Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2014) **23**, 1314–1323

A Journal of Macroecology

Global Ecology and Biogeography

MACROECOLOGICAL METHODS

Generalizing the use of geographical weights in biodiversity modelling

C. Mellin[1,2*], K. Mengersen[3], C. J. A. Bradshaw[2,4] and M. J. Caley[1]

[1]*Australian Institute of Marine Science, PMB no. 3, Townsville MC, Townsville, Qld 4810, Australia,* [2]*The Environment Institute and School of Earth and Environmental Sciences, The University of Adelaide, Adelaide, SA 5005, Australia,* [3]*Queensland University of Technology, School of Mathematical Sciences, Brisbane, Qld 4001, Australia,* [4]*South Australian Research and Development Institute, PO Box 120, Henley Beach, SA 5022, Australia*

*Correspondence: C. Mellin, Australian Institute of Marine Science, PMB no. 3, Townsville MC, Townsville, Qld 4810, Australia.
E-mail: camille.mellin@adelaide.edu.au

ABSTRACT

Aim Determining how ecological processes vary across space is a major focus in ecology. Current methods that investigate such effects remain constrained by important limiting assumptions. Here we provide an extension to geographically weighted regression in which local regression and spatial weighting are used in combination. This method can be used to investigate non-stationarity and spatial-scale effects using any regression technique that can accommodate uneven weighting of observations, including machine learning.

Innovation We extend the use of spatial weights to generalized linear models and boosted regression trees by using simulated data for which the results are known, and compare these local approaches with existing alternatives such as geographically weighted regression (GWR). The spatial weighting procedure (1) explained up to 80% deviance in simulated species richness, (2) optimized the normal distribution of model residuals when applied to generalized linear models versus GWR, and (3) detected nonlinear relationships and interactions between response variables and their predictors when applied to boosted regression trees. Predictor ranking changed with spatial scale, highlighting the scales at which different species–environment relationships need to be considered.

Main conclusions GWR is useful for investigating spatially varying species–environment relationships. However, the use of local weights implemented in alternative modelling techniques can help detect nonlinear relationships and high-order interactions that were previously unassessed. Therefore, this method not only informs us how location and scale influence our perception of patterns and processes, it also offers a way to deal with different ecological interpretations that can emerge as different areas of spatial influence are considered during model fitting.

Keywords
Biodiversity, macroecological modelling, method, non-stationarity, prediction, spatial scale, species distribution modelling.

INTRODUCTION

Truly spatially invariant phenomena are rare in nature (e.g. Bersier *et al.*, 1999). In contrast, spatially dependent patterns and processes are common and illustrated by many examples, including the slope of species–area curves (Lyons & Willig, 2002), the shapes of species richness and altitude relationships (linear or hump-shaped; Rahbek & Graves, 2001) and the location of biodiversity hotspots (Hurlbert & Jetz, 2007). Ecological processes can vary either as a function of the location (non-stationarity; Brunsdon *et al.*, 1998) or of the spatial scale considered. In this regard, scale dependency of ecological patterns is expected to be both general (Lyons & Willig, 2002) and modified by species-specific traits such as body size and/or dispersal capacity (Rahbek, 2005).

The importance of spatial scale in ecology can be understood by considering each of its components: (1) *grain size*, the size of the elementary sampling unit such as transect length or quadrat area; (2) *sampling interval*, the average distance between neighbouring sampling units; and (3) *spatial extent*, the total area

DOI: 10.1111/geb.12203

Table 1 Comparison of geographically weighted regression (GWR) (Brunsdon *et al.*, 1998; Fotheringham *et al.*, 2002) and the spatially weighted models developed in this study.

Criterion and subcriterion	GWR	Spatially weighted models (this study)
Method formulation		
Type of predictors	Quantitative	Quantitative, qualitative, interactions
Error distributions	Gaussian (binomial, Poisson)	Gaussian, binomial, Poisson, gamma, quasi, negative binomial
Risk function	IWLS	IWLS, maximum-likelihood estimation
Spatial weighting	Fixed Gaussian, fixed bi-square, adaptive	Fixed Gaussian (other methods can be implemented)
Ease of implementation		
Software	GWR 4.0, R {spgwr}	R {stats; gbm}
Running time (simulated data)	260 minutes	70 minutes
Type of outputs	Estimated local statistics, model result summary	Flexible (includes these and other user- defined options)
Possible inferences		
Spatial stationarity	✓	✓
Model predictions	✓	✓
Nonlinear relationships	✗	✓
Random effects	✗	✓ (Only with generalized linear mixed-effect models)

IWLS, iteratively reweighted least squares.

under study (Wiens, 1989; Allen & Hoekstra, 1991). Observed variation in an ecological phenomenon over space can depend on the values adopted for each of these three components as well as on the location within the study region, and can therefore affect how well any covariate, from fine- to broad-scale, will explain this variation over space. One might expect *a priori* that variation in species richness across a fine-grain and small-extent sampling regime might be predicted well by a fine-scale physical variable, such as substratum structure (Pittman *et al.*, 2004). At the other extreme, species richness at broader spatial scales (e.g. regional or global) is likely to be captured better by broader-scale variables, such as temperature gradients, or possibly spatial predictors such as latitude and longitude (e.g. Caley & Schluter, 1997; Mellin *et al.*, 2010). Sampling and pattern scales are thus ineluctably intertwined (Hutchinson, 1953; Levin, 1992), and sampling at a scale that matches physical and biological patterns is necessary to capture the underlying mechanisms that shape communities and build better distribution models, irrespective of whether they are used for description or prediction.

Geographically weighted regression (GWR) is a useful and widely adopted method for exploring the degree to which species–environment relationships vary according to location and at different spatial scales (Brunsdon *et al.*, 1998; Fotheringham *et al.*, 2002). GWR is a local regression technique whereby an ordinary least-squares regression is fitted around a focal point (i.e. the 'regression point' defined by Fotheringham *et al.* 2002), with data closer to the focal point weighted more heavily in the local regression than data farther away. The weight assigned to each datum decreases as the distance to the focal point increases. Whereas the outputs of GWR are mostly insensitive to the weighting function used, they are sensitive to the bandwidth of the chosen weighting function, thus determining the effective area of influence around the focal point considered during model calibration (Fotheringham *et al.*, 2002) (see Fig. S1 in Supporting Information for an example of a fixed Gaussian spatial kernel). This process is repeated with all data in turn being considered as the focal point. A unique weighting scheme, defined within the neighbourhood of the focal point, results in a unique solution for each local regression.

