Developing physical surrogates for benthic biodiversity using co-located samples and regression tree models: a conceptual synthesis for a sandy temperate embayment

Zhi Huang*, Matthew McArthur, Lynda Radke, Tara Anderson, Scott Nichol, Justy Siwabessy and Brendan Brooke

Coastal, Marine and Climate Change Group, Geoscience Australia, Canberra, ACT 2601, Australia

(Received 29 August 2011; final version received 15 January 2012)

Marine physical and geochemical data can be valuable surrogates for predicting the distributions and assemblages of marine species. This study investigated the bio-environment (surrogacy) relationships in Jervis Bay, a sandy marine embayment in south-eastern Australia. A wide range of co-located physical data were analysed together with biological data, including multibeam bathymetry and backscatter surfaces and derivatives, parameters that describe seabed sediment and water column physical/geochemical characteristics and seabed exposure. Three decision tree models and a robust model selection process were applied. The models for three diversity indices and seven out of eight infaunal species explained 32–79% of the variance. A diverse range of physical surrogates for biodiversity were identified. The surrogates are presented in a conceptual model that identifies the mechanisms that potentially link to biodiversity patterns. While some surrogates may exert direct influence over organisms to exposure and chlorophyll-$a$, for example, most pointed to complex relationships between multiple biological and physical factors occurring in different process domains/zones. The reliable bio-environment relationships identified from co-located samples and conceptual models enabled a mechanistic understanding of benthic biodiversity patterns in a sandy coastal embayment that may have implications for marine environmental management.

Keywords: surrogates; benthic biodiversity; Jervis Bay; conceptual model

1. Introduction

The management of marine ecosystems requires accurate mapping of biodiversity patterns over large areas. The paucity of biological data and the prohibitive cost of directly sampling biota over large areas have led to the development of a mapping strategy based on physical surrogates (Zacharias and Roff 2000). In the marine zone, physical (geochemical) data are generally more readily measured than biological data. These physical variables can act as surrogate measures of biodiversity and be valuable in predicting the potential distributions and assemblages of marine species (McArthur et al. 2010). Importantly, modelling bio-environment relationships can help identify and verify ecological processes that link physical environmental attributes to the distribution of seabed biota (Austin 2002).

*Corresponding author. Email: zhi.huang@ga.gov.au
Many physical variables have been used as surrogates for marine biodiversity (Zacharias and Roff 2000, McArthur et al. 2010, Huang et al. 2011a) and can be grouped into three types: resource, direct and indirect variables (e.g. Austin and Smith 1989, Guisan and Zimmermann 2000, Austin 2002). Resource variables are consumed by biota. Direct variables exert a direct physical/physiological influence on biota. Indirect variables do not have physiological connections to biota but can correlate strongly with resource and direct variables that do influence biota. The resource and direct variables are preferred inputs for the modelling of bio-environment relationships, because they represent causal links (Austin 2002). The indirect variables, which are often easier to obtain, can also be useful because they may help to identify processes behind biological patterns.

There are many challenges to successfully model bio-environment relationships in a marine context. Data issues such as sample bias, inadequate sample size, missing important covariates and location and attribute errors can have significant impacts on the modelling results (e.g. Barry and Elith 2006, Hernandez et al. 2006, Pearce and Boyce 2006, Philips et al. 2006). In addition, selection of the appropriate spatial/statistical model(s) from the many that are available, optimisation of model performance and identification of potential surrogates from a model remain challenging tasks, despite the rapid development of modelling techniques.

The objectives of this study were to (i) use a modelling approach to identify the physical surrogates that can be used to map benthic biodiversity in a sandy marine embayment and (ii) develop a conceptual model that synthesises and conveys the causal mechanisms behind the surrogacy relationships in a process-orientated framework. In this study, the physical variables identified by the models as important were regarded as surrogates; and the inferred bio-environment relationships were called ‘surrogacy relationships’.

A unique feature of this study is that the models were developed using accurately co-located samples so that the derived surrogacy relationships would have greater reliability. Most previous marine ecological studies have relied on interpolated physical variables (e.g. Pitcher et al. 2007, Gogina et al. 2010, Mellin et al. 2010), which introduce some uncertainty into the modelling process/results. We also addressed several of the issues mentioned above through robust data collection and modelling strategies. First, modern marine survey technologies were used for data collection to ensure the highest position and attribute accuracy. Second, a large number of physical/environmental variables were collected and derived to reduce the chance of missing important covariates. Finally, machine learning models and a model selection process were employed to provide an appropriate and robust analysis framework.

2. Study area
Jervis Bay is situated on the south coast of New South Wales (Figure 1a). The bay entrance is 3.5 km wide, up to 30 m deep and is permanently open to ocean swell from the east to south-east. Jervis Bay is a typical example of a sandy marine embayment common to south-eastern Australia. It has significant ecological, social and economic values and forms part of the Jervis Bay Marine Park (NSW Marine Parks Authority 2010). This study concerns a $3 \times 5 \text{ km}^2$ area of seabed situated adjacent to the mouth of the bay, here termed the Darling Road Grid (DRG) (Figure 1; Anderson et al. 2009). This area was chosen for mapping and sampling because it contains most of the habitat
types identified in the bay (e.g. drift algae, polychaete hummocks, rippled and bioturbated sands, bivalve clumps and heart urchin areas) and supports a diverse fauna and flora (CSIRO 1993, Jacoby et al. 1995; Figure 1c). A previous study had used habitat categories and species assemblages as biological surrogates for evaluating marine biodiversity in Jervis Bay (Ward et al. 1999). However, physical surrogates remain unexplored in the area.

3. Materials and methods

3.1. Field methods
The DRG seabed was mapped in December 2007 and May 2008 using a Simrad EM3002 300 kHz multibeam sonar system (Kongsberg Maritime, Kongsberg, Norway), which provided bathymetry and seabed backscatter data (Anderson et al. 2009). Two sampling surveys were undertaken (24–28 June and 27–29 August 2008) to collect accurately co-located sediment samples for physical/geochemical and infaunal analyses and to make water column observations (water temperature (WT), photosynthetically available radiation (PAR) and dissolved oxygen (DO) concentrations). A Shipek-style grab was used to collect sediment samples designated for physical/geochemical analyses, and a Van Veen grab was used to collect the biological samples. Stations, with 60 samples, were allocated using a stratified-random spatial sampling design based on the a priori habitat designations (CSIRO 1993; Figure 1b).
3.2. Data processing and analytical methods

Laboratory analyses and data post-processing yielded a large number of physical/environmental variables for use in the machine learning models. These variables (detailed below) and the broad-scale benthic habitat types (CSIRO 1993) constitute the covariates (or predictors) in this study (Table 1).

