ECOLOGY LETTERS

Ecology Letters, (2017) **20**: 561–576 doi: 10.1111/ele.12757

IDEA AND PERSPECTIVE

How to make more out of community data? A conceptual framework and its implementation as models and software

Otso Ovaskainen,[1,2]*
Gleb Tikhonov,[1] Anna Norberg,[1]
F. Guillaume Blanchet,[3,4]
Leo Duan,[5] David Dunson,[5]
Tomas Roslin[6] and
Nerea Abrego[2,7]

Abstract

Community ecology aims to understand what factors determine the assembly and dynamics of species assemblages at different spatiotemporal scales. To facilitate the integration between conceptual and statistical approaches in community ecology, we propose Hierarchical Modelling of Species Communities (HMSC) as a general, flexible framework for modern analysis of community data. While non-manipulative data allow for only correlative and not causal inference, this framework facilitates the formulation of data-driven hypotheses regarding the processes that structure communities. We model environmental filtering by variation and covariation in the responses of individual species to the characteristics of their environment, with potential contingencies on species traits and phylogenetic relationships. We capture biotic assembly rules by species-to-species association matrices, which may be estimated at multiple spatial or temporal scales. We operationalise the HMSC framework as a hierarchical Bayesian joint species distribution model, and implement it as R- and Matlab-packages which enable computationally efficient analyses of large data sets. Armed with this tool, community ecologists can make sense of many types of data, including spatially explicit data and time-series data. We illustrate the use of this framework through a series of diverse ecological examples.

Keywords

Assembly process, biotic filtering, community distribution, community modelling, community similarity, environmental filtering, functional trait, joint species distribution model, metacommunity, phylogenetic signal.

Ecology Letters (2017) **20**: 561–576

INTRODUCTION

Ecology has been described as the scientific understanding of factors determining the abundance and distribution of species (Smith 1966; Begon *et al*. 1986). This understanding can hardly be achieved by studying species one by one since their abundances and distributions depend not only on their individual responses to the abiotic environment, but also on their interactions (Wisz *et al*. 2013). Thus, a key aim in modern community ecology is to gain an integrative understanding of how biotic and abiotic factors mould local species pools at different spatiotemporal scales.

Community ecology began as a descriptive science in which communities were classified based on the identities and sizes of local species pools (e.g. Clements 1936; Elton 1966). Modern community ecology is progressing from the description of patterns towards a mechanistic perspective, which seeks to understand the processes determining the identities and abundances of the species from local to global spatiotemporal scales (Agrawal *et al*. 2007; Logue *et al*. 2011). During the last few decades, experimental ecologist have used observations and experiments to assess the relative influences of stochasticity, competition and niche differentiation (see Logue *et al*. 2011), theoretical ecologists have developed models for predicting community dynamics (e.g. Tilman 1990, 2004; Holt *et al*. 1994; Bolker *et al*. 2003; Leibold *et al*. 2004; Holyoak *et al*. 2005), and statistical ecologists have developed metrics for assessing compositional changes among local communities (e.g. Gauch 1982; ter Braak & Prentice 1988; Legendre & Legendre 2012).

While a general theory to explain how communities are assembled across space and time is still lacking, community ecologists have converged towards a synthesis acknowledging that local species communities are a result of both stochastic and deterministic processes, henceforth called assembly processes (Gravel *et al*. 2006; Leibold & McPeek 2006; Stokes &

[1] *Department of Biosciences, University of Helsinki, P.O. Box 65, Helsinki FI-00014, Finland*

[2] *Department of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*

[3] *Department of Mathematics and Statistics, McMaster University, 1280 Main Street West Hamilton,Ontario L8S 4K1, Canada*

[4] *Département de biologie, Faculté des sciences, Université de Sherbrooke, 2500 Boulevard Université Sherbrooke, Québec J1K 2R1, Canada*

[5] *Department of Statistical Science, Duke University, P.O. Box 90251, Durham, USA*

[6] *Department of Ecology, Swedish University of Agricultural Sciences, Box 7044, Uppsala 75651,Sweden*

[7] *Department of Agricultural Sciences, University of Helsinki, P.O. Box 27, Helsinki FI-00014, Finland*

**Correspondence: E-mail: otso.ovaskainen@helsinki.fi*

Archer 2010; Weiher *et al.* 2011; Götzenberger *et al.* 2012). These forces encompass neutral processes, historical contingencies such as speciation, dispersal, abiotic environmental factors and biotic interactions (Vellend 2010; Weiher *et al.* 2011; Götzenberger *et al.* 2012).

As proposed by Zobel (1997) and illustrated in Fig. 1, the assembly processes can be envisaged as 'filters' operating at different scales. In this scheme, the 'global species pool' consists of all existing species, the 'regional pool' of all species able to colonise a given area, and the 'local species pool' of the set of species found at the finest scale considered (Cornell & Harrison 2014). Clearly, the species pools found at finer scales are filtered also by assembly processes acting at broader scales (Cornell & Lawton 1992). Of these hierarchical sieves, 'environmental filters' correspond to those abiotic factors which prevent the establishment or persistence of species in local communities (Kraft *et al.* 2015), and thus outline the fundamental niche of a species. 'Biotic filters' refer to interspecific and intraspecific competitive and facilitative interactions that determine the set of species in local communities (Wisz *et al.* 2013; Garnier *et al.* 2016), and thus determine their realised niches. These two types of forces may interact, as environmental filters may modify biotic interactions (e.g. Callaway & Walker 1997).

Beyond the deterministic processes selecting species from regional to local scales, stochastic processes create additional variation in the local communities. These processes – related to colonisation, extinction, ecological drift and environmental stochasticity – generate divergence among communities occupying identical environments (Chase & Myers 2011). The responses of the species to abiotic and biotic filters vary depending on species-specific characteristics known as the response traits, including e.g. their dispersal propensity and competitive abilities (Lavorel & Garnier 2002) (Fig. 1). Thus, they will determine what species reach and colonise given areas, and what species succeed in securing adequate resources (McGill *et al.* 2006; Bolnick *et al.* 2011).

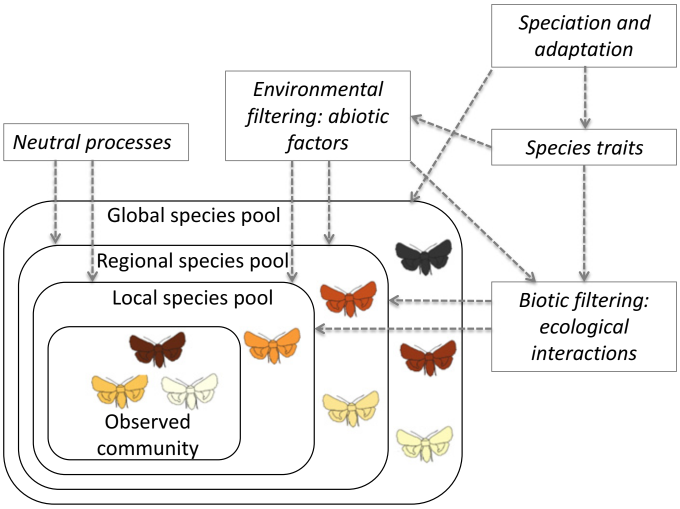

Figure 1 A conceptual diagram of the assembly processes influencing ecological communities at different spatiotemporal scales. The composition and dynamics of local, regional and global communities are influenced by the combined effects of environmental filters, biotic interactions and neutral processes. The responses of the species to these factors depend on their traits, which are ultimately shaped by evolutionary history and therefore constrained by phylogenetic relationships.

As a net outcome of the assembly processes outlined above, we find variation in the number, abundance, identities, and traits of the species present over a set of replicate communities observed in space and/or time (Fig. 2). While faced with a variety of data types, community ecologists have so far been armed with rather disparate statistical tools for connecting them with theories on community assembly. In particular, we lack a statistical frameworks that would enable us to infer actual assembly processes from community samples (Logue *et al.* 2011), and thus, a gap remains between theoretical predictions and data. Currently, the most popular tools used to study community structure are distance-based ordinations (e.g. Gauch 1982; ter Braak & Prentice 1988; Legendre & Legendre 2012) and diversity measures (see Magurran 2004). While such approaches provide insights into patterns of diversity and community composition at different spatiotemporal scales (Legendre *et al.* 2005; Dray *et al.* 2012; Legendre & Gauthier 2014), they provide little quantitative insight into the relative contributions of different assembly processes. To overcome these limitations, community ecologists are showing increasing interest in model-based approaches (Warton *et al.* 2015a,b).

Single-species distribution models have been widely used to explain and predict how different taxa respond to environmental variation (Guisan & Thuiller 2005; Elith & Leathwick 2009). To capitalise on this success, there is a growing interest in extending species distribution models to community-level models (Guisan & Rahbek 2011). One way of predicting community-level properties is simply to add predictions of single-species models into 'stacked' species distribution models (Guisan & Rahbek 2011; Calabrese *et al.* 2014). Another way is the use of joint species distribution models, which explicitly acknowledge the multivariate nature of species assemblages, allowing one to gather more mechanistic and predictive insights into assembly processes (Warton *et al.* 2015b). By simultaneously drawing on the information from multiple species, these models allow one to seek community-level patterns in how species respond to their environment (e.g. Ovaskainen & Soininen 2011; Wang *et al.* 2012; Hui *et al.* 2015; Ovaskainen *et al.* 2016b), to relate such patterns to species traits and phylogenies (Pollock *et al.* 2012; Brown *et al.* 2014; Abrego *et al.* 2017a) and to quantify co-occurrence patterns among species (Pollock *et al.* 2014; Ovaskainen *et al.* 2016a). Recent method development has made it possible to apply joint species distribution models e.g. to presence–absence data or abundance data (Hui *et al.* 2015; Clark *et al.* 2017), as well as to study designs of spatial (Thorson *et al.* 2015; Ovaskainen *et al.* 2016b), temporal (Sebastián-González *et al.* 2010) or spatio-temporal (Thorson *et al.* 2016) nature.

