

REDISCOVERING THE SPECIES IN COMMUNITY-WIDE PREDICTIVE MODELING

JULIAN D. OLDEN,^{1,3} MICHAEL K. JOY,² AND RUSSELL G. DEATH²

¹Center for Limnology, University of Wisconsin–Madison, 680 N. Park Street, Madison, Wisconsin 53706 USA

²Institute of Natural Resources–Ecology, Massey University, Private Bag 11 222, Palmerston North, New Zealand

Abstract. Broadening the scope of conservation efforts to protect entire communities provides several advantages over the current species-specific focus, yet ecologists have been hampered by the fact that predictive modeling of multiple species is not directly amenable to traditional statistical approaches. Perhaps the greatest hurdle in community-wide modeling is that communities are composed of both co-occurring groups of species and species arranged independently along environmental gradients. Therefore, commonly used “short-cut” methods such as the modeling of so-called “assemblage types” are problematic. Our study demonstrates the utility of a multiresponse artificial neural network (MANN) to model entire community membership in an integrative yet species-specific manner. We compare MANN to two traditional approaches used to predict community composition: (1) a species-by-species approach using logistic regression analysis (LOG) and (2) a “classification-then-modeling” approach in which sites are classified into assemblage “types” (here we used two-way indicator species analysis and multiple discriminant analysis [MDA]). For freshwater fish assemblages of the North Island, New Zealand, we found that the MANN outperformed all other methods for predicting community composition based on multiscaled descriptors of the environment. The simple-matching coefficient comparing predicted and actual species composition was, on average, greatest for the MANN (91%), followed by MDA (85%), and LOG (83%). Mean Jaccard’s similarity (emphasizing model performance for predicting species’ presence) for the MANN (66%) exceeded both LOG (47%) and MDA (46%). The MANN also correctly predicted community composition (i.e., a significant proportion of the species membership based on a randomization procedure) for 82% of the study sites compared to 54% (MDA) and 49% (LOG), resulting in the MANN correctly predicting community composition in a total of 311 sites and an additional 117 sites ($n = 379$), on average, compared to LOG and MDA. The MANN also provided valuable explanatory power by simultaneously quantifying the nature of the relationships between the environment and both individual species and the entire community (composition and richness), which is not readily available from traditional approaches. We discuss how the MANN approach provides a powerful quantitative tool for conservation planning and highlight its potential for biomonitoring programs that currently depend on modeling discrete assemblage types to assess aquatic ecosystem health.

Key words: artificial neural networks; assemblage types; bioassessment; biomonitoring; classification-then-modeling approach; community classification; freshwater fish; New Zealand; River InVertebrate Prediction and Classification System, RIVPACS.

INTRODUCTION

Predictive ability is viewed by many as the ultimate confirmation of theory and understanding in ecology (Pielke and Conant 2003). In conservation biology, predictive modeling has been focused predominately on individual species, partly for historical reasons in the wildlife biology antecedents of resource management and conservation, but primarily because of the perceived intractability of a more holistic approach (Caughley 1994). However, it is increasingly recognized that conservation planning and management for biodiversity cannot be met by focusing on one species at a time, but

requires the adoption of multispecies or entire-community approaches (Simberloff 1998, Margules and Pressey 2000). Indeed, landscape-scale and ecoregional approaches to conservation are a pressing research priority in ecology (Soulé and Orians 2001), and the biological community is viewed as an appropriate unit of study.

Statistical modeling of species’ distributions is playing an increasing role in applied ecology (Rushton et al. 2004), in particular for identifying biodiversity priority areas (Williams and Araújo 2000). The success of multispecies management strategies and regional conservation planning depends on the development and rigorous testing of quantitative approaches that model the entire species composition of communities. Such models would enable an understanding of how communities are structured across different environmental gradients and provide a basis for predicting community

Manuscript received 14 March 2005; accepted 8 December 2005. Corresponding Editor: P. S. Levin.

³ E-mail: olden@wisc.edu

change to future environmental alteration. Community-level models would also contribute to the analytical needs of regional conservation planning that require the prediction of species compositional similarity across different sites and taxonomic groups (Ferrier 2002).

