Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8,** 443–452 doi: 10.1111/2041-210X.12723

TECHNOLOGICAL ADVANCES AT THE INTERFACE BETWEEN ECOLOGY AND STATISTICS

Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context

Gleb Tikhonov*[,1], Nerea Abrego[2], David Dunson[3] and Otso Ovaskainen[1,2]

[1]*Metapopulation Research Centre, Department of Biosciences, University of Helsinki, P.O. Box 65, Helsinki FI-00014, Finland;* [2]*Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, N-7491 Trondheim, Norway; and* [3]*Department of Statistical Science, Duke University, P.O. Box 90251, Durham, NC, USA*

Summary

1. Joint species distribution models (JSDM) are increasingly used to analyse community ecology data. Recent progress with JSDMs has provided ecologists with new tools for estimating species associations (residual co-occurrence patterns after accounting for environmental niches) from large data sets, as well as for increasing the predictive power of species distribution models (SDMs) by accounting for such associations. Yet, one critical limitation of JSDMs developed thus far is that they assume constant species associations. However, in real ecological communities, the direction and strength of interspecific interactions are likely to be different under different environmental conditions.

2. In this paper, we overcome the shortcoming of present JSDMs by allowing species associations covary with measured environmental covariates. To estimate environmental-dependent species associations, we utilize a latent variable structure, where the factor loadings are modelled as a linear regression to environmental covariates.

3. We illustrate the performance of the statistical framework with both simulated and real data. Our results show that JSDMs perform substantially better in inferring environmental-dependent species associations than single SDMs, especially with sparse data. Furthermore, JSDMs consistently overperform SDMs in terms of predictive power for generating predictions that account for environment-dependent biotic associations.

4. We implemented the statistical framework as a MATLAB package, which includes tools both for model parameterization as well as for post-processing of results, particularly for addressing whether and how species associations depend on the environmental conditions.

5. Our statistical framework provides a new tool for ecologists who wish to investigate from non-manipulative observational community data the dependency of interspecific interactions on environmental context. Our method can be applied to answer the fundamental questions in community ecology about how species' interactions shift in changing environmental conditions, as well as to predict future changes of species' interactions in response to global change.

Key-words: biotic interaction, co-occurrence, environmental change, interaction network, interaction outcome, latent factor, species distribution model, stress-gradient hypothesis

Introduction

The assembly dynamics of species communities depend on both environmental filtering and biotic filtering (Vellend 2010; Götzenberger *et al.* 2012). While there is a large body of statistical literature on how to predict species' distributions based on abiotic variation (Elith & Leathwick 2009), the inclusion of interspecific interactions into species distribution modelling frameworks has been rather recent (Boulangeat, Gravel & Thuiller 2012; Kissling *et al.* 2012; Wisz *et al.* 2013). Thus far, this has been carried out either by adding the abundances of other species as predictors to single species models (le Roux *et al.* 2014; Mod *et al.* 2015), or by estimating species associations (residual co-occurrence patterns that remain after accounting for environmental niches) in joint species distribution models (JSDMs; Latimer *et al.* 2009; Clark *et al.* 2014; Thorson *et al.* 2015; Ovaskainen *et al.* 2016b). These statistical methods assume constant species associations in relation to the environment, yet the direction and strength of interspecific interactions in ecological communities can covary with environmental conditions (Tylianakis *et al.* 2008). For instance, when resources become scarce, competition among species might be intensified (Grime 1973; Goldberg 1990), whereas under abiotically stressful environmental conditions,

*Correspondence author. E-mail: gleb.tikhonov@helsinki.fi

facilitation might become particularly important (Brooker *et al*. 2008; He, Bertness & Altieri 2013). Changes in the outcomes of interspecific interactions in relation to changing environmental conditions have been found for a wide array of taxonomical groups (e.g. Erland & Finlay 1992; Brooker 2006). However, most studies thus far have focused only on a few species and/or simplified experimental conditions. This was partly due to the lack of statistically efficient methods that would enable one to evaluate how environmental conditions influence interspecific interactions from observational data on species-rich communities.

Ecologists have traditionally inferred the presence and strength of interspecific interactions from observational species occurrence data by examining species' co-occurrence patterns. Statistical methods for assessing species' co-occurrences include distance-based ordination approaches (Legendre & Legendre 2012), pairwise co-occurrence approaches (Veech 2014), metrics measuring species' aggregation and segregation patterns (Stone & Roberts 1990), and null model approaches (Gotelli 2000). A caveat with these methods is that they confound co-occurrence patterns generated by ecological interactions with those generated by co-variation in the species responses to abiotic variation. It is possible to examine whether the co-occurrences depend on environmental covariates (Williams, Howe & Hofmockel 2014), but this does not necessarily clarify whether the environmental covariates influence the occurrences or co-occurrences of the species. Another alternative is to use, in addition to environmental covariates, some of the species as predictors when modelling the occurrences of the other species (le Roux *et al*. 2014). Including an interaction term between the predictor species and the abiotic environmental covariates then allows one to ask whether and how abiotic conditions modify species associations (Mod, le Roux & Luoto 2014). However, this approach can be inefficient for the estimation of species associations in species-rich communities, and thus it has been typically applied for dominant species only.

Joint species distribution models have emerged in the ecological literature as a promising framework for assessing the effects of environmental filtering and species interactions in an integrated way (Warton *et al*. 2015). Joint species distribution models consider as the response variable the vector of occurrences of all species, and thus provide a model-based approach for inferring simultaneously species associations as well as species relationships to the abiotic environment (Clark *et al*. 2014; Pollock *et al*. 2014; Warton *et al*. 2015). As JSDMs control for measured environmental covariates, their estimates of species associations are more representative of true interactions than raw co-occurrence indices (Stone & Roberts 1990; Gotelli 2000), especially if inference is derived from partial correlations (Ovaskainen *et al*. 2016a) or from time-series data (Ives *et al*. 2003; Thorson *et al*. 2016). However, as non-manipulative data on species occurrence do not allow conclusive causal inference on ecological interactions, species associations estimated by JSDMs should be treated merely as hypotheses on ecological interactions (Ovaskainen, Hottola & Siitonen 2010), the validity of which should be verified in controlled experiments. In the early phase of the development of JSDMs, these approaches suffered from the curse of dimensionality, limiting the estimation of species association matrices to only few tens of species (e.g. Latimer *et al*. 2009; Ovaskainen, Hottola & Siitonen 2010). Thanks to recently introduced statistical techniques based on latent variable modelling (e.g. Bhattacharya & Dunson 2011), current JSDMs are able to estimate species association matrices for hundreds of species, including study designs with multiple hierarchical levels (Ovaskainen *et al*. 2016a) and spatially explicit data (Thorson *et al*. 2015; Ovaskainen *et al*. 2016b). However, as pointed out by Warton *et al*. (2015), an unresolved issue of current JSDMs is that they assume that the species associations are constant over environmental gradients, as well as more generally over space and time.