Table 1. The list of abiotic variables.

| Name         | Description                           | Name     | Description                                                                 |
|--------------|---------------------------------------|----------|-----------------------------------------------------------------------------|
| Bathymetry   | Seabed water depth                    | Moran2_11| Local Moran’s I of backscatter in a 11 × 11 window                         |
| Slope_3      | Slope gradient in a 3 × 3 window       | Moran2_19| Local Moran’s I of backscatter in a 19 × 19 window                         |
| Slope_7      | Slope gradient in a 7 × 7 window       | Dist_160 | Averaged seabed disturbance from 160° wave direction                        |
| Slope_11     | Slope gradient in a 11 × 11 window     | Dist_90  | Averaged seabed disturbance from 90° wave direction                         |
| Slope_19     | Slope gradient in a 19 × 19 window     | Dist_all | Averaged seabed disturbance from all directions                             |
| TR_3         | Topographic relief in a 3 × 3 window   | MGS      | Sediment mean grain size                                                    |
| TR_7         | Topographic relief in a 7 × 7 window   | Mud      | %mud                                                                        |
| TR_11        | Topographic relief in a 11 × 11 window | Sand     | %sand                                                                      |
| TR_19        | Topographic relief in a 19 × 19 window | Gravel   | %gravel                                                                    |
| SA_3         | Surface area in a 3 × 3 window         | CaCO₃    | %carbonate                                                                 |
| SA_7         | Surface area in a 7 × 7 window         | Sorting  |                                                                           |
| SA_11        | Surface area in a 11 × 11 window       | Skewness |                                                                           |
| SA_19        | Surface area in a 19 × 19 window       | Kurtosis |                                                                           |
| TPI_3        | Topographic position index in a 3 × 3 window | SSA | Sediment surface area                                                                 |
| TPI_7        | Topographic position index in a 7 × 7 window | Chl-a | Sediment chlorophyll-a concentration                                        |
| TPI_11       | Topographic position index in a 11 × 11 window | TSM | Total sediment metabolism                                                   |
| TPI_19       | Topographic position index in a 19 × 19 window | pTCO₂ | Sediment TCO₂ pool                                                          |
| Moran1_3     | Local Moran’s I of bathymetry in a 3 × 3 window | TS | Total sulphur                                                              |
| Moran1_7     | Local Moran’s I of bathymetry in a 7 × 7 window | Fe | Water temperature                                                          |
| Moran1_11    | Local Moran’s I of bathymetry in a 11 × 11 window | WT | Light attenuation                                                           |
| Moran1_19    | Local Moran’s I of bathymetry in a 19 × 19 window | DO | Water column dissolved oxygen concentration                                 |
| Backscatter  | Backscatter intensity                  | K₄       | Broad habitat classification derived from CSIRO (1993) study                |
| Moran2_3     | Local Moran’s I of backscatter in a 3 × 3 window | Habitat |                                                                           |
| Moran2_7     | Local Moran’s I of backscatter in a 7 × 7 window | Aspect | Slope directions                                                            |
3.2.1. Multibeam acoustic data and its derivatives

Processing of multibeam bathymetry data was undertaken using Caris Hips and Sips V6.1 software (developed by CARIS, New Brunswick, Canada; http://www.caris.com/products/hips-sips/) to correct for tidal variations and to remove data artefacts. The multibeam backscatter data were processed using CMST-GA MB Process v8.11.02.1 software, a multibeam backscatter processing toolbox co-developed by Geoscience Australia and the Centre for Marine Science and Technology at Curtin University of Technology, Australia. The backscatter processing included correction for transmission loss and ensonification area and removal of the system-implemented model and the angular dependence (Gavrilov et al. 2005). The angularly equalised backscatter strengths were normalised to the backscatter strength at an angle of 25° (Heap et al. 2009). The bathymetry and backscatter data were gridded at 1 m resolution.

Bathymetry (depth) is a potentially powerful physical surrogate for benthic biodiversity (e.g. Gogina et al. 2010, McArthur et al. 2010) due to its relationship with light availability, food availability and WT. The following seabed morphological variables were derived from the bathymetry data: slope, topographic position index (TPI) (Weiss 2001), surface area (SA) (Jenness 2004), topographic relief (TR), local Moran’s index (Moran1) (Anselin 1995) and aspect. With the exception of aspect, these variables describe seabed rugosity and heterogeneity that can affect species diversity and abundance (e.g. Kostylev et al. 2005, Holmes et al. 2008, McArthur et al. 2010). For example, Moran1 measures heterogeneity (spatial autocorrelation) of seabed bathymetry values within a local neighbourhood. Backscatter has been associated with substrate composition (Goff et al. 2000, Kloser et al. 2001, Siwabessy et al. 2006) and is also a potential physical surrogate of benthic biodiversity (McArthur et al. 2010). In this study, local Moran’s I was also derived from the backscatter data as a texture measure (Moran2). All the morphological variables and texture measures were calculated for four window sizes (3 × 3, 7 × 7, 11 × 11 and 19 × 19 m) to capture the effect of spatial scale (Holmes et al. 2008). Aspect was calculated from a 3 × 3 window only. Aspect can characterise potential exposure when used in conjunction with outputs from an oceanographic model and thus is also a potential surrogate for patterns of benthic biodiversity (Holmes et al. 2008).

3.2.2. Sediment and water column data

A range of sediment properties were derived from laboratory analysis and subsequent post-processing, including mean grain size (MGS), %mud (Mud), %sand (Sand), %gravel (Gravel), %carbonate (CaCO3), sorting, skewness, kurtosis, sediment surface area (SSA), chlorophyll-a (chl-a) concentrations, total sediment metabolism (TSM), total carbon dioxide pools (pTCO2) and total sulphur (TS) and Fe concentrations. These sediment variables can provide information on the nature of the interstitial environment, food availability and redox characteristics (e.g. McArthur et al. 2010, Radke et al. 2011a, 2011b). The analytical methods used to determine sediment size fractions, SSA and the geochemical concentrations are described in Radke et al. (2011a). Summary statistics for sediments were derived from laser grain-size data using the computer program GRADISTAT (Kenneth Pye Associates Ltd, Crowthorne, UK) (Blott and Pye 2001).

The light attenuation coefficient (Kd) between the surface and benthic PAR measurements was calculated from a standard equation \( K_d = \ln(lo/lz)/z \), in which \( lo \) and \( lz \) refer to PAR at the surface and at depth, and \( z \) is the water column depth. This water column property and WT and DO concentration were also used for this
study as these variables have direct influence over benthic organism distributions (Snelgrove 2001).

