Journal of Biogeography (J. Biogeogr.) (2008) **35**, 1481–1497

ORIGINAL ARTICLE

Dispersal, disturbance and the contrasting biogeographies of New Zealand's diadromous and non-diadromous fish species

J. R. Leathwick[1]*, J. Elith[2], W. L. Chadderton[3], D. Rowe[1] and T. Hastie[4]

[1]*National Institute of Water and Atmospheric Research, Hamilton, New Zealand,* [2]*School of Botany, University of Melbourne, Parkville, Vic., Australia,* [3]*The Nature Conservancy, Great Lakes Program, Chicago, IL, USA and* [4]*Department of Statistics, Stanford University, CA, USA*

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

ABSTRACT

Aim To examine the relationship between diadromy and dispersal ability in New Zealand's freshwater fish fauna, and how this affects the current environmental and geographic distributions of both diadromous and non-diadromous species.

Location New Zealand.

Methods Capture data for 15 diadromous and 15 non-diadromous fish species from 13,369 sites throughout New Zealand were analysed to establish features of their geographic ranges. Statistical models were used to determine the main environmental correlates of species' distributions, and to establish the environmental conditions preferred by each species. Environmental predictors, chosen for their functional relevance, were derived from an extensive GIS database describing New Zealand's river and stream network.

Results In terms of geography, most diadromous species occur in a scattered fashion throughout extensive geographic ranges, and occupy large numbers of catchments of widely varying size. By contrast, most non-diadromous species show relatively high levels of occupancy of smaller geographic ranges, and most are restricted to a few large catchments, particularly in the eastern South Island. In terms of environment, there is marked separation of diadromous from non-diadromous species, with diadromous species generally caught most frequently in low-gradient coastal rivers and streams with warm, maritime climates. With a few notable exceptions, most diadromous species have lower occurrence in river segments that are located above obstacles to upstream migration. Non-diadromous species are usually caught in inland rivers and streams with cool, strongly seasonal climates, typified by a low frequency of high-intensity rainfall events.

Main conclusions We interpret the contrasting biogeographies of New Zealand's diadromous and non-diadromous species as reflecting interaction between their marked differences in dispersal ability and a landscape that is subject to recurrent, often large-scale, natural disturbance. While both groups are likely to be equally susceptible to local, disturbance-driven extinction, the much greater dispersal ability of diadromous species has allowed them to persist over wide geographic ranges. By contrast, the distributions of most non-diadromous species are concentrated in a few large catchments, mostly in regions where less intense natural disturbance regimes are likely to have favoured their survival.

Keywords

Boosted regression tree, diadromy, dispersal, disturbance, fish, freshwater, statistical model, species ranges.

INTRODUCTION

The role played by contemporary dispersal in the widespread geographic distributions of a number of southern, diadromous, freshwater fish species has been periodically debated ever since Charles Darwin (1872) first drew attention to the potential for substantial movement by juveniles of these species during their marine phase. More recent ideas originating particularly in the work of Croizat (e.g. Croizat *et al.*, 1974) largely discounted contemporary dispersal, instead interpreting these circum-hemispheric distributions as the remnants of an ancient Gondwanan fauna. McDowall (2002) explores the case for and against these conflicting explanations, arguing that the overwhelming weight of evidence from both ecological and genetic studies supports the importance of contemporary dispersal. While he acknowledges that a role for tectonic plate movement cannot be ruled out, he proposes that the most parsimonious explanation for the extensive distribution of a species such as *Galaxias maculatus*, which occurs in Australia, New Zealand and its offshore islands, southern Chile and Argentina, and the Falkland Islands, is the widespread dispersal of its juveniles, which spend several months at sea before returning to spend their adult lives in fresh water.

As part of his case for the role played by dispersal, McDowall (2002) draws attention to the contrasting local distributions of New Zealand's diadromous and non-diadromous fish species. He argues that while the diadromous species are geographically widespread, occurring across the full range of latitudes that are potentially available to them, their non-diadromous counterparts have relatively restricted distributions, many of which show evidence of having been shaped by the effects of well-documented events in New Zealand's recent geological history. In this paper we use an extensive set of data describing the distributions of New Zealand's freshwater fish species and their environments to quantitatively examine these assertions. In particular, we contrast the geographical and environmental distributions of 30 species, 15 of which are diadromous, and the remainder non-diadromous. We also explore the manner in which the environmental niches of these species play out in geographic space, recognizing that the distributions of species, while largely reflecting competition-mediated responses to environment (Pulliam, 2000), are also subject to influence by processes such as disturbance and dispersal that may have localized effects geographically (e.g. Leathwick, 1998).

Analytically we extend our previous descriptions of the distributions of New Zealand's diadromous fish species (Leathwick *et al.*, 2005, 2006) by working with a substantially expanded set of distributional data, and including a mix of both diadromous and non-diadromous fish species. In addition, we use the relatively recently developed statistical modelling method of boosted regression trees (BRT) (Friedman, 2001, 2002). This machine-learning based technique provides significant gains in predictive performance over traditional regression methods through its ability to fit non-linear responses and its automatic detection and fitting of interaction effects between predictors (Leathwick *et al.*, 2006; Elith *et al.*, in press).

MATERIALS AND METHODS

Data

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 in 2006 held fish distribution records for approximately 22,500 sites throughout New Zealand. A subset of 13,369 records was selected for this analysis by including only sites sampled after January 1980, and for which all species were identified (Fig. 1). Approximately 75% of these records were sampled from 1990 onwards. Data were extracted for 15 diadromous and 15 non-diadromous species (Table 1) for which presences were recorded for at least 30 sites. A small number of data points from repeated visits to the same sites, usually sampling different habitats or carried out during different seasons, were given a weight equal to $1/n$ where n was the number of repeats. This prevented their lack of complete independence having an undue influence on model fitting and subsequent estimation of predictive performance. Sites sampled only once were given a weight of one. We also excluded sites from tidal rivers, lakes or other still waters, and sites sampled using methods appropriate for only some species or life stages (e.g. whitebait nets, plankton nets or diving). Although most sites selected for analysis were sampled by electric fishing (76%), other techniques used included fyke nets (7%), spotlighting (5%) and traps (6%), some of which were baited. Combinations of two or more techniques were used for a small proportion of samples (6%). Samples from smaller rivers and streams greatly outnumbered those from larger rivers, in part reflecting the reduced effectiveness of techniques such as electric fishing in water greater than 0.5 m depth (Minns, 1990; Chadderton & Allibone, 2000). Although records of fish abundances were available for many sites, all data were converted to presence–absence form for this analysis because of difficulties in correcting for differing catch rates by different capture methods and/or variations in the areas fished.

Environmental predictors

Review of the literature describing habitat requirements of freshwater organisms (e.g. Biggs *et al.*, 1990; Townsend & Hildrew, 1994; Poff, 1997; Lamouroux & Souchon, 2002) indicates that the most important environmental determinants of species occurrence at any site are likely to include factors such as the amount of water flow and its temporal variability, water temperature and chemistry, terrestrial inputs of energy, and river substrate size and sorting. The

Figure 1 Distribution of fish sample sites used in the analysis (open circles). Only rivers with mean annual flows greater than 10 $m^3 s^{-1}$ are shown.

Table 1 Six letter codes and scientific names of fish species included in the analysis, along with their number of presences and the number of catchments in which they occur. An asterisk is used to indicate diadromous species. Taxonomic authorities follow McDowall (1990) for all species except *Galaxias* 'species D' and *G.* 'species N', which are undescribed entities formerly included in *Galaxias vulgaris* (McDowall, 2006). Also shown are the number and median size of catchments occupied by each species.

Code	Species name	Presences	No. catchments	Median catchment size (km^2)
Angaus*	*Anguilla australis*	2670	554	408.7
Angdie*	*A. dieffenbachii*	6650	924	608.0
Chefos*	*Cheimarrichthys fosteri*	1250	273	705.9
Galano	*Galaxias anomalus*	63	7	5702.7
Galarg*	*G. argenteus*	387	158	169.6
Galbre*	*G. brevipinnis*	1453	387	620.8
Galcob	*G. cobitinis*	34	6	895.9
Galdep	*G. depressiceps*	170	5	5702.7
Galdiv	*G. divergens*	348	46	902.3
Galeld	*G. eldoni*	59	1	5702.7
Galfas*	*G. fasciatus*	1649	639	20.4
Galgol	*G. gollumoides*	78	13	1566.2
Galmac*	*G. maculatus*	1372	564	75.2
Galmar	*G. macronasus*	30	8	719.4
Galpau	*G. paucispondylus*	167	20	1649.8
Galpos*	*G. postvectis*	348	136	135.6
Galpro	*G. prognathus*	62	8	1649.8
Galpul	*G. pullus*	51	3	5702.7
Galspd	*G.* 'species D' – Clutha flathead galaxias	110	15	1040.3
Galspn	*G.* 'species N' – northern flathead galaxias	83	6	2388.6
Galvul	*G. vulgaris*	624	51	726.3
Geoaus*	*Geotria australis*	325	117	705.9
Gobbas	*Gobiomorphus basalis*	723	110	957.2
Gobbre	*G. breviceps*	1760	156	1003.5
Gobcot*	*G. cotidianus*	2182	508	440.0
Gobgob*	*G. gobioides*	159	122	10.9
Gobhub*	*G. hubbsi*	661	193	499.4
Gobhut*	*G. huttoni*	2313	662	82.0
Retret*	*Retropinna retropinna*	509	177	761.8
Rhoret*	*Rhombosolea retiaria*	78	40	540.0

volume of river flow is a critical factor determining biological composition, as, together with the river gradient, this determines both the water velocity and depth, factors shown to be strongly correlated with the local-scale geographic distributions of both fish (Lamouroux & Jowett, 2005) and invertebrate species (e.g. Lamouroux *et al.*, 2004). Flow variability is also important, and is influenced by the size and nature of the upstream catchment, its elevation relative to any particular segment, the permeability of its underlying rocks, and its rainfall, slope and vegetation cover (Jowett & Duncan, 1990). Variation in flow has been identified as an important determinant of the biological character of rivers and streams (Poff, 1997), and can successfully predict a range of biological features, including the distributions of individual fish species (Jowett, 1990; Poff *et al.*, 1997; Richter *et al.*, 2003) and community structure in fish (Poff & Allan, 1995), invertebrates (Quinn & Hickey, 1990; Death & Winterbourn, 1995) and periphyton (Biggs, 1995).