In this paper, we aim to fill the above-described gap in JSDMs and thus enable ecologists to use observational data for addressing the question of whether and how the environment influences species associations. We present an approach that extends latent factor based JSDMs to account for the dependency of species associations on environmental variables. In latent factors models, species associations are generated by factor loadings that stand for the responses of the species to latent factors, which can be interpreted as unmeasured environmental variables that generate the species associations. While in standard latent factors models the factor loadings are constant, we add another hierarchical layer that treats the factor loadings as linear regressions to measured environmental variables. After presenting the general statistical framework, we test its performance against two kinds of simulated data (data simulated by the statistical model itself and data simulated from an individual-based two-species competition model) and apply it to a real data set on plant presence–absence.

Materials and methods

DESCRIPTION OF THE STATISTICAL FRAMEWORK

We consider the typical kind of data acquired in community ecological studies: a set of species surveyed in a set of sites, of which the latter we refer to as sampling units. We denote sampling units by index $i = 1...n_y$ and the species by index $j = 1...n_s$, where n_y is the number of surveyed sampling units and n_s is the total number of recorded species. Our model is an extension of generalized mixed linear models, and thus it can be applied to various kinds of data (presence/absence, count, biomass, cover, etc.) through different link functions and error distributions. We model the recorded data as $y_{ij} \sim D(L_{ij}, \sigma_j^2)$, where D is the statistical distribution pertinent to the kind of data, L_{ij} is the linear predictor which models the expectation, and σ_j^2 is a variance parameter. For some distributions such as Poisson or Bernoulli, σ_j^2 is excluded, and for other cases such as normal distribution or over-dispersed Poisson distribution, it is included. The linear predictor is modelled as

$$L_{ij} = \sum_{k=1}^{n_c} x_{ik}\beta_{jk} + \varepsilon_{ij}, \quad \text{where } \varepsilon_{ij} = \sum_{h=1}^{n_f} \eta_{ih}\lambda_{jh}(\boldsymbol{x}_i^*.) \qquad \text{eqn 1}$$

The first term of this equation represents standard linear regression, where x_{ik} is the measured covariate $k = 1...n_c$ in the sampling unit i, and β_{jk} is the regression coefficient representing the linear response of species j to this covariate. In this model, the intercept is included to the

regression part by setting $x_{i1} = 1$ for all sampling units i, so that the number of measured environmental covariates is $n_c - 1$. To enable the parameterization of the model with sparse data or rare species, we model the regression coefficients as $\boldsymbol{\beta}_{j\cdot} \sim \mathrm{N}(\boldsymbol{\gamma}, \mathbf{V})$, where vector $\boldsymbol{\gamma}$ and the variance–covariance matrix $\mathbf{V}$ are community-level parameters to be estimated (Ovaskainen & Soininen 2011). The vector $\boldsymbol{\gamma}$ models the mean response over the species to the measured covariates, and the matrix $\mathbf{V}$ models the variation among individual species around the expectation. The dot notation in $\boldsymbol{\beta}_{j\cdot}$ (used also for other variables) means that $\boldsymbol{\beta}_{j\cdot}$ is a vector consisting of the β_{jk} values with the index k ranging over $k = 1\ldots n_c$, whereas the value of species index j is fixed.

The species associations are modelled through the term ε_{ij}, which is a sum of products of latent factors (η_{ih}) and latent loadings (λ_{jh}). The term η_{ih} is the latent factor $h = 1\ldots n_f$ for the sampling unit i, and $\lambda_{jh}(\boldsymbol{x}^*_{i\cdot})$ is the factor loading of species j to latent factor h, given a vector of predictors $\boldsymbol{x}^*_{i\cdot}$ at sampling unit i. The predictors $\boldsymbol{x}^*_{i\cdot} = (x^*_{i1}, \ldots, x^*_{in^*_c})$ on which the species associations are assumed to depend can be arbitrary. In practice, they usually consist of a subset of the predictors x_{ik} which are used to model how expected species occurrence depends on covariates.

In previous works (Bhattacharya & Dunson 2011; Warton *et al.* 2015; Ovaskainen *et al.* 2016a), the latent loadings have been assumed to be constant, i.e. not dependent on the $\boldsymbol{x}^*_{i\cdot}$. In that case, assuming $\eta_{ih} \sim N(0,1)$ as is usually done in factor analysis (see Jolliffe 2005), the species-to-species covariance structure is $\boldsymbol{\varepsilon}_{i\cdot} \sim N(0,\boldsymbol{\Omega})$, where $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$, and $\boldsymbol{\Lambda}$ denotes the latent loading matrix $\boldsymbol{\Lambda} = (\lambda_{jh})$. Here, we extend the previous works by modelling the factor loadings further as a linear regression of the x^*_{ik},

$$\lambda_{jh}(\boldsymbol{x}^*_{i\cdot}) = \sum\nolimits_{k=1}^{n^*_c} x^*_{ik}\lambda_{jhk} \qquad \text{eqn 2}$$

With this extension, which borrows from recent developments in the statistical literature (Hoff & Niu 2012; Fox & Dunson 2015), the among species covariance structure becomes a function of the environmental condition as $\boldsymbol{\varepsilon}_{i\cdot} \sim N(0, \boldsymbol{\Omega}(\boldsymbol{x}^*_{i\cdot}))$, where $\boldsymbol{\Omega}(\boldsymbol{x}^*_{i\cdot}) = \boldsymbol{\Lambda}(\boldsymbol{x}^*_{i\cdot})\boldsymbol{\Lambda}(\boldsymbol{x}^*_{i\cdot})^{\mathrm{T}}$, and $\boldsymbol{\Lambda}(\boldsymbol{x}^*_{i\cdot})$ is the matrix of latent loadings, which now depends on the environmental conditions x^*_{ik}. In the same way as in eqn (1), the intercept is included by setting $x^*_{i1} = 1$ for all sampling units i. We note that the baseline model, which assumes constant species associations, can be recovered by including the intercept as the only predictor in eqn (2).

We parameterized the model in the Bayesian framework. As we have assumed here straightforward modifications of the prior distributions and MCMC sampling schemes applied in previous works (Bhattacharya & Dunson 2011; Ovaskainen *et al.* 2016b), we provide the full description of these in the Appendix S1, Supporting Information only. Briefly, to incorporate the dependency of the latent loadings on the environmental variables (eqn 2), we assumed the multiplicative gamma shrinkage prior of Bhattacharya & Dunson (2011) for the λ_{jhk}, independently for each k, and derived a Gibbs sampling scheme for sampling of the λ_{jhk} from their full conditional distribution. As described in more detail in Bhattacharya & Dunson (2011), this prior allows for uncertainty in the number of factors and sparsity structure. The amount of shrinkage increases approximately exponentially with factor index k, and thus, the model assigns non-negligible factor loadings only for those factors that are strongly supported by the data.

INTERPRETATION OF MODEL PARAMETERS

After fitting the model to data, the parameterized model can be used both for inference and prediction in the same ways as JSDMs in general (Warton *et al.* 2015; Ovaskainen *et al.* 2016a). Here we focus on the novel components of the present approach, i.e. how to infer whether and how species associations depend on the environmental conditions $\boldsymbol{x}^*$, and how to use this information when generating predictions.

