



RESEARCH  
PAPER

# Environmental and spatial predictors of species richness and abundance in coral reef fishes

C. Mellin<sup>1,2\*</sup>, C. J. A. Bradshaw<sup>2,3</sup>, M. G. Meekan<sup>4</sup> and M. J. Caley<sup>1</sup>

<sup>1</sup>Australian Institute of Marine Science, PMB no. 3, Townsville MC, Townsville, Qld 4810, Australia, <sup>2</sup>The Environment Institute and School of Earth and Environmental Sciences, University of Adelaide, SA 5005, Australia, <sup>3</sup>South Australian Research and Development Institute, PO Box 120, Henley Beach, SA 5022, Australia, <sup>4</sup>Australian Institute of Marine Science, The University of Western Australia (MO96), Botany Building, 35 Stirling Highway, Crawley WA 6009, Australia

## ABSTRACT

**Aim** We developed predictive models of coral reef fish species richness and abundance that account for both broad-scale environmental gradients and fine-scale biotic processes, such as dispersal, and we compared the importance of absolute geographical location (i.e. geographical coordinates) versus relative geographical location (i.e. distance to domain boundaries).

**Location** Great Barrier Reef, Australia.

**Methods** Four annual surveys of coral reef fishes were combined with a 0.01°-resolution grid of environmental variables including depth, sea surface temperature, salinity and nutrient concentrations. A principal component-based method was developed to select candidate predictors from a large number of correlated variables. Generalized linear mixed-effects models (GLMMs) were used to gauge the respective importance of the different spatial and environmental predictors. An error covariance matrix was included in the models to account for spatial autocorrelation.

**Results** (1) Relative geographical descriptors, represented by distances to the coast and to the barrier reef, provided the highest-ranked single model of species richness and explained up to 36.8% of its deviance. (2) Accounting for spatial autocorrelation doubled the deviance in abundance explained to 71.9%. Sea surface temperature, salinity and nitrate concentrations were also important predictors of abundance. Spatially explicit predictions of species richness and abundance were robust to variation in the spatial scale considered during model calibration.

**Main conclusions** This study demonstrates that distance-to-domain boundaries (i.e. relative geographical location) can offer an ecologically relevant alternative to geographical coordinates (i.e. absolute geographical location) when predicting biodiversity patterns, providing a proxy for multivariate and complex environmental processes that are often difficult or expensive to estimate.

## Keywords

**Abundance, biodiversity, generalized linear mixed-effect model, Great Barrier Reef, reef fish, spatial autocorrelation, species distribution model, species richness.**

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

## INTRODUCTION

The development and application of species distribution models, including models of species richness and abundance, has proliferated in the last few decades (Guisan & Thuiller, 2005). Originally used to test ecological theory and detect species–environment relationships, such models are now used regularly in conservation planning (Guisan & Zimmermann,

2000; Austin, 2002). They have been used to assess spatial patterns in biodiversity (e.g. Ferrier, 2002), disentangle natural and anthropogenic influences on the structure of biological communities (e.g. Mellin *et al.*, 2008) and forecast community responses to environmental change (Guisan & Thuiller, 2005). Although widely applied in terrestrial ecology, species distribution models are still in their infancy in marine ecosystems, particularly in coral reef ecosystems at spatial scales relevant to

environmental monitoring and management (Beger & Possingham, 2008; Mellin *et al.*, 2009). The development of such models in coral reefs could, however, provide a new insight into species distribution modelling and offers a novel opportunity to investigate the relationships between biological and environmental variables, and their spatial structure, in a system made of highly fragmented habitat patches subjected to broad-scale environmental gradients.

Physical variables that might be useful in predicting spatial patterns of species distributions can be categorized broadly as environmental or spatial. The influence of environmental variables partly relies on the species–energy relationship, a climatically based hypothesis that postulates that energy availability generates and maintains gradients of species richness in both terrestrial (see review by Hawkins *et al.*, 2003) and marine ecosystems (e.g. Roy *et al.*, 1998; Allen *et al.*, 2002). On coral reefs, thermodynamic and mechanical forces with the potential to shape fish assemblages include water temperature (Mora & Robertson, 2005; Sandin *et al.*, 2008), salinity (Fraser & Currie, 1996), nutrients and phytoplankton production (Mora & Robertson, 2005; Sandin *et al.*, 2008), wave exposure (Depczynski & Bellwood, 2005; Fulton & Bellwood, 2005) and ocean currents (M. J. Emslie *et al.* unpublished; Cowen *et al.*, 2006). Such environmental gradients can also indirectly influence fish assemblages by shaping benthic communities and reef geomorphology and thus the structure of fish habitats (e.g. Madin & Connolly, 2006; Finkl & Andrews, 2008). To date, however, most studies have only considered these environmental variables in isolation, so any hierarchy of their respective effects on the species richness and abundance of coral reef fishes remains unclear. In addition, analysis at restricted spatio-temporal scales (e.g. 10 km and 2 months; Depczynski & Bellwood, 2005) precludes the development of general rules about the influence of environmental factors on the structure and composition of reef fish communities and the distributions of their constituent species at broader scales likely to be important to the functioning of these ecosystems.

Spatial context can also influence the structure of biological communities, with the mechanisms involved depending on the spatial scale considered. Legendre & Legendre (1998) defined the trend in biological variables, or ‘true’ gradient, as a broad-scale spatial pattern arising from the influence of spatially structured environmental characteristics varying as a function of geographical location. By contrast, fine-scale spatial autocorrelation can generate a ‘false’ gradient in response to contagious biotic processes such as dispersal, thereby becoming a function of the distance between sites. Coral reef fish communities are likely to provide a unique system for investigating the respective role of each spatial scale. At a scale of > 10 km, coral reef fish distributions can reflect strong latitudinal (e.g. Mora *et al.*, 2003) or cross-shelf (e.g. Williams & Hatcher, 1983) differences.

The origin of these gradients remains controversial and several non-environmental drivers have been invoked, such as the mid-domain effect (Connolly *et al.*, 2003; Bellwood *et al.*, 2005). Under the mid-domain model, random placement of

species ranges is expected to result in the greatest species richness occurring in the middle of the spatial domain considered. Spatial gradients in the structure of tropical reef fish assemblages can also correlate closely with environmental gradients such as latitudinal variation in temperature (Roy *et al.*, 1998) or cross-shelf gradients in salinity and chlorophyll (Liston *et al.*, 1992; Burrage *et al.*, 2002). Therefore, both theory and empirical observations suggest that latitudinal and/or cross-shelf position might provide a cost-effective surrogate for the combined effects of different environmental variables when predicting the assemblage structure of coral reef fishes at broader spatial scales.

In species distribution modelling to date, broad-scale spatial patterns have always been accounted for by incorporating longitude and latitude as predictors (sometimes transformed into a trend surface), thereby providing an index of the absolute geographical location (e.g. Lichstein *et al.*, 2002; Greve *et al.*, 2008). In contrast, the distance-to-domain boundaries, representing a relative measure of geographical location along environmental gradients, might instead provide a better predictor of biodiversity patterns, independent of any mid-domain effect. The utility of such an approach, however, has not yet been investigated.

