

METHOD

A working guide to harnessing generalized dissimilarity modelling for biodiversity analysis and conservation assessment

Karel Mokany¹  | Chris Ware² | Skipton N. C. Woolley³ | Simon Ferrier¹ | Matthew C. Fitzpatrick⁴

¹CSIRO, Canberra, Australian Capital Territory, Australia

²CSIRO, Sandy Bay, Tasmania, Australia

³School of Ecosystem and Forestry Science, The University of Melbourne, Parkville, Victoria, Australia

⁴Appalachian Lab, University of Maryland Center for Environmental Science, Frostburg, Maryland, USA

Correspondence

Karel Mokany, CSIRO, PO Box 1700, Canberra, ACT 2601, Australia.
Email: Karel.Mokany@csiro.au

Funding information

Commonwealth Scientific and Industrial Research Organisation; National Science Foundation, Grant/Award Number: 1461868 and 1656099

Editor: Volker Bahn

Abstract

Aim: Generalized dissimilarity modelling (GDM) is a powerful and unique method for characterizing and predicting beta diversity, the change in biodiversity over space, time and environmental gradients. The number of studies applying GDM is expanding, with increasing recognition of its value in improving our understanding of the drivers of biodiversity patterns and in implementing a wide variety of spatial assessments relevant to biodiversity conservation. However, apart from the original presentation of the GDM technique, there has been little guidance available to users on applying GDM to different situations or on the key modelling decisions required.

Innovation: We present an accessible working guide to GDM. We describe the context for the development of GDM, present a simple statistical explanation of how model fitting works, and step through key considerations involved in data preparation, model fitting, refinement and assessment. We then describe how several novel spatial biodiversity analyses can be implemented using GDM, with code to support broader implementation. We conclude by providing an overview of the range of GDM-based analyses that have been undertaken to date and identify priority areas for future research and development.

Main conclusions: Our vision is that this working guide will facilitate greater and more rigorous use of GDM as a powerful tool for undertaking biodiversity analyses and assessments.

KEYWORDS

beta diversity, compositional turnover, distance, GDM, generalized dissimilarity modelling, site-pair

1 | INTRODUCTION

Scientific understanding of the amazing diversity of life on Earth continues to develop and improve. Early ecology focused substantial

effort on understanding why the number of species in an assemblage (alpha diversity) varies in space and time (Gleason, 1922; MacArthur & Wilson, 1967). However, it has become increasingly apparent that changes in the composition of species (beta diversity) are just as, if

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 Commonwealth Scientific and Industrial Research Organisation. *Global Ecology and Biogeography* published by John Wiley & Sons Ltd.

not more, important and informative of ecological pattern and process (Anderson et al., 2011; Ferrier, 2002; Hillebrand et al., 2018; Legendre et al., 2005; Whittaker, 1960). In addition to taxonomic beta diversity, there is increasing interest in improving our understanding of what drives genetic, phylogenetic, and functional diversity across space and through time (Jay et al., 2012).

While beta diversity is sometimes quantified as a single metric for a whole region (Tuomisto, 2010), considering beta diversity as differences between site-pairs is potentially more information rich, and is the perspective considered here. Such pairwise analyses of beta diversity typically require a distance based statistical framework (Anderson et al., 2011). Statistical assessment of pairwise dissimilarities in ecology have developed over time. Earlier approaches include the Mantel test of correlation between distance matrices (Legendre, 1993), ordination of assemblages based on their pairwise compositional differences [e.g., non-metric multidimensional scaling (NMDS); Prentice, 1977], and linear matrix regression (Manly, 1986), where the level of difference between pairs of assemblages is related linearly to differences in environment or space. While each of these approaches have their benefits and continue to be applied, they fail to account for two key nonlinearities often apparent in pairwise dissimilarities (Ferrier et al., 2007). First, many ecological dissimilarity measures are constrained between 0 and 1 and therefore saturate at a maximum value of 1 once pairs of assemblages are completely different (i.e., further environmental distance between sites in a pair cannot increase ecological dissimilarity beyond a value of '1'). Second, change in assemblage composition can occur more rapidly at some points along environmental gradients than others (Faith & Ferrier, 2002; Fitzpatrick et al., 2013; Oksanen & Tonteri, 1995).

To better account for these two important nonlinearities in statistically relating beta diversity to environmental gradients, Simon

Ferrier led the development of generalized dissimilarity modelling (GDM; Ferrier, 2002; Ferrier et al., 2004, 2007). This approach to modelling pairwise distances is an extension of generalized linear modelling (GLM), and was stimulated by the very practical need for improved information on the distribution and drivers of biodiversity to inform conservation decisions. While initial implementation of GDM was enabled by customized applications (ARCGIS extension, windows software package, unofficial R package), the release of a CRAN R package in 2015 (*gdm*: Fitzpatrick et al., 2021; R Core Team, 2021) provided broad access and transparency of GDM for the research community.

Application of GDM to improve our understanding of biodiversity patterns and processes has increased steadily since its development (Figure 1). GDM has now been used in studies from terrestrial systems to the marine and aquatic realms (Koubbi et al., 2010; Leathwick et al., 2010), and spanning regional, to continental and global scales (Ferrier et al., 2004; Hoskins, Harwood, Ware, Williams, et al., 2020; Prober et al., 2012). GDM has proven particularly useful in improving knowledge for highly diverse taxonomic groups for which there may be many species known from only a small number of locations, but for which there are sufficient data to characterize changes in assemblage composition (e.g., invertebrates, tropical plants; Chesters et al., 2019; Mokany et al., 2014). While most applications of GDM have focused on modelling changes in assemblage composition, it has also been applied to model genetic (Fitzpatrick & Keller, 2015; Murray et al., 2019) and trait (Thomassen et al., 2010) differences between populations within species and phylogenetic differences between assemblages (Rosauer et al., 2014). GDM has mainly been used for plot-based datasets (Jones et al., 2013), but also has shown promise for use with presence-only inventory data

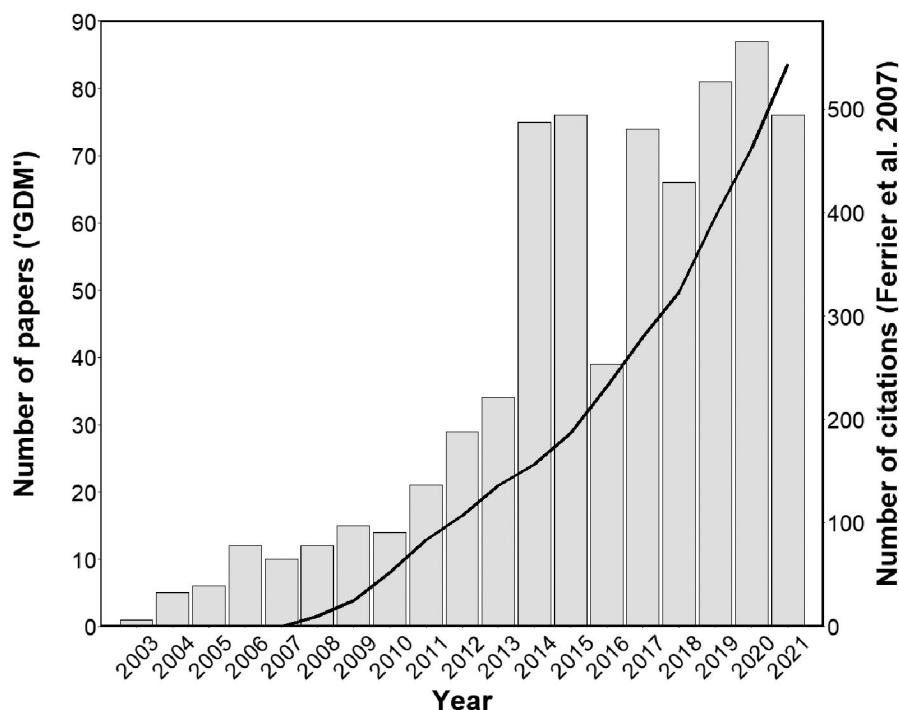


FIGURE 1 The number of papers per year using the phrase “generalized dissimilarity model*” (or “generalised dissimilarity model”) shown with grey bars, and the cumulative number of citations of the primary reference for generalized dissimilarity modelling (GDM; Ferrier et al., 2007) shown as a line. Data are from queries in Web of Science

(Fitzpatrick et al., 2011, 2013; Hoskins, Harwood, Ware, Williams, et al., 2020; Ware et al., 2018). The capacity for GDM to predict patterns in turnover across entire spatial grids (e.g., Overton et al., 2009) is one of its core strengths, enabling a diverse range of spatial analyses and applied conservation assessments (Williams et al., 2010). It is worth noting that other spatial modelling approaches have also emerged to consider and account for beta diversity, including gradient forests (Ellis et al., 2012) and joint species distribution models (Pollock et al., 2014), which each have their own benefits and limitations (Fitzpatrick & Keller, 2015; Wilkinson et al., 2019).

While the application of GDM is expanding, there is limited guidance available on fitting, interpreting, and applying GDM for spatial assessments. Here we aim to address this shortfall by providing a working guide to GDM, with reference to R-based implementations using the package *gdm* (Fitzpatrick et al., 2021). We present a simple description of how GDM fitting works, then describe the key steps in a typical workflow: data preparation, model fitting, model interpretation, model selection, predicting dissimilarity and spatial analyses. For each step in the workflow we discuss important issues requiring consideration, with a supporting vignette providing a guide to the more practical implementation in R (see Supporting Information Appendix S1). We illustrate key points using data and a GDM developed for vascular plant community compositional dissimilarity in Tasmania, Australia (Mokany et al., 2012). We conclude with a broad overview of studies that have extended GDM in novel and interesting ways, point to where research is required in further improving the rigour of GDM applications, and highlight some of the most productive areas for new application of GDM.

1.1 | Explanation of generalized dissimilarity modelling

A comprehensive statistical description on using GDM to model dissimilarities was presented by Ferrier et al. (2007), which remains the core reference for the GDM fitting approach. Here we provide a brief overview of the underlying model fitting procedure as it is currently implemented in the *gdm* R package (Fitzpatrick et al., 2021; R Core Team, 2021).

We start with the goal of modelling biological differences between pairs of sites as is typically measured in ecology using a dissimilarity metric. A wide range of different biological dissimilarity metrics could be considered, at almost any level of biological organization, notably compositional (taxonomic), phylogenetic, or genetic dissimilarity (Table 1). Because the response variable is a measure of biological distances between sites, the predictors in this model must also take the form of site pairwise distances, such as differences in temperature between sites, or the geographic distance between them. As different predictors are measured on different scales (e.g., temperature in degrees, precipitation in mm), they are transformed as part of model fitting, such that the transformed distance between

TABLE 1 Alternative ecological response data types that have been modelled using generalized dissimilarity modelling

Response	Data source	Preparation	Index	Example
Compositional dissimilarity	Community surveys	Calculate pairwise dissimilarities (automated in <i>formatSitepair</i>)	Typically Sorensen's, though many alternatives	Mokany et al. (2014)
	Presence-only occurrence data	Aggregate data to grid cells, select grid cells with sufficient records to constitute a community sample, then process as community survey data	As above	Fitzpatrick et al. (2013)
Phylogenetic dissimilarity	Community data and an associated phylogeny	As for compositional dissimilarity, but also incorporate a phylogeny in calculating dissimilarity	Phylogenetic Sorensen's	Rosauer et al. (2014)
Trait dissimilarity	Trait measurements	Calculate pairwise dissimilarity in morphological /functional traits	Functional beta diversity metrics	Thomassen et al. (2010)
Genetic dissimilarity	DNA sequencing	Calculate pairwise genetic distances	Population-level genetic distance (e.g., F_{ST})	Fitzpatrick and Keller (2015)
Observation pair difference	Presence-only occurrence data	Randomly sample pairs of presence-only observations, ensuring sufficient inclusion of observation pairs of the same species	Binary (0,1) indicating whether the two observations are from the same (0) or different (1) species	Ware et al. (2018)
Single species occurrence	Presence-only or presence-absence for a single species	Generate background absence for presence-only. Prepare all or a sample of occurrence pairs	Binary (0,1) indicating whether the two observations are both presences (0) or are a presence & absence (1)	Elith et al. (2006)

a pair of sites for different predictors can be meaningfully compared and combined.

GDM formulates the relationship between biological dissimilarity and environmental and/or spatial distance using GLM with a link function of the form:

$$d_{ij} = 1 - e^{-\eta} \quad (1)$$

where d_{ij} is the biological dissimilarity between sites i and j , and η is the transformed environmental distance between those sites, which has been termed 'predicted ecological distance' (Faith et al., 1987; Ferrier et al., 2007; η is called the linear predictor in GLM parlance). In virtually all applications of GDM to date, the link function applied has been a negative exponential, which makes the useful assumption that dissimilarity between site-pairs increases monotonically from 0 to 1, and in a saturating manner with predicted ecological distance (Figure 2). Model fitting is therefore focused on generating the best estimate of predicted ecological distance, which is achieved by appropriately transforming each predictor variable p .

