Freshwater Biology (2005) **50,** 2034–2052 doi:10.1111/j.1365-2427.2005.01448.x

Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish

J. R. LEATHWICK,* D. ROWE,* J. RICHARDSON,* J. ELITH† AND T. HASTIE‡
*National Institute of Water and Atmospheric Research, Hamilton, New Zealand
†School of Botany, The University of Melbourne, Parkville, Victoria, Australia
‡Department of Statistics, Stanford University, CA, U.S.A.

SUMMARY

1. Relationships between probabilities of occurrence for fifteen diadromous fish species and environmental variables characterising their habitat in fluvial waters were explored using an extensive collection of distributional data from New Zealand rivers and streams. Environmental predictors were chosen for their likely functional relevance, and included variables describing conditions in the stream segment where sampling occurred, downstream factors affecting the ability of fish to move upriver from the sea, and upstream, catchment-scale factors mostly affecting variation in river flows.
2. Analyses were performed using multivariate adaptive regression splines (MARS), a technique that uses piece-wise linear segments to describe non-linear relationships between species and environmental variables. All species were analysed using an option that allows simultaneous analysis of community data to identify the combination of environmental variables best able to predict the occurrence of the component species. Model discrimination was assessed for each species using the area under the receiver operating characteristic curve (ROC) statistic, calculated using a bootstrap procedure that estimates performance when predictions are made to independent data.
3. Environmental predictors having the strongest overall relationships with probabilities of occurrence included distance from the sea, stream size, summer temperature, and catchment-scale drivers of variation in stream flow. Many species were also sensitive to variation in either the average and/or maximum downstream slope, and riparian shade was an important predictor for some species.
4. Analysis results were imported into a Geographic Information System where they were combined with extensive environmental data, allowing spatially explicit predictions of probabilities of occurrence by species to be made for New Zealand's entire river network. This information will provide a valuable context for future conservation management in New Zealand's rivers and streams.

Keywords: distribution, environment, fish, fresh-water, MARS

Introduction

Understanding geographical variation in biological patterns and the processes that give rise to them has long been a central problem in ecology (e.g. Levin, 1992; Buckland & Elston, 1993; Lawton, 1996; Scott, Heglund & Morrison, 2002). While early descriptive

Correspondence: John Leathwick, National Institute of Water and Atmospheric Research, PO Box 11115, Hamilton, New Zealand.
E-mail: j.leathwick@niwa.co.nz

© 2005 Blackwell Publishing Ltd

studies of such patterns were driven largely by scientific curiosity, the growing recognition of our urgent need to conserve both species and communities is now driving an increased demand for detailed, predictive knowledge of relationships between environment and the distributions of biota (e.g. Ferrier *et al*., 2002; Olden, 2003). During the last decade, statistical modelling has become increasingly popular for analysing such patterns (see review by Guisan & Zimmerman, 2000), typically for relating the abundance or occurrence of a species to some set of environmental and/or geographic predictors. It is now used for purposes ranging from the testing of ecological hypotheses (e.g. Austin, 2002) or processes (Leathwick & Austin, 2001) to the prediction of species distributions across geographically extensive areas for conservation (e.g. Gregr & Trites, 2001; Elith & Burgman, 2002) and/or resource management (e.g. Borchers *et al*., 1997).

While statistical modelling has perhaps been applied less in studies of freshwater biota than in terrestrial and marine settings, several recent analyses have used artificial neural nets (e.g. Ripley, 1996) to relate the distributions of fish species to environment (e.g. Lek *et al*., 1996; Brosse & Lek, 2000; Olden & Jackson, 2001). In two such studies, neural nets were used to simultaneously predict the distributions of multiple species using a single analytical model (Olden, 2003; Joy & Death, 2004). However, although neural nets allow fitting of non-linear relationships between species and their predictors, their degree of automation and 'black-box' character allow minimal control over model fitting (Venables & Ripley, 1999), and they are prone to becoming computationally intractable if large numbers of predictors are used (e.g. Moisen & Frescino, 2002; Friedman & Meulman, 2003). More limited use has been made in freshwater studies of generalised additive models (GAM; Hastie & Tibshirani, 1990), a technique that is widely applied in both terrestrial (Guisan & Zimmerman, 2000) and marine environments (e.g. Gregr & Trites, 2001). In several European studies, GAM have been used to analyse the distributions of freshwater macrophytes (Lehmann, 1998; Schmieder & Lehmann, 2004), benthic macro-invertebrates (Castella *et al*., 2001), and fish (Brosse & Lek, 2000).

In this study, we use an alternative technique, multivariate adaptive regression splines (MARS; Friedman, 1991), to analyse the environmental relationships of fifteen diadromous fish species using distributional data from New Zealand rivers and streams. MARS is capable of fitting complex, non-linear relationships between species and predictors, and in one of its implementations can be used to fit a model describing relationships between multiple species and their environment (Hastie & Tibshirani, 1996). In a study of the comparative performance of different modelling techniques using the same data as in this study (J. Leathwick, J. Elith & T. Hastie, unpublished data), a MARS multi-species analysis gave comparable performance to models fitted individually (i.e. species by species) using both MARS and GAM. Such multi-species models may offer advantages by their identification of a set of environmental predictors that best recover overall variation in species composition (Olden, 2003).

Diadromous fish comprise a highly distinctive component of New Zealand's indigenous freshwater fauna (McDowall, 1990, 1993, 1999). The majority are from the families Galaxiidae (five species) and Eleotridae (four species), with three from the Anguillidae, two from the Retropinnidae, and one each from the Geotriidae, Pinguipedidae, and Pleuronectidae. Most of these species spend the majority of their lifespan in fresh water, rather than in the sea. Most are widely distributed in New Zealand, including on its offshore islands, and a number also have a wider regional distribution, including on islands in the Pacific, and in Australian and South American fresh waters. Previous analyses of New Zealand's freshwater fish fauna, including their relationships with environment, can be found, for example, in Minns (1990) and Jowett & Richardson (1995, 1996, 2003). Joy & Death (2004) present results of an analysis of fish: environment relationships in one region of New Zealand using neural nets, and Broad *et al.* (2001) describe a logistic regression model for one diadromous species at a regional scale.

Methods

Fish distribution data

Fish distribution data were drawn from the New Zealand Freshwater Fish Database (McDowall & Richardson, 1983, http://www.niwa.co.nz/services/nzffd/), which now holds fish distribution records for approximately 22 500 sites throughout New Zealand.

A subset of 9866 records was selected for this analysis by including only sites sampled after January 1980, and for which all species were identified (Fig. 1). Repeated samples from the same site were treated as independent records. We also excluded sites from tidal rivers, lakes or other still waters, sites sampled using methods appropriate for only some species or life stages (e.g. whitebait nets, plankton nets, diving or spotlighting), and sites upstream from significant obstructions such as dams, large culverts or impassable cascades or waterfalls that are likely to impede passage by diadromous fish. Most sites selected for analysis were sampled by electric fishing (75%), but a variety of other techniques were also used including nets of various construction (17%), and traps (8%), some of which were baited. Samples from smaller rivers and streams greatly outnumber those from larger rivers, in part reflecting the reduced effectiveness of techniques such as electric fishing in the latter (Minns, 1990). Although records of fish abundances are available for many sites, all data were converted to presence/absence form for this analysis because of difficulties in correcting for differing catch rates for different capture methods and/or variation in the area fished. This paper deals with these observations as records of occurrence, although strictly speaking they are records of capture. We recognise the potential for confounding between detectability, capture method, and environmental relationships, but are generally satisfied that the main trends we model reflect environmental effects on occurrence. A further step would be to include detectability in the models (e.g. MacKenzie *et al*., 2002), but this is a complex undertaking.



Fig. 1 Sample sites from the New Zealand Freshwater Fish Database used in the analysis (open circles). Only rivers with an annual mean flow >10 $m^3 s^{-1}$ are shown. Those shown in light grey were excluded from the analysis because of known significant downstream obstructions to fish migration to/from the sea.

Data were extracted for 15 diadromous species (Table 1) that occurred in the dataset with a capture frequency of 0.5% or above. The anguillids and *Rhombosolea retiaria* are catadromous, whereas *Geotria australis* and some retropinnid stocks are anadromous. The remaining species are amphidromous, meaning that adults remain resident in freshwater, but larval fish are carried out to sea where they spend a short period before migrating back to freshwater to grow to adulthood. The scope of freshwater habitat accessible

Table 1 Fish species included in the analysis, and their prevalence, i.e. the proportion of sample sites at which they were recorded

Code	Species name	Authority	Prevalence
Angaus	*Anguilla australis*	Richardson 1848	0.233
Angdie	*A. dieffenbachii*	Gray 1842	0.577
Galarg	*Galaxias argenteus*	Gmelin 1789	0.034
Galbre	*G. brevipinnis*	Günther 1866	0.099
Galfas	*G. fasciatus*	Gray 1842	0.137
Galmac	*G. maculatus*	Jenyns 1842	0.118
Galpos	*G. postvectis*	Clarke 1899	0.025
Gobcot	*Gobiomorphus cotidianus*	McDowall 1975	0.183
Gobgob	*G. gobioides*	Valenciennes 1837	0.012
Gobhub	*G. hubbsi*	Stokell 1959	0.065
Gobhut	*G. huttoni*	Ogilby 1894	0.211
Geoaus	*Geotria australis*	Gray 1851	0.031
Chefos	*Cheimarrichthys fosteri*	Haast 1874	0.121
Rhoret	*Rhombosolea retiaria*	Hutton 1873	0.007
Retret	*Retropinna retropinna*	Richardson 1848	0.042

to amphidromous species is therefore influenced by their ability to migrate upriver.

