Selecting traits that explain species–environment relationships: a generalized linear mixed model approach

Jamil, T., Ozinga, W. A., Kleyer, M., & ter Braak, C. J. F.

Jamil, T., Ozinga, W. A., Kleyer, M., & ter Braak, C. J. F. (2013). Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. Journal of Vegetation Science, 24(6), 988-1000. https://doi.org/10.1111/j.1654-1103.2012.12036.x

This is the pre-peer reviewed version of the following article:

Jamil, T, Ozinga, WA, Kleyer, M and ter Braak CJF (2012) Selecting traits that explain species-environment relationships: a Generalized Linear Mixed Model approach. Journal of Vegetation Science. Doi: 10.1111/j.1654-1103.2012.12036.x.

which has been published in final form at [http://dx.doi.org/10.1111/j.1654-1103.2012.12036.x].

Selecting traits that explain species-environment relationships: a Generalized Linear Mixed Model approach

Tahira Jamil[1,5], Wim A. Ozinga[2,3], Michael Kleyer[4] and Cajo J. F. ter Braak[1*]

[1] Biometris, Wageningen University and Research Centre, Box 100, 6700 AC Wageningen, the Netherlands.
[2] Alterra, Wageningen University and Research Centre, Box 47, 6700 AA Wageningen, The Netherlands.
[3] Radboud University Nijmegen, Department Ecology, Toernooiveld 1, NL-6525 ED Nijmegen, NL
[4] Landscape Ecology Group, University of Oldenburg. Oldenburg, 26111, Germany.
[5] COMSATS Institute of Information Technology, Department Mathematics, Park Road Islamabad, Pakistan.

[*] Corresponding author. E-mail: cajo.terbraak@wur.nl

The name and complete mailing address of corresponding author:
Prof. Cajo J.F. ter Braak
Biometris, Building 107, Room W2Aa085
Wageningen University
Droevendaalsesteeg 1
6708 PB Wageningen, the Netherlands
Phone 0031 317 480803
Fax 0031 317 483554
email: cajo.terbraak@wur.nl

Running title: trait-species-environment modeling by GLMM

Abstract

Question: Quantification of the effect of species traits on the assembly of communities is challenging from a statistical point of view. A key question is how species occurrence and abundance can be explained by the traits values of the species and the environmental values at the sites.

Methods: Using a sites x species abundance table, a site x environment data table and a species x trait data table, we address this question by a novel Generalized linear mixed model (GLMM) approach. The GLMM overcomes the problem of pseudo-replication and heteroscedastic variance by including sites and species as random factors. The method is equally well applicable to presence-absence data as to count and multinomial data. We present a tiered forward selection approach for obtaining a parsimonious model and compare the results with the fourth corner method and RLQ ordination.

Results: We illustrate the approach on a presence-absence version on two well-known data sets. In the Dune Meadow data species presence is parsimoniously explained by moisture and manure of the meadows in combination with seed mass and specific leaf area, respectively. In the Grazed Grassland data species presence is parsimoniously explained by the grazing intensity and soil phosphorous in combination with the C:N ratio and flowering mode, respectively.

Conclusions: Our GLMM approach can be used to identify which species traits and environmental variables best explain the species distribution, and which traits are significantly correlated with environmental variables. The method is better suited for providing an interpretable and predictive model than the fourth corner method and RLQ.

Key-words: community assembly; environmental gradient; trait-environment relationship; fourth corner; functional ecology; generalized linear mixed model; RLQ; species traits

Introduction

A central focus of community ecology is to understand and explain where and when particular species or groups of species occur and thrive, and where and when not. Species differ in what they require from the environment and environmental conditions vary in space and time. Differences in traits of species and differences in the environment must thus be part of the explanation. The role of species traits in community assembly has received much recent interest (Cornwell et al. 2009, He 2010, Lavorel et al. 2002, Shipley et al. 2006, Statzner et al. 2004, Weiher et al. 1998). Quantification of the effect of traits on the assembly of communities turns out to be challenging from a statistical point of view (Dray et al. 2008, Kleyer et al.).

Typical data in community ecology are arranged in two data tables: a table **Y** recording the occurrence and abundance of numerous species in sites and a table **X** recording habitat and other site characteristics, *i.e.* the values or states of numerous environmental variables at the sites (Fig.1). Such data are commonly used to study the relationships between species and environmental conditions, such as in species distribution models (Guisan et al. 2005, Guisan et al. 2000) and direct and indirect gradient analysis (ter Braak et al. 2004). Such models are powerful tools in investigating the possible consequences of changes in land-use and climate change on the distribution of species (Guisan et al. 2005, Raxworthy et al. 2003, Thuiller et al. 2005). They are also an important ingredient of conservation planning and management (Carroll et al. 2001, Johnson et al. 2004, Raxworthy et al. 2003).

Additional insight in why the species are distributed the way they are and why the species respond to changes in the way they do might be gained by adding information on traits of species, that is by adding a third table **Z** (Fig. 1), a matrix with values and states of numerous species traits (Dray et al. 2008, Legendre et al. 1997). A trait is a well-defined property of organisms that is usually measured at the organism level and used comparatively across species (McGill et al. 2006). On neglecting the intra-species variability when it is small compared to the inter-species variable (Garnier et al. 2001), a trait is a species property (Kleyer et al. 2008). Intra-species variability can be considerable (e.g. Albert et al. 2011, de Bello et al 2011) but we do not consider it in this paper. If traits are important in structuring communities, then the composition of local communities should be a non-random sample from the regional species pool (Ozinga et al. 2005a, Shipley et al. 2006). Environmental conditions, such as nutrient availability and soil moisture for plants, can act as filters that alter the probabilities of species to enter a local community according to their trait states (Cornwell et al. 2009, Ozinga et al. 2004, Ozinga et al. 2005b, Weiher et al. 1998). Many empirical studies have shown that species traits are associated with habitat conditions (Pöyry et al. 2008, Townsend et al. 1997, Townsend et al. 1994) in accordance with the theory that species in the evolutionary process adapt to the habitat and landscape characteristics in which the species occur (Ackerly 2003, Southwood 1977). Our interest is in efficient joint analysis of the three data tables (Fig. 1) using modern mainstream statistical methods. We will do so by adding species traits to models that relate species to the environment.

The key questions when modeling species, traits and environments are: (a) how does the expected abundance of species depend on trait and environmental values and (b)

which traits and environmental variables best explain the distribution of abundance in space and time and (c) to what extent are traits associated / correlated with environmental variables (Legendre et al. 1997). For modeling different approaches are used. Key question (c) can be address at the community (site) level or the species level (Kleyer et al. 2012). At community level, the sites × species table **Y** and species × trait table **Z** can be combined in to a sites × trait table which is then related to the sites × environment table **X** by standard statistical methods (Díaz et al. 1992, Sonnier et al. 2010). At the species level, the sites × species table **Y** and sites × environment table **X** can be combined in to species × environment table which is then related to the species × trait table **Z** by standard statistical methods (Kleyer et al. 2012). Legendre et al. (1997) and Dray & Legendre (2008) integrated these two steps in to one, the fourth-corner problem, in which they fill the trait × environment corner that is missing in Fig. 1. The entries of the missing corner table are Pearson correlations between traits and environmental variables, when quantitative, calculated from an inflated table. This approach naturally combines the community and species level approaches (Dray & Legendre 2008). Statistical testing is however a challenge. What is the unit of analysis, site or species or even, as the vectorized version of the fourth corner problem of Fig. 1 in Dray & Legendre (2008) suggests, each (non-zero) entry of table **Y**? So what to permute in permutation tests? Dray & Legendre (2008) examined six permutation-based methods to test the statistical significance of the trait-environment relationship, but none of them truly controlled the type I error. This defect was recently repaired (ter Braak et al. 2012). The multivariate version of the fourth-corner problem is the RLQ ordination (Dolédec et al. 1996, Choler 2005) which has been used for selecting the best traits in plant functional trait analyses (Bernhardt-Romermann et al. 2008). These methods focus on key question (c); they demonstrate trait-environment relationships but can hardly be used for predicting community composition from specified traits and environmental values.

A focus on key question (a) can be found in Shipley et al (2006) and Ozinga et al. (2005). Shipley et al. (2006) used the above mentioned sites × trait table in a novel way as a macroscopic feature of communities to predict species abundance in sites by the maximum entropy principle. The result is a logistic model relating abundances to traits, as used in logistic regression but fitted in a different way (He 2010) and without environmental variables. Shipley (2010) extended his model by adding environmental variables to deal with questions (b) and (c).

Ozinga et al. (2005a) started from the multiple logistic regression method and used it to quantify the effect of functional traits in a way that accounts for spatial variation in the composition of the local species pool. In this approach, the matrix **Y** is vectorized, that is, each entry is taken as a separate unit of analysis. An early example of such an approach is given by Nygaard & Ejrnæs (2004) and recent one by Pollock et al. (2012). Their methods assume that species records within sites are independent (Hosmer et al. 2000), thus commits pseudo-replication (Crawley 2002, Hurlbert 1984). In applying generalized linear models, such as logistic regression, researchers often ignore the hierarchical structure of the data thereby producing incorrect variance estimates and increasing the likelihood of committing type I error (Gillies et al. 2006, Wagner et al. 2005).

To address all three key questions, we develop a dedicated Generalized Linear Mixed Model (GLMM), very much in line with the very nice recent paper of Pollock et al. (2012), of which we were unaware, but we take into account the pseudo-replication introduced by vectorizing **Y**. GLMMs are as very general powerful class of statistical

models in ecology and elsewhere (Bolker et al. 2009, Gelman et al. 2007, Zuur et al. 2009). We introduce our GLMM as the result of integrating a two-step procedure into one, so obtaining a GLMM with main effects for traits and environmental variables as well as interaction effects between them. The GLMM utilizes species trait data efficiently and overcomes the problem of pseudo-replication (Paterson et al. 2003). The main advantage of using a GLMM approach is that standard software and methodology for model selection and model checking becomes available to address the key questions. We will illustrate the model and model selection in the main text using presence-absence versions of two data sets, one of which was analyzed by seven rival methods in Kleyer et al (2012). Count and multinomial data can be used as well, as we show in Appendix S1 in Supplementary Information, and provide more information on the relative fitness of the species (Shipley 2011). For fitting the model we use the library lme4 (Bates et al. 2011) in the free software package R (R Development Core Team 2010). Other statistical packages with good GLMM facilities include SAS proc glimmix (Stroup 2011) and Genstat (http://www.vsni.co.uk/software/genstat/).

Methods

The data sets

The first data set is the Dune Meadow data (Jongman et al. 1995). This is a small data set of 28 higher plants in 20 sites with five environmental variables and five species traits (Appendix S2), originating from the Dutch island Terschelling. Four environmental variables are treated as quantitative (abbreviation between parentheses): thickness of the A1 horizon (A1), moisture content of the soil (Moisture), agriculture grassland use (Use) and quantity of manure applied (Manure). One environmental variable was categorical: grassland management type (Mag) with four classes (SF: standard farming, BF: biological farming, HF: hobby farming and NM: nature conservation management). Four traits are quantitative : Specific leaf area (SLA) canopy height of a shoot (Height), leaf dry matter content (LDMC), Seed mass. One trait is categorical: life span with categories annual and perennial. Traits were taken from the LEDA database (Kleyer et al. 2008) and Lienin & Kleyer (2011).

The second dataset is the grazed semi-natural grassland data from NE Germany, taken from Kleyer et al (2012) and comprised 50 plant species in 43 sites, with environmental variables grazing intensity (Grazing), soil water holding capacity (Water), extractable phosphorous (Soil P) and the following species traits: canopy height, specific leaf area (SLA), seed mass (log-transformed) (Seed mass), leaf C : N ratio (C : N ratio), onset of flowering date (Onset), flowering mode (Polycarpic/Monocarpic). For detail see Kleyer et al (2012).

The generalized linear mixed model

In this section, we derive our generalized linear mixed model (GLMM) from a two-step approach. The data we consider is a binary data table $\mathbf{Y} = [y_{ij}]$ recording the presence (1) -absence (0) of m species (columns) in n sites (row), an environmental variable $\mathbf{x} = [x_i]$ with quantitative measurements in the n sites, and a quantitative trait $\mathbf{z} = [z_j]$ with quantitative values for the m species. The subscripts i and j refer to site i and species j, respectively.

A natural way to study the relationship between a trait and an environmental variable on the basis of species presence-absence data is in two steps, consisting of

1. fitting, for each species separately, a logistic regression of its presence-absence against the environmental variable *x* and
2. regressing parameters retrieved from the *m* logistic regressions on to the trait *z*.

In its simplest form, the first step involves a linear-logistic regression and models the probability of occurrence as a function of the environmental variable. The first stage of two stage approach assumes that

$$\text{logit}(p_{ij}) = \alpha_j + \beta_j x_i \,,\; i = 1,2,.....,n \,,\; j = 1,2,.....,m \,, \tag{1}$$

with p_{ij} = Pr(y_{ij}=1) the probability of occurrence of species *j* in site *i*, α_j and β_j the intercept and slope for j^{th} species with respect to environmental variable *x* and $\text{logit}(p_{ij}) = \log(p_{ij}/(1-p_{ij}))$, the logistic function. Extensions of this simple model will be discussed later. This equation can be fitted to the presence-absence data of each species separately, resulting in *m* separate models for the probability of occurrence of the species as a function of the environmental variable *x*. In this model, the relationship of a species with the environment is summarized by the slope β_j. Its sign indicates whether the probability of occurrence increases or decreases with increasing value of *x* and its size how strongly. In its simplest form, the second step involves a (possible weighted) linear regression of the estimated regression slope coefficients $\{\beta_j\}$ on to the trait with the model

$$\beta_j = b_0 + b_1 z_j + \varepsilon_{\beta j} \;,\; j = 1,2,.....,m \;, \tag{2}$$

with b_0 and b_1 intercept and slope respectively and error $\varepsilon_{\beta j}$, normally distributed with zero mean and variance σ_β^2, *i.e.* $\varepsilon_{\beta j} \sim N(0, \sigma_\beta^2)$. The subscript β is added to the error term to distinguish it from other error terms later on. (The weights are the inverse of the squared standard errors of estimate of $\{\beta_j\}$ in step 1). Another way of expressing equation (2) is that the slope coefficient of the species *j* with trait value z_j is normally distributed with mean $b_0 + b_1 z_j$ and variance σ_β^2, *i.e.*

$$\beta_j \sim N(b_0 + b_1 z_j, \sigma_\beta^2) \,. \tag{3}$$

But, equations (1) and (3) together form an example of a generalized linear mixed model (GLMM) and can thus be integrated and estimated simultaneously.