Unfortunately, beyond the prediction of single species or metrics the field of ecology has seen limited progress in modeling the species membership of entire biological communities because such investigations are not directly amenable to traditional approaches. Frustrated by the need to accommodate more than one species (i.e., more than one response variable) in the modeling process, the use of analytic shortcuts is now widespread in conservation biology. Partly in response to the growing emphasis on monitoring species diversity, ecologists often use traditional statistical approaches to model a single descriptor of the community (e.g., species richness, diversity index) or carefully chosen indicator, focal, keystone, umbrella, or flagship species that serve to represent the larger membership of the community (Caro and O'Doherty 1999). These so-called surrogate metrics are problematic for modeling community structure because they do not capture the identities of the entire species membership and as a result their use in conservation continues to be questioned (Simberloff 1998).

A "classification-then-modeling" approach has also been widely used to model community composition and is increasingly popular in conservation planning and decision making (Ferrier et al. 2002). This approach first groups species into "assemblage types" using deductive (i.e., co-occurring groups based on expert knowledge) or inductive (i.e., data driven using supervised or unsupervised cluster analysis) methods and then models these types using traditional classification approaches. While this approach has proven useful in the past (e.g., Whittaker 1978), it assumes that communities exist as discrete entities, a debate that has raged in ecology for many years and remains unresolved (see review by McIntosh [1995]). Despite the fact that species exhibit individualistic responses to environmental gradients and that assemblage types have often been difficult to identify (e.g., Angermeier and Winston 1999) and predict (e.g., Heino et al. 2003), "classification-then-modeling" approaches continue to provide the foundation for many conservation programs. In addition to this fundamental ecological issue, there are numerous methodological problems associated with defining assemblage types. These include difficulties in group recognition (Anderson and Clements 2000), the subjective choice of similarity metric used to define the groups (Jackson et al. 1989), the degree of similarity required to identify a group, the specific species that are expected to be present in the community when the assemblage type is predicted to occur, and the effects of taxonomic resolution and sampling effort on group structure (e.g., Cao et al. 2002).

Throughout the recent history of conservation biology, it is clear that traditional statistical approaches to modeling biological communities have unfortunately fostered a practice of shoehorning complex, multispecies problems into simple, single-variable solutions. Such approaches will likely provide unsure footing in conservation and resource management. Fortunately the past decade has seen tremendous progress in advancement of more sophisticated modeling techniques in ecology (Guisan and Zimmermann 2000). Artificial neural networks, for example, have recently shown considerable promise for addressing complex pattern recognition problems because of their robustness to the various numerical characteristics frequently encountered in ecological data (i.e., nonlinearity) that commonly limit the utility of traditional approaches (Lek and Guégan 2000). Moreover, while many traditional statistical techniques are limited to modeling a single predictor variable, a neural network can simultaneously model multiple dependent variables. Recently, it was this flexibility coupled with neural networks' efficiency with handling complex, nonlinear data that independently prompted Olden (2003) and Joy and Death (2004) to be the first to use a multiresponse artificial neural network (MANN) to develop predictive models for entire biological communities. In both studies, neural networks facilitated the development of a single, integrative model that predicted the species composition of an entire fish community while inherently respecting that species show different relationships with features of the environment. The flexibility of such an approach would appear to provide a powerful analytic tool for modeling communities that are typically both discrete and individualistic in their species composition across space (Shipley and Keddy 1987).

Although neural networks are attractive for modeling biological communities (Olden 2003), a comparison of their predictive performance to the more traditional modeling approaches highlighted above is needed to evaluate their potential utility in applied ecology. This study examined the modeling performance of three approaches to predicting species composition of stream fish communities in the North Island of New Zealand. Modeling approaches included those commonly used in the ecological literature for conservation planning and biodiversity prioritization and for assessing aquatic ecosystem health: (1) a species-by-species approach in which predictions are combined from individual logistic regression models (LOG) constructed for each of the fish species in the species pool; (2) a classification-then-modeling approach in which assemblage types are defined using a clustering algorithm (here, we used two-way indicator species analysis) and modeled as a function of a set of environmental variables using a classification method (here, we used multiple discriminant analysis [MDA]) in order to determine community composition based on group membership of the predicted assemblage type; and (3) a multiresponse

TABLE 1. Families, scientific and common names, and prevalence of the 16 fish species found in the 379 stream sites across the southwest region of the North Island, New Zealand.