Environmental data

Our selection of environmental predictors for this analysis reflected our view that models are more likely to be robust when predictor variables have strong functional relevance to the physiological and behavioural attributes of the species whose distributions are being analysed (Austin, 2002). Here, we combined a conceptual model of environmental factors driving variation in freshwater ecosystems (Biggs *et al*., 1990) with knowledge of the migratory behaviour of diadromous fish, to identify a set of predictors that are likely to have strong functional links with their distributions (Rowe, 1991). For example, there is often a correlation between elevation and fish distributions, but water temperatures and/or downstream slopes or falls that hinder upstream passage are more likely to be the functional variables underlying this relationship, and were used in this study.

The predictors we used can be divided into factors describing the character of the river segment within which the sampling site was located, downstream factors affecting the ability of diadromous fish to migrate from the sea to that river segment, and upstream/catchment-scale factors affecting environmental conditions at the sampling site. As the modelling methods used are potentially sensitive to correlated variables, the final set of candidate variables was restricted to those with pair-wise correlations of <0.7, with two variables normalised (see below) to reduce their correlation with other variables.

Because of the limited and sometimes inconsistent nature of the environmental data collected at the time of fish capture, all environmental predictors were extracted from two broader scale descriptions of New Zealand's rivers and streams. The first of these was a controlling-factors classification of New Zealand rivers and streams, the River Environments Classification (Snelder & Biggs, 2002), while the second consisted of an expanded set of environmental descriptors currently being prepared to enable the development of a comprehensive multivariate classification of New Zealand's rivers and streams (T. Snelder, NIWA, New Zealand, pers. comm.). Using location data recorded at the time of sampling, all sites were linked to the river-segment in which sampling occurred, using a Geographic Information System (GIS)-based spatial database containing all the required environmental variables, and from which the relevant data were subsequently extracted. River segments consisted of a section of a river or stream

Table 2 Environmental predictors used to analyse fish occurrence

Variable	Mean (range)
Segment scale predictors	
SegJanT – summer air temperature (°C)	16.6 (9.5 to 19.8)
SegTSeas – winter air temperature (°C), normalised with respect to SegJanT – see text	0.75, (−2.6 to 4.1)
SegFlow – segment flow (m^3 sec^{-1}), fourth root transformed	0.82 (0.1 to 5.0)
SegShade – riparian shade (proportion)	0.44 (0 to 0.8)
SegSlope – segment slope (°), square-root transformed	2.2 (0 to 5.6)
Downstream predictors	
DSDist – distance to coast (km)	51.5 (0.03 to 329.5)
DSAveSlope – downstream average slope (°)	0.27 (0 to 14.5)
DSMaxSlope – maximum downstream slope (°)	17.6 (0 to 56.5)
Upstream/catchment scale predictors	
USAvgTNorm – average temperature (°C) normalised with respect to SegJanT	−0.027 (−6.2 to 2.2)
USRainDays – days/month with rain greater than 25 mm	1.29 (0.21 to 3.30)
USSlope – average slope in the upstream catchment (°)	13.9 (0 to 41.0)
USIndigForest – area with indigenous forest (proportion)	0.334 (0 to 1)
USPhos – average phosphorus concentration of underlying rocks, 1 = very low to 5 = very high	2.35 (1 to 5)
USCalc – average calcium concentration of underlying rocks, 1 = very low to 4 = very high	1.46 (1 to 4)
USHard – average hardness of underlying rocks, 1 = very low to 5 = very high	3.05 (1 to 5)
USPeat – area of peat (proportion)	0.007 (0 to 1)
USLake – area of lake (proportion)	0.002 (0 to 1)

of variable length extending upstream from any river junction to either the next upstream junction, or to a point where the stream was no longer recognisable at a mapping scale of 1 : 50 000. Segments averaged approximately 740 m in length, with a range from 30 to 15.4 km.

Segment-scale predictors

Predictors relevant to the river segment in which sampling occurred (Table 2) describe the average January (mid-summer) air temperature (SegJanT), seasonal variation in air temperature (SegTSeas), the stream flow (SegFlow), the degree of riparian shading (SegShade), and the segment slope or gradient (SegSlope). Although we explored the estimation of average water temperatures for each segment, we had insufficient data to develop fully and verify such an approach. We therefore used estimates of air temperature in both summer and winter, acknowledging that while shaded rivers and streams are likely to remain substantially in equilibrium with air temperatures, more open waters are likely to have higher temperatures because of heating by solar radiation, particularly in summer (Rutherford *et al.*, 1997). Both sets of temperature estimates were derived from thin-plate splines fitted to meteorological station data (Leathwick & Stephens, 1998). Because the summer and winter temperatures were so highly correlated, the winter estimates were normalised to create a seasonality index, i.e.

$$\mathrm{SegTSeas} = \left(\frac{W - \overline{W}}{\sigma_w} - \frac{S - \bar{S}}{\sigma_s}\right) \times \sigma_w$$

where W is the winter temperature associated with a particular segment, $\overline{W}$ is the average winter temperature across all segments, σ_w is the standard deviation of the winter temperature estimates, S is the summer temperature, and so on. Values indicate the deviation of winter temperatures (°C) from the value expected given the summer temperature, i.e. negative values indicate strong seasonal variation and positive values indicate more muted seasonal variation.

Estimates of mean flow (SegFlow) for each river segment were derived by summing estimates of the excess of average rainfall over average evaporation for cells occurring in a 100 m resolution grid in the upstream catchment. Because these values were highly skewed (samples from small rivers and streams are much more common than those from large rivers), values were subject to a fourth root transformation, after which the values can be expected to be linearly related to variation in water velocity (Jowett, 1998). The degree of riparian shade for each segment (SegShade) was calculated by first overlaying a satellite image-based, digital description of New Zealand's land cover (Land Cover Database – Ministry for the Environment, Wellington) over the river network and calculating the relative proportions of different vegetation classes (i.e. native forest, exotic forest, scrub, etc.) occurring within 100 m alongside each river segment. Shading was then estimated for each stream segment based on its average width (m), and the riparian vegetation cover and its estimated height (m) as follows. The average width of the stream or river channel in metres was computed using the relationship width = 7.8289 × average annual flow$^{0.4777}$ (I. Jowett, NIWA, N.Z., pers. comm.), and the total riverbed width was estimated as channel width/0.75 (Davies-Colley & Quinn, 1998). As no nation-wide mapping of vegetation canopy heights was available, we estimated these from available descriptions of vegetation structure and summer temperatures, with which they are closely linked. The potential canopy height of most native vegetation was estimated from the January air temperature using the relationship canopy height = −47.96 + (7.2987 × JanTemp) − (0.1635 × JanTemp^2), based on a regression fitted to vegetation profile data collated in Wardle (1991). Predicted canopy heights, assuming mature vegetation, ranged from approximately 30 m on warm northern, lowland sites to 12 m at approximately 11 °C (treeline). Exotic forest was allocated a canopy height equal to half that of native forest, assuming an approximately even mix of stand ages from recently planted through to mature, and scrub was allocated a canopy height of 0.3 times the expected mature forest height. Canopy heights were fixed at a nominal height of 2 m for coastal dune and wetland vegetation, and 1 m for pasture and tussock grassland. Following Davies-Colley & Rutherford, 2005, shading under diffuse light conditions was then estimated as $\cos^2\Phi$, where Φ is the angle from the river centre line to the top of the adjacent canopy. All vegetation was assumed to have a density of 80% so that shading values range from 0 (unshaded) to 0.8 in the most heavily shaded reaches. Segment slopes

(SegSlope) were computed from the difference in the elevation at both ends of the segment, along with the total segment length, with the required values extracted from the GIS database.

Downstream predictors

Three predictors were used to describe variation in factors that influence the ability of diadromous fish species to move to a particular river segment from the sea. The distance to the coast (DSDist) was calculated from the GIS database as the sum of the lengths of the individual river segments located between the downstream end of a particular segment and the coast. The average downstream slope (DSAveSlope) was then calculated from the estimated distance to the coast and the elevation at the downstream end of each river segment. Finally, the maximum downstream slope (DSMaxSlope) was calculated by intersecting all downstream river segments with a 30 m resolution grid describing slopes, and finding the maximum intersected value. Values for this predictor were generally much higher than for the average downstream slope, reflecting the local occurrence of features such as rapids or cascades.

