Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8,** 985–995 doi: 10.1111/2041-210X.12710

Characterising uncertainty in generalised dissimilarity models

Skipton N.C. Woolley*[,1,2], Scott D. Foster[3], Timothy D. O'Hara[1], Brendan A. Wintle[2] and Piers K. Dunstan[3]

[1]*Museums Victoria, GPO Box 666, Melbourne, VIC 3001, Australia;* [2]*Quantitative and Applied Ecology Group, School of Biological Sciences, University of Melbourne, Melbourne, VIC 3010, Australia; and* [3]*CSIRO, Hobart, TAS 7000, Australia*

Summary

1. Generalised dissimilarity modelling (GDM) is a statistical method for analysing and predicting patterns of turnover in species composition, usually in response to environmental gradients that vary in space and time. GDM is becoming widely applied in ecology and conservation science to interpret macro-ecological and biogeographical patterns, to support conservation assessment, predict changes in species distributions under climate change and prioritise biological surveys.

2. Inferential and predictive uncertainty is difficult to characterise using current implementations of GDM, reducing the utility of GDM in ecological risk assessment and conservation decision-making. Current practice is to undertake permutation tests to assess the importance of variables in GDM. Permutation testing overcomes the issue of data-dependence (because dissimilarities are calculated on a smaller number of observations) but it does not give a quantification of uncertainty in predictions. Here, we address this issue by utilising the Bayesian bootstrap, so that the uncertainty in the observations is carried through the entire analysis (including into the predictions).

3. We tested our Bayesian bootstrap GDM (BBGDM) approach on simulated data sets and two benthic species data sets. We fitted BBGDMs and GDMs to compare the differences in inference and prediction of compositional turnover that resulted from a coherent treatment of model uncertainty. We showed that our BBGDM approach correctly identified the signal within the data, resulting in an improved characterisation of uncertainty and enhanced model-based inference.

4. We show that our approach gives appropriate parameter estimates while better representing the underlying uncertainty that arises when conducting inference and making predictions with GDMs. Our approach to fitting GDMs will provide more realistic insights into parameter and prediction uncertainty.

Key-words: Bayesian bootstrap, beta-diversity, community ecology, generalised additive models, generalised linear models

Introduction

Describing patterns of diversity from biological data and geospatial data have seen a major push in the development of statistical methods that link environmental predictors to biological data. These methods aim to describe patterns of species distributions (Guisan & Zimmermann 2000; Leathwick & Elith 2009), several aspects of biodiversity (Foster & Dunstan 2010; Dunstan, Foster & Darnell 2011; Dunstan & Foster 2011) and infer biogeographical patterns (Ferrier *et al.* 2002; Foster *et al.* 2013). The ability to describe diversity patterns is increasingly relevant to decision-making, especially during planning phases. Improved understanding of the uncertainty of the information provided allows more confidence in the decisions that are made (Wintle *et al.* 2003; Polasky *et al.* 2011). Here we focus on the strategy of modelling patterns of diversity, rather than discrete entities such as individual species or community types. We model compositional turnover of species for the purposes of describing biogeographical patterns. We base our method on Ferrier *et al.* (2002) and Ferrier *et al.* (2007), and follow the approach of modelling the change in species composition across spatial and environmental gradients.

Modelling compositional turnover (beta-diversity) of species is gaining momentum via the accessibility of methods such as generalised dissimilarity modelling (GDM: Ferrier *et al.* 2002, 2007). GDM uses the compositional dissimilarity between all site-pairs (samples) and models these dependant variables as a function of the difference in environmental and geographical covariates between site-pairs. GDMs are closely related to generalised linear models (GLMs), where the explanatory variables are constructed as the absolute value of the difference in the I-spline basis values (see Ramsay 1988) at the two site locations, and the associated parameters are constrained to be positive (Ferrier *et al.* 2007). I-splines are monotonic spline

*Correspondence author. E-mail: swoolley@museum.vic.gov.au

© 2016 The Authors. Methods in Ecology and Evolution © 2016 British Ecological Society

functions useful for non-negative and non-decreasing fits (Ramsay 1988). Within the GDM framework, the use of I-splines assumes that increasing distance in environmental or geographical space will be related to an increased compositional dissimilarity between sites (Ferrier *et al.* 2007).

The methodological issues associated with the current implementation of GDM originate from using dissimilarities as independent observations. This problem of pseudo-replication also applies to a number of other existing techniques, such as matrix regression (Smouse, Long & Sokal 1986) and DISTLM (Anderson 2004). The observations in these models are not independent as each sample is used to calculate more than one dissimilarity. For example, if we have $n = 100$ survey sites, for a GDM, we are modelling $n(n-1)/2 = 4950$ dissimilarities. GDM uses these $n(n-1)/2$ dissimilarities as independent data, but the likelihood values obtained, and particularly the curvature of the likelihood, will lead to estimates that are too precise (underestimated standard errors) and will also lead to poorly selected models (via information criteria based on the likelihood). The magnitude of the problem will increase with the number of sites. The issue of non-independence in GDM has been recognised from the outset, and researchers have used permutation methods to deal with this problem in hypothesis tests (Ferrier *et al.* 2002; Anderson 2004; Fitzpatrick *et al.* 2013; Jones *et al.* 2015). Permutation testing enables hypothesis testing only, and falls short of characterisation of uncertainty in predictions of dissimilarities (through interval estimation methods). Despite the previous use of permutation methods to test for significance of variables (Ferrier *et al.* 2002), a more general approach for quantifying uncertainty and dealing with the issue of non-independence when modelling dissimilarities in statistical inference is required. In particular, it is important to quantify uncertainty in the predictions of dissimilarity, from which the model is interpreted. We propose an extension of GDM that characterises uncertainty with respect to the amount of information in the site data. Our approach is to incorporate the Bayesian bootstrap (BB: Rubin 1981) into the GDM framework. The BB approach differs from the standard bootstrap (see Davison 1997) as a site's data are never completely removed from the bootstrap sample. This has the fortuitous effect of ensuring that all dissimilarities are observed in each of the bootstrap samples, that no pairwise dissimilarity between bootstrap samples is identically zero, and that the range of biological and environmental gradients is preserved. Additionally, we can estimate credible intervals for parameters, and produce improved diagnostics. We refer to this extension to GDM as Bayesian bootstrap generalised dissimilarity modelling (BBGDM).

Materials and methods

We describe the GDM in terms of a generalised linear model (GLM) [as formulated in McCullagh & Nelder (1989) and as previously described by Ferrier *et al.* (2007)] and define estimation of parameters, diagnostics and interpretation of GDM in a GLM framework. Within a GLM, we assume the realised dissimilarities, y_{ij} with $i, j = 1, \ldots, n$, are independent observations/realisations of a response variable. The assumption of independence is accounted for in Section 'Bayesian bootstrap'. We take this dissimilarity to be the number of species *not* shared between the sites but note that *any* dissimilarity could be used with appropriate changes to the GLM set-up. For the purposes of this paper, we assume that the dissimilarities can be well represented using a binomial process, at least for the first two moments. The working model is

$$Y_{ij} \sim Bn(n_{ij}, \pi_{ij}), \qquad \text{eqn 1}$$

where Y_{ij} is a vector representing the number of species not shared between sites i and j and n_{ij} is the union of species between the two sites i and j, so Y_{ij}/n_{ij} is the observed binomial proportion, which is essentially Jaccard's dissimilarity (Jaccard 1912), and π_{ij} is the expected dissimilarity.