In some terrestrial ecosystems, strong spatial autocorrelation in species richness and abundance resulting from population connectivity occurs at scales < 10 km, which overlays the effects of both environmental and spatial variables (e.g. Selmi & Boulinier, 2001; Lichstein *et al.*, 2002). Despite potentially major effects on observed and predicted patterns of coral reef fish biodiversity, spatial autocorrelation has, to our knowledge, never before been considered explicitly in predictive models of assemblage patterns of reef fishes at a scale relevant to conservation and management decisions. By not incorporating spatial autocorrelation, existing models might overestimate environmental effects on species abundance (Lichstein *et al.*, 2002; Wintle & Bardos, 2006) due to a lack of independent errors, leading to inflated degrees of freedom and thus potentially biased estimates of Type I error (Legendre, 1993; Legendre & Legendre, 1998). It also seems important, regardless of the ecosystem under study, to combine fine-scale spatial autocorrelation with large-scale spatial structure informed through distance-to-domain boundaries as an alternative to geographical coordinates.

Here we use a regional-scale, 4-year dataset of abundances of coral reef fishes on the Great Barrier Reef (Australia) combined with a 0.01°-resolution grid of environmental and spatial variables to: (1) identify the best set of environmental predictors of coral reef fish assemblages, (2) assess the extent to which broad-scale spatial variables, including distances to domain boundaries, can act as effective proxies for these environmental variables in predictive models, and (3) provide a modelling framework that also accounts for the effects of fine-scale spatial autocorrelation on assemblage structure. The resulting model provides spatially explicit, biogeographical predictions of coral reef fish diversity and abundance on the Great Barrier Reef, and a conceptual model that can be applied to other ecosystems.

**METHODS**

**Study reefs**

The Great Barrier Reef (GBR) consists of more than 2900 reefs extending over 2300 km between 9 and 24°S latitude and covering approximately 350,000 km<sup>2</sup>. Since 1993, reef communities of the GBR have been monitored annually by the Australian Institute of Marine Science's (AIMS) Long Term Monitoring Program (LTMP; Sweatman *et al.*, 2005). Fish communities have been intensively surveyed across 46 reefs in six sectors (Cooktown/Lizard Island, Cairns, Townsville, Whitsunday, Swains, Capricorn Bunkers) spanning much of the GBR (Fig. 1). In each sector (with the exception of the Swains and Capricorn Bunker sectors) at least three reefs were sampled in each of three shelf positions (i.e. inner, mid and outer).

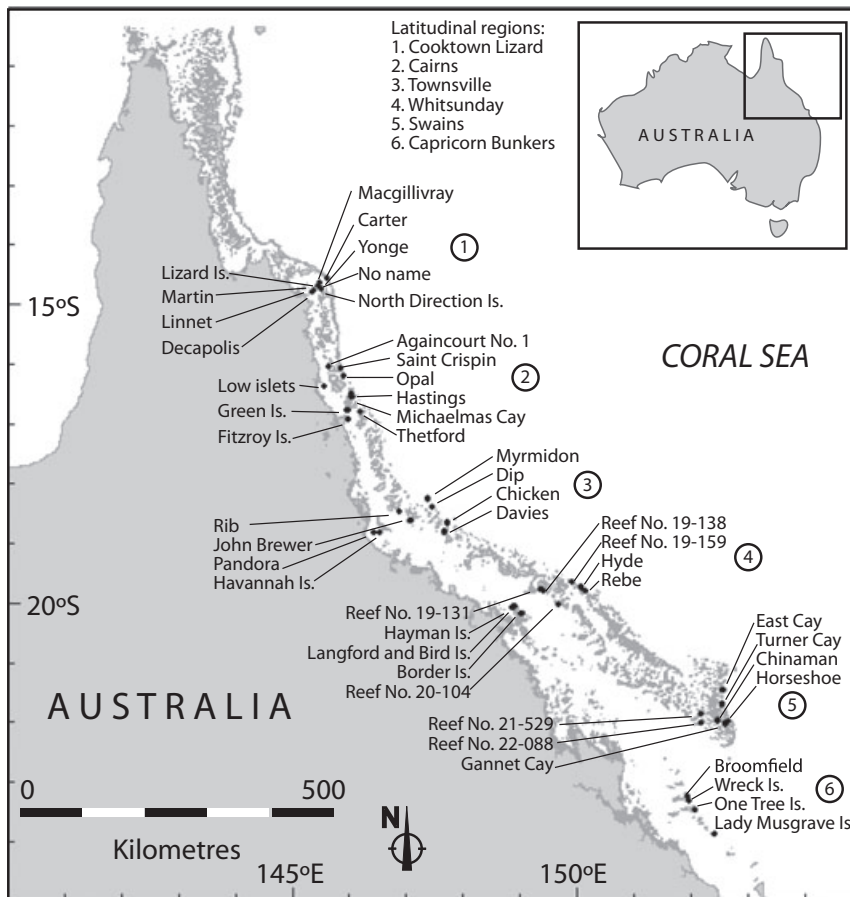
**Survey methods and data collection**

On each reef, three sites in a single habitat (the first stretch of continuous reef on the north-east flank of the reef, excluding vertical drop-offs) and separated by > 250 m were selected for sampling. Within each site, five randomly selected and permanently marked 50-m long transects were deployed roughly parallel to the reef crest, each separated by 10–40 m along the 6–9 m

depth contour. The sampling was evenly distributed among years and transects except for 2005 (see below). Within these transects, abundances of 251 species representing 10 families were estimated each year. These 251 species were chosen because they can be surveyed consistently at different sites and under different conditions. This set of species excludes those that are cryptic or nocturnal, and which therefore have a low (or highly variable) probability of detection. The number of species surveyed during these censuses provides a surrogate for the total number of species present at a site, estimated using destructive methods (C. Mellin *et al.* unpublished). Small site-attached species such as damselfishes (Pomacentridae) were counted in a 1-m wide strip along the transect while transect width was increased to 5 m for larger, mobile species (see Halford & Thompson, 1996 for detailed methods). Only adult fish (> 1 year old) were recorded, these being distinguished from juveniles by their size and coloration. Sites were sampled by different divers and annual calibration exercises were undertaken to ensure consistency among divers (i.e. there was no systematic bias). The methods used to calibrate of divers has proven effective (Halford & Thompson, 1996).

