One Toolbox, Many Tools: A Practitioner's Guide to Latent Variable Modelling for Community Ecology

Audun Rugstad,*[1], Bert van der Veen,[1,3], Robert Brian O'Hara,[1],
and Anne Catriona Mehlhoop,[2]

[1]Department of Mathematical Sciences, Norwegian University of Science and Technology, NO
[2]Norwegian Institute for Nature Research, NO
[3]University of Bayreuth, DE

February 5, 2026

In this article, we present the case for Generalized Linear Latent Variable Models (GLLVMs) as a go-to choice of statistical method for any community ecologist wanting to tackle a range of present-day ecological research questions. GLLVMs bring tools and capabilities from classic (mixed-effects) regression models to multivariate community analysis, providing a number of novel ways to tailor models specifically to one's study questions and data properties not available when using non-model-based multivariate methods. In order to facilitate further adoption of these methods by community ecologists, we provide 1) a practitioner-focused and practical overview of the advantages the GLLVM framework brings to the table when addressing different core ecological questions, 2) a number of concrete suggestions for how GLLVMs best can be incorporated into the analytical workflow of community ecologists, and 3) two illustrative worked examples of this workflow in action on real-world data.

Keywords: Ecological modelling, Multispecies data, Community ecology, Community modelling, Ordination, Data exploration, Model selection, Model-based workflow, Latent variable modelling, Invasive species, Ecological restoration

---

*Corresponding author. Contact: audun.rugstad@ntnu.no

Using different types of data is becoming increasingly important to improve our understanding of the nature and dynamics of ecological communities in a range of real-world scenarios. Examples include assessing restoration success (Ribeiro et al., 2023), the impacts of invasive species (Souza-Alonso et al., 2022; Herrmann et al., 2022), and the modelling of community responses to climate change (Sahade et al., 2015). In all of these cases, how well one's ecological research questions can be addressed depends not only on data, but also on the selection of appropriate tools and methods for analysis. And while the statistical toolbox available to ecologists today is large, it is also fragmented, which can make it difficult to chose a set of methods to address the relevant research questions in a study in a way that is both coherent, streamlined and reproducible.

One important example of this is the fact that community ecologists today often find themselves juggling two quite different methodological "schools" when addressing different kinds of ecological questions. On the one hand, questions about univariate data, such as predation rates, breeding success, or the abundance of individual species in different habitats, are typically tackled in a model-based framework, using "standard", statistically well-established regression models within the overarching framework of Generalized Linear Mixed Models (GLMM) (Bolker et al., 2009; Zuur et al., 2009). However, the same type of model-based framework has historically not been available to study differences in patterns of species composition and structure within or between communities. In these cases, where the data are multivariate, i.e. each sample is the abundance of several different species, and where the patterns of correlation between species or sites is the focus, researchers have typically used different forms of *ordination* to analyse the data. That is, distance-based or algorithmic methods such as Non-Metric Multidimensional Scaling (NMDS), Principal Component Analysis (PCA) or Correspondence Analysis (CA) (ter Braak and Prentice, 2004).

Due to their ability to effectively condense and visualize patterns in multivariate species data, traditional ordination methods have historically been very important for studying ecological communities (ter Braak and Šmilauer, 2015). However, the fact that they do not in and of themselves allow for true statistical inference have also led many to argue that their use for answering ecological questions outside of data exploration and hypothesizing is limited (Warton et al., 2012, 2015; Jupke and Schäfer, 2020). Unlike regression models for univariate data, these methods do not, for instance, include estimates of uncertainty, incorporate random effects, or provide reliable tools for checking whether key properties of ecological data, such as the mean-variance relationship, are accounted for (Warton and Hui, 2017). On a more conceptual level, because distance-based and algorithmic methods rely on extensive transformation and "collapsing" of data prior to the analysis, the link between the actual data and the results is more obscure than with model-based methods. Overall, this makes ecological inferences from these methods harder to assess.

The last decade has, however, seen a number of new model-based methods being developed to analyse multivariate community data in a more statistically informative manner (Hui et al., 2015; Niku et al., 2019; Ovaskainen et al., 2017). Most of these fall under the umbrella of the Generalized Linear Latent Variable Modeling (GLLVM) framework. In essence, GLLVMs allow for model-based counterparts to traditional ordination methods, based on Generalized Linear Mixed Models. They allow users to fit models that explain patterns of species co-occurrence by assuming that they are the result of a few underlying *latent*, or unobserved, explanatory variables (i.e., ordination axes in the classical terminology). These latent variables can be inferred from both the species composition data itself (Hui et al., 2015), as well as environmental variables (van der Veen et al., 2023).

The fact that GLLVMs are an extension of the Generalized Linear Mixed Modeling (GLMM) framework to multivariate data means that model-based ordination is situated in the more familiar context of other regression models designed to predict species occurrence and/or abundances. As Figure 1 shows, this also makes it possible to combine model-based ordinations directly with other models – such as multivariate (i.e. ”stacked”) GLMs or environment-trait interaction (”fourth corner”) models (Niku et al., 2021), opening up several new avenues of statistical analysis.

GLLVMs are currently implemented in several software packages. The `gllvm` `R` package (Niku et al., 2025) is aimed at community ecologists, and currently contains by far the richest toolbox for this purpose. The other main feature-rich `R` package is `hmsc` (Tikhonov et al., 2025), which is focused on GLLVMs for Joint Species Distribution Models, and thus has a similarly full toolbox geared at understanding how the environment affects the distributions of individual species. Other noteable software implementations include `ecoCopula` (Popovic et al., 2019), `boral` (Hui, 2025), `VGAM` (Yee, 2025), and `glmmTMB` (McGillycuddy et al., 2025).

Despite the availability of user-friendly software, as well as several examples of GLLVMs being used successfully in the ecological literature (see e.g. Lam-Gordillo et al., 2025; Daudt et al., 2025; Wong et al., 2026), the uptake of these methods in areas of community ecology where ordination has typically been common has so far been slow: at least going by the ratio of downloads of the classical `vegan` `R`-package to more packages that implement model-based ordination (see Appendix S1). In our opinion, two potential barriers for improved uptake seem especially important. The first is a lack of accessible arguments and evidence for why GLLVMs make it possible to obtain better and more reliable ecological inference from one’s data as compared to traditional, non-model based methods. The second is a lack of instructive real-world examples that show the full capability of the framework in action on real ecological data.

This article sets out to help remove these two barriers by providing a focused and practically oriented guide to the tools and capabilities of the GLLVM framework, aimed at the types of ecological questions that may be especially relevant to current users of traditional ordination methods. The text is divided into four parts: 1) An overview of what we consider to be the most important fundamental advantages of using GLLVMs in community ecology, 2) how the methods can be used more concretely to address different types of ecological questions; both with and without observed environmental covariates, 3) a suggestion for a general modelling workflow when using GLLVMs to address these questions, and 4) a demonstration on this workflow on two relevant, real-world data sets.

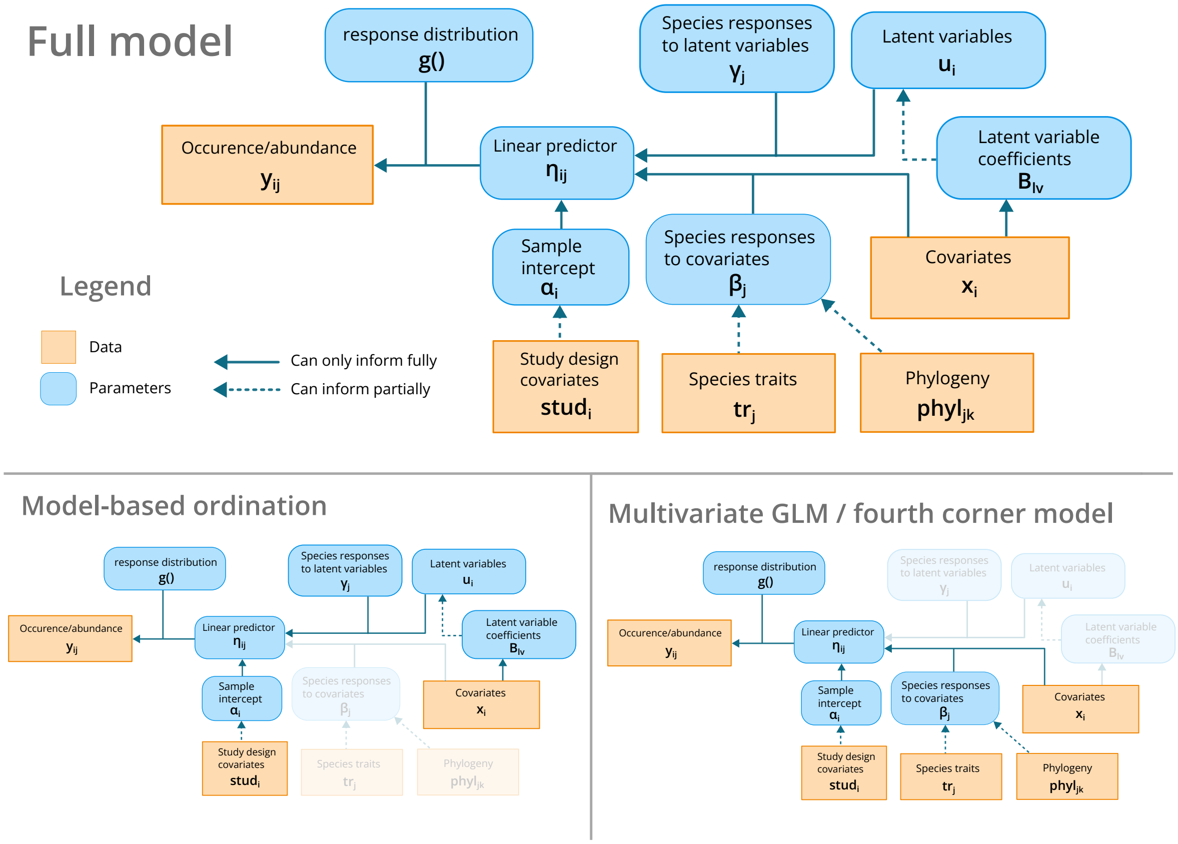

Figure 1: Graphical overview of the model structure of Generalized Linear Latent Variable models (GLLVMs), as implemented in the gllvm R package. Model components are named according to the model formulations by Niku et al. (2021) and van der Veen et al. (2023). The figure is inspired by Figure 4 from Ovaskainen et al. (2017).

1 Fundamental advantages of the GLLVM framework

The fact that GLLVMs are an extension of the Generalized Linear Mixed-effects Modelling Framework means that they offer the same options for specifying, fitting, interpreting and comparing models as classic GL(M)Ms. Here, we highlight six of the most substantial advantages that this brings to the analysis of multi-species community data. These advantages should be applicable regardless of the specific ecological questions asked.

1. **Accounting for different types of data** The GLLVM framework lets community ecologists analyse data as is, without data transformation or manipulation. As with GLMs, this done by specifying a suitable response distribution for the data, and by specifying the model's structure to match the study system or experimental design at hand. Most GLLVM software includes a variety of different response distributions, making it possible to model data recorded as presence-absence, counts, percentage cover, cover classes, biomass, and more (see e.g. Korhonen et al., 2025). Traditional multivariate methods (e.g. NMDS) offer ways to account for non-normality e.g. through the use of distance metrics, but these make the link to the ecological processes more opaque, confound results (Warton and Hui, 2017), and make assessment of fit to the data harder; whereas GLLVMs can use established tools for diagnostics (see point 1).