The predicted ecological distance η is derived as the sum across all predictor variables of the absolute differences in the model transformed predictor values $f_p(x_p)$ between sites i and j in a pair:

$$\eta = b + \sum_{p=1}^n |f_p(x_{pi}) - f_p(x_{pj})| \quad (2)$$

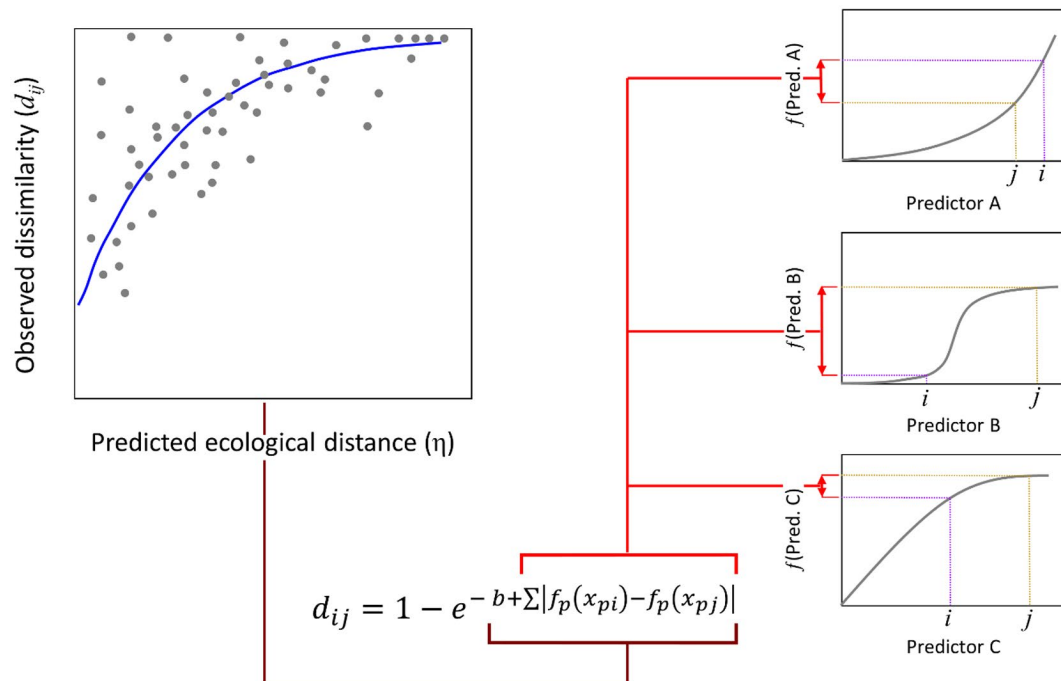


FIGURE 2 Conceptual depiction of the model fitting procedure of generalized dissimilarity modelling (GDM). To best represent the observed dissimilarity (d_{ij}) between sites i and j , GDM uses I-spline basis functions to transform each of the predictor variables (e.g., $f(\text{Pred. A})$), so that summed absolute difference in the transformed predictor values, combined with the model intercept (b) provide a predicted ecological distance (η). The negative exponential link function ensures predicted dissimilarities increase and saturate with increases in the predicted ecological distance

where b is the x-intercept added to account for a background level of observed dissimilarity when the values of all predictor variables are identical at sites i and j . In other words, b translates to the amount of dissimilarity expected between sites that have identical environments and therefore environmental distances between them equal to zero. To transform each predictor variable, GDM uses a linear combination of I-spline basis functions (Ramsay, 1988), fit using non-negative least squares regression (Figure 2 and Supporting Information Figure S1). Hence, the overall spline function $f_p(x_p)$ for each predictor variable is relatively flexible in shape, but constrained to increase monotonically. This constraint underlies a key assumption of GDM: that dissimilarity can only increase as two sites become more different in terms of the values of their predictor variables.

The degree of flexibility in transforming each predictor variable is determined through the number of I-spline basis functions applied (m_p). For each I-spline basis function for a predictor variable, a non-negative coefficient is fit:

$$f_p(x_p) = \sum_{k=1}^{m_p} a_{pk} l_{pk}(x_p) \quad (3)$$

where l_{pk} is the k th I-spline for variable x_p and a_{pk} is the fitted coefficient for l_{pk} . The position (or knot) of each I-spline basis function along the predictor variable axis is also flexible (Supporting Information Figures S1 and S2), with the default in the *gdm* R package being three knots, positioned at the minimum, median

and maximum of the observed values of a predictor. As with any GLM-based model, a variance function is also integral to the fitting process. Given that dissimilarities in GDM are constrained between 0 and 1 and can generally be described as proportions, a binomial variance function is used in model fitting (Dobson, 2001; Ferrier et al., 2007).

1.2 | Preparing the biodiversity data

The response data applied in GDM can be any dissimilarity or distance with values in the range 0 to 1 (inclusive; Table 1). Here we focus primarily on the common case of modelling dissimilarity in species composition between pairs of locations.

1.2.1 | Community survey data

The most suitable data for modelling compositional dissimilarity are from species community surveys undertaken using a consistent methodology over the region of interest (Ferrier & Guisan, 2006). Compositional data can be input in either a 'wide' format (site \times species table), a 'long' format (species \times observed occurrence location list) or through a precomputed site \times site dissimilarity matrix (using `gdm::formatsitepair`; see Supporting Information Appendix S1 – section 1.2). A precomputed dissimilarity matrix would be the required format for fitting genetic, phylogenetic, or functional distance metrics. A variety of alternative distance metrics can be specified (or used) to derive the pairwise compositional dissimilarities that will be modelled, including those based on abundance data. Sorensen's dissimilarity index (Sørensen, 1948) based on species presence-absence data (and equivalent to deriving the Bray-Curtis index for such data) has been commonly used in published studies (and is the default in `gdm::formatsitepair`).

As with any analysis, care is required in preparing the data. For community survey data, it is possible that alternative levels of survey effort (e.g., area or duration) have been used to record community composition in different sites, with sites surveyed less intensively being likely to result in artificially inflated site-pair dissimilarities. To minimize the impact of this inflation of dissimilarities, sites may be filtered so that they are only included if they have similar levels of survey effort, or the dissimilarities between site-pairs may be scaled based on the survey effort at each site (Mokany et al., 2014). Statistical methods to account for differential survey effort directly in GDM fitting through an offset (such as in GLM) have yet to be developed.

1.2.2 | Communities from presence-only data

While standardized community survey data are the most appropriate type of biological data to use in modelling compositional dissimilarity, for many regions and many taxonomic groups, such data

are non-existent or insufficient in spatial or environmental coverage (Amano et al., 2016). In such cases, building community composition data from aggregated presence-only data, such as widely available through online biodiversity databases like the Global Biodiversity Information Facility (GBIF), may be the only viable approach for modelling compositional dissimilarity (Bush et al., 2019; Fitzpatrick et al., 2011). To apply this approach, it is important to first acknowledge an important challenge presented by the use of presence-only data. In essence, the use of presence-only data for community-level modelling requires the assumption that any location where the species has not been observed can be treated as absence (Ferrier & Guisan, 2006). The validity of this assumption will vary with the number of presence-only observations in a location, with under-sampled locations being more problematic.

Because the response variable is a dissimilarity, incomplete sampling will generally artificially inflate pairwise dissimilarities between sites. In other words, the dissimilarity will increase simply because species have not been observed, not because they are absent. Biases associated with the artificial inflation of pairwise dissimilarities can be reduced by (a) removing sites with few species and/or (b) weighting site-pairs by richness of the component sites during model fitting. Using a coarser grained spatial grid is also helpful. In contrast to these spatial effects, systematic bias in the identities of species recorded in presence-only data (Troudet et al., 2017) may act to artificially reduce calculated dissimilarities, though this will depend on the distributional attributes of the species more commonly recorded.

The first step after obtaining the occurrence records for the focal group of species is to filter them based on their attributes (e.g., spatial uncertainty). The retained records then need to be aggregated to common spatial locations, which is simplest to undertake using cells of a spatial grid. If a coarse grid resolution is applied, records can simply be allocated the location of the centroid of the grid cell in which they occur. If a finer resolution grid is used, aggregating records to the grid cell they occur in may result in few cells with sufficient records to be considered a community sample, thereby exacerbating the artificial inflation of dissimilarities. In this case, it may be preferable to aggregate records based on a specified radius around each grid cell, carefully considering the spatial uncertainty of the record locations, and ensuring each record is assigned to only a single aggregate location (Bush et al., 2019).

The next important step is deciding which grid cells have sufficient species recorded to be considered as an adequate sample of the assemblage likely to occur in that area. A simple approach is to use a threshold minimum number of species recorded and to use all grid cells above that level (Fitzpatrick et al., 2011). Alternatively, sites may be weighted based on the number of observed species recorded (Williams et al., 2012; using 'weightType' in `gdm::formatsitepair`). These approaches assume that the background level of species richness is more or less constant across the region being analysed. If that is not the case, then a simple threshold may result in missing data from areas that are naturally less species rich (leading to spatial/environmental bias), and including under-sampled grid cells from species-rich areas (inflating compositional dissimilarities). If

spatially explicit information on species richness is available (e.g., a predicted species richness grid) then that may be used, otherwise it is possible to filter grid cells based on the most species-rich grid cell occurring within a specified distance (Bush et al., 2019). This approach assumes that at least one grid cell is relatively comprehensively sampled within the specified radius, which can be used as a proxy for richness. Once performed, this filtering will provide data that approximate community survey data, though with the caveats and deficiencies noted above.

1.2.3 | Other response data types

Phylogenetic dissimilarities between communities have also been used as a response variable in GDM (Rosauer et al., 2014). A key decision is the depth in the phylogeny to use in generating the phylogenetic dissimilarities, given different choices focus the analyses on different aspects of eco-evolutionary processes and biogeographic history. Phylogenetic GDM has substantial value to add to a purely compositional based analysis.

An increasing number of studies are using GDM to understand and predict genetic turnover within species (Fitzpatrick & Keller, 2015; Gougherty et al., 2021). Given well-established techniques exist for quantifying genetic dissimilarity between a pair of samples, model fitting and spatial analyses can be undertaken in a similar manner to a compositional GDM.

An alternative approach has also emerged recently in harnessing presence-only species observations (e.g., GBIF) to model compositional dissimilarity using GDM (Hoskins, Harwood, Ware, & Ferrier, 2020; Hoskins, Harwood, Ware, Williams, et al., 2020; Ware et al., 2018). This observation-pair (or 'obs-pair') approach models pairs of individual observations of species, where the response value is '0' if the two observations are of the same species, or '1' if the two observations are of different species. While the obs-pair approach to modelling compositional turnover has been trialled for modelling biodiversity at continental (Ware et al., 2018) and global scales (Hoskins, Harwood, Ware, & Ferrier, 2020; Hoskins, Harwood, Ware, Williams, et al., 2020), further work is required in developing techniques to account for the inherent biases in presence-only data, such as directly incorporating observer effort, and in more directly translating the resulting response variable into meaningful community-level metrics.

GDM has also been used to model the occurrence of individual species, based on occurrence observations. In this case the response variable is based on observation pairs for a single species, with a dissimilarity of '0' for a pair of presences, and a dissimilarity of '1' for a presence-absence pair. Under this application, the default negative exponential link function of GDM has little ecological justification, hence it is potentially more robustly applied using the GLM based fitting of GDM (Hoskins, Harwood, Ware, & Ferrier, 2020), where a more appropriate link function can be specified, such as the logistic link. An additional application of GDM to model the distributions of individual species involves

fitting a GDM to community compositional dissimilarity, then using the GDM transformed environmental layers as inputs, along with species occurrences, to kernel regression (Elith et al., 2006) or support vector machine analysis. One of the main benefits of this approach is that very few model parameters are required to predict the distribution of a species, potentially suiting the modelling of rarer species where there may be only a small number of occurrence observations.