Local regression techniques such as GWR that use spatial weighting are useful for providing a first assessment of non-stationarity and the effects of spatial scale in ecological data (Da Silva Cassemiro *et al.*, 2007; Hawkins, 2012). By modifying the bandwidth used in GWR, i.e. the extent to which distant locations contribute to the model, one can assess whether species–environment relationships are contingent on the area of spatial influence considered during model fitting. Another possible, although somewhat less explored, option is to use GWR to explore how model support varies across spatial scales as the area of spatial influence is altered. However, because it is based on ordinary least squares, GWR as usually defined and implemented remains constrained by limiting assumptions such as its inability to include qualitative factors, account for nonlinear relationships (Austin, 2007) or handle non-Gaussian error distributions (but see example extensions to Poisson and binomial distributions in Fotheringham *et al.*, 2002, and GWR logistic regression models in Osborne *et al.*, 2007). Moreover, the issue of multicollinearity among GWR coefficients associated with different predictors (Wheeler & Tiefelsdorf, 2005) remains unresolved. A possible way to overcome these limitations would be to implement the concept underpinning GWR in more flexible modelling techniques, such as those based on machine learning (Table 1).

Here we present a generalized method for investigating how species–environment relationships vary as the effective area of spatial influence is altered, by generalizing the concepts of local regression and spatial weighting to virtually any modelling technique that can accommodate differential weights among observations, including machine learning methods. This method can be applied to species distribution models (or niche models) as

well as macroecological models of, for example, species richness or total abundance (Terribile *et al.*, 2009; Guisan & Rahbek, 2011). We thus jointly refer to both types of model as 'biodiversity modelling' techniques. While observation weighting is broadly applicable across all biodiversity modelling techniques, here we illustrate its application to two of the most commonly used and powerful species distribution modelling techniques, namely generalized linear models (GLMs) and boosted regression trees (BRTs). Of all the statistical techniques used to model biodiversity, these are some of the most popular and easy to implement (Li & Wang, 2013), making them good candidates for illustrating the generalization of the use of spatial weights in this context. Using a simulated dataset, we show (1) how implementing this spatial weighting method in a classic least-squares (linear) model framework successfully reproduces GWR results and (2) how, by implementing it in more flexible frameworks, greater information can be recovered about ecological complexity such as scale-dependent interactions among environmental covariates or nonlinear species–environment relationships.

METHODS

Overview of the method, its development and application

We developed a spatially explicit weighting procedure to simulate a varying area of influence around a focal observation. We defined the focal datum as the observation assigned the greatest weight, using each datum in the full dataset successively as the focal observation. We used a fixed Gaussian density function to downweight progressively the influence of observations the greater their distance from the focal observation, with the rate of decay defined by the bandwidth (b) of the Gaussian function, and within models fitted using all observations (Fig. S1). For any given focal observation, increasing b simulates flattening of the surface of influence around this point by evening the weights across the grid. In this way, the effective area of influence or size of the observation window centred on this point varies; on a completely flat surface all points contribute equally to model fitting, whereas on a highly peaked surface only closely neighbouring points will exert much influence.

The procedure involves a separate local regression for each focal observation, fitted using all (weighted) data, the (fixed) weighting scheme being defined individually for each bandwidth (Appendix S1, Fig. S2). Model performance indices are then aggregated across local regressions at each scale and the importance of different predictors is compared among scales. We first develop the method within a least-squares framework using simulated data, for which the scale-specific influence of each predictor is known, and compare the outcomes of the resulting spatially weighted linear models (LMs) with those of a GWR. We then extend the use of spatial weights to other types of regression, namely GLMs and BRTs as examples, and provide the R code to apply the approach to any type of regression that can accommodate uneven observation weighting.

Simulated dataset

We simulated a spatially explicit dataset consisting of three predictors including seabed slope (Slope), benthic irradiance (Irradiance) and sea surface temperature (Temperature) that respectively captured local, subregional and regional variation of a single biological response variable, species richness (S). To achieve this, we first modelled S as a function of Slope, Irradiance and Temperature for a set of observations ($n = 163$) and used the resulting model to predict species richness across the study area, thereby defining our simulated dataset across a regular grid based on known influences of all predictors.

The scale of spatial variation for each predictor was captured by the correlogram of Moran's I autocorrelation coefficient as a function of increasing distance between observations, exceeding the 0.05 threshold up to a distance of 0.06° latitude for Slope, to 0.09° latitude for Irradiance and 0.13° latitude for Temperature (Fig. 1). These patterns indicate that Slope varies over fine spatial scales, Irradiance over intermediate spatial scales and Temperature over broad spatial scales. We derived all variables across the same 0.5° × 0.5° grid ($n = 2500$ observations at a 0.01° resolution) from observed data across the Torres Strait, Australia, and scaled these values between 0 and 1 to standardize comparisons between them. The Torres Strait (Fig. S3) covers approximately 60,000 km^2 between Papua New Guinea and Australia, and comprises a total of 1295 individual reefs (Haywood *et al.*, 2007).

We then predicted species richness (S) for each cell across the same grid as a function of Slope, Irradiance and Temperature in addition to a Gaussian random error. To do this, we first obtained S, the observed total number of species sampled per site, from epibenthic sled samples collected at 163 locations over the study area (see Pitcher *et al.* (2007) for a detailed description) and that sampled 15 sessile phyla, mostly represented by Porifera, Cnidaria and Chlorophyta (67% of the total biomass). We then compared linear, quadratic and cubic relationships between observed S and each predictor (Slope, Irradiance and Temperature) at the sampled locations, and combined the top-ranked GLM based on Akaike's information criterion corrected for small sample sizes (AIC_c) (Burnham & Anderson, 2002, 2004) into a model set including the null model, paired combinations and the sum of all predictors (full model). We derived the model-averaged predictions of species richness (based on AIC_c weights, $wAIC_c$), added a random error $\varepsilon \sim N\,[0, 5]$ and used this to predict for every grid cell over the study area, thereby defining continuous S over the same grid ($n = 2500$), with known influences of all predictors.