Upstream/catchment-scale predictors

Upstream, catchment-scale predictors were computed by combining the GIS database with various gridded environmental data layers, and with all mean catchment values calculated by weighting values for each grid cell by the contribution of that cell to river flow. Two variables were used to describe upstream climatic conditions, i.e. the average mean annual temperature (USAvgTNorm) and the average number of days per month when rainfall exceeded 25 mm (USRainDays). As the first of these was highly correlated with the segment-scale January air temperature, we normalised it with respect to January air temperature (as described for SegTSeas above). Negative values indicate segments in which the upstream catchment experiences colder conditions than expected given the summer temperature occurring in the segment, and are typically rivers and streams with montane headwaters. Positive values indicate segments in which the catchment experiences warmer conditions than expected from the segment summer temperature. Estimates for the catchment of the number of days per month on which rainfall exceeds 25 mm and the average upstream slope (USSlope) provide an estimate of the likely hydrological 'noisiness' or variability in flow of a river segment (e.g. Jowett & Duncan, 1990). Variation in the amount of upstream indigenous forest (USIndigForest) is likely to have a complex influence on both flow and water characteristics, with high peak flows more likely to be buffered in forested catchments with a high capacity for canopy interception and rainfall storage (e.g. Blake, 1975). In forested catchments, water temperatures are also likely to be lower because of the reduced heating through solar radiation (Rutherford *et al.*, 1997), and woody debris is likely to provide more in-stream cover.

The effect of catchment rock type on river characteristics is also complex. Here, we use three variables to describe variation in the physical and chemical character of underlying rocks, i.e. their concentrations of phosphorus (USPhos) and calcium (USCalc), and their induration or physical resistance to weathering (USHard) (see Leathwick *et al.*, 2003). Although we are unaware of any studies that link phosphorus availability and variation in fish assemblages in New Zealand rivers and streams, terrestrial systems in New Zealand are most nutrient-limited by the availability of phosphorus, and phosphorus is an important macro-nutrient for aquatic organisms (e.g. Wetzel, 2001). The amount of calcium in catchment rocks is likely to influence both water chemistry

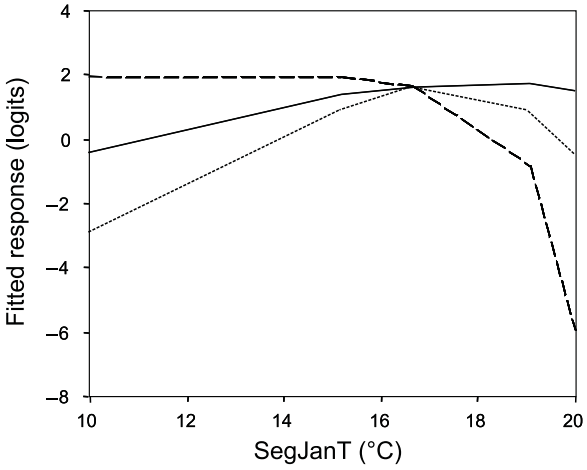


Fig. 2 Comparison of responses fitted for *Galaxias fasciatus* (solid line), *Anguilla dieffenbachii* (dotted line), and *Galaxias brevipinnis* (dashed line) in relation to segment January air temperature (SegJanT). Knots are fitted at values of 15.2, 16.7 and 19.0.

(Close & Davies-Colley, 1990) and the composition of biological communities (e.g. Wetzel, 2001). Variation in the hardness of different rock substrates has important effects on flow variability, reflecting differences in their water storage and transmissivity (Jowett & Duncan, 1990). Along with rainfall and tectonic activity, it also affects both the magnitude of river sediment loads, and their particle size distribution (Hicks, Quinn & Trustrum, 2004), and this in turn can affect fish abundance (Richardson & Jowett, 2002). Estimates of the extent of a catchment occupied by lakes (USLake) or peat (USPeat) indicate the degree of buffering of river flows, while the occurrence of peat also indicates the current and historic distribution of wetlands that provides habitat for particular species.

Model fitting

MARS is a technique in which non-linear responses between a species and an environmental predictor are described by a series of linear segments of differing slope (e.g. Fig. 2), each of which is fitted using a basis function (see Appendix, and Hastie, Tibshirani & Friedman, 2001). Breaks between segments are defined by a knot in a model that initially over-fits the data, and which is then simplified using a backwards/forwards stepwise cross-validation procedure to identify terms to be retained in the final model. All MARS models were fitted in R (R Development Core Team, 2004) using functions contained in the 'mda' library (Hastie & Tibshirani, 1996). As these functions currently only allow models to be fitted assuming ordinary least squares, we analysed our presence/absence data by fitting an initial MARS model, extracting its basis functions, and then fitting a generalised linear model (GLM; McCullagh & Nelder, 1989) that related species occurrence to these, assuming a binomial error distribution. This latter step insures that values of the response variable are constrained between 0 and 1 – in all other respects the model is essentially a MARS model. In addition, we utilised an option in the mda implementation of MARS that allows simultaneous analysis of multiple species by identifying a common set of basis functions, with knots (and therefore variables) selected according to their improvement in predictive power of the model, averaged across all species (Appendix). Separate GLM models were then used to model the occurrence of each species in relation to these basis functions, and the fitted regression coefficients were extracted for subsequent prediction of species distributions in a GIS (see below).

Model performance for each species was assessed using the area under the receiver operating characteristic curve (ROC; e.g. Fielding & Bell, 1997), a statistic that indicates the ability of a model to discriminate between sites where a species is present, versus those where it is absent. A score of 0.5 indicates that a model has no discriminatory ability, while a score of 1 indicates that presences and absences are perfectly discriminated. The area under the ROC curve can be interpreted as indicating the probability that, when a presence site and an absence site are drawn at random from the population, the first will have a higher predicted probability than the second. ROC areas were calculated using a bootstrap procedure that evaluates the performance of a model when predictions are made to omitted portions of the data set (see Appendix and Efron & Tibshirani, 1993, 1997), providing a robust method for assessing predictive performance at new sites.

Summarising the environmental relationships indicated by these various models was complicated by the uneven distribution of sample points in the multidimensional space defined by the predictor variables. This largely reflects the complex patterns of correlation between variables, such that many combinations of environment do not occur in the real world (e.g. large rivers at high elevation). In this setting, consideration of the fitted model functions in isolation from the underlying data can lead to erroneous conclusions being drawn about species: environment relationships. Given the relatively comprehensive sampling by our database of both geographic and environmental space, we therefore based our summaries of the preferred environment for each species on visual interpretation of the fitted probabilities of occurrence. These assessments were then compared with probabilities predicted from each species regression for the entire river network, including river and stream segments not represented by sample points. The latter was accomplished by exporting a tabular summary of the MARS models from R, which was then imported into a proprietary GIS (ArcView, ESRI, CA, USA) where an Avenue script was used to calculate fitted probabilities for each river segment.

Table 3 Summary of contributions of predictors to the MARS regression models

	δ-deviance	Rank
SegJanT	133.7	1
SegTSeas	24.9	10
SegFlow	68.0	4
SegShade	25.2	9
SegSlope	NF	–
DSDist	125.0	2
DSAveSlope	32.9	6
DSMaxSlope	29.4	7
USAvgT	27.0	8
USRainDays	43.8	5
USSlope	68.5	3
USIndigForest	16.6	11
USPhos	14.3	12
USCalc	NF	–
USHard	NF	–
USPeat	NF	–
USLake	11.3	13

Table entries indicate the mean change in residual deviance when dropping that variable from final models, averaged across all species – 'NF' indicates variables that were not included as significant terms by the MARS analysis.

Table 4 Deviance explained and discriminatory power of statistical models relating species presence/absence to environment for fifteen species

Species	Deviance explained	ROC
Anguilla australis	2628.1 (24.6)	0.825 (0.009)
Anguilla dieffenbachii	1626.7 (12.1)	0.712 (0.013)
Galaxias argenteus	696.8 (23.8)	0.856 (0.019)
Galaxias brevipinnis	1725.7 (27.0)	0.856 (0.012)
Galaxias fasciatus	2707.2 (34.4)	0.886 (0.008)
Galaxias maculatus	1661.5 (23.2)	0.835 (0.011)
Galaxias postvectis	510.4 (22.2)	0.847 (0.020)
Gobiomorphus cotidianus	1634.5 (17.4)	0.782 (0.011)
Gobiomorphus gobioides	412.7 (31.6)	0.907 (0.015)
Gobiomorphus hubbsi	1473.3 (31.2)	0.888 (0.012)
Gobiomorphus huttoni	2869.9 (28.2)	0.847 (0.008)
Geotria australis	380.8 (13.9)	0.772 (0.024)
Cheimarrichthys fosteri	1753.0 (24.0)	0.836 (0.012)
Rhombosolea retiaria	369.1 (42.8)	0.943 (0.011)
Retropinna retropinna	839.7 (24.4)	0.858 (0.017)
Average	1419.3 (25.4)	0.843 (0.013)

The deviance explained indicates the reduction in deviance for each species compared with a null model, with the percentage of the total deviance explained shown in brackets. All regressions were fitted using 29 degrees of freedom. Model discrimination was assessed using the area under the receiver operator characteristic curve (ROC) estimated by bootstrap re-sampling. Standard errors are shown in brackets.