1.3 | Preparing predictor data

As with other biodiversity modelling approaches, there are two broad modes of applying GDM, depending on the predictor data that are used – namely site-only and gridded. The first involves using one or more predictor variables that were measured directly, and only, at each survey location. This mode is generally applied with the main goal of better understanding the drivers of dissimilarity across a set of study sites (Jones et al., 2013). The second and more common mode of application involves only applying spatially complete gridded environmental data as predictors, which then enables projection of the model across the entire grid, and subsequent spatial analyses (Ferrier et al., 2007; Gibson et al., 2015). While providing information on potential drivers of dissimilarity, this mode of application also makes it possible to map predicted patterns of dissimilarity and undertake a wide variety of spatial biodiversity and conservation assessments. The trade-off in using spatially complete gridded environmental variables as predictors is that they generally will not be as accurate or have as direct a relationship to species biology as field measurements in quantifying a given variable of interest (Johnston et al., 2017), potentially resulting in models with lower explanatory power and greater uncertainty.

While predictor variables are typically a single value for each site, the nature of modelling pairwise dissimilarities as the response variable with GDM also enables the use of customized pairwise dissimilarities as predictor variables. One possible implementation of dissimilarities as predictors would be to use the compositional dissimilarity of one group (e.g., trees) to help predict compositional dissimilarity for another group (e.g., ferns) (Jones et al., 2013; van Schalkwyk et al., 2019). Another application would be to use customized geographic distance variables (e.g., least cost paths) instead of Euclidean distance between sites (the default approach) to account for the effect of spatial separation on observed dissimilarities (Thomassen et al., 2010).

Continuous data are preferred as predictors in a GDM, though ordinal categorical data may be applied, with more categories preferable to fewer. Nominal categorical predictors, such as geological types, are generally unsuitable for use in GDM, and even if converted to a customized binary distance value (1 = same category; 0 = different categories), they may create problems in fitting and interpreting a GDM. This is particularly the case where each category is not equally different, or where gridded model predictions are required.

Transforming predictor variables prior to model fitting is commonly undertaken in biodiversity modelling, though its value when applying GDM may be limited. This is due to automated variable transformation being a key component of the model fitting procedure. To date there has been no systematic assessment of the effects of predictor variable transformation on GDM fitting. However, in the Supporting Information (see Appendix S1 – section 3.1) we demonstrate how transforming predictor variables that are highly skewed can provide the GDM fitting procedure with more flexibility and help to explain more deviance in the response data.

1.4 | Preparing site-pairs for modelling

Fitting a GDM requires first preparing the input data into a specific format, with the predictors provided as either a site \times predictors table, raster layers of gridded predictor data, one or more site \times site distance matrices (using *gdm::formatsitepair*). The format of the input table for GDM has been developed to enable modelling under a wide variety of possible circumstances, importantly allowing the option to fit models without using all possible site-pairs. Each pairwise dissimilarity value being included in the model is represented by a row in the input table, with the columns being: observed dissimilarity, a weight, x - and y -coordinates of the first site in the site-pair, x - and y -coordinates of the second site in the site-pair, followed by each of the predictor variables for site one, and each of the predictor variables for site two (in the same order as for site one; Figure 3).

It is useful to understand the distribution of biological dissimilarity values (such as through plotting a histogram), as this will influence the capacity of the selected predictor variables to reflect expected patterns of dissimilarity and could reveal inflated dissimilarities due to sampling issues (e.g., if many of the values are near 1). To aid in model convergence, we have found that it can be beneficial in some cases to re-scale the dissimilarities, most notably for genetic distances (i.e., F_{ST}) that commonly have values $\ll 1$. The weight column in the input table allows observed dissimilarity values to be given different levels of importance in the model fitting procedure. These weights also enable implementation of the Bayesian bootstrap GDM (Woolley et al., 2017), which helps to clarify the uncertainty around the model response functions, and to account for the inflated degrees of freedom in a standard GDM, by returning the model to the correct amount of information in the observed data (rather than the pairwise dissimilarities).

One important consideration in preparing pairwise input data for modelling with GDM is whether all possible site-pairs should be used (the default in *gdm::formatsitepair*), or a subsample of possible pairs. Given that using the same site in multiple pairwise comparisons introduces non-independence in the data being modelled (Smouse et al., 1986), subsampling site-pairs would limit the degree to which this assumption of the underlying GLM methodology is being violated (Supporting Information Appendix S1 – section 3.2). Subsampling site-pairs may also be a computational necessity in situations where there are many sites. The number of possible pairwise comparisons increases in a n -squared manner with the number of sites (i.e., 100k sites results in c. 5 billion pairs), hence increasing numbers of site-pairs will eventually exceed the RAM capacity of most computers.

both sites		site 1		site 2		site 1				site 2			
distance	weights	s1.xCoord	s1.yCoord	s2.xCoord	s2.yCoord	s1.isotherm	s1.radiation	s1.min_temp	s1.precip_evap	s2.isotherm	s2.radiation	s2.min_temp	s2.precip_evap
0.39	1	147.76	-40.06	147.88	-40.32	0.45	24.27	6.82	0.57	0.46	23.65	6.61	0.56
0.79	1	147.76	-40.06	148.18	-40.54	0.45	24.27	6.82	0.57	0.45	23.16	6.89	0.57
0.91	1	147.76	-40.06	144.84	-40.73	0.45	24.27	6.82	0.57	0.49	23.71	6.71	0.90
1.00	1	145.04	-40.92	147.12	-41.45	0.51	23.04	5.31	1.10	0.50	22.96	2.30	0.61
0.84	1	145.04	-40.92	148.26	-41.45	0.51	23.04	5.31	1.10	0.52	22.12	3.80	0.64
0.94	1	145.04	-40.92	145.69	-41.46	0.51	23.04	5.31	1.10	0.47	20.65	1.56	2.13
0.75	1	147.88	-40.32	148.18	-40.54	0.46	23.65	6.61	0.56	0.45	23.16	6.89	0.57
0.89	1	147.88	-40.32	144.84	-40.73	0.46	23.65	6.61	0.56	0.49	23.71	6.71	0.90
0.71	1	147.88	-40.32	144.72	-40.86	0.46	23.65	6.61	0.56	0.49	23.53	6.74	0.97
0.74	1	148.18	-40.54	144.84	-40.73	0.45	23.16	6.89	0.57	0.49	23.71	6.71	0.90
0.61	1	148.18	-40.54	144.72	-40.86	0.45	23.16	6.89	0.57	0.49	23.53	6.74	0.97
0.74	1	148.18	-40.54	148.14	-40.87	0.45	23.16	6.89	0.57	0.46	23.15	6.82	0.61
0.65	1	144.84	-40.73	144.72	-40.86	0.49	23.71	6.71	0.90	0.49	23.53	6.74	0.97
0.73	1	144.84	-40.73	148.14	-40.87	0.49	23.71	6.71	0.90	0.46	23.15	6.82	0.61
0.83	1	144.84	-40.73	145.04	-40.92	0.49	23.71	6.71	0.90	0.51	23.04	5.31	1.10
0.66	1	144.84	-40.73	144.67	-40.95	0.49	23.71	6.71	0.90	0.48	23.33	6.60	1.06
0.77	1	144.84	-40.73	147.39	-40.99	0.49	23.71	6.71	0.90	0.50	23.58	4.82	0.64

FIGURE 3 Illustration of the input data format for GDM fitting, based on the example data for vascular plants in Tasmania (Mokany et al., 2012). Each site-pair is represented as a row. The first six columns are consistently applied, being the observed dissimilarity to be modelled, the weighting to apply to each site-pair, and the spatial coordinates of both sites in the pair (site '1' and site '2'). Subsequent columns vary with the predictor variables selected, but the values for all the predictors for the first site in the pair are specified, then the values for the same predictors for the second site in the pair. This input table format provides a large degree of flexibility in terms of customizing the site-pairs used in model fitting, and allows incorporation of different types of predictors into a single input table

While randomly selecting site-pairs will avoid computational constraints, it will retain the basic attributes of the full dataset, which may be undesirable in some situations. For example, where a dataset covers a very broad environmental or spatial extent or where beta diversity is very high, it may contain a very high proportion of site-pairs with complete dissimilarity ($d_{ij} = 1$). These data will become a dominant force in the model fitting, at the expense of non-complete dissimilarities ($d_{ij} < 1$), where much of the ecological interest often lies. Subsampling that includes a higher proportion of site-pairs from more similar environments or spatial locations will tend to result in a higher proportion of non-complete dissimilarities, though while this will help to better characterize lower values of dissimilarity it may not be reflected in a greater model deviance explained (Supporting Information Appendix S1 – section 3.3). Further assessment of site-pair sampling strategies for GDM would be valuable, as would generic functions for implementing those alternative site-pair sampling strategies.

1.5 | Model fitting

GDM fitting will be based on all the predictor variables included in the input data table. Geographic distance is an optional additional predictor, and if used, it will be incorporated as the Euclidean distance between the two sites in a pair, based on the provided x- and y-coordinates. There are important ecological mechanisms under which geographic distance may directly and independently result in dissimilarity between site-pairs, including neutral processes (Hubbell, 2001) and biogeographic history (Lu et al., 2019). If geographic distance is used, it is preferable that the x- and y-coordinates are in a projected coordinate system, as using geographic coordinates (degrees latitude and longitude) typically introduces disparity between x and y axis distances.

Given that the link and variance functions are fixed in the R-based GDM fitting function, the primary choice when fitting a model is whether to use the default number and position of I-spline basis functions for each predictor, or whether to specify alternative values for these (Supporting Information Figures S1 and S2). For many ecological applications, three splines for each predictor appears to offer a reasonable degree of flexibility in capturing nonlinear change in dissimilarities along a predictor gradient, while remaining ecologically interpretable. In the Supporting Information (Appendix S1 – section 3.4) we demonstrate how increasing the number of splines for a predictor can influence model performance assessed through cross-validation. This suggests that specifying more splines for a predictor may be warranted in certain cases, though it can risk overfitting and loss of ecological interpretability (Supporting Information Figure S2; Appendix S1 – section 3.4).

1.6 | Model interpretation

There are a number of metrics generated in fitting a GDM that can help in interpreting how well the model explains the observed

dissimilarities (shown using *gdm::summary*). The deviance explained is a key metric, being the percent deviance (or variation) in the observed dissimilarities that is explained by the model, which characterizes how well the model predicts the data used to train it. Typical percent deviance explained values for models of compositional dissimilarity range from 20 to 50%, and as with most modelling approaches, increasing the number of observed data points tends to reduce the deviance explained. Including a larger proportion of complete dissimilarities ($d_{ij} = 1$) will often increase the amount of deviance explained, though the model may actually be less informative for lower values of dissimilarity (Supporting Information Appendix S1 – section 3.4).

The model intercept is also informative, indicating the mean dissimilarity between site-pairs when they have identical predictor values (i.e., environmental distance between sites equals zero). Higher values of an intercept infer relatively large dissimilarities, even where predictor variables for two sites are identical, which may be a real phenomenon derived through stochastic ecological processes, or may be an artefact of incomplete sampling and/or missing environmental predictor variables. Changes in the value of the intercept for alternative models using a given dataset may be informative of how well low levels of dissimilarity have been characterized.

A summary of the position and predicted spline coefficient for each knot for a predictor variable provides basic information on which predictors are most important in explaining observed dissimilarity. Predictors with the greatest influence on predicted dissimilarity have the greatest summed coefficient values.

Graphical representation of a GDM can also aid in interpreting the nature of a fitted model (using *gdm::plot*; Supporting Information Appendix S1 – section 1.5). A plot of observed dissimilarities (y axis) as a function of model-derived ecological distance (x axis) shows each site-pair as a point, with a line representing the fitted prediction of dissimilarity as a function of ecological distance (Figure 4a). A plot of observed versus model-predicted dissimilarities is also informative, with each point being a site-pair, and a line of equality showing perfect model prediction (Figure 4b). For both plots, deviation of observed values from the model predicted values (lines) provides information on the predictive capacity of the model.

A plot of the spline function for each predictor variable shows the continuous representation of the fitted coefficients. These plots for each predictor (Figure 5) show the predictor variable on the x axis, and the model transformed value of that predictor on the y axis. Spline functions that reach a higher maximum transformed value (the sum of the fitted coefficients) play a stronger role in influencing predicted dissimilarity (Figure 5) and therefore a greater role of that predictor in patterns of beta diversity. A greater slope of the spline function at a given point along the predictor variable axis indicates more rapid increase in dissimilarity (and hence greater rate of turnover in composition) when the predictor values for a pair of sites straddle that point (Figure 6).