Family and scientific name	Common name	No. sites	Prevalence
Anguillidae			
<i>Anguilla australis</i>	shortfin eel	127	0.34
<i>Anguilla dieffenbachii</i>	longfin eel	299	0.09
Galaxiidae			
<i>Galaxias argenteus</i>	giant kokopu	29	0.08
<i>Galaxias brevipinnis</i>	koaro	105	0.28
<i>Galaxias divergens</i>	dwarf galaxias	28	0.07
<i>Galaxias fasciatus</i>	banded kokopu	44	0.12
<i>Galaxias maculatus</i>	inanga	52	0.14
<i>Galaxias postvectis</i>	shortjaw kokopu	27	0.07
Geotriidae			
<i>Geotria australis</i>	lamprey	16	0.04
Eleotridae			
<i>Gobiomorphus cotidianus</i>	common bully	75	0.20
<i>Gobiomorphus hubbsi</i>	bluegill bully	31	0.08
<i>Gobiomorphus huttoni</i>	redfin bully	178	0.47
<i>Gobiomorphus breviceps</i> or <i>G. basalis</i>	Cran's or upland bully	72	0.19
Pinguipedidae			
<i>Cheimarrichthys fosteri</i>	torrentfish	36	0.10
Retropinnidae			
<i>Retropinna retropinna</i>	common smelt	9	0.02
Salmonidae			
<i>Salmo trutta</i>	brown trout	155	0.41

Notes: All species are native to the region, except brown trout. Prevalence refers to the frequency of site occurrence in the data set.

artificial neural network (MANN) that models the entire community. Based on the results we discuss the manner in which the neural network approach provides a powerful quantitative tool for conservation planning and highlight its potential for biomonitoring programs that currently depend on modeling discrete assemblage types to assess aquatic ecosystem health.

MATERIALS AND METHODS

Fish community data and multiscaled environment variables

The study area was the Wellington Region of the North Island, New Zealand (41° S, 175° E). Fish distribution data was sourced from the New Zealand Freshwater Fish Database (McDowell and Richardson 1983), which contained fish assemblage data derived from electrofishing surveys on 259 stream reaches (delineated as the watercourse between two confluences with tributaries) sampled between 1980 and 2002. This database was supplemented with additional data for 120 stream reaches from a recent survey by M. K. Joy during the austral summer of 2002 (see Joy and Death [2004] for details). The final database contained records for a total of 16 fish species, ranging in prevalence from 0.024 to 0.789 (Table 1). Here, we refer to a fish assemblage as those individuals

that occur together in a “locality” with spatial dimensions such that individual fish have a reasonable probability of encountering one another in the course of daily activities (sensu Matthews 1998). Although it is difficult to rigorously define, the stream segment is commonly used as an operational unit representing a fish assemblage that results from a single sampling event using conventional ichthyological methods.

We examined a series of reach-scale and catchment-scale descriptors that are believed to be related to critical habitat requirements of fish in New Zealand (Jowett and Richardson 2003). Habitat descriptors were derived from a geographic information system (GIS) and included variables categorized into four groups: reach-scale habitat, catchment-scale habitat, catchment-scale land use, and catchment-scale geology (Table 2). Annual mean precipitation and air temperature for each catchment was estimated from thin plate splines fitted to meteorological station data (1950–1980). Catchment-scale geology and land use for the study area region were obtained from the River Environment Classification (Snelder and Biggs 2002). The geology classifications were sourced from the New Zealand Land Resources Inventory (Newsome 1990), and land cover classifications were obtained from the New Zealand Land Cover Database (*available online*).⁴ All analyses were conducted using ArcGIS (Environmental Systems Research Institute, Redlands, California, USA).

A series of diagnostic analyses and data transformations were conducted prior to statistical modeling in an attempt to ensure that multivariate normality was met for the traditional statistical approaches (note that neural networks do not depend on these assumptions). Given intercorrelations among land use and geology variables, we separately subjected each of these two variable groups to a principal-components analysis to derive a reduced set of orthogonal composite variables that summarized the major gradients in land use and geology. The statistical significance of the principal-component axes was evaluated using the broken-stick rule, in which the observed eigenvalues are compared to eigenvalues from random data (Jackson 1993). For both the geology and land use analyses the first six principal components were found to be significant and were subsequently used as predictor variables during modeling. The final data set consisted of 379 sites containing fish community data for 16 species and 24 descriptors of reach- and catchment-scale features.

Statistical methods

Modeling individual species using logistic regression analysis.—Predictions of community composition were obtained by constructing logistic regression models relating species occurrence to the suite of environmental variables for each of the 16 fish species (the traditional

⁴ (<http://www.mfe.govt.nz/>)

decision threshold of 0.50 was used to assign species' presence) and then combining the site predictions of all the models. Logistic regression analysis is a class of linear models that are parameterized using a maximum likelihood principle and are based on a logistic transformation of the response variable with a linear combination of the independent variables. Logistic regression analysis is commonly used in the ecological literature to model species occurrence (Guisan and Zimmermann 2000).