To measure the associations at a scale which is convenient for interpretation, we transform the covariance matrices $\boldsymbol{\Omega}$ into correlation matrices $\mathbf{R}$ by defining $R_{j_1j_2} = \Omega_{j_1j_2}/\sqrt{\Omega_{j_1j_1}\Omega_{j_2j_2}}$ for each pair of species $j_1, j_2 = 1\ldots n_s$. We remark that the model allows one to predict $\mathbf{R}(\boldsymbol{x}^*)$ for any environmental conditions $\boldsymbol{x}^*$, not just for those present in the training data, denoted above by $\boldsymbol{x}^*_{i\cdot}$ for sampling unit i. As we parameterize the model using Bayesian inference, the species associations for given environmental conditions $\boldsymbol{x}^*$ are characterized by the joint posterior distribution of the entire matrix $\mathbf{R}(\boldsymbol{x}^*)$, from which e.g. marginal posterior means and credible intervals can be computed for each pair of species. In addition to predicting the species associations for specific environmental conditions, we are interested in quantifying the level of statistical evidence on whether the associations vary with different environmental conditions. To do so, we define the species-to-species matrix of posterior probabilities $\boldsymbol{S}(\boldsymbol{x}^*_1, \boldsymbol{x}^*_2) = \Pr(\boldsymbol{R}(\boldsymbol{x}^*_1) > \boldsymbol{R}(\boldsymbol{x}^*_2))$. A value of $S_{j_1j_2}(\boldsymbol{x}^*_1, \boldsymbol{x}^*_2)$ close to 1 indicates that there is a high level of statistical support that the co-occurrence pattern among species j_1 and j_2 is more positive under the environment $\boldsymbol{x}^*_1$ than under the environment $\boldsymbol{x}^*_2$, whereas a value of $S_{j_1j_2}(\boldsymbol{x}^*_1, \boldsymbol{x}^*_2)$ close to 0 indicates that the opposite is true.

The MATLAB code that implements the statistical method can be found at https://github.com/gtikhonov/HMSC-DepAssoc.

TESTING THE PERFORMANCE OF THE STATISTICAL FRAMEWORK WITH SIMULATED DATA

We used simulated data to test whether the developed framework successfully estimates the dependency of species associations on the environmental context. We generated presence–absence data on species occurrence along a single environmental gradient, which we call altitude for the sake of illustration. We denote by $x_{i1} = 1$ the intercept and by x_{i2} the altitude of the sampling unit i. We assumed a study design where the sampling units were located evenly along the altitudinal gradient according to their index i, ranging from $x_{12} = -1$ to $x_{n_y 2} = 1$. We generated data with two kinds of models: in the null model, the species associations were constant along the altitudinal gradient, and in the full model, the species associations varied along the altitudinal gradient. In both cases, we assumed that the marginal occurrence probabilities of the species varied along the altitudinal gradient. Thus, in the null model, we assumed the linear predictor

$$L_{ij} = \beta_{j1} + x_{i2}\beta_{j2} + \sum\nolimits_{h=1}^{n_f} \eta_{ih}\lambda_{jh1} \qquad \text{eqn 3}$$

whereas in the full model, we assumed the linear predictor

$$L_{ij} = \beta_{j1} + x_{i2}\beta_{j2} + \sum\nolimits_{h=1}^{n_f} \eta_{ih}(\lambda_{jh1} + x_{i2}\lambda_{jh2}) \qquad \text{eqn 4}$$

While continuous-scale data are generally more informative for estimating covariance, we generated here presence–absence data, as it is a very common data type in ecology. For this we used the probit link function, implemented as $y_{ij} = 1_{L_{ij}+\epsilon_{ij} > 0}$, where $\varepsilon_{ij} \sim N(0,1)$ and $1_{x>0}$ is a binary indicator function, which takes the value of 1 if $x > 0$ and is otherwise 0. We sampled β_{j2} from $N(0,1)$, so that the species varied in their responses to the altitudinal gradient. To mimic typical sparse community ecological data where only few species are common and most are rare, we sampled β_{j1} from $N(-2,1)$. With these assumptions, the fraction of occupied sampling units is in the range [0·1,1] for 28% of the species, [0·01,0·1] for 42%, [0·001,0·01] for 21%, and <0·001 for the remaining 9%.

To mimic a biologically relevant case study, we assumed that the species associations depend on altitude in a manner that is in line with stress-gradient hypothesis (SGH). Presuming that low altitude represents mild and high altitude harsh environmental conditions, we assumed that the interspecific associations shift towards positive with increasing abiotic stress and thus with increasing altitude (Callaway *et al*. 2002). To generate such variation, we fixed the number of latent factors $n_f = 2$, and sampled λ_{jhk} according to the formulas $\lambda_{jh1} = 0{\cdot}4(a_{jh} + b_{jh})$ and $\lambda_{jh2} = 0{\cdot}4b_{jh}$, where $a_{jh} \sim N(0,1)$ and $b_{jh} \sim N(0{\cdot}9,10^{-2})$. Under this parameterization, the expected proportions of positive and negative associations at low altitude are equal, while at high altitude most of associations are positive. We refer to the primary (e.g. β_{jk} and λ_{jhk}) and derived (e.g. $\mathbf{R}(\boldsymbol{x}^*)$) parameters that were used to generate the data as 'true', to distinguish them from the estimated values of the same parameters. We call simulated data sets generated by the null model (eqn 3) and the full model (eqn 4) as null data and full data respectively.

We fixed the number of species to $n_s = 50$. To test how the statistical power of the approach depends on the data size, we varied the number of sampling units as $n_y = 100, 200, 400, 800, 1600, 3200$. We generated the data for the largest study design only ($n_y = 3200$), and obtained the data for the smaller study designs by subsampling these data. We generated 30 replicate data sets of each type and size.

For each generated data set, we fitted the JSDM with structure equal to the full model (eqn 4), and hence allowed the species associations to vary with altitude, regardless of whether they did so in reality or not. We evaluated whether the model was able to capture the variation on the species associations along the environmental gradient. To do so, we computed for each species pair the level of statistical evidence $S_{j_1j_2}(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ that their association was more positive at high altitude than at low altitude, where $\boldsymbol{x}_1^* = (1, -1)$ and $\boldsymbol{x}_2^* = (1, 1)$. For each species pair, we classified the inference obtained from the fitted model as 'correct' if there was no change in the true association (null data) and the inferred association showed no change, or if there was a change in the true association (full data) and the inferred association indicated a change with the correct direction. We classified the inference as 'misleading' if there was no change in the true association but the inferred association indicated a change, or if there was a change in the true association but the inferred association indicated a change into the opposite direction. The remaining cases in which there was a change in the true association but the inferred association indicated no change were classified as 'lack of statistical power'. The estimated association for each species pair was considered to vary with altitude if the value of $S_{j_1j_2}(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ was outside the range [0·05,0·95].

