SCIENCE FOR CONSERVATION 323

Predicting the distribution and relative abundance of fishes on shallow subtidal reefs around New Zealand

Adam N.H. Smith, Clinton A.J. Duffy and John R. Leathwick

newzealand.govt.nz

Department of Conservation
Te Papa Atawhai

Cover: John Dory (*Zeus faber*), near Arid Cove Archway, Rakitu Island, Great Barrier Island. *Photo: Kim Westerskov.*

Science for Conservation is a scientific monograph series presenting research funded by New Zealand Department of Conservation (DOC). Manuscripts are internally and externally peer-reviewed; resulting publications are considered part of the formal international scientific literature.

This report is available from the departmental website in pdf form. Titles are listed in our catalogue on the website, refer www.doc.govt.nz under *Publications*, then *Science & technical*.

ISSN 1177–9241 (web PDF)
ISBN 978–0–478–15000–1 (web PDF)

This report was prepared for publication by the Publishing Team; editing by Helen O'Leary and layout by Lynette Clelland. Publication was approved by the Deputy Director-General, Science and Capability Group, Department of Conservation, Wellington, New Zealand.

Published by Publishing Team, Department of Conservation, PO Box 10420, The Terrace, Wellington 6143, New Zealand.

In the interest of forest conservation, we support paperless electronic publishing.

CONTENTS

Predicting the distribution and relative abundance of fishes on shallow subtidal reefs around New Zealand

Adam N.H. Smith[1,2], Clinton A.J. Duffy[3] and John R. Leathwick[4,5]

[1] National Institute of Water & Atmospheric Research Ltd (NIWA), Private Bag 14901, Hataitai, Wellington 6241, New Zealand

[2] Present address: Massey University Albany, Private Bag 102904, North Shore 0745, Auckland, New Zealand.
Email: anhsmith@gmail.com.

[3] Department of Conservation, Private Bag 68908, Newton, Auckland 1145, New Zealand

[4] National Institute of Water & Atmospheric Research Ltd (NIWA), PO Box 11115, Hamilton 3216, New Zealand

[5] Present address: Department of Conservation, Private Bag 3072, Hamilton 3204, New Zealand.

Abstract

Statistical models of the distribution and abundance of 72 species of rocky reef fishes were developed using boosted regression trees and a set of environmental, geographic and dive-specific variables. The models were used to predict and map the occurrence and relative abundance of the selected species on shallow coastal reefs around New Zealand, including the remote Kermadec and Chatham Islands, at the scale of a 1-km^2 grid. A cross-validation method indicated that the models were able to explain between 8% (*Notoclinops caerulepunctus*) and 86% (*Chromis dispulis*) of the deviance in species abundances, with a mean of 43%. The most widespread species were predicted to be from the family Tripterygiidae, such as *Forsterygion flavonigrum*, *F. malcolmi*, and *F. varium* (predicted at 99%, 99%, and 93% of predicted reef sites, respectively), and also *Caesioperca lepidoptera* (95%) of Serranidae and *Notolabrus fucicola* (91%) of Labridae. The models provided here are a valuable source of information for both fish ecologists and managers. The models identify the environmental variables that are ecologically important for these species, and provide insight into the nature of broad-scale relationships between reef fishes and their environment. In particular, a minimum wintertime sea-surface temperature threshold was suggested for many northern species. Moreover, while the biogeography of New Zealand reef fishes has previously been described by reference to regional spatial scales, this is the first known prediction of the distribution of reef fishes for such a broad geographic extent and fine spatial resolution. The predictions from these models thus provide important, spatially-explicit data for use in the management of coastal biodiversity, particularly in the area of marine spatial planning and the identification of high priority areas for conservation.

Keywords: biogeography, boosted regression trees, species distribution modelling, rocky reef, reef fish, relative abundance

This paper may be cited as:

Smith, A.N.H.; Duffy, C.A.J.; Leathwick, J.R. 2013: Predicting the distribution and relative abundance of fishes on shallow subtidal reefs around New Zealand. *Science for Conservation 323*. Department of Conservation, Wellington. 25 p. + 2 supplements.

1. Introduction

The New Zealand Biodiversity Strategy (DOC & MfE 2000) aims to protect a full range of natural marine habitats and ecosystems in order to effectively conserve New Zealand's indigenous marine biodiversity. This includes the implementation of a representative network of Marine Protected Areas (MPAs) (DOC & MfE 2000). In order to achieve this aim, central and local government agencies with management responsibility for marine ecosystems, and local communities require detailed information on the spatial distribution of marine resources, and biodiversity values of specific areas and habitat types.

Shallow, coastal rocky reefs are focal points for customary, recreational and commercial use, as well as scientific research. It is unsurprising therefore that the majority of marine reserves located around mainland New Zealand are centred upon, or contain, both intertidal and shallow subtidal reef systems (Enderby & Enderby 2006). Fishes form a prominent and taxonomically and ecologically diverse component of these systems (Russell 1983; Paulin & Roberts 1992, 1993; Francis 1996; Clements & Zemke-White 2008). To date, research on distributional patterns of New Zealand reef fishes has been focused at the bioregional scale (Paulin & Roberts 1992, 1993; Francis 1996). Effective marine spatial planning, including development of representative MPA networks, requires information with much finer spatial resolution than this but, as yet, no nationally consistent data sets suitable for this purpose have been published for any New Zealand reef organisms.

Unfortunately, it is often prohibitively costly and time-consuming to conduct biological surveys with sufficient fine scale spatial resolution and broad enough geographic coverage for use in systematic marine spatial planning. However, the relationships between species distributions and environmental variables can be determined using statistical models, allowing predictions to be made for areas where observational data are lacking. A variety of modelling techniques are available for this purpose, including an ensemble approach known as Boosted Regression Trees (BRTs, Friedman 2001; Hastie et al. 2001; Elith et al. 2008). Leathwick et al. (2006a) recently used BRTs to model the distributions of demersal fishes at shelf and upper slope depths throughout New Zealand's Exclusive Economic Zone (EEZ).

We used BRTs to model the distribution and relative abundance of rocky reef fishes around the New Zealand coast line using data obtained from diver surveys, and a suite of environmental, geographic and dive-specific predictor variables. The primary objective of this work was to provide fine-scale, nationally consistent data layers that could be used in conjunction with Geographic Information Systems (GIS) and spatial planning tools to inform the management of coastal areas, specifically the identification of priority areas for the conservation and protection of reef fish biodiversity (Leathwick et al. 2008a, b).

2. Methods

2.1 Datasets

2.1.1 Diver survey of the abundance of reef fishes

Data on the relative abundance of reef fishes were obtained from 467 SCUBA dives made around the coast of New Zealand over an 18 year period from November 1986 to December 2004. The majority of the data used in this study were collected by C. Duffy, with a small number collected by A. Smith. The median length of dive was 46 minutes, and the median maximum depth was 17 m. The locations of the sites are shown in Fig. 1.

At each site, a thorough search for all species of fish was undertaken. The relative abundance of each species of fish observed was recorded on a scale of 0 to 4 (Table 1). This scale broadly represents orders of magnitude of abundance, and is similar to the so-called 'Roving Diver Technique' used by Schmitt & Sullivan (1996), Schmitt et al. (2002) and Semmens et al. (2004).

The original dataset contained 212 species. Pelagic, highly cryptic or primarily soft sediment species were excluded from the analysis. With the exception of *Anampses elegans*, *Aplodactylus etheridgii*, *Epinephelus daemelii*, *Odax cyanoallix*, *Trachypoma macracanthus* and *Zanclistius elevates*, species that were too rare to be effectively modelled (i.e. recorded from less than 20 sites) were also excluded from the analysis. Those rare species that were modelled were included at the request of the Department of Conservation (DOC). The final dataset contained 72 species (Table 2; Appendix 1).

2.1.2 Predictor variables

Fifteen variables were available to the models, each falling into one of three categories: environmental, geographic and dive-specific (Table 3). Some predictor variables were log-transformed to enable easier visualisation of the results (Table 3). Monotonic transformations have no effect on BRT models or their predictions (Friedman & Meulman 2003).

Environmental predictors

Environmental variables were obtained as GIS raster layers. Most were developed as part of the New Zealand Marine Environment Classification (MEC) (Snelder et al. 2004; Snelder et al. 2007). The MEC variable layers of bathymetry, freshwater fraction and orbital velocity were not used because they were considered too inaccurate in shallow (Leathwick, et al. 2004; Snelder et al. 2005; Smith 2006). The latter predicts the velocity of water at the sea bed as induced by swell waves but did not take into account sheltering or refraction by land (Smith 2006). Instead, average fetch, a geographically derived proxy, was used for wave exposure (see below).

Geographic predictors

The two geographic variables used were the shortest distance to land and average fetch. The shortest distance to land was calculated using ArcGIS 9.2 and was used as a proxy to represent the complex influence the land has on processes such as sedimentation, primary productivity and larval dispersal. Average fetch is essentially the average distance to land in all directions, and was used here as a surrogate for wave exposure. It was calculated using the method developed by E. Villouta and R. Pickard (described by Fletcher et al. 2005), where the distance to land was measured along 36 radial lines radiating from a point at 10 degrees intervals. Where land was not encountered the lines were cropped at 10 km. Where Fletcher et al. (2005) used the sum of the distances in each direction, we instead used the average distance. This modified index is equivalent, but we consider it more easily interpreted as the average distance to land. This approach to approximating wave exposure has been extended and validated by Burrows et al. (2008).