**Data organization**

We were primarily interested in spatial patterns in fish diversity and how they relate to environmental data collected during the



**Figure 1** The Great Barrier Reef, with the position of the 46 reefs monitored yearly since 1993. Numbers indicate the location of each sector (i.e. latitudinal region).

recent past, so only fish data recorded since 2003 were considered. The dataset collected during 2005 was also discarded as many reefs were not sampled during that year due to bad weather. A total of 137 sites (i.e. three sites at each reef, except for one reef of the Cooktown/Lizard Island sector with only two sites) that were sampled in 2003, 2004, 2006 and 2007 were thus included in the analysis. Reefs were sampled during the entire year, although half were surveyed during summer months (October to January), and the same site could be surveyed in different months in consecutive years. Fish species richness ( $S$ ) was defined as the total number of fish species sampled at each site and was obtained by pooling species across the five transects within sites and across the four years. Similarly, for each site, abundances were pooled across transects and across the four years. Mean yearly estimates of total fish abundance at each site ( $N$ ) were calculated as the total abundance of fish recorded at each site divided by the number of years sampled (four in all cases). Both  $S$  and  $N$  were therefore defined for each site and across years to match the resolution of environmental variables, which were estimated across a  $0.01^\circ$  grid (see next section) and not replicated in time. Pooling samples across years was appropriate because a Kruskal–Wallis test revealed no evidence of any consistent year effect for fish species richness ( $P > 0.21$ ) or abundance ( $P > 0.77$ ).

### Environmental variables

A large set of environmental variables (see Table S1 in Supporting Information) was collated across a  $0.01^\circ$  grid at a national scale as part of the Commonwealth of Australia's Environment Research Facility (CERF) Marine Biodiversity Hub (<http://www.marinehub.org/>). These environmental variables included: annual mean estimates of nitrate, oxygen, phosphate, silicate, temperature and salinity, bathymetry, percentage cover of sediment components, and multiple indices of the strength and frequency of the combined wave–current bed shear stress. Among these indices of bed shear stress, the percentage exceedance was defined as the percentage of time for which the bed shear stress was  $> 0.4$  Pa, a threshold over which bed shear stress represents an important driver of coral reef communities as determined by preliminary exploratory analyses (e.g. Pitcher *et al.*, 2007). The ratio was defined as the bed shear stress in excess of 0.4 Pa as a proportion of the total stress integrated over time. Indices of ocean productivity using ocean colour data were estimated by the Sea-viewing Wide Field-of-view Sensor (SeaWiFS). All data sources and references for spatial interpolation are annotated in Table S1. Each fish sampling site was assigned the environmental variables of the closest node on the  $0.01^\circ$  grid. In addition, spatial variables, including longitude and latitude and the shortest distances to the coast and to the ocean at the edge of the GBR lagoon, were calculated for each fish sampling site using a geographical information system. We used the great-circle distance, that is, the shortest distance between two points on the surface of the earth, for all distance measures.

### Analysis

The environmental dataset gathered a total of 32 variables, and most of them were intercorrelated. For this reason, we used a variable-reduction procedure to identify, among all available environmental variables (Table S1), the subset of candidate predictors that minimized multicollinearity and maximized correlation with fish species richness and abundance. This was done in two steps: (1) a principal components analysis (PCA) in conjunction with the analysis of the correlation matrix to identify groups of correlated environmental variables, and (2) within each group of correlated variables, a selection of either one or two candidate predictors that maximized the percentage deviance explained in species richness and abundance. This second step was done using generalized linear models (GLM) assuming a Poisson distribution and predicting successively  $S$  and  $N$  as a function of each single environmental variable and its quadratic term. When both fish variables ( $S$  and  $N$ ) were best explained by the same environmental variable, only that environmental variable was selected; when fish variables were best explained by different environmental variables, those two environmental variables were selected as candidate predictors. A second PCA was then performed with these candidate predictors as explanatory variables and other environmental variables as illustrative to verify that the variable reduction procedure had not altered the ordination of variables and individuals (i.e. fish sampling sites).

### Modelling

To gauge the relative importance of spatial and environmental variables on reef fish assemblages, we fitted a second series of GLMs to  $S$  and to  $N$ . For each response variable, a Poisson distribution with a log link was assumed and the normal distribution of model residuals was checked using the normal scores of standardized residual deviance (Breslow, 1996). All possible combinations of spatial and candidate environmental predictors and their quadratic terms (in order to detect possible curvilinear relationships), including cross products as well as power terms, were considered. Asymptotic indices of information loss were used to assign relative strengths of evidence to the different competing models, with both Akaike's information criterion corrected for small sample sizes ( $AIC_c$ ) as an index of Kullback–Leibler (K-L) information loss and the dimension-consistent Bayesian information criterion (BIC) (an approximation of the Bayes factor given no informative prior information on relative model support; Burnham & Anderson, 2002) used for model comparisons. These indices identify the relative evidence of model(s) from a candidate set, with the relative likelihoods of candidate models calculated using  $AIC_c$  and BIC weights (Burnham & Anderson, 2002). The K-L prior used to justify  $AIC_c$  weighting can favour more complex models when sample sizes are large (Burnham & Anderson, 2004; Link & Barker, 2006), so we considered BIC weights to determine the contribution of the most important variables, and  $AIC_c$  weights to compute the weighted-average model predictions. Weighted-



average model predictions were made over the set of models considered (i.e. the sum of the predicted response variables multiplied by the model's  $AIC_c$  weight) which did indeed maximize prediction accuracy (Burnham & Anderson, 2004; Link & Barker, 2006). For each model, the percentage deviance explained (%DE) was used as a measure of a model's goodness-of-fit.

Spatial autocorrelation can systematically invalidate the assumption of independent model errors, potentially distorting parameter estimation and inflating the probability of Type I errors (Legendre, 1993; Diniz-Filho *et al.*, 2003). Spatial correlograms assessing autocorrelation in species richness and abundance (raw data or GLM residuals) as a function of the distance between sites were generated using Moran's  $I$  (Diggle & Ribeiro, 2007). Evidence for spatial autocorrelation was assessed at five distance classes using Bonferroni correction (Legendre & Legendre, 1998). The upper limits for these distance classes (i.e. lags) were defined at 30, 63, 105, 143 and 203 km to maximize the similarity of the number of observations in the different classes (Diniz-Filho *et al.*, 2003). When a spatial structure was evident in GLM residuals, five spatial correlation structures (linear, log, log-linear, exponential, spherical) were successively considered and the structure that best fitted the observations was identified (Diggle & Ribeiro, 2007; Bivand *et al.*, 2008). This correlation structure was then incorporated through the error covariance matrix of spatial generalized linear mixed-effects models (GLMMs) coding *Site* as a random effect. GLMMs were optimized using penalized likelihood (Venables & Ripley, 2002). Note that this method prevents the computation of the likelihood and hence asymptotic indices of information loss, which must then be approximated by those obtained from the non-spatial model. Therefore, for each response variable, the non-spatial GLMs were refitted as spatial GLMMs. We computed the weighted-average model predictions based on  $AIC_c$  weights to maximize prediction accuracy (Burnham & Anderson, 2002; Link & Barker, 2006). Spatial correlograms of model residuals were plotted to test the assumption that GLMM residuals were not spatially autocorrelated.

### Prediction

For each response variable, the mean prediction error of spatial GLMMs was assessed using a 10-fold cross-validation (Davison & Hinkley, 1997). This bootstrap resampling procedure estimates a mean prediction error for 10% of observations that were randomly omitted from the calibration dataset; this procedure was iterated 1000 times. Spatial GLMMs were then used to generate and average predictions at a regional scale over the 0.01° grid, i.e. for the GBR and within the perimeter of the sampled reefs used to calibrate the models.