We compared the performance of our method with methods used earlier in the context of single species distribution models (SDMs). To do so, we fitted SDMs to the same simulated data, separately for each of the species pairs. In these models, the response variable was the occurrence of the first species, and predictors included altitude, the occurrence of the second species, and the interaction between the second species' occurrence and altitude. We fitted the probit regression models with the *glmfit* function in MATLAB, thus assuming the maximum likelihood paradigm. We considered that the species associations shifted towards the direction determined by the sign of the regression coefficient of the interaction term if this term was significant with $P < 0{\cdot}05$. We classified the inference for each ordered species pair as 'correct', 'misleading' or 'lack of statistical power' as we did with the JSDM.

We examined how much accounting for species associations influenced the predictive powers of the models. We mimicked a situation in which the ecologists would have surveyed the occurrences of all species for the sampling units that make up the training data (used for model parameterization), but only the 10 most dominant species (based on their abundances in the training data) for additional sampling units that make up the validation data. Now the question is that how well different models are able to predict the occurrences of the remaining 40 species (which are mostly rare) in the validation data. We compared the predictive powers of SDMs and JSDMs that do or do not account for species associations. The SDMs that do not account for species associations had altitude as the sole predictor, whereas the ones that account for species associations had also the 10 most dominant species and their interaction terms with altitude. The predictions by JSDMs that do or do not account for species associations were made by the same full version of the JSDM that included altitude-dependent species associations. The ones that do not account for species associations are unconditional predictions, whereas the ones that do account for species associations are conditional predictions. The latter were generated using the method of Ovaskainen *et al*. (2016a), modified to account for environmental-dependent associations. We used these four models to predict the occurrences of the 40 non-surveyed species in the validation data, and evaluated the models' predictive power with two measures that we averaged over the species: (i) Tjur's R^2, i.e. mean predicted occurrence probability over occupied sites minus mean predicted occurrence probability over unoccupied sites (Tjur 2009), and (ii) deviance, i.e. the negative of twice the likelihood of observing the validation data, given the model prediction.

To examine the generality of the results, we repeated the above described analyses in three different ways. First, to test the influence of the sparsity of the data, we used a different parameterization for the model intercepts (the β_{j1} parameters), so that instead of a community dominated by rare species, we assumed a community dominated by common species. Second, to test the sensitivity of the results to the prior assigned to the latent factors, we repeated the analyses by assuming that the researcher had prior information about the structure of the association network. Third, to test the sensitivity of the results to the structure of the association network, we used a different parameterization for the factors loadings. Instead of assuming the relatively simple structure of the association network generated by two latent factors, we assumed a more complex structure generated by five factors. Furthermore, instead of the directional change corresponding to the SGH, we assumed that the associations changed in a non-directional way along altitude. These alternative scenarios are described in more detail in the Appendix S1.

As an additional case study, we examined the robustness of the statistical approach by applying it to data simulated by an individual-based model, which data may violate the structural assumptions made by the JSDM. The description and results for this simulated example are provided in the Appendix S1.

TESTING THE PERFORMANCE OF THE STATISTICAL FRAMEWORK WITH REAL DATA

We reanalysed the plant cover data from Mod, le Roux & Luoto (2014). In the original study, Mod, le Roux & Luoto (2014) modelled separately the cover of 17 plant species, including the logarithm of geomorphological disturbance, soil moisture, logarithm of the dominant species' cover (*Empetrum*), and all interaction terms between *Empetrum* and the environment variables as predictors. As explained in more detail in Mod, le Roux & Luoto (2014), soil disturbance and moisture represent stress factors for plant growth.

We modelled the presence–absence of all 18 plant species with a JSDM where we included soil moisture, geomorphological

disturbance and their interaction as predictors, and allowed the species associations at the sampling unit level to vary with both variables. As the data had been collected by hierarchical sampling design, with sampling units located within sites, we followed Ovaskainen *et al.* (2016a) to estimate random variation in species occurrences and co-occurrences also at the site level, at which level we assumed the associations to be constant (for exact mathematical description of the model structure see Appendix S1). We predicted the associations $\mathbf{R}(\boldsymbol{x}^*)$ at different combinations of disturbance and soil moisture (low and high), and computed the level of statistical support $S(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ of difference in species associations along the disturbance and soil moisture gradients.

Results

The species association networks estimated for both the null data (Fig. 1d–f) and the full data (Fig. 1g–i) closely resembled the true association networks (Fig. 1a–c). Overall, the estimates for the null data (Figs. 1d–f and 2c) successfully inferred that the associations did not vary with altitude (grey lines in Fig. 2c), with only a small fraction of misleading cases (coloured lines in Fig. 2c). The fitted model generally replicated the true behaviour also for the full data (Figs. 1g–i and coloured lines in Fig. 2d), though as expected for the sparse