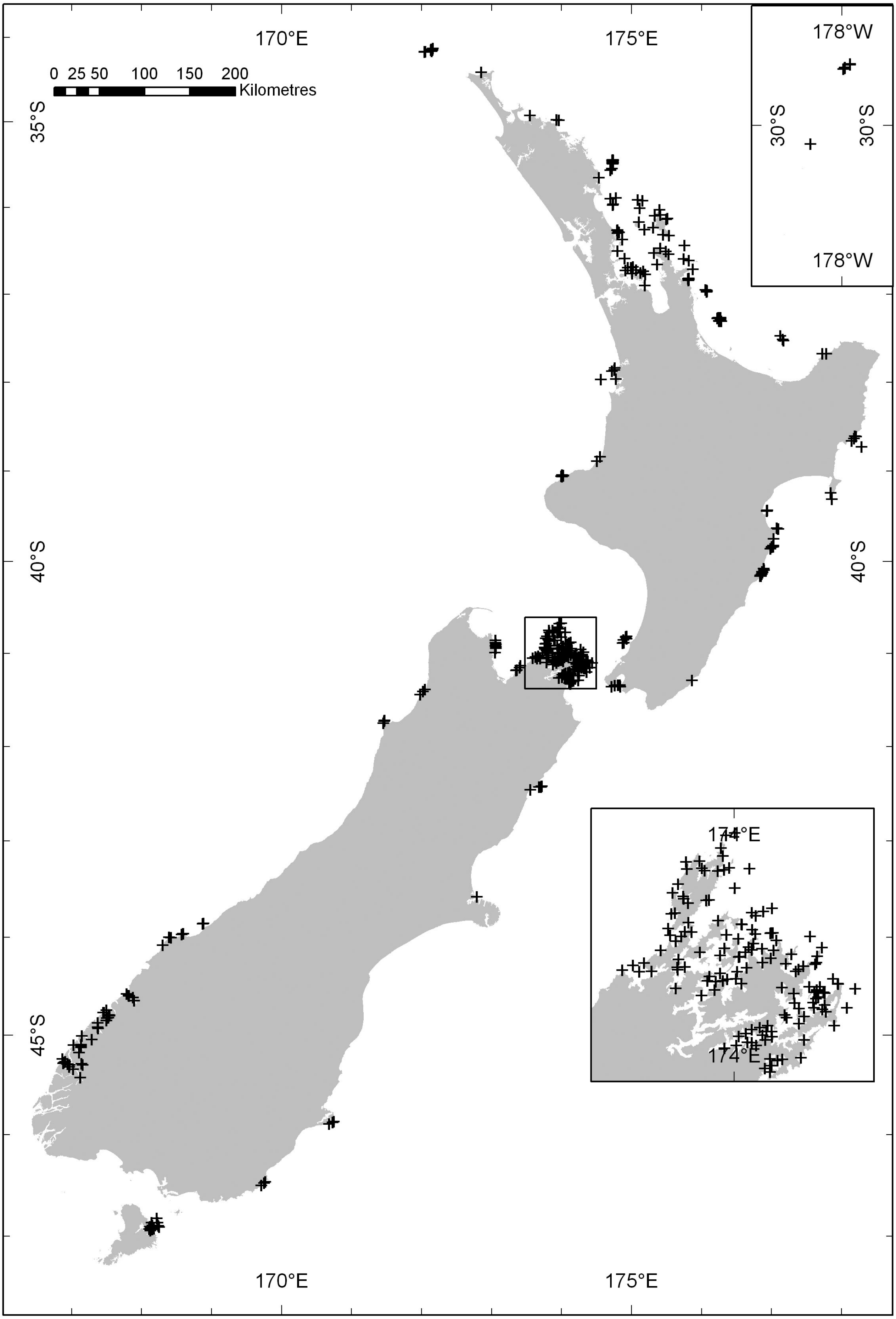

Figure 1. Locations of 467 dive survey sites. The top right and lower right insets show the Kermadec Islands and Marlborough Sounds, respectively. There were no sampling sites at any of the other island groups that are not shown here, including the Chatham Islands.

Table 1. Ordinal scale of relative abundance of fish recorded at the 467 dive survey sites which comprised the biological data used in this study. This scale broadly corresponds to a y' = ln(y + 1) transformation of the number of individuals of that species seen per dive.

VALUE	NAME	NUMBER OF FISH OBSERVED
0	Absent	0
1	Single	1
2	Few	2–10
3	Many	11–100
4	Abundant	>100

Dive-specific predictors

The level of effort and depths of survey dives were not standardised, creating a potential source of bias in the abundance estimates. To control for this, some dive-specific variables were included as predictors in the models and given fixed values for the predictions. These variables were the dive duration, visibility and the minimum and maximum depths surveyed.

The response functions fitted to these variables were forced to be monotonically positive for duration, visibility and maximum depth, and monotonically negative for minimum depth. The assumptions being that poor visibility decreases observed abundance, and that terminating a survey dive some distance below the surface should not increase the observed abundance of a species but may decrease the probability of observing species confined to very shallow water. Allowing these functions to fluctuate freely may allow the model to falsely attribute noise to these dive-specific variables, or falsely attribute variation to them that may be better explained by other variables.

Table 2. Details and performance of the boosted regression tree models for each species. The number of predictor variables and trees relate to the complexity of the models. The deviance explained is the proportion of the total deviance in the abundance of each species that was explained by the model, as evaluated by cross-validation. The proportion of the predictive reef sites at which the species was predicted (i.e. non-zero relative abundance) is an index of the degree to which the species is widespread.

SPECIES	NUMBER OF PREDICTOR VARIABLES	NUMBER OF TREES FITTED	DEVIANCE EXPLAINED (%)	PROPORTION OF PREDICTIVE SITES OCCUPIED (%)
Aldrichetta forsteri	8	820	15	8.2
Amphichaetodon howensis	3	905	58	0.9
Anampses elegans	2	9110	78	0.8
Aplodactylus arctidens	8	1825	24	74.1
Aplodactylus etheridgii	2	5165	79	1.4
Atypichthys latus	4	2845	46	1.6
Bodianus unimaculatus	5	2045	77	19.8
Caesioperca lepidoptera	11	6205	46	95.2
Canthigaster callisterna	3	1295	44	0.8
Caprodon longimanus	3	1965	72	6.8
Centroberyx affinis	4	5870	41	9.4
Cheilodactylus spectabilis	8	2310	44	65.5
Chironemus marmoratus	7	990	26	20.7
Chromis dispilus	5	2495	86	21.9
Conger verreauxi	10	1035	8	35.1
Coris sandageri	2	1545	75	6.3
Decapterus koheru	5	1215	39	11.5
Epinephelus daemelii	2	4505	69	0.6
Forsterygion flavonigrum	9	2585	43	99.0
Forsterygion lapillum	6	1835	51	45.7
Forsterygion malcolmi	6	2170	39	98.5
Forsterygion varium	12	1440	38	93.2
Girella cyanea	2	2705	44	1.2
Girella tricuspidata	6	1080	21	24.7
Grahamina gymnota	8	835	11	19.1

Continued on next page

Table 2 continued

SPECIES	NUMBER OF PREDICTOR VARIABLES	NUMBER OF TREES FITTED	DEVIANCE EXPLAINED (%)	PROPORTION OF PREDICTIVE SITES OCCUPIED (%)
Gymnothorax prasinus	4	945	34	10.9
Helicolenus percoides	8	2640	41	63.6
Hypoplectrodes huntii	9	2330	31	88.2
Hypoplectrodes sp.B	3	1720	67	10.2
Karalepis stewarti	10	1090	16	20.0
Kyphosus sydneyanus	6	690	13	12.8
Latridopsis ciliaris	8	1870	38	46.1
Latris lineata	3	1720	47	33.5
Lotella rhacina	10	1160	12	90.0
Mendosoma lineatum	6	2145	30	8.1
Nemadactylus douglasii	5	1550	64	18.6
Nemadactylus macropterus	6	1335	23	55.4
Notoclinops caerulepunctus	8	700	8	72.1
Notoclinops segmentatus	9	1670	30	73.7
Notoclinops yaldwyni	5	1340	40	56.1
Notolabrus celidotus	10	3420	62	72.4
Notolabrus cinctus	5	1390	43	30.7
Notolabrus fucicola	11	4215	48	91.1
Notolabrus inscriptus	4	2945	38	1.9
Obliquichthys maryannae	7	1755	39	70.0
Odax cyanoallix	3	5360	84	0.2
Odax pullus	12	2110	31	72.6
Optivus elongatus	5	1365	31	40.1
Pagrus auratus	5	1605	67	14.5
Parablennius laticlavius	5	1770	60	14.8
Parapercis colias	7	2575	60	84.0
Paratrachichthys trailli	7	1020	14	37.9
Parika scaber	8	3565	41	85.5
Parma alboscapularis	4	1240	62	11.0
Pempheris adspersa	7	2975	63	20.9
Plagiotremus tapeinosoma	2	905	46	4.4
Pseudocaranx dentex	4	745	22	16.8
Pseudolabrus luculentus	2	1245	83	7.2
Pseudolabrus miles	7	2950	50	87.3
Pseudophycis barbata	5	1340	11	52.3
Ruanoho whero	7	1670	32	82.1
Scorpaena cardinalis	2	2260	74	2.7
Scorpaena papillosus	5	1370	36	68.5
Scorpis lineolatus	3	1355	35	53.6
Scorpis violaceus	4	2010	62	15.5
Seriola lalandi	6	990	33	30.0
Suezichthys aylingi	3	1560	47	2.1
Trachurus novaezelandiae	4	840	12	13.9
Trachypoma macracanthus	2	3565	54	0.9
Upeneichthys lineatus	5	2010	45	53.9
Zanclistius elevatus	2	720	13	3.0
Zeus faber	4	675	11	7.5

Table 3. List of the 15 variables offered to the models, the number of models in which each was chosen (out of *n* = 72 models, i.e. one for each species) and the average contribution of the variable to the models in which it was used.

TYPE	NAME	EXPLANATION	UNITS	NUMBER OF MODELS IN WHICH THE VARIABLE WAS USED	AVERAGE CONTRIBUTION
Environmental	sstwint[a]	Wintertime sea surface temperature	°C	70	33.6
	seabedsal	Salinity at the sea bed	psu	43	18.7
	sstanamp[a]	Annual amplitude of sea surface temperature	°C	32	18.1
	logdisorgm[a, b]	Log of dissolved organic matter	Dimensionless	32	14.9
	logtidalspeed[e]	Log of tidal speed		31	12.0
	logsuspartmat[a, b]	Log of suspended particulate matter	Approx. g/m^3	29	12.2
	sstanom[a]	Sea surface temperature anomaly	°C	29	18.8
	logsstgrad[a]	Log of sea surface temperature gradient	°C/km	25	11.7
	chla2[a]	Concentration of chlorophyll *a*	ppm	22	16.2
Geographic	avefetch[c]	Average fetch	m	48	12.6
	dcoast[d]	Shortest distance to land	m	7	11.8
Dive-specific	dmax	Maximum depth of dive	m	17	12.9
	dur	Duration of dive	min	12	7.6
	dmin	Mimimum depth of dive	m	10	5.9
	vis	Visibility of dive	m	1	5.1

[a] Developed for the MEC (Hadfield et al. 2002; Snelder et al. 2004; Snelder et al. 2007)

[b] Pinkerton & Richardson (2005)

[c] Produced using a program developed by E. Villouta and R. Pickard (Fletcher et al. 2005)

[d] Calculated using ArcGIS

[e] Provided by J. Sturman

2.2 Statistical analysis

2.2.1 Fitting BRT models

Independent models were used to model the abundance of each of the 72 species of reef fish. All statistical analyses were undertaken in R (R Development Core Team 2007) using a package named 'gbm' (Ridgeway 2006) and code developed by Leathwick et al. (2006a, b). The models were built using BRT method. This approach combines many individual regression trees to form a single ensemble model. The regression trees are produced iteratively, gradually improving the overall fit by giving more weight to those sites that are poorly fitted by the previous trees. More complete descriptions of the BRT method can be found in Elith et al. (2008); Friedman (2001); Hastie et al. (2001); Ridgeway (2006). The specifications used in this study were to fit five trees at a time with a learning rate of 0.002 and a tree depth of 5 (see Leathwick et al. 2006a). A Gaussian error distribution was used, as it produced a better overall fit and residual pattern than the alternatives. Because the ordinal scale used in this study to quantify the abundance of fish (Table 1) is roughly logarithmic, this is analogous to using a log-normal approach to modelling abundance, which is common in ecology (Legendre & Legendre 1998).