Results

Examination of the marginal changes in deviance when dropping individual predictors from the various final models (Table 3) indicated that a relatively small set of predictors plays a dominant role in explaining variation in the probability of occurrence for most diadromous fish species. These included correlates or drivers of key functional aspects of stream character, including summer temperature (SegJanT), distance from the sea (DSDist), stream size (SegFlow), and catchment-scale drivers of variation in water temperature and flow variability (i.e. USAvgT, USSlope and USRainDays). The first two of these variables (SegJanT and DSDist) had nearly twice the explanatory power of any of the other variables, and four variables (SegSlope, USCalc, USHard and USPeat) were not selected by the final model because of their failure to improve significantly its predictive performance at a community level. ROC scores computed using bootstrap re-sampling to simulate model performance with independent data varied between 0.712 and 0.943 (Table 4). Species of low prevalence (e.g. *R. retiaria*, *Gobiomorphus gobioides*) generally had higher ROC scores than species occurring at a high proportion of sites, probably reflecting the greater ease with which the regression models for species of restricted range correctly identified the extensive environments from which they were absent.

Examination of probabilities fitted for the individual species by the MARS analysis indicated that there were marked differences among the environments where they most frequently occurred. For example, species varied widely in the distances that they penetrate upstream from the coast (Fig. 3a), with species such as *G. gobioides* and *R. retiaria* predicted to occur only in relatively coastal river segments. By contrast, species such as *Galaxias fasciatus*, *Retropinna retropinna*, *Gobiomorphus cotidianus*, *Anguilla australis*, *Galaxias brevipinnis*, *Anguilla dieffenbachii* and *Cheimarrichthys fosteri*, although occurring most frequently at sites located within 50 km or less of the coast, also penetrate inland for considerable distances. Moderate levels of occurrence were predicted for these species at sites 150 km inland or more. However, the degree to which different species penetrate inland is also influenced by river slope, with both the average and maximum downstream slope important for many

species. For example, species such as *R. retiaria*, *R. retropinna*, *Galaxias argenteus* and *G. maculatus* occurred most frequently in river and streams with minimal gradients (Fig. 3b), with *G. maculatus* most tolerant of higher stream gradients. By contrast, *G. fasciatus*, *G. brevipinnis*, *G. postvectis* and *Gobiomorphus huttoni* occurred most frequently in relatively high gradient streams. Similar patterns were evident in relation to maximum downstream slopes, with *G. brevipinnis* and *A. dieffenbachii* the two species most capable of passing major stream obstacles.

Strong sorting was also apparent in the distributions of species in relation to the summer temperature gradient (Fig. 3c), with optima distributed over a range of nearly 6 °C. Although all fifteen species were predicted to occur with at least moderate frequency (probabilities of occurrence >0.2) in rivers and streams with warm summer temperatures (16 °C or more), species varied widely in the range of temperatures over which they frequently occurred, with *R. retropinna* occupying the narrowest range and *G. huttoni* the widest. Similar sorting occurred in relation to average temperatures in the upstream catchment (Fig. 3d), with species such as *G. argenteus*, *G. postvectis* and *A. australis* most frequent in river segments in which the contributing catchments had warmer temperatures than expected, while species such as *G. brevipinnis*, *G. cotidianus*, *R. retiaria* and *C. fosteri* occurred most frequently in river segments in which the upstream catchments were typified by cool climatic conditions.

Although sorting of fitted values in relation to river flow was a little more muted (Fig. 3e), there was a marked contrast between species whose preferred habitat is large rivers (*Gobiomorphus hubbsi*, *R. retiaria* and *C. fosteri*), versus the remainder that have a strong preference for small to moderate sized streams and rivers. Note however, that some of these latter species also occurred with moderate frequency in larger rivers. Stream riparian shade is an important factor for some species (Fig. 3f), more so for those that occur in smaller streams than in large rivers, where streambank vegetation is less able to provide substantial shade. Although most species may occur in streams with widely varying degrees of riparian shade, the fish fauna were segregated into two groups; species with a marked preference for shaded streams, and those more frequent in streams with little riparian shade.

The influence of wider, catchment-scale drivers of variability in river-flow on fish distribution is illustrated by variation in fitted values in relation to the average upstream catchment slope (Fig. 3g). Species typical of catchments having a predominance of low slopes, i.e. mostly streams and smaller rivers in lowland areas, include *G. argenteus* and *A. australis*, while at the opposite extreme, species such as *G. brevipinnis*, *C. fosteri*, *G. hubbsi* and *G. huttoni* occur most frequently in catchments having a sizeable proportion of steeper slopes, usually in steep hill-country or mountainous areas. *G. brevipinnis* occurred at sites where upstream slopes were highest, and this reflects its mainly inland, high altitude distribution. Similarly, species such *R. retiaria*, *G. cotidianus*, and *A. australis*, although widespread, were most frequent in river segments in which the contributing catchments had a relatively low frequency of high rainfall days, i.e. low flood environments that are susceptible to summer drought (Fig. 3h), whereas *G. argenteus*, *G. brevipinnis* and *G. postvectis* occurred most frequently in environments where high rainfall days are a regular occurrence and flows are more reliable through the year.

Consideration of relationships between fitted probabilities and the predictors that make a lower contribution to the explanation of deviance (variables ranked 10 or lower in Table 3) also revealed some interesting patterns. For example, the majority of species were most likely to occur in environments with only moderate seasonal variation in temperature (Table 5), with only two species, *C. fosteri* and *R. retiaria*, occurring most frequently in rivers with strong seasonal variation in temperature. Most species are also more common in rivers that have minimal lake buffering, with *A. dieffenbachii* and *G. cotidianus* the only species to show a positive preference for rivers with substantial lake influence. Only a few species showed strong patterns of occurrence in relation to either USIndigForest or USPhos with *G. brevipinnis* and *G. postvectis* more likely in catchments with extensive native forest cover and *R. retiaria* more common in rivers with non-forested catchments.

Probabilities of occurrence calculated within ArcView from the MARS analysis and using environmental data for the entire river network are shown for *G. brevipinnis* and *G. huttoni* for part of the western North Island (Fig. 4). Highest probabilities of occurrence for *G. brevipinnis* are predicted for small, inland

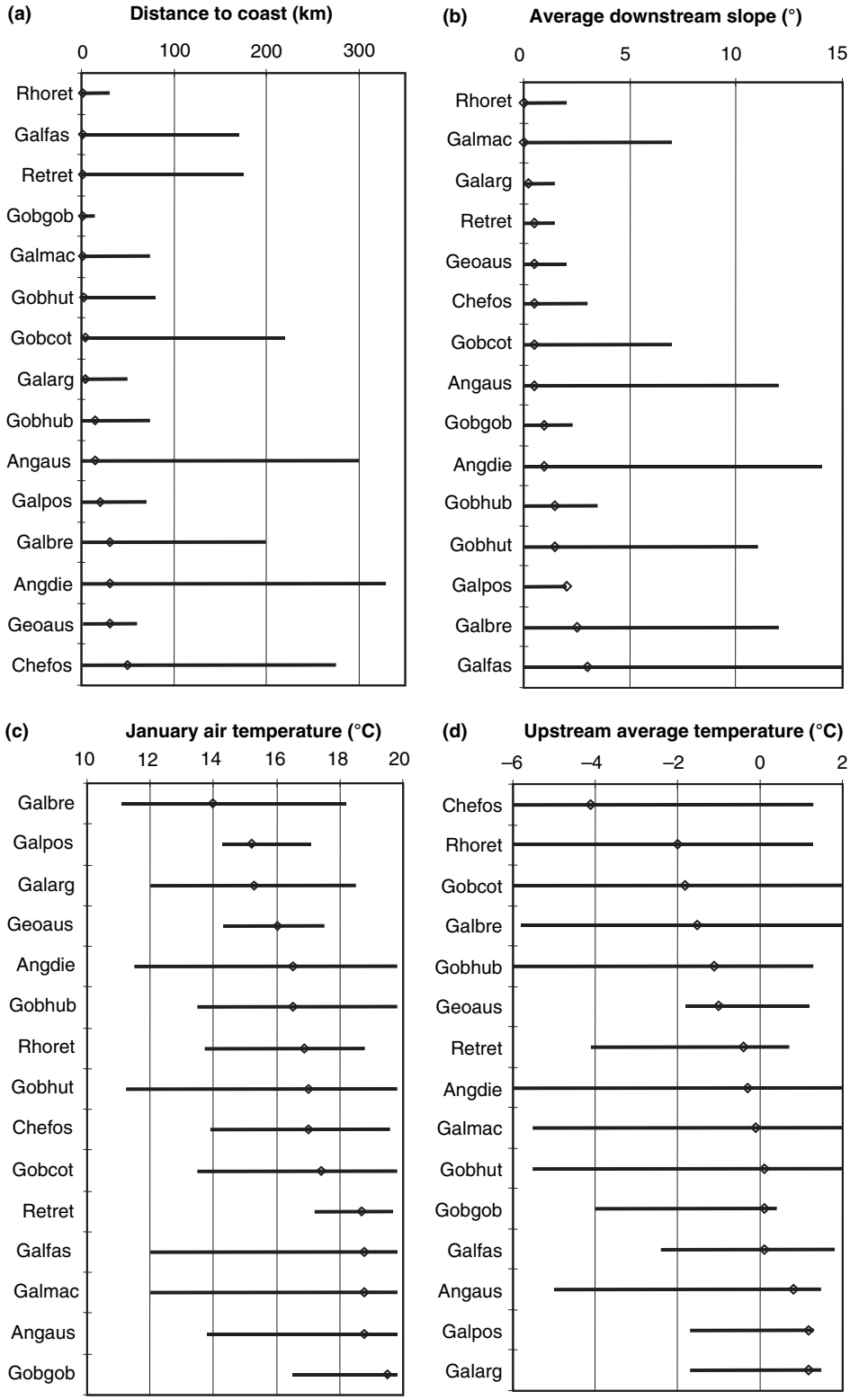


Fig. 3 Distributions of species in relation to major environmental predictors (a–h) as indicated by fitted probabilities of occurrence from the MARS community model. Values at which fitted probabilities reach a maximum are shown for each environmental variable by a diamond, and the range over which fitted values exceed 0.2 are indicated by horizontal lines. Species are sorted by their indicated optima. Codes for species consist of the first three letters of the generic and specific names as listed in Table 1.