Owing to the ongoing revolution in sequencing technology, the development of statistical methodology for macro-organisms (reviewed above) has been paralleled by the rapid development of statistical methodology for microbial community

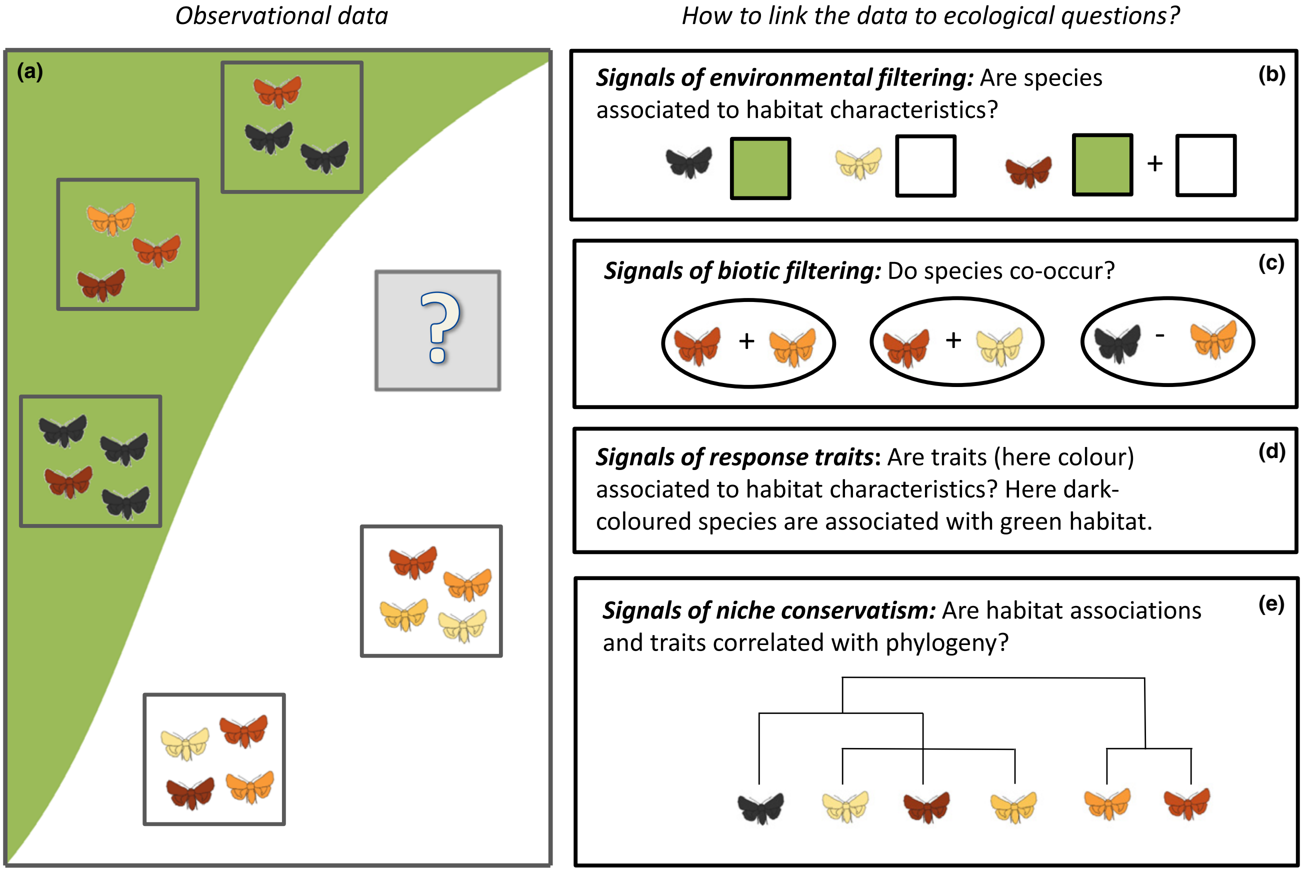
Observational data
How to link the data to ecological questions?
(a)
?
(b)
Signals of environmental filtering: Are species associated to habitat characteristics?
+
(c)
Signals of biotic filtering: Do species co-occur?
+
+
-
(d)
Signals of response traits: Are traits (here colour) associated to habitat characteristics? Here dark-coloured species are associated with green habitat.
(e)
Signals of niche conservatism: Are habitat associations and traits correlated with phylogeny?

Figure 2 A conceptual illustration of some key questions in community ecology. The green and white colours represent differences in the environmental conditions, the butterflies with different colours represent different species, and the small boxes represent sampling units. In this paper, we ask how species occurrence data (a) can be used to understand how the processes depicted in Fig. 1 structure the community under study, and how one can predict communities under new conditions (the sampling unit with question mark). In terms of species niches and environmental filtering (b), the black-coloured species only occurs under green environmental conditions and the lightest-coloured species under white environmental conditions, whereas the red coloured species occurs under both kind of environmental conditions. In terms of biotic interactions (c), some species pairs are found to co-occur more or less often by random. However, part of this variation can be explained simply by habitat associations, whereas in some cases there is non-random co-occurrence also beyond that explained by habitat association. In this example, the black-coloured species appears to dominate the local communities when present, indicating that it may be a competitively superior species. In terms of response traits, species colour appears to co-vary with environmental conditions, with the dark-coloured species being associated with the green habitat and the light-coloured species with the white habitat (d). As colour does not appear to be phylogenetically structured (e), the species provide essentially independent data points about the influence of colour on occurrence, suggesting an adaptive response rather than an artefact due to phylogenetic constraints.

ecology. Here, much of the focus has been on inferring association networks from sequence count data (Steele *et al.* 2011; Weiss *et al.* 2016), and on asking how such association networks relate to environmental conditions or ecosystem processes (e.g. Guidi *et al.* 2016). An important challenge with the analysis of such data is that the information is in relative rather than in absolute sequence counts. If not properly accounted for, this feature will lead to spurious correlations (Friedman & Alm 2012). As is the case with macro-organisms, inferring association-networks from sparse data on species rich communities is one of the main methodological challenges in current microbial community ecology (Weiss *et al.* 2016).

The aim of this paper is to facilitate the integration of theoretical and empirical approaches in community ecology, and of methods previously developed separately for micro- and macro-organisms. We suggest that recent advances in joint species distribution models can be transformed into a general framework for modern statistical analysis of community data. As the framework is based on a hierarchical joint species distribution approach, we call it *Hierarchical Modelling of Species Communities* (HMSC). HMSC integrates much of the recent methodological progress on joint species distribution models, and provides a unified platform which we hope will facilitate further amalgamation. To help community ecologists select a specific model which fits the nature of their data and the aims of their study, we offer a practical guide: we first describe the typical data types collected by empirical community ecologists, subsequently introduce the statistical HMSC framework, and then show with examples how the framework can be used to address topical questions in community ecology. Finally, we discuss some limitations of the framework, as well as challenges to be targeted by future research.

A STATISTICAL FRAMEWORK FOR HIERARCHICAL MODELLING OF SPECIES COMMUNITIES (HMSC)

Typical community data include observations on the occurrence of species in a set of temporal and/or spatial replicates, henceforth called occurrence data and referred to as the **Y** matrix (Fig. 3). Depending on the study/experimental design, on our objectives and the subject organisms, the occurrence of the species can be recorded in various ways, such as through direct visual or audial encounters, indirect cues such as tracks or droppings, or molecular identification of environmental samples. The occurrence matrix may thus describe e.g. presences and absences of the species, species counts, a percentage covered by each species or an estimate of its biomass. The amount and nature of observation error also depends on the method used. Common problems of species surveys are imperfect detection, creating false negatives, and misidentification, creating false positives (Guillera-Arroita 2017).

The occurrence data are usually accompanied by environmental data consisting of a set of measured covariates that the ecologist hypothesised to be important in explaining community composition (**X** matrix, Fig. 3). Beyond the effects of these environmental covariates, the spatiotemporal context may generate a structure to the data. In studies where the data have been collected in a hierarchical way (e.g. plots within sites), we call the finest scale (i.e., each row of the data matrices **X** and **Y**) the 'sampling unit'. In studies treating space and/or time as continuous, the study design may be described by spatial or temporal coordinates.

To relate community-level responses to environmental variation to response traits, one may wish to include data on species-specific traits (**T** matrix, Fig. 3). These data may range from morphological traits such as body size, or physiological traits such as tolerance to salinity, to functional traits such as feeding type, or to the actual position of the species within the surrounding food web. Furthermore, we may combine trait data with phylogenetic data (**C** matrix, Fig. 3). The availability of phylogenetic data is rapidly increasing, allowing the construction of quantitative matrices of phylogenetic correlations within many organisms groups. Where quantitative phylogenies are lacking, data on taxonomic identity (at the level of genus, family, order, class, phylum…) can be used as a proxy of phylogenetic relatedness (but see Whitfeld *et al.* 2012).

The statistical HMSC framework is illustrated graphically in Fig. 4 and described in more detail below. We start by modelling the occurrence (e.g. presence–absence, count or biomass) of each species (denoted as j, where $j = 1, \ldots, m$) in each sampling unit (denoted as i, where $i = 1, \ldots, n$), i.e. the data summarised by matrix **Y** in Fig. 3. For this, we use a generalised linear model,

$$y_{ij} \sim D\left(L_{ij}, \sigma_j^2\right). \tag{1}$$

Here, D is a statistical distribution tailored to the kind of data being used, L_{ij} is the linear predictor, and σ_j^2 is a variance term, which is excluded for some distributions (e.g. probit or Poisson) and included for others (e.g. normal or over-dispersed Poisson).

The linear predictor L_{ij} is modelled with the help of fixed (F) and random (R) parts as $L_{ij} = L_{ij}^F + L_{ij}^R$. The fixed effects are modelled as the regression

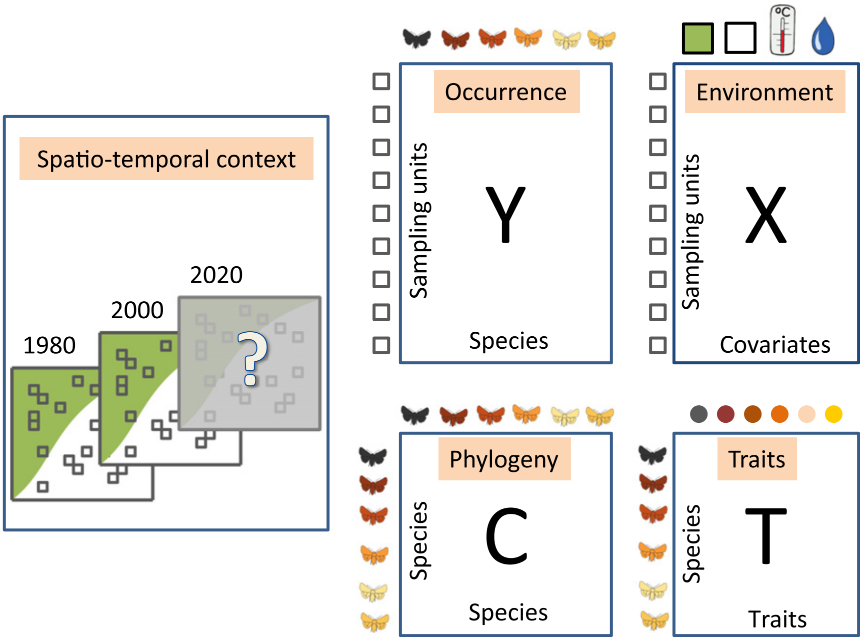

Figure 3 Data typically collected in community ecology. The occurrence data (denoted as the **Y** matrix) includes the occurrences of the species recorded in a set of temporal and/or spatial sampling units. The environmental data (denoted as the **X** matrix) consists of the environmental covariates measured over the sampling units. The traits data (denoted as the **T** matrix) consists of a set of traits measured for the species present in the **Y** matrix. To account for the phylogenetic dependencies among the species, we can include a fourth matrix consisting of the phylogenetic correlations among the species (denoted as the **C** matrix). The spatiotemporal context includes location and time information about the samples.