So far the second step only modeled the slopes, because of the particular interest in the trait-environment relationship, but we may also be interested in the influence of the trait on the overall probability of occurrence of a species. The intercept α_j in equation (1) plays such a role, in particular when the environmental variable *x* is centered prior to the analysis, as $\text{logit}^{-1}(\alpha_j) = \exp(\alpha_j)/(1+\exp(\alpha_j))$ is the probability of occurrence at mean *x*. Analogously to equation (2), we could linearly regress the estimated intercepts $\{\alpha_j\}$ on to the trait with the model

$$\alpha_j = a_0 + a_1 z_j + \varepsilon_{\alpha j} \;,\; j = 1,2,.....,m \tag{4}$$

with a_0 and a_1 intercept and slope, respectively and $\varepsilon_{\alpha j}$ normally distributed with zero mean and variance σ_α^2. As in equation (3) we rewrite this as

$$\alpha_j \sim N(a_0 + a_1 z_j, \sigma_\alpha^2) . \quad (5)$$

Equations (1), (3) and (5) together form another example of a generalized linear mixed model (GLMM). As a GLMM this model still has two shortcomings. First, it assumes that the intercept α_j and slope β_j are independent. This makes the model dependent on the scale of the environmental variable (centered or non-centered) which is undesirable, so we complete the model with a correlation ρ between them. Second, it assumes that the presence-absences of different species at the same site (given their trait values and x_i) are independent. The usual way to introduce correlation among them is with a common site-specific parameter γ_i that is assumed to be normally distributed with mean zero and variance σ_γ^2. With this parameter included, the GLMM equations become

$$\text{logit}(p_{ij}) = \alpha_j + \beta_j x_i + \gamma_i , \; i = 1,2,.....,n , \; j = 1,2,.....,m , \quad (6)$$

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} a_0 + a_1 z_j \\ b_0 + b_1 z_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right),$$

$$\gamma_i \sim N(0, \sigma_\gamma^2)$$

This completes our derivation of the GLMM that models the species presence as a function of both the environmental variable *x* and trait variable *z*. In the GLMM literature the model is called a random intercept and random slope model. This GLMM combines both steps of the two-step approach into a single model and avoids pseudo-replication by including site as a random effect.

Testing and interpreting the trait-environment relationship

Here we show that the trait-environment relationship is an interaction term in the model that can be tested for statistical significance using standard software.

By inserting equations (4) and (2) in equation (6) we obtain

$$\begin{aligned} \text{logit}(p_{ij}) &= (a_0 + a_1 z_j + e_{\alpha j}) + (b_0 + b_1 z_j + e_{\beta j}) x_i + \gamma_i \\ &= a_0 + a_1 z_j + b_0 x_i + b_1 z_j x_i + e_{\alpha j} + e_{\beta j} x_i + \gamma_i \end{aligned} \quad (7)$$

with fixed coefficients in Roman and random coefficients in Greek. This model for the probability of occurrence contains main effects for the trait *z* and the environmental variable *x* and an interaction *z.x* between them. This interaction represents the trait-environment relationship. The model also contains random terms for species ($\varepsilon_{\alpha j}$), sites (γ_i) and the environment-by-sites interaction ($\varepsilon_{\beta j} x_i$). We need to specify all effects and random terms to fit the model to data. With the lme4 library of the software package R the specification of equation (7) is

```
M1 <- glmer(y ~ z + x + z:x +(1+x|species)+(1|sites),
      family=binomial(link="logit"), data)
```

with y, z and x vectors with *nm* elements and species and sites factors with *m* and *n* levels respectively. The terms within brackets are random, the others are fixed. Library lme4 uses vector notation, *i.e.* y is the matrix $\mathbf{Y} = [y_{ij}]$ written as a vector; the species and site factors code to which species and site each element of y belongs; the value x_i of the environmental variable repeated at all *m* elements that code for site *i* and the value z_j of the trait repeated at all *n* elements that code for species *j* (Appendix S2). To test the trait-environment interaction (with null-hypothesis: $b_1 = 0$), we also fit the model without this term by

```
M0 <- glmer(y ~ z + x +(1+x|species)+(1|sites),
      family=binomial(link="logit"), data)
```

and then compare the two models by an analysis of variance statement anova(M0,M1), resulting in a P-value for the likelihood ratio (LR) test of model M1 against M0.

The estimates of the variance σ_β^2 in model M0 and M1 can be usefully compared to express the contribution of the trait to the inter-species variance in the slope parameter by the coefficient (Grosbois et al. 2009, Lahoz-Monfort et al. 2011)

$$C_\beta = 1 - \frac{\hat{\sigma}_\beta^2(\text{res})}{\hat{\sigma}_\beta^2(\text{total})} \tag{8}$$

where $\hat{\sigma}_\beta^2(\text{res}) = \hat{\sigma}_\beta^2$ in model M1 and $\hat{\sigma}_\beta^2(\text{total}) = \hat{\sigma}_\beta^2$ in model M0. The rationale is that σ_β^2 is the residual variance in equation (2), the inter-species variance of the slope parameter after taking account of the trait and therefore denoted as $\sigma_\beta^2(\text{res})$. In model M0, $b_1 = 0$ in equation (2), so that σ_β^2 represents the total variance denoted by $\sigma_\beta^2(\text{total})$.

We investigated the type I error and power of the statistical tests on trait-environment interaction. We simulated 1000 new datasets of the same size and the same environment and trait values as the Dune Meadow data. The data $\{y_{ij}\}$ were simulated using the GLMM model of equation (7) with parameters and variance components equal to the estimated ones, i.e. those of model M0 and M1 for the type I error and power calculations, respectively. We did not observe much difference between the test based on the z-statistic and the LR test and report the latter only.

So far, the environmental variable and the species trait were both quantitative. GLMM can also be applied when both are qualitative or when one is quantitative and the other qualitative (Appendix S3). A difference is that, for a qualitative environmental variable, each class beyond the first comes with its own variance component and the trait-environment interaction then consist of more than one regression parameter, but neither difference presents a problem to LR testing and further interpretation. Models with a single trait and a single environmental variable generate simple trait-environment effects as opposite to the conditional trait-environment effects in multi-trait multi-environment models of the next subsection.

Model selection with many environmental variables and traits

The GLMM of equation (7) can readily be extended to more traits and environmental variables by including a) main effects for all traits and environmental variables, b) interactions between each trait and each environmental variable, and c) species-dependent random terms for each environmental variable. Conceptually such a model can still be viewed as one with slope coefficients with respect to each of the environmental variables, which are then each (separately) regressed on to the traits. GLMM does a joint fit of such a model. Such a GLMM has far less problems with sparsity of the data (few presences, many absences) as the usual GLM, because regression coefficients, when made random, are shrunken towards zero. With two environmental variables (x1 and x2) and three traits (z1, z2 and z3), this model that can be specified in lme4 by

```
glmer(y ~ (z1+z2+z3)*(x1+x2)+(1+x1+x2|species)+(1|sites),
     family=binomial(link="logit"), data)
```

This model contains trait and environmental variable main effects and their interactions, (correlated) species-dependent random effects for all environmental variables and independent random effects for species and sites. Traits and environmental variables can be a mix of being quantitative and/or qualitative (see Appendix S3 for an example). This approach works as long as the number of environmental variables is small (particularly when compared with the number of sites) and the number of traits is smaller than the number of species. The method will benefit from many species as it makes the estimation of variance components robust. Within these constraints, we experienced limited convergence problems. To circumvent these constraints we present a model selection procedure.

With many traits and environmental variables, a natural question is to select a minimal model that describes the species occurrences data well and, related to this, to select the traits that explain the species response to relevant environmental variables. For an RLQ approach to the latter see Bernhardt–Römerman et al. (2008). These questions can be solved by model selection (Diggle et al. 2002, West et al. 2006). The number of candidate models increases exponentially with the number of predictors (traits, environmental variables, interactions and variance components) so an exhaustive search is feasible only for low numbers of predictors. Alternatives are forward and backward selection. Backward selection would start with the model with all terms included. This 'full' model may be difficult to fit, due to convergence problems, unless the number of environmental variables is small. For example we were unable to fit the full model to the Dune Meadow data using lme4. Therefore Jamil et al. (2012) proposed a tiered forward selection approach that we now describe. But first we must settle the criterion to rank models with different numbers of parameters. The most commonly used information criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (Broman et al. 2002) which is defined as minus two log-likelihood plus c times the number of degrees of freedom (df) with $c = 2$ for AIC and $c = \log(N)$ for BIC. We must thus look for models with the lowest information. The problem with BIC in GLMMs is what to choose for N, the number of observations as the observations are no longer assumed to be independent. Should it be the number of sites, the number of species or

their product? Jamil et al. (2012) used a variant, SigAIC, which multiplies df by $c = \chi^2_{1(0.05)} = 3.84$ (Broman et al. 2002). With SigAIC, the addition of a single parameter to a model will result in a lower SigAIC value if and only if that parameter is significant at the 5% level as judged by the LR test.

The tiered forward selection of Jamil et al. (2012) starts with the null model with only random effects for species and sites and then adds in each step the environmental variable for which the species-dependent random terms most decreases the information. So, in the first tier the model is that of equation (6) with random coefficients α_j and β_j. This process of adding random environmental terms is continued until information no longer decreases. At this first tier, the main effects of traits and environmental variables are not considered because the random species and site effects can already partly take account of them. After this first tier, the choice for the random part of the model is complete. In the second tier, we consider only the trait-environment interactions of environmental variables that were selected in the first stage. The reason is that the importance of the trait-environment coefficient (b_1) can only be judged against the unexplained variation in the slope coefficients $\{\beta_j\}$, as can best be seen from equation (2). At the start of the second tier, the main effects of all environment variables selected in the first tier are added. Thus the environmental variables are then component of both fixed effects and random effects. In each subsequent step we then search for the trait-environment interaction that most decreases the information. When the associated trait main effect is not yet in the model, it is added jointly with the interaction term and the resulting information is evaluated. This process is continued until the information does no longer decrease. In a final third tier any non-significant interaction effects are sequentially removed. R code for this method of model selection is given in Appendix S5.

Fourth corner and RLQ

Dray & Legendre (2008) present four interpretations of the fourth corner statistics. The one that is closest to our context is that, for quantitative variables,

it calculates a weighted Pearson correlation between the trait and the environmental variable in an inflated data table, that is, by using all species-site combinations as cases, weighted by abundance, and by assigning to each case the trait and the environmental value of the combination. This inflation is the same as used for fitting a GLMM with one difference. As absence implies zero abundance and zero weight, absences in table **Y** do not count and can be disregarded in fourth corner (and RLQ), whereas they cannot be disregarded in GLMM.

RLQ is a general three-table ordination method based on a eigen space analysis (Dolédec et al. 1996). In the usual notation the three tables are **R**, L and **Q** which correspond to **X**, **Y** and **Z**, respectively, in our notation. We used the most commonly used version of RLQ, namely the one based on a correspondence analysis of the central table **Y**. It is this version that is closely related to the fourth corner statistics (Dray & Legendre 2008). For quantitative trait and environmental variables, the essential part of RLQ reduces to a singular value decomposition of the table of fourth corner correlations. An often neglected interpretation of the usual RLQ ordination diagram of traits and environmental variables (Kleyer et al. 2012) is that it is thus an (unweighed least-squares) biplot (Jongman et al. 1995) of the table of fourth corner correlation. RLQ and fourth corner were carried out with R-package ade4 (Dray & Dufour 2007) with permutation tests as described in ter Braak et al. (2012).

Results

Models with one trait-environment term

Table 1 illustrates the GLMM results for the Dune Meadow data (Jongman et al. 1995) using Manure as environmental variable and SLA as trait value. The row of prime importance is that of the interaction Manure:SLA. The estimate of b_1 is positive (0.06) showing that species with high SLA have a higher slope coefficient with respect to Manure than species with low SLA. The occurrence probability of species with high SLA thus increases more with Manure than that of species with low SLA. The associated z-value (estimate/standard error= 3.28) indicates that the interaction is statistically significant ($P<0.01$, despite the small sample size), so that the true interaction is unlikely to be zero. The LR test (Table S1) confirms that the interaction is significant ($P < 0.01$). In a model with an interaction, the size and sign of main effects and their significance depend on the scales of the variables and we explain the interpretation in Appendix S1. It is thus of little importance that the coefficient for SLA in Table 1 is not significant.

