Received: 4 August 2017 | Accepted: 25 October 2017

DOI: 10.1111/2041-210X.12936

REVIEW

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models

Diego Nieto-Lugilde[1,2] | Kaitlin C. Maguire[3] | Jessica L. Blois[3] | John W. Williams[4,5] | Matthew C. Fitzpatrick[1]

[1]Appalachian Laboratory, University of Maryland Center for Environmental Science, Frostburg, MD, USA

[2]Departamento de Botánica, Ecología y Fisiología Vegetal, Universidad de Córdoba, Córdoba, Spain

[3]School of Natural Sciences, University of California, Merced, CA, USA

[4]Center for Climatic Research, University of Wisconsin, Madison, WI, USA

[5]Department of Geography, University of Wisconsin, Madison, WI, USA

Correspondence
Diego Nieto-Lugilde
Email: dnietolugilde@gmail.com

Present address
Kaitlin C. Maguire, USGS Forest and Rangeland Ecosystem Science Center, Boise, ID, USA

Funding information
National Science Foundation, Grant/Award Number: DEB-1257033, DEB-1257164 and DEB-1257508

Handling Editor: Pedro Peres-Neto

Abstract

1. Community-level models (CLMs) consider multiple, co-occurring species in model fitting and are lesser known alternatives to species distribution models (SDMs) for analysing and predicting biodiversity patterns. Community-level models simultaneously model multiple species, including rare species, while reducing overfitting and implicitly considering drivers of co-occurrence. Many CLMs are direct extensions of well-known SDMs and therefore should be familiar to ecologists. However, CLMs remain underutilized, and there have been few tests of their potential benefits and no systematic reviews of their assumptions and implementations. Here, we review this emerging field and provide examples in R to fit common CLMs. Our goal is to introduce CLMs to a broader audience, and discuss their attributes, benefits and limitations relative to SDMs.
2. We review (1) statistical implementations and applications of CLMs, (2) their advantages and limitations, and (3) comparative analyses of CLMs and SDMs. We also suggest directions for future research.
3. We identify seven CLM algorithms with similar data structures and predictive outputs as SDMs that should be most accessible to ecologists familiar with species-level modelling, including five methods that predict assemblage composition and individual species distributions and two methods that model compositional turnover along environmental gradients. Community-level models have been applied to numerous taxa, regions, and spatial scales, and a variety of topics (e.g. studying drivers of community structure or assessing relationships between community composition and functional traits). Studies suggest that the relative benefits of CLMs and SDMs may be case specific, especially in terms of predicting species distributions and community composition. However, CLMs may offer advantages in terms of computational efficiency, modelling rare species, and projecting to no-analog climates. A major shortcoming of CLMs is their reliance on presence–absence community composition data.
4. Studies are needed to assess the relative merits of SDMs and CLMs, and different CLM algorithms, with a focus on three key areas: (1) under which circumstances CLMs improve predictions for rare species, (2) how CLMs perform under different

community compositions (e.g. relative abundance of rare vs. common species), including the extent to which co-occurrence patterns are structured by biotic interactions, and (3) ability to project across time/space.

KEYWORDS
biodiversity, biotic interactions, community assembly, community composition, ecological niche models, large datasets, macroecology, rare species, spatial modelling

1 | INTRODUCTION

The threats to natural systems posed by global change, combined with the growing availability of georeferenced species occurrence data, have generated interest in modelling geographic patterns of biodiversity, from individual species distributions to community composition (Guisan & Thuiller, 2005; le Roux, Pellissier, Wisz, & Luoto, 2014). Species distribution models (SDMs; Elith & Leathwick, 2009), which relate the occurrence of individual species to environmental predictors, remain among the most popular methods for extrapolating biogeographical patterns across space and through time (Guisan & Thuiller, 2005; Maguire, Nieto-Lugilde, Fitzpatrick, Williams, & Blois, 2015). However, spatial modelling methods that simultaneously consider multiple co-occurring species to predict species distributions, species assemblages, and/or macroecological patterns offer an alternative to SDMs and have begun to receive greater attention (Blois, Williams, Fitzpatrick, Ferrier, et al., 2013; Bonthoux, Baselga, & Balent, 2013; D'Amen, Rahbek, Zimmermann, & Guisan, 2017; Ferrier & Guisan, 2006; Fitzpatrick et al., 2011; Maguire et al., 2016). These community-level models (CLMs) may possess practical (e.g. fitting multiple species in a single step) and theoretical benefits (e.g. improved prediction for rare species by borrowing information from common species) over SDMs (Clark, Gelfand, Woodall, & Zhu, 2014; Ferrier & Guisan, 2006; Harris, 2015). However, CLMs remain comparatively unknown and underutilized by ecologists and few studies have systematically explored their merits and limitations.

Community-level models take many forms, but can be divided into three primary groups: (1) methods that simultaneously model the distributions of multiple species using a common set of environmental variables (De'ath, 2002; Leathwick, Elith, & Hastie, 2006; Yee, 2004, 2006; Yee & Hastie, 2003), including variants that use (Bayesian) hierarchical structures to model species associations and possibly biotic interactions (Clark et al., 2014; Harris, 2015; Latimer, Banerjee, Sang, Mosher, & Silander, 2009; Warton et al., 2015), (2) models of macroecological patterns such as species richness (Rahbek & Graves, 2001) or compositional turnover (Ellis, Smith, & Pitcher, 2012; Ferrier, Manion, Elith, & Richardson, 2007), and (3) dynamic models that simulate community assembly and/or stochastic processes (e.g. dispersal; Mokany, Harwood, Williams, & Ferrier, 2012). In this review, we focus on methods within the first two categories (see D'Amen et al., 2017 for additional methods beyond the scope of this review) that meet two basic criteria: (1) models that, like SDMs, are inherently static and correlative (as opposed to dynamic, mechanistic, or stochastic models) and (2) simultaneously combine information for multiple co-occurring species (in the form of a site × species matrix or some derivative thereof) to generate predictions of individual species distributions and/or community-level attributes through space and/or time. We focus on these methods as they share several commonalities with "traditional" SDMs, most notably data structures and predictive outputs (from species distribution to community composition), and therefore are highly accessible to ecologists already familiar with species-level modelling. We do not cover CLMs that predict only species richness and provide no information about species identity or composition (D'Amen et al., 2017).

To remain consistent with published terminology (Baselga & Araújo, 2010; Ferrier & Guisan, 2006; Maguire et al., 2015, 2016; Mokany et al., 2012; Olden, Joy, & Death, 2006) and to distinguish them from SDMs, we refer to the models reviewed here as community-level models. Note, however, that this term also has been used to describe approaches not considered here (e.g. Dynamic Global Vegetation Models; D'Amen et al., 2017) and that terms such as "multiresponse," "multispecies," and "joint species distribution" models also have been used as equivalents (indeed, we refer to a subset of the models reviewed here as "multiresponse" algorithms) (e.g. Elith & Leathwick, 2007; Warton et al., 2015).

Our overall goals are to introduce CLMs to a broader audience and discuss examples of their use in ecology and evolution. Specifically, we aim to review: (1) statistical implementations and applications of several common CLM algorithms; (2) the primary advantages and limitations of the community-level modelling strategy; (3) comparative analyses of CLMs and SDMs and under what circumstances CLMs may have tangible benefits over SDMs and vice versa; and (4) directions for further research. To illustrate each method and help boost the study and use of CLMs, we provide tutorials implementing several algorithms in R (R Core Team, 2017). We begin with an overview of the general strategy underlying CLMs before turning to descriptions of the algorithms themselves and their application to ecological questions.

1.1 | The community-level modelling strategy

In an early review, Ferrier and Guisan (2006) described three strategies for predicting community attributes in space and time, using empirical modelling. The first strategy, "assemble-then-predict," involves fitting models to pre-defined community-types such as vegetation classes,

assuming that assemblages are static combinations of co-occurring species and ignoring individualistic species responses (Clementsonian concept of communities; D'Amen et al., 2017). We do not discuss the assemble-then-predict strategy further given the evidence that community types do not remain constant through time, including the formation of no-analog assemblages (Williams et al., 2013), which argues against the use of this approach for extrapolations through time. The second and most common approach involves first modelling species individually and independently of each other using SDMs to produce a separate predicted distribution map for each species, followed by classification, ordination, or aggregation (e.g. stacking) of these individual predictions to map community-level attributes such as species composition or richness. This "predict-then-assemble" strategy (Ferrier & Guisan, 2006) assumes that species exist in isolation and do not limit one another (i.e. their occurrence probabilities are uncorrelated, except insofar as they are co-determined by similar environmental tolerances; Gleasonian concept of communities; D'Amen et al., 2017) and disregards other drivers (e.g. biotic interactions or stochastic processes). In contrast to modelling and deriving community-level predictions as two distinct steps, CLMs employ a third "assemble-and-predict-together" strategy to simultaneously model multiple co-occurring species within a single integrated process (Ferrier & Guisan, 2006). By treating community structure as an emergent and continuous function of species co-occurrence and environmental variables, the assemble-and-predict-together strategy captures shared climatic requirements of species and (implicitly) any other process driving co-occurrence patterns, including biotic interactions, without discarding information on species-level responses to the environment. Methods employing this strategy are the focus of this review.

2 | OVERVIEW OF CLMs

To identify the most commonly used "assemble-and-predict-together" algorithms, we performed a literature search using Scopus (http://www.scopus.com/search/form.url; see Appendix S1 for search criteria). This search identified 244 articles from which we identified seven of the most common CLM algorithms (Table 1) and several emerging methods (Dunstan, Foster, Hui, & Warton, 2013; Harris, 2015; Hui, Warton, Foster, & Dunstan, 2013; Ovaskainen, Abrego, Halme, & Dunson, 2016; Ovaskainen, Roy, Fox, & Anderson, 2016). Next, we performed targeted searches for each algorithm (Appendix S1). After reviewing the algorithm-specific searches, we retained only ecological publications (algorithms underpinning CLMs are frequently used in other fields such as toxicology, biochemistry, or cell biology). This resulted in a final set of 89 articles that were the basis of the literature review (see Data Accessibility section).