Applying spatial weighting to simulated data

We applied local models to the simulated dataset to assess the potential non-stationarity in species–environment relationships resulting from a combination of (1) the inherent uncertainty (standard deviation) in model-averaged predictions and (2) the random error purposely embedded in the predictions. We constructed a spatially weighted LM and GLM with S (predicted

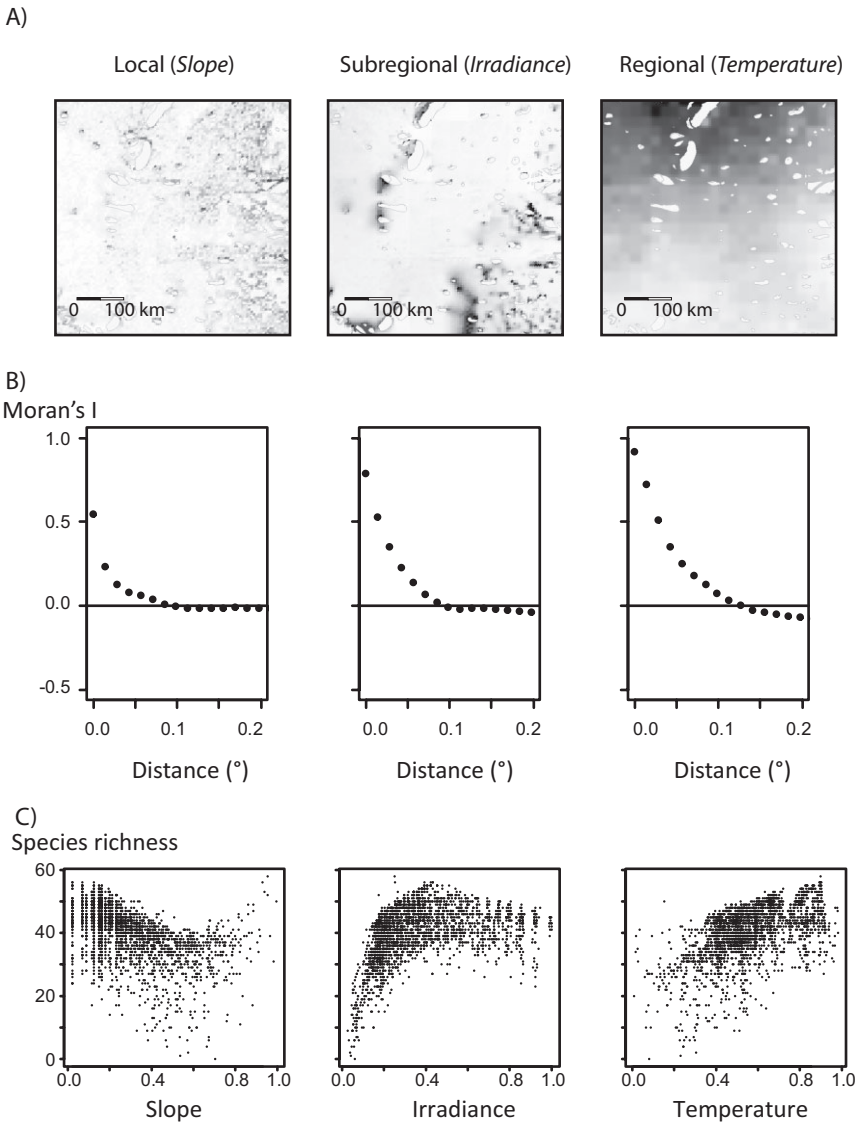

Figure 1 (A) Predictors of the simulated dataset, with seabed slope (*Slope*) varying at a local scale, benthic irradiance (*Irradiance*) varying at a subregional scale and sea surface temperature (*Temperature*) varying at a regional scale across this particular section of the Torres Strait, Australia. All predictors were scaled between 0 (light) and 1 (dark). (B) Correlogram (Moran's *I* autocorrelation coefficient) depicted as a function of increasing distance (in degrees of latitude) classes. (C) Relationships between each predictor and species richness.

species richness) as the response variable and a combination of Slope, Irradiance and Temperature as predictors based on the spatial weighting procedure described above (Appendix S1; see also R codes in Appendix S2). Given the extent of the study area ($0.5° \times 0.5°$) and the linear unit used (degrees) in this case, we used the bandwidths $b \in \{0.05; 0.1; 0.15; \ldots ; 0.5°\}$. A seven-model set consisted of one model for each predictor Slope, Irradiance and Temperature (linear terms only), pairwise combinations of predictors, the full model and the null model. For the spatially weighted GLM, we assumed a Poisson distribution with a log-link (as commonly expected for count data) and checked the normal distribution of model residuals using the normal scores of standardized residual deviance (Breslow, 1996). Model performance indices included AIC_c to provide an index of Kullback–Leibler (K-L) information loss that we used to assign relative strengths of evidence to the different competing models, and the percent deviance (*De*) in *S* explained by the model, which provided an index of the model's goodness-of-fit.

We used the 'weights' option of the lm and glm function in the package stats in R (R Development Core Team, 2013) for the spatially weighted LM and GLM, respectively, to take into account different weights for different observations in the models.

We compared results of the spatially weighted LM and GLM with those given by GWR for each model in the set using the same spatial weighting scheme (i.e. fixed Gaussian) and the same bandwidth range. The spatially weighted LM and GWR are equivalent in terms of model formulation, the main difference being the software package used to run them; however, we maintain this terminology for the purpose of determining whether or not using differential weights in the lm function results in similar estimates as based on the GWR package. Both in the spatially weighted GLM and the GWR, we log-transformed and assumed a Poisson distribution for the response variable *S*. Following Fotheringham *et al.* (2002), we defined the 'best' bandwidth as the bandwidth that minimized AIC_c. For the best

bandwidth, and for each technique, we computed and compared local parameter estimates and the distribution of model residuals using the normal scores of standardized residual deviance (*Q*–*Q* plots). We also ran the same analysis with even weighting and we compared the results of the 'global' GWR with those of the classic (unweighted) LM and GLM. We ran GWR using the software package GWR 4.0 (Nakaya *et al.*, 2005) with options (other than described above) set to the default values.