It was our preference to use the logit link function, as it is the canonical link function for binomial GLMs and is subsequently used most often for binomial data, and unlike the negative exponential (negexp) link function, it maps the linear predictor to the unit interval [0,1]. However, a small simulation study in the supporting information (see Appendix S1, Supporting Information) shows that there is not much difference between the fits of the models, and the choice of link function should be assessed on a case-by-case basis.

For a GDM, the expected dissimilarity is modelled as the absolute difference between I-spline functions (see Appendix S2). Like Ferrier *et al.* (2007), we constrain I-spline base value parameter estimates to be positive as this reflects the assumption that dissimilarities will increase with greater environmental and geographical distances. The constraint is implemented by estimating the log of the parameters. The intercept is not constrained.

In GLMs generally, the maximised log-likelihood (or minimised deviance) can be used for many purposes. Most commonly, it can be used to aid model selection, through calculation of information criteria (e.g. AIC: Akaike 1974). The curvature of a GLM's log-likelihood surface also provides information about the sampling distribution of the parameter estimates. If we are to continue the analogy of GDMs and GLMs then an obvious candidate to estimate the uncertainty in a GDM is through the curvature of the log-likelihood surface, which is immediately available through most software routines to calculate GLMs. We return to the topic of variance in parameter estimates shortly.

BAYESIAN BOOTSTRAP

We extended GDM by adding a Bayesian bootstrapping (BB: Rubin 1981) step to address the non-independence of dissimilarities. The BB is used to create new bootstrap data sets through the re-weighting of the initial data (Rubin 1981). We re-weight the set of dissimilarities based on the number of sites, and so carry forward their inherent correlation through their dependence on the site data. To achieve this, we generated B sets of bootstrap weight vectors, $\{\mathbf{w}_b\}_{b=1}^{B}$, where each weight vector $(\mathbf{w}_b)$ is a symmetric Dirichlet variable of length n (the number of sites). A Dirichlet variable is a statistical distribution of a series of numbers that sum to one (Gelman *et al.* 2013). We now have weights for sites and to turn them into weights for dissimilarities we used the upper triangle of $\mathbf{w}_b\mathbf{w}_b^{\top}$. This re-weighting scheme is based on the assumption that the probability of two sites being both sampled in the bootstrap sample is equal to the product of their individual probabilities. We then fitted a GDM using these weights, saved all the resulting parameter estimates, and repeated for the B sets of bootstrap weights. The resulting set of parameter estimates is an empirical average of the distribution of the estimates (the sampling distribution of the estimates). For example,

the Bayesian bootstrap estimate of $\hat{\mu}$ for a logit regression can be represented as:

$$\hat{\mu}_{ij} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{1+\exp(-\mathbf{x}_{ij}^{\top}\hat{\boldsymbol{\beta}}^{(b)})}, \qquad \text{eqn 2}$$

where $\mathbf{x}_{ij}$ is the vector of differences in I-spline bases for sites i and j and $\hat{\beta}^{(b)}$ is the estimated parameter vector. Credible intervals are the relevant quantiles of this distribution. They take into account the variation in the observed data and do not assume that the dissimilarities, which are correlated by construction, are independent.

The BB and the case-resampled bootstrap (see Davison 1997) have very similar properties – operationally they are similar and the results tend to be similar too (Rubin 1981). The main theoretical difference is that the BB draws a sample for the posterior distribution of the statistic under study, whereas the case re-sampled bootstrap draws from the sampling distribution. This means that in the case-resampled bootstrap some observations (pairwise dissimilarities) are not included in the bootstrap sample. Hence, they are assigned a zero weight in the bootstrap sample. This will never happen with Bayesian bootstrap as there is zero probability of observing a zero in the weight vector. Practically however, and for GDMs in particular, having non-zero weights implies that the design matrix for the GLM does not change and hence there are no abrupt changes to the structure of the GLM, and parameters are thus identifiable when derived from the Bayesian bootstrap samples.

MODEL SELECTION AND INFERENCE

Model selection tools such as pseudo-R^2 (deviance explained), Akaike information criterion (AIC: Akaike 1974) and Bayesian information criterion (BIC: Schwarz 1978), are all based on the log-likelihood. In GDM and BBGDM, the model log-likelihood is based on dissimilarities, which erroneously assumes that dissimilarities are independent. This feature of GDM means that statistics directly estimated or derived from the log-likelihood are unreliable (McCullagh & Nelder 1989). The implication is that model selection needs to be performed without directly using the log-likelihood value. This subsequently rules out some common model selection tools, like AIC, BIC and deviance explained. Typically, variable selection is achieved using permutation significance tests (Ferrier *et al.* 2007). We proceed by inspecting the size of a variable's parameters in relation to its posterior variance (obtained from the BB). In particular, we use a 'Wald-like' test, named due to its similarity with the Wald test (Kodde & Palm 1986), to perform such a comparison. The 'Wald-like' test can be used to assess the parameter estimates from a statistical model based on the sample estimate. For variable p, the test statistic is $W_p = \hat{\boldsymbol{\beta}}_p^{\top}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\beta}}_p$, where $\hat{\boldsymbol{\beta}}_p$ is the BB estimate of the regression spline parameters and $\hat{\boldsymbol{\Sigma}}$ is the BB estimate of the variance of these parameters. We assume that this statistic follows a chi-squared distribution with Q degrees of freedom. We suspect that Q is an upper-bound for the degrees of freedom and, if true, then the test will be conservative. That is, it will produce P-values that are larger on average than they should be. In a model-building context (determining which covariates to include in the model), this translates to see if W_p is greater than the percentile corresponding to the nominal type-1 error rate (generally the 0·95 percentile for the 0·05 type-1 error). This is a two-sided test, picking out both positive and negative coefficients, as the test statistic is a quadratic form. If it is greater than the percentile, then the covariate is included in the GDM. In our data analysis, this process is used to select the best subset of parameters that seek to explain compositional turnover modelled within our geographic region.

MODEL DIAGNOSTICS

The use of model diagnostics plays a vital role in model building, checking and inference. One key approach, used in all areas of applied statistics, is assessing the behaviour of residuals. Due to the non-normal nature of dissimilarities, it is difficult to properly interpret standard residuals as their finite sampling distribution is not known. Thus, we require residuals that take into account the non-normality of dissimilarity data, which is assumed to be binomial (as a working model). We use random quantile residuals (Dunn & Smyth 1996) whose distribution is known, when the model provides an adequate fit to the data. We calculate these residuals based on the mean estimates of the parameters from B BB samples. However, dissimilarities are algorithmic abstractions of species observations so they are not true data observations (Warton, Wright & Wang 2012). This means the mean–variance relationship of a dissimilarity is unknown. Thus, residuals calculated on dissimilarities could be misleading due to the correlation between site-pairs and their unknown probability distribution.