From the seven CLM algorithms identified, five can be described as either direct multiresponse extensions of common SDM algorithms or recently developed hierarchical models that explicitly consider associations between species (sometimes called joint-SDMs or JSDMs, Clark et al., 2014; Warton et al., 2015), including: multivariate regression trees (MRT), multivariate adaptive regression splines (MARS), constrained ordinations (CO), multiresponse artificial neural networks (MANN), and hierarchical Bayesian models (HBM). These methods share the approach of simultaneously relating the distribution of all modelled species to a common set of environmental variables to predict individual species distributions, community composition, and other community-level properties. The remaining two algorithms comprise methods that model variation in community composition along environmental and spatial gradients, using continuous nonlinear functions and predict compositional dissimilarity between assemblages rather than the distributions of individual species. These methods include generalized dissimilarity modelling (GDM) and gradient forests (GF).

Our literature search revealed an increase in the number of CLM-related publications after the early 2000s (Figure 1a). The most commonly used methods in the final subset of 89 articles are GDM and MRT (Figure 1b). In contrast to other community-level extensions of SDMs that have appeared frequently in the literature, HBMs appear to be rapidly gaining popularity despite their more recent development (see publication years below and Appendix S2).

2.1 | Overview of CLM algorithms

This section briefly introduces each method, emphasizing the differences between the single-species (SDM) algorithm and its multiresponse (CLM) counterpart. Additional details, including underlying statistical methods and key references, are provided for each method in Appendix S2. Furthermore, we illustrate each method with an exploratory or diagnostic plot of each model that was fit on an artificial dataset of 10 species (sp1, … sp10) and three environmental variables ($x1$, $x2$, and $x3$). See Appendix S3 for more details and code to fit the seven CLMs in R.

2.1.1 | Multiresponse algorithms

Multivariate regression trees

Multivariate regression tree is a multiresponse extension of classification and regression trees (CARTs; De'ath, 2002). Classification and regression trees model species occurrence by splitting the data into increasingly homogeneous groups and identifying where along environmental gradients (i.e. the predictor variables) the greatest changes in species occurrence/abundance occur. Multivariate regression trees are built in the same way but instead splits between groups are chosen to minimize the species compositional dissimilarity within groups (De'ath, 2002). By doing so, MRTs identify and characterize where along environmental gradients the greatest changes in community composition occur. These break points are represented as nodes of the tree, whereas terminal branches correspond with the most probable community compositions given the environmental conditions (Figure 2).

Multiresponse multivariate adaptive regression splines

Like MRT, MMARS and its single species implementation (multivariate adaptive regression splines; MARS) identify key points (termed knots in MARS parlance) along environmental gradients where the greatest changes in species composition occur. However, MARS and MMARS

TABLE 1 Overview of the seven reviewed community-level models (CLMs), including their species distribution model (SDM) counterpart(s), how they address species associations, their primary output, and key literature references

CLM		SDM counterpart	How species association are addressed	Output	R packages	Main references
Constrained ordinations (CO)	Constrained linear ordination (CLO)	Generalized linear model (GLM)	Indirectly: (1) co-occurrence	Community matrix	VGAM	Yee and Hastie (2003)
	Constrained quadratic ordination (CQO)	GLM with quadratic responses	Indirectly: (1) co-occurrence	Community matrix	VGAM	Yee (2004)
	Constrained additive ordination (CAO)	Generalized additive model (GAM)	Indirectly: (1) co-occurrence	Community matrix	VGAM	Yee (2006)
Generalized dissimilarity model (GDM)			Indirectly: (1) co-occurrence	Dissimilarity	GDM	Ferrier et al. (2007)
Gradient forest (GF)		Random Forest (RF)[a]	Indirectly: (1) co-occurrence	Dissimilarity	GRADIENTFOREST (in R-Forge)	Ellis et al. (2012)
Hierarchical Bayesian models (HBM)		HBM[b]	Directly: (1) cross-species correlation matrix (2) joint distribution	Community matrix	BORAL rosalia spBayes BayesComm HMSC (in R-Forge) Code in article	Latimer et al. (2009) Ovaskainen et al. (2010) Ovaskainen and Soininen (2011) Pollock et al. (2012) Pollock et al. (2014) Clark et al. (2014)
Multiresponse artificial neural network (MANN)		Artificial neural network (ANN)	Indirectly: (1) co-occurrence[c]	Community matrix	NNET neuralnet	Olden (2003)
Multiresponse multivariate adaptive regression splines (MMARS)		Multivariate adaptive regression splines (MARS)	Indirectly: (1) co-occurrence	Community matrix	MDA EARTH	Leathwick et al. (2006)
Multivariate regression trees (MRT)		Classification and regression trees (CARTs)	Indirectly: (1) co-occurrence	Community matrix	MVPART (discontinued)	De'ath (2002)

R packages implementing the different models are also provided; examples on how to fit the packages in **bold** are provided as a vignette in the supporting information (Appendix S3).

[a]Gradient forest relies on multiple species-specific random forest models.

[b]The hierarchical structure is used to model multiple species at once; however, single species models have been fit for comparison purposes.

[c]The default implementation uses only co-occurrence patterns, but an alternative has been described to include links between taxa and incorporate explicitly biotic interactions into the model (see Larsen, Field, & Gilbert, 2012).

connect these knots with basic functions (hinge and linear) describing nonlinear relationships between species occurrence and environmental variables (see Figure 3 and Appendix S3; Leathwick et al., 2006). In MMARS, the knots are identified collectively for all the species, but the basic functions are species specific, thereby allowing the shapes of the fitted functions to differ among species and some flexibility for species-specific responses.

Constrained ordinations

Constrained ordinations encompass three different algorithms (constrained linear ordination -CLO-, constrained quadratic ordination -CQO- and constrained additive ordination -CAO-; Table 1) in which the main drivers of community composition are summarized in the form of latent variables (Yee, 2006; Yee & Hastie, 2003). Latent variables are linear and orthogonal combinations of the original set of predictor variables (Figure 4) constrained by the biological information in the community matrix (site × species). Constrained ordinations relate occurrence data for individual species to these latent variables using either a linear (CLO), quadratic (CQO; Figure 4), or additive effect (CAO; Figure 4). As such, COs are multiresponse counterparts of SDMs (Table 1; Baselga & Araújo, 2010) fitted using generalized linear models (GLMs) with either linear or quadratic terms and generalized additive models (GAM).

Multiresponse artificial neural networks

Artificial neural networks (ANN) are computational algorithms that model and predict by connecting nodes of information (called

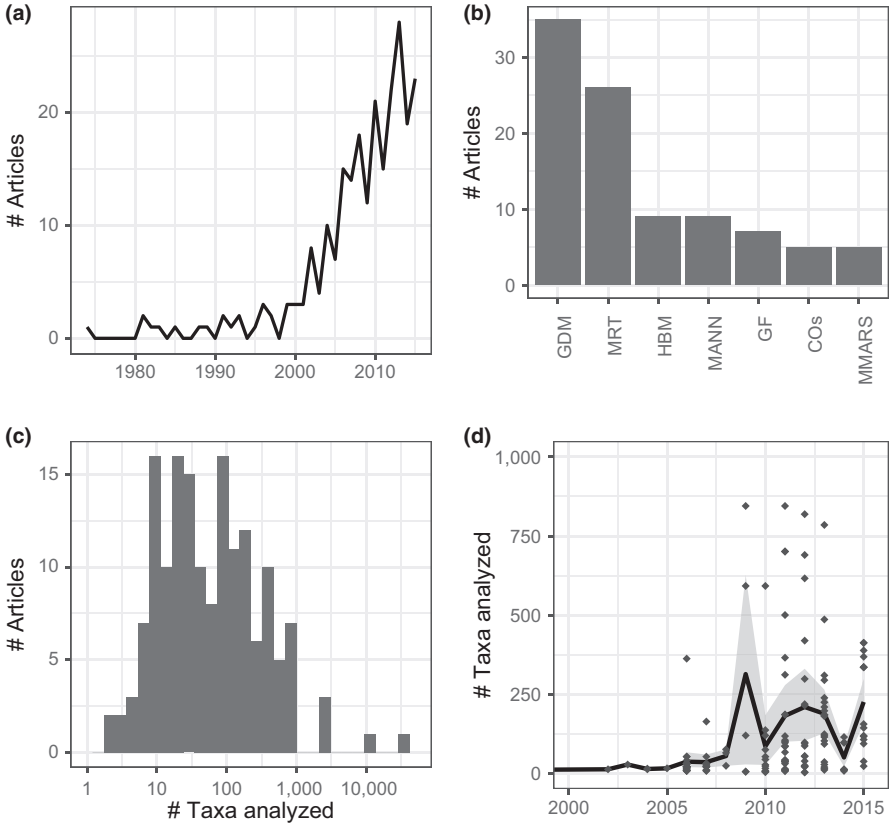

FIGURE 1 Summary plots for publications using community level models (CLMs; data from searches described in Appendix S1): (a) number of published articles through time from 1974 to 2015, (b) number of publications since 1995 by CLM algorithm described in this review, (c) total number of taxa analysed since 1995 using these algorithms; note the logarithmic scale in the *x*-axis, and (d) number of taxa analysed through time with the mean values represented as a black line and 95% confidence intervals as a grey shading. The seven classes of CLM algorithms described here are: constrained ordination (CO), multivariate regression tree (MRT), multiresponse-multivariate adaptive regression spline (MMARS), multiresponse artificial neural network (MANN), hierarchical Bayesian model (HBM), generalized dissimilarity model (GDM), and gradient forest (GF)

neurons) that imitate connections between biological neurons. Neurons are organized into three different layers: input, hidden and output. The input layer has one neuron for each predictor variable in the model and the output layer has a neuron for each response variable. Neurons in the hidden layer allow for interaction effects between predictor variables and the analyst can choose the number of interactions in the network by increasing the number of neurons in this layer as well as adding additional hidden layers. Artificial neural networks-based SDMs have one neuron in the output layer (single species approach). Multiresponse artificial neural networks, however, have multiple neurons (with one neuron for each species; Olden, 2003) and the estimated parameters for the connections between the input and the hidden layers affect all species collectively, while the connections between the hidden and the output layers are species-specific (Figure 5). Like MMARS, this approach considers community-level patterns while allowing for variation among species responses.