3.2.3. Seabed exposure data

Seabed exposure to wave and current energy can play an important role in shaping local biodiversity patterns. According to Connell (1978), the highest biodiversity is likely to occur in areas with intermediate levels of disturbance. A fine-scale hydrodynamic model, the Simulating Waves Nearshore (SWAN) model (Booij et al. 1999, Ris et al. 1999), was developed to estimate seabed disturbance for Jervis Bay. The SWAN model estimates wave propagation using the wave action equation (Hasselmann et al. 1973). The model takes into account the refraction of swell waves, plus shoaling, diffraction, dissipation and random waves, but excludes local wave generation by wind. Apart from bathymetry data, the model also required three parameters: significant wave height ($H_{\text{m0}}$), peak wave period ($T_p$) and peak wave direction ($\theta_p$). Twelve combination scenarios of these three parameters were simulated, as summarised in Table 2. The model output of maximum orbital velocity (m s$^{-1}$) was considered as the best estimate of seabed exposure. Their values were combined to generate three seabed exposure variables: averaging all 12 scenarios (Dist_all), averaging eight scenarios with $\theta_p = 160$ (Dist_160; from south-east) and averaging four scenarios with $\theta_p = 90$ (Dist_90; from east).

3.2.4. Response variables – infaunal data

Biological material retained on a 500 µm sieve (after elutriation in the field) was examined under a dissecting microscope. All animal materials were removed, identified and counted. Three biodiversity indices (total abundance, species richness and Shannon index) were calculated using PRIMER v.6 PRIMER-E Ltd, Ivybridge, UK (Clarke and Warwick 2001) and were used in our analyses. To provide a generality in species responses to benthic habitats, eight infaunal species were selected to represent taxa that were abundant (e.g. >100 individuals) but that differed by phyla and in their feeding mechanisms and life histories (Table 3).

### Table 2. Wave parameter scenarios for the SWAN model.

| Scenario | $H_{\text{m0}}$ (m) | $T_p$ (s) | $\theta_p$ (°) |
|----------|-------------------|----------|----------------|
| 1        | 2                 | 6        | 160            |
| 2        | 2                 | 8        | 160            |
| 3        | 2                 | 10       | 160            |
| 4        | 3                 | 8        | 160            |
| 5        | 3                 | 10       | 160            |
| 6        | 3                 | 12       | 160            |
| 7        | 4                 | 10       | 160            |
| 8        | 4                 | 12       | 160            |
| 9        | 3                 | 10       | 90             |
| 10       | 3                 | 12       | 90             |
| 11       | 4                 | 10       | 90             |
| 12       | 4                 | 12       | 90             |

Note: SWAN, Simulating Waves Nearshore; $H_{\text{m0}}$, significant wave height; $T_p$, wave period; and $\theta_p$, wave direction.
Table 3. Eight selected infaunal species and their properties.

| Taxon                     | Class      | Habit       | Feeding                | Total count | Min | Max | Mean | SD  |
|---------------------------|------------|-------------|------------------------|-------------|-----|-----|------|-----|
| Mesochaetopterus minutus  | Polychaeta | Tube builder| Suspension feeder      | 1820        | 0   | 732 | 28.51| 108.38 |
| Unciolidae                | Gammaridea | Shell grit  | Plankton feeder        | 519         | 0   | 63  | 6.39 | 11.23 |
| Rhinocetes robustus       | Gammaridea | Shell grit  | Plankton feeder        | 468         | 0   | 129 | 7.05 | 21.71 |
| Venerupis galactites      | Bivalvia   | Shallow infauna | Suspension feeder  | 327         | 0   | 108 | 5.39 | 14.80 |
| Owenia australis          | Polychaeta | Tube builder| Selective deposit feeder| 320         | 0   | 32  | 3.49 | 6.56  |
| Diopatra dentata          | Polychaeta | Tube builder| Unknown                | 255         | 0   | 36  | 3.86 | 8.10  |
| Nephtys inornata          | Polychaeta | Shallow infauna | Predator      | 183         | 0   | 11  | 2.73 | 3.02  |
| Echinocardium cordatum    | Echinodermata | Shallow infauna | Non-selective deposit feeder | 113         | 0   | 18  | 1.76 | 3.15  |

3.3. Numerical analyses/modelling

3.3.1. Selection of covariates

Having too many covariates in a model can reduce prediction performance (Huang et al. 2011c). Therefore, we undertook exploratory analyses to reduce the number of covariates in the individual models. First, scatter plots illustrating the relationships between covariates and response variables were generated and fitted with the curve that provided the best explanatory power (indicated by $R^2$). The following five functions were examined, where possible, for least squares fits: linear, polynomial, logarithmic, exponential and power. Covariates with $R^2$ values greater than 0.05 in the exploratory analysis were deemed suitable for the modelling and were included in the selection. Second, the correlation coefficient ($\rho$) between a pair of covariates was calculated and ranked. Lower-ranked variables (i.e. with smaller $R^2$ values) in strongly correlated pairs ($|\rho| > 0.75$) were removed from the selection. In the final step, the variables habitat and aspect were added to the selection to obtain a set of covariates for the modelling (denoted Set₁). It should be noted that Set₁ was different for individual response variables. The results of the exploratory analyses are reported in the supplementary materials available online.

3.3.2. The statistical modelling process

Decision tree models were used for the regression problem because they have (1) the flexibility to handle both continuous and categorical variables; (2) the capability to handle high-dimension (often correlated) explanatory variables; (3) the ability to handle missing data in explanatory variables; (4) the ability to simulate complex non-linear relationships between explanatory and response variables; (5) the ability to fit interactions between
explanatory variables; (6) the ability to identify and rank explanatory variables; (7) the ease of interpretation and thus the ability to discover relationships between explanatory and response variables; and (8) robust performance (e.g. De’ath and Fabricius 2000, Shan et al. 2006, Hochachka et al. 2007, Pitcher et al. 2007, Holmes et al. 2008, Pittman et al. 2009). We implemented three decision tree models (single decision tree (DT) (Breiman et al. 1984), boosted decision tree (BDT) (Friedman 1999) and random forest decision tree (RFDT) (Breiman 2001)) using DTREG software (http://www.dtreg.com). Although not shown in this article, these models can also be used to generate prediction maps.

The selection of model parameters and covariates both can influence prediction performance. This study used a series of steps to find a ‘best’ performing model for each response variable. The following were the three main steps in the modelling procedure:

(1) **Search model parameters.** This step used all the covariates in Set1. An intensive experimental process was undertaken to search for the combination of model parameters that would obtain the highest prediction $D^2$ (percent variance explained, see Section 3.3.3). For BDT, we varied four model parameters. The parameter of ‘maximum number of trees in series’ was varied between 400 and 1000 with an increment of 100; the parameter of ‘depth of individual trees’ was varied between 4 and 6 with an increment of 1; the parameter of ‘minimum size node to split’ was varied as either 2 or 3; and the number of cross-validation folds was varied between 2 and 10 with an increment of 1. All other model parameters were set at default. For RFDT, we varied three model parameters. The parameters of ‘depth of individual trees’ and ‘minimum size node to split’ were varied as BDT. The ‘number of trees’ was varied between 100 and 1000 with an increment of 100. DT does not have the ‘number of trees’ parameter and the three model parameters were varied the same as BDT. In addition, for DT and BDT, pruning (to minimise cross-validated error) was implemented to improve performance and generalisation.

