Received: 25 October 2017 | Revised: 12 April 2018 | Accepted: 18 April 2018

DOI: 10.1111/ddi.12781

BIODIVERSITY METHODS

WILEY Diversity and Distributions

Multifaceted biodiversity modelling at macroecological scales using Gaussian processes

Matthew V. Talluto[1,2] | Karel Mokany[3] | Laura J. Pollock[2] | Wilfried Thuiller[2]

[1]Leibniz Institute for Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany

[2]Laboratoire d'Écologie Alpine (LECA), CNRS, Université Grenoble Alpes, Grenoble, France

[3]CSIRO, Canberra, ACT, Australia

Correspondence
Matthew V. Talluto, Leibniz Institute for Freshwater Ecology and Inland Fisheries (IGB), Müggelseedamm 310, 12587 Berlin, Germany.
Email: talluto@igb-berlin.de

Funding information
Unitatea Executiva pentru Finantarea Invatamantului Superior, a Cercetarii, Dezvoltarii si Inovarii, Grant/Award Number: 15/310 01.01.2014; Agence Nationale de la Recherche, Grant/Award Number: ANR-13-ISV7-0004 and ANR-16-CE93-0004; Executive Agency for the Financing of High Education, Research, Development and Innovation, Grant/Award Number: 15/310 01.01.2014; People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme, Grant/Award Number: FP7/2007-2013; REA, Grant/Award Number: 659422

Editor: Risto Heikkinen

Abstract

Aim: Modelling the response of β-diversity (i.e., the turnover in species composition among sites) to environmental variation has wide-ranging applications, including informing conservation planning, understanding community assembly and forecasting the impacts of climate change. However, modelling β-diversity is challenging, especially for multiple diversity facets (i.e., taxonomic, functional and phylogenetic diversity), and current methods have important limitations. Here, we present a new approach for predicting the response of multifaceted β-diversity to the environment, called Multifaceted Biodiversity Modelling (MBM). We illustrate the approach using both a plant diversity dataset from the French Alps and a set of simulated data. We also provide an implementation via an R package.

Location: French Alps.

Methods: For both the French Alps and the simulated communities, we compute β-diversity indices (e.g., Sørensen dissimilarity, mean functional/phylogenetic pairwise distance) among site pairs. We then apply Gaussian process regression, a flexible nonlinear modelling technique, to predict β-diversity in response to environmental distance among site pairs. For comparison, we also perform similar analyses using Generalized Dissimilarity Modelling (GDM), a well-established method for modelling β-diversity in response to environmental distance.

Results: In the Alps, we observed a general increase in taxonomic (TD) and functional (FD) β-diversity (i.e., site pairs were more different from each other) as the climatic distance between site pairs increased. GDM performed better for TD and FD when fitting to calibration data, whereas MBM performed better for both when predicting to a validation dataset. For phylogenetic β-diversity, MBM outperformed GDM in predicting the observed decrease in phylogenetic β-diversity with increasing climatic distance.

Main conclusions: Multifaceted Biodiversity Modelling provides a flexible new approach that expands our capacity to model multiple facets of β-diversity. Advantages of MBM over existing methods include simpler assumptions, more flexible modelling, potential to consider multiple facets of diversity across a range of diversity indices, and robust uncertainty estimation.

KEYWORDS

beta diversity, biodiversity modelling, functional diversity, Gaussian processes, macroecology, phylogenetic diversity

1 | INTRODUCTION

Recent increases in the availability of biodiversity data have driven interest in understanding how biodiversity varies in response to environmental variables, land use and human impact (D'Amen, Rahbek, Zimmermann, & Guisan, 2017; Granger et al., 2015; Newbold et al., 2016). Modelling these relationships and projecting them into space have varied applications, including conservation planning, illuminating community assembly processes, understanding how evolutionary history shapes contemporary communities, and climate change planning (Bässler et al., 2015; Kraft et al., 2015; Lavergne, Mouquet, Thuiller, & Ronce, 2010; Yuan et al., 2016).

Historically, community-level diversity models have focused on species diversity, applying metrics such as species richness or Simpson's diversity that relates to taxonomic diversity within a single site (i.e., α-diversity) (D'Amen et al., 2017). However, recent studies have demonstrated that taxonomic diversity (TD) alone is insufficient to capture all aspects of biodiversity (Cadotte, Albert, & Walker, 2013; Cadotte et al., 2010). Instead, adding consideration of functional diversity (FD) and phylogenetic diversity (PD) provides a more holistic view of biodiversity by capturing elements of ecological processes and evolutionary history.

Additionally, compared to α-diversity, β-diversity—representing turnover in diversity among sites—can provide more detailed information regarding how biodiversity structure varies in space, and can better capture the diversity of an entire landscape (Weinstein et al., 2014). For instance, contrasting functional β-diversity patterns to expectations from neutral evolutionary models can illuminate the historical processes producing species' ecological niches, which is important for understanding the origins of biodiversity (Mazel et al., 2017). Modelling and analysing the response of β-diversity to key environmental drivers has a strong role to play in providing important new insight into these biodiversity patterns and processes. However, while modelling the distribution of α-diversity has a long history of theoretical development and strong statistical models (D'Amen et al., 2017; Thuiller, Midgley, Rougeti, & Cowling, 2006), understanding taxonomic, phylogenetic and functional β-diversity (hereafter, we use "β-diversity" to refer to all 3 facets) is more challenging as it implies pairwise comparisons among localities. These comparisons can introduce statistical non-independence among modelled data points (where each point represents a pair of sites). Moreover, there is little theoretical basis for understanding how ecological processes producing patterns of β-diversity should be related to the environment, which, for a pair of points, could be represented by the environment at either site or the difference in the environments between the sites.

A first challenge in modelling β-diversity is developing a robust statistical framework. The different facets of β-diversity can be expressed by a range of indices, most of which are derived from a combination of species presence/abundance, phylogenetic and trait data, and which generally do not have clear prior expectations regarding their statistical distribution (compared with, e.g., Poisson or negative binomial expectations for species richness). Further complicating modelling may be lack of clear mechanistic connections between available environmental variables and the target index. This can make specifying the form of a regression difficult, and can result in poorly specified models, a poor fit to the data and violated assumptions when fitting indices that may have nonlinear and highly complex responses to the environment.

There are a number of techniques that have been developed to model pairwise β-diversity, though each has limitations. Most basic is linear matrix regression (Manly, 1986), which assumes a linear response of β-diversity to environmental distance, an assumption that is commonly violated in ecological datasets. To overcome this limitation, Ferrier, Manion, Elith, and Richardson (2007) developed generalized dissimilarity modelling (GDM), a powerful technique that accounts for nonlinearity in pairwise β-diversity across environmental gradients. However, the current implementation of GDM enforces somewhat limited forms for the relationship between diversity and the environment, and in particular forces β-diversity to increase monotonically as a function of environmental distance. This assumption can be problematic when modelling FD and PD. Another technique currently used to model β-diversity is gradient forests (GF) (Ellis, Smith, & Pitcher, 2012), an extension of random forests (Breiman, 2001) which is a machine-learning regression tree approach. While GF is more flexible than GDM (for example, there is no a priori assumption of increasing diversity with increasing environmental distance) (Ellis et al., 2012), it has no means of incorporating geographic distance and may be susceptible to overfitting.