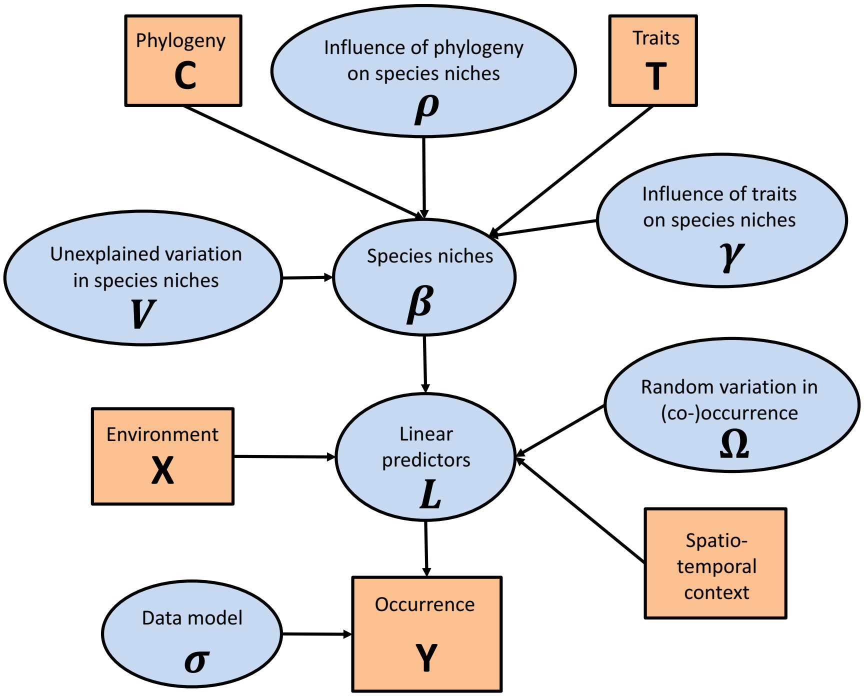
Phylogeny
C
Influence of phylogeny on species niches
ρ
Traits
T
Influence of traits on species niches
γ
Unexplained variation in species niches
V
Species niches
β
Random variation in (co-)occurrence
Ω
Environment
X
Linear predictors
L
Spatio-temporal context
Data model
σ
Occurrence
Y

Figure 4 A graphical summary of the HMSC statistical framework. In this Directed Acyclic Graph (DAG), the orange boxes refer to data, the blue ellipses to parameters to be estimated, and the arrows to functional relationships described with the help of statistical distributions.

$$L^F_{ij} = \sum_k x_{ik}\beta_{jk}, \tag{2}$$

where the term x_{ik} denotes the environmental covariate k measured at site i (i.e., the information summarised in matrix **X**, Fig. 3; with $x_{i1} = 1$ modelling the intercept), and the regression parameter β_{jk} denotes the response of species j to covariate k. As the regression parameters measure how the occurrence of the species depends on the environmental conditions, we interpret these as describing species' environmental niches. To allow the statistical framework to generate a community-level synthesis of how species respond to their environment, we assume that their responses to the environment (i.e. their regression parameters) adhere to a multivariate normal distribution,

$$\boldsymbol{\beta}_{j\cdot} \sim N\left(\boldsymbol{\mu}_{j\cdot}, \mathbf{V}\right). \tag{3}$$

We use a dot to single out a row or a column in a matrix, so that $\boldsymbol{\beta}_{j\cdot}$ denotes the vector of regression coefficients for species j. As $\boldsymbol{\beta}_{j\cdot}$ describes how species j responds to environmental covariates, it characterises its environmental niche. The expected environmental niche of species j is denoted by vector $\boldsymbol{\mu}_{j\cdot}$, and variation around this expectation is captured by the variance-covariance matrix **V** (Ovaskainen & Soininen 2011). The expected niche $\boldsymbol{\mu}_{j\cdot}$ can either be assumed to be the same for all species, or alternatively it can model the influence of species-specific traits on species' responses. In the latter case, we write $\mu_{jk} = \sum t_{jl}\gamma_{lk}$, where t_{jl} is the value of trait l for species j (matrix **T**, Fig. 3; with $t_{j1} = 1$ modelling the intercept) and the parameter γ_{lk} measures the effect of trait l on response to covariate k (Abrego *et al.* 2017a). We may use this model also to ask what percentage of variation in species' environmental niches can be attributed to species' traits (see Supporting Information for details).

To account for phylogenetic relationships (summarised by matrix **C**, Fig. 3), we model the covariance structure of the multivariate normal distribution as

$$\boldsymbol{\beta}_{\cdot\cdot} \sim N(\boldsymbol{\mu}_{\cdot\cdot}, \mathbf{V} \otimes [\rho\mathbf{C} + (\mathbf{1} - \rho)\mathbf{I}]), \tag{4}$$

where the symbol ⊗ stands for the Kronecker product and $0 \leq \rho \leq 1$ measures the strength of the phylogenetic signal. From eqn 4 it follows that for $\rho = 0$ the residual variance is independent among the species (as described by the identity matrix **I**), implying that closely related species do not have more similar environmental niches than do distantly related ones. When ρ approaches $\rho = 1$, species' environmental niches are fully structured by their phylogeny, with related species having more similar niches than expected by random, implying niche conservatism.

Let us then turn to the random term L^R_{ij}, which models the variation in species occurrences and co-occurrences that cannot be attributed to the responses of the species to the measured covariates. If the study design consists of sampling units without any hierarchical, spatial or temporal structure, L^R_{ij} will simply be $L^R_{ij} = \varepsilon^S_{ij}$, where the superscript S refers to a random effect ε that operates at the level of the sampling unit. These random effects are modelled as $\varepsilon^S_{i\cdot} \sim N(0, \boldsymbol{\Omega}^{\mathrm{S}})$, where $\boldsymbol{\Omega}^{\mathrm{S}}$ is a residual species-to-species variance-covariance matrix. Here, the word 'residual' refers to the fact that we have removed the influences of environmental covariates by the fixed effect part of the model. The diagonal element $\boldsymbol{\Omega}^{\mathrm{S}}_{jj}$ describes the amount of random variation that species j shows at the level of the sampling unit, whereas the off-diagonal element $\Omega^{S}_{j_1 j_2}$ describes the amount of covariation among the two species j_1 and j_2. For hierarchical study designs, the model may include several association matrices Ω (see Ovaskainen *et al.* 2016a), each of which may have a spatial or temporal structure.

With m species, an association matrix Ω has $m(m+1)/2$ parameters, making its parameterisation with sparse data on species-rich communities challenging. To facilitate the estimation of such matrices, we use a latent variable approach, which allows a parameter-sparse representation of the matrix Ω through latent factors and their loadings (for mathematical details see Warton *et al.* 2015b; Ovaskainen *et al.* 2016a,b). The factor loadings themselves do not have a straightforward interpretation in terms of ecological interactions, but are useful in revealing patterns where two species either co-occur more often than expected by chance: if the factor loadings have the same sign, the species respond to the latent variable in the 'same' way and thus increase in concert, whereas if the factor loadings have opposite signs, the species respond in 'different' ways, and thus one species declines when the other increases. We represent the species-to-species association network by the correlation matrix **R** defined by $R_{j_1 j_2} = \Omega_{j_1 j_2} / \sqrt{\Omega_{j_1 j_1} \Omega_{j_2 j_2}}$. The correlation $R_{j_1 j_2}$ measures to what extent species j_1 and j_2 are found together more or less often than expected by chance, *after* controlling for the environmental covariates. We note that one could alternatively measure species-to-species associations based on the precision matrix (inverse of Ω), which is more likely to identify direct links among species than the correlation matrix, as the latter one is also influenced by indirect links (Biswas *et al.* 2016; Ovaskainen *et al.* 2016a). As a further alternative, we note that instead of the latent variable approach, the correlation matrix could be parameterised through a mixture modelling approach (Pledger & Arnold 2014).

Let us note at this point that the much-used multivariate ordination approaches (Legendre & Legendre 2012) are also based on patterns of species co-occurrences. If we exclude the environmental covariates **X** from the analysis, then the latent variables behind an association matrix can be viewed as a model-based ordination (Warton *et al.* 2015b). Species close to each other in the ordination space show positive co-occurrence in the association matrix, whereas species at the opposite ends of the ordination space show negative co-occurrence. If environmental covariates are included in the analysis, then an association matrix corresponds to a residual ordination, which describes those co-occurrences that cannot be explained by shared responses to environmental covariates (Hui *et al.* 2015; Warton *et al.* 2015b).

The mathematical structure of the current model provides a convenient mapping from environmental similarity to community similarity. If the environmental covariates for sites 1 and 2 are described by the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, then the covariance between species occurrences (at the level of the linear predictor) is given by

$$\mathrm{Cov}\big(\boldsymbol{L}(\boldsymbol{x}_1), \boldsymbol{L}(\boldsymbol{x}_2)\big) = \boldsymbol{x}_1^{\boldsymbol{T}}\big(\boldsymbol{\gamma}^{\boldsymbol{T}}\mathrm{Cov}(\mathbf{T})\boldsymbol{\gamma} + (1 - \rho m(\mathbf{C}))\mathbf{V}\big)\boldsymbol{x}_2 \qquad (5)$$

where $m(\mathbf{C})$ denotes the mean value of the off-diagonal elements of the matrix **C** (see Supporting Information for the derivation of eqn 5). In eqn 5, the terms $\boldsymbol{\gamma}^{T}\mathrm{Cov}(\mathbf{T})\boldsymbol{\gamma}$ and (1-ρm(**C**))**V** partition interspecific variation in responses to environmental covariates to components that can and cannot be explained by the measured traits, respectively.

Let us finally note that, as with any statistical model, it is important to validate the above-described model in terms of its structural assumptions as well as the generality of the parameter estimates. In particular, to exclude the possibility of spurious inference due to model overfitting, we recommend examining the model's predictive power through a cross-validation approach, as we will illustrate below in some of the examples.

In the Supporting Information, we provide a Bayesian approach for parameterising the model in a computationally efficient manner which allows the analyses of large data sets. We also provide an implementation of the statistical framework as R- and Matlab-packages, which include a user manual.

APPLYING HMSC TO TOPICAL QUESTIONS IN COMMUNITY ECOLOGY

The key value of the modelling framework laid out above (Fig. 4) is in how its different components relate to processes of community assembly (Figs 1 and 2). To illustrate how they can be extracted from real data sets, and how they can be interpreted in relation to key questions in community ecology, we next turn to three contrasting case studies. These studies were specifically chosen to bring out the different strengths and uses of HMSC. The design of the first study is spatially hierarchical, the second study is spatially explicit, and the third study involves time-series data. The first case study also illustrates the utility of HMSC in addressing applied questions. To illustrate the comprehensiveness of the HSMC approach, we frame our treatise as a series of topical questions in community ecology which relate to the general theory outlined in the Introduction as well as a series of applied questions. These questions are summarised in Table 1, which also describes how HMSC can been applied to derive an answer. We note that this list is not exhaustive, and that the line between fundamental and applied questions in Table 1 is somewhat blurred. Furthermore, we remind the reader that the statistical framework of HMSC is correlative, and that the possibility of confounding effects should always be considered. Most importantly, the framework assumes (1) that environmental filtering is properly captured by the environmental covariates included in the model, (2) that biotic filtering is properly captured by residual species-to-species correlation matrices, or – in the case of time-series data – by matrices modelling the influence of each species in the previous year on the occurrence of each other species in the following year, (3) that any residual variation (i.e., variation not predicted by environmental or biotic filtering) can be attributed to random processes, such as dispersal limitation, environmental stochasticity or ecological drift. In the Discussion, we return to these assumptions and limitations of the modelling framework, and describe future challenges for its further development.