Modeling "assemblage types" using multiple discriminant analysis.—A classification-then-modeling approach was used to predict community composition in which species were first grouped into assemblage types using two-way indicator species analysis (TWINSPAN) and then these types were discriminated according to the set of environmental variables using multiple discriminant analysis. TWINSPAN (Hill 1979) is based on reciprocal averaging and simultaneously classifies species and samples in ordination space with the goal of identifying discrete assemblage types. Following Jowett and Richardson (2003), who classified fish communities of New Zealand, we set the minimum group size at 10 and we defined groups in five levels giving 32 potential site groups. No attempt was made to use fish abundance data in TWINSPAN due to potential operator bias in the New Zealand Freshwater Fish Database. In cases in which species compositions were similar or the group sizes were small, assemblage delineations were made at higher divisions. Importantly, while some have argued that other clustering algorithms may perform better than TWINSPAN (see Belbin and McDonald 1993), most if not all algorithms are seen to have some form of limitation (e.g., Jackson et al. 1989, Anderson and Clements 2000), and TWINSPAN continues to have merit as a tool in ecological investigations (e.g., Heino et al. 2003).

Multiple discriminant analysis (also called canonical discriminant analysis) is a standard multivariate method that seeks a linear combination of the independent variables to maximally separate between-class means (in this case n classes representing each of the TWINSPAN assemblage types) relative to the within-class variance. Predictions of fish assemblage types from the MDA were translated to predicted community compositions using a two-step procedure. First, we calculated the frequency with which individual species occurred in each TWINSPAN group (i.e., the relative group frequency) as the number of sites in which that taxon occurs divided by the total number of sites in the group. Second, the overall probability of a species occurring at a site was calculated as the relative group frequency multiplied by the probability of membership in each of the TWINSPAN groups from the MDA, and a decision threshold of 0.5 was used to assign species' presence (see Appendix A for an example calculation). This process gives a weighting to all TWINSPAN groups based the discriminant analysis probabilities rather than just the group with highest probability (see Wright 1995). This assign-

TABLE 2. Reach- and catchment-scale habitat variables used to model fish communities.

Habitat descriptors	Mean
Reach-scale habitat	
Latitude	44.1
Stream order (Strahler)	2.9
Distance from sea (km)	27.5
Upstream reach elevation (m a.s.l.)	141.3
Downstream reach elevation (m a.s.l.)	103.3
Reach length (m)	1116.9
Catchment-scale habitat	
Catchment area (km ²)	32.3
Mean catchment elevation (m)	333.6
Mean catchment slope (m/km)	28.5
Mean annual catchment rainfall (mm)	1530.2
Mean annual catchment temperature (°C)	19.2
Area of catchment comprised of lakes (km ²)	16.0
Catchment-scale land use	
Native forest, exotic forest, pastoral, scrub, urban, tussock, coastal, bare ground, and other	
Catchment-scale geology	
Surface rock: peat, loess, alluvium, sand, mudstone, calcareous, and other	
Base rock: peat, loess, alluvium, sand, mudstone, calcareous, windblown, greywacke, and other	

Notes: To meet the data assumption of multivariate normality for the traditional approaches we $\log_e(x)$ -transformed values of catchment area, lake catchment area, distance to sea, and downstream section elevation, and we arcsine square-root transformed the nine catchment-scale land use and 16 geology categorical variables that were expressed as proportions.

ment procedure is used in the River InVertebrate Prediction and Classification System (RIVPACS; Wright et al. 2000) and other river assessment schemes (e.g., Smith et al. 1999). In a series of analyses that we do not report in detail, we found that model performance for predicting community composition was lower when using the conventional approach of assigning each site to the single TWINSPAN group with the highest probability and when using different decision thresholds (e.g., using the total proportion of sites in which the species was present as the threshold).

Modeling the community using an artificial neural network.—Fish community composition was modeled using a multiresponse artificial neural network. An artificial neural network is an information-processing system that was inspired by the manner in which biological nervous systems, such as the mammalian brain, assimilate information (Bishop 1995). The key element of this modeling approach is the novel structure of the information-processing system, which is composed of a large number of highly interconnected processing elements (neurons) working in unity to solve specific problems. In addition to the flexibility of neural networks to model multiple dependent variables (see Olden 2003), they have a number of advantages over traditional parametric approaches, including their ability to accurately model nonlinear data and accommo-