We explored the extent to which a spatially weighted BRT could provide an improvement over GWR and the two linear model variants by detecting nonlinear relationships between the response variable and the predictors, or interactions among predictors. We first ran the spatially weighted BRT using the same bandwidth range as defined above using the 'site.weights' option of the gbm.fixed function in brt.functions.R provided by Elith *et al.* (2008) (R package {gbm}), and assessed the percentage contribution of each predictor Slope, Irradiance and Temperature at each bandwidth. For the best bandwidth, we pooled results across local BRT and derived the mean and confidence interval of the marginal effect of each predictor and their interactions at that particular bandwidth, the percentage contribution of each predictor in explaining variance in *S* and their interactions (Elith *et al.*, 2008). BRTs can be computationally intensive, so we fixed the number of trees at 1000 to ensure a reasonable processing time. We limited the tree complexity to three to avoid overfitting and used a learning rate of 0.001. The optimal number of trees was defined using cross-validation (Elith *et al.*, 2008).

We tested for edge effects on model performance indices that could result from truncated weight distributions as focal observations approached the edge of a grid (Appendix S1). Finally, we constructed a simulation to examine how the degree of spatial clustering influenced the performance of the method and its results (Appendix S1) using spatially weighted GLMs.

RESULTS

Construction and spatial scale analysis of the simulated dataset

Species richness (*S*) (40.0 ± 8.5 species, mean ± standard deviation) was best explained by a combination of all three variables, with quadratic and cubic terms for Slope and Irradiance, and only quadratic for Temperature (Table 2), confirming the strong curvilinear relationships observed between *S* and each predictor (Fig. 1). The full model explained 15.7% of the deviance in *S* and received the highest support based on $wAIC_c$ (*c.* 1), and therefore contributed most to model-averaged predictions of *S*.

When we applied a spatially even weighting scheme in models predicting *S*, global GWR provided the same results as the corresponding LM in terms of deviance, *De*, AIC_c and $wAIC_c$ (results not shown) and similar results to those given by the corresponding GLM in terms of *De*-, $wAIC_c$- and AIC_c-based model ranking (Table 3), although raw deviance and AIC_c differed between techniques.

Table 2 Summary of generalized linear models (GLM) used to generate the simulated dataset and predict species richness (*S*) as a function of spatial predictors at local (seabed slope; Slope), subregional (benthic irradiance; Irr) and regional (sea surface temperature; Temp) scales. We considered both second- or third-degree polynomial functions (Poly) of each predictor. We ranked models based on Akaike's information criterion corrected for small sample sizes (AIC_c).

Model	*k*	LL	$wAIC_c$	*De*
$S \sim \text{Poly}(\text{Slope}^3) + \text{Poly}(\text{Irr}^3) + \text{Poly}(\text{Temp}^2)$*	9	−1391.1	1.000	15.7
$S \sim \text{Poly}(\text{Slope}^3) + \text{Poly}(\text{Irr}^3)$	7	−1409.6	< 0.001	14.1
$S \sim \text{Poly}(\text{Irr}^3) + \text{Poly}(\text{Temp}^2)$	6	−1429.8	< 0.001	12.3
$S \sim \text{Poly}(\text{Irr}^3)$	4	−1452.7	< 0.001	10.3
$S \sim \text{Poly}(\text{Slope}^3) + \text{Poly}(\text{Temp}^2)$	6	−1490.9	< 0.001	6.9
$S \sim \text{Poly}(\text{Slope}^{3)})$	4	−1506.1	< 0.001	5.6
$S \sim \text{Poly}(\text{Temp}^2)$	3	−1545.0	< 0.001	2.2
$S \sim 1$	1	−1570.3	< 0.001	–

k, number of parameters; $wAIC_c$, AIC_c weight; LL, maximum log-likelihood; *De*, percentage deviance explained (a measure of the structural goodness of fit of the model). Model sequences are ordered by increasing $wAIC_c$.

*Model formula (including model coefficients) is: $S \sim 2.95 + 0.13 \times \text{Slope} - 3.27 \times \text{Slope}^2 + 4.10 \times \text{Slope}^3 + 2.99 \times \text{Irr} - 4.53 \times \text{Irr}^2 + 2.08 \times \text{Irr}^3 + 1.24 \times \text{Temp} - 0.69 \times \text{Temp}^2$.

Analysis of the simulated dataset using spatially weighted LM and GLM

We obtained comparable patterns in *De* and AIC_c for GWR and the spatially explicit linear models as bandwidth (*b*) increased (Fig. 2). In both cases, and for all models, AIC_c was smallest at the smallest bandwidth ($b = 0.05$) and gave the same model ranking across increasing bandwidths, although raw AIC_c differed. The full model received the strongest support at all bandwidths, whereas the deviance explained by the Slope + Irradiance model decreased with increasing bandwidth, and that explained by the Irradiance + Temperature model increased. We obtained contrasting results for the spatially explicit GLM; while AIC_c remained minimal at the smallest bandwidth, the deviance explained by the full model decreased with increasing bandwidth. According to *De*, the scale-specific spatial pattern in each individual predictor and its influence on *S* was more appropriately captured by this model than the previous one: the deviance explained by Slope rapidly decreased, that of Irradiance decreased but at larger bandwidths, whereas that of Temperature increased with increasing bandwidth. However, this pattern was less readily detected based on AIC_c.

For the best bandwidth ($b = 0.05$), the examination of the normal scores of standardized residual deviance revealed that the residuals of GWR and the spatially weighted LM were strongly skewed towards high values, whereas those of the spatially weighted GLM were closer to a normal distribution (Fig. 2). Local estimates of model parameters were similar for all three modelling techniques, apart from minor (both positive

Table 3 Comparison of unweighted (i.e. global) geographically weighted regressions (GWR) and generalized linear models (GLMs) used to predict species richness (S) as a function of spatial predictors at local (seabed slope; Slope), subregional (benthic irradiance; Irr) and regional (sea surface temperature; Temp) scales using the simulated dataset. We ranked models based on Akaike's information criterion corrected for small sample sizes (AIC_c).