Assessment of the parameter uncertainty associated with a fitted GDM can be obtained using a Bayesian bootstrap (Woolley et al., 2017, using the *bbgdm* R package: Woolley et al., 2015). This

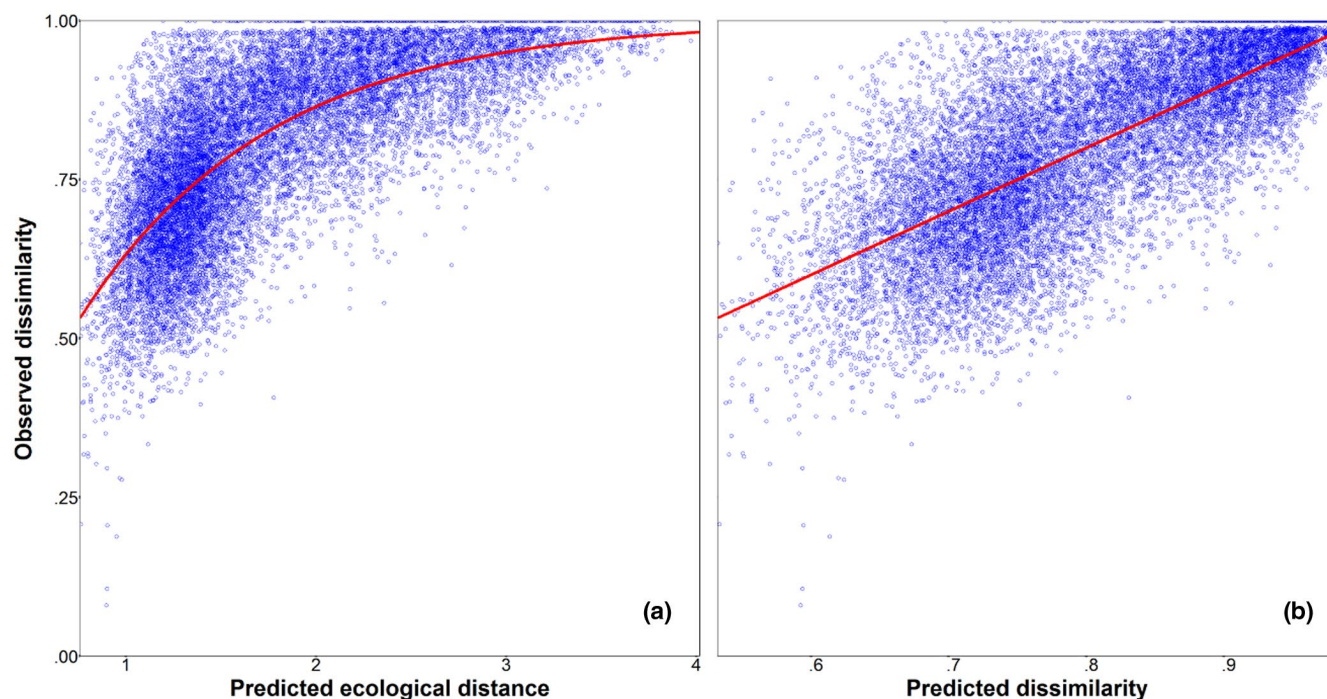


FIGURE 4 The main plots for a GDM, here based on an example for vascular plant community compositional dissimilarity across Tasmania (Mokany et al., 2012). (a) Observed dissimilarity as a function of GDM-predicted ecological distance, with each site-pair represented as a point, and the line representing the GDM-predicted dissimilarity. (b) Observed dissimilarity as a function of GDM-predicted dissimilarity, with the line of equality provided. This GDM applies all 14,535 site-pairs from 175 sites with plant compositional data, using four environmental predictors (Figure 5), explaining 53.0% deviance in observed dissimilarities, with a model intercept of 0.76

approach fits many models using the same observed dissimilarities and predictor data, but alternative weights for each site-pair, derived from a Dirichlet distribution (Rubin, 1981). Plots generated from this analysis show the confidence intervals around the fitted spline function for each predictor.

Assessing the statistical significance of a full GDM, or the individual predictors in a GDM, can be undertaken using a permutation test (using *gdm::gdm.varImp*). This procedure compares the deviance explained from the fitted model to that derived when either an individual predictor variable, or all predictor variables, are randomly permuted between sites (Fitzpatrick et al., 2013; Mokany et al., 2014). The resulting *p*-value is the proportion of models where the randomized predictor data perform better in explaining deviance than the actual predictor data. Using a specified threshold maximum *p*-value (typically .05) can indicate which variables are not significant in the model, helping inform model refinement and simplification. Note that a similar permutation procedure could be used to test for the significance of using more than three I-spline basis functions, though this is not currently implemented in the *gdm* R package, nor are we aware of any study that has done so.

Cross-validation is another important tool that can be used in assessing the accuracy of a model (using *gdm::gdm.crossvalidation*). This analysis randomly removes a specified proportion of sites from the data used to train the GDM, then applies these sites to test how well the model predicts dissimilarities on independent data. Given this is a random procedure, it is typically repeated multiple times, and

summarized across these randomizations. A variety of metrics can be used to assess model predictive accuracy under cross-validation, including deviance explained (in the test data), mean absolute error (MAE), and root mean square error (RMSE). It is also possible to consider how model predictive accuracy varies with the observed dissimilarity value, using a customized metric, equalized-RMSE, that averages RMSE over equal bands of observed dissimilarity (returned by *gdm::gdm.crossvalidation*). Model error in a GDM is typically greater for lower values of dissimilarity.

It may also be informative to assess the independent and shared contribution of different types of predictors in explaining observed dissimilarity. For example, climate, substrate, topography and geographic distance may play different roles in explaining dissimilarity, with varying degrees of overlap in their explanatory power. Given the additive nature of the predictors in a model contributing to predicting dissimilarity, combined with correlation of spatial predictors, a variance partitioning analysis can be a useful way to better understand contributions of variable sets (Jones et al., 2013; using *gdm::gdm.partition.deviance*).

1.7 | Model selection

Given that data may be available for a substantial number of potential predictor variables, refining this to a suitable subset requires model selection. Undertaking model selection for GDM shares the

same issues as other modelling approaches, such as generalized additive modelling (GAM). Given the additive nature in which predictor variables are combined in GDM, the order in which variables are included in the model (through the data input table) is irrelevant. Backward elimination variable selection is typically applied, though once all predictor variables remaining in the model are statistically significant, it is at the discretion of the user as to whether further variable refinement is undertaken in the pursuit of parsimony, potentially including the use of the Akaike information criterion (AIC; Akaike, 1973; Supporting Information Appendix S1 – section 3.6).

As with many regression modelling approaches, the most suitable approach to dealing with correlation between predictor variables in a GDM requires careful consideration. For other modelling approaches, the presence of strong linear correlation between predictor variables may violate the assumption of independence in the data used for modelling. For GDM, it is really the correlation between site-pair distances of predictor variables that might be of concern. Hence, it is worthwhile exploring correlation between predictors and using this information to inform variable selection. While standard linear correlation between predictor variables may be useful in selecting sets of predictor variables for consideration in modelling, the implications for modelling of site-pair dissimilarities requires further statistical assessment (Supporting Information Appendix S1 – section 3.5). Correlation between predictor variables can also create unforeseen repercussions in multivariate models such as GDM, as predictor variables that perform only modestly by themselves can become strong relative contributors to a multivariate GDM.

1.8 | Predicting dissimilarity

Once a GDM has been fitted, it can be used to predict dissimilarities between pairs of sites. Predictions from a GDM can be made in two main ways. The first approach is to tabulate predictor data for target site-pairs in an identical format to the input data used to fit the model, with predictions of dissimilarity being made only for those specific site-pairs (using *gdm::predict.gdm*). This method is commonly used in cross-validation analysis of a GDM, or where there is interest in making predictions for only a relatively small number of site-pairs, as making predictions for a large number of site-pairs could be computationally prohibitive.

The second approach to making predictions from a GDM is applied where the predictor data are available as spatial grids, and where predicting dissimilarities for every possible pairwise combination of grid cells is either undesirable or infeasible. In this situation, prediction from the GDM is undertaken in two steps. First, the grids for the predictors used in the GDM are transformed from their native units using the fitted spline function for each predictor (using *gdm::gdm.transform*; Supporting Information Figure S3). Then the set of transformed grids for all predictors can be used to predict the dissimilarity between any pair of grid cells, using the model formula (Equation 1; Supporting Information Appendix S1 – section 2.3). This second step of making predictions of site-pair dissimilarities

from transformed predictor grids is undertaken as desired by the researcher (independent of the *gdm* R package).

Predicting dissimilarities under climate change, through space-for-time substitution, can also be implemented simply by transforming alternative climate layers using the same model (and its coefficients). When generating GDM transformed spatial layers, either for the modelled present-day environment or for alternative climate futures, it is important to consider extrapolation beyond the range of the predictor data used to generate the model. Spatial or temporal extrapolation may be relevant for any predictor variable, and is currently performed by continuing the average slope of the spline function within the 10th quantile (for extrapolation below the minimum observation) or within the 90th quantile (for extrapolation beyond the maximum observation) of the data range used in model fitting. As with other spatial modelling approaches, users are also able to control extrapolation by transforming the input grids used in the model prediction step, such as implementing 'clamping' by setting all gridded values that are greater than the maximum used to fit the model to that maximum value.

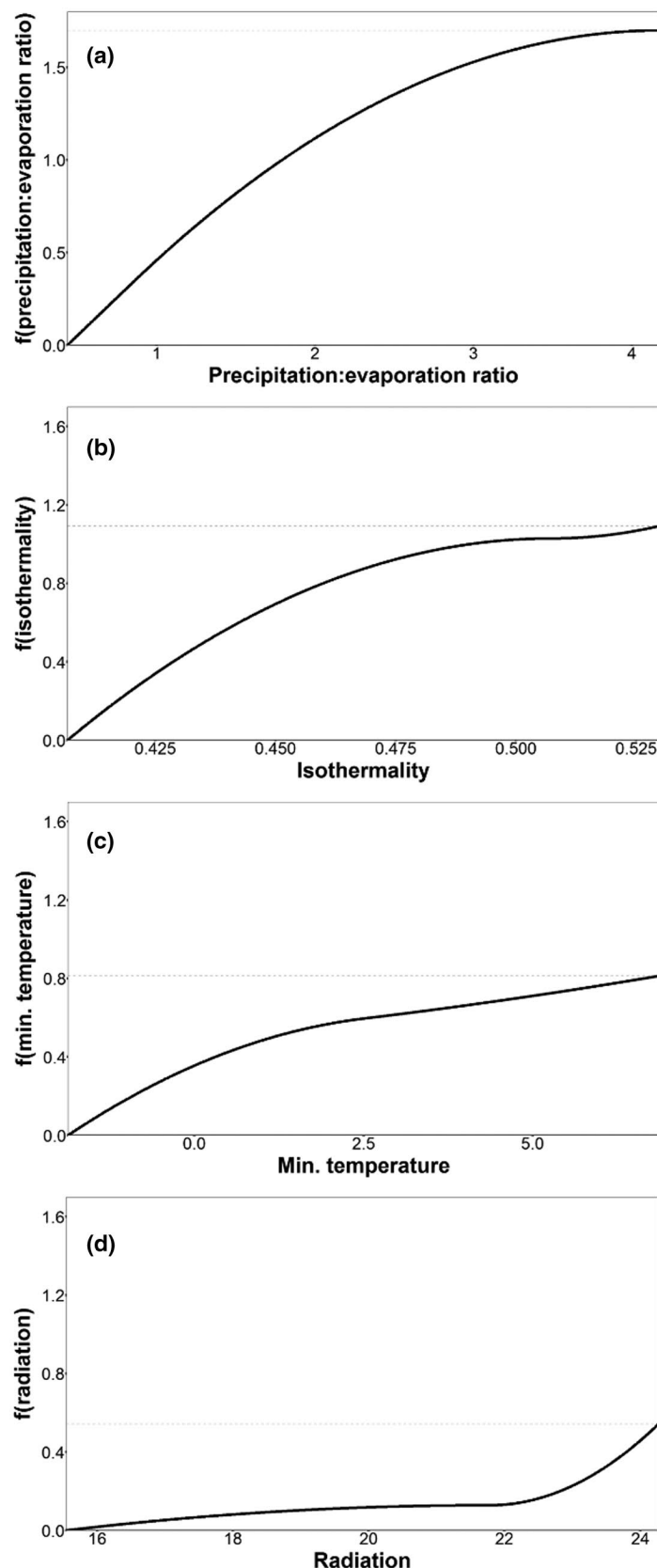
Finally, a special approach is required to deal with spatial predictions of dissimilarity where geographic distance is included as a predictor variable. This is because it is not possible to generate a single spatial layer that incorporates the predicted nonlinear effect of geographic distance on dissimilarity. The most common (default) approach applied to date has been to simplify the spline function for geographic distance to a linear function, and then translate this into two spatial grids that when combined approximate the geographic distance effect: one for the x-dimension and one for the y-dimension. The assumptions inherent in this approach will become more tenuous as the spline function for geographic distance becomes less linear. In such cases, calculating the model predicted transformed ecological distance for each pair of grid cells assessed is a more rigorous approach, though requires more computational resources.