2. **Assessing model fit to the data** Sound ecological inference requires one's modelling assumptions to be met. To ensure this, the fit of any GLLVM can be assessed using diagnostic plots and metrics familiar from the GLMM framework, such as residual versus fitted plots or Q-Q plots. Specifically, the metrics used are randomized quantile residuals, similar to the `DHARMa` package (Hartig, 2024). As with classic GLMs, this is particularly relevant for assessing whether one's selected response distribution fits the data being analysed, e.g. if there are non-linear structures or overdispersion in the data that are not accounted for by the model. For example, when the observed data type are counts, residual or QQ-plots will indicate if a Poisson distribution is applicable. If the model predicts too few zeros relative to the data, it might be more reasonable to switch to a zero-inflated Poisson distribution or a negative-binomial distribution. As there is no clear way of evaluating whether the model assumptions are met simply by looking at the resulting ordination, this is not generally recommended as a way of assessing the fit. This is not to say that model misspecification cannot have a profound impact on the ordination, which it certainly can (see Warton and Hui, 2017, for the case of the mean-variance relationship and NMDS/DCA)

3. **Accounting for different study designs** In general, GLLVMs offer the same tools as GLMMs to account for properties of the sampling and study design, such as block- and hierarchical sampling designs, or differences in the read depth of samples in the case of DNA meta-barcoding data, which are not available for traditional multivariate methods. This can be done through fixed and random effects, nesting of effects, offsets or other changes to the model's structure. For example, blocks in a randomized block design can

be included as a random effect outside of a model-based ordination, to separate its effects from patterns of interest in the ordination (see the model formulation in Figure 1).

4. **Model comparison** The model-based nature of GLLVMs also allows for the use of a range of different goodness-of-fit statistics to compare the relative fit and predictive power of different models for species composition. For ecologists, Information Criteria like AIC and BIC, or area under the curve (AUC), will perhaps be the most familiar of these. Depending on the goal of the analysis, AIC or BIC can be used to determine the ideal set of observed predictor variables, or to determine the number of unobserved latent variables that best represent the data. Traditional counterparts to this are e.g. the use of stress to determine the number of dimensions in an NMDS ordination, or the use of pseudo-AIC in methods such as Canonical Correspondence Analysis (CCA) and Redundancy Analysis (RDA); see e.g. Dexter et al. (2018).

5. **Estimation and visualisation of uncertainty** Because GLLVMs are fitted using either (marginal) Maximum Likelihood estimation or with Bayesian methods, all parameters and fitted values estimated by the model have an associated measure of uncertainty. These uncertainties can be used to make statements about statistical significance, or alternatively, the "strength of evidence", of different model components (Muff et al., 2022). These uncertainties can then be visualized, e.g. by plotting confidence or prediction regions in an ordination diagram or intervals in a coefficient plot. In this regard, the uncertainties can serve the same purpose as multivariate permutation tests like PERMANOVA (Anderson, 2001), but are more versatile and interpretable, in the same way that confidence and prediction intervals in conventional statistical models are.

6. **Prediction** As statistical models, GLLVMs can also be used to predict or forecast, with associated uncertainty. This opens up many new possibilities for community ecologists, not available when applying traditional ordination methods. For example, one can predict how community composition is expected to change under different climate change scenarios (keeping all other predictors constant), or to validate how well the predicted species community of a given habitat type fits with newly collected data (see also Worked Example 2).

2 Using the framework to answer ecological questions

The main strength of the GLLVM framework for ecologists lies in its capability to provide in-depth answers to questions about the composition and structure of ecological communities. This includes questions about which species co-occur and which factors (habitat types, climatic variables, time etc.) best explain observed patterns of composition or co-occurrence. Among the most important tools to help researchers address these questions are the many options to effectively visualize model outputs that the GLLVM framework provides. Depending on the model and the goals of analysis, these can combine information from environmental-, species- and sample- specific parameters related to the latent variables. Figure 2 provides a general overview of the most relevant types of visualisations of the different model parameters shown in Figure 1.

This section is grouped into two parts: The first part focuses on questions that can be addressed by models only considering species observations, the second section focuses on questions involving measured environmental variables and ecological communities. However, it is important to bear in mind that contemporary community ecology studies often address multiple ecological questions simultaneously, sometimes by including both analyses on species composition alone and species composition in combination with environmental predictors. As such, the methods in the literature examples given between Section 2.1 and Section 2.2 will sometimes overlap.

2.1 Species composition data

When information on the environment is absent, GLLVMs can be a powerful tool for exploring basic patterns in a multispecies dataset. As with traditional methods, an unconstrained model-based ordination can be fitted to the species data alone, and patterns can be inferred from visualisation of the results. This basic GLLVM will return scores for each sample (traditionally called site scores) and species (similarly called loadings). These can then be used to make inferences about site conditions, transitions between community types, and which species associations drive these patterns. Conceptually, if we view the latent variable(s) as estimates of unobserved environmental gradients, the species loadings represent the slopes, or the species response, of each species to the gradient(s), similar to their response to predictor variables in a standard regression. The site scores then represent the specific values of these unobserved predictor variables, calculated for each sample. As such, the latent variables are similar to observed measures of the environment, e.g. pH or soil moisture; the difference being that they are estimated from the data rather than being measured in the field (Niku et al., 2019).

Visual inspection of GLLVM scores and loadings can be done the same way as with the results produced by other unconstrained ordination methods, such as NMDS or CA. Compared to traditional methods, GLLVMs have been shown to better capture both dataset properties and underlying ecological gradients in community data (Warton and Hui, 2017; Jupke and Schäfer, 2020; van der Veen et al., 2023). In addition, GLLVMs have two other important tools for visual inference which traditional ordination methods lack.

The first tool is a correlogram, or correlation plot. The sums of the square of the species loadings in a GLLVM are statistical estimates of the overall correlation between pairs of species in the data, which can be visualized in a correlogram (see Figure2). Together with ordination plots, correlograms can be effective tools to construct an overview of species co-occurrence patterns in one's data (Ovaskainen et al., 2017), although ordination plots makes it possible to also visualize the relationship between species scores and the samples or sites.

The second tool is uncertainty estimates — i.e. prediction and confidence intervals — for both the site scores and species loadings. These allow researchers to meaningfully evaluate

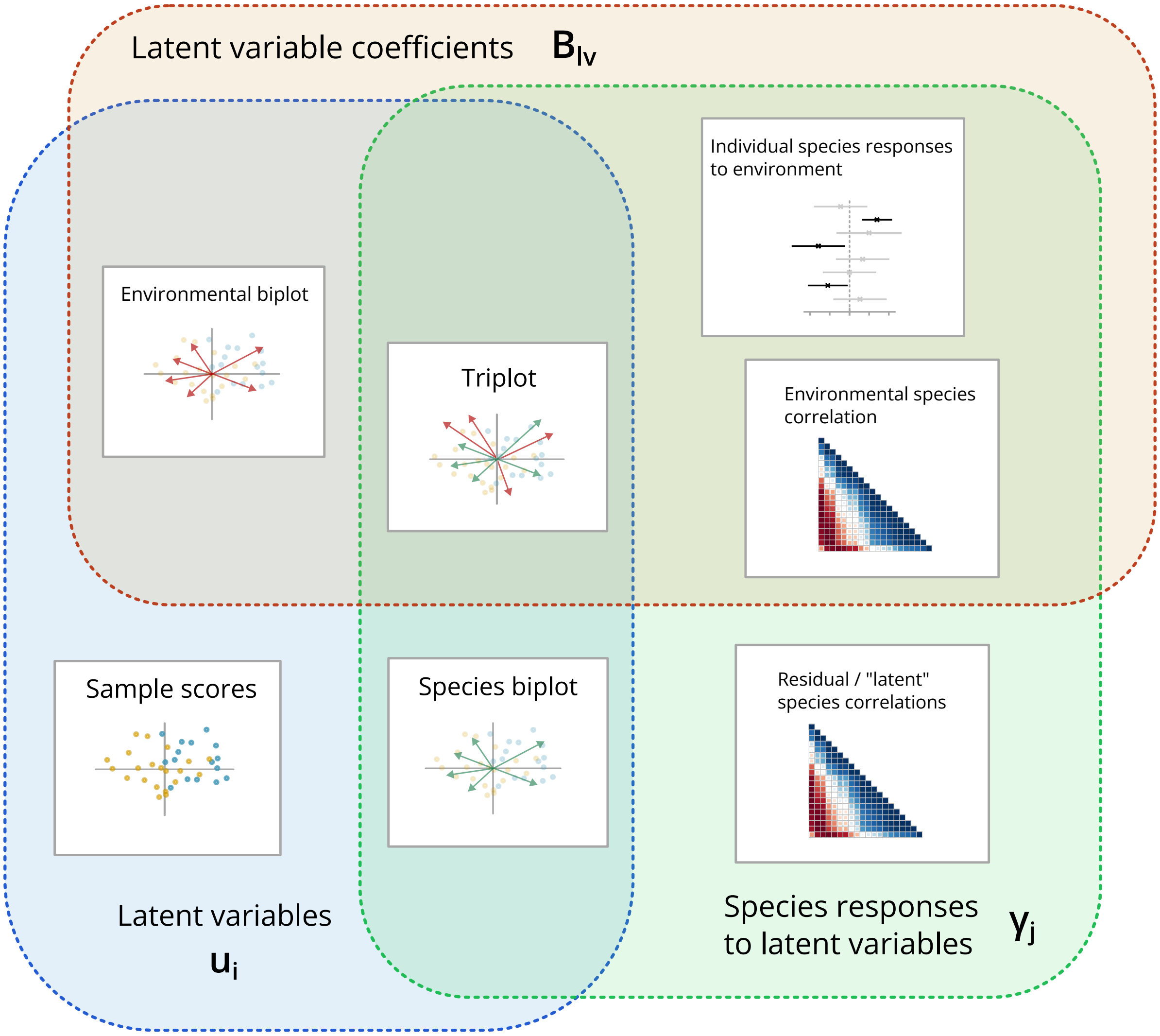

Figure 2: Overview of the different visualisations available for a GLLVM with latent variables. The colored areas represent the different model parameters introduced in Figure 1.

the statistical strength of evidence for the patterns observed in the data. For instance, if the prediction intervals of two site scores are clearly separated, it can be interpreted as the model being confident that the species compositions at these two sites are in fact different, and are expected to remain so if both sites were to be re-surveyed. The same logic holds for the species loadings, where uncertainties can be used to determine if two species are expected to co-occur.

These two tools, together with options for combining unconstrained ordinations with other forms of regression, allow a number of exploratory community ecology questions to be addressed in a single model-based framework. A selection of examples are presented in Table 1, although some may be considered exploratory before they are tackled by using information about the environment directly in the model. This will be discussed further in section 2.2.

Table 1: Examples of ecological questions that can be investigated in an exploratory manner using unconstrained ordination. The questions are broadly divided into fundamental (F) and applied (A) questions. Recent examples refer to studies in which these questions have recently been addressed using traditional methods for unconstrained ordination.

Question	Recent examples
F1: Does species composition change along one or more biotic or abiotic gradients (e.g. elevation, forest age, water salinity)	Handegard et al. (2024); Maunsell et al. (2013); Mulders et al. (2022)
F2: Are there seasonal patterns in community composition within a habitat?	Li et al. (2022); Naz et al. (2024)
F3: Are there characteristic clusters of species that tend to occur together in different sites, that can be interpreted as distinct communities?	Shembo et al. (2024); Lourenço et al. (2024)
F4: Are there associations between species in the community that are independent of associations accounted for by environmental predictors, which can be interpreted as biotic interactions?	Suárez-Tangil and Rodríguez (2023); Wang et al. (2025)
A1: How does the species composition of communities differ between different habitats or land management practices?	Larson et al. (2024); Fanfarillo et al. (2022); Graser et al. (2025); Pedley et al. (2023); Hu et al. (2024)
A2: Is there a difference between species composition of sites undergoing different ecological restoration treatments, and between those sites and undisturbed reference vegetation?	Brasil Neto et al. (2025); Helbing et al. (2023); Reis et al. (2022); see also worked example 2

Question	Recent examples
A3: How do alien species occur together with native species in an invaded community?	Hejda et al. (2023); Lanta et al. (2022); Reeve et al. (2022); see also worked example 1

2.2 Explaining species composition data using environmental predictors

When environmental predictors are available, the GLLVM framework offers even more tools to make inference about species-environment relationships. One approach is to use the environmental predictors to explain the distribution of each species individually, with the latent variables modelling any residual co-variation between the species (Ovaskainen et al., 2017). However, with large numbers of species, especially species that occur infrequently, this approach will quickly involve too many parameters to accurately estimate. A more parsimonious approach in line with ecological theory (ter Braak and Prentice, 1988; Legendre and Legendre, 2012), is to assume that species' distributions are explained by a few underlying latent variables that are, in turn, explained by environmental predictors.