(2) **Search explanatory variables.** This step used the ‘best’ combination of model parameters provided by Step 1. In the first iteration, only one covariate was added to the model. This was repeated for all covariates. The variable which provided the highest prediction $D^2$ was retained for the next iteration; and the $D^2$ value was recorded. Each of the subsequent iterations selected a different covariate until all covariates had been added to the model. The combination of covariate(s) which achieved the highest prediction $D^2$ value was selected as the ‘best’ set of covariates (denoted Set2).

(3) **Refine the model.** The covariates in Set2 were removed from the model one by one. The variable which, upon removal, gave the highest prediction $D^2$ was not included in the model. The model was then refined by tuning the model parameters like in Step 1; and the $D^2$ value was recorded. This procedure was repeated until all variables were removed from the model. The combination of the variables with the highest prediction $D^2$ was selected as the final combination of covariates (denoted Set3).

The ability to identify and rank the importance of covariates is critical for identifying physical surrogates for biodiversity. The three decision tree models were able to do this by summing the improvement in accuracy gained by each split that used the covariate. The importance score for the most important covariate in Set3 is set to a value of
100. Other covariates will have lower scores scaled accordingly (Sherrod 2008). In this study, the covariate with an importance score greater than 80 was regarded as a (physical) surrogate.

3.3.3. Assessment of model performance

The statistical performance of the models was evaluated using either a cross-validation procedure (e.g. DT and BDT) or a bagging process (e.g. RFDT) with two statistics: $D^2$ and mean absolute error (MAE). $D^2$, also called $R^2$, measures the percent variance explained by the model. MAE is an average of the absolute errors between the prediction and the true values. $D^2$ values in marine ecological applications generally range from $\sim$0.3 to 0.7 (e.g. Dzeroski and Drumm 2003, Leathwick et al. 2006, Pitcher et al. 2007, Pittman et al. 2009, Hill et al. 2010). In this study, $D^2$ values in the ranges $0.3–0.5, 0.5–0.7, 0.7–0.9$ and $>0.9$ are regarded as ‘acceptable’, ‘moderate’, ‘good’ and ‘excellent’, respectively.

3.3.4. The prediction of response curves

Response curves were constructed to help interpret the bio-environment (surrogacy) relationships identified by the modelling process. We constructed an artificial data set for each identified surrogate. If the surrogate was a continuous variable (e.g. bathymetry), the artificial data set would have 30 rows (data records) and the number of columns (data attributes) equal to the number of covariates. Using the bathymetry variable as an example, the column for the bathymetry attribute has values ranging from its minimum to its maximum sample values with equal increments. The value of any other columns was kept as a constant which was either the mean value (in cases of continuous variables) or the most frequent value (in cases of categorical variables) of the respective covariate. If the surrogate was a categorical variable (e.g. habitat), the artificial data set was given the same number of rows as there were categories (e.g. types of habitat). The values for all other columns were similarly kept as constants.

The artificial data set was then passed to the selected model to predict the values for the selected response variable. The values for the surrogate (between minimum and maximum of the sampled values, as the $x$-axis) were then plotted against the predicted values of the response variable (as the $y$-axis) to construct a predicted response curve.

4. Results

The statistical results of the decision tree models are summarised in Table 4. There was no single best performing model amongst the decision trees according to the $D^2$ and MAE results. However, DT was the best overall model for individual species and BDT was the best performing model for the diversity indices. The statistical modelling performance of Shannon index and species richness was moderate ($D^2$ between 0.5 and 0.7), while the statistical performance for total abundance was acceptable ($D^2$ between 0.3 and 0.5). In terms of the eight selected species, the model for *Diopatra dentata* had good statistical performance ($D^2$ between 0.7 and 0.9); the models for *Rhinocetes robustus* and *Owenia australis* had moderate statistical performance; the models for *Nephtys inornata*, *Mesochaetopterus minutus*, *Echinocardium cordatum* and Uncioliidae had acceptable statistical performance; and the model for *V. galactites* had poor statistical performance ($D^2 < 0.3$).
Table 4. The results of the surrogacy analysis.

|                       | $D^2$       | MAE       |
|-----------------------|-------------|-----------|
|                       | BDT | RFDT | DT | BDT | RFDT | DT |
| Total abundance       | 0.366| 0.27  | 0.238| 76.34| 80.5 | 83.08 |
| Species richness      | 0.519| 0.423 | 0.439| 7.87 | 8.57 | 8.07 |
| Shannon index         | 0.637| 0.443 | 0.381| 0.29 | 0.35 | 0.36 |
| *Mesochaetopterus minutus* | 0.289| 0.349 | 0.311| 28.52| 32.35| 27.99 |
| *Echinocardium cordatum* | 0.112| 0.214 | 0.316| 1.66 | 1.86 | 1.78 |
| *Venerupis galactites* | 0.228| 0.214 | 0.316| 4.75 | 5.06 | 5.39 |
| Uncioliidae           | 0.221| 0.313 | 0.333| 5.9  | 5.6  | 6.12 |
| *Diopatra dentata*    | 0.79 | 0.71  | 0.785| 2.09 | 2.53 | 1.97 |
| *Rhinocetes robustus* | 0.374| 0.312 | 0.526| 6.79 | 8.58 | 5.58 |
| *Nephtys inornata*    | 0.389| 0.388 | 0.412| 1.78 | 1.72 | 1.7  |
| *Owenia australis*    | 0.601| 0.503 | 0.633| 2.41 | 2.72 | 2.44 |

Notes: MAE, mean absolute error; BDT, boosted decision tree; RFDT, random forest decision tree; DT, single decision tree.
The best performance indicators are shown in bold.

In general, the predicted response curves (Figures 2 and 3) correspond well with the observed bio-environment relationships (Figures S1 and S2, supplementary materials). The broad benthic habitat types were found to be the best surrogate for the total abundance of infaunal organisms. $K_d$ and DO and TS concentrations may also be useful surrogates (Table 5; Table S2 available online). The predicted response curves (Figure 2a–c) show that higher total infaunal abundance occurs in the bioturbated sand, bivalve clump and polychaete hummock habitats, and locations with higher light availability and low to intermediate total sulphur concentrations. A range of surrogates were found for species richness, including SA_11, Mud, chl-$\alpha$, Fe, habitat and TSM (Table 5; Table S3 available online). Species richness was also predicted to be highest in bivalve clumps and lowest in drift algae and ripple sands. Interestingly, the results for the food indicators were different. There was a general increase of species richness with chl-$\alpha$ concentrations and a general decrease with total sediment metabolism. For the Shannon index, Moran1 was the strongest surrogate followed by Gravel (Table 5, Table S4). Higher Shannon diversity was found in relatively heterogeneous seabed (in terms of bathymetry) with low gravel concentrations. The gravel size fraction in these samples was due to the polychaete tubes and bound debris.