Environmental predictors were derived from data stored in a Geographic Information System (GIS) data base representing New Zealand's rivers and streams as a network topology, with each river or stream section between adjacent confluences represented by a unique segment. We used two variables to describe river flow and its variability at the segment scale: *SegLowFlow* describes the 7-day average mean annual low flow (Pearson, 1995), with a fourth-root transformation giving values that tend to be linearly related to water velocity (Jowett, 1998); and *SegFlowStability* describes the ratio of the mean annual low flow to the mean average flow, with high values indicating stable flow regimes and low values, fluctuating flows. The segment slope (*SegSlope*) is used as an indirect indicator of local topographically driven variation in river velocity, which in turn influences the development of local habitat differentiation. A square root transformation was used to compensate for varying rates of change in habitat conditions along the slope gradient, i.e. on steep slopes a given change in slope results in smaller change in habitat conditions than on sites with low slopes.

Water temperature has also been identified as an important habitat factor for many freshwater organisms (e.g. Richardson *et al.*, 1994), and was an influential predictor in previous analyses using a subset of these data (Leathwick *et al.*, 2005). In modelling species distributions, we would ideally use direct estimates of mean water temperature and its seasonal variation. Although these can be estimated by combining estimates

of river flow with climatic variables such as the incoming solar radiation, temperature, wind and humidity (e.g. Theurer, 1982), we currently lack the spatially and temporally extensive measurements of water temperature required to verify such an approach nationally. In this study we used instead the average daily air temperature in summer (January – *SegSumT*) and winter (July – *SegTSeas*), which are likely to approximate seasonal variation in river temperatures. As these two variables were highly correlated, we normalized winter temperatures relative to the summer temperatures (formula in Table 2). The resulting variable indicates for any site the deviation in winter temperature from that expected given the summer temperature; negative values indicating continental or seasonal climates, while positive values indicate more maritime climates.

Riparian vegetation has important effects on in-stream conditions through its provision of shade, its effects on bank stability, and its contribution of organic debris that provides in-stream cover, and sources of carbon and energy (e.g. Gregory *et al.*, 1991). In addition, lower temperatures are expected in

Table 2 Environmental variables used in the analysis.

	Mean and range
Segment scale predictors	
SegLowFlow – segment mean annual 7-day low flow ($m^3\ s^{-1}$), fourth root transformed, i.e. $(\text{low flow} + 1)^{0.25}$	1.092, 1.0 to 4.09
SegFlowStability – annual low flow/annual mean flow (ratio)	0.18, 0 to 0.58
SegSumT – summer air temperature (°C)	16.3, 8.9 to 19.8
SegTSeas – winter air temperature (°C), normalized with respect to *SegSumT*, i.e. $SegTSeas = \left(\left(\frac{W-\overline{W}}{\sigma_w}\right) - \left(\frac{S-\overline{S}}{\sigma_s}\right)\right) * \sigma_w$ where W is the winter temperature for a segment, $\overline{W}$ is the average winter temperature for all segments, σ_w is the standard deviation of winter temperature, S is the summer temperature, and so on	0.36, −4.2 to 4.1
SegShade – riparian shade (proportion)	0.41, 0 to 0.8
SegSlope – segment slope (°), square-root transformed (slope+1)	1.59, 1 to 5.5
SegN – log10 of nitrogen concentration (p.p.m.)	−0.19, −2.04 to 1.88
Reach-scale predictors	
ReachSubstrate – weighted average of proportional cover of bed sediment using categories of 1 – mud, 2 – sand; 3 – fine gravel; 4 – coarse gravel; 5 – cobble; 6 – boulder; 7 – bedrock	3.77, 1 to 7, 2347 missing values
ReachHabitat – weighted average of proportional cover of local habitat using categories of 1– still; 2 – backwater; 3 – pool; 4 – run; 5 – riffle; 6 – rapid; 7 – cascade	4.04, 1 to 7, 2284 missing values
Downstream predictors	
DSDist – distance to coast (km)	73.8, 0.01 to 432.8
DSAvgSlope – average slope (°), square-root transformed (slope+1)	1.19, 1 to 7.24
DSMaxSlope – maximum downstream slope (°)	7.9, 0 to 40.8
DSDam – presence of known downstream obstruction, mostly dams	Absent – 10,978; Present – 2391
DSDist2Lake – distance to a lake (km), where no lake present, set to 500, i.e. greater than the maximum river length	28.6, 0.04 to 129 (for lakes present)
Upstream/catchment scale predictors	
USAvgT – average air temperature (°C), normalized with respect to *SegJanT*	−0.38, −7.7 to 2.2
USRainDays – days/month with rainfall greater than 25 mm	18.0, 1.2 to 103.4
USSlope – average slope in the catchment (°)	14.3, 0 to 41.0
USNative – area with indigenous vegetation (proportion)	0.57, 0 to 1
USPhosphorus – average phosphorous concentration of underlying rocks, 1 = very low to 5 = very high	2.44, 1 to 5
USCalcium – average calcium concentration of underlying rocks, 1 = very low to 4 = very high	1.48, 1 to 4
USHardness – average hardness of underlying rocks, 1 = very low to 5 = very high	3.11, 1 to 5
USPeat – area of peat in catchment (proportion)	0.006, 0 to 1
USLake – area of lake in catchment (proportion)	0.002, 0 to 1
USGlacier – area of glacier in catchment (proportion)	0.003, 0 to 0.53
Fishing method	
Method – fishing method in five broad classes	Electric – 10,155; net – 986; spot – 634; trap – 763; mixture – 831
TCos, TSin – month of year fitted as sine and cosine transforms of the Julian day for the middle of the month in which fishing occurred	*TCos* – 0.26, −1 to 1 *TSin* – 0.18, −1 to 1

shaded streams compared with those of comparable size but exposed to heating by the sun (Rutherford *et al.*, 1997). We estimated the amount of riparian shading (*SegShade*) with the method described by Leathwick *et al.* (2005), in which a national, satellite image-based vegetation classification based on imagery captured during 1996/97 was used to identify riparian vegetation for each segment. The degree of shading of the water surface was then estimated from the river or stream size and the vegetation height expected given the summer temperature (see Leathwick *et al.*, 2005, for details). Estimates of stream nitrogen load (*SegN*) were based on a leaching model combined with a regionally-based regression model, implemented within a catchment framework (Woods *et al.*, 2006).

Local (reach-scale) habitat variability reflects both variation in the size and nature of sediments being transported within a river system, and effects of the flow regime on its degree of sorting (e.g. Jowett & Duncan, 1990; Rosgen, 1994), and can have important effects on fish distribution (e.g. Gregory *et al.*, 1991). Effects of local habitat variation on fish occurrence were assessed using estimates at each site of the modal sediment size (*ReachSubstrate*) of the river or stream bed, and of the predominant local habitat (*ReachHabitat*). The first of these variables was created by converting field descriptions of the relative cover of different sized sediments (see Table 2) into a single ordinal value that indicates the weighted average of the cover contributed by each sediment class. Similarly, for *ReachHabitat* we calculated a single variable describing the relative proportions of pools, runs, riffles, etc., as described in Table 2.

Several variables were used to summarize aspects of the river segment environment that are affected by conditions in the upstream catchment. The likely frequency of elevated flows was indicated by a variable (*USDaysRain*) that describes the frequency of days with significant rainfall (> 25 mm) in the upstream catchment. The variable describing average temperatures in the upstream catchment (*USAvgT*) was normalized with respect to the segment summer air temperature, because of their high correlation. Since temperature and elevation are strongly correlated, this variable also describes elevation differences between a segment and its upstream catchment. Negative values indicate rivers and stream segments for which the upstream catchments have colder temperatures (higher elevations) than average. This implies steeper upstream river gradients with the consequent higher water velocities resulting in colder temperatures (there is less time for water temperatures to equilibrate to ambient air temperatures) and an abundance of energy for sediment transport. Positive values indicate rivers and streams with warmer, lower elevation catchments than average; water temperatures in these are likely to be more closely in equilibrium with segment-scale air temperatures and less energy will be available for sediment transport, resulting in greater deposition of fine sediments. Variables describing catchment-driven modification of flow variability included the upstream average slope (*USAvgSlope*), and the proportion of land covered with native vegetation cover (*USNative*) (Jowett & Duncan, 1990). Variation in geological substrates, which affects both flow variability and water chemistry, was quantified with estimates of rock hardness (*USHardness*), and availability in surface rocks of phosphorus (*USPhosphorus*) and calcium (*USCalcium*) as described in Leathwick *et al.* (2005). Two variables were used to describe the local buffering of river flows in the upstream catchment by lakes (*USLake*) and wetlands (*USWetland*), with wetlands potentially also altering water chemistry by lowering the pH (Collier *et al.*, 1990). Finally, a variable was used to indicate the proportional cover of glaciers in the upstream catchment (*USGlacier*), these altering both water temperature and the nature of sediments, with strong consequent effects on biological communities (e.g. Castella *et al.*, 2001).