We also investigated whether the spatial extent of the calibration dataset influenced model predictions by recalibrating  $S$  and  $N$  models using half the dataset, defined as the 50% of sites located closest to a given sector, including the sites from that sector. We then compared the resulting predictions for that sector with those computed from the original models. This

analysis of spatial-scale effects was conducted on a central sector with large reefs (Townsville) and a southern sector with small reefs (Swains; Fig. 1).

## RESULTS

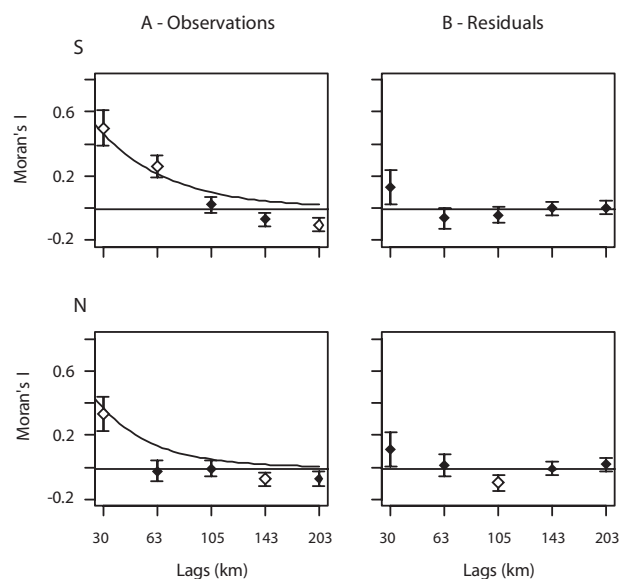
Observed fish species richness ( $S$ ) and abundance ( $N$ ) averaged ( $\pm$  standard deviation)  $72 \pm 13$  species and  $744 \pm 341$  individuals per site, respectively. The most common species were *Chlorurus microrhinos* and *Chlorurus sordidus* (Scaridae) occurring at more than 99% of the surveyed sites at least once during the four years of surveys considered here. *Pomacentrus lepidogenys*, *Pomacentrus moluccensis* and *Neopomacentrus azysron* (Pomacentridae) were the most abundant species with mean abundances of  $97 \pm 92$ ,  $70 \pm 145$  and  $66 \pm 89$  individuals per site, respectively.

The variable-selection procedure identified a set of 12 environmental variables that minimized multicollinearity and maximized the correlation with  $S$  and/or  $N$  (Table S1, Fig. S1a). The first principal component, which explained 26.1% of the environmental variation within the subset of environmental variables, primarily reflected a latitudinal gradient (see Fig. S1b). The same result was observed when all environmental variables were included in the PCA (not shown). These 12 environmental variables were thus considered as candidate predictors as well as the four spatial variables in our GLMs of  $S$  and  $N$ . For  $S$ , the best-fitting model according to BIC included only the distance to the ocean and to the coast, and explained 36.8% of the deviance (Table 1). For  $N$ , the best fitting model according to BIC explained 21.5% of the deviance and included the same predictors, in addition to annual mean salinity (quadratic term), nitrate and silicate concentration, and the standard deviation in annual mean K490 levels (a measure of light attenuation). For both species richness and abundance, the top-ranked model according to  $AIC_c$  included all predictors and also maximized the percentage deviance explained (i.e. 55.9 and 35.4%, respectively, Table 1; only models outperforming the null models, and null models themselves are shown). The change in model ranking according to  $AIC_c$  or BIC indicates that the most parsimonious models (i.e. top ranked according to BIC) were not the ones that maximized prediction accuracy (i.e. top-ranked models according to  $AIC_c$ ). Mean prediction errors of the top-ranked models according to  $AIC_c$  were 8.99 and 3.59% for  $S$  and  $N$ , respectively, and were the lowest across the entire model set, whereas mean prediction errors of the top-ranked models according to BIC were 14.34 and 4.06% for  $S$  and  $N$ , respectively.

We found evidence for spatial autocorrelation at lag = 30 km in observations of  $S$  and  $N$  ( $P < 0.001$ ; Fig. 2a), with the effect persisting in the residuals of non-spatial GLM ( $P < 0.006$ ; not shown). The exponential spatial correlation structure gave the best fit to the null model and was thus used in spatial GLMMs for both response variables. Accounting for spatial autocorrelation resulted in an increase of the percentage deviance explained from 55.9 and 35.4% up to 62.4 and 71.9% in  $S$  and  $N$ , respectively (see Fig. S2a), with scores of standardized residual deviance being close to normality both for spatial models (see

**Table 1** Summary of generalized linear model (GLM) comparisons using Akaike's information criterion corrected for small sample sizes ( $AIC_c$ ) and the Bayesian information criterion (BIC). Only the models outperforming the null model and the null models themselves are shown. Response variables are species richness ( $S$ ) and log-transformed abundance ( $N$ ) of reef fishes. Shown are the number of samples ( $n$ ), model maximum log-likelihood (LL), number of parameters ( $k$ ), change in  $AIC_c$  ( $\Delta AIC_c$ ),  $AIC_c$  weight ( $wAIC_c$ ), change in BIC ( $\Delta BIC$ ), BIC weight ( $wBIC$ ) and the percentage deviance explained (%DE). %DE is a measure of the structural goodness-of-fit of the model. Model sequences are ordered by increasing BIC for the two model sets. Spatial predictors include minimum distance to the coast (Dc) and minimum distance to the ocean (Do). Environmental predictors include: Tm, annual mean sea surface temperature; Ts, standard deviation in sea surface temperature; Sm, annual mean salinity; Om, annual mean oxygen concentration; Os, standard deviation in oxygen concentration; Nm, annual mean nitrogen concentration; Si, annual mean silicate concentration; Ks, standard deviation in K490 (the diffuse attenuation coefficient at a wavelength of 490 nm). See Table S1 for units. All\_env includes all candidate environmental predictors in Table S1, including four water chemistry variables, one substrate descriptor, three indices of ocean productivity and depth.