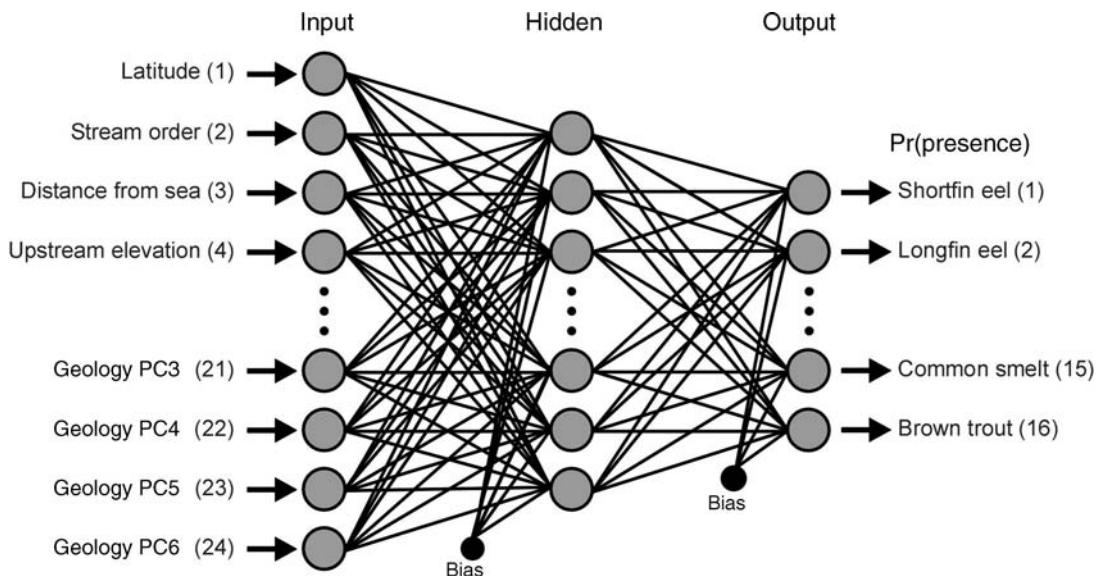


FIG. 1. A schematic of the multiresponse artificial neural network used to model stream fish communities in the North Island of New Zealand. Circles represent network neurons, and lines represent network connections whose weights are optimized during model training. The neural network in our analysis contained 24 input neurons (environmental variables; PC, principal component), 30 hidden neurons, and 16 output neurons (i.e., probability of species' presence). Values in parentheses refer to numbered environmental variables (input neurons) and species (output neurons).

date interactions among predictor variables without a priori specification (Lek and Guégan 2000).

We used a feed-forward neural network (also known as a multilayer perceptron) trained by the back-propagation algorithm to model the occurrence of the 16 fish species (Rumelhart et al. 1986). The architecture of this network consisted of a single input layer, a hidden layer, and an output layer (Fig. 1). The input layer contained one neuron for each of the environmental variables. The number of hidden neurons in the neural network was chosen by comparing the performances of different cross-validated networks, with 5–100 hidden neurons (increasing by increments of five), and choosing the number that produced the greatest network performance. The output layer contains 16 neurons representing the predicted probability of occurrence for each species. The mechanics of the neural network can be expressed mathematically as

$$y_k = \varphi_o \left[\beta_k + \sum_j w_{jk} \varphi_h \left(\beta_j + \sum_i w_{ij} x_i \right) \right] \quad (1)$$

where x_i are the input signals, y_k are the output signals, w_{ij} are the weights between input neuron i to hidden neuron j , w_{jk} are the weights between hidden neuron j and output neuron k , β_j and β_k are the biases associated with the hidden and output layers, and φ_h and φ_o are activation functions (in this case logistic functions) for the hidden and output layers. We used a traditional decision threshold of 0.5 to classify a species as present or absent.

During network training, observations are sequentially presented to the network and the back-propaga-

tion algorithm adjusts the weights in a backwards fashion, layer by layer, in the direction of steepest descent in minimizing the error function (in this case the cross-entropy error for the binary response variables). Learning rate (which controls the step size when weights are iteratively adjusted) and momentum parameters (which add inertia to the learning motion through weight space) were included during network training to ensure a high probability of global network convergence. A maximum of 500 iterations of the back-propagation algorithm were used to determine the optimal axon weights during network construction. Prior to training the network, the independent variables were converted to z scores to standardize the measurement scales of the inputs into the network. We refer the reader to the comprehensive text of Bishop (1995) for more details regarding neural networks.