Hierarchical Bayesian models of communities

Hierarchical Bayesian models are among the most recent frameworks proposed to perform community-level modelling. They encompass several algorithms that use a hierarchical structure to define latent variables from the predictor variables and introduce structured error terms in the form of a cross-correlation matrix (Figure 6) to address different problems (see Appendix S2 for specifics). For instance, error terms have been used to infer signals of biotic interactions (Ovaskainen, Hottola, & Siitonen, 2010) or improve predictions of rare species (Ovaskainen & Soininen, 2011), while others have focused on the incorporation of functional traits (Pollock, Morris, & Vesk, 2012) or spatial processes (Latimer et al., 2009). Arguably, HBMs are not categorically different from non-Bayesian hierarchical models (e.g. see COs above and Warton et al., 2015). However, because HBMs represent flexible frameworks that use error terms to incorporate factors often missing from other correlative models, they stand apart in important ways from non-Bayesian counterparts and are receiving increasing attention (Figure 1).

2.1.2 | Models of compositional turnover

The remaining two methods model community patterns using continuous functions that describe variation in species composition along environmental gradients.

Generalized dissimilarity modelling

Generalized dissimilarity modelling (Ferrier et al., 2007) combines matrix regression and GLM to relate dissimilarities in species composition between pairs of sites (as a biological distance devoid of species-specific information) to how the sites differ in environmental conditions (environmental distance) and how isolated they are from one another (geographical distance). To accommodate nonlinearity, GDM transforms the predictor variables using I-spline basis functions (Ramsay, 1988). These functions are then evaluated as predictors using non-negative least squares regression. The splines

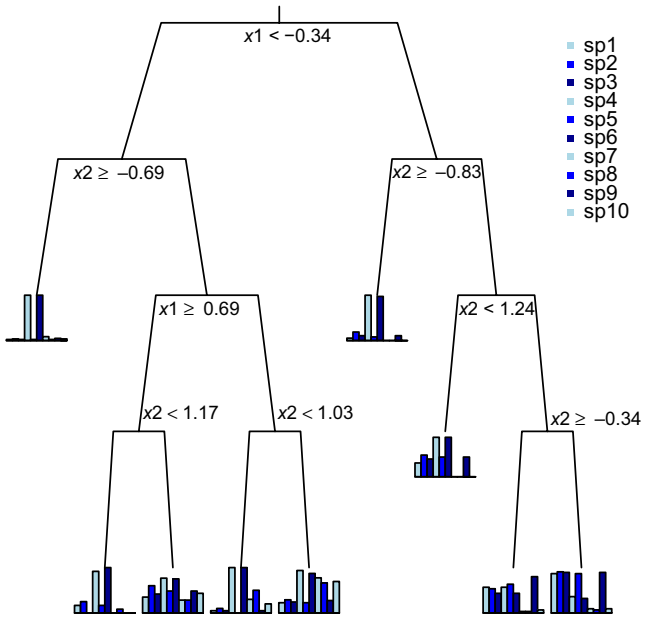

FIGURE 2 Structure of a multivariate regression tree (MRT) fitted using a simulated dataset (code in Appendix S3) of 10 artificial species at 400 sites and constrained by two variables (*x*1 and *x*2), and a third variable (*x*3) that has no effect on any of the species. Each split of the tree shows the split criteria (e.g. *x*1 < −0.34 in the first split), whereas bar charts on the terminal branches represent the typical assemblage for that set of conditions. Assemblages are represented as bar plots with 10 bars (one for each species) representing probability of membership. The MRT correctly identifies *x*1 and *x*2 as the most important drivers of community composition (no node for *x*3) and reveals how species composition changes between branches according to the environmental conditions (nodes)

indicate the overall magnitude of compositional turnover for each gradient and where along each gradient compositional change is most pronounced (Figure 7). Predictions from GDM can be used to visualize spatial variation in community composition, though without information on individual species (Ferrier et al., 2007; see Appendix S3).

Gradient forest

Gradient forest (Ellis et al., 2012) can be considered a community-level extension of Random Forest (RF). Unlike GDM, which uses a distance-based curve-fitting approach to model compositional turnover, GF builds turnover functions directly by partitioning the occurrence data (presence-absence or abundance) of individual species into different bins, with partitions occurring at numerous split values along each environmental variable. These split values are cumulatively summed along each gradient to construct nonlinear turnover functions that provide the same inference as the splines from GDM—namely the magnitude and rate of compositional turnover along each environmental gradient (Figure 7). However, GF builds turnover functions for each species, which are then aggregated across all species to provide an overall community-level turnover function weighted by the magnitude of the individual species responses. Hence, GF differs from GDM because it informs about both community-level and species-specific turnover.

3 | PROS AND CONS OF THE ASSEMBLE-AND-PREDICT-TOGETHER STRATEGY

Several studies have suggested that the assemble-and-predict-together strategy may confer a number of benefits over fitting and aggregating individual SDMs (Clark et al., 2014; Ferrier & Guisan, 2006; Harris, 2015). Many of these potential benefits arise through the leveraging of additional information that can be gleaned from co-occurrence patterns. However, information on co-occurrence does not come for free (e.g. CLMs require community-level data, and, unlike some SDMs, most cannot be fit using presence-only data). Next, we discuss the primary pros and cons of CLMs.

3.1 | Advantages

3.1.1 | Improved predictions of low-prevalence species

Because CLMs fit multiple species simultaneously, information from one species influences the parameter estimates of the others. In

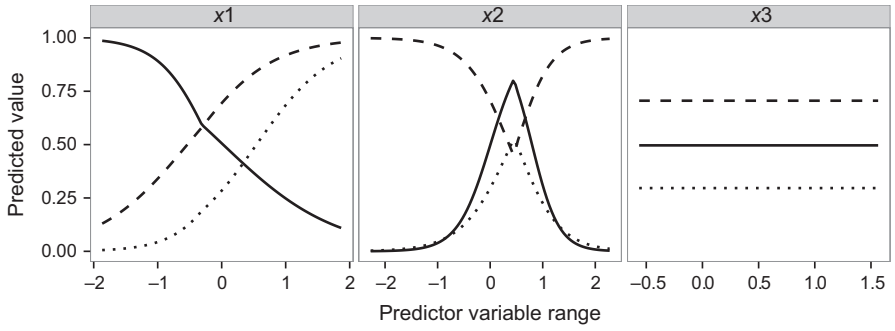

FIGURE 3 Partial response curves from a multiresponse-multivariate adaptive regression splines (MMARS) model fitted using the same simulated dataset described for Figure 2. For simplicity, we only show three species with contrasting responses. Each line type (solid, dashed, and dotted) represents the probability of occurrence along gradients for one of the three species. MMARS identifies the primary role of *x*1 and *x*2 in determining community composition, but not *x*3 (flat responses of three species at their prevalence values), and reveals the differential response of the three species (especially notable in *x*2)

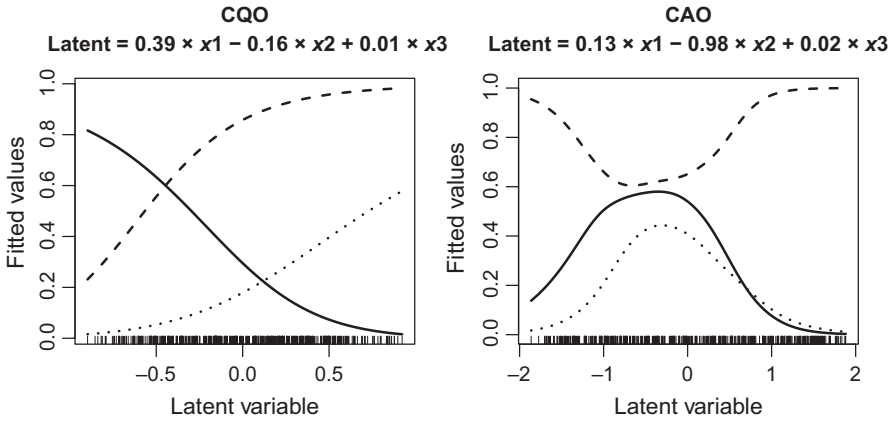

FIGURE 4 Response curves of a constrained quadratic ordination (CQO) model and a constrained additive ordination (CAO) model fitted in the same simulated dataset described for Figure 2. For simplicity, we only show three species (solid, dashed, and dotted lines) with contrasting responses. Modeled species respond to a single latent variable derived from the three predictor variables as a linear combination of the variables (formula in the title), while accounting for their effect on community composition. Black bars at the bottom represent observations of community composition along the latent variables. CQO and CAO estimate quadratic (sigmoid) and additive (humped) responses of species, respectively, to the latent variables. CQO identify *x*1 as the primary contributor to the latent variable, whereas CAO identify *x*2. This explains the similarity in the response curves between these two models and those observed from the MMARS model for *x*1 and *x*2 (Figure 3). Both CQO and CAO identify the small contribution of *x*3 to the latent variable