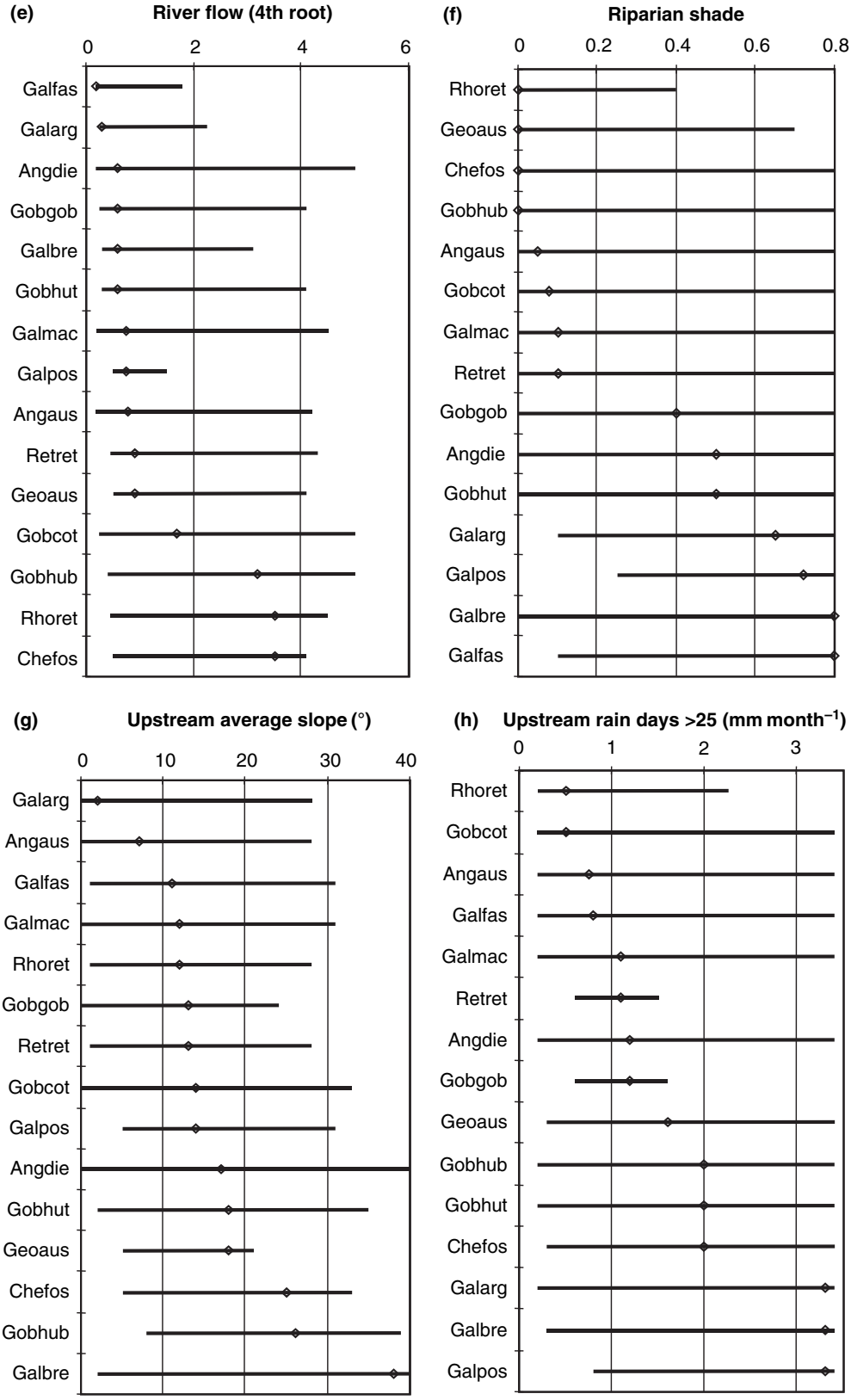


Fig. 3 *(Continued)*

Table 5 Species showing significant variation in their fitted probabilities in relation to less important environmental predictors

Variable	Negative	Positive
SegTSeas	*Rhombosolea retiaria, Cheimarrichthys fosteri*	*Anguilla australis, Anguilla dieffenbachii, Galaxias argenteus, Galaxias brevipinnis, Galaxias fasciatus, Galaxias maculatus, Galaxias postvectis, Gobiomorphus cotidianus, Gobiomorphus gobioides, Gobiomorphus hubbsi, Gobiomorphus huttoni, Retropinna retropinna*
USLake	*Anguilla australis, Galaxias argenteus, Galaxias brevipinnis, Galaxias fasciatus, Galaxias maculatus, Galaxias postvectis, Gobiomorphus gobioides, Gobiomorphus hubbsi, Cheimarrichthys fosteri, Rhombosolea retiaria, Retropinna retropinna*	*Anguilla dieffenbachii, Gobiomorphus cotidianus*
USIndigForest	*Rhombosolea retiaria*	*Galaxias brevipinnis, Galaxias postvectis*
USPhos	*Galaxias argenteus, Gobiomorphus hubbsi, Retropinna retropinna*	*Rhombosolea retiaria*

Species are listed whose fitted values vary by 50% or more with progression along the environmental predictors, and are separated into those showing negative versus positive responses.

streams at high elevation, and with moderate to high downstream slopes. By contrast, *G. huttoni* is predicted to occur most frequently in small to moderate sized coastal streams and rivers, with higher probabilities predicted for those with good riparian shade.

Finally, results of an analysis in which a factor variable indicating the capture method was added to each of the MARS individual models (Table 6), showed that there were marked differences for some species in their probabilities of capture between different sampling methods, with electric fishing generally the most effective technique for most species. Differences were most significant for the two eel species, with addition of capture method providing a substantial improvement in model discrimination, particularly for *A. dieffenbachii*. Modest improvements in model discrimination also occurred for *G. australis* and *C. fosteri*.

Discussion

Results from this analysis highlight the insights to be gained from the use of statistical modelling to relate data describing species distributions to a functionally relevant set of environmental predictors. In part the results reflect the comprehensive nature of the data available, i.e. an extensive set of fish distributional data collected using specified sampling procedures, as well as a comprehensive set of environmental predictors. At one level, results provide a robust description of the distributions of species and their sorting in environmental space (as in Austin & Smith, 1989), this information providing an important underpinning for individual species management (e.g. Levin, 1992). However, the availability of comprehensive environmental data for the entire river network also enabled the prediction of species occurrences throughout New Zealand, including for many rivers and streams for which no species sampling has been carried out. Such information is critical to the assessment of the conservation value of rivers, providing both a framework for the robust identification of representative sites for protection (e.g. Scott *et al*., 1993; Pressey *et al*., 2000), and a context for interpreting results from ongoing monitoring (Oberdorff *et al*., 2001; Joy & Death, 2004; Schmieder & Lehmann, 2004).

In terms of the ecology of these species at a collective level, the descriptions of species: environment relationships that emerge from this modelling approach accord strongly with previously published accounts of diadromous fish ecology at various scales of study. For example, the importance of distance from the coast and stream gradient as determinants of the degree to which different diadromous species penetrate inland has already been described nationally by McDowall (1990, 1993, 1999 and demonstrated numerically in several studies at national (e.g. Minns, 1990; Jowett & Richardson, 2003) and regional scales (e.g. Hayes, Leathwick & Hanchet, 1989; Jowett *et al*., 1998). Similarly, Jowett & Richardson (1995) describe varying preferences among species for different velocities and depths of water that accord strongly with