A stepwise, 10-fold, cross-validation procedure was employed to objectively determine the number of trees to be fitted in each model, thus reducing the risk of over-fitting. This approach divides the dataset into 10 subsets, each withheld in turn while models are fitted to each group of 90% of remaining sites. The holdout deviance is then calculated from the average of the prediction errors of the models to the respective withheld subsets. The final number of trees is given by that which minimises the holdout deviance. Goodness-of-fit statistics were calculated from the cross-validation routine, by taking the mean and standard error of the correlation between the observed and predicted values for the holdout sites. See Hastie et al. (2001) for a more detailed description of the cross-validation method.

2.2.2 Grooming models

The models were originally fitted using all available predictor variables. Although the cross-validation process goes some way to ensure that the models are parsimonious in terms of the number of trees fitted to the data, over-fitting can also occur by including more predictor variables than are necessary. To ameliorate this risk, the global models (those with all predictor variables included) for each species in turn were subjected to a simplification process wherein variables were removed from the models, and then the final models were created by refitting with the reduced variable set. Although the simplification process was essentially subjective, in that it was not done automatically, it was informed by some objective criteria. First, the relative contributions of each variable, in terms of deviance explained, was noted. Second, a procedure was used whereby the lowest contributing variables were sequentially removed from the model, before the model was refitted. The change in deviance explained that resulted from removing a variable was then examined.

2.2.3 Weighting of sites

The geographic placement of the dive surveys was neither random nor representative.

In fact, their placement was highly skewed, with many more sites occurring in areas where the principal collector (C. Duffy) had done intensive surveys (e.g. the Marlborough Sounds and the Poor Knights Islands). To avoid these areas having a disproportionately high influence on the models, sites were given a weighting that reflected the prevalence of sites with similar environmental characteristics. Sites with environmental characteristics that were poorly represented in the samples were weighted higher, and those with environmental characteristics that were over-represented were down-weighted.

To achieve this, a BRT model was used to calculate the probability of a sampling site being present in parts of the environmental space into which predictions are made. For input to this model, the predictive environmental space of interest was represented by 1000 points that were randomly generated from within the predictive domain (produced using Hawth's Tools for ArcGIS; Beyer 2004). Each of the random points was assigned a value of 0 and each sampling site was assigned a value of 1, and this variable was used to produce a binary response variable that was modelled using the environmental variables. The fitted values were transformed (by the log of the inverse, i.e. $\log(1/x)$, where x is the fitted value) and used as the weighting in the predictive models of the abundances of reef fishes. As we were primarily concerned with the environmental representativeness of the sites, only the environmental predictors were used in this model.

The majority of the randomly allocated 'zero' points were located immediately adjacent to the coast. This was done by selecting them according to a spatially-explicit probability distribution that decreased with distance from the coast. Although, for the final models, predictions were made beyond coastal areas area in the BRT models, the coast is where the most accuracy is required because this is where the majority of sampling sites were located. As a result, the models were more strongly weighted towards, and probably more accurate for, coastal areas.

2.2.4 Predictions

A geographic area for which predictions were made (the predictive domain) was delineated. The latitudinal extent of the predictions was from the Kermadec Islands in the north to Stewart Island/Rakiura in the south. Although no surveys were conducted at the Chatham Islands, predictions were made for this remote island group. However, these predictions should be treated with caution as it is known that some mainland species of reef fish have not been recorded there.

The predictive domain was first produced as a 1-km^2 grid, in which a pixel was included if it satisfied at least one of the following conditions: it was within the 50-m depth contour; within 1 km from the shore; or within 1 km of a sample site. This grid was then overlain with a shape file showing the positions of subtidal reefs inferred from navigational charts, and those grid cells that contained no reef were removed. This predictive domain was then converted into a 1-km^2 grid of points. Values for the environmental and geographic variables were extracted for each point, and then the BRT models were used to make predictions for each point according to the environmental and geographical conditions.

For the dive-specific variables, fixed values were assigned to the predictive points. For duration and visibility, the median values from the surveys were assigned, specifically 46 min and 7 m, respectively. With no reliable bathymetric information available (see section 2.1.2—Environmental predictors), arbitrary depths had to be assigned to the predictive domain. The maximum depth was fixed to 30 m for the entire domain. The minimum depth was fixed to 0 m for points adjacent to the coast (i.e. within 1 km) and 10 m for offshore points.

Eight of the predicted species distributions in preliminary modelling showed presences far southwards of their known ranges as given by Francis (2001). These species tended to have a very low number of presences in the dataset. This problem was managed by restricting both the data that were used to produce the model, and the predictions, to northwards of specific latitudes. These species, with their southern latitudinal limits given in parentheses, were *Anampses elegans* (38°S), *Atypichthys latus* (38°S), *Bodianus unimaculatus* (40°S), *Girella tricuspidata* (40°S), *Nemadactylus douglasii* (42°S), *Notolabrus inscriptus* (39°S), *Optivus elongatus* (43°S) and *Trachypoma macracanthus* (40°S).

2.2.5 Scaling the predictions

Two issues with the raw predictions from the models needed to be overcome. Firstly, because a Gaussian error distribution was used and the fact that predictions from the BRT method are the average taken from many models, the output from these models was on a continuous scale, rather than the ordinal one in the input data. Secondly, the results for many species were poorly scaled, so that the predictions at the lower end of the scale were overestimated and those at the upper end were underestimated.

To correct these issues, a second step was used to rescale the predictions to match the original scale of abundance used in the raw data. This involved fitting a single classification tree (using the rpart library in R; see Breiman et al. 1984) to the observed values using the predicted values. The predictions from the entire domain were then rescaled using this model and rounded to one decimal place. Although the rounding to one decimal place meant that the predictions were decimals rather than whole numbers, this method meant that the predicted abundance was unbiased and the original scale of relative abundance was preserved.

2.3 Uncertainty

2.3.1 Bootstrapped confidence intervals

Bootstrapped confidence intervals (Manly 1997) were produced to obtain estimates of certainty for each prediction. This was done by refitting the model to each of 500 bootstrap samples of the original data and making predictions (including scaling, see section 2.2.5) for each species. The 0.025 and 0.975 quantiles were taken for every predicted point to provide 95% confidence intervals. These confidence intervals are not presented in this report because of the space required but are available upon request.

2.3.2 Coverage of the environmental space by samples

The ‘environmental space’ is the multidimensional space conceived when each environmental variable is treated as a dimension. The samples and predictive sites can be projected into this space according to their values on each environmental variable. Some parts of this environmental space will contain many samples and thus be considered well covered by the biological data. However, because of correlations among the variables (i.e. multicollinearity), other parts of the environmental space will not be well covered by samples. Furthermore, there may be areas in the environmental space for which predictions are made, but which are underrepresented by samples. Here, predictions are considered less reliable and should be treated with a higher degree of scepticism than predictive sites in areas that are well covered by samples.

We used a novel approach to quantifying the degree to which the environmental conditions of each predictive site was covered by the samples. The approach was similar to that used in the weighting of sites (see section 2.2.3). A sample of 50 000 random values was taken from the environmental space and assigned a value of 0, indicating that these were ‘false’ sample sites. These were combined with the true samples, to which a value of 1 was assigned. A BRT model was then used to model this new variable that contained 0s for false (random) samples and 1s for true samples, using the Bernoulli error distribution. Predictions using this model then yielded estimates of the probability of a site occurring in each part of the environmental space. Five-hundred trees with an interaction depth of two were used, so that only pair-wise combinations of the environmental variables were regarded. Predictions were then made for the predictive domain using this model, generating values between 0 and 1, according to how well each predictive site was represented by the samples. This may be used as a spatially explicit index of the degree of confidence that can be placed at each cell of the predictive domain, given its environmental conditions.

3. Results

3.1 Model outputs

The primary output of these models is a set of spatially explicit maps that quantify the estimated relative abundance of 72 reef fishes inhabiting shallow coastal waters from Stewart Island/Rakiura north. Predictions were initially made for 52 110 gridded locations, but this was reduced to 9605 when locations that did not contain rocky reef were removed (see section 2.2.4). Presentation of the predictions for each of the 72 species in this document was not possible due to restrictions on space. However, national-scale maps are provided for all modelled species in Supplement 1, and the spatial data underlying these are available on the DOC website associated with this publication (see Appendix 2 for details). The predicted abundance of the marblefish (*Aplodactylus arctidens*) is presented and discussed in section 3.3. This species had an interesting predicted distribution and illustrates how the broader results may be used and interpreted.

The stepwise routine fitted between 675 and 9110 trees to the models, and took between 48 seconds and 46 minutes of computation time (median: 1698 trees and 4 min. respectively). For each species, the deviance in abundance that was explained by the model was evaluated using cross-validated data that were systematically withheld from the modelling process. This is a more robust and conservative method of evaluating goodness-of-fit of a model than using the same data with which the model was trained (Hastie et al. 2001). As assessed by this method, the models were able to explain between 8% (*Notoclinops caerulepunctus*) and 86% (*Chromis dispilis*) of the deviance in species abundances (Table 2), with a mean of 43%.