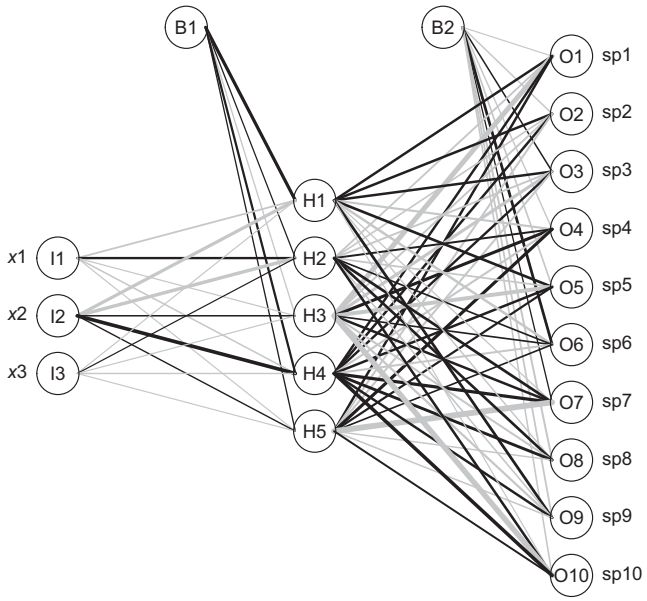

FIGURE 5 Structure of a multiresponse artificial neural network (MANN) model using the same simulated dataset described for Figure 2. The network consists of three layers of neurons. The input layer (left column) has three nodes or neurons (one for each predictor variable; *x*1, *x*2, and *x*3). The hidden layer (middle column) in this example has five neurons, whose neurons receive and combine information from all the neurons in the input layer and serve to model interactions between the three predictor variables. The output layer (right column) has 10 neurons (one for each species). B1 and B2 are bias nodes that apply constant values, similar to intercept terms in a regression model. Connections between nodes represent coefficients in which shading indicates their sign (grey = negative, black = positive) and width represents their strength. Thin connections (low value) between *x*3, compared to *x*1 and *x*2, and neurons in the hidden layer, suggest a small contribution from this variable. The combined effect of environmental variables is defined in the hidden layers that affect each species independently (connections between the hidden layer and the output layer)

some cases, this could improve predictive performance if correlations across species provide the ability to better quantify shared environmental responses than would be possible by fitting models independently for each species. This argument underlie suggestions that CLMs may have advantages over SDMs for modelling rare (or rarely recorded) species, some of which may be beyond the reach of SDMs due to low sample size (Baselga & Araújo, 2009; Chapman & Purse, 2011; but see Ovaskainen & Soininen, 2011; Fitzpatrick et al., 2011). From a practical standpoint, an ability to model rare species is desirable because conservation assessments based on SDMs are inherently limited to relatively well-sampled species and regions (i.e. they exclude most species and the tropics; Feeley & Silman, 2011), which are often of least concern from a degree-of-threat perspective. In addition to improving predictions of some species using information from others, CLMs may also be less sensitive to stochastic responses of individual species and, hence provide more robust parameter estimates than SDMs (Ovaskainen & Soininen, 2011). Furthermore, CLMs approaches have been proposed to deal with the zero-inflated nature of rare species datasets (Clark et al., 2014). However, there is no reason to believe that CLMs will universally lead to improved predictions. For instance, if rare species respond differently to the environment than the broader community, CLMs could estimate biased parameters. This may explain why studies examining whether CLMs or SDMs better predict rare species largely have been equivocal and collectively suggest that performance gains for rare species may be context dependent (see section SDMs vs. CLMs comparisons).

3.1.2 | Biotic interactions and other drivers of community assembly

As mentioned, building community-level predictions by aggregating individual SDMs assumes species exist in isolation and do not limit one

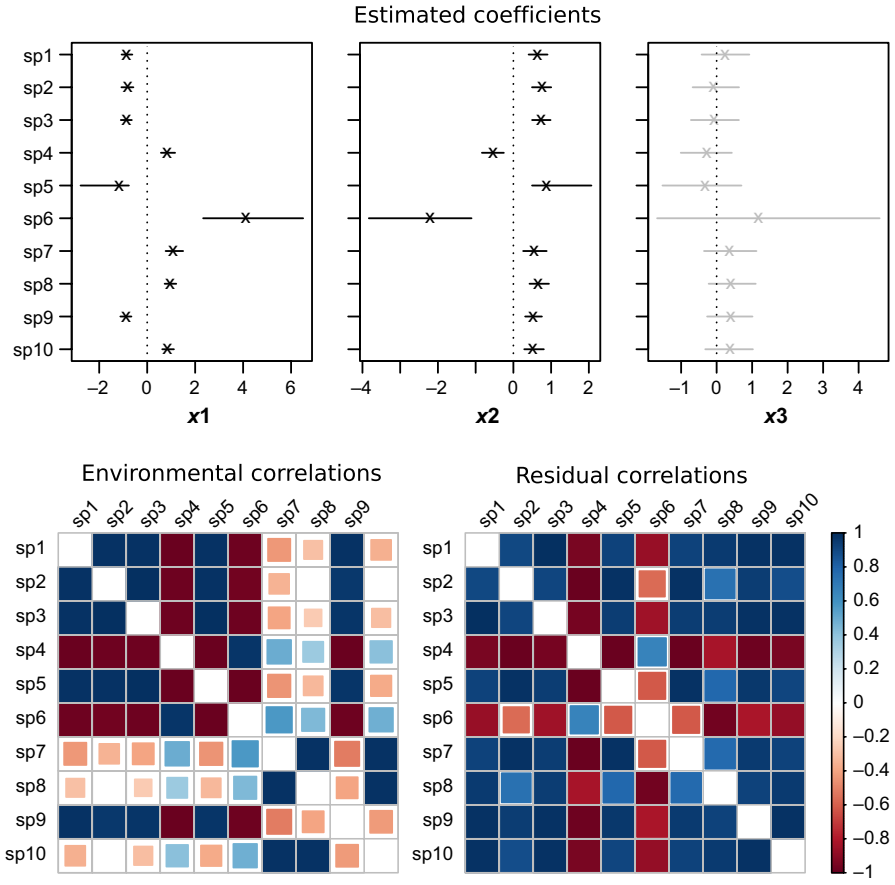

FIGURE 6 Estimated coefficients (top row of panels) for each of the three predictor variables and each species in the same simulated dataset described in Figure 2 from a hierarchical Bayesian model (HBM). Crosses show the median value for each parameter in the Monte Carlo Markov Chain, while lines represent the 0.95 confident intervals. Two cross-correlation matrices (bottom row of panels) estimated from the model. The environmental correlation matrix shows when two species have similar (blue) or contrasting (red) environmental responses, according to the predictor variables. The residual correlation matrix shows when two species co-occur more (blue) or less (red) frequently than expected by random given the predictor variables. Consistent with having no relationship with species occurrences, the confidence intervals for variable *x*3 overlap zero, whereas the confidence intervals for *x*1 and *x*2 do not. In the correlation matrices, species with similar coefficients have similar environmental responses (blue) in the environmental cross-correlation matrix

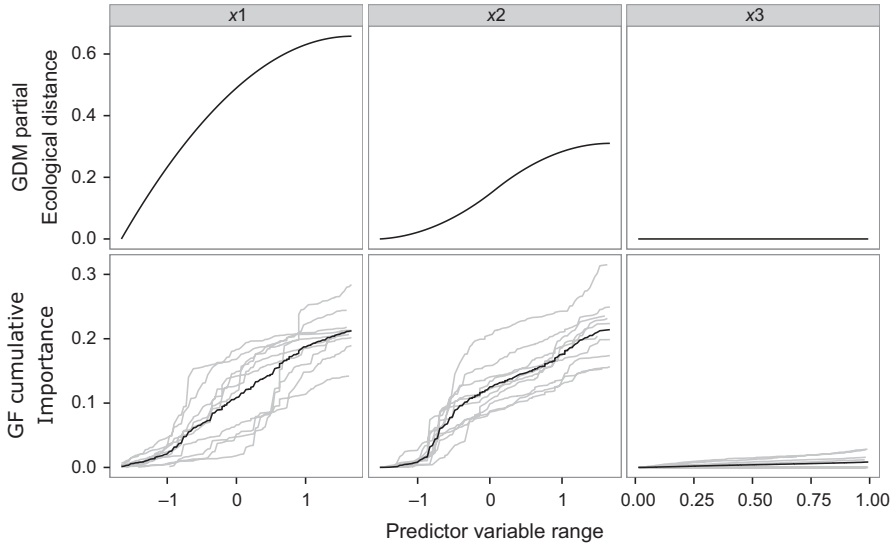

FIGURE 7 Contribution of each predictor variable in explaining compositional turnover in the same simulated dataset described for Figure 2. Generalized dissimilarity model (GDM; upper panels) creates a single I-spline for each variable. The slope of the splines at each point represents the turnover rate at that particular point along the gradient; the maximum height of the spline indicates the importance of that variable in the model. Gradient forest (GF), however, plots the cumulative contribution of each variable to explain species presence-absence along the gradients for each species (grey lines) and then use a weighted aggregation of these species-level curves to calculate the cumulative contribution for the entire community (black lines). GDM indicates *x*1 as the primary driver of community composition (the higher the curve the higher the importance) followed by *x*2, whereas *x*3 has no effect. Comparatively, GF estimate a similar effect of *x*1 and *x*2, while the effect of *x*3 remains negligible

another. However, both the distribution and abundance of species covary, because of (dis)similar environmental requirements (Harris, 2015; Pollock et al., 2014), interactions between species (Ovaskainen et al., 2010), finite limits to shared space (Clark et al., 2014), other drivers (unmeasured environmental variables, evolutionary history, or human impact), or stochastic processes including dispersal limitation. Failure to include biotic interactions is considered a major limitation of SDMs (Kissling & Schleuning, 2014; Wisz et al., 2013) and may explain why stacked SDMs often do not accurately predict community-level properties, such as species richness, community composition, or compositional turnover (Calabrese, Certain, Kraan, & Dormann, 2014; Clark et al., 2014; Mateo, Mokany, & Guisan, 2017). By fitting a model to a co-occurrence matrix, CLMs should be able to implicitly accommodate any process that generates co-occurrence patterns (Baselga & Araújo, 2010). However, none of the CLM algorithms reviewed here explicitly include interactions between organisms or other biotic drivers and therefore inferences regarding the relative importance of different processes in determining patterns of co-occurrence should be made with care, if at all with correlative methods (see Zurell et al., 2016 for alternative approaches with dynamic models). While patterns within the cross-species correlation matrices fitted in some HBMs are often interpreted as signals of biotic interactions, they could also represent other unmeasured drivers. Furthermore, co-occurrence patterns might reflect disequilibrial dynamics or historical legacies (Gill, Williams, Jackson, Lininger, & Robinson, 2009; Svenning, Fløjgaard, & Baselga, 2011) rather than direct responses to contemporary environmental drivers. Pollock et al. (2014) showed how to decompose a cross-correlation matrix into correlation due to similar environmental responses and residual correlation due to factors not included in the model, but isolating true signals of biotic interactions from other possibilities remains a major challenge (Clark et al., 2014; Harris, 2015). In summary, the main difference between SDMs and CLMs is the use of co-occurrence information when fitting CLMs, which may or may not be an indicator of species interactions (Gotelli, Graves, & Rahbek, 2010).