Model	k	Deviance	De	AIC_c	$wAIC_c$
GWR					
$S \sim$ Slope + Irr + Temp	4	2183.3	57.8	2191.3	1
$S \sim$ Irr + Temp	3	2861.5	44.7	2867.6	< 0.001
$S \sim$ Slope + Irr	3	3325.8	35.7	3331.9	< 0.001
$S \sim$ Temp	2	3853.2	25.5	3857.2	< 0.001
$S \sim$ Slope	2	3969.1	23.3	3973.1	< 0.001
$S \sim$ Irr	2	4525.6	12.5	4529.6	< 0.001
$S \sim 1$	1	5172.5	0.0	5174.5	–
GLMs					
$S \sim$ Slope + Irr + Temp	4	2141.4	57.9	15615.6	1
$S \sim$ Irr + Temp	3	2813.7	44.7	16285.8	< 0.001
$S \sim$ Slope + Irr	3	3270.3	35.8	16742.5	< 0.001
$S \sim$ Temp	2	3799.0	25.4	17269.2	< 0.001
$S \sim$ Slope	2	3906.8	23.3	17376.9	< 0.001
$S \sim$ Irr	2	4456.3	12.5	17926.5	< 0.001
$S \sim 1$	1	5092.5	0.00	18560.7	–

k, number of parameters; $wAIC_c$, AIC_c weight; LL, maximum log-likelihood; De, percentage deviance explained (a measure of the structural goodness of fit of the model). Model sequences are ordered by increasing $wAIC_c$.

and negative) differences (Fig. 3). Although parameter estimates were multicollinear for all three techniques, absolute values for the Spearman's correlation coefficients were lower for the spatially weighted LM and GLM than for GWR in most cases (Table S1).

Analysis of the simulated dataset using a spatially weighted BRT

The use of a spatially weighted BRT resulted in the same pattern of change in predictor ranking with increasing spatial scale, as with a spatially weighted GLM, even though the scales at which transitions occurred in the contribution of each model differed slightly, and Irradiance only outperformed Temperature for $b \approx 0.2$ (Fig. 4). However, using a spatially weighted BRT allowed us to detect nonlinear relationships between S and Slope (42% contribution to the total deviance explained; Fig. 4), and to a lesser extent, between S and Irradiance (34% contribution). Although we detected some interactions among individual predictors, in all cases they contributed < 2% of the total deviance explained. We found no evidence for edge effects or for effects of the degree of spatial clustering on the performance and results of the method (Appendix S1).

DISCUSSION

Recognizing the importance of non-stationary and scale-dependent ecological processes and building on the local regression concept of the geographically weighted regression, our generalized spatial weighting procedure is simple and broadly applicable to most biodiversity modelling techniques (see scripts adaptable to other techniques in Appendix S2). Based on a simulated dataset with curvilinear relationships between the response and the predictors – as is often the case with ecological data (Austin, 2007) – this method was able to account for such ecological complexities while providing an assessment of spatial non-stationarity and scale effects. In particular, spatially weighted BRTs detected scale-specific nonlinear relationships between the predictors and the response and potential interactions among predictors, even though these effects were weak in the simulated dataset. Given these findings, we suggest that: (1) GWR is an efficient and straightforward local regression technique for assessing spatial non-stationarity and scale effects as long as residuals are normally distributed; (2) alternatively, or if other error distributions are expected, spatially weighted GLMs can be more useful; and (3) if scale-dependent nonlinear relationships or high-order interactions are expected, spatially weighted BRTs should be the preferred option. In a hypothesis-testing framework, however, GLMs are often more appropriate than BRTs; this might also apply to spatially weighted GLMs compared with spatially weighted BRTs. Using the most appropriate method for the data available and the question being asked should help circumvent competing ecological interpretations of variable support that can emerge as the region under study and/or area of influence is altered.

Results of spatially weighted GLMs differed from those given by GWR, probably due to the use of different fitting algorithms: maximum likelihood estimation in GLMs (McCullagh & Nelder, 1989) and the iteratively reweighted least-squares algorithms in GWR (Fotheringham *et al.*, 2002). In this case, results of spatially weighted GLMs appear more relevant, with the top-ranked models based on AIC_c also explaining the most deviance in the response, and a distribution of residuals closer to normality, although the 'best' model based on AIC_c was less discernible than for the GWR. Different model performance indices have different properties and limitations; for example, AIC_c can be affected by spatial autocorrelation (Diniz-Filho *et al.*, 2008), so we recommend considering multiple indices while selecting the most appropriate modelling technique. A more surprising result, given that spatially weighted LMs and GWR are based on the same algorithm, was the difference between outputs from these two techniques, which is likely to be a consequence of using difference statistical programs. Furthermore, in the calculation of AIC_c, GWR uses the effective number of parameters – a function of the trace of the hat matrix (Fotheringham *et al.*, 2002) – which might explain the difference in absolute AIC_c between spatially weighted LMs and GWR. Finally, despite multicollinearity among parameters for all the spatially weighted LMs, GLMs and GWR we examined here, the absolute