The core model in this case is the *concurrent ordination*, where the latent variables depend on both environmental predictors and additional variation outside of the predictors (van der Veen et al., 2023). Concurrent ordination works by estimating latent variable coefficients (also called canonical coefficients; B_{lv} in Figure 1 and 2), that explain how a change in the latent variable (and thus the species composition) is associated with a change in each environmental variable (specifically, how much a latent variable changes following a one-unit change in a given environmental variable, all other variables being equal). In addition to the latent variable coefficients, the latent variables estimated by the models can also have a residual, or unexplained, component (for more detail see van der Veen et al., 2023). This means that the model can provide estimates not only of the degree to which the main patterns of species composition are explained by the environmental factors, but also to what degree there are additional unobserved factors driving species composition. The relative importance of the environmental and unobserved factors can then be disentangled by variance partitioning. In this regard, concurrent ordination addresses a longstanding problem with the use of unconstrained and constrained ordination (Økland, 1996; ter Braak and Šmilauer, 2015), as it simultaneously facilitates exploring species co-occurrence patterns and species-environment relationships.

Specifying the concurrent ordination to have no residual variation, i.e. assuming that the latent variables are completely explained by the environmental predictors, corresponds to what is traditionally called *constrained* or direct ordination, for which popular traditional methods include Canonical Correspondence Analysis (ter Braak, 1986) and Redundancy analysis (Legendre and Legendre, 2012). However, both of the aforementioned methods make strong assumptions about the distribution of the data, whereas GLLVMs are flexible enough to accommodate any data type found in community ecology (see Section 1).

Modeling communities with constrained or concurrent GLLVMs presents a number of additional features and tools for statistical inference over traditional methods: (1) As in the unconstrained case, the latent variable coefficients will have an uncertainty, and thus a confidence interval, associated with them. These confidence intervals can be used to make inference about the strength of evidence for the effect different environmental predictors, site scores and species loadings in the model. (2) Although the predictors affect the latent variables, they can be easily translated to predictor effects for individual species, making it straightforward to connect movement along environmental gradients to changes in individual species' abundances. As shown in Figure 2, the individual species effects, extracted from the model, are typically plotted using a caterpillar plot, while the latent variable coefficients are typically represented in an ordination biplot or triplot. (3) Predictor effects for the latent variables can be specified as either fixed or random effects (inside or outside the ordination), allowing for greater flexibility in the types of models that can be fitted. Non-linear effects such as splines can also be included in the model. (4) The relative importance of the different model components in explaining the responses of the different species can be assessed through variance partitioning. This includes assessing the importance of residual variation of the unexplained part of the latent variable(s) in a concurrent ordination, the effects of predictor variables both within and outside ordinations, and other model components, such as site intercepts, traits etc. (see Figure 1) in explaining the linear predictor for each species. Proportions of variance can be calculated to estimate the relative contributions of each model component in explaining each species' response.

Table 2 outlines some examples of ecological questions where models with concurrent or constrained latent variables would be relevant to answer ecological questions, as well as examples from the recent literature where they have been approached using mostly traditional methods.

Table 2: Examples of ecological questions that can be investigated using latent variable models with predictors, divided into fundamental (F) and applied (A) questions.

Question	Recent examples
F1: How do different environmental gradients (e.g. elevation, climate, water depth) explain differences in the community composition between sites?	Cheng et al. (2023); Young et al. (2022); Askeyev et al. (2023); Matavelli et al. (2022)
F2: Are specific species in a community indicators of changing environmental conditions?	Andrew-Priestley et al. (2022); Korolyuk et al. (2024)
A1: What is the effect of antropogenic vs. non-antropogenic factors in terms of explaining community composition?	Christman et al. (2022), Sanchez et al. (2023)
A2: Do certain environmental factors explain the prevalence of alien species in an ecosystem?	Kalusová et al. (2019), see also worked example 1

Question	Recent examples
A3: How does a community respond to different restoration treatments?	Crouch et al. (2022), see also worked example 2
A4: How will the composition of a community shift in response to changing climate?	Forte et al. (2024)

3 Guidelines for a GLLVM modeling workflow

Guidelines for other model-based analyses have been outlined by Warton et al. (2015), Zuur et al. (2010) and Zuur and Ieno (2016), among others, and the same recommendations generally hold for GLLVMs. Based on these, we present a five-step workflow, specifically geared toward the effective and sound application of GLLVMs in community ecology. The workflow outline is primarily adapted from Warton et al. (2015), and is summarized in Figure 3. Section 4 demonstrates the workflow on two relevant real-world data sets.

1. **Formulate the biological question as a statistical question** After the biological and ecological questions of the study are clarified, the first step in any model-based workflow should be to formulate them as concretely as possible in statistical terms. This means clarifying why a GLLVM is the right tool for the problem, and how exactly the model will be used to answer the ecological questions (e.g. which parameters should be included in the model).

   Ideally this first step should be undertaken before collecting data, in order to make sure that the study design and sampling strategies are geared towards getting the data needed to answer the ecological questions of interest (for a further discussion of this, see Warton et al., 2015).

   For example, if the main interest of a study is in making inference about how a species community changes along a temperature gradient, care should be taken to sample the environmental variables along that gradient so that they capture enough variation in the environment to meaningfully answer that question. Similarly, if the goal is to investigate the response of one or more specific focal species within the community to environmental and biotic changes, one should make sure to collect data on a wide enough range of conditions where they might be expected to occur and not occur (i.e., their niches should be well-sampled), in order to actually obtain enough data to make meaningful statistical inferences about their relationship to the environment and/or other species (see also worked example 1, as well as the Section 5). These considerations might occasionally also need to be balanced with strategies for ensuring sample representativeness, for example by deploying sampling methods that have some way of quantifying detectability (see e.g. Jeliazkov et al., 2022), as long as it is is consistent with the broader objectives of the study.

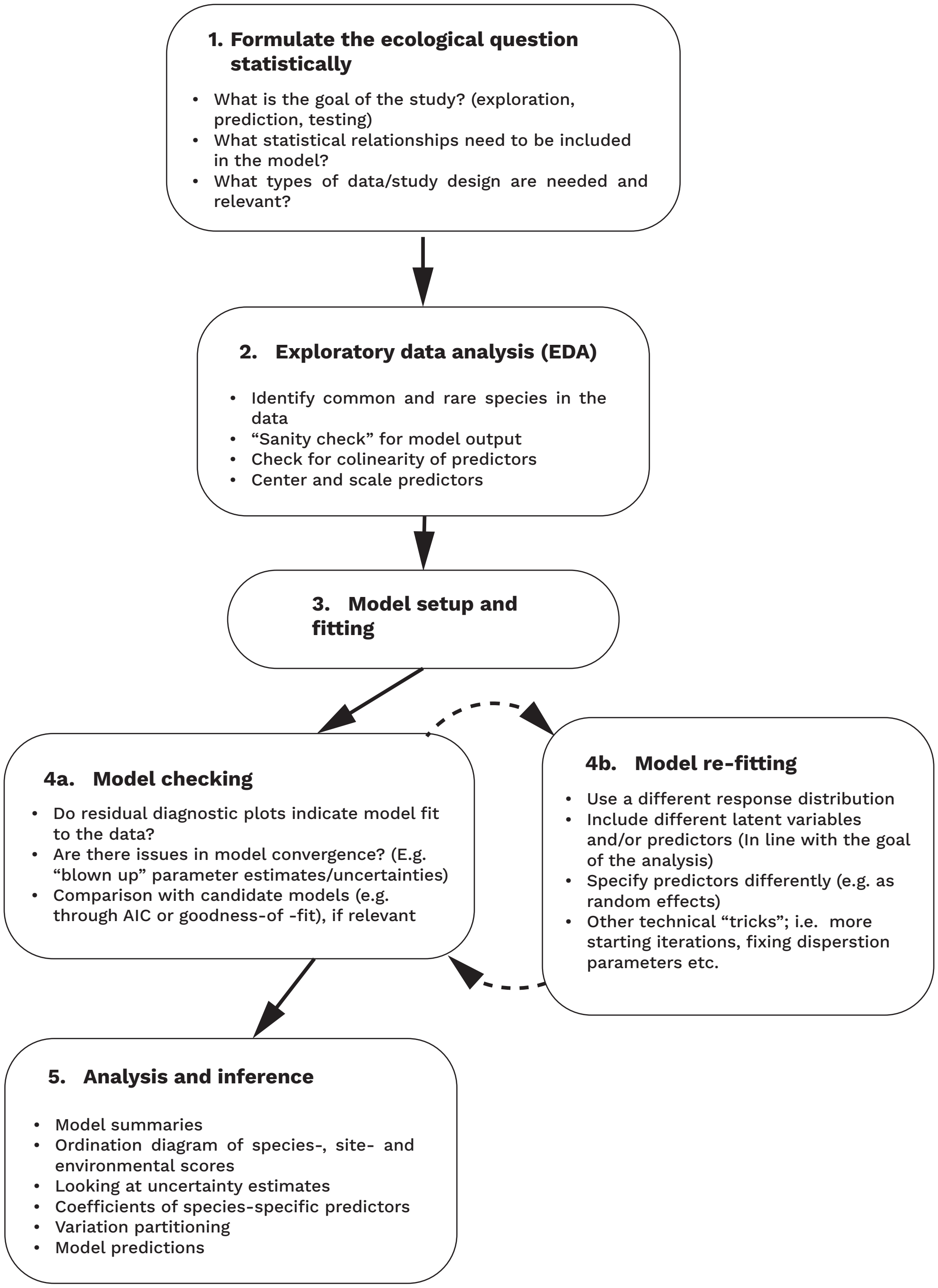

Figure 3: Visual representation of the analytic workflow suggested for modeling ecological communities with latent variables with GLLVMs. Adapted from Figure 1 in Warton et al. (2015).

This step also includes considering which type of model is best suited to answer one's research questions and represent the ecological relationships of interest. For example, if gathering data on environmental or habitat-type variables is part of the study, representing these in a concurrent ordination will often be a natural choice.

Clarifying whether the objective of one's study is primarily exploratory, confirmatory or predictive is arguably another important part of this step (Shmueli, 2010), particularly for guiding choices around the inclusion of predictor variables and model selection. If the goal is prediction, i.e. to find the GLLVM with the combination of predictor variables that most accurately predicts either community composition or the occurrence of specific focal species, optimizing one's model for this purpose through model selection, using e.g. AIC or similar tools, can be a meaningful strategy. However, if the goal of the analysis is rather to explore or make inference about the species community or communities in the data, as in the example above, variable selection for prediction could lead to biased inference and should in general be avoided (Sainani, 2014). Instead, all variables that are of interest should be included in the model (as the best statistical representation of the ecosystem), and the results of the fitted model should be explored as is. AIC or BIC might still be useful for determining the number of latent variables that best fit the data. Model selection of predictor variables based on optimizing for prediction should however be avoided, especially if the aim of the study is confirmatory, i.e. testing specific hypotheses about ecological relationships rather than exploring them more generally. Although in general, confirmatory analyses might be less common for the types of community ecology questions considered here.

In general, our modeling philosophy is the same as that of Ovaskainen et al. (2017): whenever possible, the aim should be to fit a single comprehensive model which can be used to address all relevant research questions, rather than analysing different models in parallel. This helps to streamline and making the analysis more reproducible, as well as preventing data dredging and ensuring that uncertainties are handled correctly.

2. **Exploratory data analysis (EDA)** After collecting data, and before fitting a GLLVM, exploratory inspection and visualisation of the raw data should always be done in order to get a better understanding of the dataset and to act as a sanity check on the model output. Relevant dataset properties to consider for GLLVMs are largely the same as for other models in the GLM family, and we generally recommend the same strategies proposed by Zuur et al. (2010).