3.1.3 | Fitting large numbers of species and sites

One practical advantage of CLMs is their ability to model multiple species at once, potentially making the task of modelling a large number of species more efficient, at least for some algorithms. Due to data limitations, early applications of SDMs typically considered one or up to several dozen species. Today, however, biodiversity databases contain hundreds of millions of records for millions of species (e.g. GBIF; O Tuama & Braak, 2011). At the same time, methods such as metagenomics and environmental DNA are providing insight into species assemblages at a depth and breadth not previously possible. For example, it is now possible through metagenomics to identify tens of thousands of taxa within microbial communities (Fierer et al., 2012). As the number of modelled entities grows, individually fitting, evaluating and assembling SDMs into communities or macroecological patterns becomes increasingly inefficient. Although CLMs allow simultaneous fitting of multiple species, some algorithms have more demanding computational requirements than SDMs. For instance, COs are slower than their SDMs counterparts or other CLMs (e.g. MRT, MMARS, or MANN; personal observation). Although further developments could improve computation efficiency, other algorithmic boundaries could be limiting (i.e. cross-correlation matrices in HBMs scale with the square of species number). Other CLMs, such as GDM, can efficiently model hyper-diverse communities by analysing compositional turnover between sites, using a pairwise site dissimilarity matrix (sites × sites; Ferrier et al., 2007). For example, Landesman, Nelson, and Fitzpatrick (2014) used GDM to model nearly 40,000 soil microbial taxa. However, GDM may ultimately be limited by the number of sites, since the size of the dissimilarity matrix increases with the square of the number of sites. Although subsampling can reduce this limitation, we are not aware of studies evaluating the effect of such subsampling on model estimation and/or its predictions. So far, CLMs have been fit on a relatively limited number of species (between 10 and a few 100 species; Figure 1c,d) and only a few methods have been tested with a very large number of species (>500; Chapman & Purse, 2011; Ovaskainen & Soininen, 2011; Landesman et al., 2014; Mokany, Thomson, Lynch, Jordan, & Ferrier, 2015; Jones et al., 2016), so the computational advantages remain to be fully demonstrated. Analysis of the topology of interaction networks may help determine the number of species that should be modeled and/or reduce the number of model parameters (Morales-Castilla, Matias, Gravel, & Araújo, 2015).

3.1.4 | Shifting niches and no-analog climates

Climate change studies require flexible methods capable of projecting ecological responses to new times and places, including no-analog climates (Williams et al., 2013). Because SDMs and CLMs assume that species and communities respond in a consistent manner to environmental gradients through time, the predictive ability of both methods would be compromised if biological responses to environmental drivers change between fitting and projecting time periods or places (see Blois, Williams, Fitzpatrick, Ferrier, et al., 2013; Maguire et al., 2016). In addition, the occurrence data upon which SDMs and CLMs are based provide little information to predict how species may respond under novel environments, and no-analog climates in the past are associated with no-analog communities (and likely will be in the future as well). But which method may be more robust to these challenges? Many CLMs use the combined response of multiple species to constrain their responses to environmental gradients, while also allowing individualistic responses at specific points along environmental gradients or to the main components of environmental variation. As such, CLMs could be exploiting some of the strengths of SDMs, while not assuming community types will remain fixed through time (D'Amen et al., 2017; Ferrier & Guisan, 2006) and this ability to balance community and individualistic responses could make them less sensitive to novel climates than SDMs (Ferrier & Guisan, 2006). This contention is supported by a recent multimodel comparison (Maguire et al., 2016), in which CLMs generally outperformed SDMs when extrapolating to dissimilar contexts. Although both SDMs and CLMs performed poorly,

CLMs gain a slight advantage over SDMs for projecting species distributions and community composition under novel climate.

3.2 | Limitations

3.2.1 | Data requirements

Perhaps the single biggest downside of CLMs is their reliance on community-level data (note that CLMs analyse co-occurrence patterns and therefore information on presence-absence of all species at every site is preferred), and the use of presence-only data with pseudo-absences requires further study (but see Ferrier et al., 2007). Presence-only records are the most common type of biodiversity data (Elith & Leathwick, 2007) and SDM-based methods for using such data are well developed (Barbet-Massin, Jiguet, Albert, & Thuiller, 2012; Phillips, Anderson, & Schapire, 2006), which in part may explain the much greater implementation of SDMs to date. Although the use of pseudo-absences to complement presence-only data and their use in community level approaches has been suggested previously (Ferrier & Guisan, 2006; Ferrier et al., 2007) and has been implemented (Elith & Leathwick, 2007; Fitzpatrick et al., 2011), to our knowledge the reliability and accuracy of CLMs fitted in this manner have not been evaluated beyond the level of sensitivity analyses (Fitzpatrick et al., 2011). However, the increasing availability of community-level data, facilitated in part by metagenomics and environmental DNA techniques, may help support wider use of CLMs in the future.

4 | APPLICATIONS OF CLMS TO ECOLOGICAL QUESTIONS

While not as broadly as SDMs, CLMs have been used to address a range of ecological questions. Among the most common uses of CLMs is to identify drivers of community structure (De'ath, 2002; Matabos et al., 2011). For instance, Heino and Alahuhta (2015) used MRTs to determine that community composition of beetles in Europe is more strongly related to climate than land cover or geographical gradients. CLMs have been also used to classify community-types or biogeographical regions (Snelder et al., 2012). Leathwick et al. (2011) used GDM to assess compositional dissimilarity between New Zealand rivers and to classify them. Such classifications can be mapped (Bachraty, Legendre, & Desbruyères, 2009) to inform sampling efforts (Ashcroft et al., 2010) or to prioritize conservation areas (Thomassen et al., 2011). Other applications of CLMs include identifying biodiversity indicators (Claudet, Pelletier, Jouvenel, Bachet, & Galzin, 2006). For instance, Beck, Pfiffner, Ballesteros-Mejia, Blick, and Luka (2013) used GDM to test whether compositional turnover and its relationship with abiotic drivers are interchangeable among ground beetles, rove beetles and spiders. Although SDMs can perform similar analyses, CLMs were preferred given their ease of implementation and because they inherently integrate information across multiple species.

With growing frequency, CLMs are being used to examine the role of biotic interactions in determining species distributions and to incorporate them into predictions of responses to climate change (Kissling & Schleuning, 2014; Maguire et al., 2016; Wisz et al., 2013). Such efforts largely have been limited to HBMs that include a cross-correlation matrix between species as a residual term, which can be interpreted as possible signals of either biotic interactions or unmeasured predictor variables. Following this logic, Ovaskainen et al. (2010) estimated the cross-correlation matrix between species of wood-decaying fungi in Finland to identify a number of previously unknown species associations and generate new hypotheses about direct and indirect species interactions. Nonetheless, we urge caution when inferring species interactions from co-occurrences. We expand on this point below (see section "Biotic interactions and other drivers of community assembly" and the following references for alternative approaches: Morales-Castilla et al., 2015; Harris, 2016; Morueta-Holme et al., 2016).

Like SDMs, CLMs are also increasingly being used to extrapolate biodiversity patterns across space and through time (Maguire et al., 2016) and—in some cases—from one taxonomic group to another. For example, studies have tested for congruency of compositional turnover among taxonomic groups (Thomson et al., 2014) or between paleoecological proxies and their parent taxa (Nieto-Lugilde, Maguire, Blois, Williams, & Fitzpatrick, 2015). Other studies have examined whether compositional turnover–climate relationships remain stable across climate change events (Blois, Williams, Fitzpatrick, Ferrier, et al., 2013) and how different historic processes such as post-glacial migration and climate stability shape compositional patterns in different regions (Fitzpatrick et al., 2013). Common themes of these studies are examining the extent to which one pattern or dataset can serve as a proxy for another and similarities in how different taxa respond to environmental gradients in space and time. Extending this idea to space-for-time substitution, Blois, Williams, Fitzpatrick, Jackson, and Ferrier (2013) fitted GDM models using an extensive dataset of fossil pollen since the Last Glacial Maximum to show that relationships found using spatial patterns can be used to predict changes in community composition through time.