The relative contribution of the environmental variables in the neural network depends primarily on the magnitude and direction of the weights connecting the input and output neurons via the hidden layer (see review by Olden et al. [2004]). Following the methodology of Olden and Jackson (2002) we quantified variable importance as the product of the input–hidden and hidden–output connection weights between each input neuron (environmental variable) and output neuron (fish species) and then summing the products across all hidden neurons. Positive values represent positive environment–species associations, whereas negative values represent negative associations. The relative contributions of the environmental variables were calculated by dividing the absolute value of each

variable contribution by the grand sum of all absolute contributions, and their statistical significance was assessed using a randomization test. All neural network analyses were conducted using computer macros written in the MatLab (MathWorks, Natick, Massachusetts, USA) programming language.

Model construction, validation, and performance

To generate model predictions and assess classification performance of the LOGs, MDA, and MANN, N -fold cross-validation was used. This validation method excludes one observation, constructs the model with the remaining $n - 1$ observations, predicts the response of the excluded observation using this model, and repeats the procedure n times. The degree of concordance between predicted and observed fish community composition was assessed using the simple matching coefficient and Jaccard's similarity coefficient (Krebs 1999), both of which range from 0 (no species in common between observed and predicted community composition or 0% similarity) to 1 (all species in common between observed and predicted community composition or 100% similarity). A simple matching coefficient gives equal weighting to correctly predicted presences and absences from the models and was used to provide an overall assessment of the predictability of the fish communities. Jaccard's coefficient quantifies community similarity excluding the cases in which species were correctly predicted absent and therefore accounts for the influence of low species prevalence in the data set for inflating the predictability of species absence. After generating predictions of community composition the next question to address is whether the model performed greater from expectations based on chance. Given that the performance of a statistical model is strongly influenced by the prevalence of species occurrence in the data set (Manel et al. 2001), we used a randomization test to assess whether model predictions of fish community composition were greater than expectations based on chance. For each site, the test involved randomly assigning the presence of each species with probability Pr_k , where Pr equaled the prevalence of species k in the data set (see Table 1), calculating the simple matching coefficient and Jaccard's similarity coefficient between the randomized community composition and the observed community composition, and repeating the entire procedure 9999 times. Significance levels were then calculated as the proportion of trials in which the percentage of similarity between the randomized and the observed community was equal to or greater than the percentage of similarity between the observed and predicted community matrix. A critical alpha of 0.05 was used to determine whether the community at a particular site was correctly predicted (i.e., a significant proportion of the community). Lastly, by summing the predicted occurrences across the individual species, we are able to compare the performances of the approaches for predicting species richness.

RESULTS

The logistic regression models exhibited high predictive power, on average, correctly classifying individual species' presence or absence for 84% of sites, ranging between 67% correct classification for bluegill bully (255 out of 379 sites) and 93% for common smelt (351 out of 379 sites; see Table 1 for scientific names). The TWINSpan analysis grouped the 379 sites into 11 assemblage types comprised of characteristic suites of fish species that we viewed as representative of the study region. The primary division separated sites containing predominantly large galaxiids (groups G–K, Fig. 2a) from sites containing a more diverse assemblage (groups A–F). Of the sites containing mainly species in the Galaxiidae family, a secondary division distinguished between those sites containing diadromous species (i.e., regularly migrating between freshwater and saltwater) and the sites containing non-diadromous species (i.e., dwarf galaxies, group G). Of the sites not containing galaxiids, a secondary division separated sites containing or lacking brown trout (groups D–F vs. A–C), and of the sites with brown trout, a tertiary division separated sites with non-diadromous bullies (i.e., Cran's or upland bully, group D) from the sites with the remaining bullies that are diadromous. Importantly, common species were members of multiple assemblage types (e.g., shortfin eel, longfin eel, redbull, brown trout), whereas rare species were often not a member of any assemblage type. Assemblage type centroids showed good separation according to the first two canonical functions from the MDA using the environmental data (Fig. 2b). We found clear separation between galaxiid (types A–F) and non-galaxiid assemblages (types G–K) along the first canonical function, whereas trout (types D–F) vs. non-trout assemblages (types A–C) were separated along the second canonical function. The MDA correctly predicted fish assemblage group membership for 64% of the stream sites (range 47–92%). The five most important variables contributing to the MDA were, in order of importance (according to F -to-enter values based on the overall Wilks' lambda): catchment area ($F_{1,10} = 15.5$, $P < 0.001$), distance from sea ($F_{2,10} = 15.0$, $P < 0.001$), catchment elevation ($F_{3,10} = 14.9$, $P < 0.001$), mean annual precipitation ($F_{4,10} = 14.4$, $P < 0.001$), and mean annual air temperature ($F_{5,10} = 13.8$, $P < 0.001$). These groupings and their discriminant functions were then used to predict the expected communities using the assignment approach.