INTERPRETATION

Bayesian bootstrap GDM aims to estimate turnover of species across geographic and environmental space. The incorporation of a BB enables us to interpret parameters in light of their uncertainty. Typically, the posterior distribution represents the parameter estimates with respect to the data. For GDM and BBGDM we can present a simple example. Consider the model

$$g\big(\mathrm{E}(y_{ij})\big) = \beta_0 + \sum_{k=1}^{K}|I_k(\text{temperature}_i) - I_k(\text{temperature}_j)|\beta_{1k} + \sum_{k=1}^{K}|I_k(\text{oxygen}_i) - I_k(\text{oxygen}_j)|\beta_{2k}, \qquad \text{eqn 3}$$

where $\mathrm{E}(y_{ij})$ is the expectation of the dissimilarity between sites i and j, $g(\cdot)$ is the link function, and the right-hand side of the equation specifies the dissimilarity as a link-linear function of an intercept β_0 and a set of explanatory covariates, temperature and oxygen. The intercept value is the baseline estimate of species turnover when covariates temperature and oxygen are equal at both sites. We can interpret the relative contribution of temperature and oxygen in driving species turnover based on the contribution of the sum of I-spline differences for each covariate. For example, if the estimates of the temperature coefficients ($\boldsymbol{\beta}_1$) had values twice those of the oxygen coefficients ($\boldsymbol{\beta}_2$), then a twofold overall change in oxygen would be required to drive the same amount of turnover as that of temperature. As creation of the I-spline basis gives all covariates the same units (within the particular data set), this relative increase is in relation to the range of observed covariates and not the units in which the covariates are measured. The B_k parameters are inferred from the data. The parameter estimates are those given by the BB; a reasonable point summary is usually the median of this distribution. GDM is a single realisation of the BB re-sampling (with weights all equal). A reasonable way to interpret the posterior distribution of regression coefficient estimates is to visually compare partial effect plots of I-spline contributions and their associated credible intervals derived from BB posterior estimates. Thus, important regression coefficients will be ones with larger magnitudes (absolute values) and smaller credible intervals around these coefficients.

FITTING BBGDM IN R

As a companion to this paper, we have made available an R package, *bbgdm*, that is available on 'GitHub' (https://github.com/skiptoniam/bbgdm/) and can fit the models specified in this article. The method requires the formula for the explanatory variables expected to drive compositional turnover of species. The approach uses a maximum likelihood estimation, with the Bayesian bootstrap to provide an empirical estimation of parameters. For example, this code will run a simple model:

```
fm1 <- bbgdm(~1 + x, sp.dat = spdata, env.dat
= envdat, link = 'logit', nboot = 100, geo = FALSE,
splinetype = 'ispline')
```

The coefficients and their respective uncertainties will be returned from this function call. We provide an R script that runs the simple model presented here, and includes the use of functions that run diagnostics, 'Wald-like' tests, and plotting of partial response plots (see our online example at: https://github.com/skiptoniam/bbgdm/).

SIMULATION

Thus far, we have discussed the methodological extension and uses of BBGDM. To illustrate how our BBGDM approach extends on GDM, we consider the following simulated data sets. Simulation of dissimilarities is quite difficult; unlike species occurrence (on which dissimilarities are constructed), we cannot directly assign known means and variances. Thus, we generated species occurrences with known means and variances, and translated these generated occurrences into a reasonable ecological community with species turnover across a gradient. To ensure that we had a strong gradient correlated with species turnover, we generated our community using a mixture of multivariate normal distributions to describe the species coefficients to the environmental gradient, following a similar approach proposed in Ovaskainen, Hottola & Shtonen (2010). We generated a community of $n = 200$ species inhabiting a set of $m = 50$ sites. We generated 1000 random realisations of this data set and then used these simulated data sets to compare the expectation of dissimilarity, as calculated from the known probability of species at each site, and the sample mean (as derived from BBGDM and GDM). Comparisons of the sampling mean (BBGDM) and population mean (GDM) were undertaken to see if either method adequately captures the expectation of dissimilarity. Further details on simulation of species occurrences are presented in the Supporting Information (see Appendix S3).

We further tested if incorporating a Bayesian bootstrap into the GDM framework would help capture the variance in parameter estimates, and if the standard GDM underestimated uncertainty. Using our simulated data sets, we compared the observed variance in parameters (obtained from taking the standard deviation of the GDM parameters on each of the simulated data sets) against the estimate of variance obtained from the model (measured by the mean of the standard errors of the parameters over each of the simulated data sets) for BBGDM and GDM. If the variance estimates obtained from BBGDM or GDM are appropriate they should match the distribution of means produced during simulation. We generated the variance for GDM via the covariance matrix of the parameter estimates, which is derived from the inverse of the negative of the Hessian matrix (McCullagh & Nelder 1989). We took the standard errors in parameters as the square roots of the diagonal elements of the covariance matrix and in turn 95% confidence intervals were generated. We refer to this as the 'naive variance and confidence intervals', as it treats dissimilarities as independent (which they are not). Non-independence has been addressed for hypothesis testing (using permutation tests Ferrier *et al.* 2002), but not for predictions and not the contributions of individual environmental gradients. This latter assessment of uncertainty is given by the BBGDM presented in this work.

CASE STUDIES

Tasmanian marine invertebrates

We applied our method to a real-world data set of benthic invertebrates from the continental margin and seamounts of Tasmania, Australia. Samples of fauna from the megabenthos (specimens ≥5 mm in size were retained) were collected with an epibenthic sled (Lewis 1999). The data were collected via surveys on Australia's Marine National Facility vessel *Southern Surveyor* during 2007 (SS200702). The region sampled is topographically complex, comprised of distinct continental margin structure (Koslow *et al.* 2001). Adjacent to the continental margin are a number of extinct volcano seamounts. Seamount peaks range from 700–1400 m depth, and are characterised by elevation ranges of approximately 200–300 m from base to summit.

Biological data. A total of 39 sampling sites were used in analyses, 17 sites were located on the continental slope, within the continental slope surveys, 4 were on the upper-slope (200–500 m depth) and 13 the mid-slope (600–1200 m). The remaining 22 sites were located on the sides of seamounts that occur away from the continental margin at depths greater than 700 m (Fig. 1) [see Williams *et al.* (2010b) and Dunstan *et al.* (2012b) for further detail]. Megabenthos specimens were all identified by taxonomic authorities to species level [or operational taxonomic units (OTUs), if undescribed species]. The most common taxa were ascidians, decapods, echinoderms, molluscs, octocorals and sponges. A total of 493 unique species were identified from the 39 samples, a presence–absence matrix of sites and species was used for the GDM and BBGDM analyses.

Environmental and physical data sets. We incorporated a set of environmental and spatial covariates for the Tasmanian continental margin and seamount data. Two data sources were used to represent the abiotic environment of the benthos across the study extent: the CSIRO Atlas of Regional Seas (CARS) climatology (Ridgway, Dunn & Wilkin 2002) for physical oceanography and the MARS sediment data base (GA 2009). The MARS sediment data base comprises sediment records from over 40 000 samples from around Australia; we used three interpolated sediment composition layers: sand percentage, gravel percentage and mud percentage. Ten oceanographic predictors were selected from the CARS data, including: annual mean and standard deviations of temperature, salinity, oxygen, phosphorus and nitrogen. These 13 covariates, plus depth at the sea-floor of survey sites were used in GDM and BBGDM model building. Both data sets are temporally integrated, and are spatially interpolated to 0·01° grid cells.