Six species that were present at fewer than 20 sites were included in this study because they were of particular conservation interest. Surprisingly, the models for four these species were very successful. Models for these four species (*Epinephelus daemelii*, *Aplodactylus etheridgii*, *Odax cyanoallix* and *Anampses elegans*) explained between 69% and 84% of the deviance in abundance, while the remaining two of the rare species (*Trachypoma macracanthus* and *Zanclistius elevatus*) were among the poorest in terms of deviance explained (Table 2).

The predicted percentage of (rocky reef) sites where a species occurs (i.e. predicted abundance was greater than zero) can be used as a measure of the prevalence or 'widespreadness' of a species (Table 2). By this criterion, the most widespread species were from the families Tripterygiidae (*Forsterygion flavonigrum*, *F. malcolmi*, *F. varium*, *Ruanoho whero*), Serranidae (*Caesioperca lepidoptera* and *Hypoplectrodes huntii*), Labridae (*Notolabrus fucicola* and *Pseudolabrus miles*), Monacanthidae (*Parika scaber*), Moridae (*Lotella rhacina*), and Pinguipedidae (*Parapercis colias*), all of which were predicted at greater than 80% of sites. Most of the rarest species were excluded from the analysis, but the least widespread of those that were modelled were *O. cyanoallix* (endemic to the Three Kings Islands), *E. daemelii*, *Anampses elegans*, *Canthigaster callisterna*, *Amphichaetodon howensis* and *T. macracanthus*, all of which were predicted to occur at less than 1% of sites. The distributions of these species are generally restricted to offshore islands in northern New Zealand.

3.2 Influence of the predictor variables

A set of partial dependence plots is provided for all modelled species in Supplement 2. These plots show the marginal effect of each predictor variable on the abundance of the species, after accounting for all the other predictor variables in the model (Elith et al. 2008). Although the partial dependence plots contain a great deal of information about the environmental conditions in which each species is found, there are too many species to discuss them in detail here. However, the overall patterns and importance of the variables across all models are discussed below.

The number of variables selected in the 72 species models ranged from 2 to 12, with a median across species of 5 variables (Table 2). The overall importance of each predictor variable can be quantified by the number of species models in which it was chosen and its average contribution to those models (Table 3). The most consistently important variable for predicting the abundance of species was wintertime sea surface temperature (*sstwint*), which was selected in 70 (of 72) species models. It was dominant in 30 of these models, with an average of 33% contribution in those models in which it was selected (Table 3).

One very consistent pattern noted in the partial dependence plots (see Supplement 2) is that the shape of the response to *sstwint* forms a threshold for many northern species. Below the threshold the response is near zero or negative and above it is positive. These thresholds are likely to represent the minimum temperature requirements for the species. For example, *Chromis dispilus* appears to prefer temperatures above 14°C (Fig. 2). There were only two species—*Forsterygion lapillum* and *Trachurus novaezelandiae*—for which *sstwint* was not selected. The next most important variables were average fetch and salinity at the seabed (Table 3).

The four dive-specific variables were seldom included in the models, mostly due to their poor predictive power. However, maximum depth of dive was the best predictor for *Caesioperca lepidoptera* and *Forsterygion flavonigrum*. This is consistent with these species' preference for deeper reef habitats (Francis 2001).

3.3 Results for *Aplodactylus arctidens*

The results for *Aplodactylus arctidens* (marblefish) are presented in more detail to serve as an example of how to interpret the results for the 72 species. The deviance explained by the model (a measure of the overall success of the model at explaining variation in abundance) was 24% (Table 2). This is much lower than the mean of 43% across species, which suggests that the abundance of *A. arctidens* is less related to the predictor variables than most other species modelled here, and thus less confidence can be placed in the predictions for this species.

The predicted distribution of *A. arctidens* spans a wide area from Manawatawhi/Three Kings Islands in the north to Stewart Island/Rakiura in the south and the Chatham Islands in the east (Fig. 3). This is consistent with the known range of this species (Francis 2001). The highest abundance predicted on the ordinal scale was 2, corresponding to up to 10 individuals seen per dive. This species is predicted to occur at relatively high abundance on several stretches of

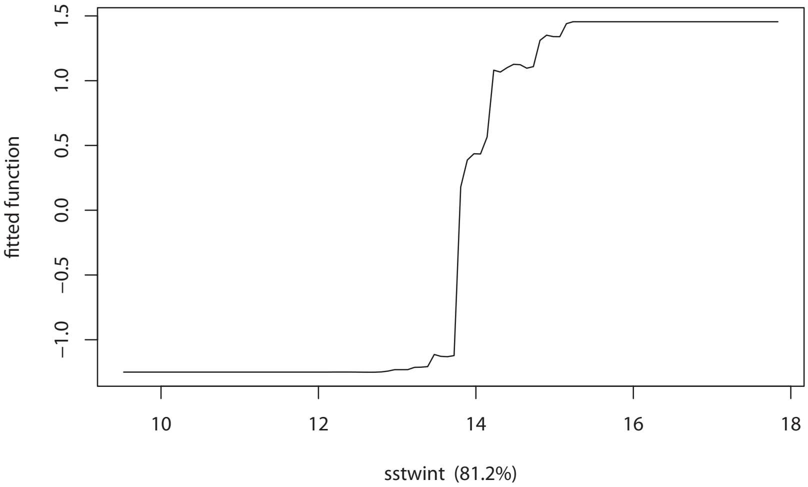

Figure 2. A partial dependence plot of the effect of wintertime sea surface temperature on the abundance of *Chromis dispilus* (two-spot demoiselle). This is shown as an example of a temperature threshold that drove the predictions for many northern species. It appears that *C. dispilus* is rarely found in water colder than 14°C.

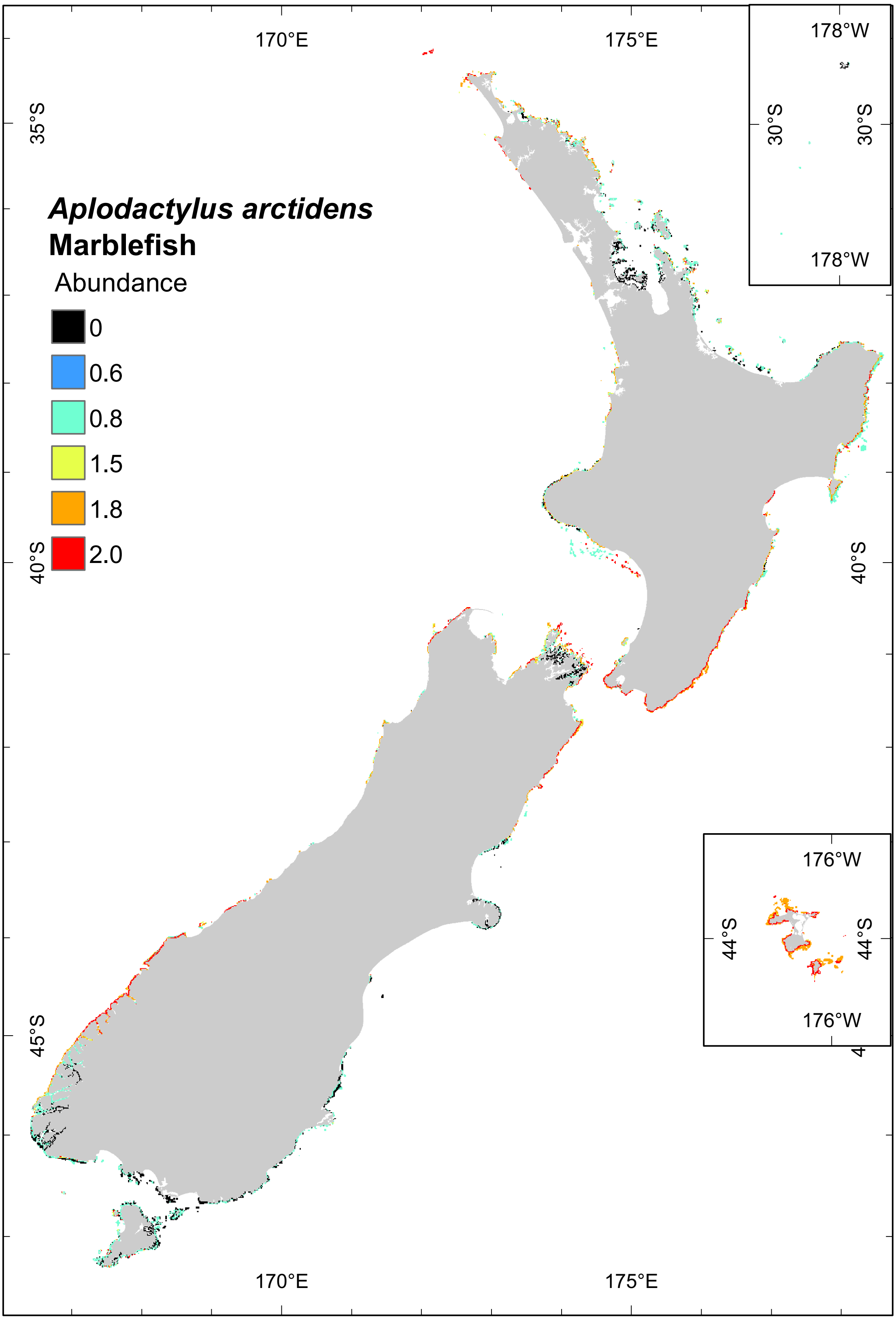

Figure 3. The predicted abundance of *Aplodactylus arctidens* (marblefish) on rocky reefs around New Zealand. Abundance was measured on an ordinal scale according to how many individuals were seen on a dive, shown on Table 1. The predicted values were rounded to one decimal place and can be interpreted on the original scale. For example, a predicted abundance of 0.6 means that the expected score on the original scale is between 0 (absent) and 1 (one individual), so that on average a single individual is seen roughly 60% of the time. Likewise, a predicted abundance of 1.5 means that one would expect a score of between 1 (one individual) and 2 (2–10 individuals) at that site, each occurring roughly 50% of the time. The top right and lower right inserts show the Kermadec and Chatham Islands, respectively.

coastline, particularly along the west and northeast coasts of the South Island and the southern part of the North Island. North of Taranaki and Hawke Bay, its predicted abundance is more variable, being predicted to occur in very few areas of the Bay of Plenty and Hauraki Gulf, but becoming more abundant in the far north and at the Three Kings Islands.