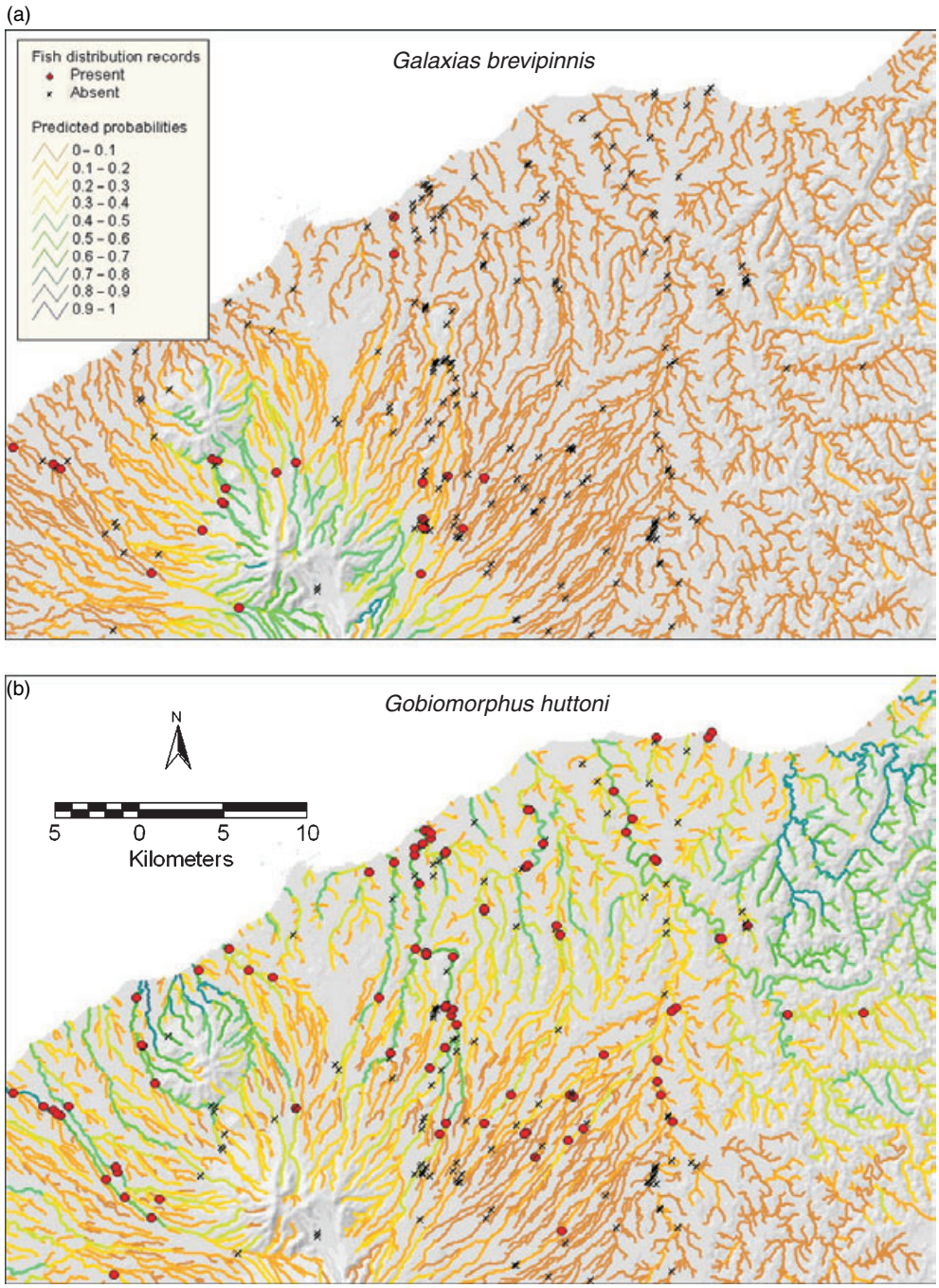


Fig. 4 Predictions from the MARS community model for (a) *Galaxias brevipinnis* and (b) *Gobiomorphus huttoni* for rivers and streams in northern Taranaki, North Island, New Zealand.

our descriptions of species occurrences with respect to river size (see also Jowett & Richardson, 2003). At a broader catchment-scale, Jowett & Duncan (1990) describe strong links between variability in river flows and variation in catchment-scale factors such as precipitation, catchment geology and slope, forest cover and the proportion of a catchment occupied by lakes. They in turn link this with variation in fish habitat availability related to the degree of development of pool/riffle structure, the nature of periphyton

Table 6 Variation in catchability of species by different capture methods as tested by adding to the original environment-only regression, a four level category variable describing capture methods

Species	Scaled deviance	Improvement in ROC	Most effective method
Anguilla australis	158.5	0.026	Electric fishing
Anguilla dieffenbachii	260.3	0.050	Electric fishing
Galaxias argenteus	29.2	0.009	Electric fishing
Galaxias brevipinnis	23.0	0.004	Electric fishing
Galaxias fasciatus	(0.8)	–	–
Galaxias maculatus	19.6	0.005	Mixture
Galaxias postvectis	5.3	0.003	Electric fishing
Gobiomorphus cotidianus	39.0	0.008	Electric fishing
Gobiomorphus gobioides	11.2	0.012	Traps
Gobiomorphus hubbsi	55.9	0.013	Electric fishing
Gobiomorphus huttoni	74.2	0.011	Electric fishing
Geotria australis	23.6	0.028	Electric fishing
Cheimarrichthys fosteri	97.5	0.022	Electric fishing
Rhombosolea retiaria	(1.8)	–	–
Retropinna retropinna	4.3	0.002	Nets

Table values indicate the scaled change in deviance, the marginal improvement in ROC score, and the most effective fishing method for each species. Changes in deviance are distributed approximately as for a *F*-statistic, with values >2.60 and 3.78 significant at $P > 0.05$ and 0.01 respectively. Non-significant changes are enclosed in brackets.

communities, and the distribution and abundance of aquatic invertebrates. Several studies have also highlighted both the importance of riparian shade for species such as *G. fasciatus*, *A. dieffenbachii*, *G. huttoni* and/or *G. argenteus*, and the preference shown by *A. australis* and *Galaxias maculatus* for less-shaded waters (e.g. Hanchet, 1990; Rowe *et al*., 1999; Jowett & Richardson, 2003). Rowe *et al*. (2002) describe links between the presence of upstream indigenous forest and the occurrence of native fish species in downstream habitats.

Less information is available about the importance of temperature for these fish species. Based on a synthesis of field data, Rowe (1991) indicates sorting of species in a similar order along a generalised water temperature gradient similar to that indicated in this study, but results from a laboratory study of the preferred and lethal temperatures for a range of fish species (Richardson, Boubée & West, 1994) show only moderate agreement. This may reflect differences in the ability of species to tolerate extreme summer temperatures, with stenothermal species likely to be much less tolerant of widely varying temperatures than eurythermal species. However, this result may also reflect our lack of proximate measurements of water temperature, and the wide variation in water temperature that can occur with variation in riparian shade (Rutherford *et al*., 1997).

Consideration of the environmental correlates of particular species or groups of species also reveals some interesting insights into their niche requirements. For example, three species that occur most frequently in large rivers, where high water velocities can be expected, are all well adapted to resist displacement by strong currents. *G. hubbsi* and *C. fosteri* both penetrate moderate distances inland and show morphological adaptations that allow them to live in rocky rapids (McDowall, 1990). By contrast, *R. retiaria*, a flounder, generally penetrates only small distances inland, and is adapted to resist river and tidal currents in lower gradient, more gently flowing lower reaches. Similar contrasts can be seen in the environmental preferences of *G. fasciatus* and *G. brevipinnis*, the two Galaxiids most capable of climbing waterfalls and other obstructions. Although the first of these two species extends into streams with steeper gradients than the former, it is also much more coastal in its distribution. While it has a wide tolerance of summer temperatures, *G. fasciatus* is much less tolerant of cold winter temperatures than *G. brevipinnis*. In addition, *G. brevipinnis* also occurs most frequently in catchments with frequent high rainfall, and this may indicate either a greater tolerance of frequent flood events, or sensitivity to the low flows that occur in lower rainfall areas where *G. fasciatus* is more common. Alternatively, the greater

size of returning juveniles of *G. brevipinnis* may endow them with a greater ability to migrate inland than *G. fasciatus*. Finally, the positive association between probabilities of occurrence for *G. cotidianus* and lake influence is likely to reflect its known ability to breed in lacustrine environments, with downstream dispersal of individuals likely.

The relationships between local-scale variation in pool/riffle structure and stream biology described by Jowett & Duncan (1990) highlights an important issue related to the dominant spatial scales at which our analysis was carried out. That is, our study focussed on the broader, segment- and catchment-scale drivers of variation in probabilities of occurrence for different species. Our ability to explore links between this and variation in both pool/riffle structure and substrate size was limited not by lack of relevant local-scale data from the sample sites, but by our lack of comprehensive data describing variation in these factors throughout the river network. As a consequence, although we could have provided a statistical description of relationships between fish occurrence and variation in these factors, we would have been unable to incorporate this into our subsequent predictions of occurrence throughout the New Zealand river network, and so we chose to omit it from this analysis.

Choice of analytical method

To our knowledge, this is the first published use of MARS to analyse species distributions in freshwater environments. Comparative studies in terrestrial (Moisen & Frescino, 2002; Muñoz & Felicísimo, 2004) and aquatic environments (Leathwick *et al*., in review) suggest that this technique offers a similar level of performance to other non-linear modelling techniques such as GAM, artificial neural nets, and classification and regression trees (Breiman *et al*., 1984). Given this relative lack of difference in performance between several methods that fit non-linear relationships, we argue that other issues including computational demands (Leathwick *et al*., in review), ability to use analysis results to make more extensive predictions in a GIS (Muñoz & Fellicísimo, 2004), ease of inspection of fitted relationships for their ecological sensibility (Austin, 2002), and sensitivity to inclusion of non-discriminatory predictors (Friedman & Meulman, 2003) are probably the most important features governing choice of analytical technique. MARS not only measures up well against these criteria, but its ability to perform unified analyses of community data also offers advantages for applications focusing on community level conservation management (Olden, 2003).