Model	LL	$k$	$\Delta AIC_c$	$wAIC_c$	$\Delta BIC$	$wBIC$	%DE
<i>S</i> ( $n = 137$ )							
Dc + Do	-344.0	4	7.7	0.016	0.0	0.989	36.8
Tm <sup>2</sup> + Ts + Sm <sup>2</sup> + Om + Os	-339.2	9	7.6	0.016	9.6	0.008	43.2
Tm <sup>2</sup> + Ts + Sm <sup>2</sup> + Om + Os + Dc + Do	-334.8	11	6.6	0.028	15.4	0.003	48.3
All_env	-331.2	13	2.8	0.180	17.7	0.000	51.5
All_env + Dc + Do	-326.8	15	0.0	0.760	18.8	0.000	55.9
Null	-367.1	2	47.3	0.000	31.7	0.000	0.0
<i>logN</i> ( $n = 137$ )							
Nm + Sm <sup>2</sup> + Dc + Do + Si + Ks	-47.6	8	6.5	0.042	0.0	0.529	21.5
Nm + Sm <sup>2</sup> + Si + Ks	-52.6	6	11.8	0.004	0.3	0.455	17.8
Nm + Sm <sup>2</sup> + Dc + Do + Si + Ks + Om + Os	-47.0	10	9.9	0.008	8.2	0.010	22.4
All_env + Dc + Do	-35.7	15	0.0	0.946	9.4	0.006	35.4
Null	-73.3	2	44.5	0.000	22.7	0.000	0.0

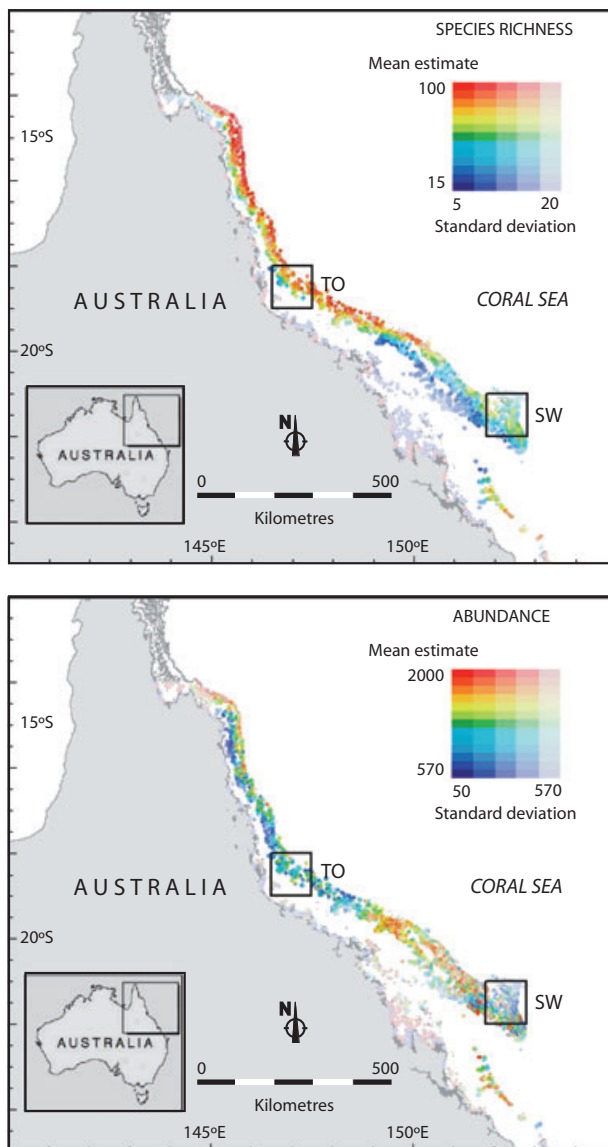


**Figure 2** Spatial correlograms of Moran's  $I$  as a function of increasing distance between sites (lags), respectively, up to 30, 63, 105, 143 and 203 km. (A) Spatial autocorrelation in observations of fish species richness ( $S$ ) and abundance ( $N$ ). The continuous line represents the exponential spatial structure assumed in the spatial generalized linear mixed-effects models (GLMMs). (B) Spatial autocorrelation in GLMM residuals. Error bars represent standard deviations in Moran's  $I$ . White diamonds show Moran's  $I$ -values for which zero was excluded from the 95% confidence interval after a Bonferroni correction.

Fig. S2b) and non-spatial ones (not illustrated). Except for  $N$  at lag = 105 km, there was much weaker evidence of spatial autocorrelation in GLMM residuals (all  $P > 0.061$ ; Fig. 2b). Mean prediction errors of spatial GLMMs assessed using 10-fold cross-validation were 7.4 species and 245.5 individuals per site, respectively.

Mapping model predictions over the GBR highlighted different spatial patterns for the two response variables. The greatest predicted values of  $S$  were obtained for the northern sectors and on the outer-shelf reefs (Fig. 3). Standard deviation in model predictions was the highest for coastal reefs and between the Whitsunday and Swains sectors. By contrast,  $N$  was predicted to be greatest on the southernmost reefs of the GBR, but overall had a bimodal pattern with higher values on northern and southern reefs. Again, the greatest standard deviation in model predictions was observed on coastal reefs and between the Whitsunday and Swains sectors. Finer-scale spatial variation was observed for  $N$  than for  $S$ .

When the spatial extent was reduced to half the dataset around a focal sector, the same models maximized the percentage deviance explained in response variables, even though model rankings based on BIC scores could differ from analysis using the entire dataset (not shown). For the Townsville and Swains sectors, small-scale models explained 80.5 and 71.5% of deviance in  $S$ , and 64.2 and 77.0% of deviance in  $N$ , respectively. Generally, similar spatial patterns were observed for predictions computed from fine-scale or broad-scale models (Fig. 4).



**Figure 3** Spatial predictions of species richness and abundance of coral reef fishes on the Great Barrier Reef. A two-dimensional colour key is used to represent both mean estimate (hue) and standard deviation (saturation) of predictions. Insets show sectors selected for the spatial scale analysis (Fig. 4). TO = Townsville, SW = Swains.