The effect of each predictor variable on the predicted abundance of *A. arctidens* is shown in Fig. 4. The most important variable in the model for this species was average fetch (*avefetch*), with a contribution of 18% relative to the other variables (see Elith et al. 2008). As *avefetch* is an index of exposure, the corresponding graph in Fig. 4 indicates a positive relationship between exposure and the abundance of *A. arctidens*. This is reflected in Fig. 3, with the species predicted to be absent from most sheltered reefs, such as those in the Marlborough Sounds, Hauraki Gulf and inner Fiordland. Almost as important as *avefetch* was wintertime sea surface temperature (*sstwint*), with the plot suggesting a preference for areas of between 10° and 14°C (particularly 12°C). This species showed negative associations with sea surface temperature anomaly (*sstanom*) and sea surface gradient (*logsstgrad*), and less-clear relationships with dissolved organic matter (*logdisorgm*), tidal speed (*logtidalspeed*) and suspended particulate matter (*logsuspartmat*). The strongly decreasing function associated with the minimum depth of the dive (*dmin*) indicates that this species was mostly seen on dives that included reef in less than 10 m of water (i.e. *dmin* < 10 m). This is consistent with observations that this herbivorous species usually occupies and forages in reef habitats shallower than 15 m depth (Russell 1983; Francis 2001; Clements & Zemke-White 2008).

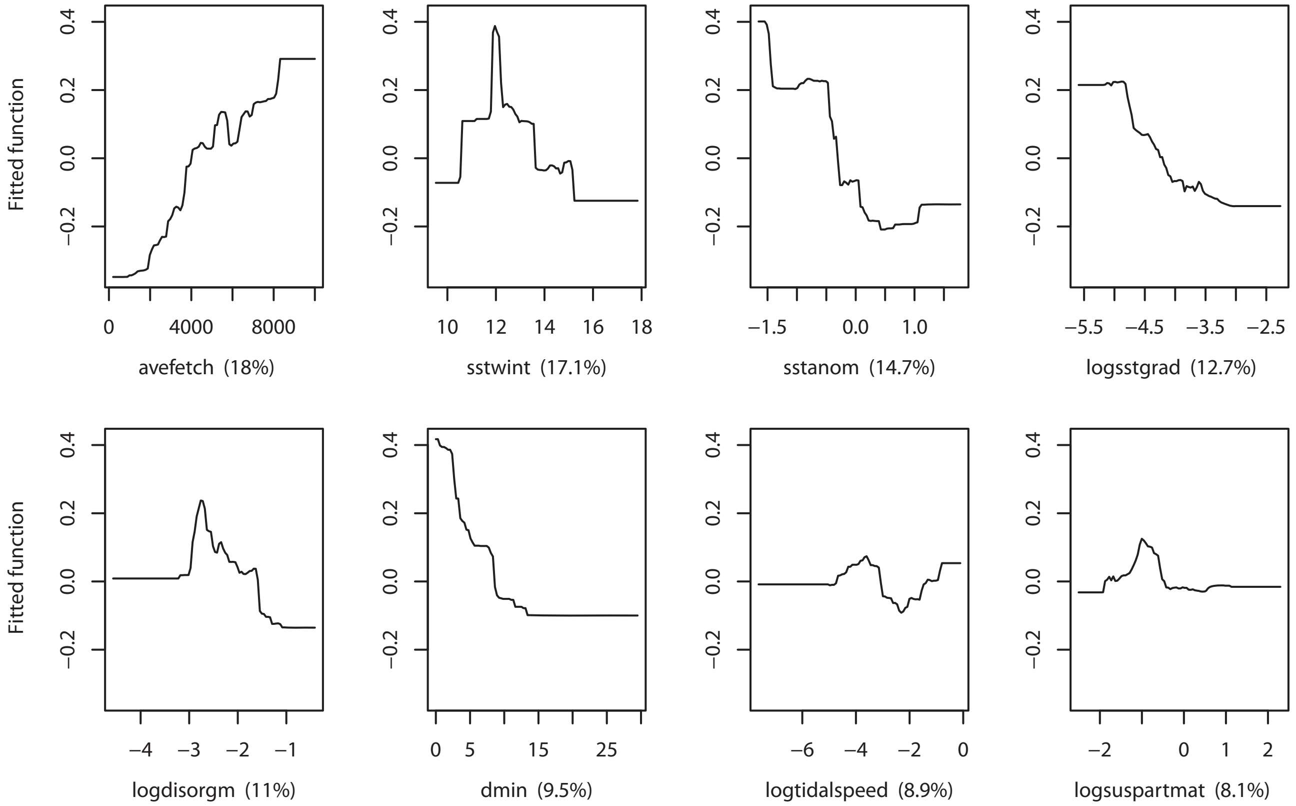

Figure 4. Partial dependence plots showing the relationships between the predictor variables and the abundance of *Aplodactylus arctidens* (marblefish) as observed by scuba on rocky reefs and modelled by a boosted regression tree ensemble model. Each plot represents one predictor variable as indicated on the x-axis label. The percentage values in parentheses indicate the relative contribution of the variable to the model. The shape of the function on the graphs shows the overall effect of the variable on the abundance of *A. arctidens*, shown on a relative scale. The x-axis represents the range of the predictor variable, and the y-axis (the 'fitted function') corresponds to the influence of the variable. A positive value on the y-axis at a given value on the x-axis indicates a positive effect on abundance of this value of the predictor variable, and the converse is true for negative values on the y-axis. For example, the abundance of as *A. arctidens* generally increases with increasing values of avefetch (an index of exposure). See Elith et al. (2008) for more details.

3.4 Geographic variation in the reliability of predictions

The modelling of the coverage of the environmental space by the samples produced a spatially explicit layer indicating the areas where predictions extrapolated beyond the environmental characteristics of the input data (Fig. 5). Poorly covered areas included the inner Hauraki Gulf, the east and south coasts of the South Island, most of Stewart Island/Rakiura (with the exception of Paterson Inlet/Whaka a Te Wera) and the Chatham Islands. Areas where coverage was good included the rest of northeastern North Island, East Cape, Taranaki, southern North Island, the outer Marlborough Sounds and Fiordland.

The relative importance of each variable in this model indicates which variables have significant parts of their range that are not well represented by samples. Chlorophyll *a* (*chla2*) was the most important variable in the model for distinguishing between sampled and non-sampled parts of the environmental space (Fig. 6). This suggests that some part of the range of this variable (as found in the predictive domain) is poorly represented in the samples. The partial dependence plots in Fig. 6 show how well the specific parts of the range of each variable are represented in the samples. In the case of *chla2*, the positive values of the fitted function between zero and one indicate that a high proportion of the samples have values of *chla2* in this range. The negative values greater than two indicate that while these values exist in the predictive domain this part of the range of *chla2* is relatively poorly sampled. For other variables, samples appear to be well represented in areas with high gradients of *logsstgrad*, moderate levels of *sstwint*, *logtidalspeed*, *logsuspartmat*, *logdisorgm*, salinity at the seabed (*seabedsal*), *sstanom*, annual amplitude of sea surface temperature (*sstanamp*), and *avefetch*.

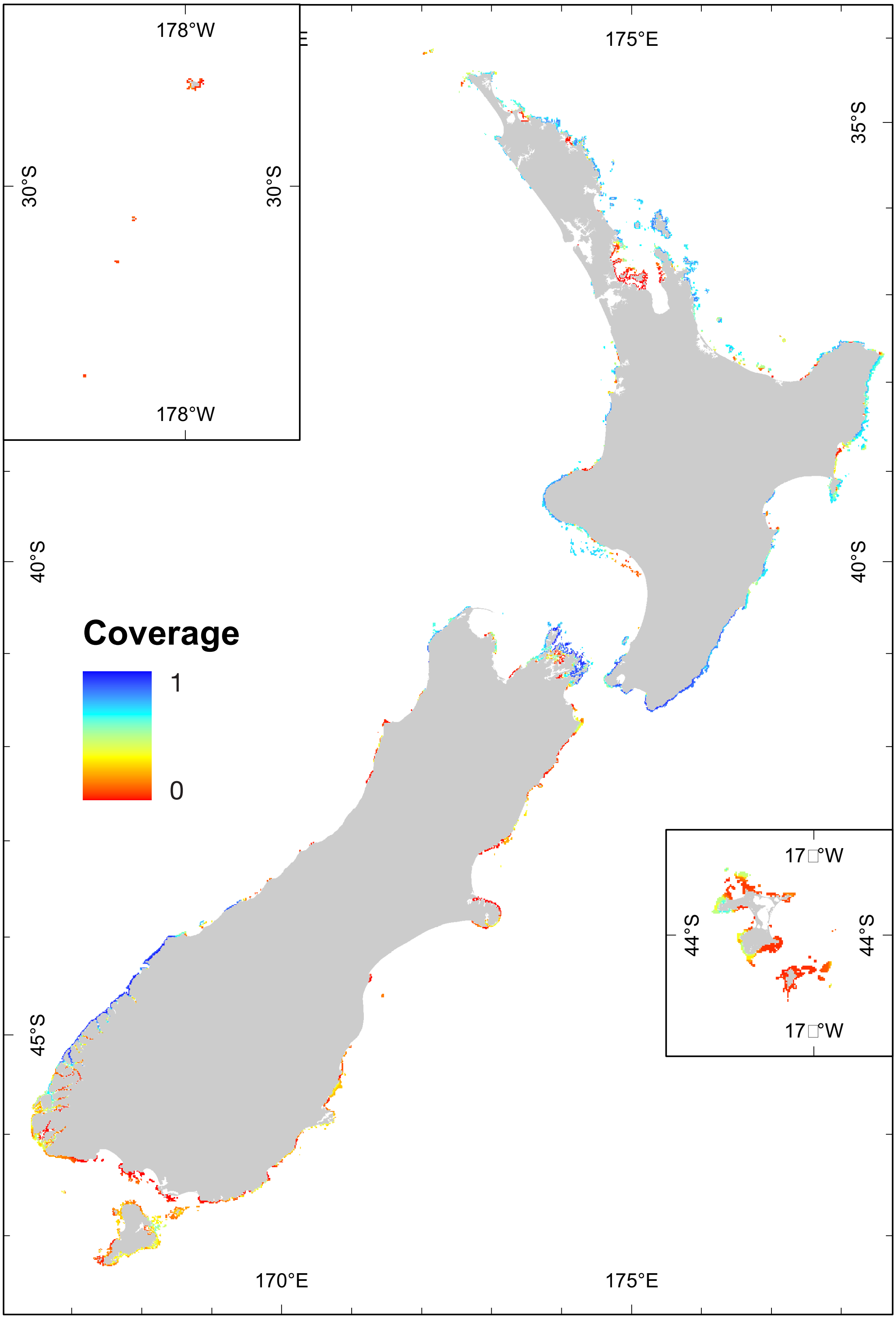

Figure 5. Map of the predictive domain showing how well the environmental conditions of each pixel were represented by the survey sites (see section 2.3.1). The top left and lower right insets show the Kermadec and Chatham Islands, respectively. This measure of environmental coverage provides an index of how much confidence can be placed in the predictions, and is based on the probability of a sample site occurring at each location given the environmental conditions there. It takes potential values between zero (i.e. no samples in the dataset with those environmental conditions) and one (i.e. many samples with those environmental conditions).