Fig. 2 displays how the fitted occurrence probability depends on Manure for some selected species with and without usage of the trait SLA in the model (without SLA, $a_1 = 0$ and $b_1 = 0$ in equation (7)). In general the fitted curves of the two model differ little, because both include the species-dependent random slope (β_j) with respect to manure. The largest difference occurs for the species which have few presences and extreme SLA. For example, due to its low SLA value (10.3) the curve for *Eleocharis palustris* with using SLA is stronger decreasing than without SLA. Fig. 3 plots the species-dependent random slopes (β_j) against SLA and shows the scatter of the around the fitted line according to model M1 (equations (3) and (6)). The slopes fitted by GLM have a much wider scatter and the fitted line in this two-step approach is consequently less steep. The GLM-slopes are extremely large in absolute value (>3) for species with few occurrences. In the GLMM approach the slopes are shrunken towards the common regression line; the vertical deviations from the line are summarized by the parameter $\sigma_\beta = 0.36$ in equation (6). In model M0 (with the manure-SLA interaction) σ_β is estimated as 0.49. According to equation (8), SLA accounts for 46% of the inter-species variance in the species response to Manure.

The simulated type I error was 3.5% and 1% for significance levels 5% and 1%, respectively. The power of the test was reasonably high (80% at a signifance level of 5% and 59% at 1%). Appendix S4 describes a small simulation study comparing GLMM and the two step approach in more detail and shows that the standard error of the Manure:SLA interaction in the GLMM well represents the variability seen across parametric bootstrap samples.

Tables 3 list the sign and significance of single trait-environment terms as found by GLMM and fourth corner in the Dune Meadow. In Table 3, GLMM and fourth corner agree on three relationships: SLA with Management type and with Manure, and Seed mass with Moisture). In addition, GLMM detects SLA-Moisture and fourth corner detects hLDMC-Use. In a similar table (Table 4) for the Grazed Grassland data, GLMM and fourth corner agree on six relationships. GLMM finds two others that are marginally non-significant in fourth corner (both P=0.07). Fourth corner finds one other term significant which is non-significant (P = 0.24) in GLMM.

Models with multiple trait-environment terms

We now build a parsimonious regression model from all available trait and environment variables using tiered forward selection of environmental variables and traits. Table 2 illustrates the results for the Dune Meadow data. In the first tier Moisture and Manure are selected, where after no environmental variable decreases SigAIC. At the start of the second tier the main effects of Moisture and Manure are added. The best interaction to add was that between Manure and SLA (Manure:SLA). One more interaction (between Moisture and Seed mass) further decreased SigAIC. As all interaction terms were significant at the end of the second tier, nothing was deleted in the third tier. The sign of the interaction terms in the final model (Table S2) show that SLA is positively related to Manure ($P < 0.001$) and Seed mass negatively related to Moisture ($P < 0.02$). The variance estimates of the random slopes with respect to Moisture and Manure in models with and without trait-environment interactions shows that the traits account for 21% and 22% of the variance in species response to Moisture and Manure, respectively. The value for Manure is surprisingly low as in the comparable model without Moisture and Seed mass, SLA accounted for 45% of the variance. This presumably due to the correlation (0.55) between the random effects of Moisture and Manure in the final model. For diagnostic checks we made a Q-Q plot of the random effects (random slopes for species with respect to environmental variables as deviations from their common slope) to check normality. Some non-normality is visible but not seriously (Fig. S1).

In the Grazed Grassland data, the first tier selected Soil P first and then Grazing, with drops in SigAIC of 117 and 102, respectively. The variable Water did not improve the model (it increased SigAIC by 1.9). The second tier selected the interaction between Soil P and Polycarpic and then that between Grazing and C : N ratio, with drops in SigAIC of 25 and 8, respectively. The next interaction to enter would have been Grazing with Polycarpic, but that addition increased SigAIC by 2. Both selected interactions had a negative sign and were very significant ($P < 0.0001$; Table S3). The traits account for 60% of the variance in species response to Soil P and 25% of the variance in Grazing. The Q-Q plot of the random effects highlights some non-normality for the random slope deviations with respect to Soil P.

RLQ

The two main axes of RLQ explain applied to the Dune Meadow data explains 63% and 27% of the co-inertia (together 90%). The RLQ ordination diagram of traits and environmental variables (Fig. 4) is a biplot of the trait × environment table of fourth corner statistics (missing corner in Fig. 1). By their long arrows the classes of management type are shown to have strong correlations with SLA, Seed mass and perhaps Height. Table 3 showed that on their correlations with SLA were statistically significant (P = 0.004) but the others were not (Seed mass: P= 0.11 and Height:

P=0.27). Moisture and Manure also have relatively long arrows and show the positive correlation between SLA with Manure and the negative correlation between Moisture and Seed mass.

The RLQ of the presence-absence version of the Grazed Grassland data (Fig. S3) looked similar to the one based on abundance, shown in Fig. 2 of Kleyer et al. (2012), although Onset stands out less. The arrows for Grazing and Soil P (or their distance from the origin) are twice or more the length of the arrow for Water, thus showing the traits are more related to grazing and phosphorus than to the water holding capacity. The biplot nicely shows the strong negative correlation between Soil P and Polycarpic and between Grazing and C : N ratio (Table 4). It also show that, of all traits, C : N ratio has correlation closest to zero with Soil P, but it is of course not clear that it is the only one that is non-significant (P = 0.17).

Discussion

In this paper, we showed how GLMM can be applied for modeling and explaining species response along environmental gradients by species traits. It is based on a sound statistical model that allows, as a standard by-product, questions to be answered about which traits and environmental variables are significantly related and which best explain the species response in a parsimonious model.

GLMM accounts for pseudo-replication and heteroscedastic variance by including sites and species as random factors. Our GLMM approach can be understood as a two-step approach executed at once. In the first step species response is related to the environment and in the second step the (multivariate) outcome of the first step is related to the trait data. The integration of these two steps into one has several advantages: GLMM models directly the variable of interest (occurrence probability, expected abundance), it automatically weighs the different kinds of information for an optimal model fit and standard statistical significance testing and it provides consistent estimates of the between-species variance of (slope) parameters, without introducing unnecessary random variation by replacing the (slope) parameters by their estimates as in the two step approach and it can be applied with small sample size.

In comparison with separate regressions for each species (as in the first step of the two-step approach), the GLMM regression coefficients for each species tend to be pulled inward toward a common value; they are a compromise between the coefficients from a per-species fit and the population average. Such estimates are called shrinkage estimates (Pinheiro et al. 2000). The shrinkage is particularly evident for the species that have few presences. The estimates for these species lead to abnormally high estimates in the GLM fit (Fig. 3). The pooling of species in the GLMM estimation gives a certain amount of robustness to species with few occurrences in the data.

Our GLMM starts with a logistic linear model (Fig. 2) and is therefore most suitable along short environmental gradients. Such data sets are common in our experience. Moreover, the random component for sites (γ_i) allows for any common non-linearity with as prime exaple the niche model with equal niche width (de Rooij 2007, Ihm et al. 1984, ter Braak 1988), as we show now. Consider the simplest unimodal curve, the Gaussian logistic curve (ter Braak and Looman 1986)

$$\text{logit}(p_{ij}) = a_j - 0.5(x_i - u_j)^2 / t_j^2 \qquad (9)$$

with a_j a coefficient related to maximum probability of occurrence, u_j the species optimum and t_j the tolerance of species *j*. On assuming that the species have equal tolerance ($t_j = t$), expanding the square in equation (9) and setting $\alpha_j = a_j - 0.5u_j^2 / t_j^2$, $\beta_j = u_j / t^2$ and $\gamma_i = -0.5x_i^2 / t^2$, we obtain

$$\text{logit}(p_{ij}) = \alpha_j + \beta_j x_i + \gamma_i,$$

which is as equation (6) and can be viewed as GLMM. GLMMs can thus model unimodal species composition data without the need of squared or other nonlinear terms. By consequence, the trait model for the slope β_j of equation (2) thus implies a model for the optimum u_j in case species have (near) equal tolerance.

An alternative is to convert quantitative environmental variables to qualitative and model how the occurrence probabilities in the newly formed environmental categories depend on the traits, being either quantitative or qualitative. This approach fits in our proposed framework as illustrated in Supplementary Material. Another alternative, adding polynomial terms as random component to the model, is less attractive as it leads to coefficients that lack a clear interpretation.

We found a fair agreement between the simple trait-environment effects of GLMM and the one-by one relations found by fourth corner (Tables 3 and 4). Perhaps we were lucky. More disagreement can be expected when both the environmental gradients are longer and the species are more variable in niche width. The RLQ ordination diagram illustrated that RLQ is simply an ordination of the fourth corner statistics and therefore does not build a truly multi-trait multi-environment model in the sense of regression analysis and GLMM. GLMM model selection resulted in parsimonious models. For example, when in the Grazed Grassland data the interaction between flowering mode (polycarpic/monocarpic) and phosphorous is taken into account, the remaining traits to do not contribute much to explain the species response to phosphorous. It is unclear to us how such simplification could be achieved with RLQ-based methods. Such simplification may help ecological interpretation.

So far we did neither considered phylogeny, which puts constraints on the way traits may evolve in evolutionary time (Prinzing et al. 2008), nor the spatial configuration of the sites, which set constraints on dispersion (Dray et al. 2008, Ozinga et al. 2004). Both aspects can be modeled in a GLMM through random terms for species and sites whose correlation depends on either phylogenic association or spatial distance, respectively. In a recent paper, Ives & Helmus (2011) investigated the phylogenetic structure in community data in combination with either a single environmental variable or a single trait variable or no external data. Interestingly, Ives & Helmus (2011) did not account for phylogeny while testing trait effects. The pseudo-replication due to phylogenetic close species is thus not taken care of in their analyses (neither in ours). It would be of interest to extend their analysis to the case with both environmental and trait variables and to account for phylogeny in statistical tests of trait-environment interaction. These extensions merit further research, also in terms of practical software implementation.

Species traits are likely to have much predictive value for where and when a particular species or group of species appear or disappear. Our model-based approach makes this predictive usage practical and allows the selection of the traits and environmental conditions that matter.

Acknowledgements
We thank Bill Shipley, Francesco de Bello and two anonymous reviewers for constructive comments. Jamil's research was supported by a grant from Higher Education Commission of Pakistan through NUFFIC (The Netherlands).

References

Ackerly, D.D. 2003. Community assembly, niche conservatism, and adaptive evolution in changing environments. *International Journal of Plant Sciences* 164: S165-S184.

Albert C.H., de Bello F., Boulangeat I., Pellet G., Lavorel S. & Thuiller W. 2011. On the importance of intraspecific variability for the quantification of functional diversity. *Oikos*, doi: 10.1111/j.1600-0706.2011.19672.x

Bates, D., Maechler, M. & Bolker, B. 2011. lme4: Linear mixed-effects models using S4 classes. *R package version 0.999375-39, http://CRAN.R-project.org/package=lme4*.

Bernhardt-Romermann, M., Romermann, C., Nuske, R., Parth, A., Klotz, S., Schmidt, W. & Stadler, J. 2008. On the identification of the most suitable traits for plant functional trait analyses. *Oikos* 117: 1533-1541.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.S.S. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution* 24: 127-135.

Broman, K.W. & Speed, T.R. 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society Series B* 64: 641-656.

Carroll, C., Noss, R.F. & Paquet, P.C. 2001. Carnivores as focal species for conservation planning in the rocky mountain region. *Ecological Applications* 11: 961-980.

Choler, P. 2005. Consistent Shifts in Alpine Plant Traits along a Mesotopographical Gradient. *Arctic, Antarctic, and Alpine Research* 37: 444–453.

Cornwell, W.K. & Ackerly, D.D. 2009. Community assembly and shifts in plant trait distributions across an environmental gradient in coastal California. *Ecological Monographs* 79: 109-126.

Crawley, J.M. 2002. *Statistical Computing: An Introduction to Data Analysis using S-Plus*. John Wiley & Sons, New York.

de Rooij, M. 2007. The distance perspective of generalized biadditive models: Scalings and transformations. *Journal of Computational and Graphical Statistics* 16: 210-227.

de Bello, F., Lavorel, S. Albert, C. H., Thuiller, W., Grigulis, K., Dolezal, J. Janecek, S., &. Leps, J. 2011. Quantifying the relevance of intraspecific trait variability for functional diversity. *Methods in Ecology and Evolution* 2:163-174.

Díaz, S., Acosta, A. & Cabido, M. 1992. Morphological Analysis of Herbaceous Communities under Different Grazing Regimes. *Journal of Vegetation Science* 3: 689-696.

Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S. 2002. *The Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Dolédec, S., Chessel, D., ter Braak, C.J.F. & Champely, S. 1996. Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics* 3: 143-166.
Dray, S. & Dufour, A.B. 2007. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software* 22: 1-20.
Dray, S. & Legendre, P. 2008. Testing the species traits environment relationships: The fourth-corner problem revisited. *Ecology* 89: 3400-3412.
Garnier, E., Laurent, G., Bellmann, A., Debain, S., Berthelier, P., Ducout, B., Roumet, C. & Navas, M.L. 2001. Consistency of species ranking based on functional leaf traits. *New Phytologist* 152: 69-83.
Gelman, A. & Hill, J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
Gillies, C.S., Hebblewhite, M., Nielsen, S.E., Krawchuk, M.A., Aldridge, C.L., Frair, J.L., Saher, D.J., Stevens, C.E. & Jerde, C.L. 2006. Application of random effects to the study of resource selection by animals. *Journal of Animal Ecology* 75: 887-898.
Grosbois, V., Harris, M.P., Anker-Nilssen, T., McCleery, R.H., Shaw, D.N., Morgan, B.J.T. & Gimenez, O. 2009. Modeling survival at multi-population scales using mark-recapture data. *Ecology* 90: 2922-2932.
Guisan, A. & Thuiller, W. 2005. Predicting species distribution: Offering more than simple habitat models. *Ecology Letters* 8: 993-1009.
Guisan, A. & Zimmermann, N.E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* 135: 147-186.
He, F.L. 2010. Maximum entropy, logistic regression, and species abundance. *Oikos* 119: 578-582.
Hosmer, W.D. & Lemeshow, S. 2000. *Applied Logistic Regression, 2nd Edition*. Wiley, New York.
Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187-211.
Ihm, P. & van Groenewoud, H. 1984. Correspondence analysis and Gaussian ordination. *Compstat Lectures* 3: 5-60.
Ives, A. & Helmus, M. 2011. Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs* doi:10.1890/10-1264.1: null.
Jamil, T., Opdekamp, W., van Diggelen, R. & ter Braak, C.J.F 2012. Trait-Environment Relationships and Tiered Forward Model Selection in Linear Mixed Models. *International Journal of Ecology*. doi:10.1155/2012/947103
Johnson, C.J., Seip, D.R. & Boyce, M.S. 2004. A quantitative approach to conservation planning: Using resource selection functions to map the distribution of mountain caribou at multiple spatial scales. *Journal of Applied Ecology* 41: 238-251.
Jongman, R.H.G., ter Braak, C.J.F. & van Tongeren, O.F.R. 1995. *Data analysis in community and landscape ecology*. Cambridge University Press, Cambridge.
Kleyer, M., Bekker, R.M., Knevel, I.C., Bakker, J.P., Thompson, K., Sonnenschein, M., Poschlod, P., van Groenendael, J.M., Klimes, L., Klimesova, J., Klotz, S., Rusch, G.M., Hermy, M., Adriaens, D., Boedeltje, G., Bossuyt, B., Dannemann, A., Endels, P., Gotzenberger, L., Hodgson, J.G., Jackel, A.K., Kuhn, I., Kunzmann, D., Ozinga, W.A., Romermann, C., Stadler, M., Schlegelmilch, J., Steendam, H.J., Tackenberg, O., Wilmann, B., Cornelissen, J.H.C., Eriksson, O., Garnier, E. & Peco, B. 2008. The LEDA Traitbase: A

database of life-history traits of the Northwest European flora. *Journal of Ecology* 96: 1266-1274.
Kleyer, M., Dray, S., Bello, F., Lepš, J., Pakeman, R. J., Strauss, B., Thuiller, W. & Lavorel, S.. 2012. Assessing species and community functional responses to environmental gradients: which multivariate methods? *Journal of Vegetation Science.* 10.1111/j.1654-1103.2012.01402.x
Lienin, P. & Kleyer, M. 2011. Plant leaf economics and reproductive investment are responsive to gradients of land use intensity. *Agriculture, Ecosystems and Environment* 145: 67- 76.
Lahoz-Monfort, J.J., Morgan, B.J.T., Harris, M.P., Wanless, S. & Freeman, S.N. 2011. A capture–recapture model for exploring multi-species synchrony in survival. *Methods in Ecology and Evolution* 2: 116-124.
Lavorel, S. & Garnier, E. 2002. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Functional Ecology* 16: 545-556.
Legendre, P., Galzin, R.G. & Harmelin-Vivien, M.L. 1997. Relating behavior to habitat: Solutions to the fourth-corner problem. *Ecology* 78: 547-562.
McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. 2006. Rebuilding community ecology from functional traits. *Trends in Ecology and Evolution* 21: 178-185.
Nygaard, B. & Ejrnæs, R. 2004. A new approach to functional interpretation of vegetation data. *Journal of Vegetation Science* 15:49-56.
Ozinga, W.A., Bekker, R.M., Schaminee, J.H.J. & Van Groenendael, J.M. 2004. Dispersal potential in plant communities depends on environmental conditions. *Journal of Ecology* 92: 767-777.
Ozinga, W.A., Hennekens, S.M., Schaminee, J.H.J., Bekker, R.M., Prinzing, A., Bonn, S., Poschlod, P., Tackenberg, O., Thompson, K., Bakker, J.P. & van Groenendael, J.M. 2005a. Assessing the relative importance of dispersal in plant communities using an ecoinformatics approach. *Folia Geobotanica* 40: 53-67.
Ozinga, W.A., Schaminee, J.H.J., Bekker, R.M., Bonn, S., Poschlod, P., Tackenberg, O., Bakker, J. & van Groenendael, J.M. 2005b. Predictability of plant species composition from environmental conditions is constrained by dispersal limitation. *Oikos* 108: 555-561.
Paterson, S. & Lello, J. 2003. Mixed models: Getting the best use of parasitological data. *Trends in Parasitology* 19: 370-375.
Pinheiro, J.C. & Bates, D.M. 2000. *Mixed-Effects Models in S and SPLUS*. Springer, Berlin.
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.
Pöyry, J., Luoto, M., Heikkinen, K.R. & Saarinen, K. 2008. Species traits are associated with the quality of bioclimatic models. *Global Ecology and Biogeography* 17: 403-414.
Prinzing, A., Reiffers, R., Braakhekke, W.G., Hennekens, S.M., Tackenberg, O., Ozinga, W.A., Schaminee, J.H.J. & van Groenendael, J.M. 2008. Less lineages - more trait variation: phylogenetically clustered plant communities are functionally more diverse. *Ecology Letters* 11: 809-819.
R Development Core Team 2010. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. www.R-project.org, Vienna.

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

Shipley, B., Vile, D. & Garnier, E. 2006. From plant traits to plant communities: A statistical mechanistic approach to biodiversity. *Science* 314: 812-814.

Shipley, B. 2010. From plant traits to vegetation structure : chance and selection in the assembly of ecological communities. Cambridge University Press, Cambridge.

Sonnier, G., Shipley, B. & Navas, M.L. 2010. Quantifying relationships between traits and explicitly measured gradients of stress and disturbance in early successional plant communities. Journal of Vegetation Science 21: 1014-1024.

Southwood, T.R.E. 1977. Habitat, the templet for ecological strategies? *Journal of Animal Ecology* 46: 337-365.

Statzner, B., Dolédec, S. & Hugueny, B. 2004. Biological trait compostion of European stream invertebrate communities: assessing the effects of various trait filter types. *Ecography* 27: 470-488.

Stroup, W.W. 2011. Living with Generalized Linear Mixed Models. *SAS Global Forum 2011* Paper 349: 1-18.

ter Braak, C.J.F. 1988. Partial canonical correspondence analysis. In: Bock, H.H. (ed.) *Classification and related methods of data analysis*, pp. 551-558. North-Holland, Amsterdam.

ter Braak, C.J.F. & Prentice, I.C. 2004. A theory of gradient analysis. *Advances in Ecological Research* 34: 235-282.

ter Braak, C.J.F., Cormont, A. & Dray, S. 2012. Improved testing of species traits-environment relationships in the fourth corner problem. *Ecology*. 10.1890/12-0126.1

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

Townsend, C.R., Dolédec, S. & Scarsbook, M.R. 1997. Species traits in relation to temporal and spatial heterogeneity in streams: a test of habitat templet theory. *Freshwater Biology* 37: 367-387.

Townsend, C.R. & Hildrew, A.G. 1994. Species traits in relation to a habitat templet for river systems. *Freshwater Biology* 31: 265-275.

Wagner, H.H. & Fortin, M.J. 2005. Spatial analysis of landscapes: Concepts and statistics. *Ecology* 86: 1975-1987.

Weiher, E., Clarke, G.D.P. & Keddy, P.A. 1998. Community assembly rules, morphological dispersion, and the coexistence of plant species. *Oikos* 81: 309-322.

West, B.T., Welch, K.B. & Galecki, A.T. 2006. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall/CRC, London.

Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A. & Smith, G.M. 2009. *Mixed Effects Models and Extensions in Ecology with R*. Springer Berlin.

Supporting information

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

Appendix S1. The GLMM trait model for count data and multinomial data

Appendix S2. Dune meadow data in matrix form (Y, X and Z) and in vector notation for lme4.

Appendix S3. The GLMM trait model with qualitative and/or qualitative trait and environmental variable with interpretation of regression coefficients

Appendix S4. Comparison of GLMM with the two-step approach

Appendix S5. Zip file with R programs

Figure S1. Q-Q plot of random effect in the final model of the Dune Meadow data

Figure S2. A Q-Q plot of the random effects in the final GLMM model of the Grazed Grassland data.

Figure S3. RLQ biplot of the Grazed Grassland data explaining 99% of the variance in the fourth corner statistics.

Table S1. Comparison of models with (M1) and without (M0) SLA-Manure interaction by anova(M0,M1) in the Dune Meadow data.

Table S2. The final GLMM model (after model selection) for the Dune Meadow data.

Table S3. The final GLMM model (after model selection) for the Grazed Grassland data.

Table 1.

Effect of Manure and SLA on the presence-absence of Dune meadow species: parameter estimates of fixed effects from GLMM model M1 (equation (7)); z-value: estimate/standard error; *P<0.05, **P<0.01. The remaining estimates are $\sigma_\alpha = 0.85$, $\sigma_\beta = 0.36$, $\rho = -0.39$ and $\sigma_\gamma = 0.21$.

Name	Symbol	Parameter estimate	Standard error	z-value
(Intercept)	a_0	-1.75*	0.81	-2.17*
Manure	b_0	-1.21**	0.40	-3.00**
SLA	a_1	0.03	0.03	0.97
Manure:SLA	b_1	0.06**	0.02	3.28**

z-value: estimate/standard error; *P<0.05, **P<0.01.

Table 2.

Tiered forward selection of environmental and trait variables in models explaining species occurrence probability in the Dune Meadow data. The variable or interaction (indicated by :) giving the lowest SigAIC is added in each row (indicated by +). The best model in each tier is indicated in bold.

Tier	Effects	SigAIC
	Random effects	
1	(1 | species)+(1 | site)	650.9
1	(..+ Moisture | species)	590.4
1	(.. + Manure | species)	**571.1**
1	(..+ Use | species)	574.8
	Fixed effects	
2	+Moist+Manure	578.8
2	+Manure:SLA	567.5
2	+Moisture:Seedmass	**566.6**
2	+Moisture:SLA	567.9
3	No interactions deleted	**566.6**

Table 3.

Sign and significance of single trait-environment terms as found by GLMM and fourth corner in the Dune Meadow.

GLMM/4th	A1	Moisture	Mag	Use	Manure
SLA	NS/NS	- /NS	**/**	NS/NS	++/++
Height	NS/NS	NS/NS	NS/NS	NS/NS	NS/NS
hLDMC	NS/NS	NS/NS	NS/NS	NS/--	NS/NS
Seedmass	NS/NS	--/--	NS/NS	NS/NS	NS/NS
Lifespan	NS/NS	NS/NS	NS/NS	NS/NS	NS/NS

./. = sign of GLMM/ sign of fourth corner (4th) with single sign: P< 0.05, double sign: P< 0.01, NS: P > 0.05. For a categorical variable the sign is replace by *.

Table 4.

Sign and significance of single trait-environment terms as found by GLMM and fourth corner in the Dune Meadow.

GLMM/4th	Grazing	Soil P	Water
Polycarpic	NS/NS	++/++	++/NS
C:N ratio	--/--	NS/NS	- /NS
Seed mass	NS/NS	NS/-	++/++
SLA	NS/NS	+ /++	NS/NS
Height	NS/NS	- /-	NS/NS
Onset.flower	NS/NS	--/--	NS/NS

./. = sign of GLMM/ sign of fourth corner (4^{th}) with single sign: $P < 0.05$, double sign: $P < 0.01$, NS: $P > 0.05$.

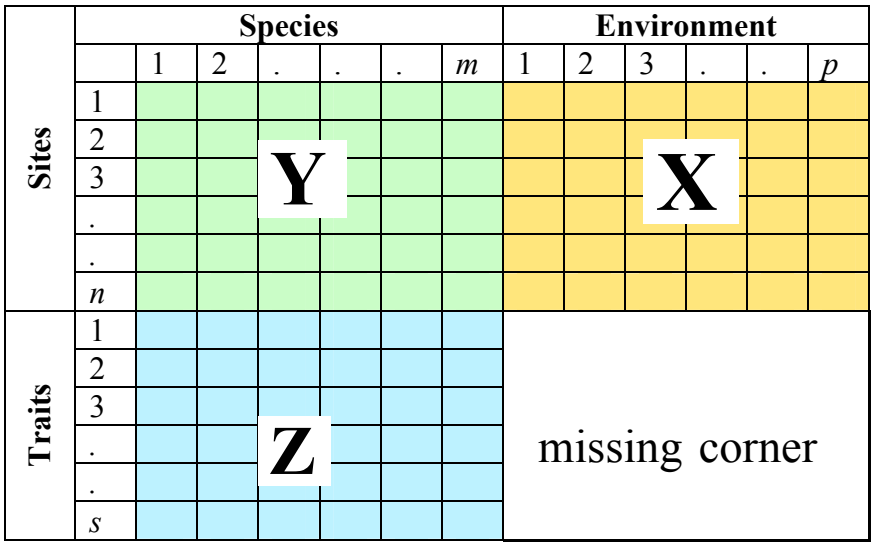

Fig. 1. A table **Y** ($n \times m$) containing the abundances of m species at n sites, a second table **X** ($n \times p$) with measurements of p environmental variables for the n sites, and a third table **Z** ($m \times s$) describing s traits for the m species. In the fourth corner method the $s \times p$ missing corner is filled with traits × environment correlations.