While existing approaches to modelling β-diversity have advanced our capacity to understand biodiversity patterns, there is significant scope for development of new techniques that help overcome current limitations. Here, we present a novel method for statistically modelling β-diversity: Multifaceted Biodiversity Modelling (MBM). The underlying approach is quite general, in that it is suitable for modelling both α- and β-diversity for all facets of diversity and for any choice of index. For simplicity, we focus here on modelling β-diversity (i.e., β TD, FD and PD). The foundation of MBM is Gaussian process regression, a highly flexible Bayesian approach to machine learning. This method improves on some of the shortcomings of other methods (e.g., assumptions of linearity, monotonicity). Computational time is reasonable, allowing for the application of MBM to large datasets. Furthermore, the analysis provides full conditional posterior estimates of all biodiversity metrics, allowing for a robust understanding of uncertainty that can be propagated when biodiversity metrics are used in downstream analyses. We present the general approach here, then describe specific modelling steps in detail with a case study of plant β-diversity in the French alps. To explore computational time and compare modelling performance with GDM, we also model a simulated dataset constructed from virtual species. An implementation is provided via an R package, *mbm*, which is freely available (http://github.com/mtalluto/mbm).

2 | METHODS

The core of our method relies on Gaussian Process (GP) regression, a widely used Bayesian machine learning technique. A full discussion of GPs is beyond the scope of this paper; we direct the reader to (Rasmussen & Williams, 2005) for a comprehensive mathematical treatment and to (Golding & Purse, 2016; Jones & Moriarty, 2012) for ecological examples. In general terms, a GP regression operates by estimating a smooth latent function $f(x)$ of a set of predictor variables x, where the value of $f(x)$ is defined by a multivariate normal distribution with mean function $\mu(x)$ and covariance function $k(x)$. We can then estimate the probability of observing a response variable y (here a metric of β-diversity) as $P(y) = D[f(x), \sigma]$, where D is some probability density function with expectation $f(x)$ and (depending on the distribution) dispersion parameter σ. Much like generalized linear modelling (GLM), generalized additive modelling (GAM), and other methods, a variety of distributions are available, and the latent function can be connected to the response via a link function. Currently, the *mbm* package implements Gaussian likelihoods with options for identity, probit, and log links, although other options will be possible in future versions. Guidelines for installation and testing of basic functionality for the *mbm* package are provided in Supporting Information Appendix S1.

GP regressions produce latent functions (i.e., predictions for the mean of y) that smoothly vary as a function of the predictor variables and, depending on the covariance function used, can take flexible shapes that do not necessarily conform to, e.g., linear or quadratic relationships between y and x. This makes them a natural choice for biodiversity modelling, where predictive inference is desired but the link between predictors and response is poorly understood and the shape of the response function may be irregular and nonlinear. Often the relationship between climate and diversity is strong, but acts indirectly via unknown intermediates (e.g., climate may influence the physiological behaviour of individuals, which may then scale to population- and species-level performance). GPs have been suggested for similar reasons for modelling species distributions (Golding & Purse, 2016).

2.1 | Model structure

Our general approach is to estimate β-diversity indices for pairs of sites and to use these indices as responses in a GP regression on a set of predictor variables. These predictors can vary, but will always include a pairwise measure of environmental distance. As with a GLM, it is important to consider the structure of the response variable and choose an appropriate likelihood and link function; we discuss these issues further in the case study. Once the model form is chosen, we estimate the model parameters and the mean and covariance functions of the GP using Laplace approximation {implemented in a Python package, GPy; gpy2014}. Because GPs are fit to data, the prior mean function $\mu(x)$ has little influence on the posterior mean of the latent function when the data coverage is adequate. Thus, in general, we fit a prior mean $\mu(x) = 0$. However, when extrapolating or when data coverage is poor, this can lead to undesired behaviour as the GP will tend to revert to the mean; therefore it may be desirable to fit an alternative prior mean. We discuss selection of the mean function in more detail in the case study.

For the covariance function, GP regression uses a kernel function centred around each data point x_i to describe covariance in the latent function based on the predictor variables. Thus, the shape of the underlying kernel will determine the shape of the predicted relationship between the predictor variables and the responses. For all of our examples, we have selected a negative exponential kernel (also sometimes called the radial basis function), which, assuming the most basic case of a single predictor variable x, defines the covariance between two points x_i and x_j as:

$$k(x_i,x_j) = \sigma_k^2 \exp\left(-\frac{1}{2}r^2\right)$$

$$r = \frac{x_i - x_j}{l}$$

The hyperparameters σ_k^2 and l indicate the kernel variance (i.e., variance in the latent function $f(x)$) and the length scale, respectively. This kernel has the desirable property that the strength of covariance decreases as sites become more dissimilar in their predictor variables and is widely used in GP regression. We therefore expect, for most biodiversity modelling applications, that this kernel will be the most useful, and it is presently the only kernel implemented in the *mbm* package. Other kernels are discussed in detail in Rasmussen and Williams (2005).

In order to fit the latent function, it is necessary to provide the hyperparameters to the kernel. If known, they can be supplied as fixed constants, but generally users will wish to estimate these parameters from the data. For many problems, it will be sufficient to use maximum likelihood estimation to find point estimates for the hyperparameters. Otherwise, hyperparameter posterior distributions can be estimated by supplying appropriate prior distributions and using a Markov chain Monte Carlo (MCMC) sampler.

A common issue when modelling β-diversity is that the relationship between β-diversity and environmental distance is not constant across an environmental gradient (Ferrier & Guisan, 2006). For example, turnover may be more rapid on the cold end of a temperature gradient than on the warm end. To address this issue, *mbm* by default includes both environmental distance (computed as the multivariate Euclidean distance in environmental space between each pair of survey points) along with the average position on the environmental gradient for each environmental variable, resulting in a model with $n + 1$ predictors for n environmental variables. By default, *mbm* centres and scales all predictor variables to zero mean and unit variance before computing the environmental distance among points. The multivariate kernel used is anisotropic, where the kernel function k is defined in $n + 1$ dimensions (with a corresponding length scale hyperparameter l for each dimension). Having a single compound anisotropic kernel provides considerable flexibility in fitting a model to various combinations of the

predictors and allows the shape of the response of β-diversity to environmental distance to vary at different locations in the parameter space. A second issue is that it may be desirable in some cases to enforce additional constraints; for instance, although we may lack the data on distant sites to fit such a constraint, we may have the prior expectation that, above a certain environmental distance, taxonomic dissimilarity saturates at 1 and does not decline. We suggest the use of a prior mean function in these cases and demonstrate this in the case study.