   When dealing specifically with the types of multivariate species data considered here, we will also recommend a few additional exploratory strategies as good practice. The first is to simply get a broad-scale overview of the data by creating a table or histogram of how many samples (rows) each species (column) is observed in, as well as the inverse (how many different species are observed in each sample). This makes it possible to get a sense of how the data is spread over the samples, which e.g. can be seen in context with the sampling design, or to identify potentially data-deficient species (see discussion in Section 5). When species data are quantitative (i.e. not simply presence/absence), visualizing the relationship between species' prevalence in the data and their average abundance in each site with an Abundance-Occupancy (AO) plot can also be a helpful tool in this regard, making it possible to see whether the data follows the classic positive relationship commonly found in ecological data sets or not (Gaston et al., 2000), and whether some species deviates notably from others in terms of their AO-relationship – either due to factors do to the sampling design or the ecological dynamics of the system, which sometimes can be challenging to untangle (Russell et al., 2005; Gaston and Blackburn, 2003), but which in any case may provide important context for interpreting the results of a model fit.

Depending on the goals of the study, fitting a simple unconstrained ordination to the data – either through an unconstrained GLLVM or a classical method like PCA – could also be a part of this exploratory phase, to be used as a simple summary of the main species co-occurrence patterns in the data, before a model specifically geared towards one's research objectives is specified in step 3.

As for the EDA of predictor variables, visualizing their pairwise co-linearity using a correlation plot or similar is a good general-purpose tool for informing decisions about predictor inclusion in the model. However, as predictor collinearity is typically associated either with properties of the study design or inherent properties of the study system (e.g. the relationship between temperature and altitude), the question of whether to include or discard predictors due to collinearity should be informed by the goals of the study, study design and one's *a priori* knowledge of the study system, rather than numerical rules of thumb. Note also that while including highly co-linear predictors in a GLLVM might lead to increased uncertainty in the coefficient estimates and potentially convergence issues, it should not in principle lead to a change in which parameter estimates are favored by the model. Scaling and centering of the predictors is also recommended here as a standard procedure to improve coefficient estimation and convergence of the model before fitting.

3. **Model setup and fitting** Following from steps 1 and 2, the relevant model(s) should have been identified, and can now be fitted to the data. Important parts of the model to specify are (a) the response distribution for the species abundances/occurrences, (b) row (i.e. site) effects to explain the total abundance of individuals in the samples (i.e. predictors that effect the abundance of all species equally), (c) the number of latent variables (of different types) fitted to the data, and (d) model formulae for latent variables and species effects. It is important to note here that transforming, scaling or otherwise changing the species response variables in order to give more desirable statistical properties is, again, not in line with the GLLVM modeling philosophy. The focus should be on specifying an appropriate statistical response distribution that describes the data that was actually collected.

4. **Model checking and re-fitting** After a GLLVM model has been fitted to the data, it should be evaluated thoroughly. If there are issues with the model fit, these should be addressed and the model re-fit, as illustrated in the flowchart in Figure 3. As with classic GLMMs,

it is important to check that the data meet the model assumptions, by visualizing the residuals in diagnostic plots, as discussed in Section 1.

It can sometimes be difficult to get good convergence and numerical stability when fitting GLLVMs. Inspecting the gradient vector of the likelihood function to see if it is close to zero, or checking for artefacts such as negative estimates for the standard error of parameter estimators, can be useful tools to get an indication of this. visualisation of model estimates and uncertainties can also be helpful, e.g. if some species have ”exploding” species loading estimates or uncertainties. This typically happens when some species occur very infrequently in the dataset or are only associated with a subset of predictors (e.g. a species only occurs in one habitat, and habitat is included as a categorical predictor). While the easiest solution from a model stability perspective in this case is to filter out the ”problem species” from the data, this needs to be considered carefully in the context of the study. See 5 for a further discussion on this.

Another route to improvement is changing the model, perhaps by using a different response distribution (e.g. a zero-inflated Poisson distribution rather than a standard Poisson distribution, see 1.1), or specifying predictor effects as random rather than fixed. Excluding or including predictors (including more or fewer latent variables) can also help, if it does not clash with the aim of the study. A number of more technical tricks can also help, such as increasing the number of starting iterations, fixing dispersion parameters for the response distribution, or reordering the species in the response data. It might also be helpful to consult other articles discussing how to deal with model convergence in mixed models, e.g. Bolker et al. (2009).

After assessing the validity of the model, assessing the quality of the model with respect to prediction or selection, depending on the goal of the study, can be done in a number of ways. Information criteria like AIC or BIC are perhaps the most well-known. As these two criteria have slightly different interpretations (Aho et al., 2014), which criterion to use will depend on the objective of the study. Other measures of model predictive quality can also be assessed, e.g. root-mean square error of the prediction, or cross-validation.

5. **Visualisation and inference** After step 4 is completed, the model can finally be explored to make inferences about the relevant ecological questions of the study. We refer here primarily to Section 2 for a discussion of the different tools that can be used to make inferences from GLLVM models in terms of different ecological questions, as well as the worked examples.

4 Worked examples

In this section, we demonstrate how the GLLVM framework can be applied in real-world settings, using two relevant case studies from the recent ecological literature. The case studies are selected in order to showcase the tools and questions discussed in Section 2.1 and Section 2.2.

In order to demonstrate different paths to visualizing the output of GLLVM models, visualisations in Example 1 (Figure 4) are produced primarily using the native plotting functionality from the `gllvm` package, using the base `R` plotting interface, while visualisations in Example 2 (Figure 5 are constructed using the `ggplot2` package with extracted model components. Walk-throughs of the complete data analyses and visualisations, including figures for model diagnostics, are available in Appendix S2.

4.1 Example 1: Invasive trees in Argentina

In the first case study, we reanalyse data from Fernandez et al. (2021). Here, the researchers were interested in how the presence and abundance of an invasive tree species, the broad-leaf privet (*Ligustrum lucidum*), impacts the native tree community in an Argentinian second-growth subtropical forest.

Data on the tree community was recorded by measuring the basal area of 20 common species (including *L. lucidum*) in 164 forest monitoring plots. In a subset of 44 of these plots, samples of four physical-chemical characteristics of the soil: soil carbon content, nitrogen content, carbon to nitrogen ratio, and soil humidity, were collected as well.

For the purposes of this article, and in order to best help us showcase the GLLVM framework, we have condensed the ecological questions from Fernandez et al. (2021) into the following two research questions: 1) How is the abundance of *L. lucidum* in an area associated with the composition of other (native) tree species, and 2) Are some soil properties associated with increased abundance of *L. lucidum* specifically, compared to the native species?

4.1.1 Formulating the statistical question

In this case, the aim of the analysis is clearly exploratory, rather than confirmatory or predictive. No specific hypotheses about species-species or species-environment relationships are tested, and the goal is not to find a model that best predicts abundances of *L. lucidum* in the ecosystem. This suggests we should aim to model the data in a way that includes all relevant predictors of interest, and that extensive model selection beyond finding the optimal number of latent variables is not relevant.

However, the fact that environmental predictors (i.e. soil properties) are only available for a small subset of the vegetation plots, does present a challenge. In order to make the most of the data, we therefore veer slightly from our ideal workflow, and fit two different GLLVMs to the data: (1) A model with only unconstrained (i.e. not predictor informed) latent variables fitted to the full dataset; this will be used to make inferences about the patterns of co-occurrence between *L. lucidum* and the other species, and (2) a model with predictor informed latent variables (i.e., a concurrent ordination), fitted to the subset of plots with environmental variables recorded, using all 4 recorded soil properties as predictors. This second model will be used primarily to answer research question 2, make inferences about potential relationships between soil conditions and the co-occurrence of *L. lucidum* with native species. If predictor variables had been available for all plots, we could most likely have addressed all of these questions with a single concurrent ordination.

As the original study does not contain or consider explicit information about the study design, we will treat each sample (i.e. site) as independent. We do this by adding random intercepts for each row in the response data (see paragraph four in Section 1) to ensure the latent variables only account for composition rather than total abundance at each site.

4.1.2 Exploratory data analysis

Aggregating and visualizing the number of occurrences of all species in the full dataset (see Appendix S2, Section 3.2.1.), we see that every species appears in more than three plots. Of the 164 plots, only five contain just a single species, and the vast majority contains three or more species. Based on this, we assume that we have enough information in our data to avoid removing samples or species.

When selecting only the subset of the plots where soil variables were measured, however, two species were absent from all of these plots, and one species only occurred once. We thus excluded these three species from model 2, as they don't hold information, and keeping them will likely hurt model convergence.

Other than filtering the data, and centering and scaling all predictor variables to mean zero and unit variance, as discussed in Section 3, no further pre-processing was done for the data.

4.1.3 Model setup

Because our observed response variables are recorded as the area of each species in a plot, we decide to fit both models using a Tweedie distribution (Jørgensen, 1987). The Tweedie distribution arises as a Poisson sum of Gamma random variables. In other words, we assume that the number of observed individuals follows a Poisson distribution, and the area of each individual follows a Gamma distribution. As well as having an intuitive derivation, the distribution can accommodate species with zero area (unlike, for example, gamma and log normal distributions), and is also appropriate for data that follow Taylor's law (Kendal, 2004).

For both of the proposed models (the unconstrained and the concurrent), we intend to find the optimal number of latent variables which best fit the data. As discussed in Section 3, we decide to do this by finding the number of latent variables with the lowest information criterion that also fitted the data. In this case we will use AIC, as it is primarily recommended for exploratory analyses (Aho et al., 2014). It is also important to stress that in this case we only selected for the number of latent variables, not the predictors, due to the exploratory nature of the study.

We fit the models using the `gllvm()` function, with the syntax shown below, commented for clarity. We initially fit the models with one latent variable each, and proceed to add latent variables to find the AIC minimum, checking the diagnostics of each new model as we go. See Appendix S2, Section 3.3. for the full model fitting code, with explanatory comments.

4.1.4 Model checking and refitting

The diagnostic plots for both the unconstrained and constrained models did not indicate any violations of the model assumptions, and the addition of more latent variables to each model did not change this (see Sup. Figures 2.3, 2,4 and 2.6). The only caveat to this is that there seemed to be a slight structure in the residuals-versus-fitted plots — where the most prevalent species had slightly more negative residual than would be expected.

In the case of the unconstrained model, there was an AIC minimum for a model with five latent variables (see Sup. Table 2.1.). However, this was not as well converged as the model with three latent variables. Because of this, and partially in order to make the analysis as parsimonious as possible, we decided to continue with the model with three latent variables for the analysis (see Appendix S2, Section 3.4.1.). For the concurrent model, there was a clear AIC optimum at the model with two latent variables, and as such, we decided to continue with this model for visualisation and inference for the second part of the example.

4.1.5 visualisation and inference

Looking at the visualized species loadings of the unconstrained ordination (model 1) in Figure 4a, we see that *L. lucidum* is a clear outlier among all the other species. The predicted abundance of *L. lucidum* is primarily summarised by the first latent variable after rotating in the direction of maximum variance, as the position along the second latent variable (the vertical axis) is close to zero. As such, we might inspect the other species' responses to the first latent variable (the horizontal axis), for indications of their co-occurrence with the invasive species. The fact that only three other species have a positive loading along the first latent variable, and most other species are associated with the other end of the diagram, clearly indicates that an increased presence and biomass of *L. lucidum* is associated with fewer occurrences and lower biomass of most other tree species. This is also supported by the confidence intervals of the species loadings, in which the C.I. of *L. lucidum* overlaps with almost no other species.