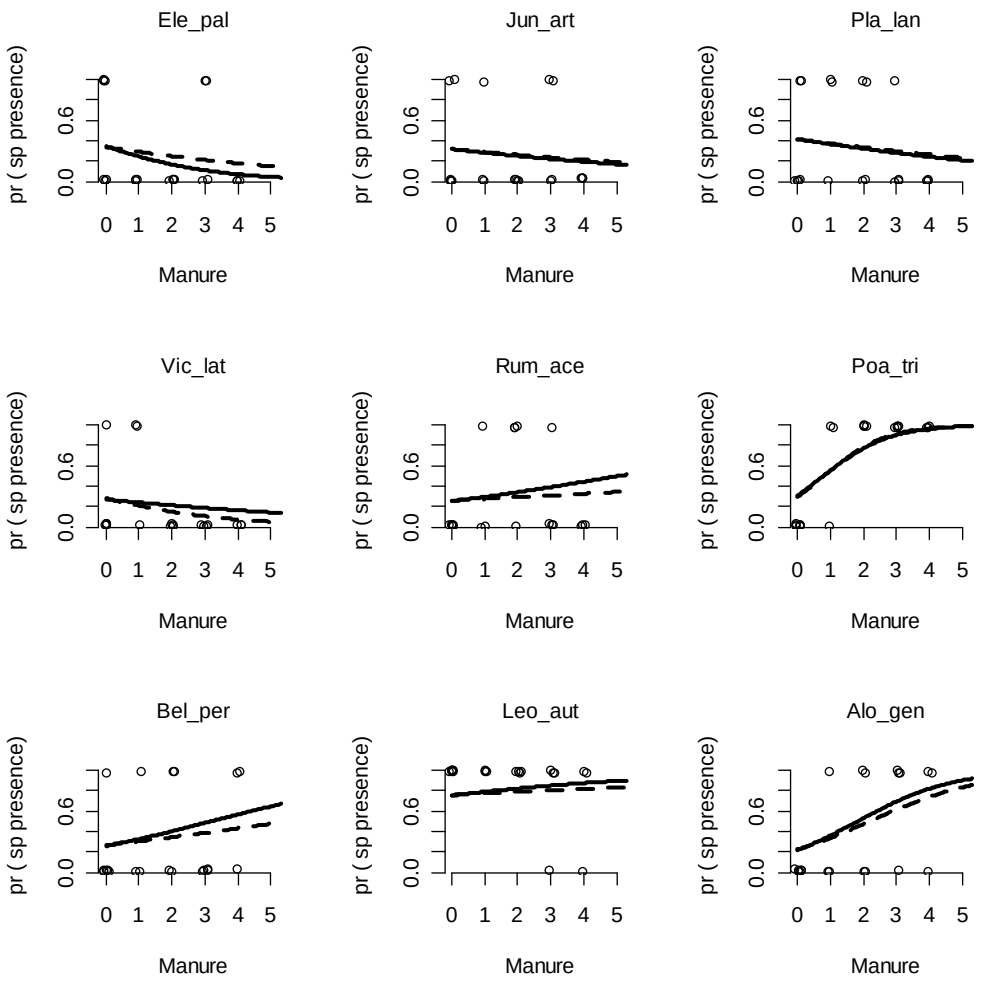

Fig. 2. Occurrence probability against Manure as fitted by GLMM for nine selected species. Both intercept and slope vary among species and either do (red-solid line) or do not (blue-dashed line) depend on the trait SLA. ◦**:** jittered presence (1) and absence (0).

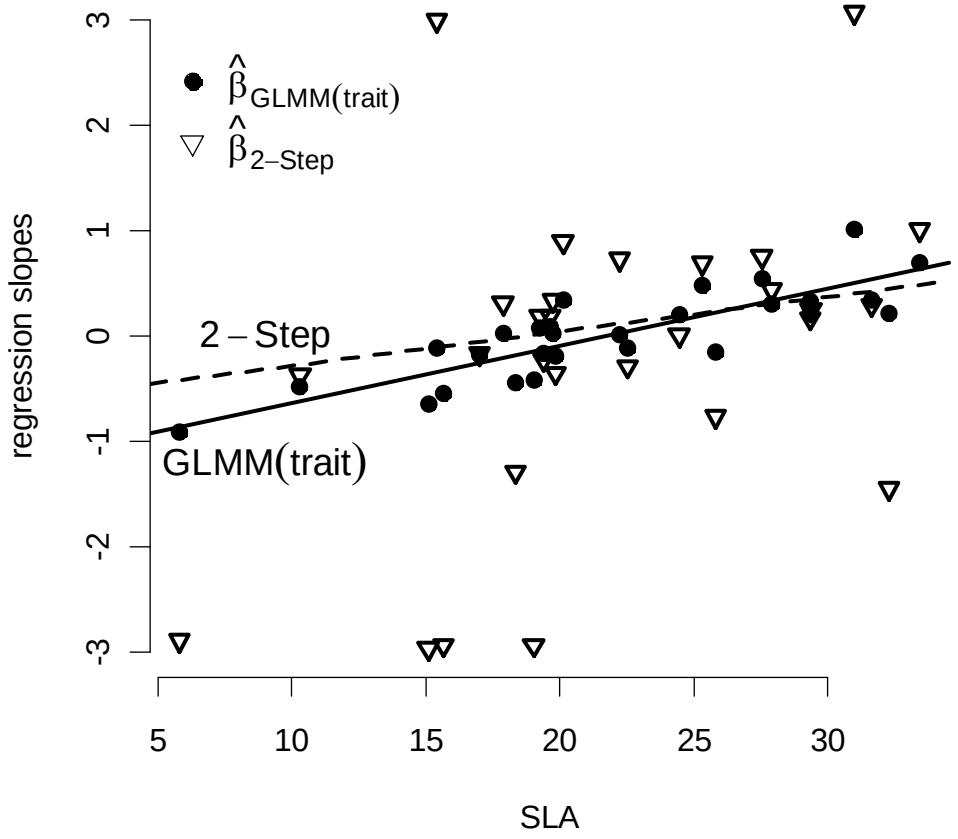

Fig. 3. Regression slopes (β_j) of species with respect to Manure plotted against trait SLA with fitted regression line for GLMM (circles with solid line) and the 2-step approach (triangles with dashed line). Slopes are truncated at 3 and -3 when estimated above 3 or below -3.

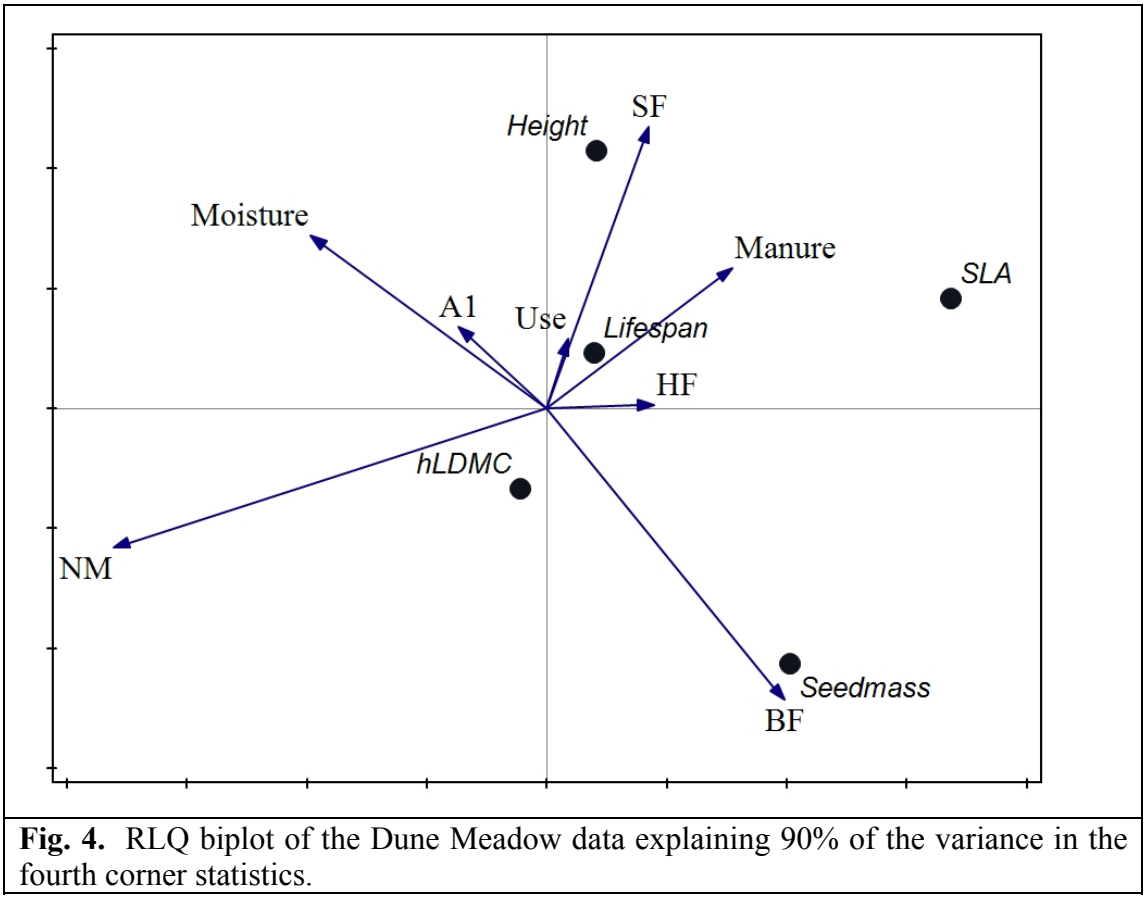

Fig. 4. RLQ biplot of the Dune Meadow data explaining 90% of the variance in the fourth corner statistics.

Supplementary figures and tables and Appendices

Fig. S1. A Q-Q plot of the random effects in the final GLMM model of the Dune Meadow data .

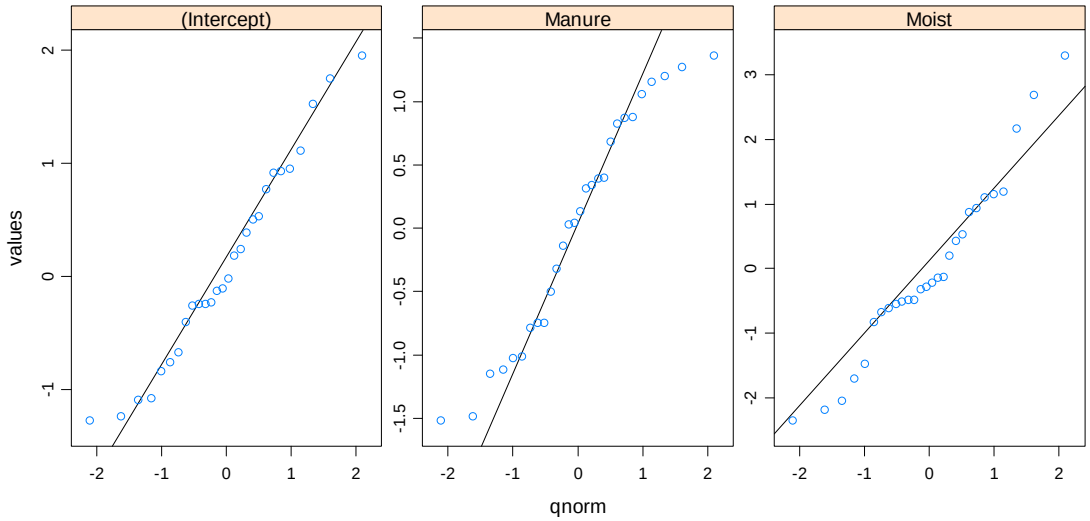

Fig. S2. A Q-Q plot of the random effects in the final GLMM model of the Grazed Grassland data.

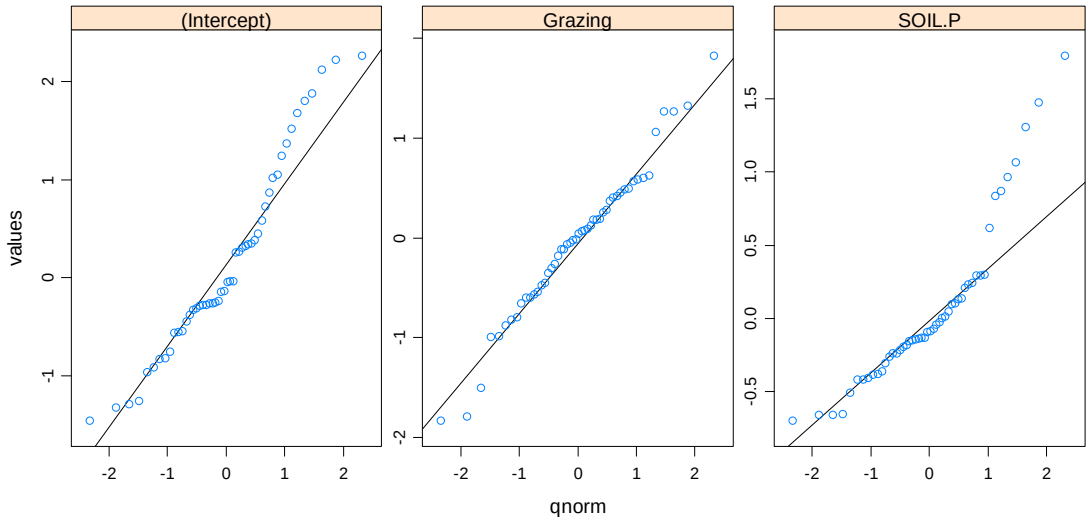

Fig. S3. RLQ biplot of the Grazed Grassland data explaining 99% of the variance in the fourth corner statistics.