Acknowledgements

This project would not have been possible without the enormous effort that went into the assembly of the New Zealand Freshwater Fish Database, which was instigated by Bob McDowall and is maintained by Jody Richardson—particular thanks are owed to numerous researchers who have contributed their data over several decades. Similar credit must be given to those who developed major databases describing environmental variation in New Zealand's rivers and streams, and in particular Ton Snelder, Mark Weatherhead and Helen Hurren. This work was funded jointly by New Zealand's Ministry for the Environment and Department of Conservation. Bob McDowall and Ian Jowett gave generously of their extensive knowledge of New Zealand freshwater ecosystems, and Rob Davies-Colley provided guidance on the estimation of stream shade. The inspiration for a comparative analysis of GAM and MARS models was generated by discussions at a workshop on statistical modelling of species distributions held in Riederalp, Switzerland, in August 2004. This work was funded by New Zealand's Foundation for Research, Science and Technology under contract C01X0305.

References

Austin M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. *Ecological Modeling*, **157**, 101–118.

Austin M.P & Smith T.M. (1989) A new model for the continuum concept. *Vegetatio*, **83**, 35–47.

Biggs B.J., Duncan M.J., Jowett I.G., Quinn J.M., Hickey C.W., Davies-Colley R.J. & Close M.E. (1990) Ecological characterisation, classification, and modeling of New Zealand rivers: an introduction and synthesis. *New Zealand Journal of Marine and Freshwater Research*, **24**, 277–304.

Blake G.J. 1975. The interception process. In: *Prediction in Catchment Hydrology* (Eds T.G. Chapman & F.X. Dunin), pp. 59–81. Australian Academy of Science, Canberra.

Borchers D.L., Buckland S.T., Priede I.G. & Ahmadi S. (1997) Improving the precision of the daily egg production method using generalized additive models. *Canadian Journal of Fisheries Aquatic Science*, **54**, 2727–2742.

Breiman L., Friedman J.H., Olshen R.A. & Stone C.J. (1984) *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Monterey, California, 358 pp.

Broad T.L., Townsend C.R., Arbuckle C.J. & Jellyman D.J. (2001) A model to predict the presence of longfin eels in some New Zealand streams, with particular reference to riparian vegetation and elevation. *Journal of Fish Biology*, **58**, 1098–1112.

Brosse S. & Lek S. (2000) Modelling roach (*Rutilus rutilus*) microhabitat using linear and nonlinear techniques. *Freshwater Biology*, **44**, 441–452.

Buckland S.T. & Elston D.A. (1993) Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology*, **30**, 478–495.

Castella E., Adalsteinsson H., Brittain J.E. *et al.* (2001) Macrobenthic invertebrate richness and composition along a latitudinal gradient of European glacier-fed streams. *Freshwater Biology*, **46**, 1811–1831.

Close M.E. & Davies-Colley R.J. (1990) Baseflow water chemistry in New Zealand rivers. 2. Influence of environmental factors. *New Zealand Journal of Marine and Freshwater Research*, **24**, 343–356.

Davies-Colley R.J. & Quinn J.M. (1998) Stream lighting in five regions of North Island, New Zealand: control by channel size and riparian vegetation. *New Zealand Journal of Marine and Freshwater Research*, **32**, 591–605.

Davies-Colley R.J. & Rutherford J.C. (2005) Some approaches for measuring and modeling riparian shade. *Ecological Engineering*, **24**, 525–530.

Efron B. & Tibshirani R.J. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, London.

Efron B. & Tibshirani R.J. (1997) Improvements on cross-validation: the 0.632+ bootstrap method. *Journal of the American Statistical Association*, **92**, 548–560.

Elith J. & Burgman M.A. (2002) Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. In: *Predicting Species Occurrence: Issues of Accuracy and Scale* (Ed. F.B. Sampson), pp. 303–314. Island Press, Covelo, CA.

Ferrier S., Watson G., Pearce J. & Drielsma M. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity: the north-east New South Wales experience. I. Species-level modelling. *Biodiversity and Conservation*, **11**, 2275–2307.

Fielding A.H. & Bell J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Friedman J.H. (1991) Multivariate adaptive regression splines. *Annals of Statistics*, **19**, 1–141.

Friedman J.H. & Meulman J.J. (2003) Multiple adaptive regression trees with application in epidemiology. *Statistics in Medicine*, **22**, 1365–1381.

Gregr E.J. & Trites A.W. (2001) Predictions of critical habitat for whale species in the waters of coastal British Colombia. *Canadian Journal of Fisheries Aquatic Science*, **58**, 1265–1285.

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

Hanchet S.M. (1990) Effect of land use on the distribution and abundance of native fish in tributaries of the Waikato River in the Hakarimata Range, North Island, New Zealand. *New Zealand Journal of Marine and Freshwater Research*, **24**, 159–171.

Hastie T. & Tibshirani R.J. (1990) *Generalized Additive Models*. Chapman and Hall, London.

Hastie T. & Tibshirani R.J. (1996) Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society (Series B)*, **58**, 155–176.

Hastie T., Tibshirani R.J. & Friedman J.H. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York.

Hayes J.W., Leathwick J.R. & Hanchet S.M. (1989) Fish distribution patterns and their association with environmental factors in the Mokau River catchment, New Zealand. *New Zealand Journal of Marine and Freshwater Research*, **23**, 171–180.

Hicks M., Quinn J. & Trustrum N. (2004) Stream sediment load and organic matter. In: *Freshwaters of New Zealand* (Eds J. Harding, P. Moseley, C. Pearson & B. Sorrel), Chapter 12. Caxton Press, Christchurch.

Jowett I.G. (1998) Hydraulic geometry of New Zealand rivers and its use as a preliminary method of habitat assessment. *Regulated Rivers: Research and Management*, **14**, 451–466.

Jowett I.G. & Duncan M.J. (1990) Flow variability in New Zealand rivers and its relationship to in-stream habitat and biota. *New Zealand Journal of Marine and Freshwater Research*, **24**, 305–317.

Jowett I.G. & Richardson J. (1995) Habitat preferences of common, riverine New Zealand native fishes and implications for flow management. *New Zealand Journal of Marine and Freshwater Research*, **29**, 13–23.

Jowett I.G. & Richardson J. (1996) Distribution and abundance of freshwater fish in New Zealand rivers. *New Zealand Journal of Marine and Freshwater Research*, **30**, 239–255.

Jowett I.G. & Richardson J. (2003) Fish communities in New Zealand rivers and their relationship to environ-

mental variables. *New Zealand Journal of Marine and Freshwater Research*, **37**, 347–366.

Jowett I.G., Hayes J.W., Deans N. & Eldon G.A. (1998) Comparison of fish communities and abundance in unmodified streams of Kahurangi National Park with other areas of New Zealand. *New Zealand Journal of Marine and Freshwater Research*, **32**, 307–322.

Joy M.K. & Death R.G. (2004) Predictive modeling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. *Freshwater Biology*, **49**, 1036–1052.

Lawton J. (1996) Patterns in ecology. *Oikos*, **75**, 145–147.

Leathwick J.R. & Austin M.P. (2001) Competitive interactions between forest tree species in New Zealand's old-growth indigenous forests. *Ecology*, **82**, 2560–2573.

Leathwick J.R. & Stephens R.T.T. (1998) *Climate Surfaces for New Zealand*. *Landcare Research Contract Report LC9798/126*. Landcare Research, Lincoln, New Zealand.

Leathwick J.R., Wilson G., Rutledge D., Wardle P., Morgan F., Johnston K., McLeod M. & Kirkpatrick R. (2003) *Land Environments of New Zealand*. Bateman, Auckland, 183 pp.

Lehmann A. (1998) GIS modeling of submerged macrophyte vegetation using Generalized Additive Models. *Plant Ecology*, **139**, 113–124.

Lek S., Delacoste M., Baran P., Dimopoulos I., Lauga J. & Aulagnier S. (1996) Applications of neural networks to modeling nonlinear relationships in ecology. *Ecological Modeling*, **90**, 39–52.

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

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.

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

McDowall R.M. (1990) *New Zealand Freshwater Fishes – a Natural History and Guide*. Heinemann Reed, Auckland, 553 pp.

McDowall R.M. (1993) Implications of diadromy for the structuring and modeling of riverine fish communities in New Zealand. *New Zealand Journal of Marine and Freshwater Research*, **27**, 453–462.

McDowall R.M. (1999) Driven by diadromy: its role in the historical and ecological biogeography of the New Zealand freshwater fish fauna. *Italian Journal of Zoology*, **65**(Suppl. S), 73–85.

McDowall R.M. & Richardson J. (1983) The New Zealand freshwater fish survey, a guide to input and output. *New Zealand Ministry of Agriculture and Fisheries, Fisheries Research Division Information Leaflet*, **12**, 1–15.