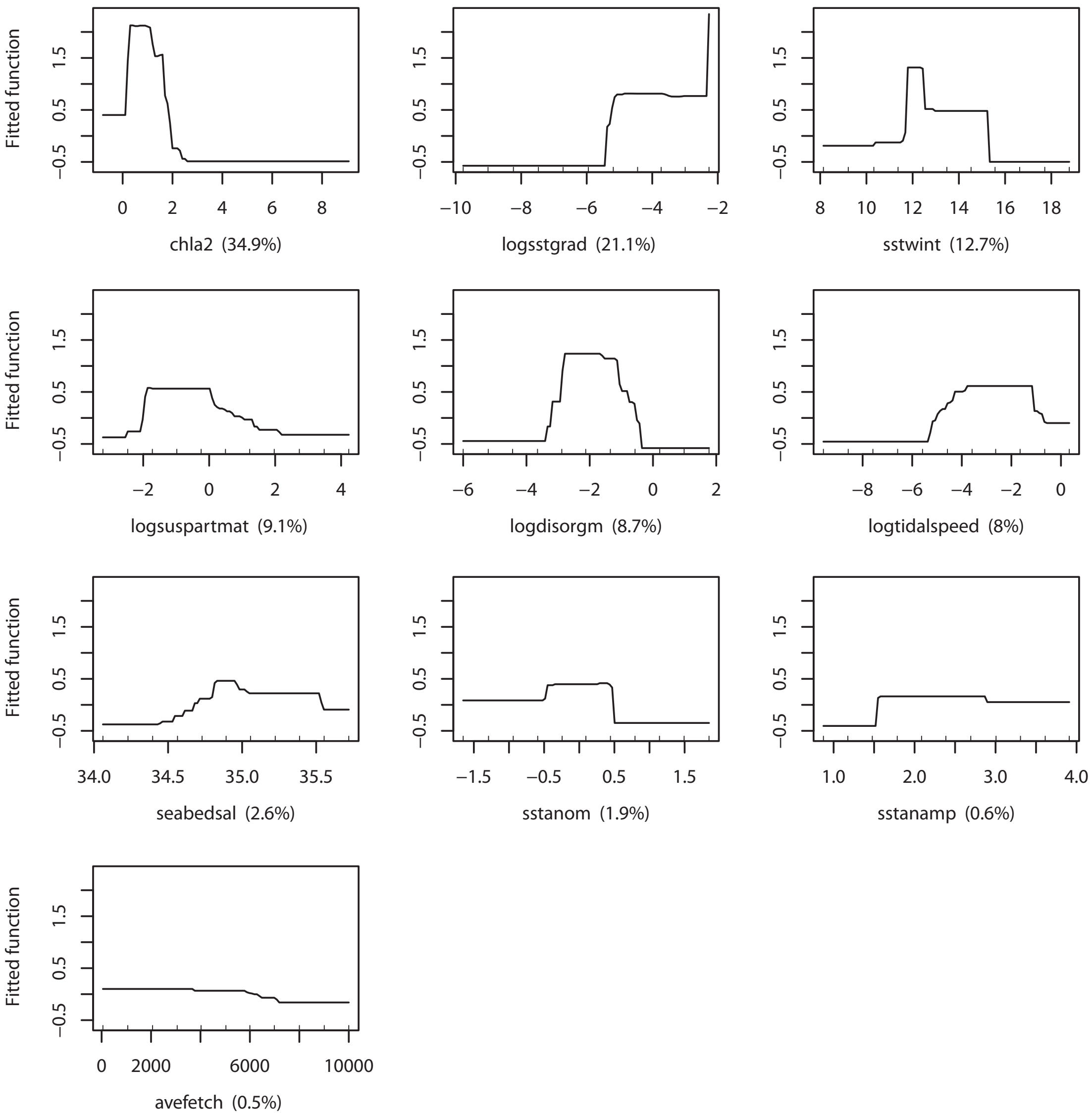

Figure 6. A set of partial dependence plots showing the contributions and fitted functions of the predictor variables in the model of the coverage of the environmental space by the survey sites (see section 2.3.1). Each plot represents a single predictor variable, as indicated in the label of the x-axis (see Table 3 for expansions of the abbreviations). The percentages in parentheses indicate the importance of that variable in the model at discriminating between sampled and unsampled parts of the environmental space. The values on the x-axes represent the range of values found in the predictive domain for each variable. The y-axis allows the graph to show the values of that variable that are well represented by samples (positive values) and the values that are not well represented (values near zero or negative). For example, there are many samples in the dataset that have values between zero and one for chla2 (Chlorophyll *a*), and few that have greater than a value of two.

4. Discussion

4.1 Predicted patterns of abundance

This study predicted patterns of distribution and abundance for 72 reef fishes around New Zealand (detailed in Supplement 1). The spatial resolution of the predictions is a 1-km grid covering shallow (< 50 m) reefs located on the continental shelf around mainland New Zealand, as well as the Kermadec and Chatham Islands. The only shallow areas in New Zealand's exclusive economic zone (EEZ) not included were those around the subantarctic islands, which are known to have a distinctive, relatively depauperate reef fish fauna (Kingsford et al. 1989). A mean (across species) of 43% of the deviance in reef fish abundance was explained by the models according to the cross-validation procedure (see raw values in Table 2). This figure suggests that a reasonable degree of confidence can be placed in the predictions of these models.

4.2 Measures of uncertainty

Quantitative measures of the uncertainty of the predictions permit the end users to know the degree of confidence that can be placed in them. The measure of mean deviance explained by the models for each species, which was based on cross-validation statistics, can be taken as an overall measure of the accuracy of the predictions for that species (Table 2). This is not spatially explicit, and is a single number that may be used to weight the species when combining them in a single analysis, as more confidence can be placed in the predicted distributions of species that have higher accuracy.

Although the cross-validation method is more robust than traditional measures (Hastie et al. 2001), it does not indicate how robust the models may be beyond the range of the biological dataset, chiefly because some areas into which we are predicting may have environmental characteristics that are not represented by the survey sites. We therefore provide a spatially explicit index of the coverage of the environmental conditions of each pixel by the survey dives (Fig. 5). This has relevance across all 72 predicted species distributions.

Finally, bootstrapped confidence intervals for each species at each predicted site were calculated to give within-species, spatially explicit measures of uncertainty. These confidence intervals are available from the authors but are not presented here because of the additional space that would be required to do so. All of these measures of uncertainty can be incorporated into the use of the data in management applications, thanks to recent developments in reserve planning software (Moilanen et al. 2006).

4.3 Limitations and assumptions

It is important to note the limitations of these predictions imposed by the input data and the methods. They are not intended to be a definitive account of where each species can be found. The predictions are subject to the restrictions imposed by the survey method, and do not take into account factors such as diver-positive or diver-negative behaviour by fish (Cole 1994), or variation in detection probabilities (MacKenzie et al. 2002; 2003). Rather, these layers represent predictions of the fish assemblages that might be seen on a typical dive at each of these locations, which can fairly safely be assumed to be correlated with true local abundance.

The samples and thus the predictions were limited to depths that could be dived safely. Although the predictive domain extended to the 50 m isobath, only 10% of the survey dives were deeper than 30 m, so predictions beyond this depth should be used with caution.

The predictions were made primarily on the basis of the suitability of the environment for each species. The models contained no explicit spatial information of any kind and disregarded biogeographic factors such as natural barriers to dispersal. Therefore, caution is needed in using the predictions for offshore island groups (such as the Chatham Islands). On the other hand, if a species does not occur at a location at which it was predicted to occur, this suggests that finer-scale processes or factors other than the environmental and geographic variables used in the model are important in determining the species' distribution.

A spatial resolution of 1 km applies to these predictions, chiefly because this was the resolution of the environmental variables provided by the MEC. Although this is a finer scale than that of any known published work on the distributions of New Zealand reef fishes, variation in abundances will occur at smaller spatial scales than this. This is due to variations in habitats between reefs that may occur within a pixel, or even within a single contiguous reef. The predictions are for known and inferred reefs but they assume that suitable habitat for each species is available on each of these reefs. Note, also, that the predictions were made for dives of 0 to 30 m depth at sites within 1 km of shore and 10 to 30 m for offshore sites. Variation in depth and its effects on species abundance obviously occurs at a far finer spatial scale than the predictions of this study. Therefore, predictions for species that are largely driven by depth, and those for species that have tight depth ranges, should also be treated with caution. For example, *F. flavonigrum* showed a strong preference for deeper water, but not all sites in the predictive domain may contain reef at the depths that this species is found.

Finally, no explicit spatial information on bathymetry or wave exposure was available at the scale and accuracy required by this study. For the latter, a geographical calculation of fetch was used instead, and this geographic approach has been shown to be a good proxy for wave exposure (Burrows et al. 2008). However, the inclusion of accurate and explicit measures for these factors would probably improve the accuracy of the models.

4.4 Rare species

Interestingly, four of the rare species which were included in this study because their distributions were of particular interest to DOC, were among the most successfully modelled in terms of the deviance in abundance that was explained by their respective models. One of these species (*Odax cyanoallix*) is largely endemic to Manawatawhi/Three Kings Islands (Choat & Ayling 1987, Francis 1996). Water surrounding this island group has a very low *sstanom* caused by local upwelling of cold water (Stanton 1973). This enabled the model to easily delineate the distribution of this species based almost entirely on this variable. The remaining rare species are all considered subtropical (Francis 1996), meaning that they require relatively high temperatures. This can result in good model performance, provided there is sufficient data to successfully establish the temperature threshold and that the species is fairly consistently present at locations above this threshold (e.g. *Chromis dispilus*, Fig. 2). This appears to be the case for *Aplodactylus etheridgii*, *Epinephelus daemelii* and *Anampses elegans*, which were very successfully modelled, but not the case for *Trachypoma macracanthus* and *Zanclistius elevatus*, for which the models performed poorly.