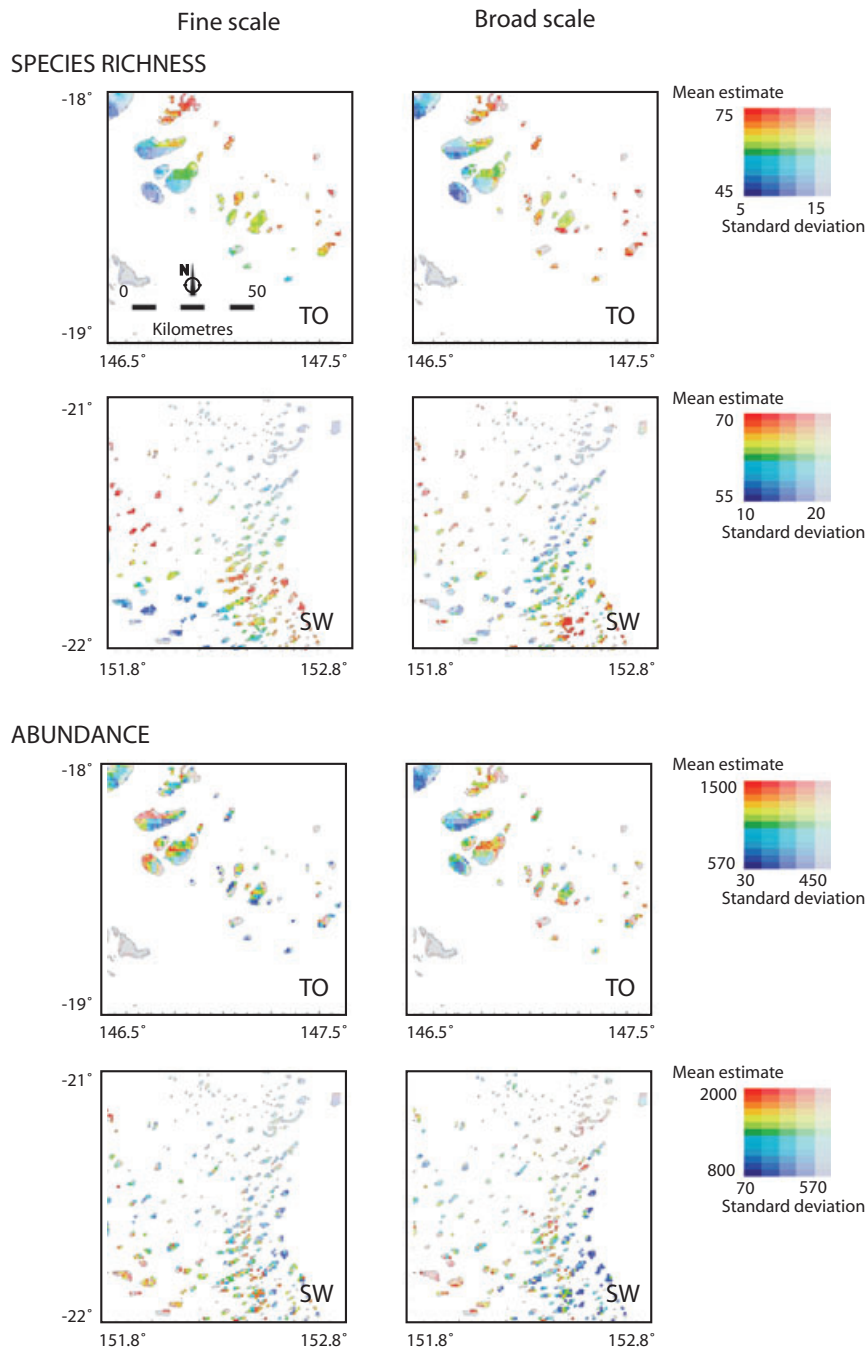
## DISCUSSION

The mid-domain effect, when applied in species distribution models, leads to the expectation of species richness being greatest at the centre of the domain (e.g. Connolly *et al.*, 2003; Bellwood *et al.*, 2005). However, because the processes that display gradients across the domain are complex and act at different scales, spatially explicit predictions of species richness and abundance differed from this pattern. Distances to domain boundaries (i.e. coast and ocean) provided the most parsimonious model of species richness and were also important predictors of abundance, suggesting that relative geographical location

alone can provide an ecologically relevant predictor in species distribution models. Whereas geographical coordinates can sometimes fail in predicting spatial patterns in biodiversity (e.g. Greve *et al.*, 2008), our results suggest that a relative measure of geographical location might be more appropriate in representing complex environmental gradients. This approach is broadly applicable across species distribution models in which variables such as the distance to mountain ranges, lake edges or oceans could each be tested as an alternative to geographical coordinates.

Most environmental variables we examined on the Great Barrier Reef displayed strong cross-shelf gradients such as salinity (Burrage *et al.*, 2002), temperature and chlorophyll (Liston *et al.*, 1992) that are influenced by freshwater inputs and terrigenous influences near the coast, and by oligotrophic oceanic waters at the shelf edge. Indeed, the distance between the coast and the ocean increases from northern to southern sectors on the Great Barrier Reef; therefore, for northern sites both distances would be short, whereas for southern sites at least one of these distances would be long. Our models highlighted these complex relationships between spatial, environmental and biological variables, providing a framework for the development of approaches that incorporate multiscale spatial and environmental variables. The most parsimonious model of fish species richness only included relative geographical location (i.e. distances to the coast and to the ocean), while a combination of spatial and environmental predictors explained up to 55.9% of the deviance in fish species richness. The most parsimonious model of fish abundance included relative geographical location, indices of ocean productivity, salinity and water composition. The analysis of spatial-scale effects showed that using half the data resulted in spatial predictions that were similar to those computed from broad-scale models, indicating that spatial patterns predicted by the model were robust with respect to the spatial extent of the calibration dataset, and suggesting that similar predictions could be achieved using half of the dataset only.

Incorporating fine-scale spatial autocorrelation into our models doubled the deviance explained in fish abundance, and substantially increased the deviance explained in fish species richness. Among ecological processes that drive population dynamics in marine ecosystems, contagiously distributed biotic processes such as larval dispersal, reproduction, competition and predation have the potential to induce spatial autocorrelation in observations (Legendre & Legendre, 1998). Such models might thus be biased unless they explicitly incorporate spatial organization of the populations and communities of interest among their predictor variables. Such effects of spatial autocorrelation, however, remain largely overlooked in marine species distribution models, especially in coral reef ecosystems. Our spatial GLMMs were efficient in coping with spatial autocorrelation, and so were able to incorporate the underlying processes leading to these spatial associations. We strongly support further developments of GLMM or other statistical techniques that account for spatial autocorrelation in marine species distribution models to predict and map realistic assemblage distribution patterns.



**Figure 4** Spatial predictions of species richness and abundance of coral reef fishes for the sectors of Townsville (TO) and Swains (SW; see location on Fig. 1) computed from fine- and broad-scale models. Fine-scale models were built using half the dataset, including the nearest sites from the sector of interest. Broad-scale models were built using the entire dataset. A two-dimensional colour key is used to represent both mean estimate (hue) and standard deviation (saturation) of predictions.

In addition to the spatial context, several environmental variables emerged as important predictors of fish species richness, including mean annual temperature and standard deviation (a proxy for seasonal temperature variability). Temperature has been suggested to act as a proxy for energy input in coral reef ecosystems (Fraser & Currie, 1996) and as a predictor of species

richness (e.g. Roy *et al.*, 1998; Macpherson, 2002; Bellwood *et al.*, 2005). However, the latter assumption had never been tested in predictive modelling. Moreover, temperature effects on fish species richness are poorly understood because both variables are spatially autocorrelated (Mora & Robertson, 2005), which can lead to spurious correlations unless spatial autocor-



relation is properly considered. For the first time, models developed in this study clearly demonstrate that once spatial autocorrelation has been accounted for, annual mean temperature and its seasonal variation still remain important additional predictors of fish distribution.

Our models also revealed the importance of mean annual salinity and nitrate concentration as important predictors. These results are consistent with previous suggestions that variation in salinity induces physiological stress for corals and associated organisms (Rosen, 1971; Kassahn *et al.*, 2007), although this influence might act at a scale of < 10 km (e.g. a river mouth or barrier reef pass; Fraser & Currie, 1996). Annual mean nitrate concentration is also an important predictor of fish abundance, supporting the idea that biogeochemical processes can be important in the structuring of reef communities, for instance through the development of algal turf consumed by herbivorous fishes (McClanahan *et al.*, 2003). Nitrate concentration was also the best predictor of the distribution of coastal fish species along latitudinal gradients in both the eastern and western Atlantic (Macpherson, 2002). Our results demonstrate that between these broad and fine spatial scales, nitrate concentration remains an effective predictor of fish species richness and abundance, thus unifying conclusions on the importance of this predictor across scales.