These co-occurrence patterns are also clearly supported by Figure 4b, albeit more nuanced, as the correlation plot uses information from all three latent variables. The correlation of *L. lucidum* with the other species resulting from the species scores are all estimated to be negative, except in three cases. The ecological interpretation of this first model, then, is that *L. lucidum* seems to either displace most native species where it occurs, or that its environmental tolerance or preference is different from most other species in our data, thus thriving in conditions that are not favorable to other species. It could also be a combination of both scenarios, as

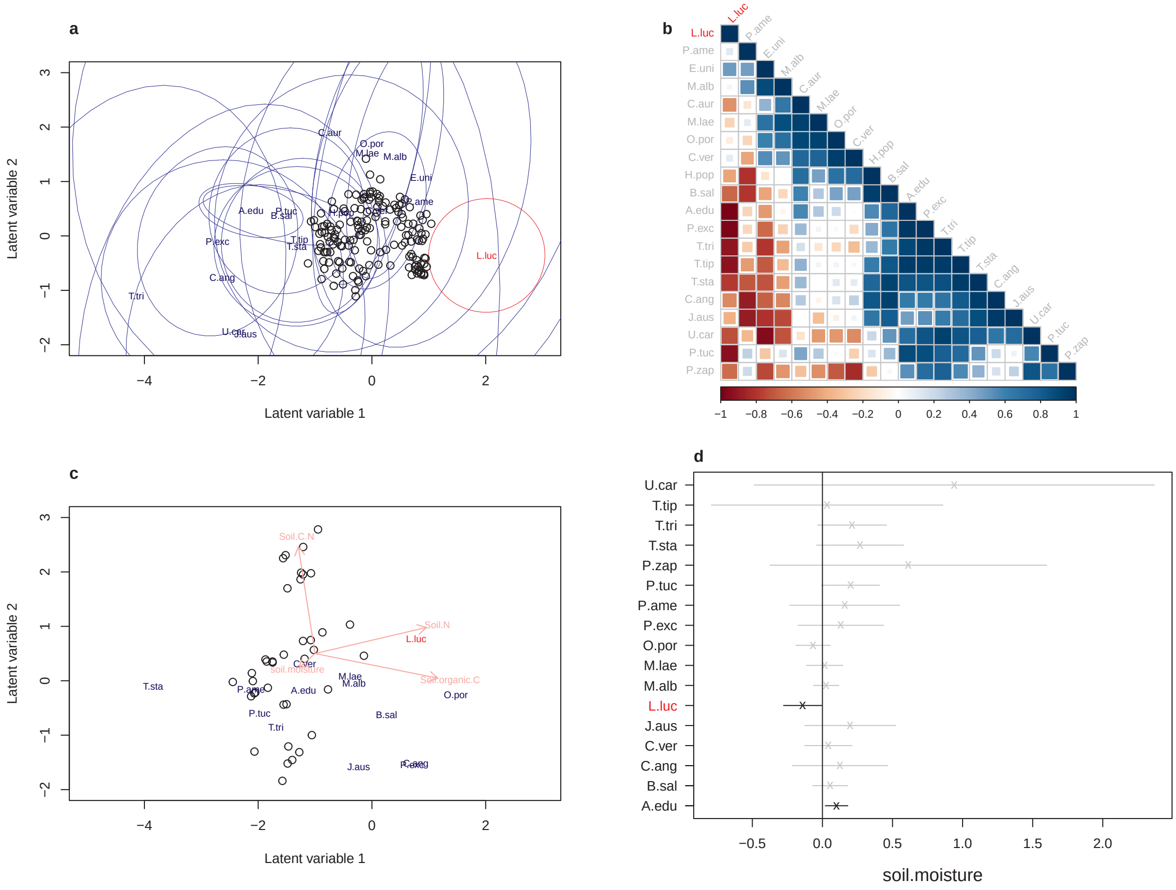

Figure 4: Selected visualisations of the estimates from the unconstrained (a,b) and concurrent (b,c) latent variable models. L.luc = *Ligustrum lucidum*, is indicated in red text in all figures. A: Site scores (black) and species scores (blue, red) for the unconstrained model with three latent variables (model 1) Ellipses represent prediction intervals for species scores. Species and site score and uncertainty ellipses of the three latent variables are all rotated in the directions of maximum variance to produce latent variable 1 and 2 using singular value decomposition, similar to a PCA rotation of an NMDS ordination. B) Correlation plot of the between-species correlations estimated from the species scores of the unconstrained model. C) Ordination diagram of site scores (black points), species scores (blue, red text) and environmental coefficients (red arrows) of the concurrent (predictor-informed) latent variable model using a subset of 40 points. Light red indicates that the 95% confidence interval of the latent variable predictor includes 0 for one or more of the latent variables, while the converse is true for the dark red arrows. D: Species specific coefficients (slopes) for the effect of soil moisture content on the abundance on the different species in the model (on the link scale), ordered from lowest to highest. Cross = coefficient estimate, line = 95% confidence intervals. Confidence intervals that cross 0 are indicated in grey.

Fernandez et al. (2021) hypothesize, in which *L. lucidum* alters the soil chemical properties where it establishes itself, making it more favorable for itself and less for the native species.

The concurrent ordination that includes environmental predictors (model 2), suggests that the observed environmental variables explain a significant portion of the community structure. As Figure 4c shows, the species scores of *L. lucidum* are separated from the others along the horizontal axis, as in the first model. Additionally, we see that it is clearly negatively associated with increasing soil moisture content, and positively associated with a larger soil carbon-to-nitrogen ratio. This is made even more clear when looking at the species-specific predictor effects in Figure 4d. *L. lucidum* is the only species which is estimated to decrease its abundance with higher soil moisture, while all other species respond either neutrally or positively to moisture. An inverse association seems to exist for the C:N ratio, although less pronounced, being shared by a few other species, as well as associated with higher uncertainties for all species (Sup. Figure 2.10). Variation partitioning also revealed soil moisture to be the variable explaining the highest mean proportion of variance for the species in the second model (Sup. Figure 2.11). However, the variance partitioning, as well as the model summary (Appendix S2, Section 3.5.2), also indicates that about 30% of the variation in the species composition was not explained by the environmental covariates , and is therefore an indication that there might be other important environmental predictors – or other dynamics in the community – that influence the species composition and which were not included in the model.

In summary, the ecological conclusion to draw from these two models seems relatively clear: The models provide a strong indication that in the ecosystem where it appears as an invasive species, *L. lucidum* is associated with a lower diversity and species abundance of most other native, common tree species. Secondly, this effect can largely be explained by *L. lucidum* either preferring, better tolerating or even facilitating drier, more nutrient-poor soils, supporting the initial hypothesis from Fernandez et al. (2021).

4.2 Example 2: Roadside Restoration in Norway

The second case study is based on data and ecological questions from Mehlhoop et al. (2022). Their aim was to assess the impact of different restoration efforts on roadside vegetation, in order to mitigate the effects of road construction. The dataset consists of the percentage cover of 164 different vascular plant species at 282 roadside plots across 3 regions in southern Norway. At each site, plots were subject to one of three restoration treatments: Re-seeding using commercial seed mixes, planting with native vegetation, or natural, i.e. unassisted re-vegetation. In addition, plots in intact reference vegetation were also sampled. Other variables, including the time since restoration (for the non-reference plots), as well as biological and environmental variables like soil organic matter content, canopy cover and grain size, were recorded at each plot to account for potential environmental factors that may influence species composition not directly related to restoration.

The primary research question of Mehlhoop et al. (2022) was how effective the three different restoration treatments were in bringing the vegetation of the impacted sites closer to the assumed natural vegetation in the reference sites. Ideally, this knowledge can then be used to inform future restoration efforts in similar nature types. As a secondary goal, chosen specifically to further showcase the capabilities of the GLLVM framework, we also ask how the vegetation in the restored sites is expected to change over the next 20 years.

4.2.1 Formulating the statistical question

The main goal of the analysis is to understand the relationship between a set of predictor variables (restoration method and time) and species occurrences in the data, and not to test any specific hypothesis. However, in contrast to the first example, the samples themselves (i.e. the restoration sites), rather than the species, are the primary unit of interest. Although the secondary goal is prediction oriented, the primary goal is explanatory in nature. As such, we decide to base our prediction on whichever model serves the explanatory purpose of the study best, rather than the other way around, even if that model might not predict optimally.

Consequently, fitting a concurrent ordination for the species composition including all potentially relevant predictors (treatment, time since restoration and the environmental variables) best aligns with the goals of this study. As the effect of time on species composition might be different for different restoration treatments, and because this potential difference is central to the ecological question in this case, we decide to include an interaction effect between the restoration treatments and time since restoration.

It is also necessary to think about how to account for the influence of our study design. In particular, there could be potential differences in the overall prevalences of species between the three different study regions that we want to separate out from the effect of restoration. To address this, we included region as a fixed row effect in the model, with additional random species-specific intercepts for each region. Within regions we might also want to account for differences in the sampling intensity between sites and plots. To do this, we add an additional random row effect for each site to account for potentially confounding differences in the total sample abundance between sites. In other words, we condition the ordination on the study design, and thus remove information about the effect of the regions and sites on the species community from the ordination.

4.2.2 Exploratory data analysis

Of the 164 species in the data, more than 50 only appear in a single plot and almost 40 appeared only two or three times. In order to reduce the chance that the final model is unduly influenced by data-deficient species, and because the focus of the study was the effect on restoration on the overall compositional differences between sites, rather than a focus on any particular species, we decided to exclude the species with three or fewer occurrences. Consequently we did not exclude any sites from our data. See Section 5 for a further discussion on the handling of data-deficient species in GLLVMs. We also scaled and centered all numeric predictor variables.

4.2.3 Model setup

Because our data is proportions with a large number of zeros, we used an ordered beta distribution as our response distribution (see Korhonen et al., 2024). We then set up the initial model following the structure outlined in the beginning of this section. To include the interaction effect between restoration treatment and time (and exclude a time interaction with the reference category), a custom model matrix was constructed where only interaction effects between the treatments and time were included (see Appendix S2, Section 4.2.).

As in worked example 1, we decide to use information criteria to determine the optimal number of latent variables. Code for the model fitting, with comments, can be found in Appendix S2, Section 4.3.

4.2.4 Model checking and re-fitting

Unlike in example 1, fitting the concurrent ordination specified above presented some numerical challenges, as models with both one, two, and three latent variables struggled to converge. We thus changed the fitting method to Extended variational approximation (Korhonen et al., 2023), and changed the ordering of the species in the input data, placing the most abundant species first. This helped to stabilise fitting of the models with one and two latent variables, however the model with three was still not able to converge. And while the diagnostic plots for both the one- and two latent variable models looked good, the model with two latent variables still showed some potential convergence issues. In particular, many of the variances of the parameter estimators calculated by the model were negative, which makes the model fit had to interpret.

As the model summaries of both models also indicated that the residual variation in the latent variables (i.e. the unexplained part), was consistently negligible (variance <e-7), we thus tried instead to fit a simpler model with constrained (i.e. fully predictor determined) latent variables, to make both the fitting and the inference easier. Still, the same lack-of-convergence problems persisted for the constrained models with two and three latent variables, in addition many species loadings being severely ”blown up” and linearly correlated in the ordination loadings, making the interpretation of the results ecologically questionable. As such, we ultimately decided to move forward with the model with one constrained latent variable for our analysis, even though the AIC was lower for the two-variable model for both the concurrent and constrained models (see Sup. Table 2.1).

4.2.5 Visualisation and inference

Our one-dimensional constrained latent variable model indicates that the different restoration treatments are the most important factor separating the species composition of the different sites (Figure 5a), with the reference vegetation sites clustering on one end of the scale, the naturally re-vegetated sites in the middle, and the planted and seeded sites on the other side. This is supported by the species loadings of the model, showing that the species most associated with the sites in the reference vegetation (left part of axis 1) are the european blueberry (*Vaccinium myrtillus*), may lily (*Maianthemum bifolium*) and oak (*Quercus robur*), all species characteristic of Norwegian south boreal forests, in which the reference plots were placed. Other tree species such as Norway spruce (*Picea abies*) were also strongly associated with the left-hand side. On the other side of the restoration axis, the plots undergoing seeding and planting were mostly associated with grasses such as red fescue (*Festuca rubra*), timothy (*Phleum pratense*) and small-reed (*Calamagrostis stricta*), plants more typical of typical of roadside vegetation and early succession, as well as some commercial seed mixes (Mehlhoop et al., 2022).

The estimated interaction between restoration treatment and time since restoration (i.e. how the effect of the restoration on species composition changes with time) is also different between treatments (Figure 5a, see also model summary in Appendix S2, Section 4.5.). Natural sites had a moderate trend towards the reference sites, while the effect was much smaller for the planted sites and absent for the seeded sites. The other measured environmental variables mostly have weak associations with the latent variable, exception for soil organic matter, which has a moderate correlation with the species composition of the intact sites.

The effect of the study design (Figure 5b) was also pronounced, explaining around 26% of the total variation in species responses according to variance partitioning (see Appendix S2, Section 4.5.). Among other things, we see that Region 3 was in general more species-rich than the other regions. This indicates that not accounting for the study design might have led to a different inference about the effect of restoration, because the distribution of the treatment groups and time since restoration is not equally distributed among the study regions (Mehlhoop et al., 2022), so the confounding could have lead to regional differences being modelled as treatment effects.