Several physical variables were identified as surrogates for individual species (Table 5, Tables S5–S12). Here, we only discuss the findings on the four species with the best modelling performance (Table 4). For *D. dentata*, Dist_160 was consistently identified as the most important physical control (Table 5). The predicted response curve (Figure 3a) shows that *D. dentata* occurs in areas of intermediate to high seabed exposure. CaCO$_3$ was consistently the best surrogate of the crustacean *R. robustus* (Table 5). Another identified surrogate was Dist_160. The predicted response curves (Figure 3b) show that higher numbers of *R. robustus* are found in sediments with low bulk carbonate content (<10%). Higher abundances of *R. robustus* also occur in areas of high seabed disturbance (Figure 3c). For *N. inornata*, an errant predator polychaete, habitat was most important (Table 5). *Nephtys inornata* occurs in high numbers in the bivalve clump and polychaete hummock habitats (Figure 3e). SA_11 was identified as the most important factor governing the distribution of the polychaete *O. australis* (Table 5), which is a selective deposit feeder. The predicted
Figure 2. Predicted response curves for biodiversity indices: (a) habitat type and total abundance (refer to Figure 1c for the correspondence between the numbers and names of the habitat types); (b) $K_d$ and total abundance; (c) TS and total abundance; (d) SA_11 and species richness; (e) %mud and species richness; (f) chlorophyll-a and species richness; (g) Fe and species richness; (h) habitat type and species richness; (i) TSM and species richness; (j) local Moran’s $I$ of bathymetry and Shannon index; (k) %gravel and Shannon index.
response curve (Figure 3d) shows that *O. australis* is found in high numbers on extremely flat seabed in the south-western part of the bay.

The response curves (Figure 2) are interpreted with reference to a conceptual model that presents the surrogates in a process-based framework (Figure 4). The conceptual model incorporates the broad habitats into three zones. From west to east these include (i) a resuspension-dominated zone; (ii) a bioturbation zone with a sub-zone noted for high *M. minutus* abundance (Figure 5a); and (iii) a wave-dominated zone. The model was informed by the results of this study (Figure 2), the raw data (Table 6) and a companion study (Radke et al. 2011a), which provided spatial context for a subset of the variables. The approximate spatial distribution of the zones is shown in Figure 5 and was determined by the magnitude of the local Moran’s *I* parameter in relation to our *a priori* understanding of the broad-scale energetics of the DRG and its bearing on the distribution of benthic habitats and processes (Radke et al. 2011a). Overall, the benthic biodiversity shows strong zonation differences across the study area (Table 6). For example, the resuspension-dominated zone has the lowest species richness, whereas the bioturbation zone has the highest species richness (Figure 5b; Table 6).

Figure 3. Predicted response curves for individual species: (a) seabed disturbance and the abundance of *Diopatra dentata*; (b) %CaCO₃ and the abundance of *Rhinocetes robustus*; (c) seabed disturbance and the abundance of *R. robustus*; (d) seabed rugosity and the abundance of *Owenia australis*; (e) habitat type and the abundance of *Nephtys inornata*. 
Table 5. The important variables (score >80) selected for the diversity indices and individual species.

| Total abundance | BDT                  | RFDT                  | DT       |
|-----------------|----------------------|-----------------------|----------|
| Species richness | Total abundance      | Habitat, $K_d$, TS    | Mud, TS, habitat, DO, $K_d$ |
| Shannon index   | Habitat, mud, SA_11, Gravel | Morlan1_11, bathmetry | Bathmetry, SA_11 |
| Mesochaetopterus minutus | Moran1_7, Gravel | Slope_3, | Habitat |
| Echinocardium cordatum | Sorting | Sorting | Sorting, pTSCO_2 |
| Venerupis galactites | Moran2_3, habitat | Moran2_3 | Sorting |
| Uncioliidae | CaCO_3 | SSA, skewness | $K_d$ |
| Diopatra dentata | Dist_160 | Dist_160 | Dist_160 |
| Rhinocetes robustus | CaCO_3, Dist_160 | CaCO_3 | CaCO_3 |
| Nephtys inornata | Habitat | Habitat | Habitat |
| Owenia australis | SA_11 | SA_11 | SA_11 |

Note: BDT, boosted decision tree; RFDT, random forest decision tree; DT, single decision tree.

Figure 4. Conceptual model of bio-environment relationships and supporting contextual information. The conceptual model has three zones and one sub-zone, which incorporate the broad habitats. From west to east these include (i) a resuspension-dominated zone; (ii) a bioturbation zone with a sub-zone noted for active mound-building by the polychaete *Mesochaetopterus minutus*; and (iii) a wave-dominated zone.
Figure 5. Positions of the zones as they relate to the study area map (Figure 1c): (a) zones overlaid on local Moran’s I of bathymetry, the abundance of *Mesochaetopterus minutus* is also shown; (b) species richness samples overlaid on the zone map.

Table 6. Actual measure values of surrogates and indices in the conceptual zones.

| Variable \ zone | Wave-dominated zone | Bioturbation zone | Mesochaetopterus sub-zone | Resuspension-dominated zone |
|----------------|---------------------|-------------------|--------------------------|----------------------------|
| Mean ! SD Species richness | 30.3 ± 9.9 | 39.3 ± 15.2 | 28.9 ± 9.5 | 16.8 ± 4.4 |
| Total abundance | 150.8 ± 100.9 | 191.1 ± 116.6 | 245.4 ± 225.2 | 63.6 ± 30.8 |
| Shannon index | 2.7 ± 0.3 | 2.8 ± 0.5 | 2.0 ± 0.9 | 2.3 ± 0.1 |
| Water depth (m) | 26.4 ± 1.8 | 19.8 ± 3.5 | 19.6 ± 1.3 | 15.8 ± 1.0 |
| SA_11 | 1.00014 ± 0.000075 | 1.00012 ± 0.00012 | 1.00019 ± 0.00013 | 1.000047 ± 0.000033 |
| Moran1_11 | 57.4 ± 31.4 | 17.3 ± 60.0 | 2.7 ± 5.1 | 29.8 ± 14.5 |
| Iron (Fe) (mg kg⁻¹) | 6.0 ± 4.6 | 2.9 ± 2.8 | 3.4 ± 2.3 | 3.2 ± 2.5 |
| TS (mg m⁻²) | 0.4 ± 0.2 | 0.4 ± 0.2 | 0.5 ± 0.1 | 0.6 ± 0.2 |
| Chl-a (µg) | 0.5 ± 0.2 | 0.9 ± 0.5 | 1.0 ± 0.5 | 0.4 ± 0.2 |
| TSM (µmol cm⁻³ d⁻¹) | 0.4 ± 0.3 | 0.5 ± 0.4 | 1.2 ± 0.8 | 0.5 ± 0.2 |
| Gravel (%) | 0.2 ± 0.3 | 0.3 ± 0.4 | 0.9 ± 0.9 | 0.1 ± 0.1 |
| Mud (%) | 0.1 ± 0.1 | 4.8 ± 5.0 | 2.9 ± 2.2 | 20.2 ± 14.8 |
| K_d (m⁻¹) | 0.11 ± 0.04 | 0.14 ± 0.02 | 0.12 ± 0.02 | 0.16 ± 0.04 |
5. Discussion