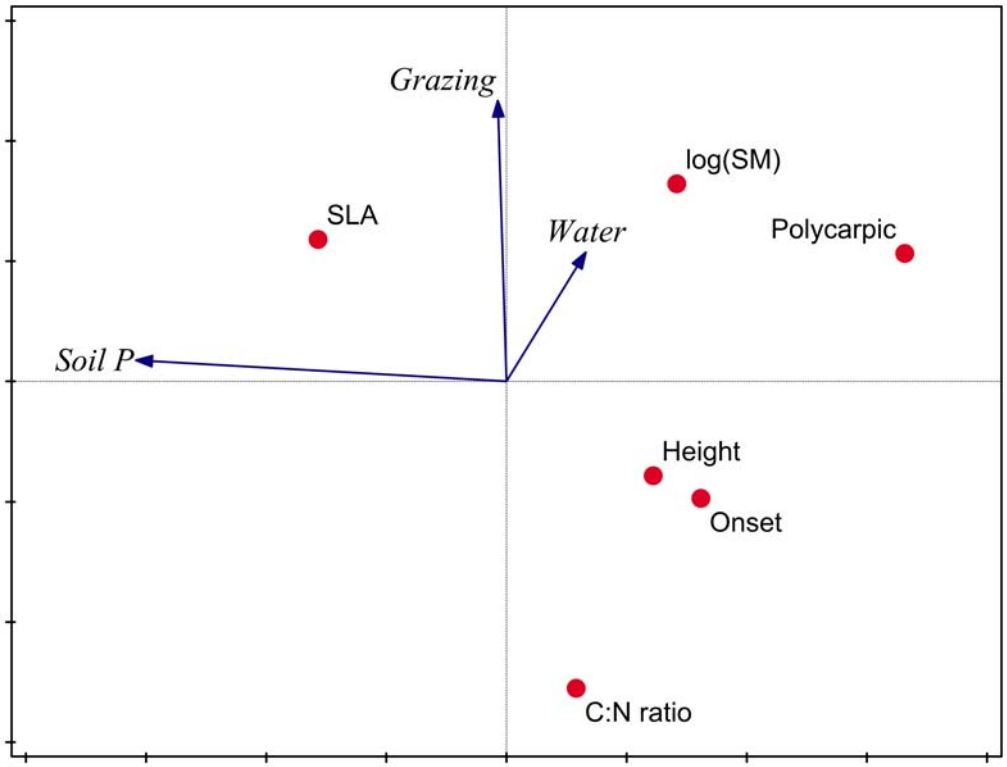

Table S1. Comparison of models with (M1) and without (M0) SLA-Manure interaction by anova(M0,M1) in the Dune Meadow data. df = degrees, logLik = loglikelihood, Chi-sq = 2*difference in logLik, Chi df = difference in df, Pr(>Chisq) = P-value.

Model	df	logLik	Chi-sq	Chi-df	Pr(>Chisq)
M0	7	-305.95			
M1	8	-301.12	9.67	1	0.0018**

Table S2. The final GLMM model (after model selection) for the Dune Meadow data.

```
Generalized linear mixed model fit by the Laplace approximation
Formula: y ~ (1 | site) + (1 + Moist + Manure | sp) + Moist + Manure
+      SLA + Seedmass + Manure:SLA + Moist:Seedmass
   Data: Data
   AIC   BIC logLik deviance
 540.9 601.4 -256.4    512.9
Random effects:
 Groups Name        Variance Std.Dev. Corr
 sp     (Intercept) 1.43789  1.19912
        Moist       2.82409  1.68050  -0.231
        Manure      1.33610  1.15590   0.299  0.511
 site   (Intercept) 0.32543  0.57046
Number of obs: 560, groups: sp, 28; site, 20

Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.618291   0.313967  -5.154 2.55e-07 ***
Moist          -0.031705   0.380945  -0.083   0.9337
Manure          0.009563   0.299437   0.032   0.9745
SLA             1.276773   0.298861   4.272 1.94e-05 ***
Seedmass       -0.274162   0.270879  -1.012   0.3115
Manure:SLA      0.904098   0.259725   3.481   0.0005 ***
Moist:Seedmass -0.824813   0.327925  -2.515   0.0119 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table S3. The final GLMM model (after model selection) for the Grazed Grassland data.

```
Generalized linear mixed model fit by the Laplace approximation
Formula: y ~ (1 | site) + (1 + SOIL.P + Grazing | sp) + SOIL.P +
Grazing +      Polycarpic + CNratio + SOIL.P:Polycarpic +
Grazing:CNratio
   Data: Data
  AIC  BIC logLik deviance
 1684 1763 -827.8     1656
Random effects:
 Groups Name        Variance Std.Dev. Corr
 sp     (Intercept) 1.38065  1.17501
        SOIL.P      0.58959  0.76785   0.351
        Grazing     0.91612  0.95714  -0.084 -0.162
 site   (Intercept) 0.46159  0.67940
Number of obs: 2150, groups: sp, 50; site, 43

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.4289     0.2187 -11.107  < 2e-16 ***
SOIL.P             -0.4298     0.1806  -2.380  0.01731 *
Grazing             0.4148     0.1940   2.139  0.03244 *
Polycarpic         -0.0267     0.1911  -0.140  0.88887
CNratio            -0.3604     0.1847  -1.951  0.05105 .
SOIL.P:Polycarpic  -0.8788     0.1376  -6.388 1.68e-10 ***
Grazing:CNratio    -0.6213     0.1714  -3.624  0.00029 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix S1. The GLMM trait model for count and multinomial data

When abundance is a count, Poisson log-linear regression analysis is a commonly used starting point. Poisson log-linear regression is part of the generalized linear model family. If the data y_{ij} are assumed to follow a Poisson distribution with mean μ_{ij}

$$y_{ij} \sim \text{Poisson}(\mu_{ij})$$

and the link function is logarithmic function, the analogue of the first part of equation 6 in the main text is

$$E(y_{ij}) = \mu_{ij} = \exp(\alpha_j + \beta_j x_i + \gamma_i)$$

which is usually written as

$$\log(\mu_{ij}) = \alpha_j + \beta_j x_i + \gamma_i$$

The other aspects of the model specification remain the same. In lme4 the GLMM trait model for counts can be fitted by simply replacing "binomial" by "poisson" and "logit" by "log":

```
M1 <- glmer(y ~ z + x + z:x +(1+x|species)+(1|sites),
     family=poisson(link="log"), data)
```

Nothing else changes.

Count data may have a larger variance than assumed by the Poisson distribution. This is called overdispersion and can be detected in the data by introducing using a data-level variance component in the GLMM (Gelman & Hill 2007). The GLMM for overdispersed count data is

$$\log(\mu_{ij}) = \alpha_j + \beta_j x_i + \gamma_i + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim \text{N}(0, \sigma_\varepsilon^2)$$

The variance component σ_ε^2 measures the amount of overdispersion and can be tested for significance by a LR test. In lme4 we specify

```
data$rows = 1:nrow(data)
M2 <- glmer(y ~ z + x + z:x +(1+x|species)+(1|sites) +
     (1|rows), family=binomial(link="poisson"), data)
```

and can test the significance of the overdispersion by

```
anova(M1,M2).
```

Multinomial data is data that is count data with a constraint sum so that only the fraction is informative. Abundance data may be modeled as multinomial data as the interest is in the relative abundance only or if the data has been sampled as such, for

example, if at each site a pre-specified number of individuals is collected. Multinomial data can be modeled as count data by adding a fixed effect for the factor sites (McCullagh & Nelder 1989)

```
M1 <- glmer(y ~ z + x + z:x + sites + (1+x|species),
     family=poisson(link="log"), data)
```

Unfortunately this specification failed to run in lme4 at the time of writing.

References

Gelman A. & Hill J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

McCullagh P. & Nelder J.A. (1989). *Generalized linear models (second edition)*. Chapman and Hall, London.

Appendix S2. Dune meadow data in matrix form (**Y**, **X** and **Z**) and in vector notation for lme4.

For the analysis in the main text the abundance data are converted to presence-absence data (0/1). For full names see Jongman et al. (1995).
Abundance data (matrix $\mathbf{Y}^T$):

											1	1	1	1	1	1	1	1	1	2
	Species	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0
1	Ach_mil	1	3	0	0	2	2	2	0	0	4	0	0	0	0	0	0	2	0	0	0
2	Agr_sto	0	0	4	8	0	0	0	4	3	0	0	4	5	4	4	7	0	0	0	5
3	Air_pra	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	0
4	Alo_gen	0	2	7	2	0	0	0	5	3	0	0	8	5	0	0	4	0	0	0	0
5	Ant_odo	0	0	0	0	4	3	2	0	0	4	0	0	0	0	0	0	4	0	4	0
6	Bel_per	0	3	2	2	2	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0
7	Bro_hor	0	4	0	3	2	0	2	0	0	4	0	0	0	0	0	0	0	0	0	0
8	Che_alb	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9	Cir_arv	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Ele_pal	0	0	0	0	0	0	0	4	0	0	0	0	0	4	5	8	0	0	0	4
11	Ely_rep	4	4	4	4	4	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0
12	Emp_nig	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
13	Hyp_rad	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	5	0
14	Jun_art	0	0	0	0	0	0	0	4	4	0	0	0	0	0	3	3	0	0	0	4
15	Jun_buf	0	0	0	0	0	0	2	0	4	0	0	4	3	0	0	0	0	0	0	0
16	Leo_aut	0	5	2	2	3	3	3	3	2	3	5	2	2	2	2	0	2	5	6	2
17	Lol_per	7	5	6	5	2	6	6	4	2	6	7	0	0	0	0	0	0	2	0	0
18	Pla_lan	0	0	0	0	5	5	5	0	0	3	3	0	0	0	0	0	2	3	0	0
19	Poa_pra	4	4	5	4	2	3	4	4	4	4	4	0	2	0	0	0	1	3	0	0
20	Poa_tri	2	7	6	5	6	4	5	4	5	4	0	4	9	0	0	2	0	0	0	0
21	Pot_pal	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0
22	Ran_fla	0	0	0	0	0	0	0	2	0	0	0	0	2	2	2	2	0	0	0	4
23	Rum_ace	0	0	0	0	5	6	3	0	2	0	0	2	0	0	0	0	0	0	0	0
24	Sag_pro	0	0	0	5	0	0	0	2	2	0	2	4	2	0	0	0	0	0	3	0
25	Sal_rep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	5
26	Tri_pra	0	0	0	0	2	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0
27	Tri_rep	0	5	2	1	2	5	2	2	3	6	3	3	2	6	1	0	0	2	2	0
28	Vic_lat	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	1	0	0

Environment data (matrix **X**):

sites	A1_hor	Moisture	Management	Use	Manure
1	2.8	1	SF	2	4
2	3.5	1	BF	2	2
3	4.3	2	SF	2	4
4	4.2	2	SF	2	4
5	6.3	1	HF	1	2
6	4.3	1	HF	2	2
7	2.8	1	HF	3	3
8	4.2	5	HF	3	3
9	3.7	4	HF	1	1
10	3.3	2	BF	1	1
11	3.5	1	BF	3	1
12	5.8	4	SF	2	2
13	6	5	SF	2	3
14	9.3	5	NM	3	0
15	11.5	5	NM	2	0
16	5.7	5	SF	3	3
17	4	2	NM	1	0
18	4.6	1	NM	1	0
19	3.7	5	NM	1	0
20	3.5	5	NM	1	0

Trait data (matrix **Z**):

	Species	SLA	Height	LDMC	Seedmass	Lifespan
1	Ach_mil	19.63	21.15	172.20	0.13	perennial
2	Agr_sto	29.35	18.40	273.55	0.03	perennial
3	Air_pra	15.66	7.00	270.29	0.16	annual
4	Alo_gen	33.40	20.00	211.95	0.37	perennial
5	Ant_odo	22.53	14.80	400.74	0.23	perennial
6	Bel_per	31.62	2.40	177.10	0.10	perennial
7	Bro_hor	27.90	38.40	260.68	1.72	perennial
8	Che_alb	22.21	48.00	164.33	0.65	annual
9	Cir_arv	15.40	87.50	141.66	1.25	perennial
10	Ele_pal	10.31	52.50	217.76	1.01	perennial
11	Ely_rep	20.12	53.80	446.49	1.88	perennial
12	Emp_nig	5.80	16.20	443.00	0.88	perennial
13	Hyp_rad	18.35	5.00	163.96	0.47	perennial
14	Jun_art	19.38	32.50	202.95	0.02	perennial
15	Jun_buf	17.87	18.00	136.50	0.03	annual
16	Leo_aut	32.27	6.70	194.33	0.61	perennial
17	Lol_per	25.31	28.80	263.78	2.01	perennial
18	Pla_lan	19.84	7.10	199.88	1.51	perennial
19	Poa_pra	27.52	32.40	281.68	0.24	perennial
20	Poa_tri	30.98	38.33	252.36	0.16	perennial
21	Pot_pal	19.02	35.00	264.86	0.84	perennial
22	Ran_fla	16.99	21.50	173.91	0.45	perennial
23	Rum_ace	29.34	47.50	102.11	0.93	perennial
24	Sag_pro	19.25	6.00	217.00	0.07	perennial

25	Sal_rep	15.09	60.00	388.06	0.04	perennial
26	Tri_pra	19.73	24.20	277.13	1.87	perennial
27	Tri_rep	24.44	12.40	217.14	0.47	perennial
28	Vic_lat	25.80	13.00	217.00	2.05	annual

Dune meadow data in vector notation for lme4

```
Dune=read.table("Dune.txt", header=TRUE)
head(Dune)
```

	site	species	sp	abun	y	A1	Moist	Mag	Use	Manure	SLA	Height	LDMC	Seedmass	Lifespan
1	1	Ach_mil	1	1	1	2.8	1	SF	2	4	19.63	21.15	172.2	0.13	perennial
2	2	Ach_mil	1	3	1	3.5	1	BF	2	2	19.63	21.15	172.2	0.13	perennial
3	3	Ach_mil	1	0	0	4.3	2	SF	2	4	19.63	21.15	172.2	0.13	perennial
4	4	Ach_mil	1	0	0	4.2	2	SF	2	4	19.63	21.15	172.2	0.13	perennial
5	5	Ach_mil	1	2	1	6.3	1	HF	1	2	19.63	21.15	172.2	0.13	perennial
6	6	Ach_mil	1	2	1	4.3	1	HF	2	2	19.63	21.15	172.2	0.13	perennial
7	. .....														