1.9 | Spatial analyses

After using a fitted GDM to generate transformed spatial layers for each predictor variable in the model, a wide variety of spatial biodiversity analyses may then be undertaken. The simplest technique for spatially visualizing a GDM does not require calculation of dissimilarities, but rather involves mapping the first three axes derived from a principal component analysis (or similar data reduction technique) performed on the set of transformed predictor layers. A red, blue and green colour dimension is assigned to each axis and mapped by combining the colours (Supporting Information Appendix S1 – section 2.2). This simplification of multiple GDM transformed spatial layers down to three principal components can be a useful way to visualize areas that are predicted to be more similar (shown as similar colours) or less similar (shown as different colours; Figure 7a).

Most spatial analyses, however, are based on predicted dissimilarities from the GDM. The core calculation for such analyses

FIGURE 5 GDM spline functions for each predictor variable, for the GDM for vascular plant community compositional dissimilarity across Tasmania (Mokany et al., 2012; Figure 4). The maximum height of the spline function (hence the maximum value of the transformed predictor), shown with a pale dashed line, indicates the importance of the predictor in explaining dissimilarities. Above, the precipitation : evaporation ratio (a) is the strongest predictor of the observed dissimilarities, while radiation (d) is the least important predictor variable



involves using the transformed values of all predictor variables to determine the predicted dissimilarity between a pair of cells on the spatial grid (Equation 2; Supporting Information Appendix S1 – section 2.3). The challenge remaining is then to devise meaningful

spatial analyses involving many pairwise predictions of the dissimilarity between grid cells, to address a specific question. The computational resources required for an analysis will therefore scale with the number of grid cell pairs considered.

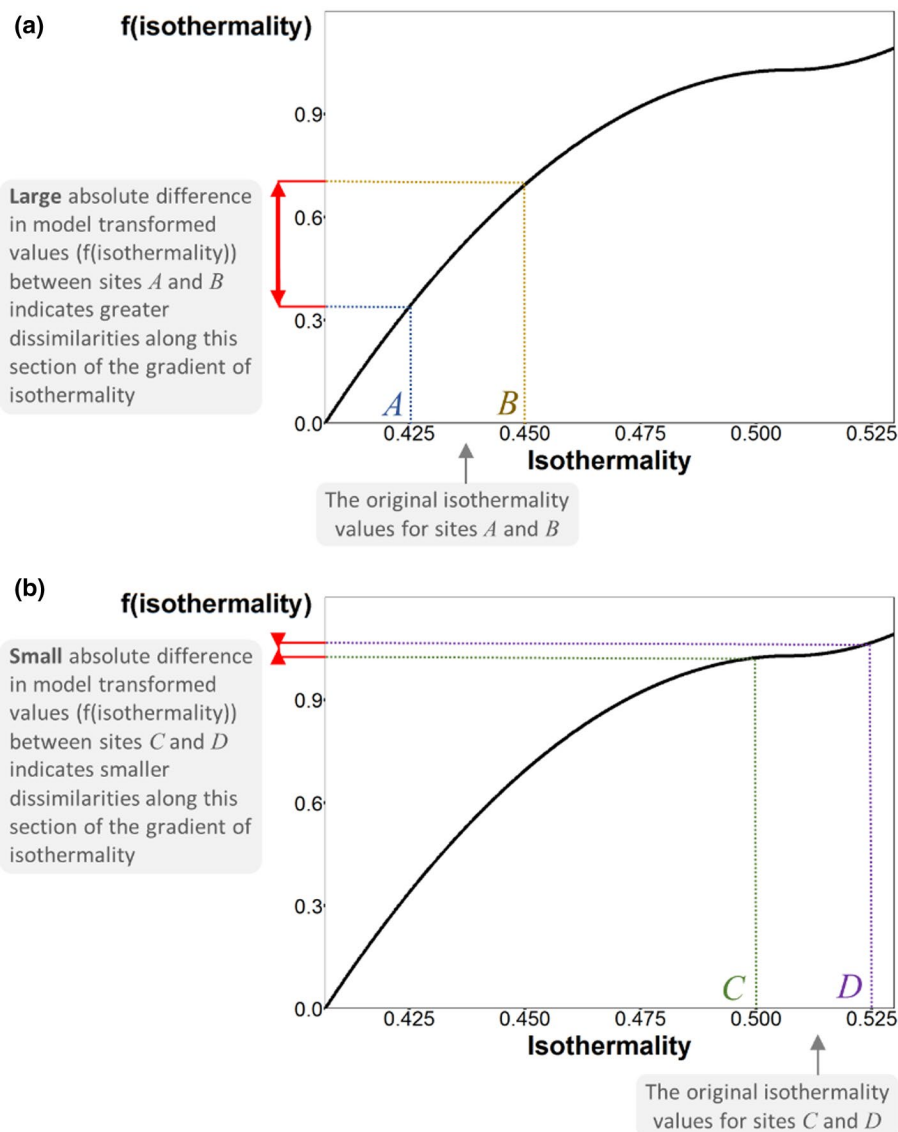


FIGURE 6 Implications of the slope of a GDM spline function for predicted pairwise dissimilarity, here using the spline function for one predictor (isothermality) from the Tasmanian vascular plant example (Mokany et al., 2012; Figures 4 and 5b). (a) Sections of the spline function with steeper slope imply greater dissimilarity per unit change in the predictor variable along that section of the gradient, illustrated by two sites, *A* and *B*. (b) Sections of the spline function— with flatter slope imply less dissimilarity per unit change in the predictor variable along that section of the gradient, illustrated by two sites, *C* and *D*. Sites *A* and *B* have the same difference in isothermality as do sites *C* and *D*, yet the absolute differences in the GDM transformed values for isothermality ($f(\text{isothermality})$) between sites *A* and *B* are much greater than between sites *C* and *D*, indicating greater predicted dissimilarity

Perhaps the simplest GDM-based spatial analysis assesses the similarity to a focal point, which aims to understand how similar every location is to a specific place of interest. It involves specifying a grid cell of interest, then predicting the similarity (the complement of dissimilarity) of every cell on the spatial grid to that focal cell (Figure 7c; Supporting Information Appendix S1 – section 2.4). This analysis can be used in various ways, such as to help visualize and interpret GDM spatial predictions (Overton et al., 2009) or to identify potential areas to restore and protect in order to offset the loss of habitat due to development activities in a particular focal location.

Another simple assessment is to consider the mean similarity within a neighbourhood, which aims to understand which places

have the most variation in the area around them. It assesses the mean similarity of all pairwise cell combinations within a specified radius around each cell on the grid. This analysis can focus on either the similarity of cells in the specified radius to the focal cell, or the similarity of all possible pairs of cells within the specified radius (Figure 7d; Supporting Information Appendix S1 – section 2.5). Depending on the neighbourhood applied, this analysis could identify locations potentially harbouring locally rare species, or indicate regions likely to support more species (larger γ -diversity), due to high average compositional dissimilarity (β -diversity) (Whittaker, 1960).

A more generic GDM-based spatial assessment that has a range of possible applications is to consider the mean similarity to

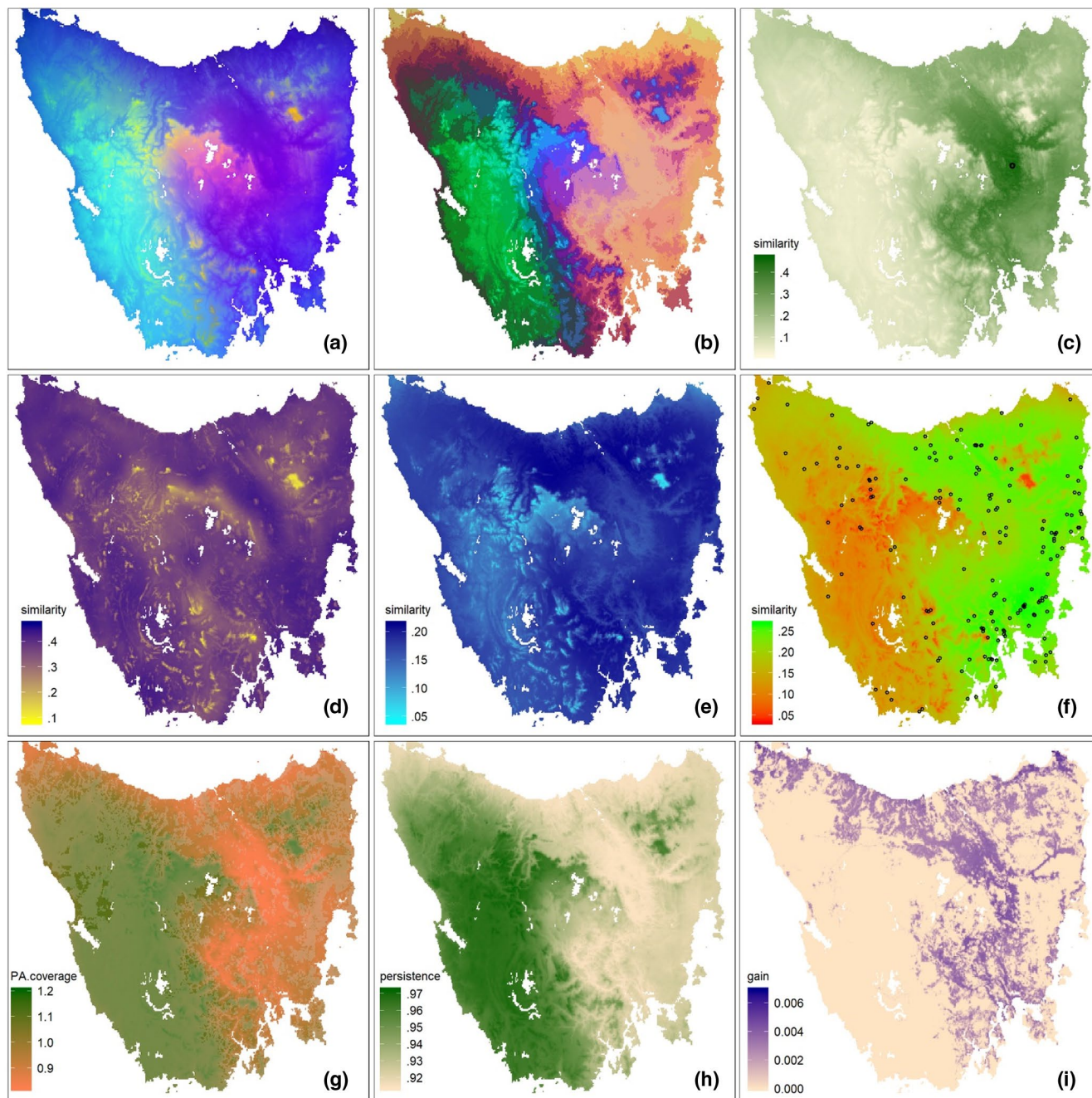


FIGURE 7 Examples of generalized dissimilarity modelling (GDM)-based spatial analyses for Tasmania, Australia, based on a GDM for vascular plant compositional dissimilarity (Mokany et al., 2012; Figures 4 and 5). (a) Primary differences in predicted community composition based on a principal component analysis (PCA) of the GDM-transformed environmental predictors, where more similar colours indicate more similar expected composition. (b) Community classification based on GDM, with similar colours indicating community types with similar composition. (c) Predicted compositional similarity to the focal point shown. (d) Local turnover – average compositional similarity within a 15-km radius. (e) Uniqueness – average similarity in relation to the entire region. (f) Survey gaps – relative compositional similarity to the survey locations shown. (g) Protected area (PA) representativeness – relative similarity to the protected areas, shown with a semitransparent mask. (h) Expected species persistence – the proportion of species originally occurring in each location expected to persist anywhere in the region over the long term given habitat loss/modification to date. (i) Marginal gain – the expected increase in species persistence for each location if all remaining habitat in that location was restored to maximum habitat condition. See Supporting Information Appendix S1 for example code implementing each analysis

a specified set of reference cells. This involves specifying a set of n reference cells j , then calculating for each cell i on the spatial grid, its mean similarity to those reference cells ($S_i = \sum s_{ij}/n$). Where the

reference cells are a random selection from the spatial grid, this analysis will determine the compositional uniqueness of each location (U_i) (Figure 7e; Supporting Information Appendix S1 – section 2.6). If

the reference cells are the locations where biological surveys have been undertaken, this analysis can be used to assess gaps in the environmental coverage of those surveys (Guerin et al., 2021), in which case the mean similarity to the survey locations should be divided by the uniqueness of each location, to account for the inherent rarity of certain environments (i.e., S_i/U_i) (Figure 7f; Supporting Information Appendix S1 – section 2.7). Alternatively, the reference locations may be a random sample of cells from within the protected areas of the region, in which case the representativeness of those protected areas can be assessed, again dividing the mean similarity of each grid cell to the reference cells by the uniqueness of that cell, to account for its inherent level of rarity (Allnutt et al., 2008; Figure 7g; Supporting Information Appendix S1 – section 2.8).