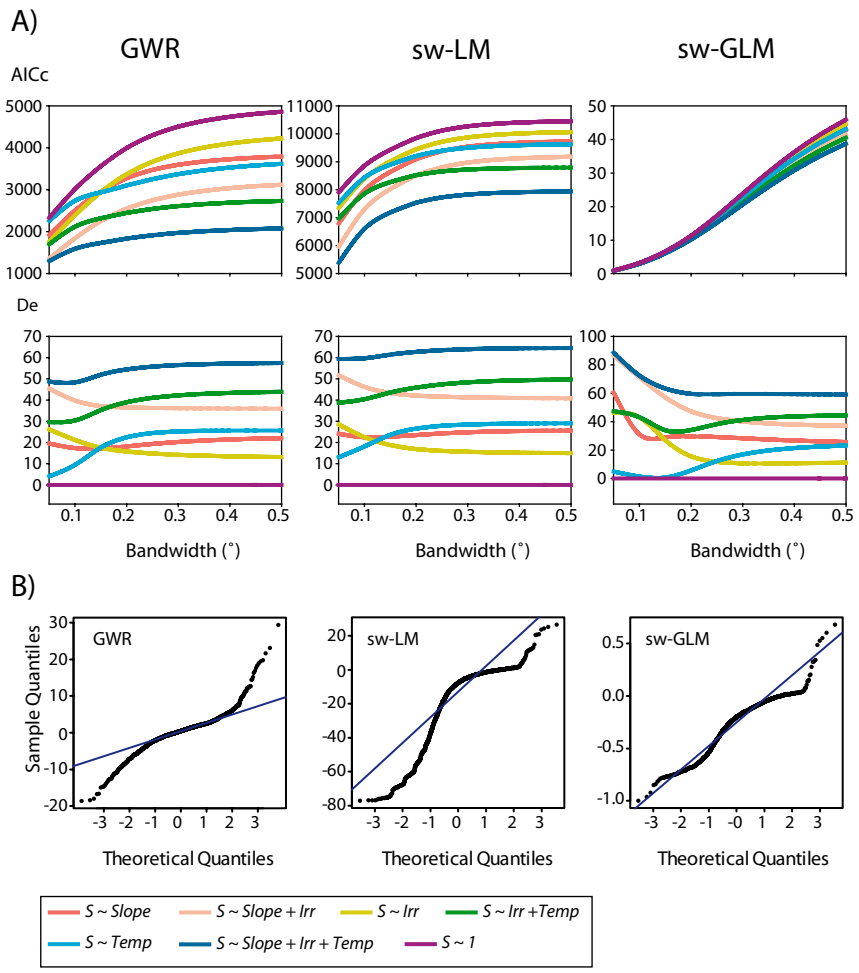

Figure 2 (A) Spatial scale analysis of the simulated dataset comparing model support from seven models using geographically weighted regression (GWR; left), a spatially weighted linear model (sw-LM; middle) and a spatially weighted generalized linear model (sw-GLM; right). Shown are Akaike's information criterion corrected for small sample sizes (AIC_c; top) and the percentage deviance explained by each model (*De*, bottom). Spatial scale varied based on a Gaussian weighting with bandwidth, as shown in Fig. S1. Model predictors include seabed slope (Slope; blue) varying at a local scale, benthic irradiance (Irradiance; green) varying at a subregional scale and sea surface temperature (Temperature; red) varying at a regional scale as depicted in Fig. 1. (B) Normal scores of standardized residual deviance (*Q*–*Q* plots) for the geographically weighted regression (GWR; top panel), spatially weighted linear model (sw-LM; middle panel) and spatially weighted generalized linear model (sw-GLM; bottom panel) done on the simulated dataset with best bandwidth $b = 0.05°$ latitude.

correlation coefficients were generally lower for LMs and GLMs, providing support for the use of these methods (Wheeler & Tiefelsdorf, 2005).

Future studies could explore alternative spatial weighting methods. For example, the adaptive spatial kernel option in GWR (Fotheringham *et al.*, 2002) or Markov random fields (e.g. Kindermann & Snell, 1980), determined either by a distance metric or a nearest-neighbour formulation, might be useful for larger or more heterogeneous regions. Such utility could arise where the distance metric approximates a step function, which is zero in non-neighbouring areas. Such an approach would have the advantage of inducing sparseness in the covariance matrices, which can reduce computational time with efficient programming (Stanaway *et al.*, 2011). The bandwidth could also be treated as a model parameter and optimized during model fitting; however, this option might be computationally challenging and detract from the simplicity of the approach we propose. Finally, different schemes could be used to aggregate the site-specific indices of model fit, reflecting different goodness-of-fit criteria. For example, a weighted-average scheme could be used to allow for preferential fits to certain geographic regions, areas of high species diversity or sites with less estimation uncertainty. This can be done regardless of the manner in which the spatial weights are generated.

We simulated a simple dataset for the purposes of exposition and development; however, one should bear in mind the constraints imposed by more complex, real ecological data and use caution when interpreting the results or making predictions for new areas. For instance, our use of three predictors for illustrative purposes is unlikely to capture complex interactions and collinearity with other predictors left out of the models. Furthermore, we assessed the extent to which site clustering over a regular grid affected the performance of our method; however, future studies should evaluate similar effects in spatially constrained sampling designs (e.g. along roads or reef edges), common in ecological datasets. Finally, when there is evidence for non-stationarity, extrapolation should be avoided because, by definition, species–environment relationships will vary locally and predictions remain highly uncertain where no information is available on the strength or direction of such relationships.

A few studies have examined the effects of the area of influence on model performance and the issue of spatial non-stationarity by using GWR (Lieske & Bender, 2009; Murphy *et al.*, 2011; Gouveia *et al.*, 2013). Local models tend to perform better than regional models, and a model's explanatory power often decreases with the area of spatial influence (Foody, 2004; Bickford & Laffan, 2006; Osborne *et al.*, 2007). This decay of