Minns C.K. (1990) Patterns of distribution and association of freshwater fish in New Zealand. *New Zealand Journal of Marine and Freshwater Research*, **24**, 31–44.

Moisen G.G. & Frescino T.T. (2002) Comparing five modeling techniques for predicting forest characteristics. *Ecological Modelling*, **157**, 209–225.

Muñoz J. & Felicísimo Á.M. (2004) Comparison of statistical methods commonly used in predictive modeling. *Journal of Vegetation Science*, **15**, 285–292.

Oberdorff T., Pont D., Hugueny B. & Chessel D. (2001) A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. *Freshwater Biology*, **46**, 399–415.

Olden J.D. (2003) A species-specific approach to modeling biological communities and its potential for conservation. *Conservation Biology*, **17**, 854–863.

Olden J.D. & Jackson D.A. (2001) Fish-habitat relationships in lakes: gaining predictive and explanatory insight by using artificial neural networks. *Transactions of the American Fisheries Society*, **130**, 878–897.

Pressey R.L., Hager T.C., Ryan K.M., Schwarz J., Wall S., Ferrier S. & Creaser P.M. (2000) Using abiotic data for conservation assessments over extensive regions: quantitative methods applied across New South Wales, Australia. *Biological Conservation*, **96**, 55–82.

R Development Core Team (2004) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Richardson J. & Jowett I.G. (2002) Effects of sediment on fish communities in East Cape streams, North Island, New Zealand. *New Zealand Journal of Marine and Freshwater Research*, **36**, 431–442.

Richardson J., Boubée J.A.T. & West D.W. (1994) Thermal tolerance and preference of some native New Zealand freshwater fish. *New Zealand Journal of Marine and Freshwater Research*, **28**, 399–407.

Ripley B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Rowe D.K. (1991) Native fish habitat and distribution. *Freshwater Catch*, **46**, 4–6.

Rowe D.K., Chisnall B.L., Dean T.L. & Richardson J. (1999) Effects of land use on native fish communities in east coast streams of the North Island of New Zealand. *New Zealand Journal of Marine and Freshwater Research*, **33**, 141–151.

Rowe D.K., Smith J., Quinn J. & Boothroyd I. (2002) Effects of logging with and without riparian strips on fish species abundance, mean size, and the structure of native fish assemblages in Coromandel, New Zealand, streams. *New Zealand Journal of Marine and Freshwater Research*, **36**, 67–79.

Rutherford J.C., Blackett S., Blackett C., Saito L. & Davies-Colley R.J. (1997) Predicting the effects of shade on water temperature in small streams. *New Zealand Journal of Marine and Freshwater Research*, **31**, 707–721.

Schmieder K. & Lehmann A. (2004) A spatio-temporal framework for efficient inventories of natural resources: a case study with submersed macrophytes. *Journal of Vegetation Science*, **15**, 807–816.

Scott J.M., Heglund P.J. & Morrison M.L. (2002) *Predicting Species Occurrences: Issues of Accuracy and Scale*. Island Press, Washington, D.C.

Scott J.M., Davis F., Csuti B. *et al.* (1993) Gap analysis: a geographical approach to protection of biological diversity. *Wildlife Monographs*, 123.

Snelder T.H. & Biggs B.J.F. (2002) Multiscale river environment classification for water resources management. *Journal of the American Water Resources Association*, **38**, 1225–1239.

Steyerberg, E.W., Harrell, F.E. Jr., Borsboom, G.J.J.M., Eijkemans, M.J.C., Vergouwe, Y. & Habbema, J.D.F. (2001) Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, **54**, 774–781.

Venables W.N. & Ripley B.D. (1999) *Modern Applied Statistics with S-PLUS*, 3rd edn. Springer-Verlag, New York, 501 pp.

Wardle P. (1991) *Vegetation of New Zealand*. Cambridge University Press, Cambridge.

Wetzel R.G. (2001) *Limnology: Lake and River Ecosystems*. Academic Press, San Diego, 1006 pp.

Wintle BA, Elith J & Potts J (2005) Fauna habitat modelling and mapping in an urbanising environment; A case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology*, **30**, 729–748.

(Manuscript accepted 4 August 2005)

Appendix: analysis details

Model fitting

Multivariate adaptive regression splines is a procedure for fitting adaptive non-linear regression that uses piece-wise linear basis functions to define relationships between a response variable and some set of predictors (Friedman, 1991). The resulting regression surface is piecewise linear and continuous. Basis functions are defined in pairs, using a knot or value of a variable that defines an inflection point along the range of a predictor, e.g.

$$\mathrm{bf_n} = \max(0, 1.0 - \mathrm{SegFlow})$$

$$\mathrm{bf_{n+1}} = \max(0, \mathrm{SegFlow} - 1.0)$$

In this example the knot takes a value of one, and the values of $\mathrm{bf_n}$ can therefore be seen to have a value of one when SegFlow is zero, declining to zero as SegFlow approaches one. They remain fixed at zero at values of SegFlow >1. By contrast, $\mathrm{bf_{n+1}}$ (the pair to $\mathrm{bf_n}$) takes a value of SegFlow – 1 when SegFlow is >1, but otherwise takes a value of zero. Coefficients applied to each of the basis functions define the slopes of the non-zero sections. The fitting of two or more such basis functions in a linear regression allows specification of different slopes within different parts of the range of a predictor variable, effectively allowing the fitting of a non-linear response between a response and its predictor. More than one knot (i.e. more than one pair of basis functions) can be specified for a predictor variable, allowing complex non-linear relationships to be fitted. Another way to think of the basis functions is to envisage a new model matrix, where each predictor variable in the original data is replaced by two or more columns that are the basis functions for that variable.

When fitting a MARS model, knots are chosen in a forward stepwise manner (Hastie & Tibshirani, 1996). Candidate knots can be placed at any position within the range of each predictor variable to define a pair of basis functions. At each step, the model selects the knot and its corresponding pair of basis functions that give the greatest decrease in the residual sum of squares. Knot selection proceeds until some maximum model size is reached, after which a backwards-pruning procedure is applied in which those basis functions that contribute least to model fit are progressively removed. At this stage, a predictor variable can be dropped from the model completely if none of its basis functions contribute meaningfully to predictive performance. The sequence of models generated from this process is then evaluated using generalised cross-validation, and the model with the best predictive fit is selected.

Two novel features are possible when using MARS. First, interactions between variables can be fitted, but rather than fitting a global interaction between a pair

of variables, these are specified using basis functions. As each basis function only describes variation for part of the range of its variable, interactions are specified locally, i.e. the interaction effect is confined to the sub-ranges of the two variables described by the non-zero parts of the basis functions, rather than across the full range of both variables. The R implementation of MARS also allows for the fitting of multiple response variables. In this case knots are selected based on their ability to reduce the residual sum of squares, averaged across all response variables. The final MARS model then uses a common set of basis functions for all response variables, but individual regressions are used to relate variation in each response variable to the final set of basis functions (i.e. to calculate unique coefficients for each basis function per species).

The current implementation of MARS in R uses least squares fitting appropriate for data with normally distributed errors. To constrain predicted values within the range 0–1, as appropriate for presence-absence data, we first fitted a MARS model using the standard R code. We then extracted the basis functions from this model and computed a GLM model(s) that related these to the presence/absence of each species.

Assessment of model performance

Ideally the predictive performance of models such as we have fitted would be assessed by making predictions for sites not used in the model fitting, with these predictions then compared with actual probabilities of occurrence at the independent sites. However, in a study such as this it is generally desirable to include all possible sites in fitting the final model. Two alternatives can be used to assess robustly the performance of the final model when making predictions to independent sites. In *k*-fold cross validation (e.g. Hastie *et al*., 2001) the data are divided into a small number (usually five or ten) of mutually exclusive subsets. Model performance is assessed by successively removing each subset, re-fitting the model to the retained data, and predicting to the omitted data. The average error when predicting to new sites can then be calculated by averaging the predictive performance across each of the subsets.

Alternatively, repeated bootstrap samples of the same size as the original data can be selected from it at random, but with replacement (Efron & Tibshirani, 1993, 1997), and on average these will include around 63% of the sites in the original dataset. When a model is fitted to such a bootstrap sampled dataset, predictions can be made to the omitted sites, and comparison of the actual and predicted occurrences can be used to assess the predictive performance of the model. Repeated implementation of this procedure (e.g. 200–300 times) allows the predictive performance to be averaged across many samples of the original data rather than five or ten subsets, so that boot-strap sampling can be seen as a smoothed or averaged form of *k*-fold cross-validation. Variations in this bootstrap methodology focus on whether predictive performance is assessed only on omitted sites, or on weighted or unweighted combinations of omitted and modelled sites. Weighted combinations tend to give the most unbiased estimates (Hastie *et al*., 2001), so we followed the 0.632+ procedure of Efron & Tibshirani (1997) as described by Steyerberg *et al*. (2001). An ecological example of the use of this procedure can be found in Wintle, Elith & Potts (2005). We selected a total of 300 bootstrap samples, and for each sample, completely re-selected the model and predicted to both the withheld data and the modelled data. The area under the ROC curve was then calculated to measure model performance, as described in the cited references to the 0.632+ bootstrap.