Because β-diversity is defined between site pairs, the size of the input dataset of site pairs can become very large. Models of such datasets can be infeasible to fit with Laplace approximation, particularly if it is necessary to optimize hyperparameters. In these cases, we suggest employing a stratified subsampling scheme prior to fitting the GP, where the model is then fit with β-diversity indices computed only for the selected sites. In addition to improving computation time, subsampling sites reduces non-independence among data points and provides holdout data against which models can be evaluated to guard against overfitting. The nature of the stratification will naturally depend on the characteristics of the dataset, but it should be constructed to ensure that full ranges of both the environmental gradient and the response variable are adequately represented.

2.2 | Case study: French Alps plants

We demonstrate the method using an intensive dataset (4,417 sites and 2,863 plant species) from a well-studied ecosystem in the French Alps. We used a 2.5 km-resolution database of plant occurrences in the French Alps (Figure 1) (Thuiller, Pollock, Gueguen, & Münkemüller, 2015) with climatic covariates selected from bioclimatic rasters derived from WorldClim data (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005). We upscaled the climatic data to 2.5 km (to match the resolution of the occurrence data) by taking the area-weighted average of cells at the starting resolution (30 s). We used four variables: temperature seasonality (i.e., the coefficient of variation of monthly average temperatures), minimum temperature of the coldest month, annual temperature range and precipitation seasonality. These variables were selected based on exploratory analyses that identified variables with strong univariate correlations to β-diversity, ruling out any variables with correlations greater than 0.7 with other predictor variables. We centred all variables to zero mean and scaled to unit variance before analysis. Because of the high degree of redundancy inherent in a full pairwise dissimilarity matrix (i.e., where all possible site pairs are represented) and to make computational time reasonable, we stratified the cells by elevation and by sampling effort. First, we removed cells with fewer than five samples, with the exception of cells above 3,000 m elevation (because high-elevation regions were less-intensively sampled, and we wanted to ensure adequate coverage of high alpine areas). Then, we divided the study area into fifteen 250-m elevational bands and randomly selected a maximum of eight cells from each band (chosen to yield a final sample size of approximately 100 cells, which testing indicated would provide reasonable computational time). Because there were few cells in the highest elevation band, we ended with a final sample size of 106 cells from the 4,417 raster cells in the original dataset, with an identical number of sites selected for validation.

Traits for all species were also extracted from (Thuiller et al., 2015). We used four traits, mean maximum vegetative height, leaf dry matter content (LDMC), seed mass and specific leaf area (SLA), that had the greatest coverage in the database and that represent classic plant strategies that are strongly tied to response to climate (Westoby, 1998). Of 2863 species in the occurrence dataset, 1054 had data for all four traits. All FD analyses were based on this subset of species. Seed mass, SLA and height were strongly right skewed, so we log-transformed them before analysis. All traits were then scaled to 0 mean and unit variance before computing functional distance. For phylogenetic diversity, we used a tree of European alpine flora resolved to the genus level (Thuiller et al., 2014). For our analyses, we selected the maximum a posteriori tree from 100 posterior samples.

For taxonomic β-diversity, we used the Sørensen dissimilarity, defined for a given pair of sites *i* and *j* as:

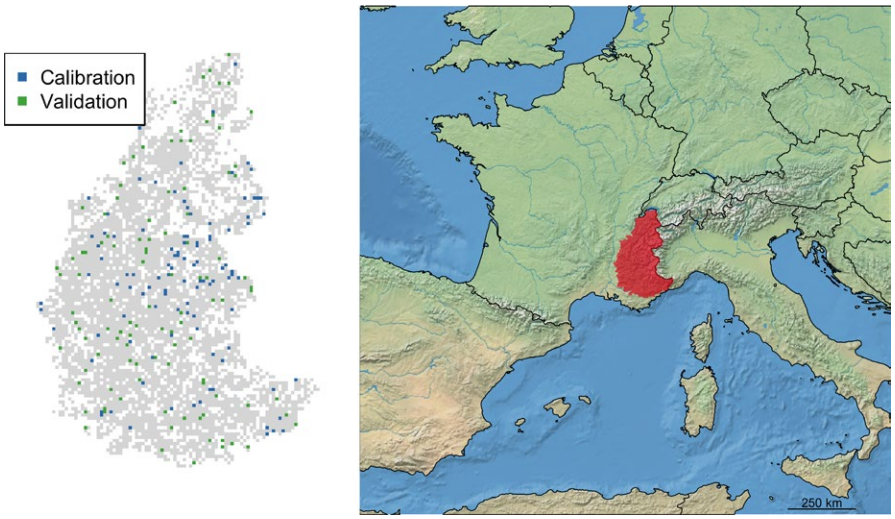

FIGURE 1 Map of study area, showing raster cells selected for calibration and validation. Empty (white) cells were excluded due to insufficient sampling. Right panel shows the location of the study area within Europe

$$S_{i,j} = 1 - \frac{2n_{ij}}{n_i + n_j}$$

where n_i is the number of species in site i, n_j is the number of species in site j, and n_{ij} is the number of species shared between the sites. Functional and phylogenetic β-diversity were computed as the mean pairwise distance (MPD) among all possible site pairs (Mouchet, Villéger, Mason, & Mouillot, 2010; Webb, Ackerly, McPeek, & Donoghue, 2002). MPD is simply the average distance (either phylogenetic or functional) between all possible pairs of species among two sites. We selected MPD because it is simple to compute and interpret and, because it is based on the distance matrix, it facilitates comparison between FD and PD. It has also been shown that MPD is sensitive to deep branching structure in phylogenetic trees and is thus less sensitive to poorly resolved phylogenies (Mazel et al., 2016). Unlike traditional measure like Faith's phylogenetic diversity, MPD is independent of species richness or species turnover (Mazel et al., 2016). Functions for computing MPD and Sorensen dissimilarity are included in the *mbm* package.