While one of the main attractions of applying GDM and associated spatial analyses is its capacity to capture continuous variation in community composition, for some purposes a discrete view of community composition is desired. Using GDM to classify and map community types may be particularly attractive, given its objective data-driven methods and the inherent scaling of environmental space to be biologically meaningful (Leathwick et al., 2011). To undertake a GDM-based classification, a sample of grid cells can be selected for 'training', the dissimilarities between all training grid cells calculated, hierarchical clustering performed on the training grid cells to generate the desired number of classes, then every grid cell can be allocated to the class of the most similar training grid cell (Figure 7b; Supporting Information Appendix S1 – section 2.9).

Using a fitted GDM to derive transformed environmental grids at two or more time points enables an even greater range of potential analyses. The simplest of these is to predict the similarity of each grid cell to itself over time, such as from the present to the future, or from the past to the present. This analysis therefore predicts which locations are expected to change most in composition over time under specific changes in environmental conditions (Fitzpatrick et al., 2011; Gougherty et al., 2021; Figure 8a; Supporting Information Appendix S1 – section 2.13).

In implementing the above GDM-based spatial analyses, it may be desirable to set the model intercept to zero when calculating the predicted dissimilarity between a pair of grid cells. The rationale for such an approach is that the model intercept may be primarily a result of incomplete sampling and/or missing predictors, and hence using it in prediction will result in higher dissimilarity between pairs of grid cells than the real level in nature. However, in some cases there may be a naturally high level of dissimilarity between two locations with very similar environmental conditions, such as where there are important stochastic processes, and/or where the grid resolution is fine. For temporal predictions, applying the model intercept in predicting dissimilarities will result in an assumed difference in composition for a grid cell over time, even if environmental conditions remain constant or the time step is very short.

Another important consideration in undertaking GDM-based spatial analyses is the computational resources required. For spatial grids that have a broad extent and/or fine resolution, there may be an extremely large number of grid cell pairs for which dissimilarity

could be calculated. Hence, depending on the analysis, calculating the predicted dissimilarity for the specified grid cell pairs could take substantial time, or require parallelization across many compute cores. These computational overheads are best dealt with through strategic planning of the desired analysis to reduce the number of pairwise grid cell comparisons required, targeted parallel processing, and potentially harnessing different coding languages/environments (e.g., python, C++).

1.10 | Other analyses based on generalized dissimilarity modelling

We have described some of the more commonly applied GDM-based spatial analyses above, which are directly implementable via existing functions in the *gdm* package, or through R code provided in the vignette associated with this paper. However, there are a variety of other more customized GDM-based analyses that have been implemented and which may be useful for different applications.

GDM spatial predictions can be used to assess the impacts of habitat change on biodiversity persistence. This analysis combines GDM predicted pairwise similarity between grid cells with a spatial layer indicating the habitat condition for every grid cell, ranging from 1 (pristine) to 0 (no value for biodiversity). This is used to estimate the proportion of species expected to persist across the whole region, given a change in habitat condition, additionally incorporating species richness predictions for all grid cells where available (Figure 7h; Supporting Information Appendix S1 – section 2.10). This approach has been used to assess the implications of habitat loss in Madagascar (Allnutt et al., 2008) and the world (Ferrier et al., 2004; Di Marco et al., 2019), and is a demonstration case-study for biodiversity accounting under the United Nations System of Environmental-Economic Accounting – Ecosystem Accounting (SEEA EA) framework (Mokany, Harwood, & Ferrier, 2019; UNCEEA, 2021; UNEP-WCMC, 2016). A global implementation of this same basic approach is also now endorsed by the Biodiversity Indicators Partnership (BIP) as a primary biodiversity indicator – the Biodiversity Habitat Index (<https://bipdashboard.natureserve.org>).

Using a similar analytical approach to that described above for assessing biodiversity persistence, it is possible to assess for each grid cell, the marginal change in biodiversity persistence following change in habitat condition. This analysis can assess both the marginal benefit for biodiversity of improving the habitat condition in a location (Figure 7i; Supporting Information Appendix S1 – section 2.11), or the marginal impact on biodiversity of reducing the habitat condition in a location. For example, a marginal benefit assessment can help to highlight priorities for habitat protection or restoration (Drielsma et al., 2012). However, this marginal benefit/loss analysis is contingent on the spatial configuration of 'current habitat' provided, hence assessments of actions in multiple cells would require an iterative cell-by-cell analysis.

GDM has also been combined with data on habitat condition to identify areas of high-value biodiversity habitat for the world's

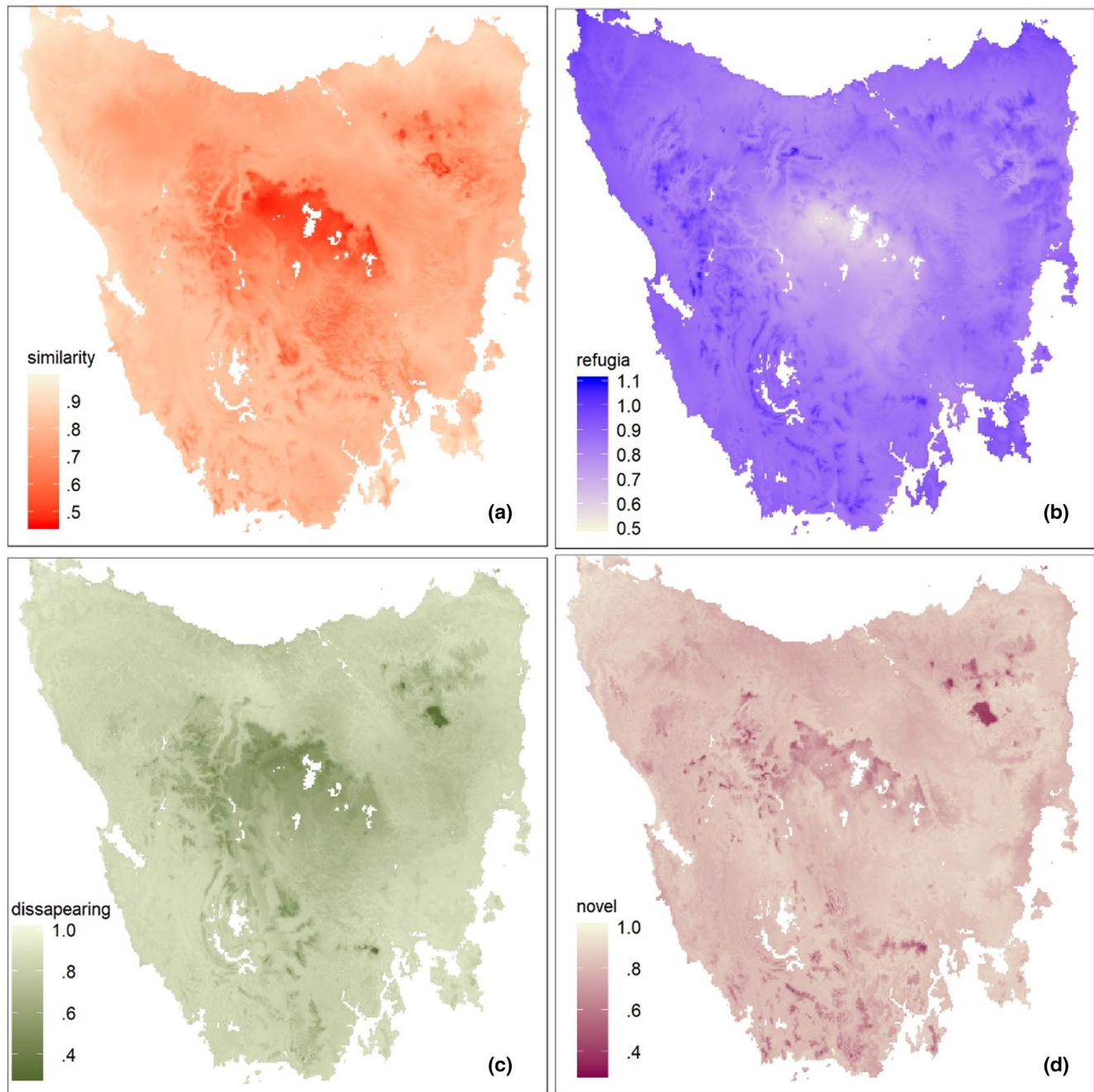


FIGURE 8 Examples of generalized dissimilarity modelling (GDM)-based spatial analyses that consider climate change, for Tasmania, Australia, based on a GDM for vascular plant compositional dissimilarity (Mokany et al., 2012; Figures 4 and 5). (a) Predicted compositional similarity for each location between current and future climates (darker = more change). (b) Predicted value of each location as refugia in the future for current biodiversity within the surrounding 15 km (darker = greater refugia). (c) Disappearing biological environments, being locations with low predicted similarity to any locations in the future (darker = more disappearing). (d) Novel biological environments, being locations in the future that have low similarity to any location in the present (darker = more novel). See Supporting Information Appendix S1 for example code implementing each analysis

terrestrial biota. This analysis of 'contextual intactness' identifies areas in better condition than most other locations that are predicted to have once supported a similar assemblage of species (based on GDM predictions; Mokany et al., 2020). The areas of high contextual intactness can be found within both intact regions as well as human-dominated landscapes.

In the context of a changing climate, GDM has been used to help identify likely areas of refugia under both past and future climates (Mokany et al., 2017; Reside et al., 2013). The basic principle is to identify areas that in the future or past have environments that are most similar to habitats that today support important biodiversity features (Figure 8b; Supporting Information Appendix S1 – section

2.14). These analyses may also incorporate the connectivity between those important environments over space and time.

Broader resilience of biodiversity in the face of habitat loss and climate change has been assessed by combining GDM predictions under current and future climates with data on habitat configuration and analyses of connectivity through least-cost paths (Ferrier et al., 2020). The resulting Bioclimatic Ecosystem Resilience Index (BERI) assesses the extent to which a given spatial configuration of natural habitat will promote or hinder climate-induced shifts in biological distributions, and is also endorsed by the BIP as a primary biodiversity indicator (<https://bipdashboard.natureserve.org>).

Climate change can also be considered in GDM-based analyses of protected area representativeness. These assessments consider how well existing protected areas are likely to support biodiversity into the future, under climate change (Dunlop & Brown, 2008; Williams et al., 2016). This is a relatively simple adaptation of the representativeness analysis under current climate (Figure 7g; Supporting Information Appendix S1 – section 2.8), but substituting the representative sample of grid cells from the whole region and protected areas with the same points from the future climate GDM transformed layers.

Under future environmental change scenarios, the likelihood of current habitats 'disappearing' can be predicted using GDM by determining the greatest predicted similarity to the current state of the reference cell within a sample of cells from the future (Ferrier et al., 2012; Figure 8c; Supporting Information Appendix S1 – section 2.15). 'Novel habitats' can be predicted in a similar manner, but in this case by determining the greatest predicted similarity for each reference cell in the future to a sample of cells in the present (Figure 8d; Gougherty et al., 2021).

GDM spatial predictions have been combined with spatial predictions of species richness, and all available species occurrence data, to predict the species composition (identities of species present) for every grid cell across a region. This approach, called DynamicFOAM, has been applied to a number of taxonomic groups in several regions (Mokany et al., 2011, 2012, 2014), and has also been used as the initial composition for dynamic metacommunity modelling of spatio-temporal changes in biodiversity under changing climate and habitat availability (Mokany et al., 2012, 2013, 2015, Mokany, Bush, & Ferrier, 2019).

There has also been an alternative approach applied in using GDM to assess environmental gaps in the coverage of biological surveys. This method relies only on the predicted ecological distance that is derived from a GDM, without undertaking a prediction of pairwise compositional dissimilarity (Ferrier et al., 2007). Applying GDM transformed environmental layers within the environmental diversity (ED) framework (Faith & Walker, 1996), this approach has been applied in a number of settings (Funk et al., 2005; Guerin et al., 2021).

An extension of GDM in terms of response data has been to apply it in modelling the dissimilarity between more than two sites (Latombe et al., 2017). This multi-site GDM approach is based on the concept of zeta diversity, which is the average number

of species shared by n sites, where n is the 'order' of zeta that is being considered (e.g., three sites are considered where $n = 3$). Modelling the dissimilarity between more than two sites focuses more on the role of common species and less on rare species in influencing compositional turnover (Latombe et al., 2017). Multi-site GDM therefore requires careful consideration of the goals of the modelling exercise.