We used four variables to describe variation in factors likely to affect the ability of diadromous fish to move to/from each segment to the sea. The downstream distance to the coast (*DSDist*) was derived using a GIS script to calculate the downstream network distance from the mid-point of each river segment to the coast. The average downstream slope (*DSAvgSlope*) was calculated from the elevation at the mid-point of the segment and the distance from the mid-point to the coast. Similarly, *DSMaxSlope* described the maximum local slope occurring between the midpoint of each segment and the coast. To quantify this, we first calculated local slopes at 100-m intervals along each river segment, and then traversed from each segment along the river network to the coast to identify the maximum. A categorical variable (*DSDam*) indicated river segments with known downstream obstructions such as dams, culverts or waterfalls that were likely to impede fish passage. A fifth and related variable (*DSDist2Lake*) was calculated for segments occurring upstream of any lakes located on the river network, allowing us to test the importance of lakes as breeding habitat for diadromous species such as *Galaxias brevipinnis* and *Gobiomorphus cotidianus*. This variable indicated the distance from the midpoint of such segments to the nearest downstream lake.

Three variables were used to describe features of the fishing method. A categorical variable (*Method*) was used to distinguish between four major capture methods (i.e. electric fishing, nets, spotlights and traps) with a fifth level used to identify fishing using some combination of these methods. Two variables (*TSin* and *TCos*) were used to assess the effects of variation in the time of year at which fishing occurred, following the method of Flury & Levri (1999) for analysis of periodic data. Trigonometric transforms were used to fit the time of year so that fitted occurrences at the end of December match those at the beginning of January.

Analysis of geographic ranges

We used several statistics to explore differences in the geographic range characteristics of diadromous and non-diadromous species. We first used a GIS to calculate the number and median size of catchments recorded as occupied by each species. For this analysis, rivers having a maximum order of six and above were split into their sixth order sub-

catchments, plus the main river stem, while catchments for rivers of fifth order or less were kept intact. We then calculated two measures of geographic range size for each species by computing alpha-hulls (Burgman & Fox, 2003) enclosing the *x* and *y* coordinates of its recorded presences. An alpha-hull is similar to a convex hull, consisting of a complex enclosing polygon that is constructed by joining together delaunay triangles connecting the locations of all the presence records. However, in constructing an alpha-hull, triangles having a mean (or maximum) side length longer than some nominated maximum length are removed, resulting in the production of a more complex enclosing polygon that more tightly encloses the underlying data points than would be constructed using a convex hull. For each species we calculated the area of alpha-hulls constructed using maximum triangle segment lengths of 200 km ($a\text{-}hull_{200}$) and 50 km ($a\text{-}hull_{50}$). These distances were selected by taking into consideration the geographic extent of New Zealand and typical catchment sizes, and therefore reflect the likely biological implications of geographic spread with and without migration. The larger alpha-hull indicates the broad limits of the geographic range, while the smaller provides a more conservative estimate of the actual area occupied within the broader geographic range (see examples in Appendix S1 in Supplementary Material).

Statistical modelling

Relationships between fish occurrence and environment were analysed using boosted regression trees (BRT), a type of regression model, technical details for which are provided in Appendix 1. Only relevant analysis settings are provided here. All analyses were carried out in R (version 2.0.1, R Development Core Team, 2004) using the 'gbm' library of Ridgeway (2004), supplemented with our own functions (Elith *et al.*, in press). All models were fitted to allow interactions, using a tree complexity of 5 and mostly using a learning rate of 0.01 – a learning rate of 0.005 was used for three species of very low prevalence. Ten-fold cross validation was used to determine the optimal number of trees for each model, i.e. that giving the maximum predictive performance – for most species this was between 1000 and 4000 trees (see Appendix S2). Because of the tendency of BRT models to over-fit the training data, the performance of all models was not assessed on the training data, but on predictions to sites that were with-held during cross-validation. Two values were calculated for each model: the predictive deviance, and the discrimination as measured by the area under the receiver operator characteristic curve (AUC) (Hanley & McNeil, 1982). Values for the predictive deviance provide a measure of the goodness-of-fit between predicted and raw values when predicting to independent data, and we expressed this as a percentage of the null deviance for each species. Values for AUC give a measure of the degree to which fitted values discriminate between observed presences and absences; values can be interpreted as indicating the probability that a presence for a species drawn at random from the data will have a higher fitted probability than an absence drawn at random. A value of 0.5 indicates that a model predicts presences and absences no better than a simple random allocation of probabilities, while a value of 1.0 indicates that presences and absences are perfectly distinguished.

The relative importance of the individual predictors was determined using a script in the 'gbm' library that sums, by predictor, reductions in error across all the individual regression tree rules. The influence of interactions between predictor variables was evaluated using purpose-written functions as described in Appendix 1. Environmental optima for each species were determined by plotting the distributions of fitted values in relation to each of the predictors.

RESULTS

Geographic ranges

Analysis of geographic range characteristics based on the fish sample sites indicates marked differences between the two species groups (Table 1, Fig. 2). For example, average broad-scale geographic range sizes for diadromous species were more than seven times larger than for their non-diadromous counterparts (horizontal axis in Fig. 2, mean $a\text{-}hull_{200}$ = 244,377 km^2 vs. 33,189 km^2, $F = 59.8$, $P < 0.001$). Differences in the average fine-scale geographic range sizes for these two groups of species were also significant but more muted (mean $a\text{-}hull_{50}$ = 65,498 km^2 vs. 15,875 km^2, $F = 8.62$, $P < 0.01$). There were also clear patterns in the ratios of fine-scale to coarse-scale range sizes exhibited by the two species groups (vertical axis in Fig. 2), these indicating the relative degree of range occupancy. In particular, for a large proportion of

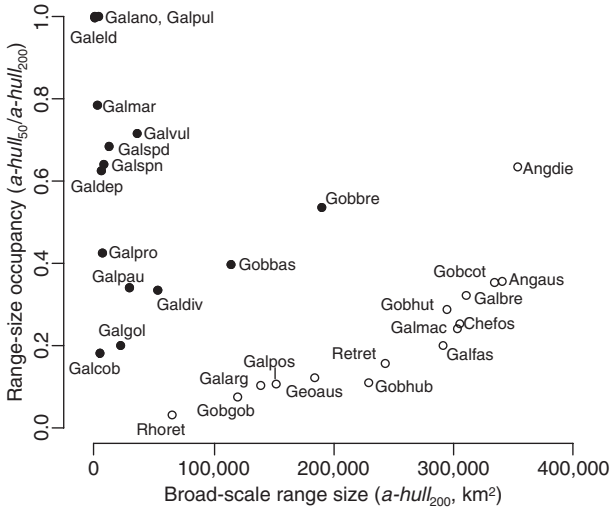

Figure 2 Broad-scale range size and occupancy for diadromous and non-diadromous fish species as estimated from geographic areas of alpha-hulls for each species calculated using threshold values of 50 and 200 km. Open circles are used to indicate diadromous species, and solid circles to indicate non-diadromous species. Six letter codes and their equivalent full scientific names are listed in Table 1.

non-diadromous species their fine-scale geographic ranges are more than half of their broad-scale ranges (upper left of Fig. 2), indicating relatively compact geographic distributions with a low frequency of outliers. By contrast, diadromous species such as *Rhombosolea retiaria*, *Galaxias argenteus*, *Galaxias postvectis* and *Geotria australis* have only scattered occurrences over large geographic areas (lower centre and lower right of Fig. 2). Finally, non-diadromous species occurred in far fewer catchments than their diadromous counterparts (Table 1; 30.3 vs. 363.6, $F = 23.5$, $P < 0.001$), and the average of the median catchment sizes occupied by non-diadromous species (2421 km^2) was more than six times the average for those occupied by diadromous species (388 km^2, $F = 13.88$, $P < 0.001$).

Summary of statistical models

On average, the predictive deviance estimated using 10-fold cross validation was substantially higher for non-diadromous species (60.8%) than for diadromous species (36.8%), and average AUC scores for the non-diadromous species (0.981) exceeded those for the diadromous species (0.914) (Table 3, Appendix S2). Inspection of the contributions of predictor variables to the individual species models indicates that access related variables (*DSAvgSlope* and *DSDist*) are the strongest correlates of diadromous distribution, while climatic variables (*SegTSeas* and *USRainDays*) are the strongest correlates of non-diadromous distributions (Table 4, Appendix S3). Despite this marked difference, four out of the five most important predictors for both groups are shared in common. *DSDist* is ranked third for non-diadromous species. *SegSumT* is the third ranked predictor for diadromous species and fifth for non-diadromous, and *DSMaxSlope* is ranked fourth for both groups.