While all our examples are based on published studies, the novelty of this section is in illustrating how a wide range of questions and data types can be analysed with the help of the encompassing statistical framework of HMSC. More

Table 1 A summary of topical questions in community ecology and an outline of how they can be addressed within the HMSC framework

	Question	How to address the question statistically?	Illustration
FQ1	How much variation in species occurrence is due to environmental filtering, biotic interactions and random processes, and how do these impacts vary across spatial and temporal scales?	By assessing the explanatory power of models and by variance partitioning among fixed and random effects at different scales (see Supporting Information for details)	All case studies
FQ2	How do species' traits and phylogenetic relationships correlate with ecological niches?	By modelling responses to environmental covariates ($\boldsymbol{\beta}$) as a function of species' traits (**T**) and phylogenetic correlations (**C**)	Bryophyte, butterfly and fungal case studies
FQ3	What are the structures of species interaction networks?	By estimating the species-to-species association matrices Ω or **A**	All case studies
FQ4	How does community similarity depend on environmental similarity and/or geographic distance?	By decomposing community similarity into similarity due to responses to environmental covariates and/or spatial covariance, eqn 5	Bryophyte and butterfly case studies
FQ5	How does community structure change over time due to predictable succession or stochastic ecological drift?	By including time since environmental perturbation as a predictor, or by including temporally varying random effects	Bryophyte and water bird case studies
AQ1	Do some species indicate the presence of others?	By testing how much the predictive power of the model increases for a focal species when accounting for the occurrences of other species	Butterfly and fungal case studies
AQ2	How can geographic areas be classified into communities of common profile?	By clustering predicted communities based on their similarity	Butterfly case study
AQ3	Which processes have been central in determining the response of a community to environmental change	By decomposing the response to environmental change to components related to species niches and random effects	Bryophyte case study
AQ4	How can species be classified in terms of their response to abiotic environment?	By clustering parameters or predictions measuring the species responses to environmental covariates	Bryophyte case study
AQ5	How is community structure predicted to change under various scenarios of e.g. environmental change	By use of scenario simulations	Bryophyte case study

All questions relate to variation in the number, abundance, identities and traits of the species present over a set of replicate communities observed in space and/or time (Fig. 2). The questions have been grouped into fundamental questions in basic science (FQ1–5) and questions of applied interest (AQ1–5), though we note that these two categories overlap. See text for further details. The fungal case study is presented in Supporting Information.

information on all case studies, including on their ecological context, can be found in the primary publications referred to below.

A spatially hierarchical study design: how do epiphytic bryophytes respond to forest management?

As community ecologists, a fundamental question to ask is what processes create observed variation in community structure (Fig. 2): what proportion of variance in species occurrence can be attributed to environmental filtering, biotic filtering and random processes, respectively (Table 1). If replicate samples are available in space or time, we may further ask how the relative importance of different assembly processes change over spatial or temporal scales. In general, biotic interactions are likely to be more important at the finest scales, where species physically meet, whereas environmental covariates (such as macroclimate and soil types) are likely to be more important at broader scales (Araújo & Rozenfeld 2014).

In a study of epiphytic bryophytes, the species were surveyed on aspen trees on a set of forest sites (Fig. 5a, for original study see Oldén *et al.* 2014). The fitted model explained 10% of species occurrences at the tree level (measured by Tjur's R^2 (Tjur 2009), averaged over species) and 60% at the site level (measured as correlation in abundance, averaged over species). Thus, while it is difficult to predict which species are present or absent on a particular tree, more predictable patterns emerge at the level of the forest site. With the intent of identifying the strength of different processes shaping the local communities, we partitioned the explained variation into components attributable to the environmental covariates vs. variation assigned to random effects operating at different spatial scales (Fig. 5b; see Supporting Information for how this was technically done). Forest management type and age had a major influence on epiphytic bryophyte composition, as on average almost half of the variation was attributed to these forest site-level covariates (Fig. 5b). These environmental aspects proved to be the main agents in structuring epiphytic bryophyte assemblages at the level of forests, since once the effects of these covariates were accounted for, the species no longer showed strong associative patterns (Fig. 5c; forest site level). In contrast, the covariate measured at the tree level (tree diameter), explained only about half of the variation (Fig. 5b; fixed vs. random effects at the tree level), and thus much of the community variation was captured by the random effects (Fig. 5c; tree level).

Species vary in their morphology, life-history, behaviour and other traits. Relating this variation to community-level responses is a key challenge in community ecology (Weiher *et al.* 2011) (Table 1, Fig. 2). Thus, we may wish to address the fourth corner problem (Legendre *et al.* 1997), i.e. to ask

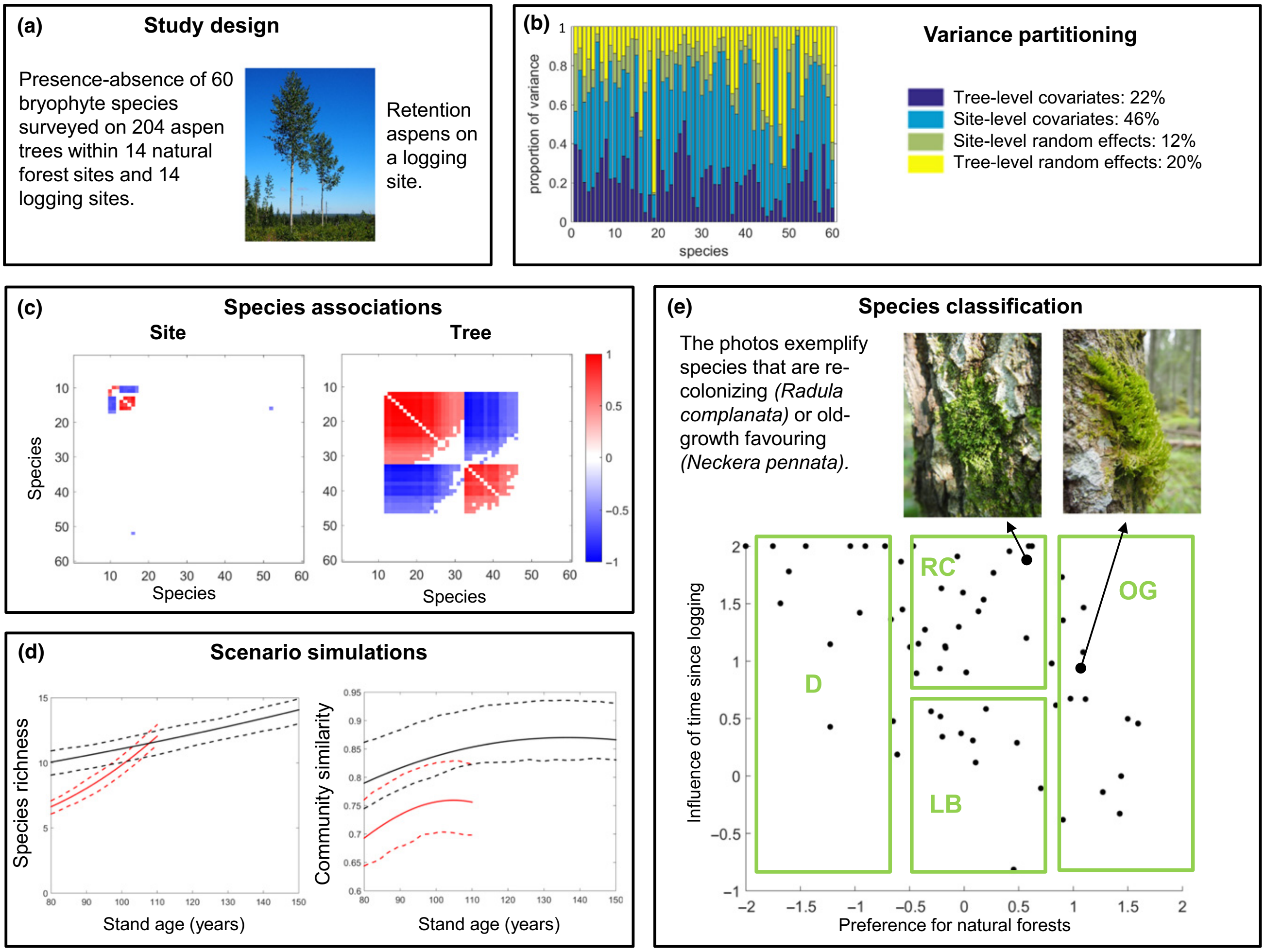

Figure 5 A case study on bryophytes, showing how the HMSC statistical framework can be applied to a spatially hierarchical study design and to questions related to biodiversity-oriented forest management. Panel (a) describes the study design. Panel (b) shows results on variance partitioning. Variation in species occurrence is partitioned into responses to covariates measured at the level of the site (natural or logged, stand age for natural and time since logging for logged sites) and the tree (diameter), as well as to random effects at these two levels. The bar-plot shows species-specific results (species ordered according to increasing prevalence so that species 1 is the rarest one) whereas the legend shows averages over species. Panel (c) shows estimates of species associations measured by residual correlation. Species-to-species association matrices identify species pairs showing a positive (red) or negative (blue) association, shown only if association has either sign with at least 75% posterior probability (the remaining cases are shown by white). The species have been ordered in a way that emphasizes the network structure. (Panel d) shows the results from scenario simulations. Here, we generated predictions for an aspen tree with a diameter of 30 cm, growing in a forest with a stand age of 80 years. We then considered two scenarios: one where the site was logged so that the aspen became a retention tree (red lines), and the other where it remained within the forest (black lines). We assumed that the aspen grows linearly so that its diameter reaches 60 cm in 70 years. The solid (dashed) lines show the mean (interquartile range) prediction for species richness and the similarity to a reference community, which we defined as the community predicted to occur on a 60 cm aspen in a natural forest with an age of 150 years. Panel (e) classifies the species into categories based on their responses to forest management, quantified with the help of the scenario simulations shown in panel (d). The *x*-axis reflects how much more likely the species is to be found in a natural forest of stand age 110 years than in a retention site 30 years after logging a forest of stand age 80 years. The y-axis shows how much more likely the species is to be found at a retention site 30 years after logging than immediately after logging (i.e. how quickly its occurrence probability recovers after logging). Unit: log-transformed odds ratio, absolute value truncated to two if greater than that. The boxes illustrate a classification of species into old growth forest specialists (OG), disturbance specialists (D), and species benefitting from retention trees through life-boating (LB) or re-colonization (RC).

to what extent and in which way some traits influence the responses of the community to environmental variation. In such an analysis, each species is essentially one data point. Due to phylogenetic relationships, related species are expected to be similar both concerning their traits and occurrence patterns, making the data points not independent (Harvey & Pagel 1991) (Fig. 2). Thus, to ask about the adaptive significance of traits, it is necessary to account for phylogenetic constraints. Accounting for phylogenetic signal can also yield further evolutionary insights, as it allows one to examine niche evolution by asking whether related species have more similar niches than expected by chance (Warren *et al.* 2008).

To assess how the traits of the bryophytes influenced their responses to the environmental covariates, we classified species by their life-form, a trait which has been previously suggested to be of functional importance (Bates 1998). However, this trait explained only 9% of the variation among species' environmental niches. The analysis yielded no strong support for the hypothesis of niche conservatism, with the posterior mean of the phylogenetic correlation parameter ρ being 0.26 (with a 95% credible interval from 0 to 0.77). These results suggest that in this particular study, the niches of bryophytes are to a limited extent related to their life-form, but mainly structured by some other traits that are not strongly correlated with their phylogeny, pointing out a knowledge gap in their functional ecology.