The MANN exhibited higher classification power for predicting community composition compared to the other approaches. Based on the simple-matching coefficient, mean similarity between predicted and observed community composition for the neural network was 91%, which was greater than the classification-then-modeling approach (85%) and species-by-species logistic regression models (84%; Fig. 3). The MANN correctly predicted a significant portion of community membership for 82% of the study sites based on the randomization test, whereas

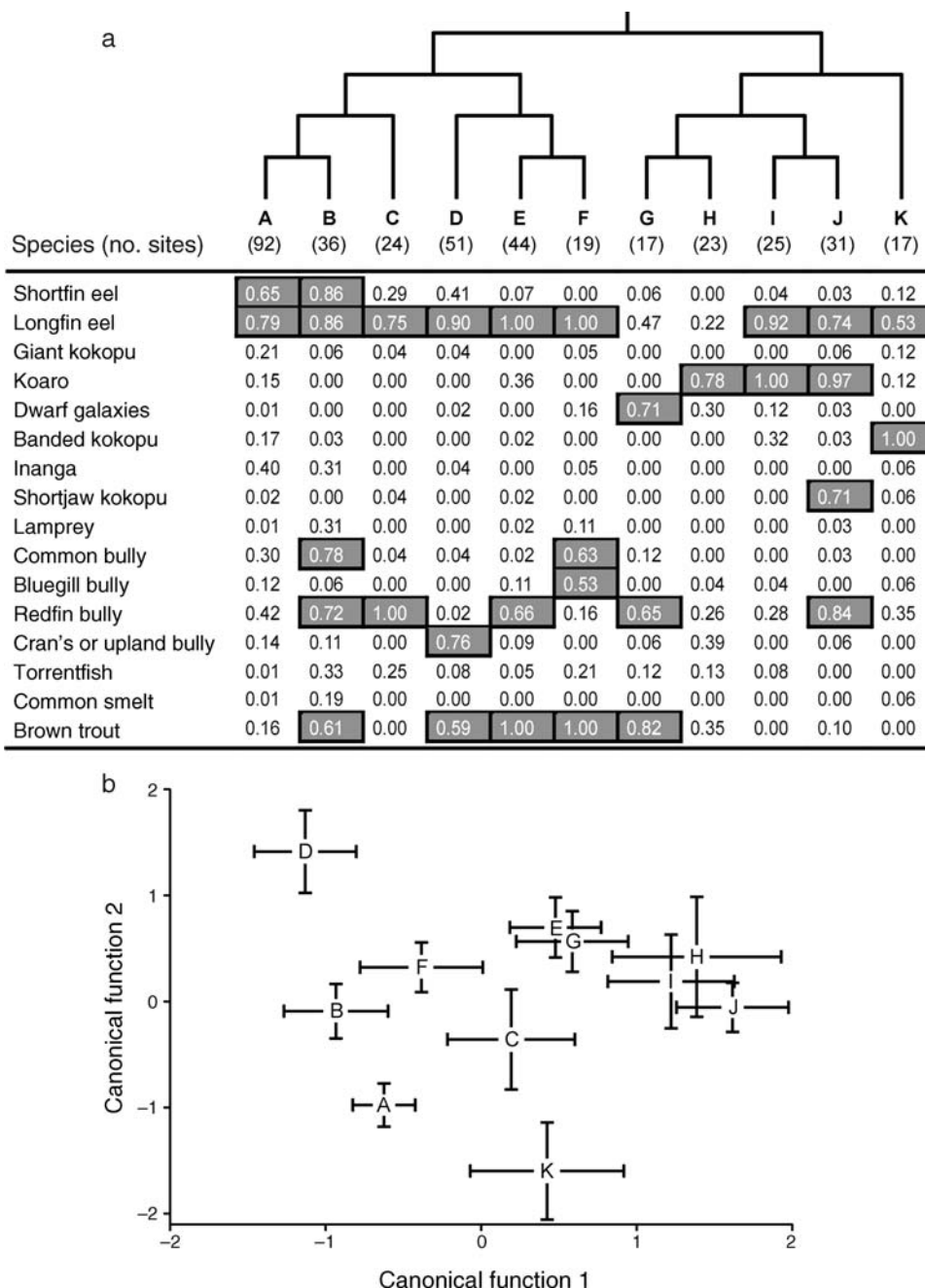


FIG. 2. (a) TWINSpan classification (using a modified convergence criteria of Hill [1979]) of the study sites into 11 assemblage types (A–K). The dendrogram shows the number of sites (in parentheses) and the prevalence of the species in each group (shaded cells indicate species' presence based on a threshold probability of 0.5). (b) Position of the mean coordinates of the 11 TWINSpan groups in discriminant space obtained by multiple discriminant analysis on the 24 environmental variables. The letter in each centroid denotes the assemblage-type group and error bars denote 95% confidence intervals.