Environmental relationships of fish species

Functions fitted by the BRT models were highly variable in shape, and were frequently non-linear, as illustrated by those fitted for the four most important predictors for *Galaxias postvectis* (Fig. 3, see Appendix S3 for functions for all species). In this example, the fitted functions indicate that *Galaxias postvectis* is caught most frequently in river segments combining a high frequency of rainfall in the upstream catchment, distances to the coast of around 20 km, and coarse substrates. The coefficient fitted for the factor predictor *Method* indicates that this species is detected most frequently using spotlighting. The fitted functions tend to be smoother in those parts of the range of each variable where data densities are higher (as indicated by the decile ticks in Fig. 3), but may become more complex or noisy in those parts of the environmental space where sample densities are low (e.g. over the range from *c.* 40–70 for *USRainDays* in Fig. 3).

Direct interpretation of these functions for some environmental predictors is complicated by interaction effects fitted by the BRT models. The interaction between *DSDist* and *DSAvgSlope* had the highest median value for diadromous species (Table 5), and strong interactions were also evident between *DSAvgSlope* and both *ReachSubstrate* and *SegTSeas*. *SegTSeas* was involved in all of the five most important interaction effects for non-diadromous species. The presence of such interactions results in sometimes marked variation in the response fitted between species catch and a particular predictor, depending on the value taken by a second predictor. For example, for *Gobiomorphus huttoni* (Fig. 4a), there is a pronounced peak of occurrence at sites that combine both low downstream average slopes and proximity to the coast. This species does penetrate some distance inland, but only when river gradients remain gentle, and also penetrates steeper gradient streams, provided that they are close to the coast. Interactions fitted for the non-diadromous *Gobiomorphus breviceps* indicates that this species is caught most frequently at sites with low upstream frequencies of days with intense rainfall, but that this response is strongly affected by

Table 3 Comparison of the predictive performance of statistical models relating species distributions to environment, averaged across (a) diadromous species, and (b) non-diadromous species. Table values indicate the mean of the predictive deviances and their standard errors (SE), the mean percentage of the total deviance explained, and the mean of the AUC scores and their standard errors. Estimates of predictive performance were calculated using 10-fold cross validation.

	Diadromous	Non-diadromous
Null deviance	0.500	0.150
Predictive deviance (SE)	0.312 (0.005)	0.068 (0.003)
Deviance explained (%)	36.5	58.2
AUC (SE)	0.916 (0.005)	0.983 (0.003)

Table 4 Average percentage contributions of the ten most important predictors for BRT models relating the distributions of diadromous and non-diadromous species to environment. Predictors are ranked by their maximum importance across both groups, with values in brackets indicating rank for the ten most important predictors for each group.

Predictor	Diadromous	Non-diadromous
DSAvgSlope	11.9 (1)	4.4 (9)
DSDist	10.6 (2)	8.2 (3)
SegTSeas	3.9 (8)	9.8 (1)
USRainDays	6.5 (5)	8.5 (2)
SegSumT	7.7 (3)	6.1 (5)
DSMaxSlope	7.2 (4)	6.8 (4)
SegFlowStability	3.7 (9)	5.3 (6)
Method	5.1 (6)	0.6
USCalcium	2.1	4.8 (7)
ReachSubstrate	4.6 (7)	3.4
USNative	2.3	4.6 (8)
USPhosphorus	2.9	4.2 (10)
SegLowFlow	3.6 (10)	2.1

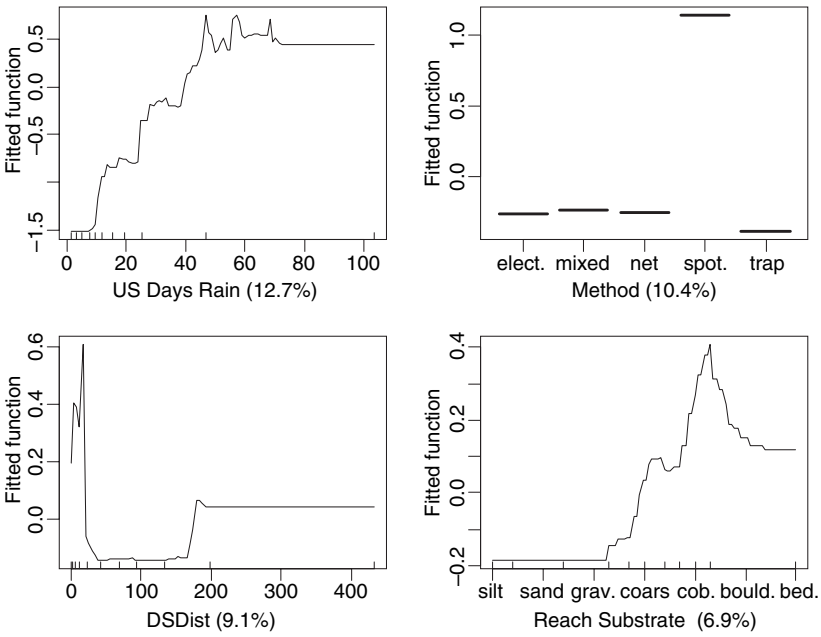

Figure 3 Functions fitted for the four most important predictors by a boosted regression trees (BRT) model relating the probability of occurrence of *Galaxias postvectis* to environment. Functions are continuous for the three continuous variables, but consist of discrete values for each level of the factor predictor, *Method*. Ticks across the bottom of each plot show the distribution of deciles for each continuous predictor variable. A common scale is used on the vertical axis for all plots.

Table 5 The five pairwise interactions between predictor variables having the highest median contributions to boosted regression trees (BRT) models relating the distributions of diadromous and non-diadromous species to environment. Table entries indicate the relative degree of departure from a purely additive effect, with a value of zero indicating that no interaction is present.

Diadromous	Non-diadromous
DSDist & *DSAvgSlope* – 4.0	*SegTSeas* & *SegLogN* – 4.4
SegFlowStability & *USDaysRain* – 3.6	*SegTSeas* & *USDaysRain* – 3.4
SegSumT & *DSMaxSlope* – 3.6	*SegTSeas* & *DSDist* – 3.1
DSAvgSlope & *ReachSubstrate* – 3.2	*SegSumT* & *SegTSeas* – 2.4
SegTSeas & *DSAvgSlope* – 3.0	*SegTSeas* & *SegFlowStability* – 2.3

temperature seasonality, with high capture rates only occurring at sites with cool winters (Fig. 4b).

Examination of the distributions of fitted values in relation to the dominant predictors indicates often marked separation in the environments occupied by diadromous and non-diadromous species. For example, all diadromous species show a marked bias towards coastal rivers and streams. Approximately two-thirds are caught most frequently within 30 km of the coast, and only three are caught most frequently at distances of 50 km or more (horizontal axis in Fig. 5a). By contrast, only one non-diadromous species, *Galaxias cobitinis*, is predicted to occur most frequently in more coastal rivers, the majority of this group being caught most frequently at sites between 70 and 200 km inland. Sorting is also evident in relation to the average down-stream river gradient (vertical axis in Fig. 5a), with diadromous species separated into those occurring either in low-gradient streams with minimal barriers to upstream movement (e.g. *Anguilla australis*, *Cheimarrichthys fosteri*, *Galaxias maculatus*, *Gobiomorphus cotidianus*, *Retropinna retropinna* and *Rhombosolea retiaria*), and a smaller number of capable climbers that preferentially occur in steep gradient streams in both coastal (e.g. *Galaxias fasciatus*) and more inland (*Galaxias brevipinnis*) locations. By contrast, approximately half of the non-diadromous species occur most frequently in rivers that are well above the line indicating the average relationship between average down-stream slope and distance from the sea.

Similar separation of these two species groups is also evident with respect to segment-scale temperatures (horizontal axis in Fig. 5b). The majority of diadromous species are caught most frequently at sites with warm summer temperatures (16°C or more), and all 15 species are caught most frequently where temperature seasonality is muted, i.e. maritime climates in which winters are milder than expected, given the summer temperature (upper right of Fig. 4b). By contrast, the majority of non-diadromous species are predicted to be caught most frequently on sites with mild to cool summers (*c.* 16°C or below), and more than half of these species are caught most frequently at sites with more continental climates, i.e. with cold winters. The marked exception to this general pattern is the non-diadromous *Gobiomorphus basalis*, which occurs on sites that combine warm summer temperatures and mild winters, suggesting a much lower tolerance of cold than that shown by other non-diadromous species.