Statistical analysis. We used GDM and BBGDM to estimate rates of community composition turnover for the Tasmanian continental margin benthic diversity data set. To fit GDMs and BBGDMs, we constructed a site-by-species matrix, where sites were the locations recorded from epibenthic sled tows on the sea-floor. A site-by-covariate matrix was used to construct I-splines within the GDM and BBGDM

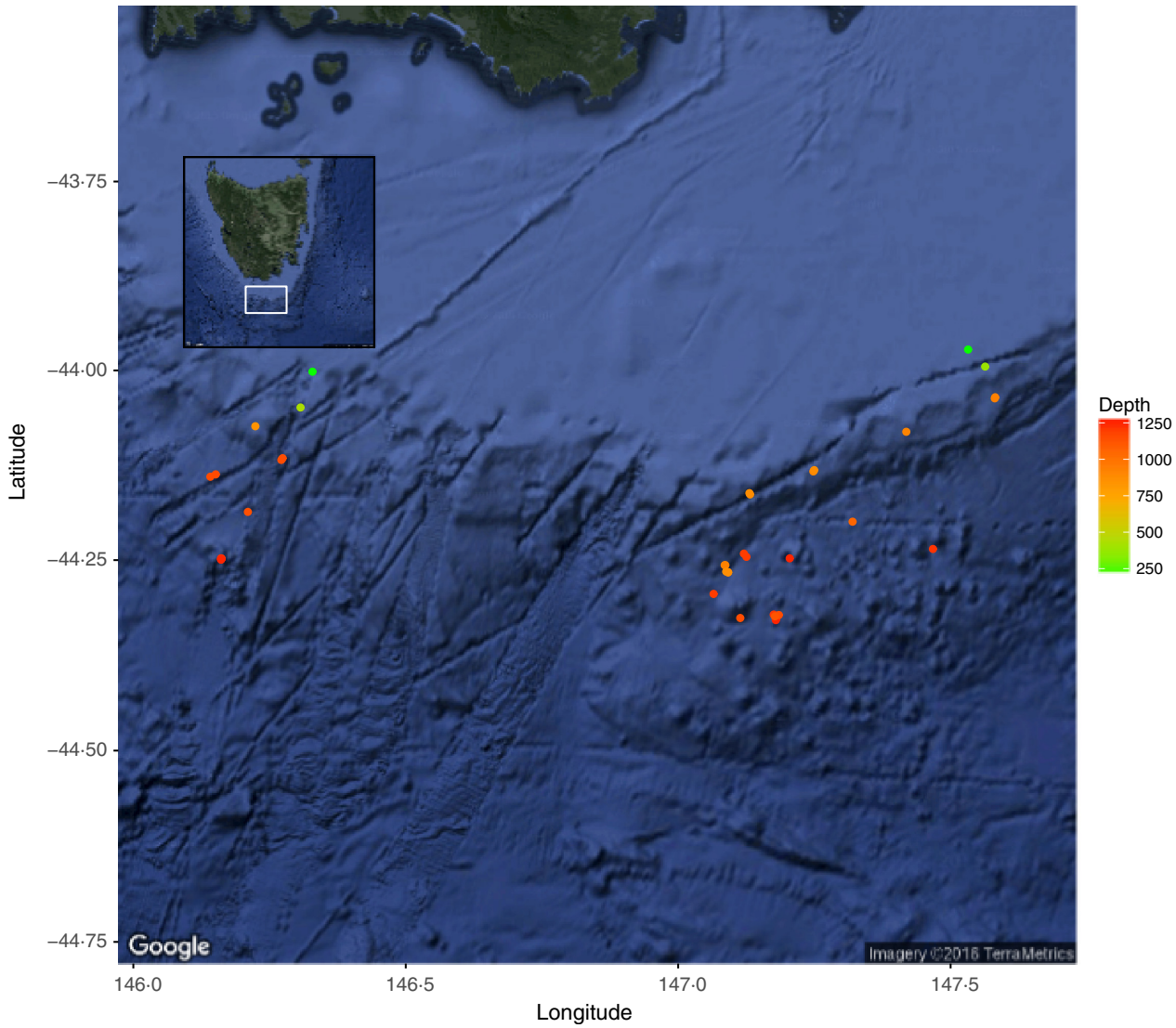

Fig. 1. Seamount and continental shelf survey sites of the study extent off Southern Tasmania. Sites are coloured green (shallow) to red (deep). Box in the top left corner shows the study extent in relation to Tasmania.

models. GDM analyses were implemented as a naïve BBGDM, comprising a single BB with all weights set to one. BBGDMs were implemented using the 'bbgdm' R statistical package presented in this paper. For each BBGDM, 1000 BB replicates were generated. Those results included, (i) a fitted I-spline for each covariate included in the models, which represents the rate of compositional turnover of species along a spatial or environmental gradient; (ii) a 'Wald-like' test to assess I-splines, and in-turn select the best subset of parameters that seek to explain beta-diversity modelled within our geographic region; and (iii) diagnostic tools, including random quantile residual plots.

Western Australian fish surveys

We present a second case study as an independent test of the BBGDM method. This was undertaken to see if we could detect a signal in covariates to explain turnover using the BBGDM method. This case study was an analysis of demersal fish species collected along the coast of Western Australia, and covered 14 latitudinal degrees and 190–1405 m of depth. While these imply large environmental gradients, there were only 65 survey sites. This means that the amount of information was relatively limited but the changes may be large. We present the results of this case study to support the validity of our BBGDM approach. The study extent, methods, results and discussion of this case study are presented in Appendix S4.

Results

SIMULATION

Point estimates for BBGDM and GDM were very similar for our simulated species across a single environmental gradient (Fig. 2). BBGDM and GDM appear to be capturing the expectation of dissimilarity from simulation. The variance of the mean estimates, from 1000 simulations, is slightly larger for the BBGDM than for the GDM (Figs. 2a and 3a), but within sampling variability.

Comparing the variance of means generated from the simulated data sets versus the expected variance, we can see that the estimated variance in BBGDM closely resembles the variance

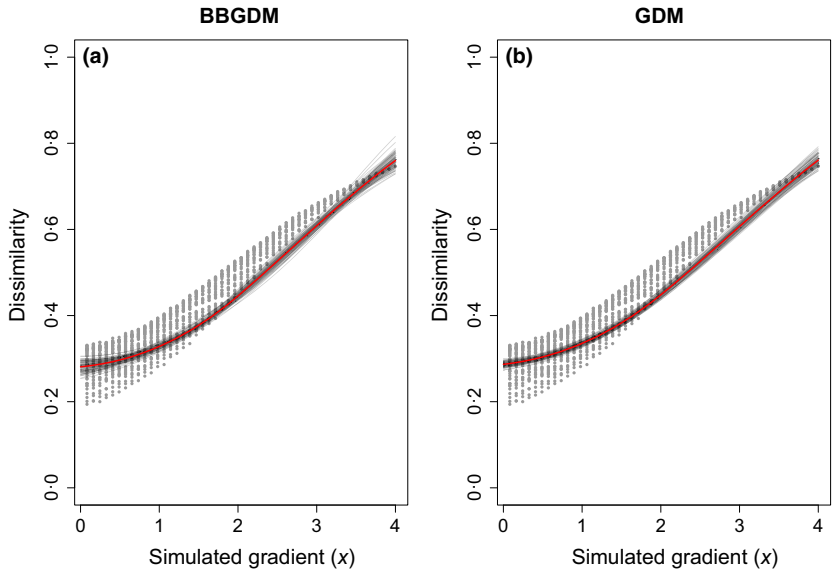

Fig. 2. Mean estimates of dissimilarity from Bayesian bootstrap GDM (BBGDM) and generalised dissimilarity modelling (GDM) simulated data sets. Within each plot the red line depicts the mean expectation of dissimilarity across all model simulations. The grey lines represent each model realisation from a simulation. Grey points are the expectation of dissimilarities between all site-pairs (see Appendix A5).