Spatial patterns in predictive maps of reef fish biodiversity generated by our models reflected strong latitudinal and cross-shelf gradients. The latitudinal gradient predicted in reef fish diversity can result from the latitudinal gradient in temperature, and it is consistent with observations that fish species richness decreases as the distance from the Indo-Australian Archipelago increases (e.g. Bellwood & Hughes, 2001). Different hypotheses have been formulated to explain this latitudinal gradient, including environmental conditions or past disturbance events (e.g. Fraser & Currie, 1996), or geometric constraints implied by boundaries of the Indo-Pacific domain (Connolly *et al.*, 2003; Bellwood *et al.*, 2005). The cross-shelf gradient is consistent with recent work suggesting that the structure of fish communities would correlate with mesoscale oceanographic features (M. J. Emslie *et al.*, unpublished). Our approach was not designed to test these hypotheses, but it reproduced well the broad-scale and well-known spatial patterns of coral reef fish biodiversity on the GBR.

The development of such multiscale species distribution models is now warranted by an increasing availability of global-scale, high-resolution remotely sensed environmental data (Mellin *et al.*, 2009), but several potentially important aspects require attention. First, even though environmental datasets can reasonably be considered as appropriate for predicting broad-scale, latitudinal and cross-shelf patterns of fish species richness and abundance, more research is required to assess their predictive power at the reef scale. Indeed, interpolated data and point-based data might perform differently as predictors depending on the spatial scale at which fish assemblages respond to environment heterogeneity. Model applicability to the entire reef should also be assessed given that (1) these fish data were collected from a single habitat (the north-east flank of each reef)

and (2) the resolution of these environmental data does not necessarily reflect the fine-scale patterns of environmental variation occurring in this habitat. In the view of investigating scale effects on model performance, a preliminary study used spatial exponential weighting to simulate an increasing spatial scale around a focal point and look at how predictor ranking and model performance change with increasing spatial scale (i.e. the spatial extent of the training dataset; C. Mellin *et al.*, unpublished). This procedure showed that the importance of geographical predictors only appeared at a medium (20-km) spatial scale, below which fine-scale variation in environment provided better predictors of species richness. This analysis also provides a tool to determine the spatial scale optimizing model cost-effectiveness (i.e. maximizing the deviance explained in the response variable when minimizing the spatial extent of the training dataset). Second, predicting species richness and abundance does not predict species composition, and sites with similar fish species richness might yield completely different species compositions. The contribution of population processes to the spatial structure identified in species richness should also be assessed in further studies, provided that population structure and demography are well documented. Ideally, the kinds of models developed here should be coupled with multivariate models of reef assemblages, including population structure and accounting for taxonomy, species life-history traits, diet and susceptibility to environmental and anthropogenic impacts.

Over the past few decades, coral reefs have suffered increasing human-induced disturbances that have affected the structure of reef communities and their potential resilience to these disturbances (Nyström *et al.*, 2000; Bellwood *et al.*, 2004). Thermal stress and ocean acidification, in particular, are expected to increase in the coming decades, with serious implications for the persistence of reef corals and associated organisms (Hoegh-Guldberg *et al.*, 2007; McClanahan *et al.*, 2007). This situation mandates the urgent need for cost-effective conservation tools for predicting changes in biodiversity to facilitate timely and effective intervention where necessary. Species distribution models can provide such tools because they allow prediction of biodiversity patterns in areas where only environmental data are available. Combining spatially explicit predictions of univariate and multivariate indices of biodiversity patterns can indeed offer efficient, pluralistic and comprehensive tools for understanding species responses to environmental variability – a critical step towards the sustainable management of biodiversity.

## ACKNOWLEDGEMENTS

Funding was provided by the Australian Commonwealth Environment Research Facilities (CERF) programme (partners include the University of Tasmania, CSIRO, Geoscience Australia, the Australian Institute of Marine Science and Museum Victoria). We thank all members of the Australian Institute of Marine Science Long-Term Monitoring Program who have contributed to data collection. We are grateful to B. Venables, R. Pitcher, N. Bax and M. Hughes for their insights and suggestions.

## REFERENCES

- Allen, A.P., Brown, J.H. & Gillooly, J.F. (2002) Global biodiversity, biochemical kinetics, and the energetic-equivalence rule. *Science*, **297**, 1545–1548.
- Austin, M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101–118.
- Beger, M. & Possingham, H.P. (2008) Environmental factors that influence the distribution of coral reef fishes: modeling occurrence data for broad-scale conservation and management. *Marine Ecology Progress Series*, **361**, 1–13.
- Bellwood, D.R. & Hughes, T. (2001) Regional-scale assembly rules and biodiversity of coral reefs. *Science*, **292**, 1532–1534.
- Bellwood, D.R., Hughes, T., Folke, C. & Nyström, M. (2004) Confronting the coral reef crisis. *Nature*, **429**, 827–833.
- Bellwood, D.R., Hughes, T.P., Connolly, S.R. & Tanner, J. (2005) Environmental and geometric constraints on Indo-Pacific coral reef biodiversity. *Ecology Letters*, **8**, 643–651.
- Bivand, R.S., Pebesma, E.J. & Gómez-Rubio, V. (2008) *Applied spatial data analysis with R*. Springer, New York.
- Breslow, N.E. (1996) Generalized linear models: checking assumptions and strengthening conclusions. *Statistica Applicata*, **8**, 23–41.
- 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.
- Burrage, D.M., Heron, M.L., Hacker, J.M., Stieglitz, T.C., Steinberg, C.R. & Prytz, A. (2002) Evolution and dynamics of tropical river plumes in the Great Barrier Reef: an integrated remote sensing and *in situ* study. *Journal of Geophysical Research*, **107**, 8016, doi:10.1029/2001JC001024.
- Connolly, S.R., Bellwood, D.R. & Hughes, T.P. (2003) Indo-Pacific biodiversity of coral reefs: deviations from a mid-domain model. *Ecology*, **84**, 2178–2190.
- Cowen, R.K., Paris, C.B. & Srinivasan, A. (2006) Scaling of connectivity in marine populations. *Science*, **311**, 522–527.
- Davison, A.C. & Hinkley, D.V. (1997) *Bootstrap methods and their application*. Cambridge University Press, Cambridge.
- Depczynski, M. & Bellwood, D.R. (2005) Wave energy and spatial variability in community structure of small cryptic coral reef fishes. *Marine Ecology Progress Series*, **303**, 283–293.
- Diggle, P.J. & Ribeiro, P.J. Jr (2007) *Model-based geostatistics*. Springer, New York.
- Diniz-Filho, J.A.F., Bini, L.M. & Hawkins, B.A. (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, **12**, 53–64.
- Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*, **51**, 331–363.
- Finkl, C.W. & Andrews, J.L. (2008) Shelf geomorphology along the south-east Florida Atlantic continental platform: barrier coral reefs, nearshore bedrock, and morphosedimentary features. *Journal of Coastal Research*, **24**, 823–849.
- Fraser, R.H. & Currie, D.J. (1996) The species richness-energy hypothesis in a system where historical factors are thought to prevail: coral reefs. *The American Naturalist*, **148**, 138–159.
- Fulton, C.J. & Bellwood, D.R. (2005) Wave-induced water motion and the functional implications for coral reef fish assemblages. *Limnology and Oceanography*, **50**, 255–264.
- Greve, M., Gaston, K.J., van Rensburg, B.J. & Chown, S.L. (2008) Environmental factors, regional body size distributions and spatial variation in body size of local avian assemblages. *Global Ecology and Biogeography*, **17**, 514–523.
- Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.
- Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.
- Halford, A.R. & Thompson, A.A. (1996) *Visual census surveys of reef fish. Long term monitoring of the Great Barrier Reef. Standard operational procedure number 3*. Australian Institute of Marine Science, Townsville.
- Hawkins, B.A., Field, R., Cornell, H.V., Currie, D.J., Guégan, J.F., Kaufman, D.M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E.M., Porter, E.E. & Turner, J.R.G. (2003) Energy, water, and broad-scale geographic patterns of species richness. *Ecology*, **84**, 3105–3117.
- Hoegh-Guldberg, O., Mumby, P.J., Hooten, A.J., Steneck, R.S., Greenfield, P., Gomez, E., Harvell, C.D., Sale, P.F., Edwards, A.J., Caldeira, K., Knowlton, N., Eakin, C.M., Iglesias-Prieto, R., Muthiga, N., Bradbury, R.H., Dubi, A. & Hatziolos, M.E. (2007) Coral reefs under rapid climate change and ocean acidification. *Science*, **318**, 1737–1742.
- Kassahn, K.S., Crozier, R.H., Ward, A.C., Stone, G. & Caley, M.J. (2007) From transcriptome to biological function: environmental stress in an ectothermic vertebrate, the coral reef fish *Pomacentrus moluccensis*. *BMC Genomics*, **8**, 358.
- Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.
- Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd English edn. Elsevier, Amsterdam.
- Lichstein, J.W., Simons, T.R., Shiner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.
- Link, W.A. & Barker, R.J. (2006) Model weights and the foundations of multimodel inference. *Ecology*, **87**, 2626–2635.
- Liston, P., Furnas, M.J., Mitchell, A.W. & Drew, E.A. (1992) Local and mesoscale variability of surface-water temperature and chlorophyll in the Northern Great Barrier Reef, Australia. *Continental Shelf Research*, **12**, 907–921.
- McClanahan, T.R., Sala, E., Stickels, P.A., Cokos, B.A., Baker, A.C., Starger, C.J. & Jones, S.H. (2003) Interaction between nutrients and herbivory in controlling algal communities and coral condition on Glover's Reef, Belize. *Marine Ecology Progress Series*, **261**, 135–147.