5.1. Model performance

The three decision tree models used in this study were well suited for investigating bio-environment relationships. Model performances were at least acceptable for all three diversity indices and for seven out of eight selected infaunal species. The values of $D^2 > 0.5$ (in 5 of 11 cases; Table 4) are indeed promising for marine biodiversity prediction (e.g. Dzeroski and Drumm 2003, Leathwick et al. 2006, Pitcher et al. 2007, Pittman et al. 2009, Hill et al. 2010). Using co-located physical and biological samples instead of interpolated data minimised positional and attribute errors and ensured the reliability of derived bio-environment relationships. However, the three decision tree models were not consistent in their performance (Table 4). There was no single best performing model, which supports our strategy of using multiple models, as has been shown elsewhere (e.g. Elith et al. 2006, Huang et al. 2011b). The predicted response curves (Figures 2 and 3) had various shapes highlighting that machine learning models are appropriate for simulating complex bio-environment relationships. However, given potential limitations with data and statistical models, the relationships identified by the models are not necessarily true and/or causal. Rather, the success of an ecological modelling process should be judged not only by model statistical performance but also by the biophysical knowledge that the model imparts (Austin 2002) (detailed in Section 5.2). It should be noted that several factors can affect the effectiveness of biodiversity surrogates and that there are other ways of evaluating their effectiveness (e.g. Grantham et al. 2010) than those used in this study.

5.2. Conceptual synthesis of surrogacy relationships and causal links

The ‘productivity-richness’ hypothesis (reviewed in Tittensor et al. (2010)) predicts a positive relationship between species richness and productivity, which is consistent with most of our observations in Jervis Bay. The broad sweep of habitats in the central DRG (e.g. bioturbated sands, bivalve clumps and polychaete hummocks), comprising the bioturbation-dominated zone in our conceptualisation (Figure 5), had the highest average values for sediment chl-a concentrations, and likewise, the largest predicted values for species richness and total abundance (Figure 2a and h). Indeed, our results are concordant with the findings of Jacoby et al. (1995) in that the predicted species richness was highest in the bivalve clumps and lowest in drift algae and rippled sands (Figure 2h).

Species richness increased with chl-a concentrations and decreased with total sediment metabolism (Figure 2f and i), which may imply that richness was supported more by freshly deposited phytoplankton and/or benthic micro-algae than by reactive bulk organic matter which may include fresh phytodetritus as well as other labile factors (e.g. dissolved organic carbon and carbohydrates, lipids, proteins and nucleic acids of plant or animal origin) that are measured as total sediment metabolism. However, highest total sediment metabolism occurred in association with a M. minutus-dominated sub-zone, because polychaetes secrete labile mucoid substances as burrow linings (Kristensen 2000). The M. minutus-dominated sub-zone had relatively low Shannon diversity index and the largest number of organisms of the zones (Table 6). The low diversity of the M. minutus-dominated sub-zone was reflected in the modelling because the Shannon index was predicted to be low in sediments with high gravel percentages and low local Moran’s $I$ values (Figure 2j and k), both of which occurred here. A potential explanation for the modelled decrease in species richness with total sediment metabolism (and %gravel and local Moran’s $I$) is thus that
some species were excluded from the polychaete mounds by either direct competition with *M. minutus* or as an indirect consequence of habitat modification by this species.

Ecosystem engineering (including bioturbation) involves the construction, maintenance and modification of habitats by organisms (Gutiérrez *et al.* 2002) and includes the modulation of abiotic materials and their fluxes (Berkenbusch and Rowden 2003). While mounds built by the polychaete *M. minutus* are the most overt features of ecosystem engineering in the DRG, the seabed of the bioturbation zone (and especially the *M. minutus* sub-zone) is also recognised for its high level of biogeochemical heterogeneity (Radke *et al.* 2011a). It is likely that the lateral extent of bioturbation activity and/or the preservation of bioturbation features is constrained by wave-dominated processes occurring to the west and east of the bioturbation zone because these seabed environments have less heterogeneity in bathymetry (i.e. lower values for local Moran’s *I*, Figure 5), thus giving definition to the resuspension- and wave-dominated zones (see also Radke *et al.* 2011a).

It is widely observed that energetic environments support specialised species and low-diversity communities (Gray 2002), and this is consistent with our results. The rippled sand habitat (comprising the wave-dominated zone in our conceptualisation) is subject to high wave energy from the open ocean (CSIRO 1993) and is predicted to support a fauna that is low in both richness and abundance (Figure 2a). Likewise, the seafloor of the drift-algae habitat can experience significant reworking by waves under modal wave conditions (resuspension-dominated zone) (Radke *et al.* 2011a) and is also predicted to have a fauna that is low in richness and abundance. Moreover, the modelled abundances of some species (e.g. *D. dentata* and *R. robustus*) were highest in the wave-dominated zone as evidenced by preferences for intermediate to high exposure to wave energy from the south-east direction (Figure 3a and c). The robust mucous/sediment and silk/sediment tubes that these taxa, respectively, build could explain their ability to recruit well to mechanically challenging habitats. Likewise, a predilection for filter feeding (as with *R. robustus*; Table 3) would be advantageous in areas where strong currents reduce the settling of organic matter.

Our results point to a potential levelling influence of *K*<sub>d</sub> and %mud on species abundances and richness (Figure 2b and e), and the response curve for mud (and likewise gravel) concentrations is contrary to the widely held view that biodiversity is higher in sediments with diverse grain sizes (Whitlatch 1977), for example, higher mud (or gravel) concentrations in a sandy matrix. The highest values of these surrogates occur in the resuspension-dominated zone (Table 6), and their potential influences on biodiversity have to be reconciled with consideration of more likely controls such as wave-induced disturbance (resuspension) and sediment anoxia (Radke *et al.* 2011a). Modelled organism abundances were found to decrease with increasing total sulphur concentrations (Figure 2c). These concentrations, which were highest in the resuspension-dominated zone (Table 6), reflect the occurrence of an anoxic process called sulphate reduction which can produce pyrite and iron monosulphides (FeS<sub>2</sub> and FeS) from reactive iron. The anoxia was patchy and was caused by degradation of the drift algae and its propensity, when abundant, to inhibit sediment resuspension and the turbulent transfer of oxygen to the seabed (Radke *et al.* 2011a).