2.3 | GP model

We selected a Gaussian likelihood for the error distribution of all diversity indices; this likelihood is the most computationally simple in GP regression, but it does require an additional hyperparameter, σ^2_N, representing the standard deviation of the error distribution. For TD, the Sørensen index is bounded between 0 and 1, so we used a probit link to enforce this restriction. We fit a total of four TD models. The first used all of the default settings, with all parameters estimated via maximum likelihood. To demonstrate the effect of changing the length scale on the model, we fit identical models with the length scale fixed to either 2.0× or 0.2× the maximum likelihood estimate. Finally, to demonstrate how the mean function of the GP can be used to enforce prior expectations about the relationship between β-diversity and climatic distance, we fit a model with a linear mean function with the slope (on the link scale) constrained to be greater than 0. This represents a prior expectation that dissimilarity saturates to 1 at large climatic distances. This also allowed for an easy comparison to the GDM package, which makes a similar assumption (Ferrier et al., 2007). For FD and PD, we used no link function, and, because we had no prior expectation about the mean functions for FD and PD, we used the default $\mu(x) = 0$. For comparison with GDM, we built GDMs for all three diversity facets using the same calibration data and set of predictors. We used the default options in GDM and dropped any predictors that had no effect (i.e., that had all I-spline parameters fixed to 0). We then computed the root mean square error (RMSE) of the MBM and GDM models for both the calibration and validation datasets to compare both fit to the data as well as the ability to predict to new data.

2.4 | Case study: simulated communities

We used simulations in order to compare MBM with GDM when the underlying "true" relationship between β-diversity and the environment is known, and to explore how computational time and model performance for both methods change as sample sizes change. Our general approach was to generate a simulated landscape using randomly generated species' niches along two environmental axes, then "sample" this landscape at intensities ranging from 25 to 2,025 sites, then compare how well a model trained on these sites predicts β-diversity at the remaining unsampled sites.

To generate the landscape, we first generated random niches for 300 species along two environmental axes. Niches were described by a bivariate Gaussian distribution, where the $p_{s,i}$, the probability of occurrence of species s at site i, is given by:

$$p_{s,i} = \rho_s \times \mathcal{N}(\mathbf{x}_i, \boldsymbol{\mu}_s, \boldsymbol{\sigma}_s)$$

where $\mathcal{N}$ is the bivariate normal density function given the location vector $\mathbf{x}_i$ (i.e., environmental values) and the species-specific mean vector $\boldsymbol{\mu}_s$ and covariance matrix $\boldsymbol{\sigma}_s$ (i.e., the centre and widths of the niche in the two environ mental dimensions). Finally, the scale parameter ρ_s was included to allow for species-specific variation in overall probability of occurrence. These species-specific parameters were drawn at random from hyperdistributions as follows: $\boldsymbol{\mu}_s$ bivariate Gaussian with mean (0,0), standard deviation of (12,12) and no covariance; $\boldsymbol{\sigma}_s$: diagonal entries (i.e., standard deviation for each environmental axis) drawn from two independent Gamma distributions with shapes (4,7) and rates (1.2, 1.2) and off diagonals (i.e., covariance) set to 0; ρ_s: Beta distribution with parameters (15,5). These hyperparameters were chosen via exploration to produce landscapes with per-site species richness varying between approximately 5 and 50 when both environmental variables varied from −5 to 5.

We then defined the landscape as a 100 × 100 grid of sites, with each dimension defined by an environmental gradient varying uniformly between the arbitrarily-chosen values −5 and 5. Each site was populated with species by computing the probability of presence for all species, then conducting Bernoulli trials where a success indicated that a species was present. For simplicity, we do not consider species interactions; thus, all species distributions were independently drawn. This procedure generated a landscape where β-diversity increased smoothly and monotonically with increasing environmental distance; this relationship is commonly observed in studies of taxonomic β-diversity and is the case for which GDM is built (Ferrier et al., 2007).

To simulate a sampling process, we used a random starting location and then selected n evenly spaced locations from that start. We used sample sizes of 25, 49, 100, 225, 529, 1024 and 2025 (sample sizes are perfect squares, so that both environmental dimensions were sampled with the same intensity). We then ran both MBM and GDM to predict β-diversity as a function of environmental distance on the selected sites. For MBM, we used the GP as described in the case study (which is appropriate for smaller sample sizes) when the number of sites was less than 100. For larger sample sizes, we used an approximation method (Hemsman, Fusi, & Lawrence, 2013) implemented in the *mbm* package via the *svgp* option (Supporting information Appendix