R-code

Code from dune_data_expand.r

```
rm(list=ls(all=TRUE))
# Transform three data tables Y (abundance), X (environment), Z (traits)
# to vector notation for GLMM
# version   May 2011
# Abundance data (20 sites x 28 species)
# Important: Look at package multitable for this
# multitable did not include the site and species/sp columns at the time of writing)
# but likely to be updated.
file_Y <- "data/dune_abundance_Y.txt"  # sites x species abundance data
file_X <- "data/dune_environment_X.txt"# sites x environment data
file_Z <- "data/dune_traits_Z.txt"     # species x traits data
Y<-read.table(file_Y, header = TRUE)
sites <- Y[,1]; Y <-Y[,-1]; species <-names(Y); rownames(Y) = sites;
#dim(Y);head(Y);names(Y);rownames(Y)
X<-read.table(file_X, header = TRUE)
sitesX <- X[,1]; rownames(X) = sitesX; X <-X[,-1]
if (!all.equal(sitesX,sites)) print("BEWARE: site names unequal?")
#dim(X);head(X);names(X);rownames(X)
Z<-read.table(file_Z, header = TRUE)
speciesZ <- Z[,1]; traits <-names(Z); rownames(Z) = speciesZ; Z <-Z[,-1]
#dim(Z);head(Z);names(Z);rownames(Z)
a <-as.character(speciesZ); b <-as.character(species)
if (!all.equal(a,b)) print("BEWARE: species names unequal?")

sitespec <- expand.grid(rownames(Y),colnames(Y))
site <-sitespec[,1]; species<-sitespec[,2]; sp <-as.numeric(species)
abun <- as.vector(as.matrix(Y))
y <- 1*(abun>0)  # transformation to presence - absence
Xvec <-  X[site,]
Zvec <-  Z[species,]
XYZ <- cbind(data.frame(site,species,sp,abun, y),Xvec,Zvec)
Dune <- XYZ
write.table(Dune,file = "Dune.txt")
```

Appendix S3. The GLMM trait model with qualitative and/or qualitative trait and environmental variable

In the main text, the species trait and the environmental variable were both quantitative. The GLMM can also be used to model when species trait and environmental variable are both qualitative or when one is quantitative and the other qualitative. Here we illustrate all the combinations with key output and interpretation of the regression coefficients using the Dune Meadow data in vector notation in Appendix S2 (dataframe Dune). The R code at the end of this appendix shows for all combinations how to compute the fitted occurrence probability with confidence bands from the estimated regression coefficients and their covariance matrix.

1-Both trait and environmental variable quantitative

As in the main text we consider here the case where both species trait and the environmental variable are quantitative. Now we fit a model, using glmer in the lme4 package, where sp codes for species and site for sites,

```
glmer(y~Manure+SLA+Manure:SLA+(1+Manure|sp)+(1|site),
family=binomial, Dune)
```

or to the same effect

```
glmer(y~Manure*SLA+(1+Manure|sp)+(1|site),
family=binomial, Dune)
```

The fixed effects estimates are in Table 1.

Table 1. Fixed effects estimated from GLMM for quantitative environmental variable and quantitative species trait.

Fixed effect	Parameter estimate	Standard error	z statistic
(Intercept)	-1.751	0.806	-2.172*
Manure	-1.210	0.403	-3.005**
SLA	0.033	0.035	0.967
Manure:SLA	0.055	0.017	3.275**

The regression equation is

$$ls = -1.751 - 1.210 \times Manure + 0.033 \times SLA + 0.055 \times Manure \times SLA .$$

The result is on logit scale and can be converted to occurrence probability (prob) by

$$prob = invlogit(ls) = 1/(1+\exp(-ls)).$$

Fig.1 shows the occurrence probability against Manure for species with different SLA (SLA = 5, 15 and 30) along with 95% confidence bands.

We now return to Table 1. In a model with an interaction, the size and sign of main effects depend on the scales of the variables and may thus be difficult to interpret. Manure runs from 0 to 4 in the data and SLA ranges from 5.8 to 33.4 (low to high). In Table 1, the main effect for Manure (-1.21) is negative and

significant showing that, if a species would have SLA = 0, it would decrease in occurrence probability with higher Manure. Such species do not occur in the data; the lowest SLA is 5.8. Species with SLA=5 still decreases (Fig 1); their slope with respect to Manure is -1.21+5*0.055= -0.935. Species with a high SLA value, for example SLA = 30, have a slope of -1.21+30*0.055 = 0.44, indicating that such species are increasing in occurrence probability with higher Manure (Fig 1). The mean SLA is ~22, giving close to 0 slope, indicating that the occurrence probability does not depend on Manure for species with mean SLA value.

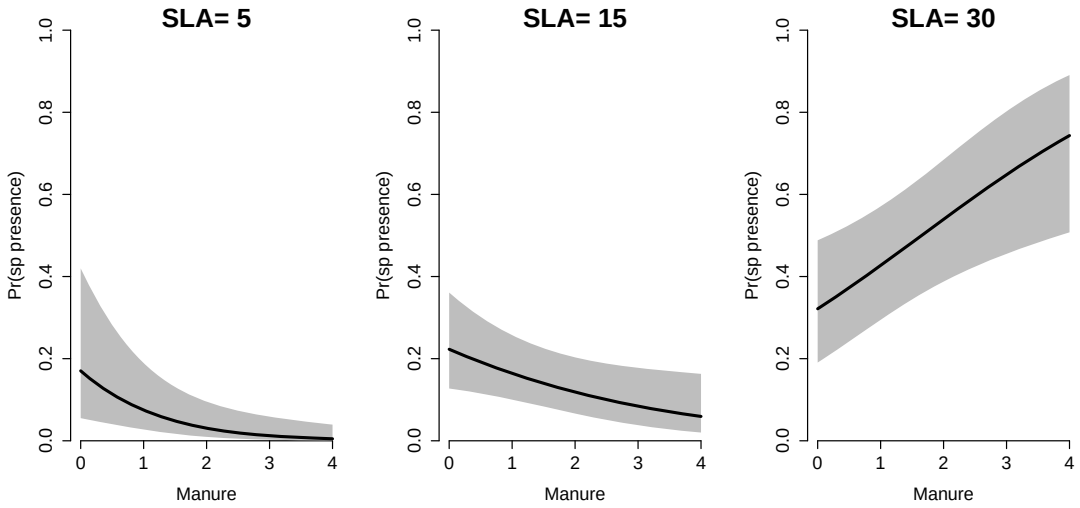

Fig. 1. Occurrence probability of species having low (SLA=5), intermediate (SLA=15) and high (SLA=30) values in relation to manure in the dune meadow from a GLMM model along with 95% confidence bands.

We now turn to the effect of trait SLA. In model M1, the main effect for SLA (0.033) is positive and nonsignificant, showing that the occurrence probability of the species slightly increases with increasing SLA in sites with Manure = 0. As the interaction coefficient is positive, the strength of the positive relation increases for higher values of Manure. Therefore, for all values of Manure, species with high SLA have higher occurrence probability than species with low SLA; this tendency is stronger the higher the value of Manure.

2- Quantitative trait and qualitative environmental variable

We consider the case where the environmental variable is qualitative and species trait is quantitative. This yields separate regression lines for each category of the environmental variable (Fig. 1). In our example Manure is a qualitative explanatory variable (i.e., a factor), we divide the range of manure into two intervals and convert them to a factor with two categories: Manure-no and Manure-yes

```
ManureLab = c("no","yes")

Dune$Manure=cut(Dune0$Manure,
```

```
breaks = c(-1,0.5,10),labels=ManureLab)
```
Now we fit a model, using glmer in the lme4 package, with the same type of statement as before

```
glmer(y~Manure*SLA+(1+Manure|sp)+(1|site),
family=binomial,Dune)
```
The fixed effects estimates are in Table 2.

Table 2. Fixed effects estimated from GLMM for qualitative environmental variable and quantitative species trait in the default parameterization.

Fixed effect	Parameter estimate	Standard error	z statistic
(Intercept)	-1.083	0.961	-1.126
Manure-yes	-3.884	1.179	-3.294***
SLA	-0.012	0.042	-0.286
Manure-yes:SLA	0.197	0.051	3.870***

The regression equation for manure-no (the first level of the factor manure) is on logit scale is straightforward

Manure-no $\rightarrow$ -1.083 – 0.012$\times$SLA

The regression equation for Manure-yes can be obtained as follows. The intercept for Manure-yes can be found by adding the coefficients for intercept and Manure-yes and the slope for Manure-yes with respect to SLA by adding the coefficients for SLA and Manure-yes:SLA . The regression equation for Manure-yes becomes on logit scale

Manure-yes $\rightarrow$(-1.083 -3.844) + (-0.012+0.197)$\times$SLA= -4.967 + 0.185$\times$SLA

Both equations can be converted to occurrence probability curves with confidence bands (Fig. 2).

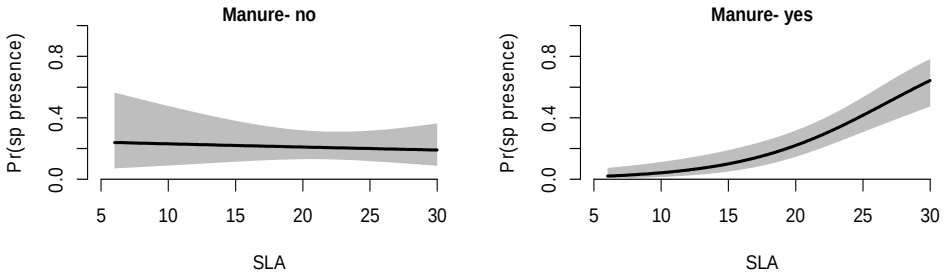

Fig. 2. Occurrence probability of a species in manure-no and manure-yes meadows, with 95% confidence band, in relation to the species trait SLA from a GLMM model where the environmental variable is a factor with two categories.

Fig. 2 shows that in manure-no meadows the probability of occurrence of species decreases very slightly with increasing SLA, whereas in manure-yes meadows the probability of occurrence of species increases with increasing SLA. The two regression lines show the interaction between Manure and SLA. In Table 2 the

interaction is represented by one regression coefficient (0.197) which is highly significant.

A trick to immediately obtain the regression model for each meadow category is to make a slight modification in the model specification:

```
glmer(y~0+Manure+Manure:SLA+(0+Manure|sp)+(1|site),family
=binomial, Dune)
```

The results for this model specification are displayed in Table 3.

Table 3. Fixed effects estimated from GLMM for qualitative environmental variable and quantitative species trait in natural parameterization.

Fixed effect	Parameter estimate	Standard error	z statistic
Manure-no	-1.083	0.961	1.126
Manure-yes	-4.967	0.896	-5.546***
Manure-no:SLA	-0.012	0.042	-0.286
Manure-yes:SLA	0.185	0.038	4.900***

Table 3 contains directly the coefficients of the regression equations for Manure-no and Manure-yes equation on logit scale:

Manure-no $\rightarrow$ -1.083 – 0.012 $\times$SLA
Manure-yes $\rightarrow$ -4.967 + 0.185 $\times$SLA

In Table 3 there seem to be two interaction terms, but the real interaction is the difference between the two. Table 2 and Table 3 use different parameterizations of the same model. In either case, a likelihood ratio (LR) test of the interaction is obtained by comparison with the model

```
glmer(y~Manure+SLA+(1+Manure|sp)+(1|site),
family=binomial, Dune)
```

using the `anova()` statement.

3- Qualitative trait and quantitative environmental variable

We consider the case where the trait is qualitative and the environmental variable is quantitative. This yields separate regression lines for each trait category (Fig. 3). In our example SLA turned into a qualitative explanatory variable (i.e., a factor), with three categories: low, middle and high.

```
SLALab = c("low","middle","high")
Dune$SLA=cut(Dune$SLA, breaks=c(0,13,25,40),labels=
SLALab) #3levels
```

Now the model specification using glmer in the lme4 package is

```
glmer(y~Manure*SLA+(1+Manure|sp)+(1|site),
family=binomial, Dune)
```

The fixed effects for the above model are given in Table 4.

Table 4. Fixed effects estimated from GLMM for qualitative trait and quantitative environmental variable.

Fixed effect	Parameter estimate	Standard error	z statistic
(Intercept)	-1.164	0.877	-1.326
Manure	-0.635	0.489	-1.297
SLA-middle	-0.054	0.925	0.058
SLA-high	0.324	0.950	0.341
	0.482	0.508	0.948
Manure:SLA-middle	1.055	0.517	2.042*
Manure:SLA-high			

The regression equation for SLA-low (the first level of the factor SLA) is on logit scale is straightforward

SLA-low $\rightarrow$ -1.164– 0.635 $\times$Manure

The regression equation for SLA-middle can be obtained as follows. The intercept for SLA-middle can be found by adding the coefficients for intercept and SLA-middle and the slope for SLA-middle with respect to manure by adding the coefficients for Manure and Manure:SLA-middle . The regression equation for SLA-middle becomes on logit scale

SLA-middle $\rightarrow$(-1.164 –0.054) + (-0.635 +0.482)$\times$Manure= -1.218 - 0.153$\times$Manure

Similarly for SAL-high is

SLA-high $\rightarrow$(-1.164 –0.324)+ (-0.635 +1.055)$\times$Manure=-1.488 + 0.42 $\times$Manure

All the three equations can be converted to occurrence probability curves (Fig. 3).