Although there is little apparent difference in the distributions of diadromous and non-diadromous species in relation to flow stability (horizontal axis in Fig. 5c), optima for non-diadromous species occur over a wider range of conditions than those for diadromous species. For example, non-diadromous species, such as *Galaxias macronasus*, *Galaxias paucispondylus* and *Galaxias prognathus*, are caught most frequently in rivers with stable flows (right of Fig. 5c),

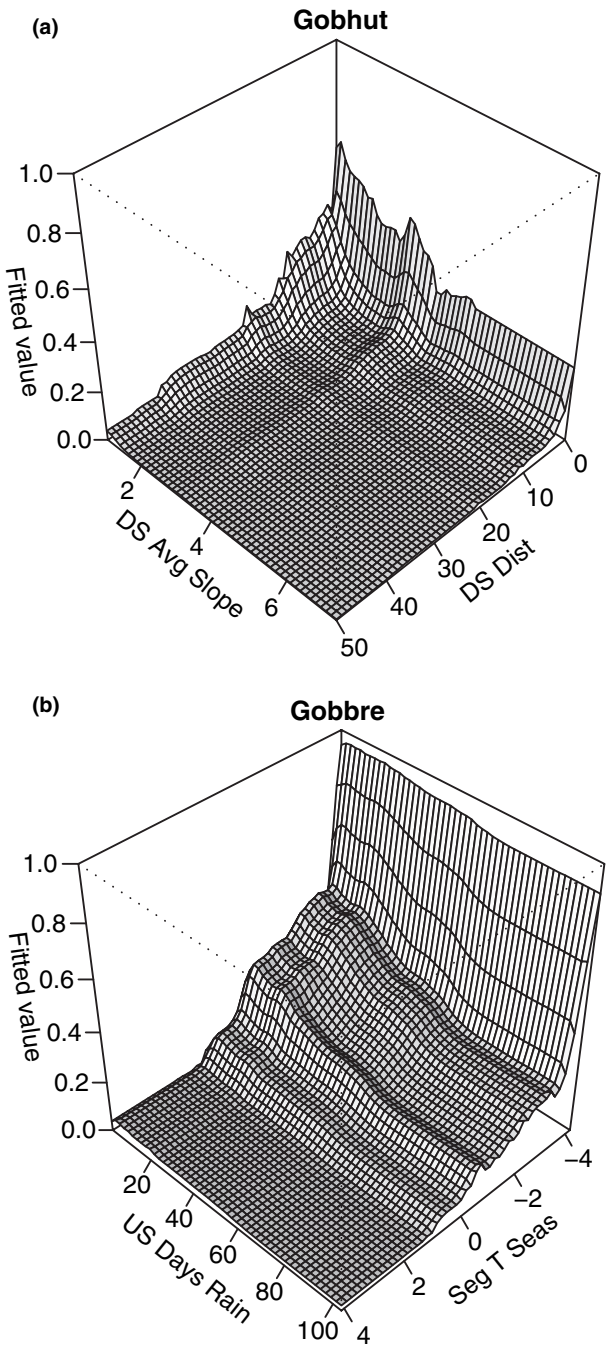

Figure 4 Typical effects of interactions between predictors on predicted probabilities of occurrence predicted for species. (a) Gobhut, *Gobiomorphus huttoni*; (b) *Gobbre, G. breviceps*.

while *Galaxias anomalus* and *Galaxias gollumoides* are caught most frequently in rivers and streams with stronger seasonal variation in flow. More marked separation between the two species groups is apparent in relation to the frequency of major upstream rain events (vertical axis in Fig. 5c), with most non-diadromous species occurring in environments in which such events are relatively infrequent (*c.* 10 per year or less). By contrast, most diadromous species are caught most frequently in catchments in which high intensity rain events occur with moderate frequency, and *Galaxias argenteus*, *Galaxias postvectis* and *Galaxias brevipinnis*, along with the non-diadromous *Galaxias prognathus*, reach their maximum levels of occurrence in catchments with a high frequency of intense rainfall events (> 30 year^{-1}). Finally, there is wide separation between species with respect to variation in sediment size and upstream average slope (Fig. 5d). Rivers and streams with fine sediments, generally those with low average upstream slopes, and in the lower elevation parts of catchments, are mostly occupied by diadromous species, with four of these caught most frequently here. By contrast, a mix of diadromous and non-diadromous species dominates river segments typified by gravels or coarser substrates and catchments with moderate to steep slopes. Non-diadromous species occur in catchments with widely varying combinations of slope and sediment size, including streams with finer sediments but steep catchments (*Galaxias prognathus* and *Galaxias paucispondylus*, upper left of Fig. 5d), and those with gently sloping catchments and coarse gravels, often of glacial origin (*Galaxias depressiceps*, *Galaxias eldoni* and *Galaxias pullus*, lower right of Fig. 5d).

Preferences in relation to less important predictors are diverse, and we provide only a brief summary for variables and species for which contributions to model outcomes are greater than 5% (Appendix S4). *DSMaxSlope* was important for 15 species, with results contrasting strongly with those for *DSAvgSlope*. In particular, most non-diadromous species occur most frequently in rivers and streams in which maximum downstream slopes do not exceed 4–5 degrees, while nearly half the non-diadromous species occur most frequently in rivers and streams in which maximum downstream slopes are 7–8 degrees or more. *SegSlope* was important for four species, and *SegN* was important for eight species. Predictors describing differences in composition of underlying catchment rocks were only important for non-diadromous species. *Method* was important for five species, four of which, *Anguilla dieffenbachii*, *Galaxias argenteus*, *Galaxias postvectis* and *Gobiomorphus gobioides*, are detected more frequently using spotlighting.

DISCUSSION

The contrasting biogeographies of diadromous and non-diadromous species

Our analysis provides strong numerical support for earlier arguments in favour of the important role that differences in dispersal ability play in determining the biogeographies of New Zealand's diadromous fish species (McDowall, 1993, 1999, 2002), and in broader terms, is strongly consistent with the wider reinstatement of the importance of dispersal in understanding the biogeography of oceanic islands (e.g. McGlone, 2005; Heaney, 2007). In particular, our results support the proposition that the superior dispersal ability of diadromous fish results in them having geographical (and environmental) distributions that are markedly different to those of the

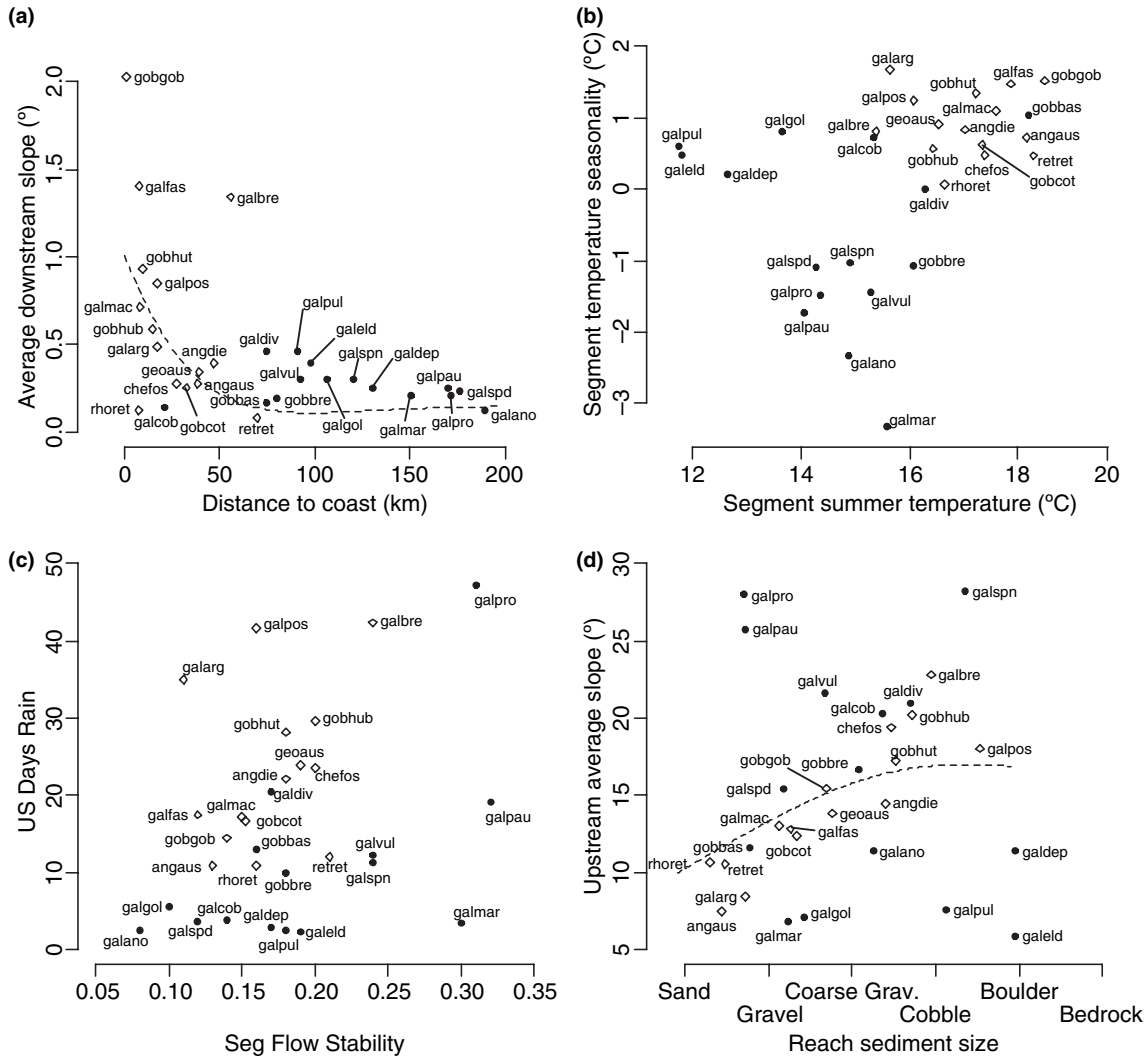

Figure 5 Environmental optima in relation to major environmental predictors as determined from model fitted values. Diadromous species are plotted using open diamonds, and non-diadromous species using solid circles. For parts a and d, dashed lines indicate the average relationship between the two environmental predictors as described by the fish samples.

non-diadromous species, whose potential for inter-catchment movement is, at best, severely constrained. In geographic terms, the majority of *diadromous* species have extensive distributions that encompass many catchments, varying widely in size. Many show relatively scattered local-scale distributions within their broader scale geographic ranges, indicating that their preferred environments have been found and occupied across a wide geographic area. The importance of marine dispersal is also emphasized both by their strong bias towards coastal locations, and by the major contributions of the two downstream slope variables, with interaction terms indicating that most diadromous species penetrate further inland where geographic barriers to upstream movement are minimal. By contrast, most of the *non-diadromous* species occur over relatively limited geographic ranges and occur mostly in large catchments. Most have a relatively compact geographic range, with minimal numbers of outlier records that might, if present indicate occasional long-distance dispersal.