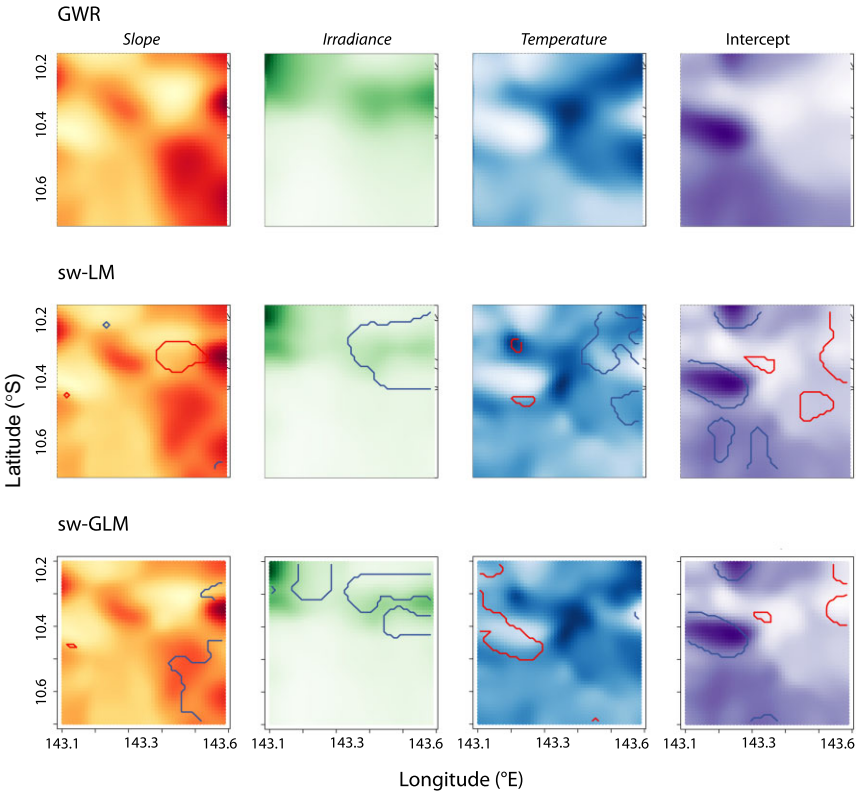

Figure 3 Local parameter estimates across the study area for predictors of the simulated dataset including seabed slope (Slope; blue) varying at a local scale, benthic irradiance (Irradiance; green) varying at a subregional scale and sea surface temperature (Temperature; red) varying at a regional scale, estimated using geographically weighted regression (GWR; top), a spatially weighted linear model (sw-LM; middle) and a spatially weighted generalized linear model (sw-GLM; bottom) of species richness for the best bandwidth ($b = 0.05°$ latitude). Parameter values obtained for each predictor and based on each method were scaled between 0 (light) and 1 (dark). For sw-LM and sw-GLM, contour lines outline areas with a difference of < -0.1 (blue) or > 0.1 (red) from GWR estimates.

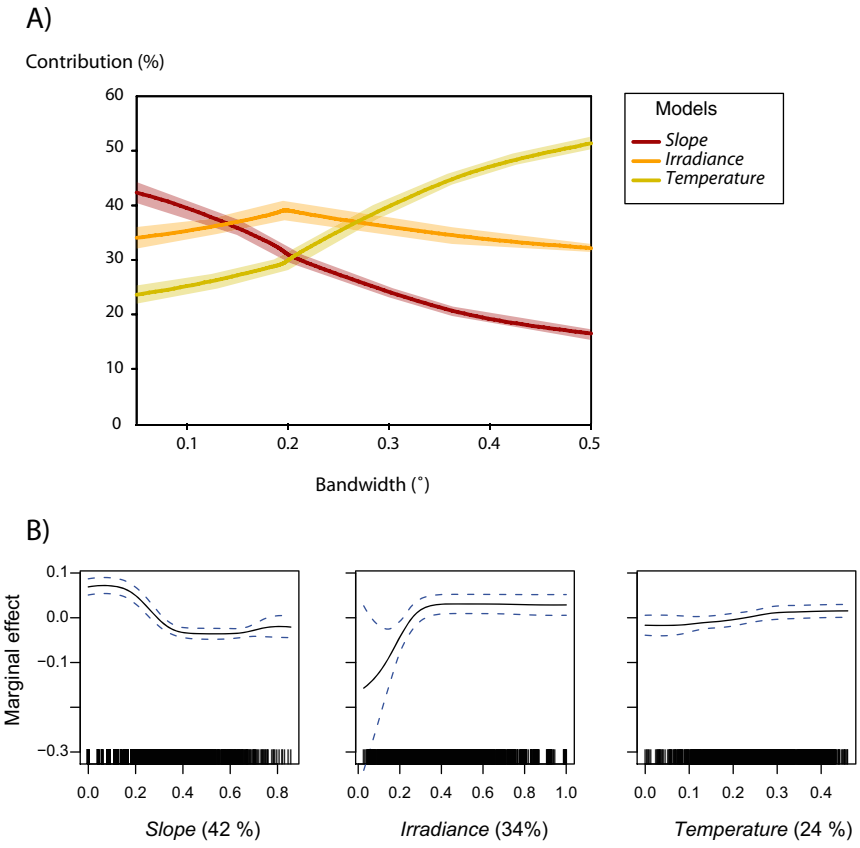

Figure 4 Spatial scale comparisons from spatially weighted boosted regression trees predicting species richness (S) as a function of the predictors Slope, Irradiance and Temperature. (A) Importance of each predictor (expressed in percentage contribution) with increasing area of influence (i.e. bandwidth). Envelopes indicate standard deviations. (B) Marginal effects estimated using a 0.05° latitude bandwidth. Numbers in brackets represent the contribution of each predictor to the total deviance explained in S.

power with increasing bandwidth corroborates our results, although the deviance explained only decreased with the area of influence for spatially weighted GLMs. However, other studies report the opposite pattern, whereby regional models accounting for large-scale gradients outperform local models (Rosa *et al.* 2008; Mellin *et al.*, 2010). These contrasting results demonstrate a variety of possible patterns, and that the optimal spatial scale for analysis will be dependent on the system and the question being asked. For example, incomplete sampling of long environmental gradients is likely to bias our perception of environmental correlates of species richness (Rahbek, 2005). While there is no universal guideline for choosing the optimal spatial scale to consider (Rahbek, 2005), spatial weights can, in practice, assist in identifying the most relevant predictors – whether the objective is to predict biodiversity patterns over entire regions (e.g. Mellin *et al.*, 2010) or achieve the highest predictive accuracy possible at finer scales (e.g. Pittman *et al.*, 2007).

ACKNOWLEDGEMENTS

This work was done for the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Research Program (NERP; http://www.nerpmarine.edu.au). We are grateful to two anonymous referees, and to W. Venables and R. Pitcher for helpful comments on the analyses and manuscript.

REFERENCES

Allen, T.F.H. & Hoekstra, W.H. (1991) Role of heterogeneity in scaling of ecological systems under analysis. *Ecological heterogeneity* (ed. by J. Kolasa and S.T.A. Pickett), pp. 47–68. Springer, Berlin.

Austin, M. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1–19.

Bersier, L.F., Dixon, P. & Sugihara, G. (1999) Scale-invariant or scale-dependent behavior of the link density property in food webs: a matter of sampling effort? *The American Naturalist*, **153**, 676–682.

Bickford, S.A. & Laffan, S.W. (2006) Multi-extent analysis of the relationship between pteridophyte species richness and climate. *Global Ecology and Biogeography*, **15**, 588–601.

Breslow, N.E. (1996) Generalized linear models: checking assumptions and strengthening conclusions. *Journal of Statistics and Applications*, **8**, 23–41.

Brunsdon, C., Fotheringham, S. & Charlton, M. (1998) Geographically weighted regression – modelling spatial non-stationarity. *Journal of the Royal Statistical Society Series D–the Statistician*, **47**, 431–443.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a practical information theoretic approach*, 2nd edn. Springer-Verlag, New York.