Forecasting 20 years in the future, assuming that all other environmental variables in the sites remain the same, the model predicts that the composition of the natural re-vegetated sites will have caught up to the composition of the reference forest, while sites in the other restoration treatment groups will have changed little (Figure 5c). Forecasting for two species which could potentially be used as indicator species, based on their species loadings and pre-existing knowledge about their ecology, *F. rubra* and *P. abies* (Figure 5d), underpins this by showing a marked difference between the different restoration treatments.

The main takeaway from this analysis is that the roadside vegetation sites that were left to naturally re-vegetate, were closer in terms of species composition to the forest reference than

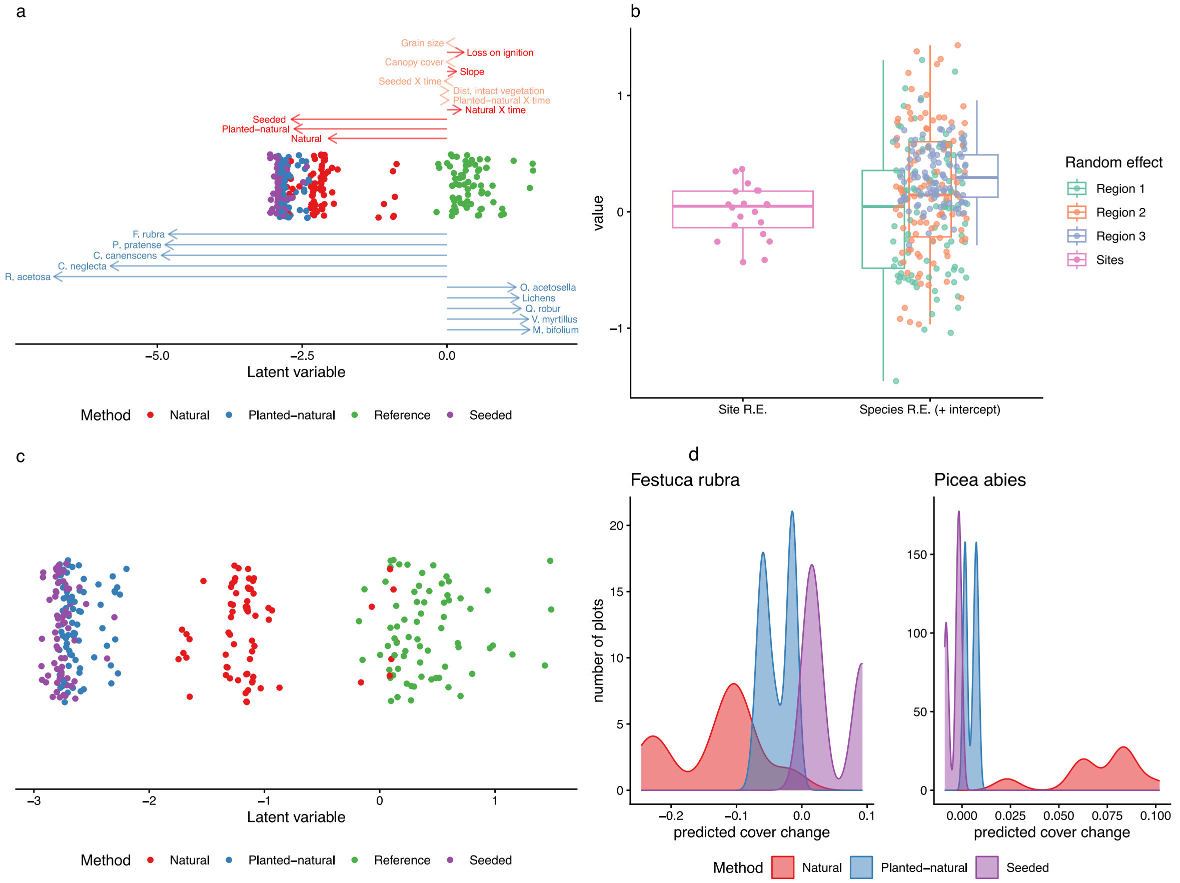

Figure 5: Visualisations of the estimates (a, b) and predictions (c,d) from the constrained latent variable model of the roadside vegetation vegetation. (a) One-dimensional diagram of the constrained latent variable. Sites are coloured by restoration treatment; red arrows indicate the latent variable coefficients of the model (X denotes an interaction effect). Dark red indicates that the 95% confidence interval of the predictor does not cross 0. Blue arrows show the species scores of the five most positively and five most negatively associated species with the latent variable. (b) Boxplot of the site-specific and species-specific random effects of the sites and different study regions in Southern Norway, respectively. For the species-region random effects, the combined effect of the fixed-effect intercepts of region 2 and 3 and the random species effect are shown (c) Predicted site scores for the latent variable with 20 years added to the site coefficients. (d) Density plot of predicted change in cover in different treatment groups for two potential indicator species, *Festuca rubra* and *Picea abies*.

the sites that had been artificially seeded. This vegetation treatment also showed a stronger response to time, which can be interpreted as a faster succession than the other treatments. Because no other variables in the model explained differences in species composition to the same degree, potential confounding effects of the study design were accounted for. Finally, because we did not estimate residual variation, we can be confident in the conclusion that the natural re-vegetation was the most effective method of restoration for the roadside vegetation communities.

5 Summary and discussion

In this article, we have provided an overview of Generalized Latent Variable models and a practical introduction to a range of their uses in community ecology. We have shown that a fully model-based methodology and workflow can produce models of ecosystems and communities that are feature rich as well as more statistically and conceptually interpretable than traditional ordination methods, e.g. by enabling features like prediction and uncertainty quantification.

We have made GLLVMs more tangible by demonstrating applications of the framework on two real world examples. In the first worked example, we showed how the impact of an invasive species on a community could be described using both unconstrained and concurrent model-based ordination. In contrast to traditional ordination methods, we could look at species loadings with uncertainties, as well as their associated between-species correlation estimates to indicate how strongly these associations were supported by the data. This let us paint more a comprehensive picture of what the data say about the associations between the native and invasive species in the community. The same was true for the effect of the predictors describing species co-occurrence in the communities, where the GLLVM option of visualizing the effect of covariates on individual species was able to identify the predictor most associated with the negative association between *L. lucidum* and the native species. This species-centered way of using GLLVMs will be readily transferable to a number of other ecological questions, such as identifying indicator species related to specific environmental variables of habitats, or identifying distinct clusters of species associations in a community (see references in Tables 1 and 2).

In the second worked example, we demonstrated how a concurrent ordination could be used to estimate the effect of different ecological restoration treatments on community composition, while accounting for a spatially grouped study design. Within a single model we could include the effect of the different treatments with parameter uncertainty, accounting for the study design, and forecast how the communities will change in the future; examples of capabilities of the GLLVM framework offer that is not possible to do in a comprehensive way with traditional methods. The use of the methods demonstrated in the example can serve as a relevant template for other sample-focused research questions. For instance, assessing the effect of different management practices on community composition, or which level of a hierarchical habitat classification system that best explains variation in community structure (see again references in Tables 1 and 2).

In both examples we explored the number of occurrences per species and site, which lead to removal of species in the second example. It is important to clarify that it is not strictly necessary to remove data deficient species prior to fitting a model-based ordination, but it can at times make the modelling process easier. Data deficiency can cause difficulties with model convergence, presenting results, or drawing inference. For example, species with only one or two occurrences on top of a mountain may exhibit extreme clustering in a (constrained) ordination diagram when the ordination axis represents elevation. Here, the model will interpret the data as the species not occurring at lower elevations at all, thus placing the species at the far end of the ordination axis. This is natural; the model has not seen any other information after all, but the results may not representative for the full niche of these species. Instead, it is an artefact of the sampling process. Still, at times a few extra species can add valuable information on the end points of an ecological gradient (i.e., serve to better inform the positions of site scores), so that removal is not always advisable. If all species on top of the mountain are data deficient, removing them will truncate the observed gradient, and impact the placement of all other sites and species in the data. Such data deficiency of species is often used as an argument for analysing the data in collapsed form, so that species identities are masked (as in e.g., NMDS), and the data are analysed on the basis of sites only. However, we argue that there is nothing inherently more complex to model-based ordination that makes it less suitable for the analysis of data deficient or rare species. The important thing is rather to distinguish between species that have few occurrences in data because they have been insufficiently sampled, so that they cannot be correctly placed in the environment, versus species that are rare for other reasons. Ideally, the pool of species being studied is clearly defined prior to data collection, so that a survey can be expanded to ensure data sufficiency for all species when necessary.

It is also important to stress that the GLLVM framework encompasses several avenues for modeling community data outside of the main use cases presented in this article. This includes the possibility of including more data types, such as species traits, in the models. Currently, traits can be incorporated into GLLVMs in two main ways: (1) using fourth-corner models, which estimate environment-trait interactions outside the ordination (Niku et al., 2021; Abrego et al., 2025), see also Figure 1), (2) reversing sites and species in a concurrent ordination, so that traits can be modelled on the latent variable(s) in the same way as environmental variables. This approach cannot include the environment as well. Alternatively, GLLVMs can been used to look at how traits covary between species, by letting species act in the place of a site, and traits as species. In other words, the GLLVM would be used to model a hypothetical lower-dimensional community trait space (Laughlin, 2014). Integrating functional traits and environmental predictors into the concurrent ordination framework, similar to approaches that have been developed for other ordination methods (ter Braak et al., 2018), is also currently an active area of development.

Other extensions that are available are using species phylogeny to inform species responses to the environment (van der Veen and O'Hara, 2025), modelling communities in time rather than space (Ovaskainen et al., 2017), and incorporating spatial autocorrelation in the latent variables (Thorson et al., 2015; Ovaskainen et al., 2017).

GLLVMs also open up a range of other avenues for modeling community ecology not possible with traditional (i.e. ordination) methods. This includes the possibility of using latent variables to model species niches (Ovaskainen et al., 2016) and niche overlap (van der Veen et al., 2024), including different niche sizes of species along latent variables (van der Veen et al., 2021), in order to model differences between generalists and specialists. Developing new types of model-based ecological indicators or classification schemes that may be useful in management settings, would be another interesting path to explore.

There has also been work done to develop methods and protocols for using pilot studies to determining the sampling effort and amount of data required to confidently answer specific ecological questions using GLLVM-type models (Maslen et al., 2023), which could potentially have a profound impact on resource- and time management when planning community ecology research.

In summary, applying GLLVMs in community ecology does not have to require a change in one's research questions, theoretical frameworks or data. Rather, it is a broad and robust toolbox that gathers a wide range of methodological tools in community ecology under the same statistical roof. It can both help you "do what you were already doing", only in more powerful and informative ways, as well as address ecological questions in new ways. As we have demonstrated, this makes the framework relevant for a number of research topics. Looking to the future, as GLLVMs become more widely adopted within community ecology, researchers will no doubt also discover new uses for the methods that their developers did not think of, which may again lead to further development of the framework. This underlines the importance of a constructive two-way collaboration between statistical developers and practitioners to address the pivotal ecological questions of the 21st century.

References

Abrego, N., P. Niittynen, J. Kemppinen, and O. Ovaskainen. 2025. Joint species-trait distribution modeling: The role of intraspecific trait variation in community assembly. *Ecology 106*(9): e70174. https://doi.org/10.1002/ecy.70174 .

Aho, K., D. Derryberry, and T. Peterson. 2014. Model selection for ecologists: The worldviews of AIC and BIC. *Ecology 95*(3): 631–636. https://doi.org/10.1890/13-1452.1 .

Anderson, M.J. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology 26*(1): 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x .

Andrew-Priestley, M., K. Newton, M.E. Platell, L. Le Strange, H. Houridis, M. Stat, R.M.K. Yu, C. Evans, Z. Rogers, J. Pallot, J. Van Den Broek, and G.R. MacFarlane. 2022. Benthic infaunal assemblages adjacent to an ocean outfall in Australian marine waters: Impact assessment and identification of indicator taxa. *Marine Pollution Bulletin* 174: 113229. https://doi.org/10.1016/j.marpolbul.2021.113229 .