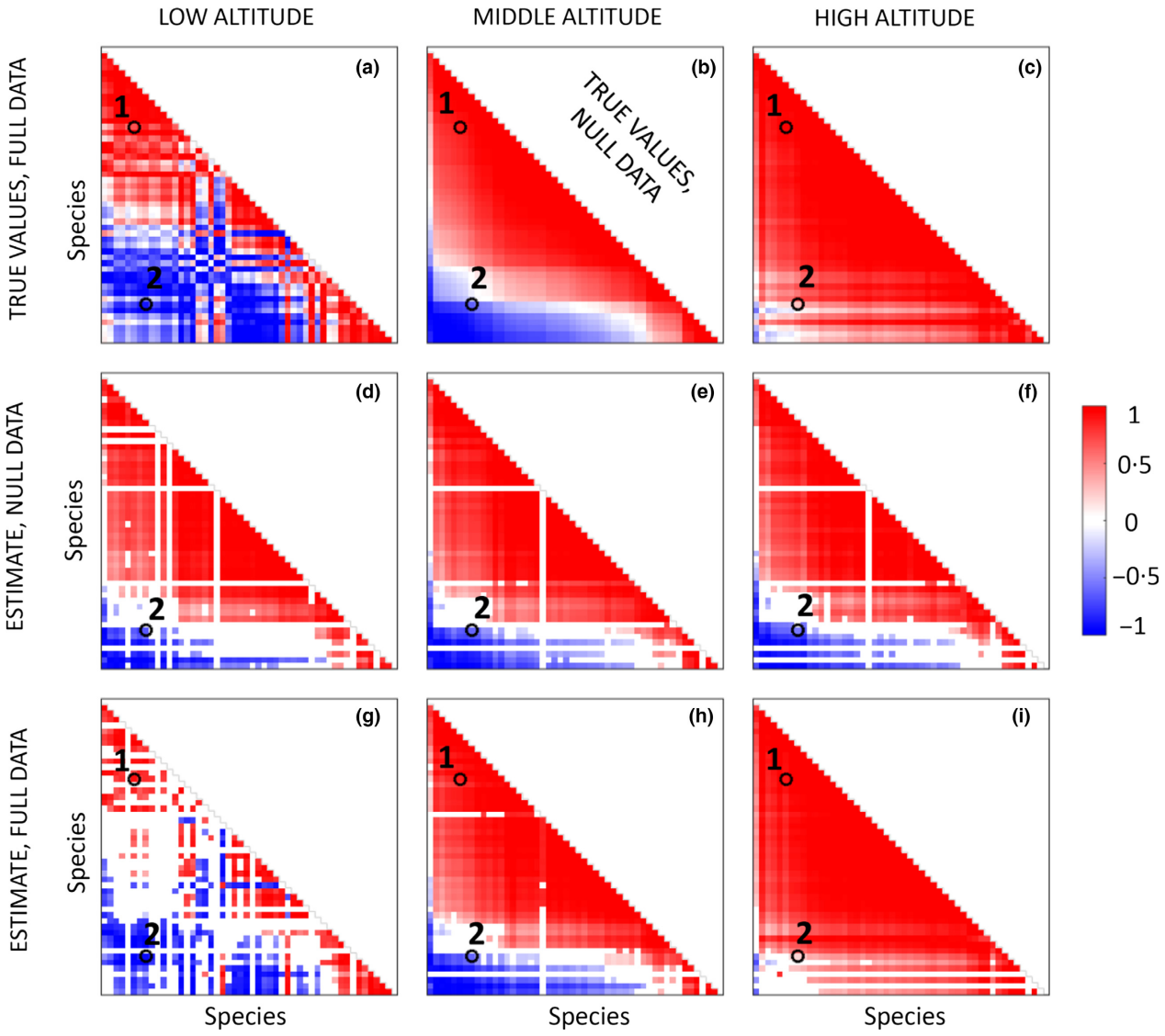

Fig. 1. True and estimated species associations measured by correlation matrices $\mathbf{R}(\boldsymbol{x}^*)$. (a–c) It shows the true associations for a single replicate of the full data at low altitude (a), middle altitude (b) and high altitude (c). (b) It also represents the true associations for the corresponding replicate of the null data, which assumes constant associations along the altitude gradient. (d–f) represent the estimated associations based on fitting the full model to 3200 sampling units of the null data and predicting the associations at low (d), middle (e), and high (f) altitudes. (g–i) It represents the estimated associations based on fitting the full model to 3200 sampling units of full data and predicting the associations at low (g), middle (h), and high (i) altitudes. Cell colour depicts the value of true correlations (a–c) or posterior mean of estimated correlations (d–i), with red standing for positive correlations and blue for negative (see the colour bar). In (d–i), cases where the 95% posterior credible interval of the correlation includes zero are shown by white. Species' order is the same in all panels, and it has been selected so that the structure of the association network in (b) is as clear as possible. In all panels, only lower triangles of the symmetric association matrices are shown. Black circles, marked with numbers 1 and 2, highlight species pairs that are analysed further in Fig. 2.

community data being analysed, there was a large fraction of species for which there was lack of statistical power (grey lines in Fig. 2d). The lower panels of Fig. 2 exemplify further the true and estimated associations for three species pairs, illustrating one case of lack of statistical power (Fig. 2e), and two cases of correct inference (Figs. 2f and g).

The ability of the model to identify changes in associations between species pairs increased with the amount of data (Fig. 3d), and with the true change in the association (Fig. S1). The JSDM approach had more statistical power than the SDM for capturing whether the associations changed with altitude, as evidenced by the red bars being higher in Fig. 3d than in Fig. 3b. The fraction of misleading inferences was in line with expectations based on the threshold criteria used, both with SDM and JSDM approaches. In terms of predictive power, unconditional SDMs and JSDMs showed almost equal performance for all amounts of data. Conditioning the predictions for each focal species on the 10 dominant species considerably improved the models' predictions in terms of the Tjur's R^2 both for SDM and JSDM, with no major differences between these two approaches (Fig. 3f). However, in terms of the deviance, the conditional predictions by the JSDM overperformed those of the SDM (Fig. 3e).

For the community dominated by common species, both SDM and JSDM approaches had more power to correctly detect the changes in the associations (Fig. S2) than for the community dominated by the rare species (Fig. 3). Joint species distribution models performed better than SDMs especially for the sparse data. The results for the alternative prior (Fig. S3) and for the alternative structure of the species association network (Figs. S4 and S5) were qualitatively similar to the results shown in Fig. 3, confirming their robustness.

Our analysis of the plant data provided partial support to previous findings of Mod, le Roux & Luoto (2014) that the plant species associations can be dependent on multiple environmental factors: for many species pairs, the association differed at different combinations of environmental factors (Fig. 4a–d). However, the statistical support for a change in the associations along at least one of the environmental gradients was greater than 95% only for 51 out of 153 species pairs (Fig. 4e and f). Furthermore, the numbers of species pairs that changed towards more positive and more negative associations were similar for both environmental gradients (Fig. 4e and f).