While the importance of distance from the coast and downstream slope as predictors for non-diadromous species was initially counter-intuitive, their distributional bias towards the inland locations with above-average downstream slopes suggests that they persist more successfully in locations where there is less competition from diadromous species. This mechanism was invoked by McDowall & Allibone (1994) to explain the loss of non-diadromous *Galaxias vulgaris* in the upper Waipori River catchment, where increased inland penetration by *Galaxias brevipinnis* was facilitated by human impoundment that enabled its inland lacustrine recruitment. These patterns are perhaps also amplified by negative interactions between many non-diadromous species and the diadromous introduced brown trout (Townsend & Crowl, 1991; McIntosh & McDowall, 2004). Some non-diadromous species now only persist in headwater catchments that brown trout (*Salmo trutta*) have been unable to colonize.

Marked differences are also apparent in the environmental distributions of these two groups of species. For example, *diadromous* species occur predominantly in climates typified by warm summers and/or mild winters, and a moderate to high frequency of days with substantial rainfall. Many show a strong preference for rivers and streams with intact riparian shade (Hanchet, 1990), and they dominate rivers with fine sediments, most of which occur in low-gradient, coastal locations that are easily accessible from the sea. By contrast, *non-diadromous* species occur most frequently in climates typified by cool summers and/or with cool winters, and more than half occur in climates with very low frequencies of significant rainfall events. Many are caught most frequently in rivers and streams with little shade and gravelly substrates, and whose catchments have low native vegetation cover.

Determining the functional significance of these environmental differences is complicated by correlations between distance from the coast and some environmental and anthropogenic factors. For example, there are consistent differences between coastal and inland climates, with the coastal locations that are most available to dispersing juvenile diadromous species generally having warmer climates and/or more muted seasonal temperature variation than the more inland climates generally occupied by non-diadromous species. Caution should therefore be applied to any interpretation of the bias of diadromous species towards warmer and/or less seasonal climates as a climatic preference. This difference may simply reflect the mostly obligate requirements of these species for access to marine environments for completion of their life cycles, and for all but the most climbing-capable and stronger swimming species, this implies at least some restriction to coastal environments. Alternatively, inland penetration by diadromous species might be partially restricted where non-diadromous species are present in the upper reaches of larger rivers.

Interpretation of distributional differences with respect to catchment-scale frequencies of heavy rainfall events and variability in river flows is less ambiguous, as both these variables shows less correlation with distances from the coast. The bias of non-diadromous species towards catchments with low frequencies of intense rainfall (and consequent higher flow variability) may reflect a bias towards environments in which there is a reduced risk of displacement during floods, particularly for juveniles. Such conditions would favour successful in-stream recruitment and population persistence of species with limited ability to recolonize any catchments from which they are removed by floods. Alternatively, populations of non-diadromous species may experience less competition from introduced salmonids (e.g. McIntosh *et al.*, 1992) in these environments, given the lower abundance of the latter in rivers and streams with more variable flows (Jowett, 1990). By contrast, the diadromous species generally occur most frequently in rivers and streams subject to recurrent flushing by intense rainfall events, and in these there is presumably less risk to the persistence of these species, given the ability of marine life stages to reoccupy any catchments from which they might occasionally be displaced. For some species, the high river flows produced by such events may even provide important cues by triggering the hatching of juveniles and aiding their dispersal to the sea (Charteris *et al.*, 2003). However, it is possible that diadromous species were once more common in the lower rainfall environments such as in the eastern South Island, but have been more reduced there than in the west because of the more intensive agricultural development that is most common in lower rainfall environments.

The role of history

Although this analysis has focused on contemporary distributions, consideration of geographic variation in historic physical disturbance is also important for understanding the distributions of these two species groups, particularly given the dynamic nature of New Zealand's landscapes (Newnham *et al.*, 1999). Although this topic has been explored extensively in other branches of ecology (e.g. White, 1979; Sousa, 1984), it has received less attention as a factor explaining the distributions of fish species (although see, for example, McDowall, 1996, 2002).

The likely importance of historical disturbance is most evident in the distributions of non-diadromous species, and particularly their restricted geographic ranges that are biased towards large catchments in the eastern South Island. Their virtual absence in the central North Island has already been described by McDowall (1996) who attributed this to the effects of a massive rhyolitic eruption from the Taupo Volcanic Centre approximately 1800 years ago. While this event would undoubtedly have affected any fish species inhabiting this region, the current absence of non-diadromous species may have even more distant origins, given that rhyolitic volcanoes there have ejected at least 10,000 km^3 of magma over the last 2 Myr (Newnham *et al.*, 1999).

A parallel explanation probably also applies in the South Island, where the strong distributional bias of non-diadromous species to catchments east of the Southern Alps and their virtual absence in the west can probably be linked to historic patterns of glaciation. While piedmont glaciers extended to coastal elevations along most of the middle and southern parts of the west coast of the South Island during the last glacial maximum, in the east glaciers were largely confined to alpine regions adjacent to the Southern Alps (e.g. Newnham *et al.*, 1999). Extensive non-glaciated areas occurred along the east coast of the South Island, and many of the rivers draining the lower elevation eastern hills would have been substantially less directly affected by glaciation than rivers west of the Alps. The existence of these more extensive, largely ice-free lowlands, coupled with large catchment sizes, probably favoured survival of non-diadromous species there, despite the undoubtedly harsh climate and extensive areas of unstable glacial gravels.

The relative lack of non-diadromous species in the north-west of the South Island is unlikely to reflect displacement by glaciation, given its limited historic extent there. One potential

explanation is that the relatively high frequency of days with intense rainfall here makes conditions unsuitable for them. Another possible explanation is that non-diadromous species are unable to survive the in-stream impacts of the large magnitude earthquakes that occur periodically in this region, a consequence of movement along the major tectonic plate boundary that runs through this region (e.g. Suggate, 1982). For example, in 1929 a M 7.7 earthquake initiated major landscape instability that produced an average debris yield of 210,000 m^3 km^{-2} across a 1200 km^2 survey area (Pearce & O'Loughlin, 1985). Landslides initiated by these large-magnitude earthquakes occasionally form lakes by damming water courses, and some of these persist for lengthy periods until in-filled by alluvial material (Adams, 1981); major disruption may also occur in water courses below such impoundments as a result of mudflows formed when the landslide is overtopped by the pent-up water.

Although the mechanisms of species displacement might vary from region to region, it would appear that the relatively limited dispersal ability of non-diadromous species (a direct reflection of their lack of a marine phase) severely limits their ability to reoccupy sites if they are displaced by disturbance. Some evidence is available of limited dispersal between catchments by non-diadromous species, mostly through headwater capture events where geological 'accidents' result in the diversion of headwater tributaries from one catchment to another, taking with them any upstream residents (e.g. Waters & Wallis, 2000). By contrast, the majority of diadromous species occur consistently in rivers and streams of widely varying size and throughout New Zealand. Their distributional patterns show strong and consistent sorting with respect to environment, and they exhibit very few if any apparent geographic disjunctions, once environmental differences are taken into account. The simplest explanation of these features is that these species are no less prone to extinction than their non-diadromous counterparts, but that their capacity for marine dispersal endows them with a much greater resilience in the face of the large-scale, catastrophic physical disturbances that are a recurrent feature of many New Zealand landscapes. This conclusion provides strong support for Heaney (2007) who, in arguing for the need for a new paradigm with which to explain island biogeography, asserts that even superficially similar groups of species can exhibit widely varying degrees of dispersal (and subsequent gene flow), dependent on their ecology.

Contemporary disturbance is probably also implicated in some of these distributional patterns, reflecting the effects of extensive clearance of native vegetation and consequent loss of riparian shade, changes in sediment and nutrient inputs with agricultural development, blocking of upstream access by dams and culverts, and alteration of competition through the introduction of alien fish species. However, it is difficult to make robust inferences about these effects without reliable time series data, and the opportunities for collecting this are now largely past. Our analysis shows that some species (*Anguilla australis* and *Galaxias maculatus*) have apparently good tolerance of the high nitrogen inputs and/or low shade typical of rivers and streams flowing through agricultural landscapes. It is possible that species such as *Anguilla australis* are favoured by these changes, either directly, or through the decline of former competitors that are less tolerant of environmental change (Rowe *et al.*, 1999). However, declines, particularly along New Zealand's east coast, in the whitebait fisheries that are based largely on the juveniles of *Galaxias maculatus* (McDowall, 1990) would suggest that the negative impacts of contemporary landscape changes on this species have been considerable.