While most studies to date have used CLMs to model species-level patterns and above, the flexibility and structure of some CLMs also makes them highly suitable for modelling biological variation below the species-level, such as genetic or functional trait variation. Several recent studies have capitalized on this capability. For example, Thomassen, Freedman, Brown, Buermann, and Jacobs (2013) used GDM to show that synchronization of reproduction with geographically-distinct seasonal rainfall cycles may contribute reproductive isolation of three African giraffes. Fitzpatrick and Keller (2015) built on these ideas and described how the turnover functions from GDM and GF can be used with genomic data to identify the primary environmental gradients driving genomic variation, understand and map current and future genomic patterns, and infer which regions of the genome may be under selection. GDM also has been used to examine trait–climate relationships (Baldassarre, Thomassen, Karubian, & Webster, 2013) and to model phylogenetic beta diversity

(Rosauer et al., 2014). Methods such as MRT have been used for similar purposes, such as analysing geographic patterns of genetic variation in forest trees (Hamann, Gylander, & Chen, 2011). Pollock et al. (2012) built a HBM to incorporate the relationship between functional traits and climate, providing key insight into how traits modulate species responses to environmental gradients. Finally, Thomassen et al. (2011) combined functional traits and genetic data in the same GDM analysis to study their relationships with climate.

5 | SDMS VS. CLMS COMPARISONS: WHICH HAS A HIGHER PREDICTIVE SKILL?

While CLMs might be better aligned with ecological theory and may help to overcome some practical challenges, the extent to which these benefits translate into improved predictions remains unclear. Here, we compare the predictive skill of CLMs and SDMs for the following response variables and operational requirements: species distributions, community composition and/or macroecological patterns (e.g. species richness), low-prevalence species, and variable selection.

5.1 | Predicting species distributions

Elith and Leathwick (2007) found that a CLM (MMARS) outperformed its SDM counterpart (MARS) when predicting species distributions using presence-only occurrence data. This result corroborates previous findings from a multi-model comparison (Elith et al., 2006), where MMARS was among the modelling techniques scoring the highest AUCs among several SDMs and CLMs. A more recent study comparing an SDM (GLM) and its CLM counterpart (CQO), which was based on two independent bird surveys in France recorded 25 years apart, concluded that both methods have similar and limited ability to predict species distributions (Bonthoux et al., 2013). In contrast, Leathwick et al. (2006) and Heinänen and Von Numers (2009) found that a CLM (MMARS) did not outperform two SDMs (MARS and GAM), and Baselga and Araújo (2009) found that an SDM (GLM) predicted species distributions better than its CLM counterpart (CQO) when predicting the distribution of 158 native tree species across Europe. Similarly, Chapman and Purse (2011) found that two SDMs (CARTs and ANN) performed slightly better than their CLM counterparts when modelling the distributions of 705 native plant species in Great Britain. These conflicting findings have been attributed to differences in data quality and the prevalence of rare taxa in the dataset (Heinänen & Von Numers, 2009; Leathwick et al., 2006); and Elith et al. (2006) argued that CLMs are preferred when modelling many rare species or when using presence-only data with pseudo-absences, but SDMs are preferred when modelling mostly common species or when using high-quality presence-absence data. Chapman and Purse (2011) argued that the benefits of CLMs (such as modelling many species in a single integrated process and resolving more realistic response curves) outweighs their slight underperformance compared to SDMs.

5.2 | Predicting community composition and/or species richness

Olden et al. (2006) compared the abilities of a CLM (MANN) versus stacked SDMs (ANN) and found that the CLM predicts community composition better. Other studies have found that neither SDMs nor their CLM counterpart could reliably predict composition. Bonthoux et al. (2013) showed that both a CLM (CQO) and its SDM counterpart (GLM) performed poorly when predicting community composition, especially under land cover change caused by individualistic species responses or because the CLM failed to accurately predict species richness and community composition. Similarly, Baselga and Araújo (2010) found that neither an SDM (GLM) nor its CLM counterpart (CQO) correctly estimate tree assemblages across Europe, which they attributed to a failure capturing the processes that drive community assembly. In contrast, Chapman and Purse (2011) found that two SDMs (CARTs and ANNs) consistently outperformed their CLM counterparts when modelling community composition and species richness.

5.3 | Predictions for low-prevalence species

Few studies have systematically evaluated how SDMs and CLMs compare in terms of their relative abilities to predict low-prevalence species. In their study, Baselga and Araújo (2009) found that a CLM (CQO) usually outperformed its SDM counterpart (GLM) when predicting narrowly-distributed species and the converse was true for widely distributed species. However, Chapman and Purse (2011) found that CLMs (MANN and MRTs) predicted rare species less accurately than their SDM counterparts.

5.4 | Improving variable selection

In the only study addressing this issue, Madon, Warton, and Araújo (2013) compared the ability of model and variable selection in GLM using the Akaike Information Criteria (AIC) calculated individually for each species (species-level approach) or averaged across all the species (community-level approach). The results were inconclusive, but suggested an advantage of the community-level approach to model rare species, as well as to solve more difficult problems of model selection by detecting noisy variables more efficiently than at the species level approach. Elith and Leathwick (2007) suggested that the CLM version of MARS (MMARS) outperforms its SDM counterpart (MARS) by improving the stability of variable selection for rare taxa, but they did not explicitly test this hypothesis.

5.5 | SDM vs. CLM summary

Existing studies comparing SDMs and CLMs do not provide clear evidence that one framework consistently outperforms the other. This is true regardless of the metric considered, though questions regarding low-prevalence species and variable selection require more study. However, while several studies have compared SDMs and CLMs, few

have compared multiple CLMs with their direct SDM counterpart or have evaluated predictions from CLMs against observed changes in species assemblages. Hence, an important, but uncontrolled source of variation in most studies is the algorithms themselves. Furthermore, comparative studies have considered a small number of taxonomical groups (mainly plants and birds), spatial scales (namely regional), and regions (mostly Europe and Oceania), and are usually constrained by the minimum number of occurrences necessary to fit SDMs, which limits comparisons to only those taxa that can be modeled using SDMs.

Maguire et al. (2016) performed one of the only comprehensive head-to-head comparisons of CLMs and their direct SDMs counterparts. Five models of each type were fitted and evaluated across numerous time periods using observed changes in fossil pollen assemblages for North America since the Last Glacial Maximum. They found that both SDMs and CLMs performed poorly when projected to times that were temporally distant and climatically dissimilar from those in which models were fit. However, CLMs slightly outperformed SDMs when extrapolating to dissimilar contexts, especially when models were fit with sparse calibration datasets (few sites), but not for rare species (few presences). Overall, their findings suggest that CLMs may have a slight advantage over SDMs for projecting through time, though novel climates presented a major challenge for both methods. More SDM-CLM comparisons are needed, but taken together, studies suggest that the primary benefits of CLMs may be in their efficiency, with generally small improvements in predictive performance over SDMs when fitting sparse datasets and projecting across time.

6 | SUMMARY AND RECOMMENDATIONS FOR FUTURE DIRECTIONS

Community-level models offer complimentary methods to SDMs and possess a number of potential benefits and limitations. Like SDMs, numerous CLMs are available to analyse and map biodiversity and many are implemented in R (see Table 1 and a tutorial in Appendix S3). Despite persistent questions about the relative merits and performance of CLMs, existing algorithms are growing in popularity and are being applied to both existing and new ecological questions. CLMs are also being applied to ecological levels of organization beyond species and communities to consider genetic variation, functional traits, and/or spatial processes.

Regardless of which method is superior in terms of performance, there are certain circumstances when CLMs might always be a better choice than SDMs (e.g. if the goal is to predict compositional patterns of hundreds or thousands of species and there is little interest in individual species per se). However, outside of broad generalities, existing studies provide little guidance regarding under which circumstances SDMs may be preferred over CLMs (or vice versa) or regarding the relative strengths of different CLM algorithms.

Next, we summarize what we see as the most promising lines of research: (1) defining the circumstances under which CLMs improve predictions for rare or low-prevalence species, (2) quantifying how CLMs perform under different types of community structure (e.g. different proportions of low and high prevalence species or under different trophic networks), (3) assessing the extent to which species interactions are captured by the models (especially in HBMs using cross-correlation matrices), and (4) measuring performance when projecting across time and space using independent spatial and temporal datasets.

These research topics also are relevant to other modelling strategies (e.g. stacking SDMs) or among CLMs algorithms. Intermodel comparison may help to disentangle the relative cost and benefits of CLMs (e.g. do CLMs predict rare species better than SDMs and under which circumstances? Which CLMs algorithms perform best?). Furthermore, comparing CLMs with other modelling approaches can shed light onto important ecological questions. For instance, comparing predictions from CLMs based on climate variables with predictions from simulation models of stochastic processes (e.g. migration and/or local extinctions) may clarify the relative importance of such drivers in determining community composition. Similarly, comparing CLMs with mechanistic or functional-trait-based models may help to identify alternative drivers (e.g. known interactions or environmental filtering). These comparisons should be tested with truly independent datasets (most of the current comparisons have been evaluated within the same study area or time period, but Maguire et al., 2016). A promising approach in this regard may include the use of virtual species (Leroy, Meynard, Bellard, & Courchamp, 2015) and/or virtual communities (Blanchet, 2015; Dunstan, Foster, & Darnell, 2016), which would allow the drivers and characteristics of species distributions and communities (e.g. biotic interactions vs. environmental requirements) to be known and controlled, without biasing the results to one approach over another (e.g. CLMs vs. SDMs).

The flexibility to incorporate additional data (e.g. functional traits, spatial context, etc.) into HBMs and the possibility to apply GDM and GF to levels of organization beyond species and communities has fuelled most of the recent developments and will likely continue to do so into the near future. We also foresee new statistical developments to improve computation speed and circumvent some current limitations (e.g. increasing number of species and sites). Due to the general interest in CLMs to model rare species and the importance of these predictions for conservation, quantifying uncertainty should be one of the main developments in any CLM algorithm (note that those are explicitly addressed in Bayesian approaches). We see potential (especially in HBMs) to incorporate further constraints from ecological theory: incorporating detectability of species (such as in occupancy models), using network information to define priors or reduce parameter space (e.g. trophic network, Morales-Castilla et al., 2015; or species associations, Morueta-Holme et al., 2016), adding saturation constraints (Mateo et al., 2017), and/or incorporating stochasticity (e.g. from complementary simulation modules). Finally, there is a need to develop and test CLMs algorithms for use with presence-only data, which would enable the use of massive amounts of data from online repositories.