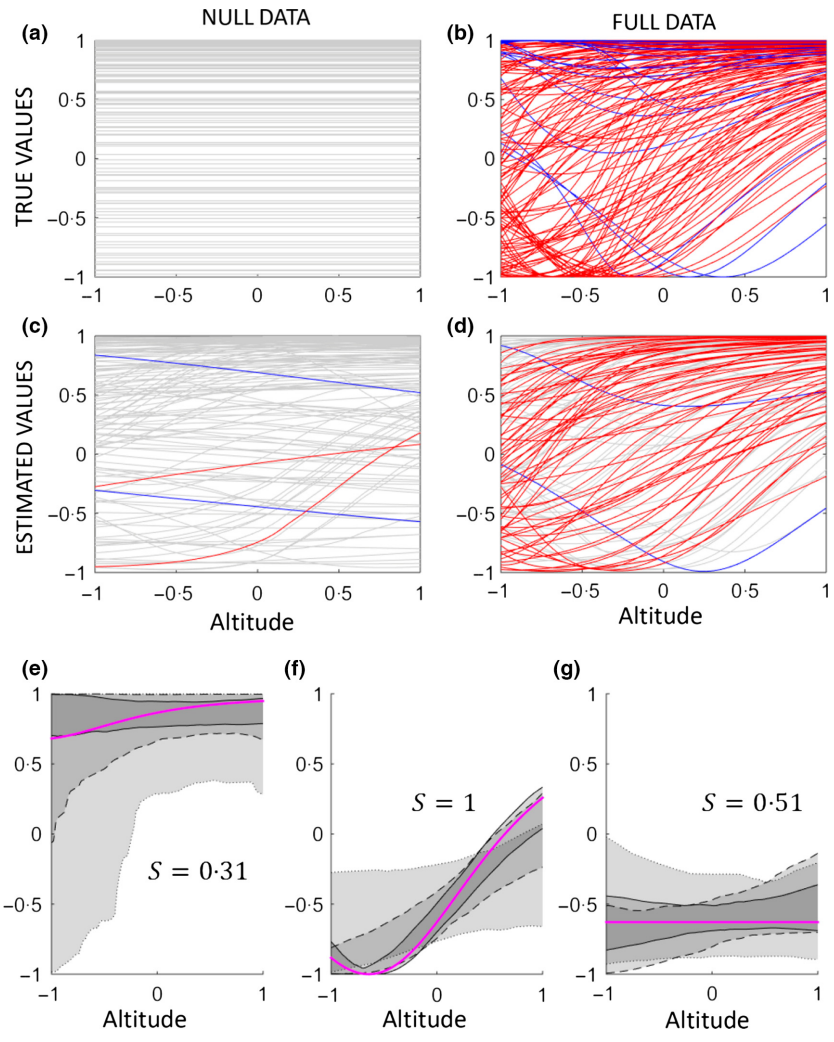

Fig. 2. Post-processed model outputs illustrating whether and how species associations depend on altitude. (a–d) It shows how the true (a and b) and estimated (c and d) associations vary with altitude among all pairs of 20 randomly selected species. The data as well as the fitted models are the same as in Fig. 1, with (a and c) corresponding to null data where the true associations do not vary with altitude and (b and d) corresponding to the full data where the true associations vary with altitude. The colours indicate whether the associations are inferred to increase (red; $S > 0{\cdot}95$), decrease (blue; $S < 0{\cdot}05$), or remain constant (grey; $0{\cdot}05 \leq S \leq 0{\cdot}95$) along the altitude gradient. For (a and b) the same colours refer to the assumed true behaviour. (e–g) It exemplifies posterior distributions of altitude-dependent associations for species pairs indicated in Fig. 1 by circles numbered by 1 (e: full data) and 2 (f: full data; g: null data). In these panels, the solid magenta lines show the true associations, and the areas with different grey intensities represent the [5%,95%] posterior quantiles estimated based on data with $n_y = 400$ (dotted lines), $n_y = 800$ (dashed lines) and $n_y = 3200$ (solid lines) sampling units. The values of statistical support S shown in these panels are based on $n_y = 3200$.

Discussion

Determining whether the outcomes of interspecific interactions depend on the environmental conditions is of major interest both for understanding the basic ecology and distribution of species, as well as for predicting changes at macroecological scales due to global change (Tylianakis *et al*. 2008; Van der Putten, Macel & Visser 2010; Hagen *et al*. 2012). The statistical framework introduced in this paper provides a new tool for ecologists interested in inferring the dependency of interspecific interactions on environmental context from non-manipulative observational community data. Compared to existing methods, the principal advantage of our method is that it enables inference about species associations and their changes from sparse data on large communities dominated by rare species.

Our results with simulated data show that the approach proposed here is capable of finding signals of changing associations in a robust and statistically efficient way. As expected, the ability to classify how species associations depend on environmental conditions increases with the size of the data set. Our results were robust for all sizes of data in the sense that the rate of false positives corresponded to the threshold level of statistical support we used to infer changing associations. Our additional simulated example (presented in Appendix S1) consisted of data generated by an individual-based model. The statistical model successfully captured how associations varied along the environmental gradient also in this example (see Appendix S1), suggesting the robustness of the method to data that violates the structural assumptions of the statistical model.

As any statistical method, also the method presented here has its shortcomings and underlying assumptions, which the researchers should keep in mind when interpreting the results. Most importantly, while JSDMs are aimed to capture biotic interactions, they are of correlative nature, and thus their results should be interpreted cautiously. For example, model miss-specifications (such as missing covariates or variation in detection probability over environmental gradients) may generate both co-occurrence patterns and their changes over environmental gradients. However, our method avoids many pitfalls of earlier approaches (see Introduction), and thus allows community ecologists to move as close to causal

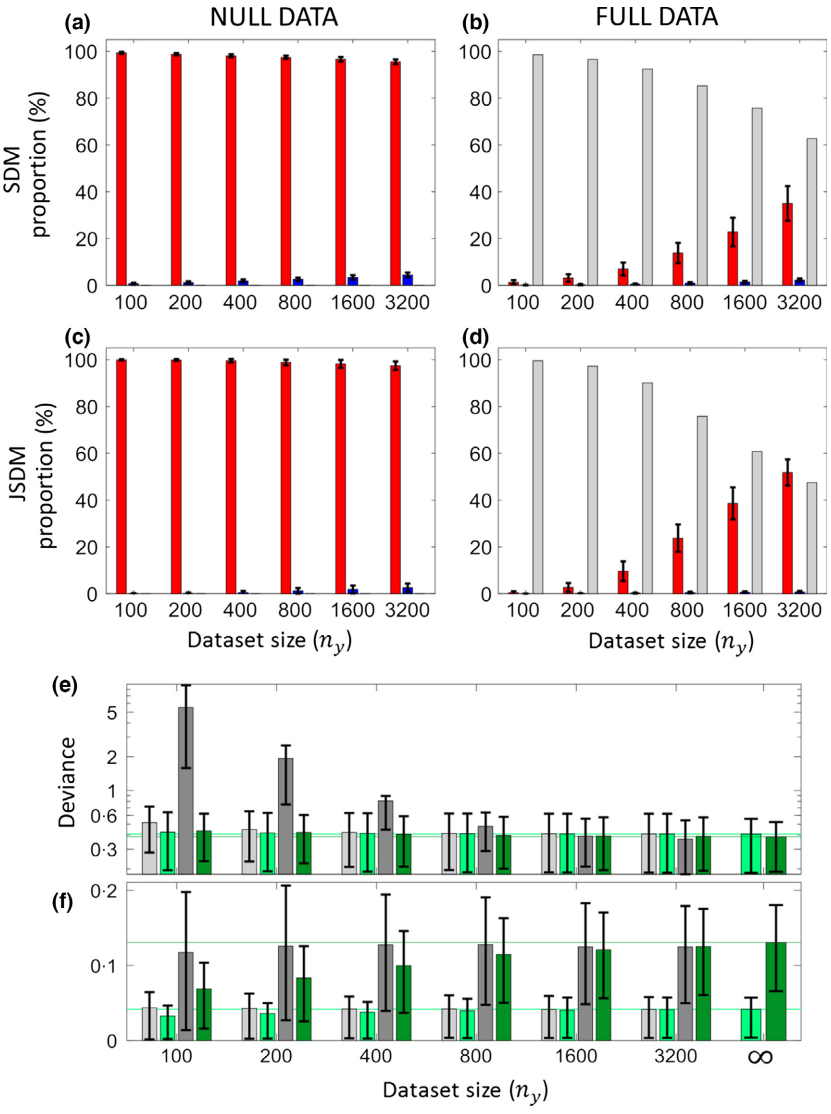

Fig. 3. A comparison between single species distribution models (SDMs) and joint species distribution models (JSDMs) in their ability to account for the dependency of species associations on environmental covariates. The heights of the bars in (a–d) show the proportions of species pairs (averaged over the replicate data sets; error bars show standard errors) with correct inference (red), misleading inference (blue), and lack of statistical power (grey). (e and f) It compares the predictive powers of the fitted models, measured by deviance (e) and Tjur R^2 (f) of the model's predictions for validation data (averaged over the replicate data sets; error bars show 25%, 75% quantiles). The colours correspond to different models: SDM (light grey) and JSDM (light green) that do not account for species associations (unconditional), and SDM (dark grey) and JSDM (dark green) that utilize the information on the 10 most dominant species (conditional). The two rightmost bars and the horizontal lines represent the unconditional and conditional JSDMs with true parameters values.

inference as is principally possible with non-manipulative observational data. Yet, the patterns revealed by our approach should be considered merely as hypotheses on species interactions and their dependency on environmental conditions. For empirical ecologists, the most rigorous way of validating results obtained by applying our method to observational data is to pick examples of species pairs with changing associations, and perform experimental tests for those species pairs. From the theoretical point of view, additional tests using simulated data generated from independent models could provide further insights on the capabilities and limitations of our method.