Askeyev, O., S. Monakhov, I. Askeyev, A. Askeyev, and T.H. Sparks. 2023. Fish assemblages in lakes along environmental gradients at the eastern edge of Europe. *Environmental Biology of Fishes 106*(6): 1265–1276. https://doi.org/10.1007/s10641-023-01414-0 .

Bolker, B.M., M.E. Brooks, C.J. Clark, S.W. Geange, J.R. Poulsen, M.H.H. Stevens, and J.S.S. White. 2009. Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution 24*(3): 127–135. https://doi.org/10.1016/j.tree.2008.10.008 .

Brasil Neto, A.B., G. Schwartz, N.C. Noronha, M.A.P. Gama, G.C. Ferreira, E.J.M. Carvalho, and N.M.d.Q.X. Brasil. 2025. Seedling planting to restore ecosystems after bauxite mining in Brazilian amazon: The need of monitoring and reviewing techniques and procedures. *Plant Ecology 226*(8): 907–920. https://doi.org/10.1007/s11258-025-01539-5 .

Cheng, Z., T. Aakala, and M. Larjavaara. 2023. Elevation, aspect, and slope influence woody vegetation structure and composition but not species richness in a human-influenced landscape in northwestern Yunnan, China. *Frontiers in Forests and Global Change* 6. https://doi.org/10.3389/ffgc.2023.1187724 .

Christman, M.E., L.R. Spears, J.P. Strange, W.D. Pearse, E.K. Burchfield, and R.A. Ramirez. 2022. Land cover and climate drive shifts in *Bombus* assemblage composition. *Agriculture, Ecosystems & Environment* 339: 108113. https://doi.org/10.1016/j.agee.2022.108113 .

Crouch, C.D., B.O. Knapp, S.A. Cohen, J.S. Glitzenstein, J.L. Walker, and G.G. Wang. 2022. Longleaf pine restoration on hydric sites: Understorey plant community responses to site preparation through 15 years. *Applied Vegetation Science 25*(1): e12637. https://doi.org/10.1111/avsc.12637 .

Daudt, N.W., G. Loh, K.I. Currie, M.R. Schofield, R.O. Smith, E.J. Woehler, L. Bugoni, and W.J. Rayment. 2025. Changing species occurrences in seasonal seabird assemblages at the Subtropical Frontal Zone. *Estuarine, Coastal and Shelf Science* 323: 109405. https://doi.org/10.1016/j.ecss.2025.109405 .

Dexter, E., G. Rollwagen-Bollens, and S.M. Bollens. 2018. The trouble with stress: A flexible method for the evaluation of nonmetric multidimensional scaling. *Limnology and Oceanography: Methods 16*(7): 434–443. https://doi.org/10.1002/lom3.10257 .

Fanfarillo, E., D. Calabrese, C. Angiolini, G. Bacaro, S. Biagiotti, P. Castagnini, S. Loppi, T. Martellini, and S. Maccherini. 2022. Effects of conventional and organic management on plant and insect communities in a traditional elephant garlic crop. *Community Ecology 23*(3): 417–427. https://doi.org/10.1007/s42974-022-00091-w .

Fernandez, R.D., P. Castro-Díez, R. Aragón, and N. Pérez-Harguindeguy. 2021. Changes in community functional structure and ecosystem properties along an invasion gradient of Ligustrum lucidum. *Journal of Vegetation Science 32*(6): e13098. https://doi.org/10.1111/jvs.13098 .

Forte, T.G.W., M. Carbognani, A. Vannini, G. Chiari, and A. Petraglia. 2024. Short-term vegetation shifts in an alpine grassland under current and simulated climate change. *Journal of Vegetation Science 35*(6): e70000. https://doi.org/10.1111/jvs.70000 .

Gaston, K.J. and T.M. Blackburn. 2003. Dispersal and the interspecific abundance-occupancy relationship in British birds. *Global Ecology and Biogeography 12*(5): 373–379. https://doi.org/10.1046/j.1466-822X.2003.00054.x .

Gaston, K.J., T.M. Blackburn, J.J. Greenwood, R.D. Gregory, R.M. Quinn, and J.H. Lawton. 2000. Abundance–occupancy relationships. *Journal of Applied Ecology 37*(s1): 39–59. https://doi.org/10.1046/j.1365-2664.2000.00485.x .

Graser, A., M. Georg, J. Kallmayer, A. Marten, C. Pertl, H. Rumpf, C. Senf, and J. Kamp. 2025. Large-scale forest disturbance and associated management shape bird communities in Central European spruce forests. *Journal of Applied Ecology 62*(2): 329–343. https://doi.org/10.1111/1365-2664.14849 .

Handegard, E., I. Gjerde, R. Halvorsen, R. Lewis, K.O. Storaunet, M. Sætersdal, and O. Skarpaas. 2024. How important is Forest Age in explaining the species composition of Near-natural Spruce Forests? *Forest Ecology and Management* 569: 122170. https://doi.org/10.1016/j.foreco.2024.122170 .

Hartig, F. 2024. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models*. R package version 0.4.7. https://doi.org/10.32614/CRAN.package.DHARMa.

Hejda, M., J. Čuda, K. Pyšková, L.C. Foxcroft, K.V. Nkuna, A. Novoa, and P. Pyšek. 2023. Impacts of invasive alien species on riparian plant communities in South African savanna. *Journal of Tropical Ecology* 39: e39. https://doi.org/10.1017/S0266467423000299 .

Helbing, F., T. Fartmann, C. Morkel, and D. Poniatowski. 2023. Rapid response of vascular plants and insects to restoration of montane grasslands. *Frontiers in Ecology and Evolution* 11. https://doi.org/10.3389/fevo.2023.1148266 .

Herrmann, A., K. Grabow, and A. Martens. 2022. The invasive crayfish Faxonius immunis causes the collapse of macroinvertebrate communities in Central European ponds. *Aquatic Ecology 56*(3): 741–750. https://doi.org/10.1007/s10452-021-09935-5 .

Hu, Z., M. Fernández-Martínez, Q. He, Z. Xu, L. Jiang, G. Zhou, J. Chen, M. Nie, Q. Yu, H. Feng, Z. Huang, and S.T. Michaletz. 2024. Fungal composition associated with host tree identity mediates nutrient addition effects on wood microbial respiration. *Ecology 105*(8): e4375. https://doi.org/10.1002/ecy.4375 .

Hui, F.K. 2025. *boral: Bayesian Ordination and Regression AnaLysis.* R package version 2.0.3. https://doi.org/10.32614/CRAN.package.boral.

Hui, F.K., S. Taskinen, S. Pledger, S.D. Foster, and D.I. Warton. 2015. Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution 6*(4): 399–411. https://doi.org/10.1111/2041-210X.12236 .

Jeliazkov, A., Y. Gavish, C.J. Marsh, J. Geschke, N. Brummitt, D. Rocchini, P. Haase, W.E. Kunin, and K. Henle. 2022. Sampling and modelling rare species: Conceptual guidelines for the neglected majority. *Global Change Biology 28*(12): 3754–3777. https://doi.org/10.1111/gcb.16114 .

Jørgensen, B. 1987. Exponential Dispersion Models. *Journal of the Royal Statistical Society. Series B (Methodological) 49*(2): 127–162. https://doi.org/10.1111/j.2517-6161.1987.tb01685.x .

Jupke, J.F. and R.B. Schäfer. 2020. Should ecologists prefer model- over distance-based multivariate methods? *Ecology and Evolution 10*(5): 2417–2435. https://doi.org/10.1002/ece3.6059 .

Kalusová, V., N. Čeplová, M. Chytrý, J. Danihelka, D. Pavel, K. Fajmon, O. Hájek, V. Kalníková, P. Novák, V. Řehořek, J. Těšitel, L. Tichý, T. Wirth, and Z. Lososová. 2019. Similar responses of native and alien floras in European cities to climate. *Journal of Biogeography 46*(7): 1406–1418. https://doi.org/10.1111/jbi.13591 .

Kendal, W.S. 2004. Taylor's ecological power law as a consequence of scale invariant exponential dispersion models. *Ecological Complexity 1*(3): 193–209. https://doi.org/10.1016/j.ecocom.2004.05.001 .

Korhonen, P., F.K. Hui, J. Niku, and S. Taskinen. 2023. Fast and universal estimation of latent variable models using extended variational approximations. *Statistics and Computing 33*(1): 26. https://doi.org/10.1007/s11222-022-10189-w .

Korhonen, P., F.K.C. Hui, J. Niku, S. Taskinen, and B. van der Veen. 2024. A comparison of joint species distribution models for percent cover data. *Methods in Ecology and Evolution 15*(12): 2359–2372. https://doi.org/10.1111/2041-210X.14437 .

Korhonen, P., F.K.C. Hui, J. Niku, S. Taskinen, and B. van der Veen. 2025. Gllvm 2.0: Fast fitting of advanced ordination methods and joint species distribution models. *PeerJ* 13: e20338. https://doi.org/10.7717/peerj.20338 .

Korolyuk, A.Y., H.F. Shomurodov, B.S. Khabibullaev, and Z.S. Sadinov. 2024. Composition and Structure of Tugai Communities in the Indication of Ecological Conditions in the Lower Amu Dar'ya. *Contemporary Problems of Ecology 17*(1): 106–111. https://doi.org/10.1134/S1995425524010074 .

Lam-Gordillo, O., G.L. Petersen, S.F. Hailes, K. Carter, N.H. Salmond, L. McCartain, M. Ferries, B. Greenfield, E.J. Douglas, R. Hattingh, T. Drylie, B. Shanahan, and A.M. Lohrer. 2025. Extreme weather event causes contrasting macrobenthic disturbance-recovery dynamics in two New Zealand estuaries. *Next Research 2*(3): 100726. https://doi.org/10.1016/j.nexres.2025.100726 .

Lanta, V., P. Liancourt, J. Altman, T. Černý, M. Dvorský, P. Fibich, L. Götzenberger, O. Hornych, J. Miklín, P. Petřík, P. Pyšek, L. Čížek, and J. Doležal. 2022. Determinants of invasion by single versus multiple plant species in temperate lowland forests. *Biological Invasions 24*(8): 2513–2528. https://doi.org/10.1007/s10530-022-02793-8 .

Larson, D.L., M. Simanonok, A. Landsman, J.L. Larson, C. Davies, and C.R.V. Otto. 2024. Bee Habitat, but Not Bee Community Structure, Varies Across Grassland Management in Four National Parks in the Mid-Atlantic, USA. *Ecology and Evolution 14*(12): e70719. https://doi.org/10.1002/ece3.70719 .

Laughlin, D.C. 2014. The intrinsic dimensionality of plant traits and its relevance to community assembly. *Journal of Ecology 102*(1): 186–193. https://doi.org/10.1111/1365-2745.12187 .

Legendre, P. and L. Legendre. 2012. Chapter 11 - Canonical analysis, In *Numerical Ecology*, eds. Legendre, P. and L. Legendre, Volume 24 of *Developments in Environmental Modelling*, 625–710. Elsevier. https://doi.org/10.1016/B978-0-444-53868-0.50011-3.

Li, S., Z. Qian, J. Yang, Y. Lin, H. Li, and L. Chen. 2022. Seasonal variation in structure and function of gut microbiota in Pomacea canaliculata. *Ecology and Evolution 12*(8): e9162. https://doi.org/10.1002/ece3.9162 .

Lourenço, Á., C.V. Souza, A.F. Mendonça, G.G. Reis, P.F. Linhares, R.P. Moura, and E.M. Vieira. 2024. Increasing fire severity alters the species composition and decreases richness of seeds potentially dispersed by small mammals. *Biotropica 56*(3): e13318. https://doi.org/10.1111/btp.13318 .

Maslen, B., G. Popovic, M. Lim, E. Marzinelli, and D. Warton. 2023. How many sites? Methods to assist design decisions when collecting multivariate data in ecology. *Methods in Ecology and Evolution 14*(6): 1564–1573. https://doi.org/10.1111/2041-210X.14094 .

Matavelli, R., J.M. Oliveira, J. Soininen, M.C. Ribeiro, and J. Bertoluci. 2022. Altitude and temperature drive anuran community assembly in a Neotropical mountain region. *Biotropica 54*(3): 607–618. https://doi.org/10.1111/btp.13074 .