Taken together, there remain numerous research frontiers for the development and application of CLMs. Given recent and pending

developments and the increasing availability of suitable datasets, we anticipate community-level modelling will continue to grow. Hence, CLMs should be considered a useful addition to the ecologist's quantitative toolbox.

ACKNOWLEDGEMENTS

We thank Benjamin Blonder, three anonymous reviewers and subject editor for helpful discussion and encouragement. This research was supported by NSF DEB-1257164 to M.C.F., DEB-1257033 to J.L.B., and DEB-1257508 to J.W.W. This is UMCES-Appalachian Laboratory Scientific Contribution No. 5423.

CONFLICTS OF INTEREST

The authors declare no conflict of interest

AUTHORS' CONTRIBUTIONS

D.N.-L. and M.C.F. conceived the idea and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

Data compiled from the literature search and the Rmarkdown code to reproduce the tutorial are deposited in the Dryad Digital Repository http://datadryad.org/resource/doi:10.5061/dryad.99dc0 (Nieto-Lugilde et al., 2017).

ORCID


Diego Nieto-Lugilde http://orcid.org/0000-0003-4135-2881

REFERENCES

Ashcroft, M. B., Gollan, J. R., Faith, D. P., Carter, G. A., Lassau, S. A., Ginn, S. G., … Cassis, G. (2010). Using generalised dissimilarity models and many small samples to improve the efficiency of regional and landscape scale invertebrate sampling. *Ecological Informatics*, *5*, 124–132. https://doi.org/10.1016/j.ecoinf.2009.12.002

Bachraty, C., Legendre, P., & Desbruyères, D. (2009). Biogeographic relationships among deep-sea hydrothermal vent faunas at global scale. *Deep-Sea Research Part I: Oceanographic Research Papers*, *56*, 1371–1378. https://doi.org/10.1016/j.dsr.2009.01.009

Baldassarre, D. T., Thomassen, H. A., Karubian, J., & Webster, M. S. (2013). The role of ecological variation in driving divergence of sexual and non-sexual traits in the red-backed fairy-wren (*Malurus melanocephalus*). *BMC Evolutionary Biology*, *13*, 75. https://doi.org/10.1186/1471-2148-13-75

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Baselga, A., & Araújo, M. B. (2009). Individualistic vs. community modelling of species distributions under climate change. *Ecography*, *32*, 55–65. https://doi.org/10.1111/j.1600-0587.2009.05856.x

Baselga, A., & Araújo, M. B. (2010). Do community-level models describe community variation effectively? *Journal of Biogeography*, *37*, 1842–1850. https://doi.org/10.1111/j.1365-2699.2010.02341.x

Beck, J., Pfiffner, L., Ballesteros-Mejia, L., Blick, T., & Luka, H. (2013). Revisiting the indicator problem: Can three epigean arthropod taxa inform about each other's biodiversity? *Diversity and Distributions*, *19*, 688–699. https://doi.org/10.1111/ddi.12021

Blanchet, F. G. (2015). HMSC: Hierarchical Modelling of Species Community.

Blois, J. L., Williams, J. W., Fitzpatrick, M. C., Ferrier, S., Veloz, S. D., He, F., … Otto-Bliesner, B. (2013). Modeling the climatic drivers of spatial patterns in vegetation composition since the Last Glacial Maximum. *Ecography*, *36*, 460–473. https://doi.org/10.1111/j.1600-0587.2012.07852.x

Blois, J. L., Williams, J. W., Fitzpatrick, M. C., Jackson, S. T., & Ferrier, S. (2013). Space can substitute for time in predicting climate-change effects on biodiversity. *Proceedings of the National Academy of Sciences*, *110*, 9374–9379. https://doi.org/10.1073/pnas.1220228110

Bonthoux, S., Baselga, A., & Balent, G. (2013). Assessing community-level and single-species models predictions of species distributions and assemblage composition after 25 years of land cover change. *PLoS ONE*, *8*, e54179. https://doi.org/10.1371/journal.pone.0054179

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, *23*, 99–112. https://doi.org/10.1111/geb.12102

Chapman, D. S., & Purse, B. V. (2011). Community versus single-species distribution models for British plants. *Journal of Biogeography*, *38*, 1524–1535. https://doi.org/10.1111/j.1365-2699.2011.02517.x

Clark, J. S., Gelfand, A. E., Woodall, C. W., & Zhu, K. (2014). More than the sum of the parts: Forest climate response from joint species distribution models. *Ecological Applications*, *24*, 990–999. https://doi.org/10.1890/13-1015.1

Claudet, J., Pelletier, D., Jouvenel, J.-Y., Bachet, F., & Galzin, R. (2006). Assessing the effects of marine protected area (MPA) on a reef fish assemblage in a northwestern Mediterranean marine reserve: Identifying community-based indicators. *Biological Conservation*, *130*, 349–369. https://doi.org/10.1016/j.biocon.2005.12.030

D'Amen, M., Rahbek, C., Zimmermann, N. E., & Guisan, A. (2017). Spatial predictions at the community level: From current approaches to future frameworks. *Biological Reviews*, *92*, 169–187. https://doi.org/10.1111/brv.12222

De'ath, G. (2002). Multivariate regression trees: A new technique for modelling species-environment relationships. *Ecology*, *83*, 1105–1117. https://doi.org/10.1890/0012-9658(2002)083[1105:MRTANT]2.0.CO;2

Dunstan, P. K., Foster, S. D., & Darnell, R. (2016). SpeciesMix: Fit mixtures of archetype species. R package version 0.3.4.

Dunstan, P. K., Foster, S. D., Hui, F. K. C., & Warton, D. I. (2013). Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of Agricultural, Biological, and Environmental Statistics*, *18*, 357–375. https://doi.org/10.1007/s13253-013-0146-x

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., & Leathwick, J. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, *13*, 265–275. https://doi.org/10.1111/j.1472-4642.2007.00340.x

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Ellis, N., Smith, S. J., & Pitcher, C. R. (2012). Gradient forests: Calculating importance gradients on physical predictors. *Ecology*, *93*, 156–168. https://doi.org/10.1890/11-0252.1

Feeley, K. J., & Silman, M. R. (2011). Keep collecting: Accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions*, *17*, 1132–1140. https://doi.org/10.1111/j.1472-4642.2011.00813.x

Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, *43*, 393–404. https://doi.org/10.1111/j.1365-2664.2006.01149.x

Ferrier, S., Manion, G., Elith, J., & Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, *13*, 252–264. https://doi.org/10.1111/j.1472-4642.2007.00341.x

Fierer, N., Lauber, C. L., Ramirez, K. S., Zaneveld, J., Bradford, M. A., & Knight, R. (2012). Comparative metagenomic, phylogenetic and physiological analyses of soil microbial communities across nitrogen gradients. *The ISME Journal*, *6*, 1007–1017. https://doi.org/10.1038/ismej.2011.159

Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, *18*, 1–16. https://doi.org/10.1111/ele.12376

Fitzpatrick, M. C., Sanders, N. J., Ferrier, S., Longino, J. T., Weiser, M. D., & Dunn, R. (2011). Forecasting the future of biodiversity: A test of single- and multi-species models for ants in North America. *Ecography*, *34*, 836–847. https://doi.org/10.1111/j.1600-0587.2011.06653.x

Fitzpatrick, M. C., Sanders, N. J., Normand, S., Svenning, J.-C., Ferrier, S., Gove, A. D., & Dunn, R. R. (2013). Environmental and historical imprints on beta diversity: Insights from variation in rates of species turnover along gradients. *Proceedings of the Royal Society B: Biological Sciences*, *280*, 20131201. https://doi.org/10.1098/rspb.2013.1201

Gill, J. L., Williams, J. W., Jackson, S. T., Lininger, K. B., & Robinson, G. S. (2009). Pleistocene megafaunal collapse, novel plant communities, and enhanced fire regimes in North America. *Science*, *326*, 1100–1103. https://doi.org/10.1126/science.1179504

Gotelli, N. J., Graves, G. R., & Rahbek, C. (2010). Macroecological signals of species interactions in the Danish avifauna. *Proceedings of the National Academy of Sciences*, *107*, 5030–5035. https://doi.org/10.1073/pnas.0914089107

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Hamann, A., Gylander, T., & Chen, P. (2011). Developing seed zones and transfer guidelines with multivariate regression trees. *Tree Genetics & Genomes*, *7*, 399–408. https://doi.org/10.1007/s11295-010-0341-7

Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, *6*, 465–476. https://doi.org/10.1111/2041-210X.12332

Harris, D. J. (2016). Inferring species interactions from co-occurrence data with Markov networks. *Ecology*, *97*, 3308–3314. https://doi.org/10.1002/ecy.1605

Heinänen, S., & Von Numers, M. (2009). Modelling species distribution in complex environments: An evaluation of predictive ability and reliability in five shorebird species. *Diversity and Distributions*, *15*, 266–279. https://doi.org/10.1111/j.1472-4642.2008.00532.x

Heino, J., & Alahuhta, J. (2015). Elements of regional beetle faunas: Faunal variation and compositional breakpoints along climate, land cover and geographical gradients. *Journal of Animal Ecology*, *84*, 427–441. https://doi.org/10.1111/1365-2656.12287

Hui, F. K. C., Warton, D. I., Foster, S. D., & Dunstan, P. K. (2013). To mix or not to mix: Comparing the predictive performance of mixture models vs. separate species distribution models. *Ecology*, *94*, 1913–1919. https://doi.org/10.1890/12-1322.1

Jones, M. M., Gibson, N., Yates, C., Ferrier, S., Mokany, K., Williams, K. J., … Svenning, J.-C. (2016). Underestimated effects of climate on plant species turnover in the Southwest Australian Floristic Region. *Journal of Biogeography*, *43*, 289–300. https://doi.org/10.1111/jbi.12628