4.5 Utility and future research

To date, only broad-scale information on the distributions of reef fishes has been published (Paulin & Roberts 1992, 1993; Francis 1996). This study presents the first attempt at modelling the distribution and relative abundance of reef fishes at such a fine spatial resolution and broad geographic extent. It follows that of Leathwick et al. (2006a) who first used this approach with marine species in New Zealand to predict the distribution and abundance of demersal fishes across the entire EEZ.

The improved predictions and understanding of the spatial patterns of reef fishes resulting from this work have a variety of applications. Their primary purpose is to inform the management of New Zealand's coastal biodiversity and, in particular, the planning of MPAs. The predictions can support this process by revealing parts of the inner shelf likely to support diverse or unique biological assemblages (e.g. Beaumont et al. 2008, 2010) and, when used in conjunction with reserve planning software (Moilanen 2007; Moilanen et al. 2005), aid the design of MPA networks that effectively protect a representative range of species (Leathwick et al. 2008a, b).

In addition to the management applications, these models produced abundant information on the relationships between the distribution and abundance of reef fishes and several environmental and geographic variables. This information includes both the relative importance of the different variables for each species and the qualitative nature of their relationships. For example, sea surface temperature often showed a clear step function, implying minimum thresholds of tolerance for many species (e.g. Fig. 2). While conclusive statements about mechanistic processes cannot be made on the basis of this work, ecological hypotheses may be derived from the outputs of the models (Supplement 2) and then tested in future research.

5. Acknowledgements

This project was funded by a contract from DOC to the National Institute for Water and Atmospheric Research (SIN 3994; NIWA project DOC08304). The authors thank David Plew and James Sturman (NIWA) for the GIS layer of tidal speed, Ross Pickard and Eduardo Villouta (DOC) for providing the fetch analysis program and assistance with it. We also thank many individuals from NIWA who helped with this project, namely Malcolm Francis for comments on preliminary predictions, James Sturman and Arne Pallentin for assistance with GIS and Alison MacDiarmid and Don Robertson for providing comments on the report. Finally, we thank Helen O'Leary, Eduardo Villouta and an anonymous reviewer for their comments and revisions of this manuscript.

6. References

DOC (Department of Conservation); MfE (Ministry for the Environment) 2000: The New Zealand biodiversity strategy—our chance to turn the tide. Department of Conservation and Ministry for the Environment, Wellington. 146 p.

Beaumont J.; Oliver, M.; MacDiarmid, A. 2008: Mapping the values of New Zealand's coastal waters. 1. Environmental values. Biosecurity New Zealand Technical Paper No. 2008/16. 89 p.

Beaumont, J.; D'Archino, R.; MacDiarmid, A. 2010: Mapping the values of New Zealand's coastal waters. A meta-analysis of environmental values. Biosecurity New Zealand Techinical Paper No. 2010/08. 70 p.

Beyer, H.L. 2004: Hawth's Analysis Tools for ArcGIS. Available at www.spatialecology.com/htools

Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. 1984: Classification and regression trees. Wadsworth & Brooks, Belmont, California. 358 p.

Burrows, M.T.; Harvey, R.; Robb, L. 2008: Wave exposure indices from digital coastlines and the prediction of rocky shore community structure. *Marine Ecology Progress Series 353*: 1–12.

Choat, J.; Ayling, A. 1987: The relationship between habitat structure and fish faunas on New Zealand reefs. *Journal of Experimental Marine Biology and Ecology 110*: 257–284.

Clements, K.D.; Zemke-White, W.L. 2008: Diet of subtropical herbivorous fishes in northeastern New Zealand. *New Zealand Journal of Marine and Freshwater Research 42*: 47–55.

Cole, R.G. 1994: Abundance, size structure, and diver-oriented behaviour of three large benthic carnivorous fishes in a marine reserve in northeastern New Zealand. *Biological Conservation 70*: 93–99.

Dey, K.; Weatherhead, M. 2005: User Manual for the New Zealand Marine Environment Classification. Unpublished report to the Ministry for the Environment. National Institute of Water and Atmospheric Research Ltd, Hamilton. Available at www.niwa.co.nz/our-services/databases/mec.

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

Enderby, J.; Enderby, T. 2006: A guide to New Zealand's marine reserves. New Holland Publishers, Auckland. 176 p.

Fletcher, D.; MacKenzie, D.; Villouta, E. 2005: Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics 12*: 45–54.

Francis, M.P. 1996: Geographic distribution of marine reef fishes in the New Zealand region. *New Zealand Journal of Marine and Freshwater Research 30*: 35–55.

Francis, M.P. 2001: Coastal fishes of New Zealand: an identification guide. Reed Publishing, Auckland, New Zealand. 103 p.

Friedman, J.H. 2001: Greedy function approximation: a gradient boosting machine. *Annals of Statistics 29*: 1189–1232.

Friedman, J.H.; Meulman, J.J. 2003: Multiple additive regression trees with application in epidemiology. *Statistics in Medicine 22*: 1365–1381.

Hadfield, M.G.; Uddstrom, M.J.; Goring, D.; Gorman, R.M.; Wild, M.; Stephens, S.; Shankar, U.; Niven, K.; Snelder, T.H. 2002: Physical variables for the New Zealand Marine Environment Classification System: development and description of data layers. Unpublished report no. CHC2002-043. National Institute of Water & Atmospheric Research Ltd, Christchurch.

Hastie, T.; Tibshirani, R.; Friedman, J. 2001: The elements of statistical learning: data mining, inference, and prediction. Springer, New York. 533 p.

Kingsford, M.J.; Schiel, D.R.; Battershill, C.N. 1989: Distribution and abundance of fish in a rocky reef environment at the subantarctic Auckland Islands, New Zealand. *Polar Biology 9*: 179–186.

Leathwick, J.R.; Image, K.; Snelder, T.; Weatherhead, M.; Wild, M. 2004: Definition and test of Marine Environment Classifications of New Zealand's Exclusive Economic Zone and the Hauraki Gulf. Unpublished report no. CHC2004-085, National Institute of Water and Atmospheric Research, Christchurch.

Leathwick, J.R.; Elith, J.; Francis, M.P.; Hastie, T.; Taylor, P. 2006a: 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.

Leathwick, J.R.; Francis, M.P.; Julian, K. 2006b: Development of a demersal fish community map for New Zealand's Exclusive Economic Zone. Unpublished report no. HAM2006-062. National Institute of Water and Atmospheric Research, Hamilton.

Leathwick, J.R.; Julian, K.; Francis, M.P. 2006c: Exploration of the use of reserve planning software to identify potential Marine Protected Areas in New Zealand's Exclusive Economic Zone. Unpublished report no. HAM2006-064. National Institute of Water and Atmospheric Research, Hamilton.

Leathwick, J.R.; Julian, K.; Smith, A. 2008a: Use of reserve planning software to identify priority sites for protection in New Zealand's inshore waters. Unpublished report to the Department of Conservation by the National Institute of Water and Atmospheric Research, Hamilton.

Leathwick, J.; Moilanen, A.; Francis, M.; Elith, J.; Taylor, P.; Julian, K.; Hastie, T.; Duffy, C. 2008b: Novel methods for the design and evaluation of marine protected areas in offshore waters. *Conservation Letters 1*: 91–102.

Legendre, P.; Legendre, L. 1998: Numerical ecology. Elsevier Science, Amsterdam. 851 p.

MacKenzie, D.I.; Nichols, J.D.; Hines, J.E.; Knutson, M.G.; Franklin, A.B. 2003: Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology 84*: 2200–2207.

MacKenzie, D.I.; Nichols, J.D.; Lachman, G.B.; Droege, S.; Royle, J.A.; Langtimm, C.A. 2002: Estimating site occupancy rates when detection probabilities are less than one. *Ecology 83*: 2248–2255.

Manly, B.F.J. 1997: Randomization, bootstrap and Monte Carlo methods in biology. Chapman & Hall, London. 455 p.

Moilanen, A. 2007: Landscape Zonation, benefit functions and target-based planning: unifying reserve selection strategies. *Biological Conservation 134*: 571–579.

Moilanen, A.; Franco, A.M.A.; Early, R.I.; Fox, R.; Wintle, B.; Thomas, C.D. 2005: Prioritizing multiple-use landscapes for conservation: methods for large multi-species planning problems. *Proceedings of the Royal Society B: Biological Sciences 272*: 1885–1891.

Moilanen, A.; Wintle, B.A.; Elith, J.; Burgman, M. 2006: Uncertainty analysis for regional-scale reserve selection. *Conservation Biology 20*: 1688–1697.

Paulin, C.D.; Roberts, C.D. 1992: The rockpool fishes of New Zealand. Museum of New Zealand Te Papa Tongarewa, Wellington, New Zealand. 77 p.

Paulin, C.D.; Roberts, C.D. 1993: Biogeography of New Zealand rockpool fishes. Pp. 191–199 in Battershill, C.N.; Schiel, D.R.; Jones, G.P.; Creese, R.G.; MacDiarmid, A.B. (Eds): Proceedings of the second international temperate reef symposium, 7–10 January 1992, Auckland, New Zealand.

Pinkerton, M.H.; Richardson, K.R. 2005: Case 2 climatology of New Zealand: final report. Unpublished report no. WLG2005-049. National Institute of Water and Atmospheric Research, Wellington. 16 p.

R Development Core Team 2007: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Ridgeway, G. 2006: Generalized boosted models: a guide to the gbm package. R Foundation for Statistical Computing, Vienna.

Russell, B. 1983: The food and feeding habits of rocky reef fish of north-eastern New Zealand. *New Zealand Journal of Marine and Freshwater Research 17*: 121–145.

Schmitt, E.; Sluka, R.; Sullivan-Sealey, K. 2002: Evaluating the use of roving diver and transect surveys to assess the coral reef fish assemblage off southeastern Hispaniola. *Coral Reefs 21*: 216–223.

Schmitt, E.F.; Sullivan, K.M. 1996: Analysis of a volunteer method for collecting fish presence and abundance data in the Florida Keys. *Bulletin of Marine Science 59*: 404–416.