To determine how biotic filters structure communities, the first step is to know which species interact with each other and how, i.e. to determine the structure of the network of interspecific interactions (Morales-Castilla *et al.* 2015) (Table 1). Such an interaction network can be used, e.g. to examine whether the links among species within the community are organised in modular or nested ways (Fortuna *et al.* 2010). While non-manipulative data on species occurrence do not allow conclusive inference on ecological interactions, the HMSC framework makes it possible to infer residual species-to-species associations, which can be considered as hypotheses of such interactions. To examine how epiphytic bryophytes respond to each other, we included community-level random effects at the two hierarchical spatial scales included in our study (tree and site), and consequently estimated the species-to-species association networks at each of these scales (Fig. 5c). The resultant networks showed clear structure especially at the tree scale, where the species sorted essentially into three groups. The first group of species (species 11–31 in Fig. 5c) tend to occur together but not with species from the second group (species 32–45); the second group of species tend to occur together but not with species from the first group, and the third group of species (the remaining species) occur essentially independent of the other species. The estimated association networks differed between the two spatial scales, suggesting scale-dependent assembly mechanisms.

For applied ecologists dealing with the management or conservation of ecological communities, there is a need to identify the processes behind community responses to environmental change (Table 1). For example, a delayed response to habitat loss may indicate an extinction debt, whereas species lacking from areas with a suitable climate may indicate dispersal limitation and thus a colonisation credit (Hanski 2000). Another primary goal of applied community ecology is to classify species in relation to their vulnerability to environmental changes such as climate warming, habitat fragmentation and pollution – as that allows one to identify species requiring special attention in conservation management (e.g. Pacifici *et al.* 2015). In this case study, we were interested in testing the potential of Green-Tree Retention cutting (GTR) for conserving epiphytic bryophytes. GTR is a modification of traditional clear-cutting, aimed at mitigating effects on biodiversity (Rosenvald & Lõhmus 2008). To evaluate its success, we fitted the statistical framework to data on epiphytic bryophytes growing on aspens found in natural forests of varying stand age, as well as on retention aspens in logged forest with variation in time since logging. The model-based approach then allowed us to predict the influence of GTR through scenario simulation (Fig. 5d). After logging, species richness is predicted to drop temporarily, but then to quickly recover to the level of uncut forests. However, community composition is unlikely to return to the state typical of old-growth forest, but rather to follow a different trajectory (Fig. 5d). Thus, while retention aspens in old loggings may host as many species as aspen trees in natural forests, the identities of these species are likely to differ.

Based on the predicted responses of the bryophyte species to forest type and age, we may identify species groups of relevance to forest management and restoration (Fig. 5e). Some species predominantly occur in natural forests and thus benefit only little from retention trees, whereas others are primarily found in logged forests, and can thus be considered as disturbance specialists of little concern for conservation. In between, there are two groups of species that benefit from retention trees: Some species (called recolonising species in Fig. 5e) appear on retention aspens only once forest has re-grown around the retention tree. Other species (called life-boated species in Fig. 5e) can persist on the retention trees immediately after logging.

A spatially explicit study design: classifying Great Britain based on butterfly communities

When comparing local communities, we may be interested in how community similarity depends on the similarity of environmental conditions, and on the distance between the communities (Soininen *et al.* 2007). In addition to repeating the questions already addressed with the bryophyte data, we next illustrate questions on community similarity with data from a butterfly survey conducted in Great Britain. From the larger survey, we utilise presence–absence data for 55 species acquired at a 10 × 10 km resolution (Fig. 6a, for original study see Ovaskainen *et al.* 2016b).

In a spatial study, the species occurrences can be expected to be spatially autocorrelated, with the similarity of observations decreasing with distance. To account for this possibility, we assumed for the butterfly data that the underlying latent factors have a spatial covariance structure, and consequently that also the species-to-species association matrix $\Omega(d)$ depends on the distance d between the sampling units (see Supporting Information for details). Thus, we assumed that species share the same set of spatially structured latent variables, but differ in their factor loadings, i.e. in their responses to the latent variables. The latent factors may represent hidden environmental covariates, and are thus also interesting *per se* (see Ovaskainen *et al.* 2016b for illustrations of such maps).

A partitioning of variance among fixed effects (Fig. 6c) revealed that on average, environmental filters related to climatic conditions were somewhat more important than those related to habitat characteristics in explaining species occurrences, but that species differed substantially in their individual responses. More than half of the variation explained was attributed to the spatially structured random effects, suggesting that location and distance have strong effects on patterns of species occurrence and co-occurrence. To examine how species' traits influence their responses to environmental factors and thus their distributions within Britain, we classified species into those that are habitat specialists vs. those that are not, and to those that are migrants vs. those that are not.

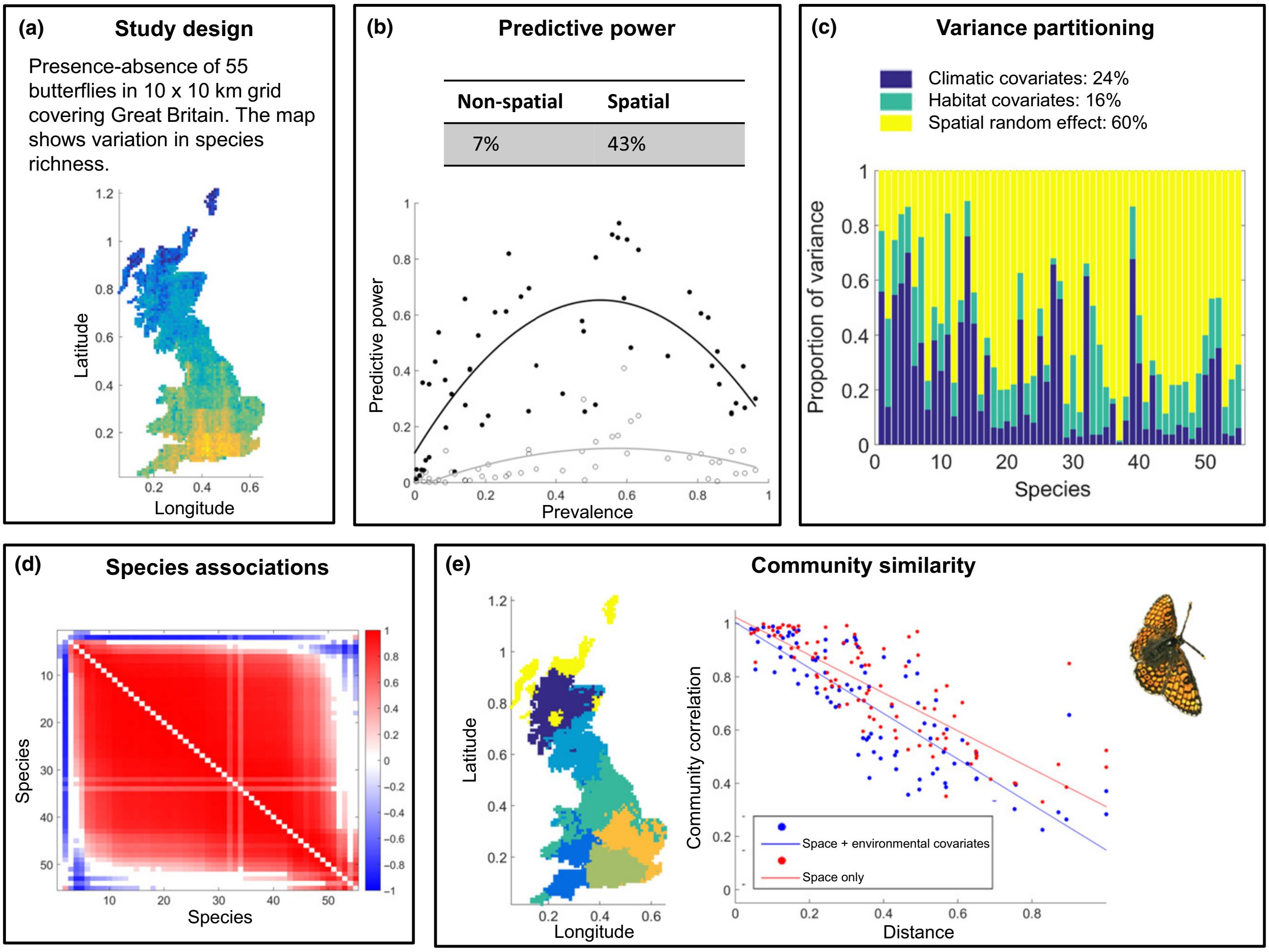

Figure 6 A case study on butterflies, illustrating the application of the HMSC statistical framework to a spatially explicit study design. Panel (a) describes the study design. Panel (b) shows analyses of predictive power, measured by Tjur (2009) R^2 for validation sites. We selected randomly 300 training sites and left the remaining 2309 sites for validation. The table shows averages over species, whereas the plots show species specific R^2 values against prevalence. Results are shown for models where the spatially structured latent factors underlying the estimation of species association matrices are excluded (Non-spatial; empty dots) or included (Spatial; black dots). Panel (c) shows results on variance partitioning, with metrics as in Fig. 5, and the covariates related to climatic and habitat variables. Panel (d) shows estimates on species associations, with contents as in Fig. 5. Panel (e) illustrates patterns of community similarity. The map classifies Great Britain into seven regions of common profiles, i.e. area with distinct butterfly communities, obtained by clustering communities based on their similarity (see Supporting Information for details). The plot shows how community similarity (eqn 5) decays with distance. The blue dots and line show the result based on the full model. For the red dots and line the environmental conditions (climate and habitat) have been standardized to their mean values, and thus they show the distance decay based on spatial variables only (see Supporting Information for details).

These categorical traits explained 19% of variation in species' environmental niches. As with the bryophyte study, there was no strong evidence for niche conservatism (posterior mean of $\rho = 0.41$, 95% credible interval 0.00–0.78). Association patterns among species were predominantly positive (Fig. 6d). This implies more variation in species richness than one would expect from independent occurrences, as species are either simultaneously present or absent from sampling units. Considering the general ecology of butterflies, where strong competition is unlikely but habitat specialisation frequent, these patterns are more likely due to shared responses to missing environmental covariates (or variation in sampling effort) than to true ecological interactions among them.

For applied purposes, we may wish to classify entire communities, to e.g. produce maps of distinct vegetation types, or more generally to define regions of common profile in terms of their community structure (Foster *et al.* 2013). Such regions can then be considered as management units, or be used in reserve selection (Margules & Pressey 2000). For the butterfly survey, regions characterised by similar communities (Foster *et al.* 2013) are shown in Fig. 6e, suggesting that British butterfly communities are primarily structured along the North–South gradient. To address how community similarity depends on the similarity of environmental conditions, and on the distance between the communities (Soininen *et al.* 2007), we may decompose the distance decay in community similarity into components that can and cannot be explained by similarity in environmental covariates (Fig. 6e). Importantly, the decay in similarity with distance gets steeper when we account for not only spatial distance but also for decreasing similarity in environmental conditions, again supporting the role of the environment in structuring these communities across space.