Kissling, W. D., & Schleuning, M. (2014). Multispecies interactions across trophic levels at macroscales: Retrospective and future directions. *Ecography*, *38*, 346–357. https://doi.org/10.1111/ecog.00819

Landesman, W. J., Nelson, D. M., & Fitzpatrick, M. C. (2014). Soil properties and tree species drive ß-diversity of soil bacterial communities. *Soil Biology and Biochemistry*, *76*, 201–209. https://doi.org/10.1016/j.soilbio.2014.05.025

Larsen, P. E., Field, D., & Gilbert, J. A. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods*, *9*, 621–625. https://doi.org/10.1038/nmeth.1975

Latimer, A. M., Banerjee, S., Sang, H. Jr, Mosher, E. S., & Silander, J. A. Jr (2009). Hierarchical models facilitate spatial analysis of large data sets: A case study on invasive plant species in the northeastern United States. *Ecology Letters*, *12*, 144–154. https://doi.org/10.1111/j.1461-0248.2008.01270.x

le Roux, P. C., Pellissier, L., Wisz, M. S., & Luoto, M. (2014). Incorporating dominant species as proxies for biotic interactions strengthens plant community models. *Journal of Ecology*, *102*, 765–775. https://doi.org/10.1111/1365-2745.12239

Leathwick, J. R., Elith, J., & Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling*, *199*, 188–196. https://doi.org/10.1016/j.ecolmodel.2006.05.022

Leathwick, J. R., Snelder, T., Chadderton, W. L., Elith, J., Julian, K., & Ferrier, S. (2011). Use of generalised dissimilarity modelling to improve the biological discrimination of river and stream classifications. *Freshwater Biology*, *56*, 21–38. https://doi.org/10.1111/j.1365-2427.2010.02414.x

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2015). virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*, 599–607. https://doi.org/10.1111/ecog.01388

Madon, B., Warton, D. I., & Araújo, M. B. (2013). Community-level vs. species-specific approaches to model selection. *Ecography*, *36*, 1291–1298. https://doi.org/10.1111/j.1600-0587.2013.00127.x

Maguire, K. C., Nieto-Lugilde, D., Blois, J. L., Fitzpatrick, M. C., Williams, J. W., Ferrier, S., & Lorenz, D. J. (2016). Controlled comparison of species- and community-level models across novel climates and communities. *Proceedings of the Royal Society B*, *283*, 20152817. https://doi.org/10.1098/rspb.2015.2817

Maguire, K. C., Nieto-Lugilde, D., Fitzpatrick, M. C., Williams, J. W., & Blois, J. L. (2015). Modeling species and community responses to past, present, and future episodes of climatic and ecological change. *Annual Review of Ecology, Evolution, and Systematics*, *46*, 343–368. https://doi.org/10.1146/annurev-ecolsys-112414-054441

Matabos, M., Plouviez, S., Hourdez, S., Desbruyères, D., Legendre, P., Warén, A., … Thiébaut, E. (2011). Faunal changes and geographic crypticism indicate the occurrence of a biogeographic transition zone along the southern East Pacific Rise. *Journal of Biogeography*, *38*, 575–594. https://doi.org/10.1111/j.1365-2699.2010.02418.x

Mateo, R. G., Mokany, K., & Guisan, A. (2017). Biodiversity models: What if unsaturation is the rule? *Trends in Ecology & Evolution*, *32*, 556–566. https://doi.org/10.1016/j.tree.2017.05.003

Mokany, K., Harwood, T. D., Williams, K. J., & Ferrier, S. (2012). Dynamic macroecology and the future for biodiversity. *Global Change Biology*, *18*, 3149–3159. https://doi.org/10.1111/j.1365-2486.2012.02760.x

Mokany, K., Thomson, J. J., Lynch, A. J. J., Jordan, G. J., & Ferrier, S. (2015). Linking changes in community composition and function under climate change. *Ecological Applications*, *25*, 2132–2141. https://doi.org/10.1890/14-2384.1

Morales-Castilla, I., Matias, M. G., Gravel, D., & Araújo, M. B. (2015). Inferring biotic interactions from proxies. *Trends in Ecology & Evolution*, *30*, 347–356. https://doi.org/10.1016/j.tree.2015.03.014

Morueta-Holme, N., Blonder, B., Sandel, B., McGill, B. J., Peet, R. K., Ott, J. E., … Svenning, J.-C. (2016). A network approach for inferring species associations from co-occurrence data. *Ecography*, *39*, 1139–1150. https://doi.org/10.1111/ecog.01892

Nieto-Lugilde, D., Maguire, K. C., Blois, J. L., Williams, J. W., & Fitzpatrick, M. C. (2015). Close agreement between pollen-based and forest inventory-based models of vegetation turnover. *Global Ecology and Biogeography*, *24*, 905–916. https://doi.org/10.1111/geb.12300

Nieto-Lugilde, D., Maguire, K. C., Blois, J. L., Williams, J. W., & Fitzpatrick, M. C. (2017). Data from: Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.99dc0

O Tuama, E., & Braak, K. (2011). GBIF Metadata Profile, Reference Guide.

Olden, J. D. (2003). A species-specific approach to modeling biological communities and its potential for conservation. *Conservation Biology*, *17*, 854–863. https://doi.org/10.1046/j.1523-1739.2003.01280.x

Olden, J. D., Joy, M. K., & Death, R. G. (2006). Rediscovering the species in community-wide predictive modeling. *Ecological Applications*, *16*, 1449–1460. https://doi.org/10.1890/1051-0761(2006)016[1449:RTSICP]2.0.CO;2

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

Ovaskainen, O., Hottola, J., & Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, *91*, 2514–2521. https://doi.org/10.1890/10-0173.1

Ovaskainen, O., Roy, D. B., Fox, R., & Anderson, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, *7*, 428–436. https://doi.org/10.1111/2041-210X.12502

Ovaskainen, O., & Soininen, J. (2011). Making more out of sparse data: Hierarchical modeling of species communities. *Ecology*, *92*, 289–295. https://doi.org/10.1890/10-1251.1

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Pollock, L. J., Morris, W. K., & Vesk, P. A. (2012). The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, *35*, 716–725. https://doi.org/10.1111/j.1600-0587.2011.07085.x

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., … McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, *5*, 397–406. https://doi.org/10.1111/2041-210X.12180

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rahbek, C., & Graves, G. R. (2001). Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences*, *98*, 4534–4539. https://doi.org/10.1073/pnas.071034898

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, *3*, 425–441. https://doi.org/10.1214/ss/1177012761

Rosauer, D. F., Ferrier, S., Williams, K. J., Manion, G., Keogh, J. S., & Laffan, S. W. (2014). Phylogenetic generalised dissimilarity modelling: A new approach to analysing and predicting spatial turnover in the phylogenetic composition of communities. *Ecography*, *37*, 21–32. https://doi.org/10.1111/j.1600-0587.2013.00466.x

Snelder, T., Ortiz, J. B., Booker, D., Lamouroux, N., Pella, H., & Shankar, U. (2012). Can bottom-up procedures improve the performance of stream classifications? *Aquatic Sciences*, *74*, 45–59. https://doi.org/10.1007/s00027-011-0194-7

Svenning, J.-C., Fløjgaard, C., & Baselga, A. (2011). Climate, history and neutrality as drivers of mammal beta diversity in Europe: Insights from multiscale deconstruction. *Journal of Animal Ecology*, *80*, 393–402. https://doi.org/10.1111/j.1365-2656.2010.01771.x

Thomassen, H. A., Freedman, A. H., Brown, D. M., Buermann, W., & Jacobs, D. K. (2013). Regional differences in seasonal timing of rainfall discriminate between genetically distinct East African Giraffe Taxa. *PLoS ONE*, *8*, e77191. https://doi.org/10.1371/journal.pone.0077191

Thomassen, H. A., Fuller, T., Buermann, W., Milá, B., Kieswetter, C. M., Jarrín-V., P., … Wang, O. (2011). Mapping evolutionary process: A multi-taxa approach to conservation prioritization. *Evolutionary Applications*, *4*, 397–413. https://doi.org/10.1111/j.1752-4571.2010.00172.x

Thomson, R. J., Hill, N. A., Leaper, R., Ellis, N., Pitcher, C. R., Barrett, N. S., & J Edgar, G. (2014). Congruence in demersal fish, macroinvertebrate, and macroalgal community turnover on shallow temperate reefs. *Ecological Applications*, *24*, 287–299. https://doi.org/10.1890/12-1549.1

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, *30*, 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Williams, J. W., Blois, J. L., Gill, J. L., Gonzales, L. M., Grimm, E. C., Ordonez, A., … Veloz, S. D. (2013). Model systems for a no-analog future: Species associations and climates during the last deglaciation. *Annals of the New York Academy of Sciences*, *1297*, 29–43. https://doi.org/10.1111/nyas.12226.

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C., … Svenning, J.-C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological Reviews*, *88*, 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x

Yee, T. W. (2004). A new technique for maximum-likelihood canonical gaussian ordination. *Ecological Monographs*, *74*, 685–701. https://doi.org/10.1890/03-0078

Yee, T. W. (2006). Constrained additive ordination. *Ecology*, *87*, 203–213. https://doi.org/10.1890/05-0283

Yee, T. W., & Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling*, *3*, 15–41. https://doi.org/10.1191/1471082X03st045oa

Zurell, D., Thuiller, W., Pagel, J., Cabral, J. S., Münkemüller, T., Gravel, D., … Zimmermann, N. E. (2016). Benchmarking novel approaches for modelling species range dynamics. *Global Change Biology*, *22*, 2651–2664. https://doi.org/10.1111/gcb.13251

SUPPORTING INFORMATION

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

How to cite this article: Nieto-Lugilde D, Maguire KC, Blois JL, Williams JW, Fitzpatrick MC. Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. *Methods Ecol Evol*. 2018;9:834–848. https://doi.org/10.1111/2041-210X.12936