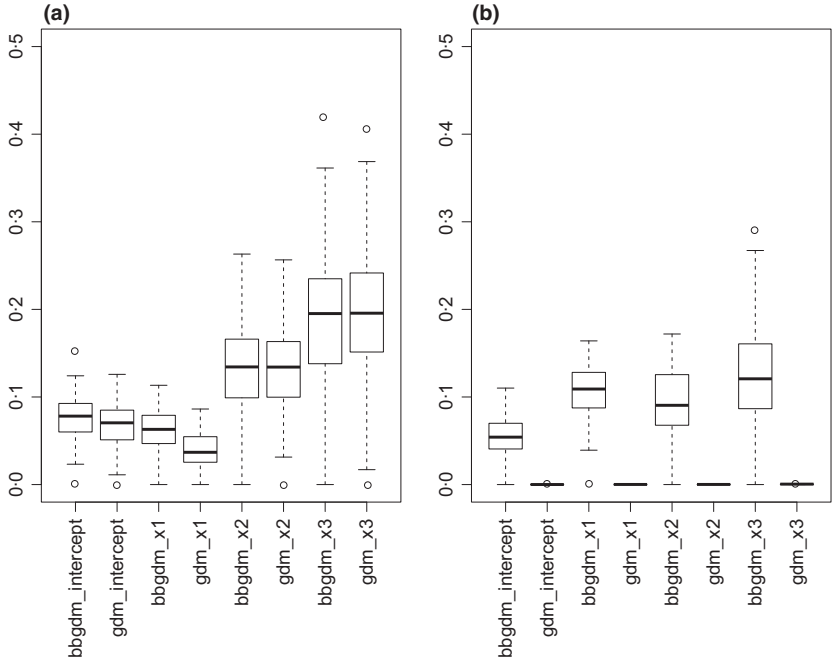

Fig. 3. Estimating variance in Bayesian bootstrap GDM and generalised dissimilarity modelling parameters derived from simulated data sets. Variance in the simulation model parameters for the intercept, and the I-spline bases: $x1$, $x2$ and $x3$ respectively: (a) the variance, over simulations, in parameter estimates; (b) the mean estimates of BBGDM variance and naive GDM variance for each parameter in our simulation models. If uncertainty is properly quantified the two plots, (a) and (b) should match.

generated from simulation (Fig. 3b), at least to within sampling variability. However, the naive GDM variance does not display this desirable behaviour, displaying estimated variances (from the model) that are very small (Fig. 3b) and do not match variance of the means, over simulated data sets (Fig. 3b).

CASE STUDIES

I-spline parameter estimates were slightly higher for GDM than BBGDM in both case studies (Figs. 4 and SI 7). Within both studies, the GDM parameter estimates were well within the sampling variability of the BBGDM (Figs. 4 and SI 7). Credible intervals of I-splines derived from BBGDM all display large variances around mean estimates, while the naive GDMs variance was small or not visible at the scale that plots were reported (Figs. 4 and SI 7). Results from the Tasmanian invertebrate case study show that naive GDM I-spline parameter estimates fail to capture the uncertainty in parameter estimates to the same degree BBGDM does (Fig. 4). Results from the Western Australian fish data set also demonstrated a greater quantification of uncertainty in parameter estimates despite the presence of strong environmental gradients (Fig. SI 7). In both case studies, the uncertainty in BBGDM parameter estimates is larger than naive GDM variance and accumulates with increasing environmental and geographic gradients (Figs. 4 and SI 7).

Interpreting parameter estimates based on the 'Wald-like' test for both case studies showed there were differences in parameters significance when comparing GDM and BBGDM

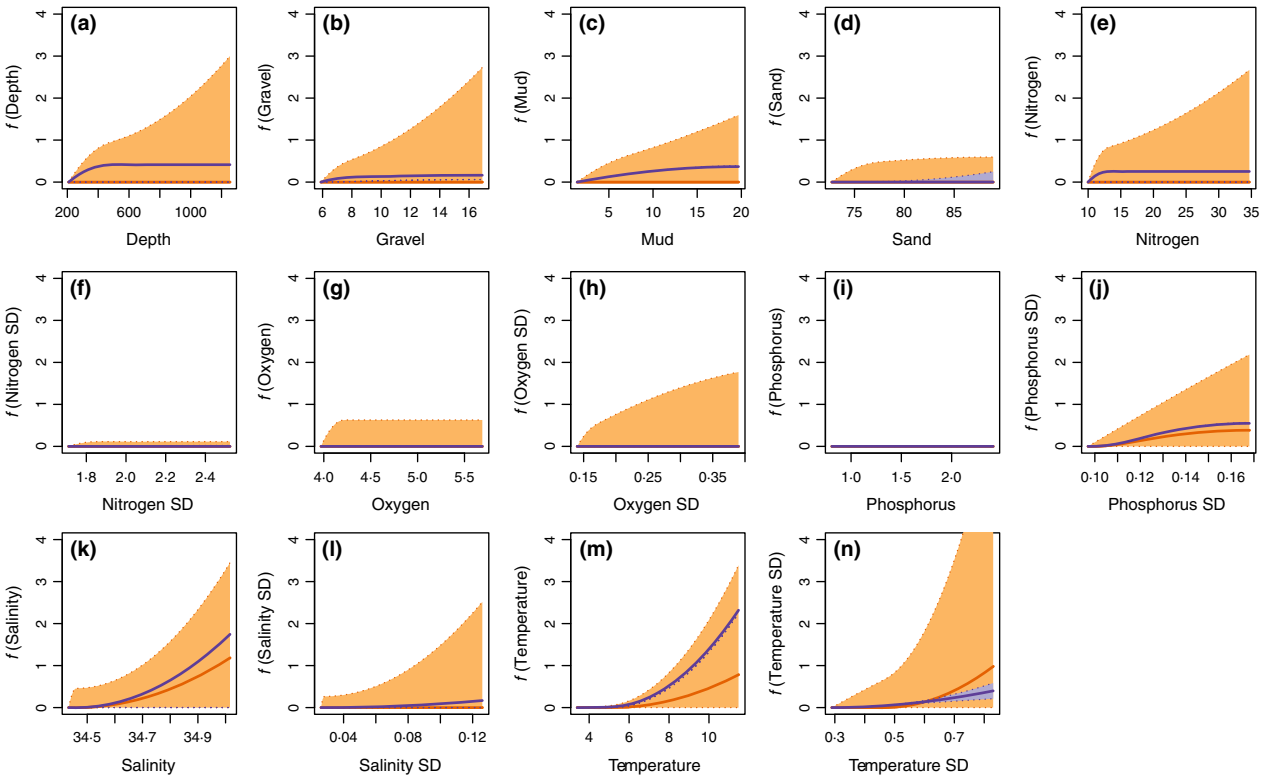

Fig. 4. Partial response plots of generalised dissimilarity modelling (GDM) and Bayesian bootstrap GDM (BBGDM) for Tasmanian benthic invertebrate data. The heavy purple line is the GDM mean estimate and the brown line is the BBGDM mean estimate. The shading is the 95% confidence interval for naive GDM variance (light purple) and 95% credible interval for BBGDM variance (light brown).