As applied ecologists, we may also wish to predict the community composition under different environmental conditions (Table 1), which may be either within the training data (interpolation) or outside it (extrapolation). For example, a scientist working with species distribution maps may wish to predict the ranges of all species based on the data points available (Guisan *et al.* 2013). The inclusion of spatially structured latent factors greatly improved the model's ability to predict butterfly communities at the validation sites (prediction for spatial model vs. prediction for non-spatial model in Fig. 6b), and thus of interpolating species distribution maps from sparse observations. This is because with spatial latent factors, the predictions are based both on measured environmental covariates and on interpolated species occurrences. Furthermore, as the spatial latent factors include information about species co-occurrence, the improved predictions not only account for the occurrences of the focal species in the nearby training sites, but also for the occurrences of all the other species at those sites.

In conservation biology, we frequently use indicator, umbrella and keystone species to indicate the presence of other species or of specific environmental conditions (Caro 2010). To identify such marker species, we should ask how well their occurrences predict the occurrences of target species or conditions. As shown in Fig. 6b (spatial vs. non-spatial prediction) and discussed above, the inclusion of other species in the model greatly increases the model's predictive power. As we may use any subset of the species as predictors and any other subset as response variables, the modelling framework permits the systematic identification of indicator species, i.e. species which are relatively common and easy to survey, but whose occurrence correlates with the occurrence of species of particular interest (Caro 2010).

A temporally explicit study design: heterospecific attraction in water birds

When replicate samples are available in time, we may also wish to address if and how local communities change, either through predictable successional pathways such as priority effects, or through stochastic ecological drift – both of which may either create new niches or result in extinctions (Dickie *et al.* 2012) (Table 1). Our third case study consists of a time-series containing data on birds' presence vs. absence across a set of 215 water ponds (Fig. 7a, for original study see Sebastián-González *et al.* 2010). Temporally replicated data provide an additional dimension for quantifying the extent to which communities are structured by biotic interactions, as we may use the occurrences of the species in the previous year as predictors of their occurrences in the following year. Such predictors can simply be included in the environmental covariate matrix **X** (Fig. 3), in which case their influences are estimated by the regression parameters $\boldsymbol{\beta}$. However, to separate species' responses to other species from species' responses to the shared environment, we denote by α_{jk} the influence of occurrence of species k in the previous year on the occurrence of species j in the present year. In this study, variation in species occurrence is explained almost equally much by biotic filters, i.e. by the occurrences of the species in the previous year, as by environmental filters, i.e., by the environmental covariates measured (Fig. 7b). Among the random effects, the most important is the one associated with the level of the sampling unit (a pond-year combination), representing heterospecific (i.e. between species) aggregation of the birds within the ponds.

To evaluate species associations over time, we examine the temporal association matrix $\mathbf{A} = (\alpha_{jk})$. As illustrated in Fig. 7c, this matrix is not necessarily symmetric, unlike the association networks measured by the covariance matrices Ω. The diagonal elements of the temporal association matrix **A** are positive, signalling that same species tend to return to the same ponds year after year. The estimated matrices also show strong support for the heterospecific attraction hypothesis (see review by Mönkkönen & Forsman 2002) both within and among years. Within a sampling unit, all species pairs are positively associated. Moreover, for many species pairs the occurrence of a species in one year increases the occurrence probability of another species in the next. For example, the Common shelduck is particularly likely to be found at the ponds where the Black-winged stilt was breeding in the previous year (Fig. 7c). Associations among species both between years and within a year may be due to unmeasured aspects of habitat quality or to the benefits that water birds may obtain from settling in sites already occupied by heterospecifics.

FUTURE DIRECTIONS AND CHALLENGES

As we have illustrated by the above three examples and a further example on wood-decaying fungi (see Supporting Information), Hierarchical Modelling of Species Communities (HMSC) offers a coherent framework for addressing a wide range of questions in community ecology, with its key advantages summarised in Box 1. However, as we discuss in this section, the core framework outlined here should be considered just a starting point, as it comes with important caveats and incitements for future work.

While conceptually our modelling framework focuses on the mechanisms behind community assembly (Figs 1–4), it is technically a hierarchical generalised linear mixed model. Hence, the results are of correlative nature, not directly implying causation. For example, we emphasise that species association networks should not be interpreted as proven ecological interactions, but that ecological interactions may have a major influence on the associations. As another example, niche or interaction-based processes do not necessarily always cause predictable community patterns (Munoz & Huneman 2016). These and other confounding issues are not a limitation of the statistical model, but of the nature of the data as such. For instance, data on species occurrence do not allow one to establish conclusively whether co-occurrences are due to ecological interactions or to other confounding factors, such as missing environmental covariates or sampling biases (Kissling *et al.* 2012). Only experimental data can bring conclusive evidence on this issue. Having said that, our modelling approach offers a particularly efficient avenue for deriving data-driven

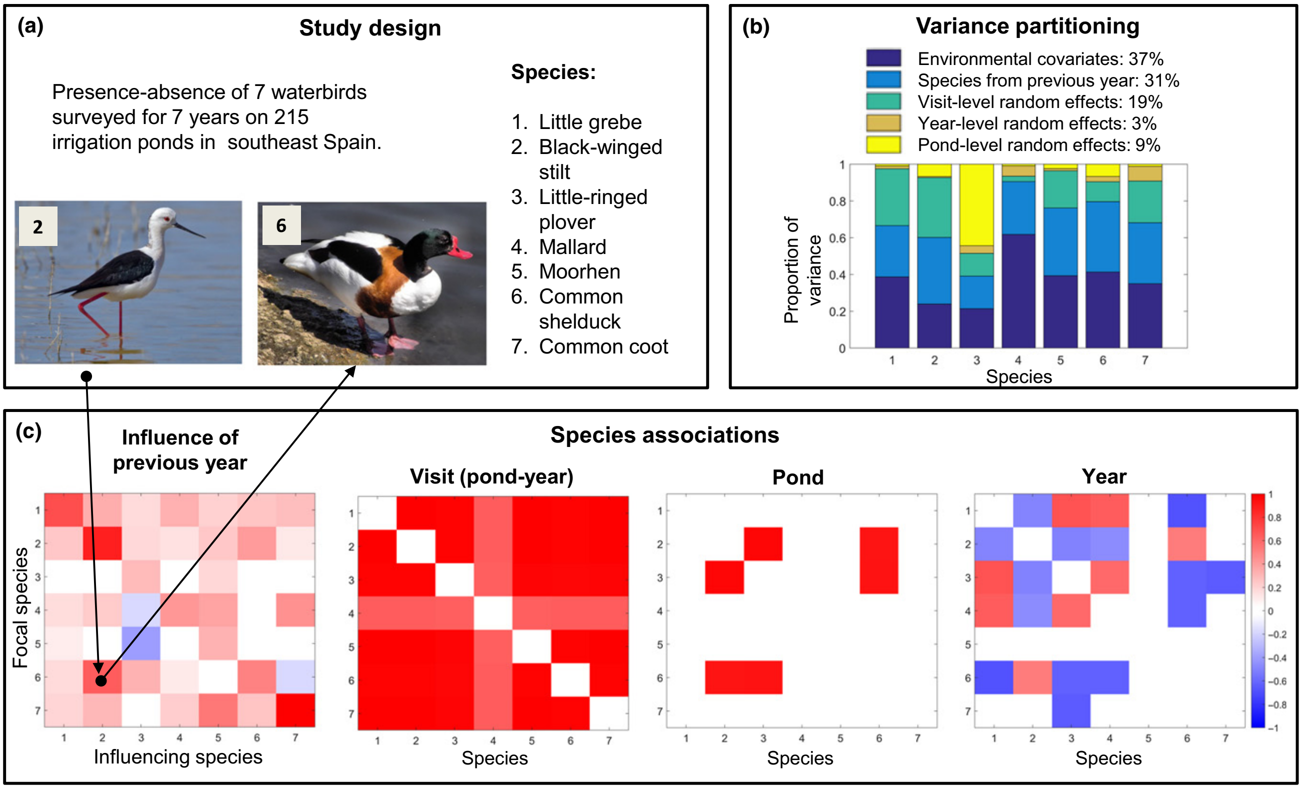

Figure 7 A case study on water birds, exemplifying the application of the HMSC statistical framework to time series data. Panel (a) describes the study design. Panel (b) shows results on variance partitioning, with metrics as in Fig. 5, except that species are not ordered according to prevalence (for order, see panel a). The environmental covariates relate to pond type and area, vegetation type, distance to natural wetlands, and connectivity. As fixed effects are also included the influence of the occurrence of each species in the previous year on species-specific occurrence in the current year. Random effects are included at the levels of year, pond and visit (pond-year combination). Panel (c) shows estimates of species associations. The matrix *influence of previous year* gives the regression coefficients measuring how the presence of a species in the previous year influences the occurrence of the focal species in the present year. The contents of the other species-to-species association matrices match those of Fig. 5, except that here, the species are shown in the same order across all panels (for order, see panel a).

hypotheses on ecological interactions, as it decomposes the observed co-occurrences into those that can be explained by the environmental covariates explicitly measured vs. those that cannot.

While the framework presented here already allows one to address some of the most fundamental questions in community ecology (Table 1), it should clearly be seen as a starting point for building the next generation of joint species distribution models. As we have based HMSC on hierarchical generalised linear mixed models, adding additional layers is both conceptually and technically straightforward. Below, we discuss four perspectives that we consider especially fruitful in this context.

The first perspective relates to the need to explicitly account for the observation process when modelling the data, thus allowing one to separate the processes of interest from biases introduced by the observer (Guillera-Arroita 2017; Beissinger *et al.* 2016; Warton *et al.* 2016). As the simplest solution, one may include covariates reflecting observational error (such as variation in sampling effort) in the **X** matrix. However, for many kinds of data there is a need to incorporate a more elaborate observation model, as is commonly done in e.g. single-species occupancy modelling (Guillera-Arroita 2017). Within the HMSC framework, the observed data $y_{ij}^{(\mathrm{obs})}$ could be modelled as a function of the underlying true occurrences y_{ij} as e.g. $y_{ij}^{(\mathrm{obs})} \sim D^{(\mathrm{obs})}\left(y_{ij}, \sigma_j^{2(\mathrm{obs})}\right)$, where the statistical distribution $D^{(\mathrm{obs})}$ describes the observation process. For example, much data on species distributions consist of presences only. Within the HMSC framework, such data could be analysed e.g. by linking the linear predictor to a spatially varying intensity of an underlying point process reflecting the sampling effort (Renner *et al.* 2015). As another example, sequence data that are commonly utilised in microbial community ecology are constrained by the total sequence count, which can create spurious correlations if not accounted for (Friedman & Alm 2012). Within the HMSC framework, such data could be modelled e.g. through multinomial logistic regression. With such an extension, HMSC would help merging the statistical methods developed thus far partly separately for micro-organisms and macro-organisms (see Introduction).