Burnham, K.P. & Anderson, D.R. (2004) Multimodel inference – understanding AIC and BIC in model selection. *Sociological Methods and Research*, **33**, 261–304.

Caley, M.J. & Schluter, D. (1997) The relationship between local and regional diversity. *Ecology*, **78**, 70–80.

Da Silva Cassemiro, F.A., De Souza Barreto, B., Rangel, T.F. & Diniz-Filho, J.A.F. (2007) Non-stationarity, diversity gradients and the metabolic theory of ecology. *Global Ecology and Biogeography*, **16**, 820–822.

Diniz-Filho, J.A.F., Rangel, T.F. & Bini, L.M. (2008) Model selection and information theory in geographical ecology. *Global Ecology and Biogeography*, **17**, 479–488.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted trees. *Journal of Animal Ecology*, **77**, 802–813.

Foody, G.M. (2004) Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. *Global Ecology and Biogeography*, **13**, 315–320.

Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2002) *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley, Chichester.

Gouveia, S.F., Hortal, J., Cassemiro, F.A., Rangel, T.F. & Diniz-Filho, J.A. (2013) Nonstationary effects of productivity, seasonality, and historical climate changes on global amphibian diversity. *Ecography*, **36**, 104–113.

Guisan, A. & Rahbek, C. (2011) SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, **38**, 1433–1444.

Hawkins, B.A. (2012) Eight (and a half) deadly sins of spatial analysis. *Journal of Biogeography*, **39**, 1–9.

Haywood, M.D.E., Browne, M., Skewes, T., Rochester, W., McLeod, I., Pitcher, C.R., Dennis, D., Dunn, J., Cheers, S. & Wassemberg, T. (2007) *Improved knowledge of Torres Strait seabed biota and reef habitats*. Marine and Tropical Sciences Research Facilities, Cairns.

Hurlbert, A.H. & Jetz, W. (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences USA*, **104**, 13384–13389.

Hutchinson, G.E. (1953) The concept of pattern in ecology. *Proceedings of the National Academy of Sciences USA*, **105**, 1–12.

Kindermann, R. & Snell, J.L. (1980) *Markov random fields and their applications*. American Mathematical Society, Providence, RI.

Levin, S.A. (1992) The problem of pattern and scale in ecology. *Ecology*, **73**, 1943–1967.

Li, X.H. & Wang, Y. (2013) Applying various algorithms for species distribution modelling. *Integrative Zoology*, **8**, 124–135.

Lieske, D.J. & Bender, D.J. (2009) Accounting for the influence of geographic location and spatial autocorrelation in environmental models: a comparative analysis using North American songbirds. *Journal of Environmental Informatics*, **13**, 12–32.

Lyons, S.K. & Willig, M.R. (2002) Species richness, latitude, and scale-sensitivity. *Ecology*, **83**, 47–58.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*, 2nd edn. Chapman and Hall/CRC, Boca Raton, FL.

Mellin, C., Bradshaw, C.J.A., Meekan, M.G. & Caley, M.J. (2010) Environmental and spatial predictors of species richness and abundance in coral reef fishes. *Global Ecology and Biogeography*, **19**, 212–222.

Murphy, H.T., VanDerWal, J. & Lovett-Doust, J. (2011) One, two and three-dimensional geometric constraints and climatic correlates of North American tree species richness. *Ecography*, **34**, 267–275.

Nakaya, T., Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2005) Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine*, **24**, 2695–2717.

Osborne, P.E., Foody, G.M. & Suarez-Seoane, S. (2007) Non-stationarity and local approaches to modelling the distributions of wildlife. *Diversity and Distributions*, **13**, 313–323.

Pitcher, C.R., Haywood, M., Hooper, J. *et al.* (2007) *Mapping and characterisation of key biotic and physical attributes of the Torres Strait ecosystem*. CSIRO/QM/QDPI CRC Torres Strait Task Final Report.

Pittman, S.J., McAlpine, C.A. & Pittman, K.M. (2004) Linking fish and prawns to their environment: a hierarchical landscape approach. *Marine Ecology Progress Series*, **283**, 233–254.

Pittman, S.J., Christensen, J.D., Caldow, C., Menza, C. & Monaco, M.E. (2007) Predictive mapping of fish species richness across shallow-water seascapes in the Caribbean. *Ecological Modelling*, **204**, 9–21.

R Development Core Team (2013) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. Available at: http://www.R-project.org/ (accessed 5 December 2013).

Rahbek, C. (2005) The role of spatial scale and the perception of large-scale species-richness patterns. *Ecology Letters*, **8**, 224–239.

Rahbek, C. & Graves, G.R. (2001) Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences USA*, **98**, 4534–4539.

Rosa, R., Dierssen, H.M., Gonzalez, L. & Seibel, B.A. (2008) Ecological biogeography of cephalopod molluscs in the Atlantic Ocean: historical and contemporary causes of coastal diversity patterns. *Global Ecology and Biogeography*, **17**, 600–610.

Stanaway, M.A., Reeves, R. & Mengersen, K.L. (2011) Hierarchical Bayesian modelling of plant pest invasions with human-mediated dispersal. *Ecological Modelling*, **222**, 3531–3540.

Terribile, L.C., Felizola Diniz-Filho, J.A., Rodriguez, M.A. & Rangel, T.F. (2009) Richness patterns, species distributions and the principle of extreme deconstruction. *Global Ecology and Biogeography*, **18**, 123–136.

Wheeler, D. & Tiefelsdorf, M. (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, **7**, 161–187.

Wiens, J.A. (1989) Spatial scaling in ecology. *Functional Ecology*, **3**, 385–397.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Figure S1 Fixed Gaussian weighting scheme.

Figure S2 Procedure used to assess changes in model support with increasing spatial scale.

Figure S3 The Torres Strait, and the position of the 220 reefs monitored in 1995 and 1996.

Table S1 Correlations (Spearman's ρ) among parameter estimates for the different modelling techniques.

Appendix S1 Supplementary text.

Appendix S2 R code for the spatial weighting procedure applied to generalized linear models.

BIOSKETCH

Camille Mellin is a Research Scientist at the Australian Institute of Marine Science, and a Visiting Researcher at the University of Adelaide. Camille is mostly interested in biogeography, coral reef ecology, ecological modelling and the impacts of global change on marine ecosystems.

Editor: José Alexandre Diniz-Filho