Maunsell, S.C., R.L. Kitching, P. Greenslade, A. Nakamura, and C.J. Burwell. 2013. Springtail (Collembola) assemblages along an elevational gradient in Australian subtropical rainforest. *Australian Journal of Entomology 52*(2): 114–124. https://doi.org/10.1111/aen.12012 .

McGillycuddy, M., D.I. Warton, G. Popovic, and B.M. Bolker. 2025. Parsimoniously fitting large multivariate random effects in glmmTMB. *Journal of Statistical Software 112*(1): 1–19. https://doi.org/10.18637/jss.v112.i01 .

Mehlhoop, A.C., A.B. Skrindo, M. Evju, and D. Hagen. 2022. Best practice—Is natural revegetation sufficient to achieve mitigation goals in road construction? *Applied Vegetation Science 25*(3): e12673. https://doi.org/10.1111/avsc.12673 .

Muff, S., E.B. Nilsen, R.B. O'Hara, and C.R. Nater. 2022. Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution 37*(3): 203–210. https://doi.org/10.1016/j.tree.2021.10.009 .

Mulders, Y., K. Filbee-Dexter, S. Bell, N.E. Bosch, A. Pessarrodona, D. Sahin, S. Vranken, S. Zarco-Perello, and T. Wernberg. 2022. Intergrading reef communities across discrete seaweed habitats in a temperate–tropical transition zone: Lessons for species reshuffling in a warming ocean. *Ecology and Evolution 12*(1): e8538. https://doi.org/10.1002/ece3.8538 .

Naz, F., M. Arif, T. Xue, and L. Changxiao. 2024. Seasonal dynamics of soil ecosystems in the riparian zones of the Three Gorges Reservoir, China. *Global Ecology and Conservation* 54: e03174. https://doi.org/10.1016/j.gecco.2024.e03174 .

Niku, J., W. Brooks, R. Herliansyah, F.K.C. Hui, P. Korhonen, S. Taskinen, B. van der Veen, and D.I. Warton 2025. *gllvm: Generalized Linear Latent Variable Models.* R package version 2.0.5. https://doi.org/10.32614/CRAN.package.gllvm.

Niku, J., F.K.C. Hui, S. Taskinen, and D.I. Warton. 2019. Gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution 10*(12): 2173–2182. https://doi.org/10.1111/2041-210X.13303 .

Niku, J., F.K.C. Hui, S. Taskinen, and D.I. Warton. 2021. Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Environmetrics 32*(6): e2683. https://doi.org/10.1002/env.2683 .

Økland, R.H. 1996. Are ordination and constrained ordination alternative or complementary strategies in general ecological studies? *Journal of Vegetation Science 7*(2): 289–292. https://doi.org/10.2307/3236330 .

Ovaskainen, O., N. Abrego, P. Halme, and D. Dunson. 2016. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution 7*(5): 549–555. https://doi.org/10.1111/2041-210X.12501 .

Ovaskainen, O., G. Tikhonov, D. Dunson, V. Grøtan, S. Engen, B.E. Sæther, and N. Abrego. 2017. How are species interactions structured in species-rich communities? A new method for analysing time-series data. *Proceedings of the Royal Society B: Biological Sciences 284*(1855): 20170768. https://doi.org/10.1098/rspb.2017.0768 .

Ovaskainen, O., G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters 20*(5): 561–576. https://doi.org/10.1111/ele.12757 .

Pedley, S.M., P. Wolstenholme, and P.M. Dolman. 2023. Plantation clear-fell patches benefit heathland arthropods. *Ecological Solutions and Evidence 4*(3): e12281. https://doi.org/10.1002/2688-8319.12281 .

Popovic, G.C., D.I. Warton, F.J. Thomson, and A.T. Moles. 2019. Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution 10*(9): 1571–1583. https://doi.org/10.1111/2041-210X.13247 .

Reeve, S., D.C. Deane, C. McGrannachan, G. Horner, C. Hui, and M. McGeoch. 2022. Rare, common, alien and native species follow different rules in an understory plant community. *Ecology and Evolution 12*(3): e8734. https://doi.org/10.1002/ece3.8734 .

Reis, B.P., K. Szitár, A. Kövendi-Jakó, K. Török, N. Sáradi, E. Csákvári, and M. Halassy. 2022. The long-term effect of initial restoration intervention, landscape composition, and time on the progress of Pannonic sand grassland restoration. *Landscape and Ecological Engineering 18*(4): 429–440. https://doi.org/10.1007/s11355-022-00512-y .

Ribeiro, L.G., H.H. Puerari, A.O. Silva, K.A. Vaz, J.V. dos Santos, C.A. Nunes, M.V. Barbosa, M.R. da Rocha, J.O. Siqueira, and M.A.C. Carneiro. 2023. Structure and composition of the nematode community in a restoration area affected by iron tailings. *Pedobiologia* 97–98: 150864. https://doi.org/10.1016/j.pedobi.2023.150864 .

Russell, L.K., S.J. Evans, L. Smith, C.M. Bevers, A.P. Luxford, W.J. Stubbs, and J.B. Wilson. 2005. Distribution/abundance relations in a New Zealand grassland landscape. *New Zealand Journal of Ecology 29*(1): 61–68 .

Sahade, R., C. Lagger, L. Torre, F. Momo, P. Monien, I. Schloss, D.K.A. Barnes, N. Servetto, S. Tarantelli, M. Tatián, N. Zamboni, and D. Abele. 2015. Climate change and glacier retreat drive shifts in an Antarctic benthic ecosystem. *Science Advances 1*(10): e1500050. https://doi.org/10.1126/sciadv.1500050 .

Sainani, K.L. 2014. Explanatory Versus Predictive Modeling. *PM&R 6*(9): 841–844. https://doi.org/10.1016/j.pmrj.2014.08.941 .

Sanchez, K.A., L. Benedict, and E.A. Holt. 2023. Landscape composition is a stronger determinant than noise and light of avian community structure in an urbanizing county. *Frontiers in Ecology and Evolution* 11. https://doi.org/10.3389/fevo.2023.1254280 .

Shembo, A.K., S.S. Ayichew, I. Stiers, A. Geremew, and L. Carson. 2024. Classification and ordination analysis of wild medicinal plants in Ada'a district, Ethiopia: Implication for sustainable conservation and utilization. *Ecological Frontiers 44*(4): 809–819. https://doi.org/10.1016/j.ecofro.2024.04.002 .

Shmueli, G. 2010. To Explain or to Predict? *Statistical Science 25*(3): 289–310. https://doi.org/10.1214/10-STS330 .

Souza-Alonso, P., Y. Lechuga-Lago, A. Guisande-Collazo, and L. González. 2022. Evidence of functional and structural changes in the microbial community beneath a succulent invasive plant in coastal dunes. *Journal of Plant Ecology 15*(6): 1154–1167. https://doi.org/10.1093/jpe/rtac026 .

Suárez-Tangil, B.D. and A. Rodríguez. 2023. Environmental filtering drives the assembly of mammal communities in a heterogeneous Mediterranean region. *Ecological Applications 33*(2): e2801. https://doi.org/10.1002/eap.2801 .

ter Braak, C.J.F. 1986. Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology 67*(5): 1167–1179. https://doi.org/10.2307/1938672 .

ter Braak, C.J.F. and I.C. Prentice. 1988. A Theory of Gradient Analysis, In *Advances in Ecological Research*, eds. Begon, M., A.H. Fitter, E.D. Ford, and A. Macfadyen, Volume 18, 271–317. Academic Press. https://doi.org/10.1016/S0065-2504(08)60183-X.

ter Braak, C.J.F. and I.C. Prentice. 2004. A Theory of Gradient Analysis, *Advances in Ecological Research*, Volume 34 of *Advances in Ecological Research: Classic Papers*, 235–282. Academic Press. https://doi.org/10.1016/S0065-2504(03)34003-6.

ter Braak, C.J.F. and P. Šmilauer. 2015. Topics in constrained and unconstrained ordination. *Plant Ecology 216*(5): 683–696. https://doi.org/10.1007/s11258-014-0356-5 .

ter Braak, C.J.F., P. Šmilauer, and S. Dray. 2018. Algorithms and biplots for double constrained correspondence analysis. *Environmental and Ecological Statistics 25*(2): 171–197. https://doi.org/10.1007/s10651-017-0395-x .

Thorson, J.T., M.D. Scheuerell, A.O. Shelton, K.E. See, H.J. Skaug, and K. Kristensen. 2015. Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution 6*(6): 627–637. https://doi.org/10.1111/2041-210X.12359 .

Tikhonov, G., O. Ovaskainen, J. Oksanen, M. de Jonge, O. Opedal, and T. Dallas 2025. *Hmsc: Hierarchical Model of Species Communities*. R package version 3.3-7. https://doi.org/10.32614/CRAN.package.Hmsc.

van der Veen, B., F.K.C. Hui, K.A. Hovstad, and R.B. O'Hara. 2023. Concurrent ordination: Simultaneous unconstrained and constrained latent variable modelling. *Methods in Ecology and Evolution 14*(2): 683–695. https://doi.org/10.1111/2041-210X.14035 .

van der Veen, B., F.K.C. Hui, K.A. Hovstad, E.B. Solbu, and R.B. O'Hara. 2021. Model-based ordination for species with unequal niche widths. *Methods in Ecology and Evolution 12*(7): 1288–1300. https://doi.org/10.1111/2041-210X.13595 .

van der Veen, B. and R.B. O'Hara. 2025. Fast fitting of phylogenetic mixed-effects models. Preprint, arXiv, https://doi.org/10.48550/arXiv.2408.05333.

van der Veen, B., R.B. O'Hara, F.K. Hui, and K.A. Hovstad. 2024. Predicting niche overlap with model-based ordination. *Ecography 2024*(4): e06938. https://doi.org/10.1111/ecog.06938 .

Wang, D., Y. Zhu, Z. Li, X. Yang, S. Kwon, Z. Shi, and T. Indree. 2025. Vegetation community reassembly changes in Eastern Eurasian degraded steppe: Roles of environmental filtering and biotic interaction. *Ecological Frontiers* 45. https://doi.org/10.1016/j.ecofro.2025.01.015 .

Warton, D.I., S.D. Foster, G. De'ath, J. Stoklosa, and P.K. Dunstan. 2015. Model-based thinking for community ecology. *Plant Ecology 216*(5): 669–682. https://doi.org/10.1007/s11258-014-0366-3 .

Warton, D.I. and F.K.C. Hui. 2017. The central role of mean-variance relationships in the analysis of multivariate abundance data: A response to Roberts (2017). *Methods in Ecology and Evolution 8*(11): 1408–1414. https://doi.org/10.1111/2041-210X.12843 .

Warton, D.I., S.T. Wright, and Y. Wang. 2012. Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution 3*(1): 89–101. https://doi.org/10.1111/j.2041-210X.2011.00127.x .

Wong, R., N. Perkins, J. Monk, M. Prall, A. Lauermann, and N. Barrett. 2026. Decadal changes in California's temperate mesophotic reef invertebrate community through the 2014–2016 northeast Pacific marine heatwave. *Marine Environmental Research* 213: 107644. https://doi.org/10.1016/j.marenvres.2025.107644 .

Yee, T.W. 2025. *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-14. https://doi.org/10.32614/CRAN.package.VGAM.

Young, E.L., K.M. Halanych, D.J. Amon, I. Altamira, J.R. Voight, N.D. Higgs, and C.R. Smith. 2022. Depth and substrate type influence community structure and diversity of wood and whale-bone habitats on the deep NE Pacific margin. *Marine Ecology Progress Series* 687: 23–42. https://doi.org/10.3354/meps14005 .

Zuur, A.F. and E.N. Ieno. 2016. A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution 7*(6): 636–645. https://doi.org/10.1111/2041-210X.12577 .

Zuur, A.F., E.N. Ieno, and C.S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution 1*(1): 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x .

Zuur, A.F., E.N. Ieno, N. Walker, A.A. Saveliev, and G.M. Smith. 2009. *Mixed Effects Models and Extensions in Ecology with R* (1 ed.). Statistics for Biology and Health. New York, NY: Springer. https://doi.org/10.1007/978-0-387-87458-6.