6. Concluding synthesis
The results of this study are promising, with acceptable to good model statistical performance for the three diversity indices and seven out of eight infaunal species. Using co-located physical and biological samples ensured the reliability of derived bio-environment relationships. Patterns of abundance and diversity were statistically related to
a wide range of different variable types, including physical sediment (e.g. mud, CaCO$_3$, gravel) and geochemical properties (e.g. Fe, chl-$a$, TSM), seabed morphometric characteristics (e.g. local Moran's $I$ of bathymetry, rugosity), seabed exposure regime (e.g. Dist$_{160}$) and water column characteristics ($K_d$). Of these, habitat, $K_d$, TS, chl-$a$, Fe, TSM, SA$_{11}$, Moran1$_{11}$, mud and gravel could be considered surrogates of biodiversity and/or species patterns.

Past studies suggest that the influences of the resource/direct variables (e.g. the exposure variables and chlorophyll) on biodiversity may be causal (reviewed by McArthur et al. (2010)). However, most surrogacy relationships identified in this study were indirect (e.g. total sediment metabolism, total sulphur, seabed rugosity/heterogeneity, gravel, mud, habitats) and pointed to complex interactions between multiple physical and biological factors. These complex interactions gave definition to different process domains/zones and in some cases produced modelled results that were contradictory to expected (e.g. species richness response to mud content, TSM, Shannon diversity response to gravel content).

While the Shannon diversity and abundance models strongly reflected conditions in the $M$. minutus hummocks and drift-algae habitats, respectively, the species richness models provided information across study area. The species richness models likely captured various dimensions of beta-diversity whereby species composition changed across space, at various scales, in accordance with the environmental preferences/tolerances of individual species to factors, such as productivity (represented by chl-$a$ concentrations), anoxia (indirectly related to Mud), seabed exposure (indirectly related to habitat type) and competitive exclusion (indirectly related to total sediment metabolism and rugosity).

In conclusion, the modelling procedure used in this study in combination with a comprehensive data set comprising co-located samples shows great potential for both identifying biodiversity surrogates and ecological process mapping. Here we developed a mechanistic understanding of benthic biodiversity patterns in a sandy coastal embayment that may have implications for marine environmental management. In addition, the decision tree models developed can be used to generate the prediction maps of benthic biodiversity and individual species in the study area. These prediction maps can provide important information for the management and monitoring of the Jervis Bay Marine Park.

Acknowledgements
This study was undertaken as part of the Surrogates Program of the Marine Biodiversity Hub, funded by the Commonwealth Environment Research Facilities (CERF) programme, an Australian Government initiative supporting world class, public good research. The CERF Marine Biodiversity Hub is a collaborative partnership between the University of Tasmania, CSIRO Wealth from Oceans Flagship, Geoscience Australia, Australian Institute of Marine Science and Museum Victoria. Support for field surveys was provided by R. De Graf, D. Holdway, S. Hodgkin, I. Atkinson, A. Hislop, C. Wintle, S. Hancock, M. Carey (Geoscience Australia) and Scott Corson, Master of RV Kimbla. We thank James Daniell and Nic Bax for comments on the draft manuscript. Constructive comments provided by two anonymous reviewers and the editor are gratefully acknowledged. This article is published with the permission of the Chief Executive Officer of Geoscience Australia.

References
Anderson, T., et al., 2009. Mapping and characterising soft-sediment habitats, and evaluating physical variables as surrogates of biodiversity in Jervis Bay, NSW [online]. Geoscience Australia, Record 2009/10. Geoscience Australia, Canberra, 43 pp. Available from: http://www.ga.gov.au/image_cache/GA14300.pdf (accessed July 2010).
Anselin, L., 1995. Local indicators of spatial association – LISA. Geographical Analysis, 27, 93–115.
Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling, 157, 101–118.

Austin, M.P. and Smith, T.M., 1989. A new model for the continuum concept. Vegetation, 83, 35–47.

Barry, S. and Elith, J., 2006. Error and uncertainty in habitat models. Journal of Applied Ecology, 43, 413–423.

Berkenbusch, K. and Rowden, A.A., 2003. Ecosystem engineering – moving away from ‘just-so’ stories. New Zealand Journal of Ecology, 27 (1), 67–74.

Blott, S.J. and Pye, K., 2001. GRADISTAT: a grain size distribution and statistics package for the analysis of unconsolidated sediments. Earth Surface Processes and Landforms, 26, 1237–1248.

Booji, N., Ris, R.C., and Holthuijzen, L.H., 1999. A third-generation wave model for coastal region 1. Model description and validation. Journal of Geophysical Research, 104, 7649–7666.

Breiman, L., 2001. Random forests. Machine Learning, 45, 5–32.

Breiman, L.H., et al., 1984. Classification and regression tree. Belmont, CA: Wadsworth International.

Clarke, K.R. and Warwick, R.M., 2001. Change in marine communities: an approach to statistical analysis and interpretation. 2nd ed. Plymouth, MA: Primer-E.

Connell, J.H., 1978. Diversity in tropical rainforests and coral reefs. Science, 199, 1302–1310.

CSIRO, 1993. Jervis Bay Marine ecological studies final report. CSIRO Division of Fisheries, Jervis Bay, Australian Capital Territory, Australia.

De’ath, G. and Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, 81, 3178–3192.

Džeroski, S. and Drummb, D., 2003. Using regression trees to identify the habitat preference of the sea cucumber (Holothuria leucosplioita) on Rarotonga, Cook Islands. Ecological Modelling, 170, 219–226.

Elith, J., et al., 2006. Novel methods improve prediction of species’ distributions from occurrence data. Ecography, 29, 129–151.

Friedman, J.H., 1999. Stochastic gradient boosting. Technical report. Stanford, CA: Department of Statistics, Stanford University.

Gavrilov, A.N., et al., 2005. Characterization of the seafloor in Australia’s coastal zone using acoustic techniques. Proceedings of the international conference in underwater acoustic measurements: technologies & results, 28 June to 1 July, Heraklion Crete, Greece.

Goff, J.A., Olson, H.C., and Duncan, C.S., 2000. Correlation of side-scan backscatter intensity with grain-size distribution of shelf sediments, New Jersey margin. Geo-Marine Letters, 20, 43–49.

Gogina, M., Glockzin, M., and Zettler, M.L., 2010. Distribution of benthic macrofaunal communities in the western Baltic Sea with regard to near-bottom environmental parameters. 2. Modelling and prediction. Journal of Marine Systems, 80, 75–70.