The second perspective relates to the need for more versatile treatments of association networks. One of the key strengths of the HMSC approach is that it allows us to estimate species association networks at different spatial or temporal scales, and to utilise the inferred associations in predictions and simulated scenarios. In the examples above, we have estimated

Box 1 Ten reasons for applying Hierarchical Modelling of Species Communities (HMSC)

(1) HMSC is a unifying framework which encompasses classic approaches such as single-species distribution models and model-based ordinations as special cases.
(2) HMSC provides simultaneous inferences at the species and community levels.
(3) HMSC offers the general advantages of model-based approaches, such as tools for model validation and prediction.
(4) HMSC overcomes previous problems of modelling communities with sparse data.
(5) HMSC overcomes the long-standing challenge in species distribution modelling of how to account for species interactions in explaining and predicting species occurrences.
(6) HMSC allows one to partition observed variation in species occurrences into components related to environmental variation measured vs. random processes (or unmeasured variation) at different spatial scales – both at the species and community levels.
(7) HMSC tackles the fourth corner problem (the influence of species traits to their occurrences, see Legendre *et al.* 1997) in a way which accounts for the phylogenetic signal in the data.
(8) HMSC can be applied to many kinds of study designs (including hierarchical, temporal or spatial) and many types of data (such as presence–absence, counts and continuous measurements).
(9) HMSC can generate predictions at the species, community or trait levels, while propagating uncertainty in parameter values to the level of the prediction.
(10) HMSC is computationally efficient, being able to analyse small data sets (with a few hundreds of sampling units and a few tens of species) in seconds, and large data sets (with some tens of thousands of sampling units and a few thousands of species) in few days.

the association matrices solely from the occurrence data. However, community ecologists will oftentimes hold interaction matrices constructed from other types of data, arising e.g. from experiments (Wootton & Emmerson 2005) or information on species traits which might influence species interactions (e.g. Schöb *et al.* 2013). In the present formulation of the model, we have assumed that species traits **T** and phylogenetic correlations **C** influence the species responses to the abiotic environment, but ignored the link from these matrices to co-occurrence. Finding the best statistical approaches to utilising such data in the HMSC framework remains a challenge. One possible solution would be to model the factor loadings underlying species co-occurrences as functions of the **T** and the **C** matrices in the same way that we have done for species niches described by the $\boldsymbol{\beta}$ parameters. Linking species co-occurrences to traits, phylogenies, and their responses to environmental variation would provide new tools for addressing several intriguing questions, e.g. whether species with few strong interactions are more vulnerable to disturbances than species with many weak interactions (Fortuna & Bascompte 2006; Aizen *et al.* 2012; Abrego *et al.* 2017b), or whether the extent to which prey species within communities are coupled by predators translate into correlated abundances in space or time (Morris *et al.* 2004; Tack *et al.* 2011; Kaartinen & Roslin 2013).

We also note that the estimated association networks are conditional on species occurrence: even if two species show a strong positive co-occurrence, the association is not realised in areas where neither of the species occurs. Less trivially, we have assumed the structure of association networks to be constant across all areas where they are realised. However, the type and strength of ecological interactions may be context-dependent (e.g. Poisot *et al.* 2015). For example, the stress-gradient hypothesis predicts that positive interactions are accentuated under stressful abiotic environmental conditions (Callaway & Walker 1997). In the context of HMSC, context dependence of the association matrices can be incorporated by modelling the underlying latent variable structure as a function of environmental covariates (Tikhonov *et al.* 2017), or more generally as functions of space or time.

The third perspective relates to the need of utilising the full potential of community-level time-series data. With our case study on water birds (Fig. 7), we illustrated time-series modelling in the context of presence–absence data – but note that abundance data are likely to yield much stronger signals on heterospecific interactions. With abundance data, one option for seeking for the effect of such interactions is to use Gompertz–type models, which account for the intraspecific interactions through density dependence (Mutshinda *et al.* 2009). One technical challenge when dealing with species rich communities is that interaction matrices between years have very high dimensionality. In close analogy with our approach to high-dimensional association matrices $\boldsymbol{\Omega}$, this problem could be solved by a latent variable approach. Here, the latent factors would correspond to community level summaries of species abundances, whereas factor loadings would model how the dynamics of individual species depend on those summaries.

The fourth and final perspective is a merger of genetic and evolutionary perspectives on community assembly. While the relevance of eco-evolutionary feedbacks is increasingly recognised in studies of single species, the quantification of such feedbacks in studies targeting the community level is still scarce (Johnson & Stinchcombe 2007). In the HMSC framework presented above, we have only touched upon evolutionary aspects by asking whether species' niches show a phylogenetic signal. An important challenge is to adopt a more micro-evolutionary perspective, e.g. by asking if and how the amount and type of genetic variation influences variation in species occurrence, either among species, or in space or through time. A related challenge is bringing the analyses to the individual level rather than operating only at the species level – as, for example, traits are often measured at the individual level (McGill *et al.* 2006). While the HMSC framework will in principle allow for this through its hierarchical structure, developing a general and computationally feasible approach that builds up communities from the individual level is not straightforward.

Having identified a need for integrating individual- and genetic-level perspectives with community ecology, we end by highlighting the status quo: for merging perspectives from genes, individuals and communities, the largest challenge is currently in the lack of appropriate data. But for the kind of community data illustrated in Fig. 3, we believe that the opposite is true: here we have data in ample supply, whereas progress has been hampered by a lack of integrative modelling frameworks to synthesise these data. We hope that the HMSC approach provides a partial solution to this problem, and that it inspires future research to overcome some of the caveats listed above.

ACKNOWLEDGEMENTS

We thank Anna Oldén for permit to use of the bryophyte data as well providing the bryophyte photographs (Fig. 5e), Ester Sebastián-González for permit to use the water bird data, Albert Phillimore for providing the butterfly phylogeny, and Andreas Trepte (www.photo-natur.de) for the photograph on Black-winged stilt (Fig. 7a). We are grateful for valuable comments provided by four anonymous referees. The research was funded by the Academy of Finland (CoE grant 284601 to OO and grant 285803 to TR), the Research Council of Norway (CoE grant 223257), the Research Foundation of the University of Helsinki (PhD grant to AN), and the LUOVA graduate school of the University of Helsinki (PhD grant for GT).

HMSC-software. The Matlab and R-versions of the HMSC-software, the user manual, and scripts for replicating the analyses presented here are found from https://www.helsinki.fi/en/researchgroups/metapopulation-research-centre/hmsc

AUTHORSHIP

OO, LD and DD developed the statistical methods. GT implemented the Matlab package, GB the R-package and AN, GT, GB, NA and OO wrote the user manual. OO, NA and TR wrote the first draft of the manuscript and all authors contributed substantially to the revisions.

REFERENCES

Abrego, N., Norberg, A. & Ovaskainen, O. (2017a). Measuring and predicting the influence of traits on the assembly processes of wood-inhabiting fungi. *J. Ecol.*, 10.1111/1365-2745.12722, in press.

Abrego, N., Dunson, D., Halme, P., Salcedo, I. & Ovaskainen, O. (2017b). Wood-inhabiting fungi with tight associations with other species have declined as a response to forest management. *Oikos*, 126, 269–275.

Agrawal, A.A., Ackerly, D.D., Adler, F., Arnold, A.E., Cáceres, C., Doak, D.F. *et al.* (2007). Filling key gaps in population and community ecology. *Front. Ecol. Env.*, 5, 145–152.

Aizen, M.A., Sabatino, M. & Tylianakis, J.M. (2012). Specialization and rarity predict nonrandom loss of interactions from mutualist networks. *Science*, 335, 1486–1489.

Araújo, M.B. & Rozenfeld, A. (2014). The geographic scaling of biotic interactions. *Ecography*, 37, 406–415.

Bates, J.W. (1998). Is 'life-form' a useful concept in bryophyte ecology? *Oikos*, 82, 223–237.

Begon, M., Harper, J.L. & Townsend, C.R. (1986). *Ecology: Individuals, Populations and Communities*, 1st edn. Blackwell Scientific Publications, Oxford, UK.

Beissinger, S.R., Iknayan, K.J., Guillera-Arroita, G., Zipkin, E.F., Dorazio, R.M., Royle, J.A. *et al.* (2016). Incorporating imperfect detection into joint models of communities: a response to Warton et al. *Trends Ecol. Evol.*, 31, 736–737.

Biswas, S., Mcdonald, M., Lundberg, D.S., Dangl, J.L. & Jojic, V. (2016). Learning microbial interaction networks from metagenomic count data. *J. Comput. Biol.*, 23, 526–535.

Bolker, B., Holyoak, M., Křivan, V., Rowe, L. & Schmitz, O. (2003). Connecting theoretical and empirical studies of trait-mediated interactions. *Ecology*, 84, 1101–1114.

Bolnick, D.I., Amarasekare, P., Araújo, M.S., Bürger, R., Levine, J.M., Novak, M. *et al.* (2011). Why intraspecific trait variation matters in community ecology. *Trends Ecol. Evol.*, 26, 183–192.

ter Braak, C.J.F. & Prentice, I.C. (1988). A theory of gradient analysis. *Adv. Ecol. Res.*, 18, 271–317.

Brown, A.M., Warton, D.I., Andrew, N.R., Binns, M., Cassis, G. & Gibb, H. (2014). The fourth-corner solution - using predictive models to understand how species traits interact with the environment. *Methods Ecol. Evol.*, 5, 344–352.

Calabrese, J.M., Certain, G., Kraan, C. & Dormann, C.F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecol. Biogeogr.*, 23, 99–112.

Callaway, R.M. & Walker, L.R. (1997). Competition and facilitation: a synthetic approach to interactions in plant communities. *Ecology*, 78, 1958–1965.

Caro, T.M. (2010). *Conservation by Proxy: indicator, Umbrella, Keystone, Flagship, and Other Surrogate Species*, 2nd edn. Island Press, Washington, Covelo, London.

Chase, J.M. & Myers, J.A. (2011). Disentangling the importance of ecological niches from stochastic processes across scales. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 366, 2351–2363.

Clark, J.S., Nemergut, D., Seyednasrollah, B., Turner, P.J. & Zhang, S. (2017). Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. *Ecol. Monogr.*, 84, 34–56.

Clements, F.E. (1936). Nature and structure of the climax. *J. Ecol.*, 24, 252–284.

Cornell, H.V. & Harrison, S.P. (2014). What are species pools and when are they important? *Annu. Rev. Ecol. Evol. Syst.*, 45, 45–67.

Cornell, H.V. & Lawton, J.H. (1992). Species interactions, local and regional processes, and limits to the richness of ecological communities: a theoretical perspective. *J. Anim. Ecol.*, 61, 1–12.

Dickie, I.A., Fukami, T., Wilkie, J.P., Allen, R.B. & Buchanan, P.K. (2012). Do assembly history effects attenuate from species to ecosystem properties? A field test with wood-inhabiting fungi. *Ecol. Lett.*, 15, 133–141.

Dray, S., Pélissier, R., Couteron, P., Fortin, M.J., Legendre, P., Peres-Neto, P. *et al.* (2012). Community ecology in the age of multivariate multiscale spatial analysis. *Ecol. Monogr.*, 82, 257–275.

Elith, J. & Leathwick, J.R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.*, 40, 677–697.

Elton, C. (1966). *The Patterns of Animal Communities*, 1st edn. Methuen; New York, John Wiley and Sons, London.

Fortuna, M.A. & Bascompte, J. (2006). Habitat loss and the structure of plant-animal mutualistic networks. *Ecol. Lett.*, 9, 281–286.

Fortuna, M.A., Stouffer, D.B., Olesen, J.M., Jordano, P., Mouillot, D., Krasnov, B.R. *et al.* (2010). Nestedness versus modularity in ecological networks: two sides of the same coin? *J. Anim. Ecol.*, 79, 811–817.