Semmens, B.X.; Buhle, E.R.; Salomon, A.K.; Pattengill-Semmens, C.V. 2004: A hotspot of non-native marine fishes: evidence for the aquarium trade as an invasion pathway. *Marine Ecology Progress Series 266*: 239–244.

Smith, A.N.H. 2006: Evaluation of the New Zealand Marine Environment Classification for shallow coastal rocky reef fish communities. Unpublished MSc thesis. University of Auckland, Auckland. 133 p.

Snelder, T.H.; Leathwick, J.R.; Dey, K.L.; Rowden, A.A.; Weatherhead, M.A.; Fenwick, G.D.; Francis, M.P.; Gorman, R.M.; Grieve, J.M.; Hadfield, M.G.; Hewitt, J.E.; Richardson, K.M.; Uddstrom, M.J.; Zeldis, J.R. 2007: Development of an ecological marine classification in the New Zealand region. *Environmental Management 39*: 12–29.

Snelder, T.; Leathwick, J.R.; Dey, K.L.; Weatherhead, M.A.; Fenwick, G.D.; Francis, M.P.; Gorman, R.; Grieve, J.M.; Hadfield, M.G.; Hewitt, J.E.; Hume, T.M.; Richardson, K.M.; Rowden, A.A.; Uddstrom, M.J.; Wild, M.; Zeldis, J.R. 2005: The New Zealand Marine Environment Classification. Ministry for the Environment, Wellington. 70 p.

Snelder, T.; Leathwick, J.R.; Image, K.; Weatherhead, M.A.; Wild, M. 2004: The New Zealand Marine Environment Classification. Unpublished report no. CHC2004-071, National Institute of Water & Atmospheric Research Ltd, Christchurch. 85 p.

Stanton, B.R. 1973: Hydrological investigations around northern New Zealand. *New Zealand Journal of Marine and Freshwater Research 7*: 85.

Appendix 1

List of reef fish species modelled

The following table lists the species of reef fish for which the distributions and relative abundance were modelled in this study. The codes were used in naming files in supplements 1 & 2.

CODE	FAMILY	BINOMIAL	COMMON NAME	NUMBER OF DIVES IN WHICH THE SPECIES WAS OBSERVED (OUT OF 467)
Apl.arc	Aplodactylidae	*Aplodactylus arctidens*	Marblefish	191
Apl.eth	Aplodactylidae	*Aplodactylus etheridgii*	Notch-head marblefish	16
Cen.aff	Berycidae	*Centroberyx affinis*	Golden snapper	29
Par.lat	Blenniidae	*Parablennius laticlavius*	Crested blenny	97
Pla.tap	Blenniidae	*Plagiotremus tapeinosoma*	Mimic blenny	33
Dec.koh	Carangidae	*Decapterus koheru*	Koheru	65
Pse.den	Carangidae	*Pseudocaranx dentex*	Trevally	65
Ser.lal	Carangidae	*Seriola lalandi*	Kingfish	95
Tra.sp.	Carangidae	*Trachurus novaezelandiae*	Jack mackerel	47
Amp.how	Chaetodontidae	*Amphichaetodon howensis*	Lord Howe coralfish	27
Che.spe	Cheilodactylidae	*Cheilodactylus spectabilis*	Red moki	212
Nem.dou	Cheilodactylidae	*Nemadactylus douglasii*	Porae	84
Nem.mac	Cheilodactylidae	*Nemadactylus macropterus*	Tarakihi	132
Chi.mar	Chironemidae	*Chironemus marmoratus*	Hiwihiwi	69
Con.ver	Congridae	*Conger verreauxi*	Common conger eel	85
Aty.lat	Kyphosidae	*Atypichthys latus*	Mado	22
Gir.cya	Kyphosidae	*Girella cyanea*	Bluefish	22
Gir.tri	Kyphosidae	*Girella tricuspidata*	Parore	26
Kyp.syd	Kyphosidae	*Kyphosus sydneyanus*	Silver drummer	23
Sco.lin	Kyphosidae	*Scorpis lineolatus*	Sweep	182
Sco.vio	Kyphosidae	*Scorpis violaceus*	Blue maomao	110
Ana.ele	Labridae	*Anampses elegans*	Elegant wrasse	12
Bod.vul	Labridae	*Bodianus unimaculatus*	Red pigfish	69
Cor.san	Labridae	*Coris sandageri*	Sandager's wrasse	74
Not.cel	Labridae	*Notolabrus celidotus*	Spotty	349
Not.cin	Labridae	*Notolabrus cinctus*	Girdled wrasse	45
Not.fuc	Labridae	*Notolabrus fucicola*	Banded wrasse	340
Not.ins	Labridae	*Notolabrus inscriptus*	Green wrasse	25
Pse.luc	Labridae	*Pseudolabrus luculentus*	Orange wrasse	61
Pse.mil	Labridae	*Pseudolabrus miles*	Scarlet wrasse	283
Sue.ayl	Labridae	*Suezichthys aylingi*	Crimson cleanerfish	46
Lat.cil	Latridae	*Latridopsis ciliaris*	Blue moki	148
Lat.lin	Latridae	*Latris lineata*	Trumpeter	21
Men.lin	Latridae	*Mendosoma lineatum*	Telescopefish	32
Par.sca	Monacanthidae	*Parika scaber*	Leatherjacket	244
Lot.rha	Moridae	*Lotella rhacina*	Rock cod	144
Pse.bar	Moridae	*Pseudophycis barbata*	Southern bastard cod	28
Ald.for	Mugilidae	*Aldrichetta forsteri*	Yellow-eyed mullet	55
Upe.lin	Mullidae	*Upeneichthys lineatus*	Goatfish	154
Gym.pra	Muraenidae	*Gymnothorax prasinus*	Yellow moray	44
Oda.cya	Odacidae	*Odax cyanoallix*	Blue-finned butterfish	15
Oda.pul	Odacidae	*Odax pullus*	Butterfish	157

Continued on next page

Appendix 1 continued

CODE	FAMILY	BINOMIAL	COMMON NAME	NUMBER OF DIVES IN WHICH THE SPECIES WAS OBSERVED (OUT OF 467)
Pem.ads	Pempheridae	*Pempheris adspersa*	Bigeye	83
Zan.ele	Pentacerotidae	*Zanclistius elevatus*	Long-finned boarfish	13
Par.col	Pinguipedidae	*Parapercis colias*	Blue cod	275
Chr.dis	Pomacentridae	*Chromis dispilus*	Demoiselle	133
Par.alb	Pomacentridae	*Parma alboscapularis*	Black angelfish	69
Hel.per	Scorpaenidae	*Helicolenus percoides*	Sea perch	63
Sco.car	Scorpaenidae	*Scorpaena cardinalis*	Northern scorpionfish	44
Sco.pap	Scorpaenidae	*Scorpaena papillosus*	Dwarf scorpionfish	154
Cae.lep	Serranidae	*Caesioperca lepidoptera*	Butterfly perch	216
Cap.lon	Serranidae	*Caprodon longimanus*	Pink maomao	60
Epi.dae	Serranidae	*Epinephelus daemelii*	Spotted black grouper	19
Hyp.hun	Serranidae	*Hypoplectrodes huntii*	Red-banded perch	80
Hyp.spB	Serranidae	*Hypoplectrodes* sp.B	Half-banded perch B	57
Tra.mac	Serranidae	*Trachypoma macracanthus*	Toadstool grouper	14
Pag.aur	Sparidae	*Pagrus auratus*	Snapper	85
Can.cal	Tetraodontidae	*Canthigaster callisterna*	Clown toado	34
Opt.elo	Trachichthyidae	*Optivus elongatus*	Slender roughy	137
Par.tra	Trachichthyidae	*Paratrachichthys trailli*	Common roughy	46
For.lap	Tripterygiidae	*Forsterygion lapillum*	Common triplefin	203
For.mal	Tripterygiidae	*Forsterygion malcolmi*	Banded triplefin	284
For.var	Tripterygiidae	*Forsterygion varium*	Variable triplefin	343
Fst.fla	Tripterygiidae	*Forsterygion flavonigrum*	Yellow-black triplefin	258
Gra.gym	Tripterygiidae	*Grahamina gymnota*	Robust triplefin	23
Kar.ste	Tripterygiidae	*Karalepis stewarti*	Scaly-headed triplefin	82
Not.cae	Tripterygiidae	*Notoclinops caerulepunctus*	Blue dot triplefin	67
Not.seg	Tripterygiidae	*Notoclinops segmentatus*	Blue-eyed triplefin	269
Not.yal	Tripterygiidae	*Notoclinops yaldwyni*	Yaldwyn's triplefin	86
Obl.mar	Tripterygiidae	*Obliquichthys maryannae*	Oblique-swimming triplefin	251
Rua.whe	Tripterygiidae	*Ruanoho whero*	Spectacled triplefin	250
Zeu.fab	Zeidae	*Zeus faber*	John dory	27

Appendix 2

Online data files

Spatial data from this project are available online at http://www.doc.govt.nz/publications/science-and-technical/products/series/science-for-conservation/ in four *.zip files associated with this publication (*Science for Conservation 323*):

EntireDomain_ASCII.zip

EntireDomain_Esri.zip

ReefOnly_ASCII.zip

ReefOnly_Esri.zip

These files are named according to the predictive domain for which predictions are made (i.e. the EntireDomain, or a subset containing the locations of ReefsOnly), and the format in which the data are provided (ASCII indicates the files are in standard *.asc grid format, and Esri indicates *.rrd raster format for ArcGIS).

Each *.zip file contains three folders. Abundance contains the estimated abundance for each of the 72 species of fish; CoverageEnvrSpace contains the estimated coverage of the environmental space (see Section 2.3.2), and SpeciesRichness contains estimated species richness at each site based on separate BRT model of the number of species seen on each dive. Lower and upper 95% confidence intervals are provided in folders named lowerCI and upperCI for the ReefOnly domain. These were not calculated for the EntireDomain because of the large computation time it would have required.

The coordinate reference system of the spatial data provided here is the same as the input variables from the Marine Environment Classification (Dey & Weatherhead 2005), specifically

Projection = Mercator

Central Meridian = 100

Standard Parallel = –46

False Easting = 0

False Northing = 0

Spheroid/Datum = Clarke 1866

Continue to supplement 1
Continue to supplement 2