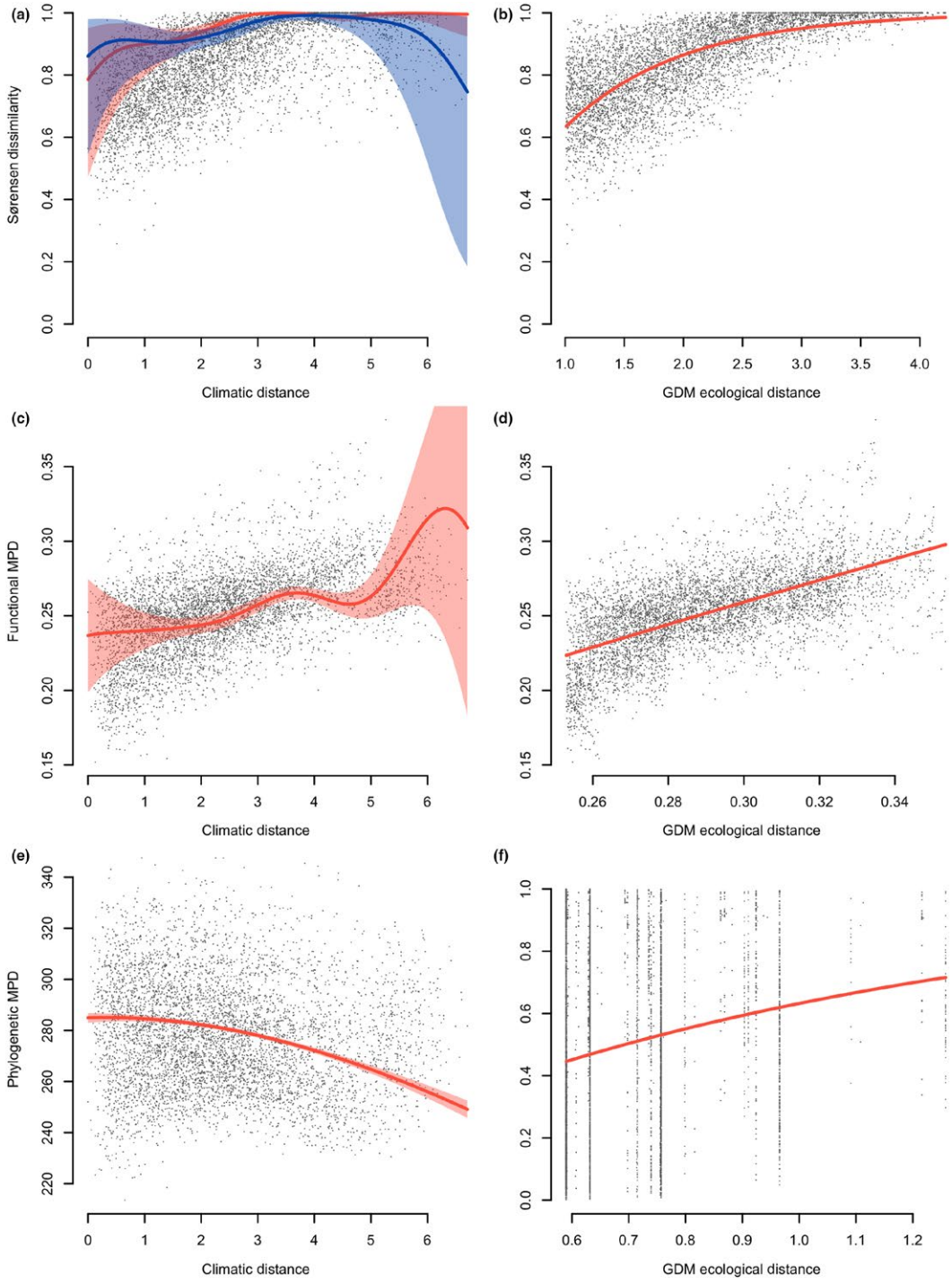

FIGURE 2 Comparison of model fits of MBM (a, c, e) and GDM (b, d, f) for taxonomic (top), functional (middle), and phylogenetic (bottom) β-diversity for the French Alps data. Also shown from MBM is the effect of including a mean function (panel A, in red) compared with no prior expectation (panel A, in blue). All MBM curves assume that, at a given distance, sites are equally spaced around the centre of the environmental gradient. GDM Ecological Distance refers to the distance between site pairs after applying the basis functions to the environmental data; see (Ferrier et al., 2007) for details. Note that the response variable for PD has been scaled between 0 and 1 for GDM fits. Uncertainty envelopes for MBM fits show 95% credible intervals. Uncertainty estimates were not available for GDM fits

S2). We recorded the computational time for each model run; although these values will vary greatly depending on the computational resources available, we used widely available hardware (a laptop with a 2-core 3.3 GHz CPU and 16 GB of RAM) and thus they can provide some guidance to users as to the size of models that can be run. Additionally, we computed predictive performance at each sample size using the RMSE computed for a set of 5,000 sites that were not included in the model calibration.

Finally, we ran each scenario ten times and report the mean and range of RMSE and computing time.

3 | RESULTS

We found a trend of increasing β TD with climatic distance for the French Alps dataset (Figure 2). However, due to sparse data at the largest distances, the default model predicted a sharp decrease in β-diversity at large distances (i.e., reverting to the prior mean of 0), contrary to our expectations. In contrast, fitting a model with a prior mean resulted in a curve that saturated to high dissimilarity at large ecological distances (Figure 2a). FD also increased with climatic distance, while PD decreased. Visualizing spatial patterns in β-diversity revealed clear clusters of similar communities (Figure 3). The influence of the Mediterranean region was apparent in all three metrics as a cluster of differentiated communities in the south. In the high mountains in the eastern portion of the study area, TD and PD were divided among northern and southern regions, while communities were functionally similar throughout high elevation regions.

Both MBM and GDM predicted similar trends for TD and FD, although the MBM models showed increased curvature (Figure 2, Supporting information Figure S1). With respect to fit to the calibration data, GDM performed slightly better for TD and substantially better for FD (Table 1). For out-of-sample prediction (i.e., with the validation dataset), MBM performed better for all three facets (Table 1). Increasing the length scale of the MBM model produced a model response curve that was very similar to the GDM curve (Figure 4). For PD, GDM performed poorly; all environmental variables except precipitation seasonality failed to converge; thus we fit a model with only this one variable. Thus, for PD, MBM provided a better fit than GDM for both the calibration and validation datasets (Table 1).

For the simulations, both MBM and GDM produced qualitatively similar fits to the subsampled sites (Supporting information Figure S1). In terms of out-of-sample root mean square prediction error (RMSE), predictions from both MBM and GDM in our simulated datasets improved as sample sizes increased up to $n = 225$, after which prediction accuracy was approximately constant (except for the largest sample size, where MBM predictions performed worse). MBM outperformed GDM at all tested sample sizes except the largest ($n = 2{,}025$), where GDM performed slightly better on average but was within the range of variability for MBM (Table 2). Computational times for GDM were considerably faster, with times 1–2 orders of magnitude faster than MBM at all sample sizes.

4 | DISCUSSION

Overall, we found MBM to be a robust approach to modelling β-diversity. Performance was similar to the existing method (GDM), with MBM generally performing slightly worse when predicting calibration data but slightly better for validation data in both an empirical dataset from the French Alps (Table 1) and a simulated dataset (Table 2). Moreover, MBM provides some notable advantages. Namely, MBM allows more varied relationships between turnover and the environment, better captures prediction uncertainty, and is extensible to incorporate a wider variety of model structures. In particular, GDM performed very poorly modelling PD, likely due to the decreasing trend of PD with climatic distance (Figure 2). Although we were able to obtain convergence with one predictor, it is clear from plots that the model is mis-specified (Figure 2). Moreover, the I-spline transformation of the predictor in GDM obscures the decrease in PD with climatic distance that we observed using MBM. Such patterns may occur if important traits evolve easily within a clade, leading to high intra-clade functional divergence and thus greater phylogenetic similarity among highly dissimilar environments (Graham & Fine, 2008). Similarly, within-community phylogenetic clustering can increase with the scale of communities, reflecting the inclusion of entire clades within communities (Cavender-Bares, Keen, & Miles, 2006). Thus, for local or regional analyses, we may expect decreasing PD with increasing distance, particularly if the overall extent of the analysis is not very large relative to the resolution. In such cases, we expect GDM and phylogenetic extensions of GDM to struggle to correctly predict how PD changes with the environment. Our approach offers an improved tool for modelling β-diversity when the assumptions of GDM may be violated; the flexibility of the covariance functions plus the prior mean function allow for a model that makes similar assumptions as GDM, but also allow for relaxing these assumptions when they are not appropriate. Other methods with potential applications to β-diversity modelling, such as gradient forests (Ellis et al., 2012), may also overcome some of these issues.