In contrast, the sparse-GDM methodological framework was designed to improve the use of high-dimensional remotely sensed data, particularly hyperspectral data, to predict community turnover with GDM (Leitão et al., 2015, 2019). It involves first transforming the remote sensed data with a sparse canonical correlation analysis (SCCA), to deal with the high dimensionality of the data, and secondly fitting the transformed data with GDM.

1.11 | Future development

GDM development and associated spatial analyses have reached a relatively mature phase, as demonstrated by the substantial and continuing increase in the number of studies applying the methods (Figure 1). However, there remains a number of avenues that warrant further research and development, to enhance the rigour and scope of GDM applications. One important extension to GDM would be developing the capacity to accommodate structured data, as in a mixed-effects model framework (Harrison et al., 2018; Schielzeth & Nakagawa, 2013). Biological data are often structured (e.g., biological observations can be grouped by geological type, or biogeographic region), which introduces a form of non-independence that can be explicitly modelled. Beyond improving model rigour, however, incorporating variation within or among different levels of structured data could be useful in its own right, allowing model parameters such as the intercept to vary regionally. Such features would then permit subsequent spatial analyses to be regionally contextualized.

GDM is most often employed to model spatial turnover in composition, with the assumption that the space-for-time substitution reasonably reflects temporal turnover. While GDM can also be used to uniquely model temporal turnover, a further key development would allow for the explicit detection of turnover that arises jointly from both spatial and temporal separation of biological communities (Blois et al., 2013). More broadly, simple approaches to including non-additive interactions between predictor variables in GDM fitting could also expand potential understanding, explanation and prediction of biodiversity.

Another basic methodological issue requiring systematic quantitative assessment is the suitability of alternative options for the link and variance functions in GDM fitting. Despite the key paper by Ferrier et al. (2007) highlighting the need for future consideration of alternative link and variance functions, there has been limited information emerging on this issue. Given code is now available to fit GDM using the core GLM functions within R (see 'BILBI_supportingFunctions.R' in Hoskins, Harwood, Ware, Williams, et al., 2020),

there is an opportunity to clarify the suitability of the link and variance functions used as defaults to date, and identify the situations in which alternatives are more suitable.

In terms of predictor data, there would be broad benefits in development of generic and efficient ways to incorporate more nuanced geographic distance metrics into both GDM fitting and spatial analyses. While Euclidean geographic distance has been a useful predictor in many studies, this often fails to capture the more complex role of biogeographic history and spatio-temporal environmental patterns in influencing how spatial separation influences compositional dissimilarity. Harnessing more complex distance metrics, such as least cost paths and circuit theory have been discussed (Ferrier et al., 2007); however, computationally efficient solutions at both the fitting and spatial analysis phases have yet to be developed.

2 | CONCLUSION

The development of GDM has enabled a variety of new analyses that have improved our understanding of the drivers of biological turnover and spatial patterns in diversity. The broad range of GDM-based spatial biodiversity assessment techniques that continue to emerge highlight the powerful and flexible opportunities enabled by GDM to inform the conservation of biodiversity. Wider application of GDM and ongoing refinement and accessibility of the methods used in model building into the future are likely to further cement an important role for GDM in biodiversity assessment.

ACKNOWLEDGMENTS

MCF acknowledges funding support from the University of Maryland Center for Environmental Science, the National Science Foundation (awards 1461868, and 1656099) and a Distinguished Visiting Researcher fellowship from the Commonwealth Scientific and Industrial Research Organisation (CSIRO) papers in this journal do not have pag.

CONFLICT OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The data for Tasmanian vascular plants used here as a case-study for GDM were derived from data extracted from Tasmania's Natural Values Atlas (www.naturalvaluesatlas.tas.gov.au), 25/2/2011, © State of Tasmania, originally presented by Mokany et al. (2012) and are available on CSIRO's Data Access Portal via: <https://doi.org/10.4225/08/5542FEA245D6F>. Data presented in the worked examples in the Supporting Information (Appendix S1) are accessible via the *gdm* R package (Fitzpatrick et al., 2021).

ORCID

Karel Mokany  <https://orcid.org/0000-0003-4199-3697>

REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Akademiai Kiado.
- Allnutt, T. F., Ferrier, S., Manion, G., Powell, G. V. N., Ricketts, T. H., Fisher, B. L., Harper, G. J., Irwin, M. E., Kremen, C., Labat, J.-N., Lees, D. C., Pearce, T. A., & Rakotondrainibe, F. (2008). A method for quantifying biodiversity loss and its application to a 50-year record of deforestation across Madagascar. *Conservation Letters*, 1(4), 173–181. <https://doi.org/10.1111/j.1755-263X.2008.00027.x>
- Amano, T., Lammig, J. D. L., & Sutherland, W. J. (2016). Spatial gaps in global biodiversity information and the role of citizen science. *BioScience*, 66(5), 393–400. <https://doi.org/10.1093/biosci/biw022>
- Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., Sanders, N. J., Cornell, H. V., Comita, L. S., Davies, K. F., Harrison, S. P., Kraft, N. J. B., Stegen, J. C., & Swenson, N. G. (2011). Navigating the multiple meanings of β diversity: A roadmap for the practicing ecologist. *Ecology Letters*, 14(1), 19–28. <https://doi.org/10.1111/j.1461-0248.2010.01552.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(23), 9374–9379. <https://doi.org/10.1073/pnas.1220228110>
- Bush, A., Catullo, R., Mokany, K., Harwood, T., Hoskins, A. J., & Ferrier, S. (2019). Incorporating existing thermal tolerance into projections of compositional turnover under climate change. *Global Ecology and Biogeography*, 28(6), 851–861. <https://doi.org/10.1111/geb.12898>
- Chesters, D., Beckschäfer, P., Orr, M. C., Adamowicz, S. J., Chun, K.-P., & Zhu, C.-D. (2019). Climatic and vegetational drivers of insect beta diversity at the continental scale. *Ecology and Evolution*, 9(24), 13764–13775. <https://doi.org/10.1002/ece3.5795>
- Di Marco, M., Harwood, T. D., Hoskins, A. J., Ware, C., Hill, S. L. L., & Ferrier, S. (2019). Projecting impacts of global climate and land-use scenarios on plant biodiversity using compositional-turnover modelling. *Global Change Biology*, 25(8), 2763–2778. <https://doi.org/10.1111/gcb.14663>
- Dobson, A. J. (2001). *An introduction to generalized linear models* (2nd ed.). Chapman and Hall.
- Drielsma, M. J., Howling, G., & Love, J. (2012). *NSW native vegetation management benefits analyses* (Technical report). NSW Office of Environment and Heritage Sydney.
- Dunlop, M., & Brown, P. R. (2008). *Implications of climate change for Australia's national reserve system: A preliminary assessment*. Department of Climate Change.
- Eliith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. McC. M., Townsend Peterson, A., ... Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2), 129–151. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>
- Ellis, N., Smith, S. J., & Pitcher, C. R. (2012). Gradient forests: Calculating importance gradients on physical predictors. *Ecology*, 93(1), 156–168. <https://doi.org/10.1890/11-0252.1>
- Faith, D. P., & Ferrier, S. (2002). Linking beta diversity, environmental variation, and biodiversity assessment. *Science*, 295, 22.
- Faith, D. P., Minchin, P. R., & Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69(1), 57–68. <https://doi.org/10.1007/BF00038687>
- Faith, D. P., & Walker, P. A. (1996). Environmental diversity: On the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. *Biodiversity & Conservation*, 5(4), 399–415. <https://doi.org/10.1007/bf00056387>

- Ferrier, S. (2002). Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? *Systematic Biology*, 51, 331–363. <https://doi.org/10.1080/10635150252899806>
- Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, 43(3), 393–404. <https://doi.org/10.1111/j.1365-2664.2006.01149.x>
- Ferrier, S., Harwood, T. D., Ware, C., & Hoskins, A. J. (2020). A globally applicable indicator of the capacity of terrestrial ecosystems to retain biological diversity under climate change: The bioclimatic ecosystem resilience index. *Ecological Indicators*, 117, e106554. <https://doi.org/10.1016/j.ecolind.2020.106554>
- Ferrier, S., Harwood, T. D., & Williams, K. J. (2012). *Queensland's biodiversity under climate change: Ecological scaling of terrestrial environmental change*. CSIRO. <https://doi.org/10.25919/5b6354b6cbc6f>
- 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>
- Ferrier, S., Powell, G. V. N., Richardson, K. S., Manion, G., Overton, J. M., Allnutt, T. F., Cameron, S. E., Mantle, K., Burgess, N. D., Faith, D. P., Lamoreux, J. F., Kier, G., Hijmans, R. J., Funk, V. A., Cassis, G. A., Fisher, B. L., Flemons, P., Lees, D., Lovett, J. C., & Van Rompaey, R. S. A. R. (2004). Mapping more of terrestrial biodiversity for global conservation assessment. *BioScience*, 54(12), 1101–1109. [https://doi.org/10.1641/0006-3568\(2004\)054\[1101:MMOTBF\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2004)054[1101:MMOTBF]2.0.CO;2)
- 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), 1–16. <https://doi.org/10.1111/ele.12376>
- Fitzpatrick, M. C., Mokany, K., Manion, G., Nieto-Lugilde, D., & Ferrier, S. (2021). *gdm: Generalized dissimilarity modeling*. R package version 1.5.0-1. <https://CRAN.R-project.org/package=gdm>
- 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(5), 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(1768), 20131201. <https://doi.org/10.1098/rspb.2013.1201>
- Funk, V. A., Richardson, K. S., & Ferrier, S. (2005). Survey-gap analysis in expeditionary research: Where do we go from here? *Biological Journal of the Linnean Society*, 85(4), 549–567. <https://doi.org/10.1111/j.1095-8312.2005.00520.x>
- Gibson, L. A., Williams, K. J., Pinder, A. M., Harwood, T. D., McKenzie, N. L., Ferrier, S., Lyons, M. N., Burbidge, A. H., & Manion, G. (2015). Compositional patterns in terrestrial fauna and wetland flora and fauna across the Pilbara biogeographic region of Western Australia and the representativeness of its conservation reserve system. *Records of the Western Australian Museum Supplement*, 78(2), 515–545. [https://doi.org/10.18195/issn.0313-122x.78\(2\).2015.515-545](https://doi.org/10.18195/issn.0313-122x.78(2).2015.515-545)
- Gleason, H. (1922). On the relation between species and area. *Ecology*, 3, 158–162. <https://doi.org/10.2307/1929150>
- Gougherty, A. V., Keller, S. R., & Fitzpatrick, M. C. (2021). Maladaptation, migration and extirpation fuel climate change risk in a forest tree species. *Nature Climate Change*, 11(2), 166–171. <https://doi.org/10.1038/s41558-020-00968-6>
- Guerin, G. R., Williams, K. J., Leitch, E., Lowe, A. J., & Sparrow, B. (2021). Using generalised dissimilarity modelling and targeted field surveys to gap-fill an ecosystem surveillance network. *Journal of Applied Ecology*, 58(4), 766–776. <https://doi.org/10.1111/1365-2664.13814>
- Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, e4794. <https://doi.org/10.7717/peerj.4794>
- Hillebrand, H., Blasius, B., Borer, E. T., Chase, J. M., Downing, J. A., Eriksson, B. K., Filstrup, C. T., Harpole, W. S., Hodapp, D., Larsen, S., Lewandowska, A. M., Seabloom, E. W., Van de Waal, D. B., & Ryabov, A. B. (2018). Biodiversity change is uncoupled from species richness trends: Consequences for conservation and monitoring. *Journal of Applied Ecology*, 55(1), 169–184. <https://doi.org/10.1111/1365-2664.12959>
- Hoskins, A. J., Harwood, T. D., Ware, C., & Ferrier, S. (2020). *BILBI—Example dataset and code*. CSIRO.
- Hoskins, A. J., Harwood, T. D., Ware, C., Williams, K. J., Perry, J. J., Ota, N., Croft, J. R., Yeates, D. K., Jetz, W., Golebiewski, M., Purvis, A., Robertson, T., & Ferrier, S. (2020). BILBI: Supporting global biodiversity assessment through high-resolution macroecological modelling. *Environmental Modelling & Software*, 132, e104806. <https://doi.org/10.1016/j.envsoft.2020.104806>
- Hubbell, S. (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton University Press.
- Jay, F., Manel, S., Alvarez, N., Durand, E. Y., Thuiller, W., Holderegger, R., Taberlet, P., & François, O. (2012). Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, 21(10), 2354–2368. <https://doi.org/10.1111/j.1365-294X.2012.05541.x>
- Johnston, M. R., Elmore, A. J., Mokany, K., Lisk, M., & Fitzpatrick, M. C. (2017). Field-measured variables outperform derived alternatives in Maryland stream biodiversity models. *Diversity and Distributions*, 23(9), 1054–1066. <https://doi.org/10.1111/ddi.12598>
- Jones, M. M., Ferrier, S., Condit, R., Manion, G., Aguilar, S., & Pérez, R. (2013). Strong congruence in tree and fern community turnover in response to soils and climate in central Panama. *Journal of Ecology*, 101(2), 506–516. <https://doi.org/10.1111/1365-2745.12053>
- Koubbi, P., Ozouf-Costaz, C., Goarant, A., Moteki, M., Hulley, P.-A., Causse, R., Dettai, A., Duhamel, G., Pruvost, P., Tavernier, E., Post, A. L., Beaman, R. J., Rintoul, S. R., Hirawake, T., Hirano, D., Ishimaru, T., Riddle, M., & Hosie, G. (2010). Estimating the biodiversity of the East Antarctic shelf and oceanic zone for ecoregionalisation: Example of the ichthyofauna of the CEAMARC (Collaborative East Antarctic Marine Census) CAML surveys. *Polar Science*, 4(2), 115–133. <https://doi.org/10.1016/j.polar.2010.04.012>
- Latombe, G., Hui, C., & McGeoch, M. A. (2017). Multi-site generalised dissimilarity modelling: Using zeta diversity to differentiate drivers of turnover in rare and widespread species. *Methods in Ecology and Evolution*, 8(4), 431–442. <https://doi.org/10.1111/2041-210X.12756>
- Leathwick, J. R., Moilanen, A., Ferrier, S., & Julian, K. (2010). Complementarity-based conservation prioritization using a community classification, and its application to riverine ecosystems. *Biological Conservation*, 143(4), 984–991. <https://doi.org/10.1016/j.biocon.2010.01.012>
- 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(1), 21–38. <https://doi.org/10.1111/j.1365-2427.2010.02414.x>
- Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology*, 74, 1659–1673. <https://doi.org/10.2307/1939924>
- Legendre, P., Borcard, D., & Peres-Neto, P. R. (2005). Analyzing beta diversity: Partitioning the spatial variation of community composition data. *Ecological Monographs*, 75(4), 435–450. <https://doi.org/10.1890/05-0549>
- Leitão, P. J., Schwieder, M., Pedroni, F., Sanchez, M., Pinto, J. R. R., Maracahipes, L., Bustamante, M., & Hostert, P. (2019). Mapping