Foster, S.D., Givens, G.H., Dornan, G.J., Dunstan, P.K. & Darnell, R. (2013). Modelling biological regions from multi-species and environmental data. *Environmetrics*, 24, 489–499.

Friedman, J. & Alm, E.J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.*, 8, 1–11.

Garnier, E., Navas, M.-L. & Grigulis, K. (2016). A functional approach to plant community structure. In: *Plant Functional Diversity: organisms Traits, Community Structure and Ecosystem Properties* (eds. Garnier, E. & Navas, M-.L. & Grigulis, K.). Oxford University Press, Oxford, UK, pp. 94–118.

Gauch, H.G. (1982). *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge.

Götzenberger, L., de Bello, F., Bråthen, K.A., Davison, J., Dubuis, A., Guisan, A. *et al.* (2012). Ecological assembly rules in plant communities-approaches, patterns and prospects. *Biol. Rev.*, 87, 111–127.

Gravel, D., Canham, C.D., Beaudet, M. & Messier, C. (2006). Reconciling niche and neutrality: the continuum hypothesis. *Ecol. Lett.*, 9, 399–409.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S. *et al.* (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532, 465–470.

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography*, 40, 281–295.

Guisan, A. & Rahbek, C. (2011). SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *J. Biogeogr.*, 38, 1433–1444.

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

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T. *et al.* (2013). Predicting species distributions for conservation decisions. *Ecol. Lett.*, 16, 1424–1435.

Hanski, I. (2000). Extinction debt and species credit in boreal forests: modelling the consequences of different approaches to biodiversity conservation. *Ann. Zool. Fenn.*, 37, 271–280.

Harvey, P.H. & Pagel, M.D. (1991). *The Comparative Method in Evolutionary Biology*, 1st edn. Oxford University Press, Oxford, UK.

Holt, R.D., Grover, J. & Tilman, D. (1994). Simple rules for interspecific dominance in systems with exploitative and apparent competition. *Am. Nat.*, 144, 741–771.

Holyoak, M., Leibold, M.A. & Holt, R.D. (2005). *Metacommunities: Spatial Dynamics and Ecological Communities*, 1st edn. The University of Chicago Press, Chicago.

Hui, F.K.C., Taskinen, S., Pledger, S., Foster, S.D. & Warton, D.I. (2015). Model-based approaches to unconstrained ordination. *Methods Ecol. Evol.*, 6, 399–411.

Johnson, M.T.J. & Stinchcombe, J.R. (2007). An emerging synthesis between community ecology and evolutionary biology. *Trends Ecol. Evol.*, 22, 250–257.

Kaartinen, R. & Roslin, T. (2013). Apparent competition leaves no detectable imprint on patterns of community composition: observations from a natural experiment. *Ecol. Entomol.*, 38, 522–530.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J. *et al.* (2012). Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *J. Biogeogr.*, 39, 2163–2178.

Kraft, N.J.B., Adler, P.B., Godoy, O., James, E.C., Fuller, S. & Levine, J.M. (2015). Community assembly, coexistence and the environmental filtering metaphor. *Funct. Ecol.*, 29, 592–599.

Lavorel, S. & Garnier, E. (2002). Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the holy grail. *Funct. Ecol.*, 16, 545–556.

Legendre, P. & Gauthier, O. (2014). Statistical methods for temporal and space–time analysis of community composition data. *Proc. R. Soc. Lond. B Biol. Sci.*, 281, 20132728.

Legendre, P. & Legendre, L. (2012). *Numerical Ecology*. 3rd edn. Elsevier, 281, 20132728.

Legendre, P., Galzin, R. & Harmelin-Vivien, M.L. (1997). Relating behavior to habitat: solutions to the fourth-corner problem. *Ecology*, 78, 547–562.

Legendre, P., Borcard, D. & Peres-Neto, P. (2005). Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol. Monogr.*, 75, 435–450.

Leibold, M.A. & McPeek, M.A. (2006). Coexistence of the niche and neutral perspectives in community ecology. *Ecology*, 87, 1399–1410.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F. *et al.* (2004). The metacommunity concept: a framework for multi-scale community ecology. *Ecol. Lett.*, 7, 601–613.

Logue, J.B., Mouquet, N., Peter, H. & Hillebrand, H. (2011). Empirical approaches to metacommunities: a review and comparison with theory. *Trends Ecol. Evol.*, 26, 482–491.

Magurran, A.E. (2004). *Measuring Biological Diversity*. 1st edn. Willey-Blackwell, Oxford, UK.

Margules, C.R. & Pressey, R.L. (2000). Systematic conservation planning. *Nature*, 405, 243–253.

McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends Ecol. Evol.*, 21, 178–185.

Mönkkönen, M. & Forsman, J.T. (2002). Heterospecific attraction among forest birds: a review. *Ornithol. Sci.*, 1, 41–51.

Morales-Castilla, I., Matias, M.G., Gravel, D. & Araújo, M.B. (2015). Inferring biotic interactions from proxies. *Trends Ecol. Evol.*, 30, 347–356.

Morris, R.J., Lewis, O.T. & Godfray, H.C. (2004). Experimental evidence for apparent competition in a tropical forest food web. *Nature*, 428, 310–313.

Munoz, F. & Huneman, P. (2016). From the neutral theory to a comprehensive and multiscale theory of ecological equivalence. *Q. Rev. Biol.*, 91, 321–342.

Mutshinda, C.M., O'Hara, R.B. & Woiwod, I.P. (2009). What drives community dynamics? *Proc. R. Soc. Lond. B Biol. Sci.*, 276, 2923–2929.

Oldén, A., Ovaskainen, O., Kotiaho, J.S., Laaka-Lindberg, S. & Halme, P. (2014). Bryophyte species richness on retention aspens recovers in time but community structure does not. *PLoS ONE*, 9, e93786.

Ovaskainen, O. & Soininen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, 92, 289–295.

Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. (2016a). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.*, 7, 549–555.

Ovaskainen, O., Roy, D.B., Fox, R. & Anderson, B.J. (2016b). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods Ecol. Evol.*, 7, 428–436.

Pacifici, M., Foden, W.B., Visconti, P., Watson, J.E.M., Butchart, S.H.M., Kovacs, K.M. *et al.* (2015). Assessing species vulnerability to climate change. *Nature Clim. Change*, 5, 215–224.

Pledger, S. & Arnold, R. (2014). Multivariate methods using mixtures: correspondence analysis, scaling and pattern-detection. *Comput. Stat. Data Anal.*, 71, 241–261.

Poisot, T., Stouffer, D.B. & Gravel, D. (2015). Beyond species: why ecological interaction networks vary through space and time. *Oikos*, 124, 243–251.

Pollock, L.J., Morris, W.K. & Vesk, P.A. (2012). The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, 35, 716–725.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M. *et al.* (2014). Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Methods Ecol. Evol.*, 5, 397–406.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J. *et al.* (2015). Point process models for presence-only analysis. *Methods Ecol. Evol.*, 6, 366–379.

Rosenvald, R. & Lõhmus, A. (2008). For what, when, and where is green-tree retention better than clear-cutting? A review of the biodiversity aspects. *For. Ecol. Manage.*, 255, 1–15.

Schöb, C., Armas, C., Guler, M., Prieto, I. & Pugnaire, F.I. (2013). Variability in functional traits mediates plant interactions along stress gradients. *J. Ecol.*, 101, 753–762.

Sebastián-González, E., Sánchez-Zapata, J.A., Botella, F. & Ovaskainen, O. (2010). Testing the heterospecific attraction hypothesis with time-series data on species co-occurrence. *Proc. R. Soc. Lond. B Biol. Sci.*, 277, 2983–2990.

Smith, R.L. (1966). *Ecology and Field Biology*, 1st edn. Harper & Row, New York, USA.

Soininen, J., McDonald, R. & Hillebrand, H. (2007). The distance decay of similarity in ecological communities. *Ecography*, 30, 3–12.

Steele, J.A., Countway, P.D., Xia, L., Vigil, P.D., Beman, J.M., Kim, D.Y. *et al.* (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J.*, 5, 1414–1425.

Stokes, C.J. & Archer, S.R. (2010). Niche differentiation and neutral theory: an integrated perspective on shrub assemblages in a parkland savanna. *Ecology*, 91, 1152–1162.

Tack, A.J.M., Gripenberg, S. & Roslin, T. (2011). Can we predict indirect interactions from quantitative food webs? - an experimental approach. *J. Anim. Ecol.*, 80, 108–118.

Thorson, J.T., Scheuerell, M.D., Shelton, A.O., See, K.E., Skaug, H.J. & Kristensen, K. (2015). Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods Ecol. Evol.*, 6, 627–637.

Thorson, J.T., Ianelli, J.N., Larsen, E.A., Ries, L., Scheuerell, M.D., Szuwalski, C. *et al.* (2016). Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecol. Biogeogr.*, 25, 1144–1158.

Tikhonov, G., Abrego, N., Dunson, D. & Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods Ecol. Evol.* in press.

Tilman, D. (1990). Constraints and tradeoffs: toward a predictive theory of competition and succession. *Oikos*, 58, 3–15.

Tilman, D. (2004). Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. *Proc. Natl Acad. Sci. USA*, 101, 10854–10861.

Tjur, T. (2009). Coefficients of determination in logistic regression models—A new proposal: the coefficient of discrimination. *Am. Stat.*, 63, 366–372.

Vellend, M. (2010). Conceptual synthesis in community ecology. *Q. Rev. Biol.*, 85, 183–206.

Wang, Y., Naumann, U., Wright, S.T. & Warton, D.I. (2012). mvabund - an R package for model-based analysis of multivariate abundance data. *Methods Ecol. Evol.*, 3, 471–474.

Warren, D.L., Glor, R.E. & Turelli, M. (2008). Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution*, 62, 2868–2883.

Warton, D.I., Forster, S.D., De'ath, G., Stoklosa, J. & Dunstan, P.K. (2015a). Model-based thinking for community ecology. *Plant Ecol.*, 216, 669–682.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. *et al.* (2015b). So many variables: joint modeling in community ecology. *Trends Ecol. Evol.*, 30, 766–779.

Warton, D.I., Blanchet, F.G., O'Hara, R., Ovaskainen, O., Taskinen, S., Walker, S.C. *et al.* (2016). Extending joint models in community ecology: a response to Beissinger et al. *Trends Ecol. Evol.*, 31, 737–738.

Weiher, E., Freund, D., Bunton, T., Stefanski, A., Lee, T. & Bentivenga, S. (2011). Advances, challenges and a developing synthesis of ecological community assembly theory. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 366, 2403–2413.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y. *et al.* (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.*, 10, 1669–1681.

Whitfeld, T.J.S., Kress, W.J., Erickson, D.L. & Weiblen, G.D. (2012). Change in community phylogenetic structure during tropical forest succession: evidence from new guinea. *Ecography*, 35, 821–830.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F. *et al.* (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol. Rev.*, 88, 15–30.

Wootton, J.T. & Emmerson, M. (2005). Measurement of interaction strength in nature. *Annu. Rev. Ecol. Evol. Syst.*, 36, 419–444.

Zobel, M. (1997). The relative of species pools in determining plant species richness: an alternative explanation of species coexistence? *Trends Ecol. Evol.*, 12, 266–269.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

Editor, Dr. Jerome Chave
Manuscript received 16 December 2016
First decision made 31 January 2017
Manuscript accepted 9 February 2017