry of species distributions at multiple spatial scales. *Ecology Letters*, **11**, 771–784.

Trivedi, M.R., Berry, P.M., Morecroft, M.D. & Dawson, T.P. (2008) Spatial scale affects bioclimate model projections of climate change impacts on mountain plants. *Global Change Biology*, **14**, 1089–1103.

Unitt, P. (2005) *San Diego County Bird Atlas*. Ibis Publishing, Temecula.
US Department of the Interior (1973) *Manual of Instructions for the Survey of the Public Lands of the United States*. US Department of the Interior, Washington, DC, USA.
Virkkala, R. (1993) Ranges of northern forest passerines: a fractal analysis. *Oikos*, **67**, 218–226.
Wilson, A., Silander, J., Gelfand, A. & Glenn, J. (2011) Scaling up: linking field data and remote sensing with a hierarchical model. *International Journal of Geographical Information Science*, **25**, 509–521.
Wright, D.H. (1991) Correlations between incidence and abundance are expected by chance. *Journal of Biogeography*, **18**, 463–466.

Received 27 July 2012; accepted 16 September 2012
Handling Editor: Robert Freckleton

Supporting Information

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

Fig. S1. Model performance of Models 1–6 when compared with the actual fine-grain presence/absence data.
Fig. S2. Relative differences in parameter estimation and predictive performance of the 5 downscaling techniques (Models 2–6) for the subset of 37 species with AUC > 0·85.

Fig. S3. Maps of species richness predicted by Models 1–6.

Fig. S4. Description of the difference in Models 1 and 6 in their predictions of species richness.

Appendix S1. BUGS code for Model 4.

Appendix S2. BUGS code for Models 5 and 6.