Finally, our descriptions of the environmental preferences shown by species that are less-tolerant of these human-induced changes may not accurately represent their historic realized niches, but merely the reduced niche spaces in which they have been able to survive following pervasive human alteration of New Zealand's landscapes and introduction of invasive fish species, particularly since the arrival of European settlers in the mid-1880s. Despite these uncertainties, this analysis provides an important context for managing the long-term persistence of these species, both in its detailed descriptions of the environmental preferences of individual species, and in its ability to provide predictions of the expected composition of fish communities in rivers and streams not yet sampled. Current work is exploring the application of these results for the development of a systematic approach to long-term conservation management in New Zealand's rivers and streams.

Statistical issues

Results from this analysis accord strongly with those from our previous numerical analyses of the distributions of New Zealand's diadromous fish species (Leathwick *et al.*, 2005) using multivariate adaptive regression splines (MARS). However, we have achieved substantial gains through use of an expanded set of distributional data, and extension of the analysis to include non-diadromous species. In addition, use of boosted regression trees gave added insight, elucidating the importance of interactions between predictors for identifying the optimal habitats of species, and more accurately identifying the relative contributions of the environmental predictors and species responses to them. In terms of model performance, these factors resulted in substantially higher AUC scores for the diadromous species common to both analyses, and which averaged 0.91 here, compared with an average of 0.84 for the MARS models used previously. In the current study, interpretation of the apparently higher predictive performance of our models for non-diadromous species compared with their diadromous counterparts is more problematic. Rather than indicating that the models for these species have superior predictive performance, we consider it more likely that the deviance and AUC for these models are particularly high because of the very low prevalence of these species (Bio, 2000; Elith *et al.*, 2006). That is, the apparently high performance of these models will be influenced in some measure by their success at predicting the absences which predominate in the data for these species.

ACKNOWLEDGEMENTS

J.L.'s contribution to this research was funded by New Zealand's Foundation for Research, Science and Technology under contract C01X0305. J.E. was funded by ARC grant DP0772671 and by the Australian Centre of Excellence for Risk Analysis. T.H. was partially supported by grant DMS-0505676 from the US National Science Foundation. The research drew heavily on environmental data layers developed during contracts funded by New Zealand's Ministry for the Environment and Department of Conservation. Substantial benefit came from discussions with colleagues including C. Baker, I. Jowett, R. McDowall, A. McIntosh, S. Parkyn, J. Quinn, J. Richardson and P. Taylor. J. Richardson extracted data from the New Zealand Freshwater Fish Database, a collection of data for which she played a key custodial role over many years.

REFERENCES

Adams, J. (1981) Earthquake-dammed lakes in New Zealand. *Geology*, **9,** 215–219.

Biggs, B.J.F. (1995) The contribution of disturbance, catchment geology and land use to the habitat template of periphyton in stream ecosystems. *Freshwater Biology*, **33,** 419–438.

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

Bio, A. (2000) *Does vegetation suit our models? Data and model assumptions and the assessment of species distribution in space*. PhD Thesis, Utrecht University, Utrecht.

Burgman, M.A. & Fox, J.C. (2003) Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conservation*, **6,** 19–28.

Cappo, M., De'ath, G., Boyle, S., Aumend, J., Olbrich, R., Hoedt, F., Colton, P. & Brunskill, G. (2005) Development of a robust classifier of freshwater residence in barramundi (*Lates calcarifer*) life histories using elemental ratios in scales and boosted regression trees. *Marine and Freshwater Research*, **56,** 713–723.

Castella, E., Adalsteinsson, H., Brittain, J.E., Gislason, G.M., Lehmann, A., Lencioni, V., Lods-Crozet, B., Maiolini, B., Milner, A.M., Olafsson, J.S., Saltveit, S.J. & Snook, D.L. (2001) Macrobenthic invertebrate richness and composition along a latitudinal gradient of European glacier-fed streams. *Freshwater Biology*, **46,** 1811–1831.

Chadderton, W.L. & Allibone, R.M. (2000) Habitat preferences and distributional patterns of native fish from an unmodified Stewart Island (New Zealand) stream. *New Zealand Journal of Marine and Freshwater Research*, **34,** 487–499.

Charteris, S.C., Allibone, R.M. & Death, R.G. (2003) Spawning site selection, egg development, and larval drift of *Galaxias postvectis* and *G. fasciatus* in a New Zealand stream. *New Zealand Journal of Marine and Freshwater Research*, **37,** 493–505.

Collier, K.J., Ball, O.J., Graesser, A.K., Main, M.R. & Winterbourn, M.J. (1990) Do organic and anthropogenic acidity have similar effects on aquatic fauna? *Oikos*, **59,** 33–38.

Croizat, L., Nelson, G.J. & Rosen, D.E. (1974) Centers of origin and related concepts. *Systematic Zoology*, **23,** 265–287.

Darwin, C. (1872) *On the origin of species by means of natural selection, or the preservation of favoured species in the struggle for life*, 6th edn. Murray, London.

Death, R.G. & Winterbourn, M.J. (1995) Diversity patterns in stream benthic invertebrate communities: the influence of habitat stability. *Ecology*, **76,** 1446–1460.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K.S., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29,** 129–151.

Elith, J., Leathwick, J.R. & Hastie, T. (in press) A working guide to boosted regression trees. *Journal of Animal Ecology*.

Flury, B.D. & Levri, E.P. (1999) Periodic logistic regression. *Ecology*, **80,** 2254–2260.

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

Friedman, J.H. (2002) Stochastic gradient boosting. *Computational Statistics and Data Analysis*, **38,** 367–378.

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

Friedman, J.H., Hastie, T. & Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, **28,** 337–407.

Gregory, S.V., Swanson, F.J., McKee, W.A. & Cummins, K.W. (1991) An ecosystem perspective on riparian zones. *BioScience*, **41,** 540–551.

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.

Hanley, J.A. & McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143,** 29–36.

Hastie, T. & Tibshirani, R.J. (1990) *Generalized additive models*. Chapman & Hall, London.

Hastie, T., Tibshirani, R. & Friedman, J.H. (2001) *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag, New York.

Heaney, L.R. (2007) Is a new paradigm emerging for oceanic island biogeography? *Journal of Biogeography*, **34,** 753–757.

Jowett, I.G. (1990) Factors related to the distribution and abundance of brown and rainbow trout in New Zealand clear-water rivers. *New Zealand Journal of Marine and Freshwater Research*, **24,** 429–440.

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.

Kawakita, M., Minami, M., Eguchi, S. & Lennert-Cody, C.E. (2005) An introduction to the predictive technique AdaBoost with a comparison to generalized additive models. *Fisheries Research*, **76,** 328–343.

Lamouroux, N. & Jowett, I.G. (2005) Generalized instream habitat models. *Canadian Journal of Fisheries and Aquatic Science*, **62,** 7–14.

Lamouroux, N. & Souchon, Y. (2002) Simple predictions of instream habitat model outputs for fish habitat guilds in large streams. *Freshwater Biology*, **47,** 1531–1542.

Lamouroux, N., Dolédec, S. & Gayraud, S. (2004) Biological traits of stream macroinvertebrate communities: effects of microhabitat, reach, and basin filters. *Journal of the North American Benthological Society*, **23,** 449–466.

Leathwick, J.R. (1998) Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science*, **9,** 719–732.

Leathwick, J.R., Rowe, D., Richardson, J., Elith, J. & Hastie, T. (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, **50,** 2034–2052.

Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T. & Taylor, P. (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, **321,** 267–281.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*, 2nd edn. Chapman & Hall, London.

McDowall, R.M. (1990) *New Zealand freshwater fishes: a natural history and guide*, revised edn. Heinemann Reed, Auckland.

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

McDowall, R.M. (1996) Volcanism and freshwater fish biogeography in the northeastern North Island of New Zealand. *Journal of Biogeography*, **23,** 139–148.

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. (2002) Accumulating evidence for a dispersal biogeography of southern cool temperature freshwater fishes. *Journal of Biogeography*, **29,** 207–219.

McDowall, R.M. (2006) *The taxonomic status, distribution and identification of the Galaxis vulgaris species complex in the eastern/southern South Island and Stewart Island*. Client Report CHCDOC2006-081. National Institute of Water and Atmospheric Research, Christchurch.

McDowall, R.M. & Allibone, R.M. (1994) Possible competitive exclusion of common river galaxiads *(Galaxias vulgaris)* by koaro *(G. brevipinnis)* following impoundment of the Waipori River, Otago, New Zealand. *Journal of Royal Society of New Zealand*, **24,** 161–168.

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.

McGlone, M.S. (2005) Goodbye Gondwana. *Journal of Biogeography*, **32,** 739–740.

McIntosh, A.R. & McDowall, R.M. (2004) Fish communities in rivers and streams. *Freshwaters of New Zealand* (ed. by J. Harding, P. Mosley, C. Pearson and B. Sorrell), pp. 17.1–17.19. New Zealand Hydrological and Limnological Societies, Christchurch.

McIntosh, A.R., Townsend, C.R. & Crowl, T.A. (1992) Competition for space between introduced brown trout (*Salmo trutta* L.) and a native galaxiid (*Galaxias vulgaris* Stokell) in a New Zealand stream. *Journal of Fish Biology*, **41,** 63–81.