- McClanahan, T.R., Ateweberhan, M., Graham, N.A.J., Wilson, S.K., Sebastián, C.R., Guillaume, M.M.M. & Bruggemann, J.H. (2007) Western Indian Ocean coral communities: bleaching responses and susceptibility to extinction. *Marine Ecology Progress Series*, **337**, 1–13.
- Macpherson, E. (2002) Large-scale species-richness gradients in the Atlantic Ocean. *Proceedings of the Royal Society B: Biological Sciences*, **269**, 1715–1720.
- Madin, J.S. & Connolly, S.R. (2006) Ecological consequences of major hydrodynamic disturbances on coral reefs. *Nature*, **444**, 477–480.
- Mellin, C., Ferraris, J., Galzin, R., Harmelin-Vivien, M., Kulbicki, M. & de Loma, T.L. (2008) Natural and anthropogenic influences on the diversity structure of reef fish communities in the Tuamotu Archipelago (French Polynesia). *Ecological Modelling*, **218**, 182–187.
- Mellin, C., Andréfouët, S., Kulbicki, M., Dalleau, M. & Vigliola, L. (2009) Remote sensing and fish–habitat relationships in coral reef ecosystems: review and pathways for systematic multi-scale hierarchical research. *Marine Pollution Bulletin*, **58**, 11–19.
- Mora, C. & Robertson, D.R. (2005) Causes of latitudinal gradients in species richness: a test with fishes of the tropical eastern Pacific. *Ecology*, **86**, 1771–1782.
- Mora, C., Chittaro, P.M., Sale, P.F., Kritzer, J.P. & Ludsin, S.A. (2003) Patterns and processes in reef fish diversity. *Nature*, **421**, 933–936.
- Nyström, M., Folke, C. & Moberg, F. (2000) Coral reef disturbance and resilience in a human-dominated environment. *Trends in Ecology and Evolution*, **15**, 413–417.
- Pitcher, C.R., Doherty, P., Arnold, P. *et al.* (2007) *Seabed biodiversity on the continental shelf of the Great Barrier Reef world heritage area*. AIMS/CSIRO/QM/QDPI CRC Reef Research Task Final Report.
- Rosen, B.R. (1971) The distribution of reef coral genera in the Indian Ocean. *Symposia of the Zoological Society of London*, **28**, 263–299.
- Roy, K., Jablonski, D., Valentine, J.W. & Rosenberg, G. (1998) Marine latitudinal diversity gradients: tests of causal hypotheses. *Proceedings of the National Academy of Sciences USA*, **95**, 3699–3702.
- Sandin, S.A., Vermeij, M.J.A. & Hurlbert, A.H. (2008) Island biogeography of Caribbean coral reef fish. *Global Ecology and Biogeography*, **17**, 770–777.
- Selmi, S. & Boulinier, T. (2001) Ecological biogeography of Southern Ocean islands: the importance of considering spatial issues. *The American Naturalist*, **158**, 426–437.
- Sweatman, H., Burgess, S., Cheal, A., Coleman, G., Delean, S., Emslie, M., McDonald, A., Miller, I., Osborne, K. & Thompson, A. (2005) *Long-term monitoring of the Great Barrier Reef. Status report number 7*. Australian Institute of Marine Science, Townsville.
- Venables, W.N. & Ripley, B.D. (2002) *Modern applied statistics with S*, 4th edn Springer, New York.
- Williams, D.M. & Hatcher, A.I. (1983) Structure of fish communities on outer slopes of inshore, mid-shelf and outer-shelf reefs of the Great Barrier Reef. *Marine Ecology Progress Series*, **10**, 239–250.
- Wintle, B.A. & Bardos, D.C. (2006) Modeling species-habitat relationships with spatially autocorrelated observation data. *Ecological Applications*, **16**, 1945–1958.

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

**Figure S1** Principal component analysis of environmental variables.

**Figure S2** Diagnostic plots for the spatial generalized linear mixed-effect models.

**Table S1** Environmental variables available at a 0.01° spatial resolution for the Great Barrier Reef, Australia.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

## BIOSKETCHES

**Camille Mellin** is a post-doctoral research associate at the Australian Institute of Marine Science and she has a joint position at the University of Adelaide. Camille is mostly interested in biogeography, coral reef ecology and the impacts of global change on marine ecosystems.

Editor: José Alexandre F. Diniz-Filho