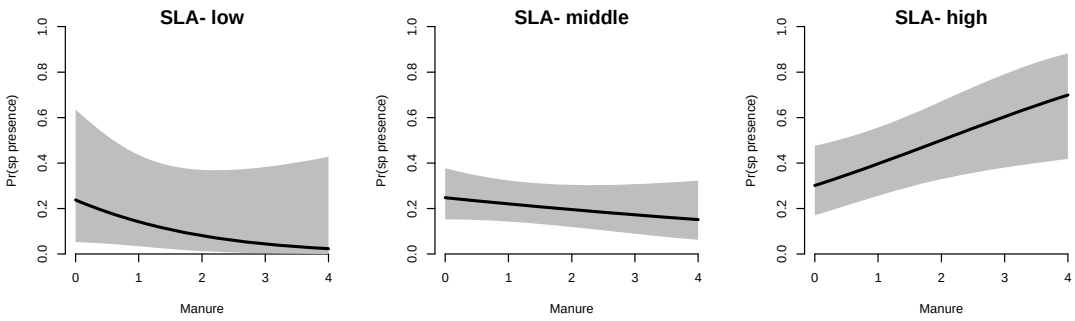

Fig. 3. Occurrence probability of species for SLA-low, SLA-middle and SLA-high, with 95% confidence band, in relation to manure in the dune meadow from a GLMM model where the trait is a factor with three categories.

The probability of occurrence of species with SLA-low and SLA-middle decreases with increasing manure, whereas the probability of occurrence of species for SLA-high increases with increasing manure (Fig. 3).

4- Both trait and environmental variable qualitative

We consider the case when both trait and environmental variable are qualitative. This yields occurrence probabilities in each class of the cross-classification of trait and environment. In our example, the species trait SLA is classified into three and environmental variable manure classified into two categories as above. The model specification in R is:

```
glmer(y ~Manure*SLA +(1+Manure|sp)+(1|site),
family=binomial, Dune)
```

Table 5. Fixed effects estimated from GLMM for qualitative environmental variable and qualitative species trait.

Fixed effect	Parameter estimate	Standard error	z statistic
(Intercept)	-0.820	0.975	-0.841
Manure-yes	-2.206	1.371	-1.609
SLA-middle	-0.492	1.040	-0.473
SLA-high	-0.674	1.087	-0.620
	2.115	1.439	1.470
Manure-yes:SLA-middle	4.082	1.480	2.758**
Manure-yes:SLA-high			

From the coefficients in Table 5 we need to construct the occurrence probabilities in each class. The reference class is the first level of manure (no) and the first level of SLA (low), Manure-no–SLA-low; we have on the logit-scale

for class Manure-no– SLA-low →-0.820
for class Manure-yes–SLA-low →-0.820-2.206 = -3.026
for class Manure-no ‑SLA-middle →-0.820-0.492=-1.312
for class Manure-yes–SLA-middle →-0.820-2.206-0.492+2.115=-1.403
for class Manure-no ‑SLA-high →-0.820-0.674=-1.494
for class Manure-yes–SLA-high →-0.820-2.206-0.674+4.082=0.383

The occurrence probabilities are the inverse logit of these values, for example

invlogit(0.383) = 1/(1+exp(-0.383))= 0.595.

Table 6 shows all probabilities and 95% confidence limits.

Table 6. Probability of occurrence and in parentheses are the corresponding confidence limits of species for two levels of manure and three levels of SLA in meadows.

		SLA		
		low	middle	high
Manure	no	**0.306** (0.061, 0.748)	**0.212** (0.117, 0.354)	**0.183** (0.080, 0.366)
	yes	**0.046** (0.005, 0.329)	**0.197** (0.112, 0.324)	**0.595** (0.395, 0.767)

5- Mix of quantitative and qualitative traits and environmental variables

Finally we consider the case with two environmental variables of which one is quantitative and the other is qualitative and also two traits, one quantitative and the other qualitative. We use Moisture and Seedmass as the quantitative variables and the Manure and SLA in their qualitative versions, as above, with two and three categories, respectively. The model specification in R is:

```
fm5<-glmer(y~Moist*Seedmass+Manure*SLA +
(1+Moist+Manure|sp)+(1|site), family=binomial, Dune)
```

Table 7. Fixed effects estimated by GLMM for a mix of quantitative and qualitative trait and environmental variables.

Fixed effect	Parameter estimate	Standard error	z statistic
(Intercept)	-1.8099	1.6331	-1.108
Moist	0.5113	0.2665	1.918
Seedmass	1.3970	0.9233	1.513
Manure-yes	-3.1052	1.9218	-1.616
SLA-middle	-1.3161	1.3542	-0.972
SLA-high	-1.4464	1.4121	-1.024
Moist:Seedmass	-0.7154	0.2710	-2.639**
Manure-yes:SLA-middle	3.1477	2.0030	1.571
Manure-yes:SLA-high	5.9773	2.0739	2.882**

The interpretation of the coefficients in Table 7 goes along the same lines as that in the previous sections. The model allows predictions to be made for each combination of value of Moisture, Manure, SLA and Seedmass. The R-code used for this example is

```
newdat <- expand.grid(Moist = c(1,2,4), Seedmass
=c(0.05,1,2), Manure=ManureLab, SLA = SLALab, y = 0)
newdat <- glPredict(fm5, newdat)
newdat[, -(5:7)]
```

The first three lines of the output are:

```
   Moist Seedmass Manure    SLA           p         plow      phigh
1      1     0.05     no    low 0.220188160 1.515774e-02 0.83819104
2      2     0.05     no    low 0.312366963 2.901418e-02 0.87351195
3      4     0.05     no    low 0.540389415 7.950060e-02 0.94119772
```

And the last three lines of the output are:

```
52     1     2.00    yes   high 0.816111789 3.488304e-01 0.97352264
53     2     2.00    yes   high 0.638929369 2.549278e-01 0.90149447
54     4     2.00    yes   high 0.219550859 4.701617e-02 0.61598359
```

6- R-code

The R code used in this Appendix is also in the R-file glmm-plot-conf-int.r. The output is in file glmm-plot-conf-int.txt.

```
rm(list=ls(all=TRUE))
library(lme4)
invlogit <- function(x){1/(1+exp(-x))}
Dune=read.table("data/Dune.txt", header=TRUE,sep=" ")
Dune0= Dune  # we will modify Dune starting from the original, Dune0
colnames(Dune)
glPredict <- function(fm1, newdat, conf = 95) {
 # Predicts occurrence probability with confidence limits from an glmer object at
 #  the points provided as rows of newdat
 # fm1 = glmer  object
 # newdat = data frame with values for predictors for which prediciton must be made
 # confidence value (in %)
 # for related code see package ez
 # Value:
 #  y, lo, hi = prediction with confidence limits on link scale
 #  p, plow, phigh =   occurrence probability with confidence limits
 frac = 1 - (100-conf)/200
 mm = model.matrix(terms(fm1),newdat)
 y = mm %*% fixef(fm1)  # prediction on link scale
 Var <- Matrix::diag(mm %*% tcrossprod(vcov(fm1),mm)) # variance on link scale
 lo = y-qnorm(frac)*sqrt(Var)
 hi = y+qnorm(frac)*sqrt(Var)
 newdat$y = y
 newdat <- data.frame(newdat, ylo = lo, yhi = hi,
  p = invlogit(y), plow = invlogit(lo), phigh = invlogit(hi))
 newdat
}

#####################################################################
# Table 1 Quantitative environmental variable; Quantitative trait
#####################################################################

fm1 = glmer (y ~ Manure *SLA +(1 + Manure | sp)+(1|site),
        family=binomial,data=Dune)

# for prediction with confidence limits
SLAval = c(5,15,30 )
newdat <- expand.grid( Manure=seq(0,4,length.out=100), SLA = SLAval, y =0)
newdat <- glPredict(fm1, newdat)
names(newdat)
# for plotting
par(bty="n")
par(mfrow=c(1,3))
for ( j in SLAval){

 data.f<- subset( newdat , SLA %in% j)
 x<- data.f$Manure
 plot(0,0,ylim=c(0,1),xlim=range(x),ylab="Pr(sp presence)" ,xlab="Manure" ,
    yaxs="i" , main="",type="n")
 mtext(paste("SLA=",j ),   font= 2, col= "black" )
 polygon(c(x, rev(x)), c(data.f$phigh, rev(data.f$plow)),col = 'gray', border = FALSE)
 points(x, data.f$p, type='l',lwd=2.5)
}
```

```
#####################################################################
# Table 2 Factor environmental variable; Quantitative trait
#####################################################################
ManureLab = c("no","yes")
Dune$Manure= cut(Dune0$Manure, breaks = c(-1,0.5,10), labels=ManureLab)
print(fm2<-glmer(y~ Manure*SLA+(1+Manure|sp)+(1|site)
            , family=binomial, Dune),corr=FALSE)
# for prediction with confidence limits
newdat <- expand.grid(Manure = ManureLab,SLA=seq(6,30,length.out=1000),y = 0)
newdat <- glPredict(fm2, newdat)
par(mfrow=c(1,2))
for ( j in ManureLab){
   data.f<- subset( newdat , Manure %in% j)

   x<- data.f$SLA
   plot(0,0,ylim=c(0,1),xlim=range(x),ylab="Pr(sp presence)" ,xlab="SLA" ,
        yaxs="i" , main="",type="n")
  # title(paste("SLA=",j ), cex.main = 1.2,   font.main= 2, col.main= "black")
   mtext(paste("Manure-",j ),   font= 2, col= "black" )
   polygon(c(x, rev(x)), c(data.f$phigh, rev(data.f$plow)),col = 'gray', border = FALSE)
   points(x, data.f$p, type='l',lwd=2.5)
}
# the alternative parametrization (Table 3)
print(fm2.B<-glmer(y~0+Manure+Manure:SLA+(0+Manure|sp)+(1|site)
            , family=binomial, Dune),corr=FALSE)
newdat.B <- glPredict(fm2.B, newdat)
all.equal(newdat.B,newdat)

fm0<-glmer(y~Manure+ SLA+(1+Manure|sp)+(1|site)
            , family=binomial, Dune)
anova(fm0,fm2)
anova(fm0,fm2.B)
#####################################################################
# Table 4 Quantitative environmental variable; Factor trait
#####################################################################
Dune = Dune0
SLALab = c("low","middle","high")
Dune$SLA= cut(Dune$SLA, breaks = c(0,13,25,40), labels= SLALab)

print(fm3<-glmer(y~Manure*SLA+(1+Manure|sp)+(1|site)
            , family=binomial, Dune),corr=FALSE)

newdat <- expand.grid( Manure=seq(0,4,length.out=100),SLA=SLALab, y =0)
newdat <- glPredict(fm3, newdat)

par(mfrow=c(1,3))
for ( j in SLALab){
   data.f<- subset( newdat , SLA %in% j)
   x<- data.f$Manure
   plot(0,0,ylim=c(0,1),xlim=range(x),ylab="Pr(sp presence)" ,xlab="Manure" ,
        yaxs="i" , main="",type="n")
   mtext(paste("SLA-",j ),   font= 2, col= "black" )
   polygon(c(x, rev(x)), c(data.f$phigh, rev(data.f$plow)),col = 'gray', border = FALSE)
   points(x, data.f$p, type='l',lwd=2.5)
}

 #####################################################################
 # Table 5 Factor environmental variable; Factor trait
```

```
 ###################################################################
ManureLab = c("no","yes")
Dune$Manure= cut(Dune$Manure, breaks = c(-1,0.5,10), labels=ManureLab)
#factor environmental variable; Quantitative trait
print(fm4<-glmer(y~Manure*SLA+(1+Manure|sp)+(1|site)
          , family=binomial, Dune),corr=FALSE)

newdat <- expand.grid( Manure=ManureLab, SLA = SLALab, y = 0  )
newdat <- glPredict(fm4, newdat)
newdat
newdat[, -(3:5)]

#####################################################################
# Table 6 mix of quantitative and qualitative traits and environmental variables
# Two Environmental variables:  quantitative and factor;  two trait: Quantitative and factor
#####################################################################

Dune = Dune0
ManureLab = c("no","yes")
Dune$Manure= cut(Dune$Manure, breaks = c(-1,0.5,10), labels=ManureLab)
SLALab = c("low","middle","high")
Dune$SLA= cut(Dune$SLA, breaks = c(0,13,25,40), labels= SLALab)
print(fm5<-glmer(y~Moist*Seedmass+ Manure*SLA+(1+Moist+Manure|sp)+(1|site)
          , family=binomial, Dune),corr=FALSE)
# for prediction with confidence limits
newdat <- expand.grid(Moist = c(1,2,4), Seedmass =c(0.05, 1,2), Manure=ManureLab, SLA =
SLALab, y = 0  )
newdat <- glPredict(fm5, newdat)
names(newdat)
newdat[, -(5:7)]
#end
```

Appendix S4 Comparison of GLMM with the two-step approach

The dashed regression line of the GLM two-step approach (Fig. 3 of the main text) is fitted by weighted least-squares (weight = inverse to variance estimate in step 1) and shows a weaker relationship than that of GLMM. A small simulation study was done to see whether that was incidental. In the 99% of the 1000 simulated data sets of the power study, the coefficient b_1 estimated by GLMM was greater than that in the two-step approach. It was also much closer to the true coefficient as judged from the root mean squared error (0.031 compared to 0.393). In GLMM, the standard deviation across simulated data sets (0.023) was close to the standard error of estimate reported by GLMM (0.017, rounded to 0.02 in Table 1), showing that this standard error of estimate is valid in this data.