One aspect that we have ignored in our case studies is that, while the association matrices $\mathbf{R}(\boldsymbol{x}^*)$ and the levels of statistical support of a change in species-to-species co-association $S(\boldsymbol{x}_1^*, \boldsymbol{x}_2^*)$ are defined for all species pairs and all environmental conditions, it is often the case that not all species are present in all environments. For example, the correlation between two species may be predicted to be positive even for a site where both species are predicted to be absent. Then the positive correlation between them means that in the very unlikely case that the species would be present in the focal environment, they would be likely to be found together in the same sampling unit. For this reason, the estimated correlation matrices can be considered as hypothetical association networks rather than realized association networks. To construct a realized association network, we propose to consider only those species pairs that are likely to be present in the focal environment. For the presence threshold, one may e.g. require that the product (over the two species) of the species occurrence probabilities exceeds a given threshold value. A species pair can then be considered to have a realized association in the focal environment if its joint presence in the focal environment exceeds such threshold value, and if the estimated association is either positive or negative with at least some specified posterior probability.

The model development presented here was partly motivated by the recent discussion in plant community ecology about to what extent there is empirical support for the SGH. In general terms, the SGH predicts that facilitation becomes more important in shaping plant communities as abiotic stress increases (Bertness & Callaway 1994). While there are many studies providing empirical support for this phenomenon (Brooker *et al.* 2008; He, Bertness & Altieri 2013), some other studies have found only weak or no evidence (Maestre, Valladares & Reynolds 2005). Some authors have suggested that there may be such a discrepancy among the results because empirical tests for SGH usually focus on a few species only, not on the whole community (Maestre

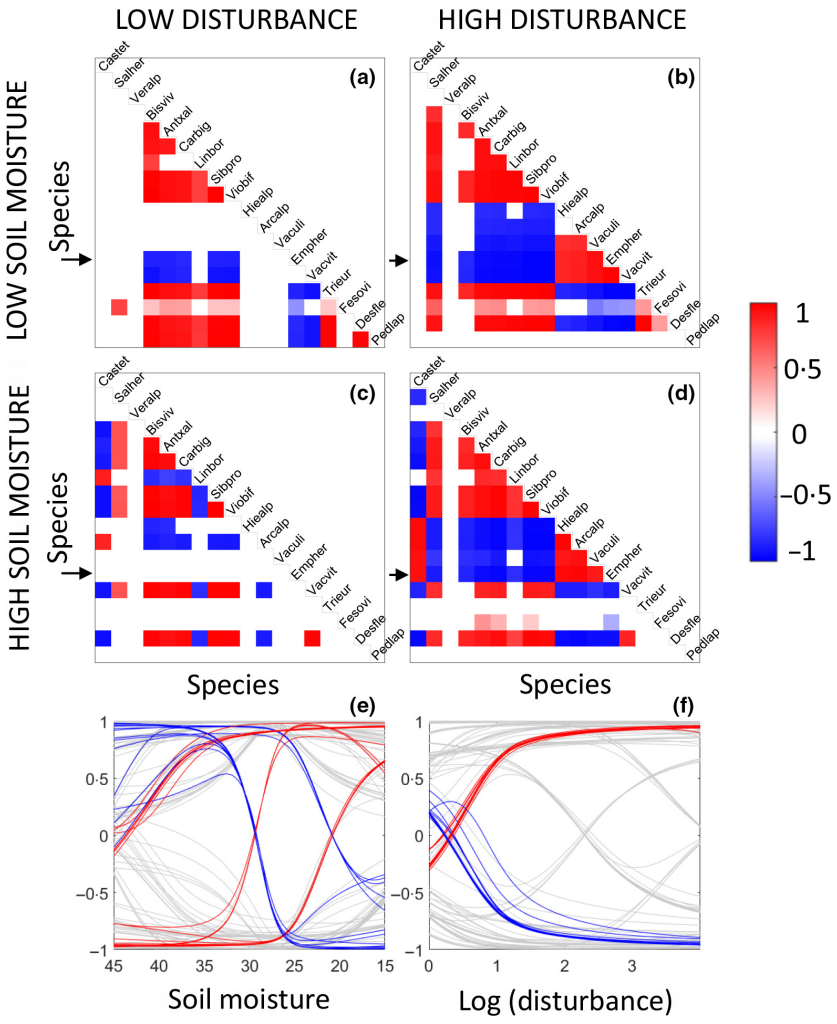

Fig. 4. Results from applying the joint species distribution model framework to plant cover data. (a–d) It illustrate the estimated species associations $R_{j_1 j_2}(\boldsymbol{x}^*)$ at different extremes of soil moisture and geomorphological disturbance: low moisture (a and b), high moisture (c and d), low disturbance (a and c), or high disturbance (b and d). (e and f) It visualize the dependence of species associations as a function of single environmental variable, with the variable not varied set to its mean value (line colours are coded as in Fig. 2a–d). The arrows indicate the rows corresponding to the dominant species (*Empetrum*) which was studied in more detail by Mod, le Roux & Luoto (2014).

et al. 2009; Soliveres, Smit & Maestre 2015). As we have illustrated using the data of Mod, le Roux & Luoto (2014), our method can be used to test the SGH using plant community data in an integrated manner. If we would have considered associations only between the most dominant species and the other species, as was done by Mod, le Roux & Luoto (2014), we would have seen only part of the bigger picture. Mod, le Roux & Luoto (2014) related their results to the SGH by noting that the positive effects of the most dominant species were generally stronger at higher levels of environmental stress. However, considering the species associations among all species within the community, the evidence for SGH that we found from this data set was rather limited, as we observed roughly equal amount of changes towards positive and negative directions. We hope that future applications of the method presented here will help to provide more synthetic insights into the SGH.

One of the reasons for the high success of SDMs in ecology and conservation biology is that they enable predicting species' distributions for different environmental scenarios (Elith & Leathwick 2009). In particular, SDMs have played an important role in the research of future effects of environmental threats on global species' expansions and extinctions (Pearson & Dawson 2003; Araújo *et al*. 2005; Hijmans & Graham 2006). Our method provides exciting possibilities for extending these lines of research, as it can additionally be used for predicting how species associations may change under different scenarios of environmental change, and how such changing associations may feed back into species' occurrence.

Authors' contributions

G.T. and O.O. conceived the ideas; G.T., D.D. and O.O. designed the methodology; G.T. performed simulations and analyses; G.T., N.A. and O.O. wrote the manuscript. All authors contributed significantly to the drafts and gave final approval for publication.