Table 1. Generalised dissimilarity modelling (GDM) and Bayesian bootstrap GDM (BBGDM) 'Wald-like' tests on the intercept and sum of I-splines ($\Sigma_{k=1}^{k}\{\beta_{pk}\}$) for Tasmanian benthic invertebrate data. Covariates with low *P*-values are those that are likely to be important in capturing the variances in compositional turnover

	DF	W GDM	*P*-value GDM	W BBGDM	*P*-value BBGDM
Intercept	1·00	0·00	1·00	23·77	0·00*
Depth	3·00	145 383·82	0·00*	0·00	1·00
Gravel	3·00	0·00	1·00	0·00	1·00
Mud	3·00	1 773 417·61	0·00*	0·00	1·00
Sand	3·00	0·00	1·00	0·00	1·00
Nitrogen	3·00	0·00	1·00	0·00	1·00
Nitrogen SD	3·00	0·00	1·00	0·00	1·00
Oxygen	3·00	0·00	1·00	0·00	1·00
Oxygen SD	3·00	0·00	1·00	0·00	1·00
Phosphorus	3·00	0·00	1·00	0·00	1·00
Phosphorus SD	3·00	51 951·26	0·00*	1·13	0·77
Salinity	3·00	0·00	1·00	1·46	0·69
Salinity SD	3·00	0·00	1·00	0·00	1·00
Temperature	3·00	131 916 547·89	0·00*	0·50	0·92
Temperature SD	3·00	3838·31	0·00*	0·24	0·97

W = 'Wald-like' statistic; DF = degrees of freedom (number of I-spline bases); SD = standard deviation.
*$P < 0·05$.

models. In the Tasmanian invertebrate case study, depth, percentage composition of mud, phosphorus SD, temperature and temperature SD were significant variables within GDM (Table 1). However, there was *no* significant spatial or environmental covariate for BBGDM (Table 1). For the Western Australian fish study, all covariates were highly significant within the GDM model (Table SI 4). While for BBGDM, latitude, depth and temperature were all significant variables in explaining compositional turnover (Table SI 4).

In both case studies, model diagnostics showed that models of compositional dissimilarity calculated using our binomial model framework provided an adequate fit to the data (Figs. 5

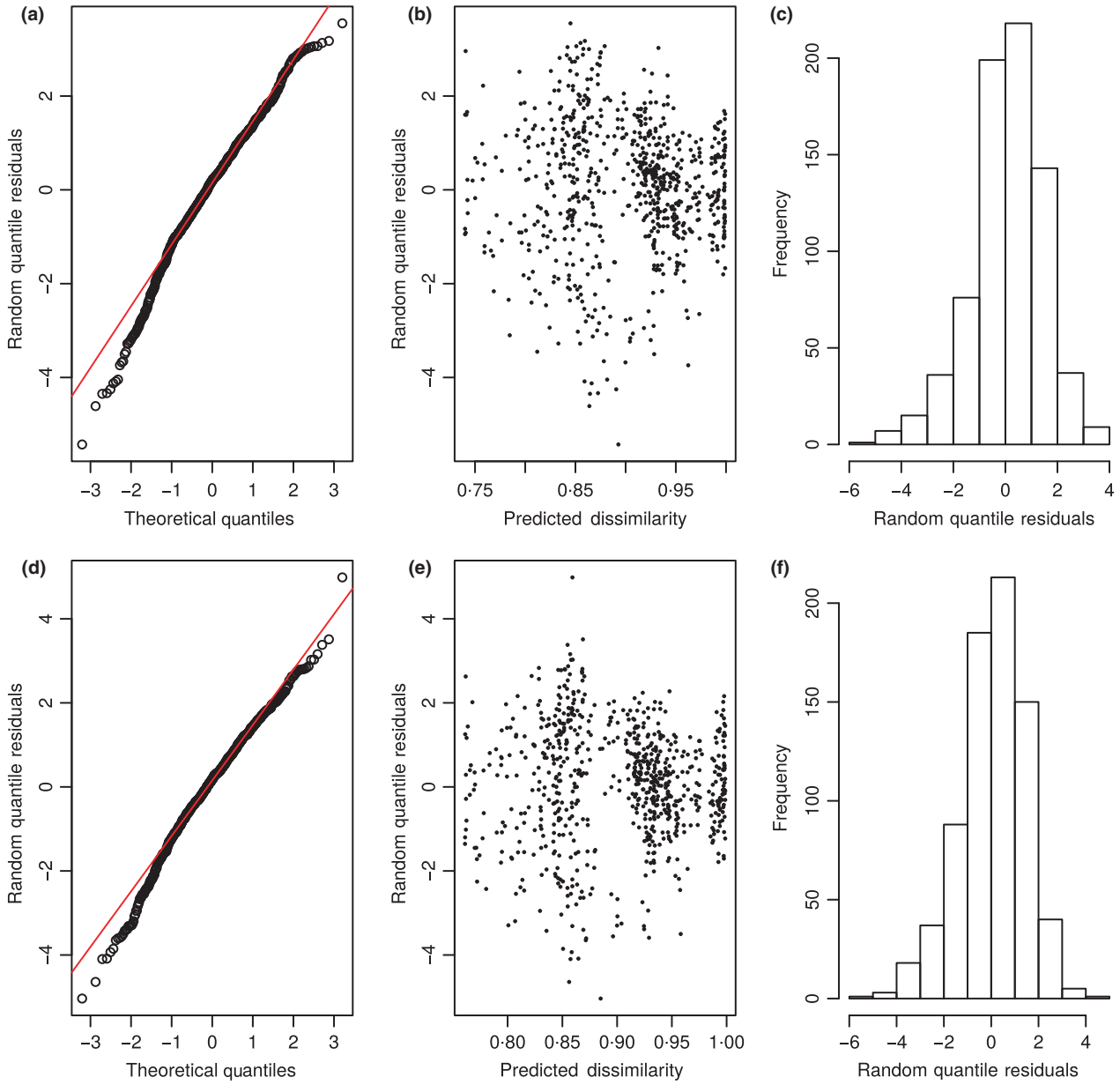

Fig. 5. Model diagnostics from generalised dissimilarity modelling (GDM) and Bayesian bootstrap GDM (BBGDM) for Tasmanian benthic invertebrate data. (a) quantile-quantile plot derived for random quantile residuals for BBGDM; (b) scatter plot of residuals versus the predicted dissimilarity (non-shared species) for BBGDM; (c) a histogram of random quantile residuals from BBGDM; (d) quantile-quantile plot for random quantile residuals for GDM; (e) scatter plot of residuals versus the predicted dissimilarity for GDM; and (f) a histogram of random quantile residuals from GDM.

and SI 5). This suggests that the binomial working model was capturing the dispersion of dissimilarities, and providing a reliable model for these data.