Grantham, H.S., et al., 2010. Effectiveness of biodiversity surrogates for conservation planning: different measures of effectiveness generate a kaleidoscope of variation. PLoS ONE, 5, e11430, doi:10.1371/journal.pone.0011430.

Gray, J.S., 2002. Species richness of marine soft sediments. Marine Ecology Progress Series, 244, 285–297.

Guisan, A. and Zimmerman, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling, 135, 147–186.

Gutiérrez, J.L., et al., 2002. Mollusks as ecosystem engineers: the role of shell production in aquatic habitats. Oikos, 101, 79–90.

Hasselmann, K., et al., 1973. Measurements of wind-wave growth and swell decay during the Joint North Sea Wave Project (JONSWAP). Ergnungsheft zur Deutschen Hydrographischen Zeitschrift Reihe, 8 (12), 95.

Heap, A.D., et al., 2009. Seabed Environments and Subsurface Geology of the Capel and Faust basins and Gifford Guyot, Eastern Australia – post survey report [online]. Geoscience Australia, Record 2009/22, 166 pp. Available from: http://www.ga.gov.au/image_cache/GA14300.pdf (accessed July 2010).

Hernandez, P.A., et al., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography, 29, 773–785.

Hill, N.A., et al., 2010. Quantifying wave exposure in shallow temperate reef systems: applicability of fetch models for predicting algal biodiversity. Marine Ecology Progress Series, 417, 83–95.

Hochachka, W.M., et al., 2007. Data-mining discovery of pattern and process in ecological systems. Journal of Wildlife Management, 71, 2427–2437.
Holmes, K.W., et al., 2008. Modelling distribution of marine benthos from hydroacoustics and underwater video. *Continental Shelf Research*, 28, 1800–1810.

Huang, Z., Brooke, B.P., and Harris, P.T., 2011a. A new approach to mapping marine benthic habitats using physical environmental data. *Continental Shelf Research*, 31, s4–s16.

Huang, Z., Brooke, B.P., and Li, J., 2011b. Performance of predictive models in marine benthic environments based on predictions of sponge distribution on the Australian continental shelf. *Ecological Informatics*, 6, 205–216.

Huang, Z., et al., 2011c. Predictive modelling of seabed sediment parameters using multibeam acoustic data: a case study on the Carnarvon Shelf, Western Australian. *International Journal of Geographical Information Science*, doi:10.1080/13658816.2011.590139.

Jacoby, C.A., et al., 1995. Deepwater fauna and flora. In: G. Cho, A. Georges, R. Stoutjesdijk, and R. Longmore, eds. *Jervis Bay: a place of cultural, scientific and educational value*. Canberra: An Australian Nature Conservation Agency Publication, 157–162.

Jenness, J.S., 2004. Calculating landscape surface area from digital elevation models. *Wildlife Society Bulletin*, 32, 829–839.

Kloser, R.J., et al., 2001. Remote sensing of seabed types in the Australian South East Fishery; development and application of normal incident acoustic techniques and associated ‘ground truthing’. *Marine and Freshwater Research*, 52, 475–489.

Kostylev, V.E., et al., 2005. The relative importance of habitat complexity and surface area in assessing biodiversity: fractal application on rocky shores. *Ecological Complexity*, 2, 272–286.

Kristensen, E., 2000. Organic matter diagenesis at the oxic/anoxic interface in coastal marine sediments, with emphasis on the role of burrowing animals. *Hydrobiologia*, 426, 1–24.

Leathwick, J.R., et al., 2006. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, 321, 267–281.

McArthur, M.A., et al., 2010. On the use of abiotic surrogates to describe marine benthic biodiversity. *Estuarine, Coastal and Shelf Science*, 88, 21–32.

Mellin, C., et al., 2010. Environmental and spatial predictors of species richness and abundance in coral reef fishes. *Global Ecology and Biogeography*, 19, 212–222.

NSW Marine Parks Authority, 2010. *Seabed habitat mapping in the Solitary Islands Marine Park and Jervis Bay Marine Park* [online], NSW Marine Parks Authority, Hurstville, NSW, Australia. Available from: http://www.mpa.nsw.gov.au (accessed November 2010).

Pearce, J.L., and Boyce, M.S., 2006. Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, 43, 405–412.

Phillips, S.J., Anderson, R.P., and Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 231–259.

Pitcher, C.R., 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. 320pp, CSIRO Marine and Atmospheric Research.

Pittman, S.J., Costa, B.M., and Battista, T.A., 2009. Using Lidar bathymetry and boosted regression trees to predict the diversity and abundance of fish and corals. *Journal of Coastal Research SI*, 53, 27–38.

Radke, L., et al., 2011a. Including biogeochemical factors and a temporal component in benthic habitat maps: influences on infaunal diversity in a temperate embayment. *Marine and Freshwater Research*, 62, 1432–1448.

Radke, L., et al., 2011b. A geochemical characterization of deep-sea floor sediments of the northern Lord Howe Rise. *Deep Sea Research II*, 58, 909–921.

Ris, R.C., Holthuijsen, L.H., and Booij, N., 1999. A third-generation wave model for coastal regions 2. Verification. *Journal of Geophysical Research*, 104, 7667–7681.

Shan, Y., Paull, D., and McKay, R.I., 2006. Machine learning of poorly predictable ecological data. *Ecological Modelling*, 195, 129–138.

Sherrod, P., 2008. *DTREG* [online]. Available from: http://www.dtreg.com (accessed July 2008).

Siwabessy, P.J.W., et al., 2006. Analysis of statistics of backscatter strength from different seafloor habitats. *Acoustics 2006*. Christchurch: Australian Acoustic Association, 507–514.

Snelgrove, P.V.R., 2001. Diversity of marine species. In: J. Steele, S. Thorpe, and K. Turekian, eds. *Encyclopedia of ocean sciences*. Oxford: Academic Press, 748–757.

Tittensor, D.P., et al., 2010. Global patterns and predictors of marine biodiversity across taxa. *Nature*, 466, 1098–1101.
Ward, T.J., et al., 1999. Selecting marine reserves using habitats and species assemblages as surrogates for biological diversity. *Ecological Applications*, 9, 691–698.

Weiss, A.D., 2001. Topographic Position and Landforms Analysis. *ESRI international user conference*, 9–13 July, San Diego, CA.

Whitlach, R.B., 1977. Seasonal changes in the community structure of the macrobenthos inhabiting the intertidal sand and mud flats of Barnstable Harbour, Massachusetts. *Biological Bulletin*, 152, 275–294.

Zacharias, M.A. and Roff, J.C., 2000. A hierarchical ecological approach to conserving marine biodiversity. *Conservation Biology*, 14, 1327–1334.