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., Freeman, E., Blackard, J., Frescino, T., Zimmermann, N.E. & Edwards, T.C., Jr (2006) Predicting tree species presence in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling*, **199,** 176–187.

Newnham, R.M., Lowe, D.J. & Williams, P.W. (1999) Quaternary environmental change in New Zealand: a review. *Progress in Physical Geography*, **23,** 567–610.

Pearce, A.J. & O'Loughlin, C.L. (1985) Landsliding during a M 7.7 earthquake: influence of geology and topography. *Geology*, **12,** 855–858.

Pearson, C.P. (1995) Regional frequency analysis of low flows in New Zealand rivers. *Journal of Hydrology (NZ)*, **33,** 94–122.

Poff, N.L. (1997) Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. *Journal of the North American Benthological Society*, **16,** 391–409.

Poff, N.L. & Allan, D.J. (1995) Functional organization of stream fish assemblages in relation to hydrological variability. *Ecology*, **76,** 606–627.

Poff, N.L., Allen, J.D., Bain, M.B., Karr, J.R., Prestegaard, K.L., Richter, B.D., Sparks, R.E. & Stromberg, J.C. (1997) The

natural flow regime: a paradigm for river conservation and restoration. *BioScience*, **47,** 769–784.

Pulliam, H.R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3,** 349–361.

Quinn, J.M. & Hickey, C.W. (1990) Characterisation and classification of benthic invertebrate communities in 88 New Zealand rivers in relation to environmental factors. *New Zealand Journal of Marine and Freshwater Research*, **24,** 387–410.

R Development Core Team (2004) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (last accessed 24 January 2008).

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.

Richter, B.D., Mathews, R., Harrison, D.L. & Wigington, R. (2003) Ecologically sustainable water management: managing river flows for ecological integrity. *Ecological Applications*, **13,** 206–224.

Ridgeway, G. (2004) *GBM: Generalized boosted regression models*. R package, version 1.3-5. http://www.i-pensieri.com/gregr/gbm.shtml.

Rosgen, D.L. (1994) A classification of natural rivers. *Catena*, **22,** 169–199.

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.

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.

Sousa, W.P. (1984) The role of disturbance in natural communities. *Annual Review of Ecology and Systematics*, **15,** 353–391.

Suggate, R.P. (1982) The geological perspective. *Landforms of New Zealand* (ed. by J.M. Soons and M.J. Selby), pp. 1–14. Longman Paul, Auckland.

Theurer, F.D. (1982) Instream water temperature model, United States Fish and Wildlife Service, Cooperative Instream Flow Group. *Instream Flow Information Paper*, **16,** 131.

Townsend, C.R. & Crowl, T.A. (1991) Fragmented population structure in a native New Zealand fish: an effect of introduced brown trout? *Oikos*, **61,** 348–354.

Townsend, C.R. & Hildrew, A.G. (1994) Species traits in relation to a habitat template for river systems. *Freshwater Biology*, **31,** 265–275.

Waters, J.M. & Wallis, G.P. (2000) Across the Southern Alps by river capture? Freshwater fish phylogeography in South Island, New Zealand. *Molecular Ecology*, **9,** 1577–1582.

White, P.S. (1979) Pattern, process, and natural disturbance in vegetation. *Botanical Review*, **45,** 229–299.

Woods, R., Elliott, S., Shankar, U., Bidwell, V., Harris, S., Wheeler, D., Clothier, B., Green, S., Hewitt, A., Gibb, R. & Parfitt, R. (2006) *The CLUES project: predicting the effects of land-use on water quality – stage II*. Client Report MAF05502. National Institute of Water and Atmospheric Research, Christchurch.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article:

Appendix S1 Examples of alpha-hulls.
Appendix S2 Summaries of predictive performance and predictor contributions for individual species models.
Appendix S3 Fitted functions for the nine most important predictors for each species.
Appendix S4 Relationships between species distribution and minor predictors.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365-2699.2007.01887.x (This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

BIOSKETCHES

John Leathwick is a principal scientist at New Zealand's National Institute of Water and Atmospheric Research (NIWA) in Hamilton, where he focuses on the analysis of broad-scale biological patterns, and their implications for conservation management.

Jane Elith is a post-doctoral research fellow at the University of Melbourne and has particular interests in the use of statistical modelling as a tool for analysing the distributions of species.

Lindsay Chadderton worked for 15 years as an aquatic ecologist for New Zealand's Department of Conservation, but now works on aquatic invasive species in the United States for The Nature Conservancy's Great Lakes Program.

David Rowe is a freshwater fish ecologist working at NIWA Hamilton, with particular interests in the ecology of New Zealand's native fish.

Trevor Hastie is currently chairman of the Department of Statistics at Stanford University, and is well known as the author of several leading texts on statistical modelling and data mining.

Editor: David Bellwood

APPENDIX 1 BOOSTED REGRESSION TREE MODELS.

Boosted regression trees, also known as stochastic gradient boosting (Friedman *et al.*, 2000; Friedman, 2001, 2002) is a relatively new addition to the range of tools available for modelling relationships between species distributions and environment (Elith *et al.*, in press). Boosted methods such as boosted regression trees (BRT) differ substantially from regression-based methods such as Generalized Linear Models (GLM; McCullagh & Nelder, 1989) and Generalized Additive Models (GAM; Hastie & Tibshirani, 1990) that have been used widely over the last decade for such analyses (Guisan & Zimmerman, 2000). While these latter methods seek to identify a single 'best' model describing relationships between the response and predictor variables, boosting progressively builds a sequence of models of increasing complexity, each one fitting the training data slightly better than its predecessor. The model-building process in BRT is referred to as 'forward stagewise', which reflects the fact that at each step a term is added to the model to slightly decrease the deviance. The terms added are in the form of small regression trees, which are fit to the gradient of the deviance (generalized residuals). Results from this ensemble of regression trees are then averaged to form a final prediction. While the strong performance of boosted methods has been known for a number of years, we are aware of only a few instances in which they have been applied to the analysis of ecological data (Cappo *et al.*, 2005; Kawakita *et al.*, 2005; Leathwick *et al.*, 2006; Moisen *et al.*, 2006).

While in theory, boosting can be applied to any model fitting method, the use of regression trees as the individual model terms in a BRT model has particular advantages. These include their resistance to outliers, tolerance of extraneous predictors, capacity to accommodate missing values, ability to automatically fit interactions between predictors, and flexibility for modelling a variety of responses (Friedman & Meulman, 2003). Fitting a BRT model requires optimization of a number of parameters, of which the first controls the complexity of the individual regression trees fitted as model terms, typically taking an integer value in the range from 1 to 10. This parameter is sometimes referred to as the interaction depth, reflecting its control over the potential fitting of interactions between predictors. An individual tree that consists of only a single decision rule (or 'split') will fit a purely additive model (Hastie *et al.*, 2001), but the potential for fitting interactions between predictors increases as the complexity of the individual trees is increased, with the maximum degree of interaction that can be accommodated by each model term equalling the number of splits. In practice, the use of more complex individual trees does not force fitting of interactions, as even quite complex trees are capable of fitting purely additive effects where these predominate in the data being analysed. As a consequence, the degree of interaction between predictors fitted by any model must be determined using diagnostic procedures such as those that we describe below. All models in this analysis were fitted using individual trees consisting of five splits or decision rules, this value being chosen as giving the best overall predictive performance after inspection of results for a number of species using values ranging from one through to seven.

Models were fitted using a purpose written script that implemented a 10-fold cross-validation procedure in which models of increasing complexity were fitted to 10 temporary data sets, each of which comprised 90% of the full data set. Each of these data sets was created by withholding one of ten mutually exclusive subsets of the full data set, each containing 10% of the data selected randomly using prevalence stratification, i.e. each of the ten data sets contained presences and absences in the same ratio as in the full data set. For each model, trees were added in sets of 50, after which predictions were formed for the points withheld from that data set. These predictions were then compared with the withheld data by calculating the predictive deviance for each subset. Tree complexity was increased until there was no further decrease in the deviance when predicting to the withheld data, averaged across all subsets, which typically reached a minimum before increasing again at higher levels of model complexity, indicating over-fitting of the data; by contrast, the training deviance typically continues to decline as the models become progressively over-fitted to the training data. We then identified the level of model complexity that gave the lowest predictive error and fitted a final model to the full data set using this number of trees.

We detected the presence of interaction effects between predictor variables using purpose written code that examined the relationship between predicted values and all possible pairwise combinations of predictors. This was achieved for each pair of predictors by creating two variables (x_1 and x_2) that consisted of values at constant intervals along their ranges, and forming predictions (y') on the logit scale for all possible combinations of these. In making these predictions, values for all other variables were set at their mean for the data set. We then used a linear model to relate the predicted values to the values of the two marginal variables, i.e. $y' \sim x_1 + x_2$, with the two predictor variables fitted as factors. Where the predicted values are formed by a largely additive combination of the two predictors, this regression object will have very low residual variance. However, where interaction effects increase the complexity of the predicted surface, a significant amount of variance in y' will be left unexplained. We recorded the residual deviance for linear models fitted separately to predictions formed using all pairwise combinations of predictors for all species, and used these to assess the relative strength of interaction effects fitted for each predictor pair by calculating its median value across all species.