Discussion

The incorporation of a Bayesian bootstrap into GDM provides an assessment of the uncertainty about the fit of the GDM that can be used in inference and prediction. Simulations showed that the estimated variance for a BBGDM was equivalent to the variance of means from multiple simulated model runs. Estimated naive variance derived from GDMs was much lower than expected across simulations (Fig. 3b). This is likely due to the pseudo-replication of observations in GDM models, resulting in an overly steep log-likelihood surface, leading to false confidence about coefficient estimates and highlighting that the variance estimated from a GDM covariance matrix under-represents uncertainty (McCullagh & Nelder 1989). This was the original motivation for undertaking Monte Carlo permutation tests (Ferrier *et al.* 2002), where randomisation was undertaken to test if a variable was significantly different from a null. Here we use the BB to extend GDM methods to include realistic uncertainty and improve inference.

Characterising uncertainty will help improve the robustness of decisions based partly on GDMs. For example, models predicting shifts in community composition under future climate scenarios currently use the mean estimates of dissimilarity, but appear to ignore the variance in predictions (Dunlop *et al.* 2012). Making decisions that are robust to realistic levels of uncertainty requires a realistic characterisation of those uncertainties (Diego *et al.* 2005; Wintle *et al.* 2011). Incorporating the Bayesian bootstrap into the GDM framework makes such a characterisation possible. Capturing model uncertainties, as presented in BBGDM, will be vital for good model selection and inference. Our solution was to assess performance of GDMs and BBGDMs based on partial effect plots, 'Wald-like' tests and model diagnostics.

For the Tasmanian marine invertebrate data set, over-fitting of parameters in GDMs leads to false confidence about covariates driving compositional turnover (Fig. 4). For BBGDM, I-spline estimates based on BB weights suggested that the uncertainty in I-spline estimates was large, making it difficult to produce reasonable inference on the covariates explaining compositional turnover (Fig. 4). 'Wald-like' tests support these findings, suggesting that, for this data set, we can only trust the intercept-only model (Table 1). Previous GDM analyses of this marine invertebrate data showed that compositional turnover was largely driven by depth and salinity (Dunstan *et al.* 2012a). These results are contrary to the findings presented in this paper. This shows that BBGDM's capacity to characterise uncertainty alters the inference made on covariates driving compositional turnover. The capacity for BBGDM to not detect trends in the data might reflect: (i) the variances in species assemblages for deep-sea invertebrate benthic communities, (ii) a lack of a gradient that shapes these differences at the spatial scale of this study (Williams *et al.* 2010a; Dunstan *et al.* 2012a) or (iii) poorly measured covariates that do not reflect what they are attempting to measure (Foster, Shimadzu & Darnell 2012; Stoklosa *et al.* 2015).

For the Western Australian fish data, we demonstrated that BBGDM is capable of capturing a signal when modelling compositional turnover against strong environmental and spatial covariates (Table SI 4). Compared to GDM, BBGDM had fewer variables that significantly contributed to driving compositional turnover (Table SI 4), demonstrating that characterising uncertainty with BBGDM is important (Fig. SI 7). For the Western Australian fish data, the importance of depth, temperature and latitude appear to be consistent with our existing knowledge of processes shaping compositional differences across this region (Williams, Koslow & Last 2001).

Not taking into account the uncertainty in parameter estimates has ramifications for applied outcomes of GDM. For this very reason, many applications of GDM have used permutation tests to assess the significance of variables in their models (Fitzpatrick *et al.* 2013; Jones *et al.* 2015). However, permutation tests assess how a variable differs from a null test, and does not propagate uncertainty through an analysis like BBGDM. For example, using predicted dissimilarities in multivariate ordination or clustering is often used to gain spatial classification or insight into bioregional distributions (e.g. Leathwick *et al.* 2011). These post-processing steps should only be undertaken if the underlying models adequately capture variance in the data. Thus, within the data presented in this manuscript, post-modelling classification would be inappropriate for Tasmanian benthic invertebrate data set. But if so desired, post-processing of predicted dissimilarities could be applied to the Western Australian fish data set, which has significant environmental covariates that drive species turnover.

Bayesian bootstrap GDMs estimate quasi-likelihoods for modelled dissimilarities, meaning that the error distribution of dissimilarities can be fitted without having to strictly conform to a binomial distribution (McCullagh & Nelder 1989). Model diagnostics and residuals based on the GDM and BBGDM working model appear to be appropriate for dissimilarities for both biological data sets (Figs. 4 and SI 5). Throughout, we present a specific form of BBGDM that uses a particular dissimilarity metric. For this metric, analytical results are available that show that the dissimilarity changes with different expectations in individual species and autocorrelation between sites (see Appendix S5). However, it is reasonable to think that researchers could test a suite of data, metrics and model configurations to better understand the model behaviours and uncertainties in their GDM applications. Like GLMs, it is reasonable to further extend the GDM approaches into Shape Constrained Additive Models (Pya & Wood 2014) and beta-regression (Ferrari & Cribari-Neto 2004). Similar to modelling count data, one would expect different data sets to drive the choice of model error distributions (Warton *et al.* 2015).

SUMMARY

This work highlights how we can characterise uncertainty when modelling dissimilarities by including a Bayesian bootstrap extension into GDM. Although this approach cannot resolve the issue of assuming dissimilarities are independent observations in models, we can better assess uncertainty in GDM and bring the model back to a framework that estimates the variance in parameters based on available data. We suggest that future applications of GDM can improve assessment of model uncertainty by using BBGDM.

Authors' contributions

S.W., S.F. and P.D. conceived the ideas and designed methodology; S.W., S.F., P.D. and T.O.H analysed the data; S.W. and S.F. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Acknowledgements

This work is an output of the project of the Marine Biodiversity Research Hub and Environmental Decisions Hub, funded through the Commonwealth National Environmental Research Program (NERP) and administered through the Australian Government's Department of Environment. We thank the Centre of Excellence for Environmental Decisions (CEED) for funding. Thanks also to Michael Bode, Gurutzeta Guillera–Arroita, Rebecca Leaper, Eric Lehmann and Karel Mokany for useful comments and suggestions. We thank Simon Ferrier and an anonymous referee for their useful and constructive comments on earlier drafts of this article.

Data accessibility

Tasmanian seamount marine invertebrate data: CSIRO MarLIN Data Server record no. 6939. http://www.marlin.csiro.au/geonetwork/srv/eng/search#!01b48376-88de-41e4-aa94-348279098c31. Western Australian demersal fish data: CSIRO MarLIN Data Server record no. 4951. http://www.marlin.csiro.au/geonetwork/srv/eng/search#!10c6dd20-45cf-4f5c-a844-4e4465bfffad. To access the relevant data on MarLIN data server, scroll down and select 'Data available via Data Trawler' link. Once you have selected 'Data available via Data Trawler', there is a 'get data' download button at the bottom of the web page. CSIRO Atlas of Regional Seas (CARS) climatology for physical oceanography: https://portal.aodn.org.au/search MARS sediment database: http://dbforms.ga.gov.au/pls/www/npm.mars.search

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Anderson, M.J. (2004) DISTLM v.5: a FORTRAN computer program to calculate a distance-based multivariate analysis for a linear model. Department of Statistics, University of Auckland, New Zealand, **10**, 2016.