Another advantage of MBM is the fully Bayesian core underlying the analysis, meaning that robust uncertainty estimation is a fundamental part of the method. By comparison, uncertainty analysis in GDM requires permutation testing or embedding the analysis in a Bayesian simulation (Woolley, Foster, O'Hara, Wintle, & Dunstan, 2017), eliminating computational time as one of the principle advantages to GDM. To improve computational time, we have presented parametric confidence intervals (based on standard errors estimated with Laplace approximation), and this is the default in the *mbm* package. However, posterior simulations are also possible, which will yield more robust uncertainty estimates and can also propagate modelling uncertainty to additional analyses based on the predictions. However, there is a cost to this flexibility, and *mbm* has two principal disadvantages. First, because it depends on third-party libraries, installation is more complex than is standard for R packages and requires the user to install Python and additional libraries before the package will function. Second, as with many Bayesian methods, performance can be much slower than non-Bayesian methods (Table 2). Thus, very large problems will require significant computational resources or may be better-suited to other methods. However, the *mbm* package includes options that

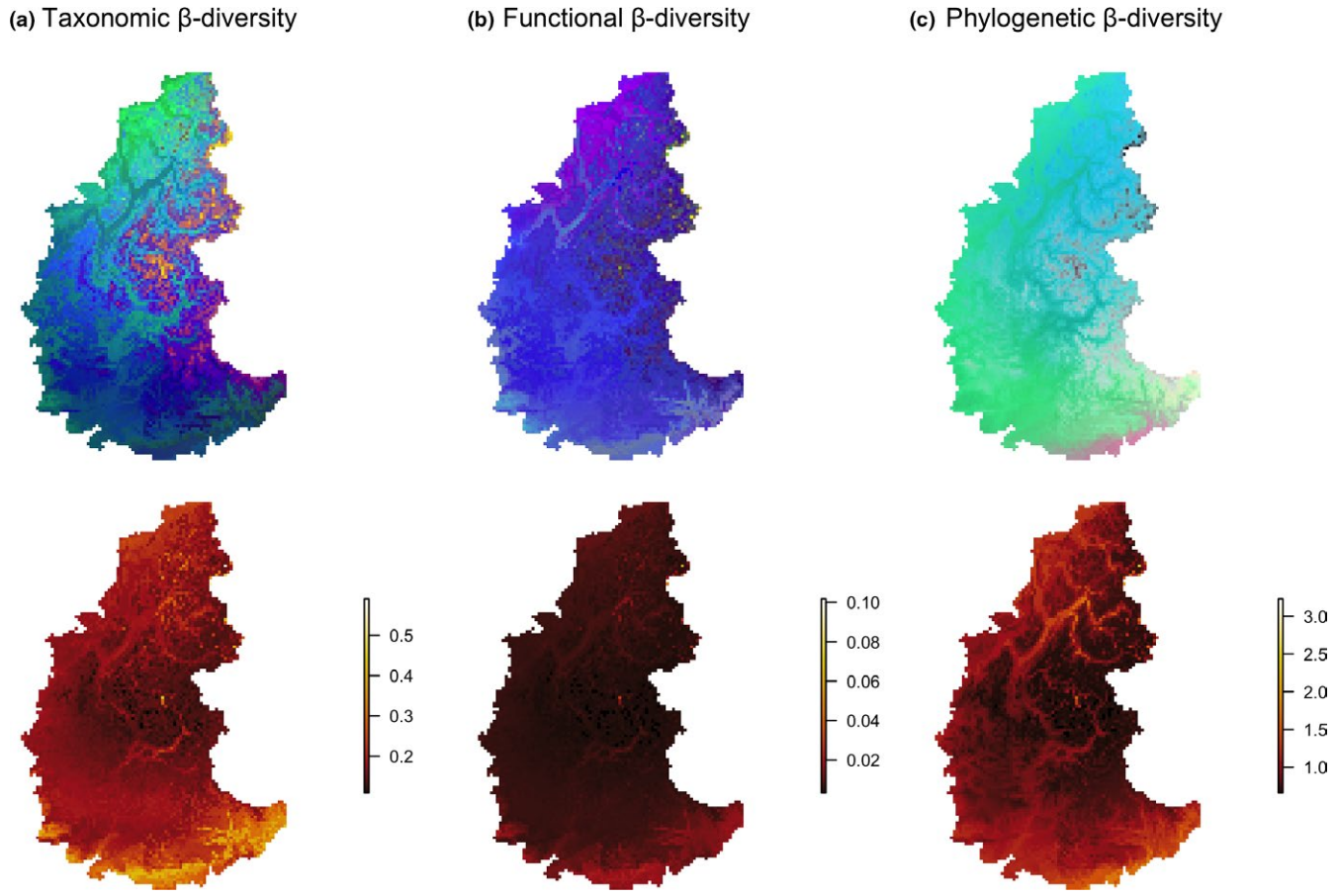

FIGURE 3 Spatial MBM predictions for three facets of diversity. In the upper row, similar colours depict similar communities (as measured by predicted Sørensen dissimilarity). The colours result from computing all possible predicted pairwise (among pixels) dissimilarity values, performing a principal components analysis on the resulting dissimilarity matrix, and using the first three axes (which described 70%–89% of the variance) as red, green and blue colour channels. For each pixel, MBM predicts β-diversity for that pixel compared to every other pixel as well as a standard error for each prediction. The average spatial uncertainty for each pixel (lower row) is then the mean of the standard errors of all predictions for that pixel

TABLE 1 Root mean square error for MBM and GDM models for calibration data (with RMSE for validation data in parentheses); smaller errors within columns (bolded) indicate better performance

	Calibration dataset			Validation dataset		
	Taxonomic	**Functional**	**Phylogenetic**	**Taxonomic**	**Functional**	**Phylogenetic**
MBM	0.113	0.0143	**18.5**	**0.107**	**0.0469**	**32.6**
GDM	**0.091**	**0.0019**	21.0	0.129	0.0654	34.5

have been optimized for large datasets. Although slower than other options, even with large sample sizes (e.g., 2,025 sites resulting in 2×10^6 unique site pairs) we obtained results in under 15 min. Finally, as with any new method, it will be necessary to continue to evaluate MBM in terms of performance and prediction accuracy across a wide range of datasets.

The application of MBM presented here has been relatively straightforward, as it is based on what can be accomplished with a simple installation of the *mbm* R package (and thus will be accessible to the widest range of users). However, the flexibility of the underlying method allows for a number of possible extensions to MBM. In particular, although our dataset had good coverage in all three facets of diversity, it will often be the case that more is known about (for example) taxonomic diversity than functional or phylogenetic diversity, and that these gaps in knowledge are spatially congruent, such as where traits are sampled well in one region but poorly in another. Although not presently implemented, it is possible to extend MBM to multiresponse models. These models use a coregionalization kernel for the covariance function, which models a given response as a function of both the covariates as well

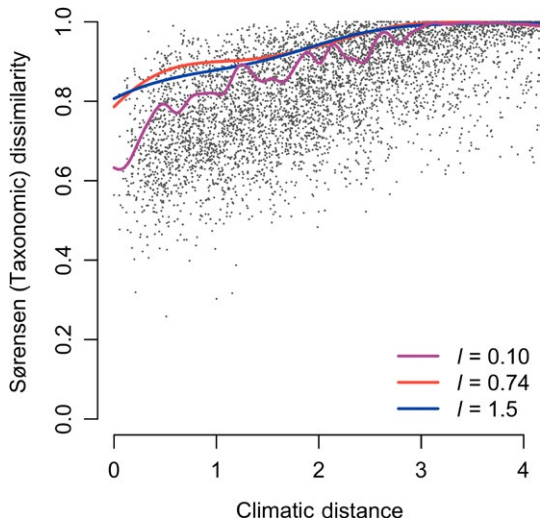

FIGURE 4 The effect of fitting different length scales (*l*). The maximum likelihood value, *l* = 0.74, represents the best fit of the model to the data. Reducing the length scale allows for greater flexibility in the curve, whereas increasing it results in a smoother relationship

as the other response variables in the model (Álvarez, Rosasco, & Lawrence, 2012). In this way, responses that are known poorly from some portions of environmental space can borrow strength from the response variables that are known well.