Acknowledgements

We thank James Thorson and an anonymous reviewer for helpful comments. We are grateful to Miska Luoto for providing the data for the empirical case study. We acknowledge funding from the Academy of Finland (grant no. 250444 to O.O.), the Research Council of Norway (CoE grant no. 223257), and the LUOVA graduate school (PhD grant for G.T.).

Data accessibility

Plant cover data from Mod, le Roux & Luoto (2014), MATLAB code that implements the statistical method, MATLAB scripts for generating simulated data, and MATLAB scripts for estimating parameters of the statistical models can be found at the GitHub repository https://github.com/gtikhonov/HMSC-DepAssoc .

References

Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. (2005) Validation of species-climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Bertness, M.D. & Callaway, R. (1994) Positive interactions in communities. *Trends in Ecology & Evolution*, **9**, 191–193.

Bhattacharya, A. & Dunson, D.B. (2011) Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291–306.

Boulangeat, I., Gravel, D. & Thuiller, W. (2012) Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters*, **15**, 584–593.

Brooker, R.W. (2006) Plant-plant interactions and environmental change. *New Phytologist*, **171**, 271–284.

Brooker, R.W., Maestre, F.T., Callaway, R.M. *et al*. (2008) Facilitation in plant communities: the past, the present, and the future. *Journal of Ecology*, **96**, 18–34.

Callaway, R.M., Brooker, R.W., Choler, P. *et al*. (2002) Positive interactions among alpine plants increase with stress. *Nature*, **417**, 844–848.

Clark, J.S., Gelfand, A.E., Woodall, C.W. & Zhu, K. (2014) More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, **24**, 990–999.

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

Erland, S. & Finlay, R. (1992) Effects of temperature and incubation time on the ability of three ectomycorrhizal fungi to colonize *Pinus sylvestris* roots. *Mycological Research*, **96**, 270–272.

Fox, E. & Dunson, D. (2015) Bayesian nonparametric covariance regression. *Journal of Machine Learning Research*, **16**, 2501–2542.

Goldberg, D.E. (1990) Components of resource competition in plant communities. *Perspectives on Plant Competition* (eds J.B. Grace & D. Tilman), pp. 27–49. Academic Press, San Diego, CA, USA; London, UK.

Gotelli, N.J. (2000) Null model analysis of species co-occurrence patterns. *Ecology*, **81**, 2606–2621.

Götzenberger, L., de Bello, F., Bråthen, K.A. *et al*. (2012) Ecological assembly rules in plant communities-approaches, patterns and prospects. *Biological Reviews*, **87**, 111–127.

Grime, J.P. (1973) Competitive exclusion in herbaceous vegetation. *Nature*, **242**, 344–347.

Hagen, M., Kissling, W.D., Rasmussen, C. *et al*. (2012) Biodiversity, species interactions and ecological networks in a fragmented world. *Advances in Ecological Research*, **46**, 89–210.

He, Q., Bertness, M.D. & Altieri, A.H. (2013) Global shifts towards positive species interactions with increasing environmental stress. *Ecology Letters*, **16**, 695–706.

Hijmans, R.J. & Graham, C.H. (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, **12**, 2272–2281.

Hoff, P.D. & Niu, X. (2012) A covariance regression model. *Statistica Sinica*, **22**, 729–753.

Ives, A.R., Dennis, B., Cottingham, K.L. & Carpenter, S.R. (2003) Estimating community stability and ecological interactions from time-series data. *Ecological Monographs*, **73**, 301–330.

Jolliffe, I.T. (2005) Factor analysis, overview. *Encyclopedia of Biostatistics*, **3**, 2nd edn. John Wiley & Sons Ltd.

Kissling, W.D., Dormann, C.F., Groeneveld, J. *et al*. (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.

Latimer, A.M., Banerjee, S., Sang, H. Jr, Mosher, E.S. & Silander, J.A. Jr (2009) Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecology Letters*, **12**, 144–154.

Legendre, P. & Legendre, L. (2012) *Numerical Ecology*, 3rd English edn. Elsevier Science BV, Amsterdam, Netherlands.

Maestre, F.T., Valladares, F. & Reynolds, J.F. (2005) Is the change of plant-plant interactions with abiotic stress predictable? A meta-analysis of field results in arid environments. *Journal of Ecology*, **93**, 748–757.

Maestre, F.T., Callaway, R.M., Valladares, F. & Lortie, C.J. (2009) Refining the stress-gradient hypothesis for competition and facilitation in plant communities. *Journal of Ecology*, **97**, 199–205.

Mod, H.K., le Roux, P.C. & Luoto, M. (2014) Outcomes of biotic interactions are dependent on multiple environmental variables. *Journal of Vegetation Science*, **25**, 1024–1032.

Mod, H.K., le Roux, P.C., Guisan, A. & Luoto, M. (2015) Biotic interactions boost spatial models of species richness. *Ecography*, **38**, 913–921.

Ovaskainen, O., Hottola, J. & Siitonen, J. (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, **9**, 2514–2521.

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 in Ecology and Evolution*, **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 in Ecology and Evolution*, **7**, 428–436.

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

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., Vesk, P.A. & McCarthy, M.A. (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, **5**, 397–406.

le Roux, P.C., Pellissier, L., Wisz, M.S. & Luoto, M. (2014) Incorporating dominant species as proxies for biotic interactions strengthens plant community models. *Journal of Ecology*, **102**, 767–775.

Soliveres, S., Smit, C. & Maestre, F.T. (2015) Moving forward on facilitation research: response to changing environments and effects on the diversity, functioning and evolution of plant communities. *Biological Reviews*, **90**, 297–313.

Stone, L. & Roberts, A. (1990) The checkerboard score and species distributions. *Oecologia*, **85**, 74–79.

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 in Ecology and Evolution*, **6**, 627–637.

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

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

Tylianakis, J.M., Didham, R.K., Bascompte, J. & Wardle, D.A. (2008) Global change and species interactions in terrestrial ecosystems. *Ecology Letters*, **11**, 1351–1363.

Van der Putten, W.H., Macel, M. & Visser, M.E. (2010) Predicting species distribution and abundance responses to climate change: why it is essential to include biotic interactions across trophic levels. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **365**, 2025–2034.

Veech, J.A. (2014) The pairwise approach to analysing species co-occurrence. *Journal of Biogeography*, **41**, 1029–1035.

Vellend, M. (2010) Conceptual synthesis in community ecology. *The Quarterly Review of Biology*, **85**, 183–206.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. & Hui, F.K.C. (2015) So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution*, **30**, 766–779.

Williams, R.J., Howe, A. & Hofmockel, K. (2014) Demonstrating microbial co-occurrence pattern analyses within and between ecosystems. *Frontiers in Microbiology*, **5**, 10.

Wisz, M.S., Pottier, J., Kissling, W.D. *et al.* (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews*, **88**, 15–30.

Received 15 August 2016; accepted 12 December 2016
Handling Editor: David Warton

Supporting Information

Details of electronic Supporting Information are provided below.

Appendix S1. Additional information and details about the statistical model, alternative simulation scenarios, testing the performance of the statistical framework with data generated from an individual-based model, and testing the performance of the statistical framework with the plant data.