Davison, A.C. (1997) Bootstrap methods and their application. *Technometrics*, **42**, 582.

Diego, S., Barbara, S., Regan, H.M., Ben-Haim, Y., Langford, B., Wilson, W.G., Lundberg, P., Andelman, S.J. & Burgman, M.A. (2005) Robust decision-making under severe uncertainty for conservation management. *Ecological Applications*, **15**, 1471–1477.

Dunlop, M., Hilbert, D., Ferrier, S. & House, A. (2012) *The Implications of Climate Change for Biodiversity Conservation and the National Reserve System: Final Synthesis*. CSIRO, Canberra, Australia.

Dunn, P.K. & Smyth, G.K. (1996) Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

Dunstan, P.K. & Foster, S.D. (2011) RAD biodiversity: prediction of rank abundance distributions from deep water benthic assemblages. *Ecography*, **34**, 798–806.

Dunstan, P.K., Foster, S.D. & Darnell, R. (2011) Model based grouping of species across environmental gradients. *Ecological Modelling*, **222**, 955–963.

Dunstan, P.K., Althaus, F., Williams, A. & Bax, N.J.(2012a) Characterising and predicting benthic biodiversity for conservation planning in deepwater environments. *PLoS ONE*, **7**, e36558.

Dunstan, P.K., Bax, N.J., Foster, S.D., Williams, A. & Althaus, F. (2012b) Identifying hotspots for biodiversity management using rank abundance distributions. *Diversity and Distributions*, **18**, 22–32.

Ferrari, S.L.P. & Cribari-Neto, F. (2004) Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**, 799–815.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and Conservation*, **11**, 2309–2338.

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.

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 Biological Sciences/The Royal Society*, **280**, 20131201.

Foster, S.D. & Dunstan, P.K. (2010) The analysis of biodiversity using rank abundance distributions. *Biometrics*, **66**, 186–195.

Foster, S.D., Shimadzu, H. & Darnell, R. (2012) Uncertainty in spatially predicted covariates: is it ignorable? *Journal of the Royal Statistical Society Series C: Applied Statistics*, **61**, 637–652.

Foster, S.D., Givens, G.H., Dornan, G.J., Dunstan, P.K. & Darnell, R. (2013) Modelling biological regions from multi-species and environmental data. *Environmetrics*, **24**, 489–499.

GA (2009) *Australian Bathymetry and Topography Grid*. Geoscience Australia, Canberra, Australia.

Gelman, A., Carlin. J., Stern, H., Dunson, D., Ventari, A. & Rubin, D. (2013) *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC, Boca Raton, FL, USA.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Jaccard, P. (1912) The distribution of the flora in the alpine zone.1. *New Phytologist*, **11**, 37–50.

Jones, M.M., Gibson, N., Yates, C., Ferrier, S., Mokany, K., Williams, K.J., Manion, G. & Svenning, J.C. (2015) Underestimated effects of climate on plant species turnover in the Southwest Australian Floristic Region. *Journal of Biogeography*, **43**, 289–300.

Kodde, D.A. & Palm, F.C. (1986) Wald criteria for jointly testing equality and inequality restrictions. *Econometrica*, **54**, 1243–1248.

Koslow, J.A., Gowlett-Holmes, K., Lowry, J.K., O'Hara, T., Poore, G.C.B. & Williams, A. (2001) Seamount benthic macrofauna off southern Tasmania: community structure and impacts of trawling. *Marine Ecology Progress Series*, **213**, 111–125.

Leathwick, J.R. & Elith, J. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology and Systematics*, **40**, 415–436.

Leathwick, J.R., Snelder, T., Chadderton, W.L., Elith, J., Julian, K. & Ferrier, S. (2011) Use of generalised dissimilarity modelling to improve the biological discrimination of river and stream classifications. *Freshwater Biology*, **56**, 21–38.

Lewis, M. (1999) CSIRO-SEBS (Seamount, Epibenthic Sampler), a new epibenthic sled for sampling seamounts and other rough terrain. *Deep-Sea Research Part I: Oceanographic Research Papers*, **46**, 1101–1107.

McCullagh, P. & Nelder, J. (1989) *Generalized Linear Models*, 2nd edn. Chapman & Hall, London, UK.

Ovaskainen, O., Hottola, J. & Shtonen, J. (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, **91**, 2514–2521.

Polasky, S., Carpenter, S.R., Folke, C. & Keeler, B. (2011) Decision-making under great uncertainty: Environmental management in an era of global change. *Trends in Ecology and Evolution*, **26**, 398–404.

Pya, N. & Wood, S.N. (2014) Shape constrained additive models. *Statistics and Computing*, **25**, 1–17.

Ramsay, J.O. (1988) Monotone regression splines in action. *Statistical Science*, **3**, 425–441.

Ridgway, K.R., Dunn, J.R. & Wilkin, J.L. (2002) Ocean interpolation by four-dimensional weighted least squares – application to the waters around Australasia. *Journal of Atmospheric and Oceanic Technology*, **19**, 1357–1375.

Rubin, D.B. (1981) The bayesian bootstrap. *The Annals of Statistics*, **9**, 130–134.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

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**, 627.

Stoklosa, J., Daly, C., Foster, S.D., Ashcroft, M.B. & Warton, D.I. (2015) A climate of uncertainty: Accounting for error in climate variables for species distribution models. *Methods in Ecology and Evolution*, **6**, 412–423.

Warton, D.I., Wright, S.T. & Wang, Y. (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, **3**, 89–101.

Warton, D.I., Foster, S.D., De'ath, G., Stoklosa, J. & Dunstan, P.K. (2015) Model-based thinking for community ecology. *Plant Ecology*, **216**, 669–682.

Williams, A., Koslow, J.A. & Last, P.R. (2001) Diversity, density and community structure of the demersal fish fauna of the continental slope off Western Australia (20 to 35S). *Marine Ecology Progress Series*, **212**, 247–263.

Williams, A., Althaus, F., Dunstan, P.K., Poore, G.C.B., Bax, N.J., Kloser, R.J. & McEnnulty, F.R. (2010a) Scales of habitat heterogeneity and megabenthos biodiversity on an extensive Australian continental margin (100–1100-m depths). *Marine Ecology*, **31**, 222–236.

Williams, A., Schlacher, T.A., Rowden, A.A. *et al*. (2010b) Seamount megabenthic assemblages fail to recover from trawling impacts. *Marine Ecology*, **31**, 183–199.

Wintle, B.A., McCarthy, M.A., Volinsky, C.T. & Kavanagh, R.P. (2003) The use of bayesian model averaging to better represent uncertainty in ecological models. *Conservation Biology*, **17**, 1579–1590.

Wintle, B.A., Bekessy, S.A., Keith, D.A. *et al*. (2011) Ecological economic optimization of biodiversity conservation under climate change. *Nature Climate Change*, **1**, 355–359.

Received 6 September 2016; accepted 22 November 2016
Handling Editor: David Hodgson

Supporting Information

Details of electronic Supporting Information are provided below.

Appendix S1. Comparison of GDM link functions.

Appendix S2. Description of I-spline bases in GDM.

Appendix S3. Description of methods used to simulate species.

Appendix S4. Western Australian fish case study.

Appendix S5. Expectation of GDM.