Multifaceted Biodiversity Modelling is particularly well suited for modelling in environments where the relationship between community composition (loosely defined to include whatever facets of diversity are of interest) and the environment is complex. Along simple environmental gradients or when community assembly is largely driven by environmental filtering, we expect that β-diversity increases both with geographic and environmental distance, and this increase should be accompanied by a reduction in variance (e.g., Supporting information Figure S1). Modelling in these environments is likely to be straightforward, regardless of the tool used. In contrast, an ecological or biogeographic process leading to convergence of widely spaced communities will lead to non-increasing relationships between turnover and distance, potentially resulting in "spikes" in variance at intermediate distances, such as if some communities continue to diverge with distance while others converge. Examples of such processes include succession, where early- and mid-successional communities may show convergence across environments even when late-successional communities diverge (Christensen & Peet, 1984; Halpern, 1988; Romme, Whitby, Tinker, & Turner, 2016). Similar convergence in FD can occur whenever similar trait values adapt species to different environmental conditions (e.g., small leaves can be adaptive in both very cold and very dry environments). We can also expect to see flat or complex relationships between β-diversity and the environment in communities that are structured by strong competition and dispersal limitation (Myers et al., 2013). In such situations, modelling the variability in β-diversity may be as interesting as the mean, especially for comparison with other environments. Finally, evolutionary processes that lead to high functional divergence within clades can produce nonlinear and even decreasing relationships between phylogenetic β-diversity with the environment (Graham & Fine, 2008). The flexible model forms and robust uncertainty estimation provided by MBM may better capture these patterns and help illuminate underlying processes driving β-diversity.

Due to the current biodiversity crisis, there is a critical need for new models targeted at understanding the distribution of diversity and the response of diversity to climate, change in human land use, and other environmental factors. This challenge is being met with an explosion in the availability of biological and environmental datasets, thus providing an opportunity to meet the crisis with new methods. Multifaceted biodiversity modelling is an extension to other common methods for modelling aggregate biodiversity (Ferrier & Guisan, 2006; Ferrier et al., 2007; Guisan & Rahbek, 2011) that is well-suited for phylogenetic and functional diversity in particular. MBM is also a Bayesian analysis; thus it is compatible with specifying prior knowledge, provides robust estimates of uncertainty, and can be used in downstream analyses that require propagation of uncertainty. Finally, the method is highly extensible at its core, allowing new theoretical developments or alternative model structures to be easily incorporated into the larger framework.

TABLE 2 Comparison of MBM and GDM on simulated data at varying sample sizes[a]

	RMSE[c]		Computational time (seconds)	
Sample size (# of sites)[b]	MBM	GDM	MBM	GDM
25 (300)	0.061 (0.059–0.064)	0.070 (0.064–0.077)	2.2 (1.6–3.6)	0.020 (0.013–0.032)
49 (1176)	0.060 (0.057–0.061)	0.064 (0.061–0.068)	10 (8.3–16)	0.020 (0.016–0.024)
100 (4950)	0.061 (0.058–0.065)	0.062 (0.060–0.064)	27 (25–30)	0.040 (0.032–0.055)
225 (25200)	0.058 (0.057–0.061)	0.060 (0.060–0.061)	32 (29–41)	0.16 (0.13–0.19)
529 (139656)	0.057 (0.055–0.060)	0.060 (0.060–0.060)	85 (80–83)	0.86 (0.80–0.93)
1024 (523776)	0.058 (0.056–0.062)	0.060 (0.060–0.060)	159 (125–231)	3.7 (3.1–4.6)
2025 (2049300)	0.062 (0.057–0.065)	0.060 (0.060–0.060)	761 (536–925)	14 (12–17)

[a]Values given are the means of 10 runs with the range in parentheses, [b]Parenthetical values are the number of unique site pairs, [c]Root mean square error.

ACKNOWLEDGEMENTS

This work was supported by the Agence Nationale de la Recherche (ANR) – France (Project ODYSSEE, ANR-13-ISV7-0004, and Project Origin-Alps, ANR-16-CE93-0004) and the Executive Agency for the Financing of High Education, Research, Development and Innovation (UEFISCDI) – Romania (Project ODYSSEE, PN-II-ID-JRP-RO-FR-2012, no. 15/310 01.01.2014). LJP acknowledges funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no. 659422.

DATA ACCESSIBILITY

Formatted data are available online via Figshare (https://doi.org/10.6084/m9.figshare.6300719). Species occurrences and traits were obtained from (Thuiller et al., 2015). Worldclim data (Hijmans et al., 2005) are freely available online. The plant phylogeny is available via (Thuiller et al., 2014). Code is available via the **mbm** R package (https://github.com/mtalluto/mbm).

ORCID

Matthew V. Talluto http://orcid.org/0000-0001-5188-7332
Karel Mokany http://orcid.org/0000-0003-4199-3697
Wilfried Thuiller http://orcid.org/0000-0002-5388-5274

REFERENCES

Álvarez, M. A., Rosasco, L., & Lawrence, N. D. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends® Machine Learning*, *4*, 195–266. https://doi.org/10.1561/2200000036

Bässler, C., Cadotte, M. W., Beudert, B., Heibl, C., Blaschke, M., Bradtka, J. H., … Müller, J. (2015). Contrasting patterns of lichen functional diversity and species richness across an elevation gradient. *Ecography*, *39*, 689–698.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Cadotte, M., Albert, C. H., & Walker, S. C. (2013). The ecology of differences: Assessing community assembly with trait and evolutionary distances. *Ecology Letters*, *16*, 1234–1244. https://doi.org/10.1111/ele.12161

Cadotte, M. W., Jonathan, Davies T., Regetz, J., Kembel, S. W., Cleland, E., & Oakley, T. H. (2010). Phylogenetic diversity metrics for ecological communities: Integrating species richness, abundance and evolutionary history. *Ecology Letters*, *13*, 96–105. https://doi.org/10.1111/j.1461-0248.2009.01405.x

Cavender-Bares, J., Keen, A., & Miles, B. (2006). Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. *Ecology*, *87*, S109–S122.