- woody plant community turnover with space-borne hyperspectral data – A case study in the Cerrado. *Remote Sensing in Ecology and Conservation*, 5(1), 107–115. <https://doi.org/10.1002/rse2.91>
- Leitão, P. J., Schwieder, M., Suess, S., Catry, I., Milton, E. J., Moreira, F., Osborne, P. E., Pinto, M. J., Linden, S., & Hostert, P. (2015). Mapping beta diversity from space: Sparse generalised dissimilarity modelling (SGDM) for analysing high-dimensional data. *Methods in Ecology and Evolution*, 6(7), 764–771. <https://doi.org/10.1111/2041-210X.12378>
- Lu, M., Vasseur, D., & Jetz, W. (2019). Beta diversity patterns derived from island biogeography theory. *The American Naturalist*, 194(3), E52–E65. <https://doi.org/10.1086/704181>
- MacArthur, R., & Wilson, E. (1967). *The theory of island biogeography*. Princeton University Press.
- 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 and Evolution*, 28, 201–218. <https://doi.org/10.1007/BF02515450>
- Mokany, K., Bush, A., & Ferrier, S. (2019). Community assembly processes restrict the capacity for genetic adaptation under climate change. *Ecography*, 42(6), 1164–1174. <https://doi.org/10.1111/ecog.03994>
- Mokany, K., Ferrier, S., Harwood, T. D., Ware, C., Di Marco, M., Grantham, H. S., Venter, O., Hoskins, A. J., & Watson, J. E. M. (2020). Reconciling global priorities for conserving biodiversity habitat. *Proceedings of the National Academy of Sciences*, 117(18), 9906–9911. <https://doi.org/10.1073/pnas.1918373117>
- Mokany, K., Harwood, T. D., & Ferrier, S. (2013). Comparing habitat configuration strategies for retaining biodiversity under climate change. *Journal of Applied Ecology*, 50(2), 519–527. <https://doi.org/10.1111/1365-2664.12038>
- Mokany, K., Harwood, T. D., & Ferrier, S. (2019). Improving links between environmental accounting and scenario-based cumulative impact assessment for better-informed biodiversity decisions. *Journal of Applied Ecology*, 56(12), 2732–2741. <https://doi.org/10.1111/1365-2664.13506>
- Mokany, K., Harwood, T. D., Overton, J. M., Barker, G. M., & Ferrier, S. (2011). Combining α - and β -diversity models to fill gaps in our knowledge of biodiversity. *Ecology Letters*, 14, 1043–1051. <https://doi.org/10.1111/j.1461-0248.2011.01675.x>
- Mokany, K., Harwood, T. D., Williams, K. J., & Ferrier, S. (2012). Dynamic macroecology and the future for biodiversity. *Global Change Biology*, 18(10), 3149–3159. <https://doi.org/10.1111/j.1365-2486.2012.02760.x>
- Mokany, K., Jordan, G. J., Harwood, T. D., Harrison, P. A., Keppel, G., Gilfedder, L., Carter, O., & Ferrier, S. (2017). Past, present and future refugia for Tasmania's palaeoendemic flora. *Journal of Biogeography*, 44(7), 1537–1546. <https://doi.org/10.1111/jbi.12927>
- Mokany, K., Prasad, S., & Westcott, D. A. (2015). Impacts of climate change and management responses in tropical forests depend on complex frugivore-mediated seed dispersal. *Global Ecology and Biogeography*, 24(6), 685–694. <https://doi.org/10.1111/geb.12292>
- Mokany, K., Westcott, D. A., Prasad, S., Ford, A. J., & Metcalfe, D. J. (2014). Identifying priority areas for conservation and management in diverse tropical forests. *PLoS One*, 9(2), e89084. <https://doi.org/10.1371/journal.pone.0089084>
- Murray, K. D., Janes, J. K., Jones, A., Bothwell, H. M., Andrew, R. L., & Borevitz, J. O. (2019). Landscape drivers of genomic diversity and divergence in woodland Eucalyptus. *Molecular Ecology*, 28(24), 5232–5247. <https://doi.org/10.1111/mec.15287>
- Oksanen, J., & Tonteri, T. (1995). Rate of compositional turnover along gradients and total gradient length. *Journal of Vegetation Science*, 6(6), 815–824. <https://doi.org/10.2307/3236395>
- Overton, J., Barker, G., & Price, R. (2009). Estimating and conserving patterns of invertebrate diversity: A test case of New Zealand land snails. *Diversity & Distributions*, 15, 731–741. <https://doi.org/10.1111/j.1472-4642.2009.00589.x>
- Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, 5(5), 397–406. <https://doi.org/10.1111/2041-210X.12180>
- Prentice, I. C. (1977). Non-metric ordination methods in ecology. *Journal of Ecology*, 65, 85–94. <https://doi.org/10.2307/2259064>
- Prober, S., Hilbert, D., Ferrier, S., Dunlop, M., & Gobbett, D. (2012). Combining community-level spatial modelling and expert knowledge to inform climate adaptation in temperate grassy eucalypt woodlands and related grasslands. *Biodiversity and Conservation*, 21(7), 1627–1650. <https://doi.org/10.1007/s10531-012-0268-4>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, 3(4), 425–461.
- Reside, A. E., VanDerWal, J., Phillips, B. L., Shoo, L. P., Rosauer, D. F., Anderson, B. J., Welbergen, J. A., Moritz, C., Ferrier, S., Harwood, T. D., Williams, K. J., Mackey, B., Hugh, S., & Williams, Y. M. & Williams, S. E. (2013). *Climate change refugia for terrestrial biodiversity: Defining areas that promote species persistence and ecosystem resilience in the face of global climate change*. National Climate Change Adaptation Research Facility. https://www.nccarf.edu.au/sites/default/files/attached_files_publications/Reside_2013_Climate_change_refugia_for_terrestrial_biodiversity.pdf
- 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(1), 21–32. <https://doi.org/10.1111/j.1600-0587.2013.00466.x>
- Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9(1), 130–134. <https://doi.org/10.1214/aos/1176345338>
- Schielzeth, H., & Nakagawa, S. (2013). Nested by design: Model fitting and interpretation in a mixed model era. *Methods in Ecology and Evolution*, 4(1), 14–24. <https://doi.org/10.1111/j.2041-210x.2012.00251.x>
- Smouse, P. E., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology*, 35(4), 627–632. <https://doi.org/10.2307/2413122>
- Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter/Kongelige Danske Videnskaberne Selskab*, 5(4), 1–34.
- Thomassen, H. A., Buermann, W., Milá, B., Graham, C. H., Cameron, S. E., Schneider, C. J., Pollinger, J. P., Saatchi, S., Wayne, R. K., & Smith, T. B. (2010). Modeling environmentally associated morphological and genetic variation in a rainforest bird, and its application to conservation prioritization. *Evolutionary Applications*, 3(1), 1–16. <https://doi.org/10.1111/j.1752-4571.2009.00093.x>
- Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, 7(1), 9132. <https://doi.org/10.1038/s41598-017-09084-6>
- Tuomisto, H. (2010). A diversity of beta diversities: Straightening up a concept gone awry. Part 2. Quantifying beta diversity and related phenomena. *Ecography*, 33(1), 23–45. <https://doi.org/10.1111/j.1600-0587.2009.06148.x>
- UNCEEA. (2021). *System of environmental-economic accounting – Ecosystem accounting, prepared by the United Nations committee of experts on environmental accounting (UNCEEA)*. United Nations Statistical Division.

- UNEP-WCMC. (2016). *Exploring approaches for constructing species accounts in the context of the SEEA-EEA*. United Nations Environment Program—World Conservation Monitoring Centre.
- van Schalkwyk, J., Pryke, J. S., Samways, M. J., & Gaigher, R. (2019). Congruence between arthropod and plant diversity in a biodiversity hotspot largely driven by underlying abiotic factors. *Ecological Applications*, 29(4), e01883. <https://doi.org/10.1002/eap.1883>
- Ware, C., Williams, K. J., Harding, J. O., Hawkins, B., Harwood, T., Manion, G., Perkins, G. C., & Ferrier, S. (2018). Improving biodiversity surrogates for conservation assessment: A test of methods and the value of targeted biological surveys. *Diversity and Distributions*, 24(9), 1333–1346. <https://doi.org/10.1111/ddi.12766>
- Whittaker, R. H. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, 30(3), 280–338. <https://doi.org/10.2307/1943563>
- Wilkinson, D. P., Golding, N., Guillera-Aroita, G., Tingley, R., & McCarthy, M. A. (2019). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, 10(2), 198–211. <https://doi.org/10.1111/2041-210X.13106>
- Williams, K. J., Belbin, L., Austin, M. P., Stein, J. L., & Ferrier, S. (2012). Which environmental variables should I use in my biodiversity model? *International Journal of Geographical Information Science*, 26(11), 2009–2047. <https://doi.org/10.1080/13658816.2012.698015>
- Williams, K. J., Ferrier, S., Rosauer, D. F., Yeates, D., Manion, G., Harwood, T. D., Stein, J., Faith, D., & Laity, T. & Whalen, A. (2010). *Harnessing continent-wide biodiversity datasets for prioritising national conservation investment*. CSIRO. <https://publications.csiro.au/rpr/pub?list=BRO&pid=csiro:EP102983>
- Williams, K. J., Harwood, T. D., & Ferrier, S. (2016). Assessing the ecological representativeness of Australia's terrestrial national reserve system: A community-level modelling approach. *A report prepared for the Australian Government, Department of the Environment and Energy*.
- Woolley, S. N. C., Foster, S. D., & Dunstan, P. K. (2015). *bbgdm: Bayesian bootstrap generalised dissimilarity models*. R package. <https://github.com/skiptoniam/bbgdm>
- 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(8), 985–995. <https://doi.org/10.1111/2041-210X.12710>

BIOSKETCH

Karel Mokany develops and applies biodiversity modelling and assessment approaches, to provide high quality information for biodiversity policy planning and management decisions.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Mokany, K., Ware C., Woolley S. N. C., Ferrier S., & Fitzpatrick M. C. (2022). A working guide to harnessing generalized dissimilarity modelling for biodiversity analysis and conservation assessment. *Global Ecology and Biogeography*, 00, 1–20. <https://doi.org/10.1111/geb.13459>