Christensen, N. L., & Peet, R. K. (1984). Convergence during secondary forest succession. *The Journal of Ecology*, *72*, 25. https://doi.org/10.2307/2260004

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 of the Cambridge Philosophical Society*, *92*, 169–187. https://doi.org/10.1111/brv.12222

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

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

Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, *7*, 598–608. https://doi.org/10.1111/2041-210X.12523

Graham, C. H., & Fine, P. V. A. (2008). Phylogenetic beta diversity: Linking ecological and evolutionary processes across space in time. *Ecology Letters*, *11*, 1265–1277. https://doi.org/10.1111/j.1461-0248.2008.01256.x

Granger, V., Bez, N., Fromentin, J.-M., Meynard, C., Jadaud, A., & Mérigot, B. (2015). Mapping diversity indices: Not a trivial issue. *Methods in Ecology and Evolution*, *6*, 688–696. https://doi.org/10.1111/2041-210X.12357

Guisan, A., & Rahbek, C. (2011). SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, *38*, 1433–1444. https://doi.org/10.1111/j.1365-2699.2011.02550.x

Halpern, C. B. (1988). Early successional pathways and the resistance and resilience of forest communities. *Ecology*, *69*, 1703–1715. https://doi.org/10.2307/1941148

Hemsman, J., Fusi, N., & Lawrence, N.D. (2013) Gaussian Processes for Big Data. Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*, 1965–1978. https://doi.org/10.1002/(ISSN)1097-0088

Jones, N. S., & Moriarty, J. (2012). Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies. *Journal of The Royal Society Interface*, *10*, 20120616. https://doi.org/10.1098/rsif.2012.0616

Kraft, N. J. B., Adler, P. B., Godoy, O., James, E. C., Fuller, S., & Levine, J. M. (2015). Community assembly, coexistence and the environmental filtering metaphor. *Functional Ecology*, *29*, 592–599. https://doi.org/10.1111/1365-2435.12345

Lavergne, S., Mouquet, N., Thuiller, W., & Ronce, O. (2010). Biodiversity and climate change: integrating evolutionary and ecological responses of species and communities. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 321–350. https://doi.org/10.1146/annurev-ecolsys-102209-144628

Manly, B. F. J. (1986). Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. *Researches on Population Ecology*, *28*, 201–218. https://doi.org/10.1007/BF02515450

Mazel, F., Davies, T. J., Gallien, L., Renaud, J., Groussin, M., Münkemüller, T., & Thuiller, W. (2016). Influence of tree shape and evolutionary time-scale on phylogenetic diversity metrics. *Ecography*, *39*, 913–920. https://doi.org/10.1111/ecog.01694

Mazel, F., Wüest, R. O., Gueguen, M., Renaud, J., Ficetola, G. F., Lavergne, S., & Thuiller, W. (2017). The geography of ecological niche evolution in mammals. *Current Biology*, *27*, 1–7.

Mouchet, M. A., Villéger, S., Mason, N. W. H., & Mouillot, D. (2010). Functional diversity measures: An overview of their redundancy and their ability to discriminate community assembly rules. *Functional Ecology*, *24*, 867–876. https://doi.org/10.1111/j.1365-2435.2010.01695.x

Myers, J. A., Chase, J. M., Jiménez, I., Jørgensen, P. M., Araujo-Murakami, A., Paniagua-Zambrana, N., & Seidel, R. (2013). Beta-diversity in temperate and tropical forests reflects dissimilar mechanisms of

community assembly. *Ecology Letters*, *16*, 151–157. https://doi.org/10.1111/ele.12021

Newbold, T., Hudson, L. N., Hill, S. L. L., Contu, S., Gray, C. L., Scharlemann, J. P. W., … Purvis, A. (2016). Global patterns of terrestrial assemblage turnover within and among land uses. *Ecography*, *39*, 1151–1163. https://doi.org/10.1111/ecog.01932

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. Cambridge, Mass: The MIT Press.

Romme, W. H., Whitby, T. G., Tinker, D. B., & Turner, M. G. (2016). Deterministic and stochastic processes lead to divergence in plant communities 25 years after the 1988 Yellowstone fires. *Ecological Monographs*, *86*, 327–351. https://doi.org/10.1002/ecm.1220

Thuiller, W., Gueguen, M., Georges, D., Bonet, R., Chalmandrier, L., Garraud, L., … Lavergne, S. (2014). Are different facets of plant diversity well protected against climate and land cover changes? A test study in the French Alps. *Ecography*, *37*, 1254–1266. https://doi.org/10.1111/ecog.00670

Thuiller, W., Midgley, G. F., Rougeti, M., & Cowling, R. M. (2006). Predicting patterns of plant species richness in megadiverse South Africa. *Ecography*, *29*, 733–744. https://doi.org/10.1111/j.0906-7590.2006.04674.x

Thuiller, W., Pollock, L. J., Gueguen, M., & Münkemüller, T. (2015). From species distributions to meta-communities. *Ecology Letters*, *18*, 1321–1328. https://doi.org/10.1111/ele.12526

Webb, C. O., Ackerly, D. D., McPeek, M. A., & Donoghue, M. J. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, *33*, 475–505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448

Weinstein, B. G., Tinoco, B., Parra, J. L., Brown, L. M., McGuire, J. A., Stiles, F. G., & Graham, C. H. (2014). Taxonomic, phylogenetic, and trait beta diversity in South American hummingbirds. *The American Naturalist*, *184*, 211–224. https://doi.org/10.1086/676991

Westoby, M. (1998). A leaf-height-seed (LHS) plant ecology strategy scheme. *Plant and Soil*, *199*, 213–227. https://doi.org/10.1023/A:1004327224729

Woolley, S. N. C., Foster, S. D., O'Hara, T. D., Wintle, B. A., & Dunstan, P. K. (2017). Characterising uncertainty in generalised dissimilarity models. *Methods in Ecology and Evolution*, *8*, 985–995. https://doi.org/10.1111/2041-210X.12710

Yuan, Z., Gazol, A., Lin, F., Wang, X., Ye, J., Suo, Y., … Hao, Z. (2016). Scale-dependent effect of biotic interactions and environmental conditions in community assembly: Insight from a large temperate forest plot. *Plant Ecology*, *217*, 1003–1014. https://doi.org/10.1007/s11258-016-0626-5

BIOSKETCH

All authors (except **KM**) belong to the research group of **Wilfried Thuiller** at the Alpine Ecology Lab (LECA) in Grenoble, France. The group focuses on understanding and predicting the spatio-temporal structure and dynamics of biodiversity across multiple scales. They use a combination of experiments and observations together with statistical and mathematical developments to achieve this goal.

Author contributions: All authors contributed to the initial framing of the paper and associated package. MT wrote the R-package and analysed the data. KM and LP beta-tested the package. All authors contributed to all phases of writing.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Talluto MV, Mokany K, Pollock LJ, Thuiller W. Multifaceted biodiversity modelling at macroecological scales using Gaussian processes. *Divers Distrib*. 2018;00:1–11. https://doi.org/10.1111/ddi.12781