the other approaches significantly predicted community composition for only 49–54% of the sites (Table 3). As a result, the neural network correctly predicting community composition for a total of 311 sites (out of 379) compared to 205 sites (MDA) and 184 sites (LOG). In addition, for 91 sites the MANN correctly predicted the entire species membership of the community (i.e., 16 out

of 16 species), whereas only 28 and 23 sites were correctly predicted based on MDA and LOG, respectively. The degree of variation in the similarity between predicted and observed fish community composition was also smallest for the neural network (Fig. 3).

The MANN also exhibited the greatest classification success in terms of predicting species' presence in the fish

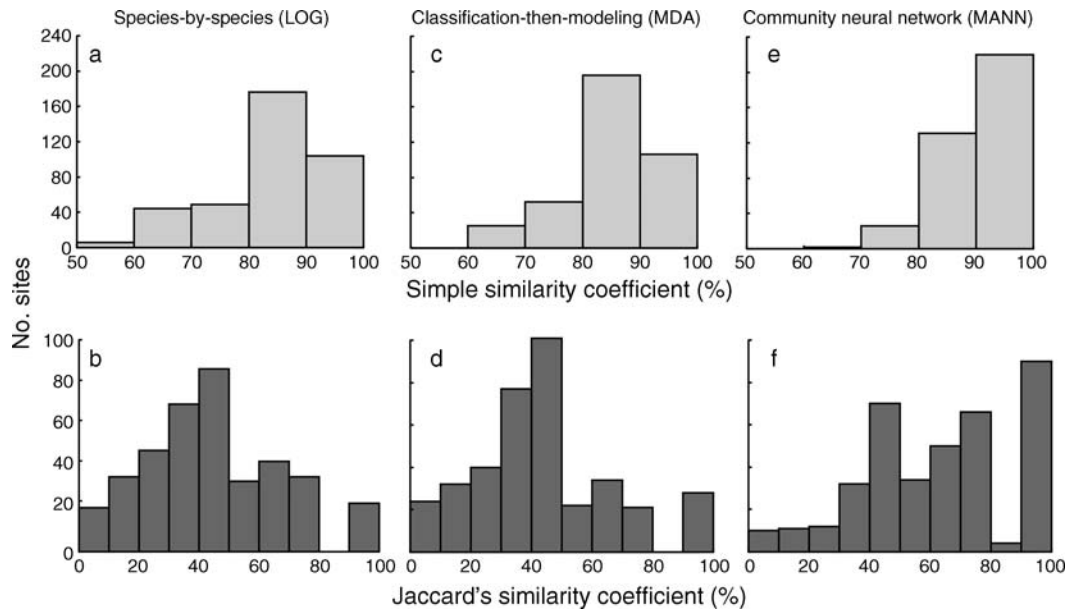


FIG. 3. Simple and Jaccard's similarity between observed and predicted fish community composition according to (a, b) the species-by-species logistic regression models, (c, d) the "classification-then-modeling" approach using TWINSpan groups and multiple discriminant analysis, and (e, f) the community multiresponse artificial neural network.

communities, thus supporting the efficiency of this approach for modeling communities containing rare species (Fig. 3). The mean Jaccard's similarity between predicted and observed community composition for the MANN was 66%, which was greater than the combined logistic regression models (47%) and the classification-then-modeling approaches (46%). This resulted in the MANN correctly predicting the community composition of 76% or 288 sites compared to LOG (173 sites) and MDA (158 sites; Table 3).

Predictions of species richness from the MANN showed the strongest correlation with actual richness ($R = 0.63$), followed by the MDA ($R = 0.32$) and the combined LOG models ($R = 0.28$). In fact, the neural network outperformed a multiple regression model that was built specifically for species richness (regression model, $F_{24,354} = 5.69$, $P < 0.05$, $R = 0.52$). Taken together, the MANN provided greater power for predicting both community composition and richness compared to traditional approaches.

Analysis of the network connection weights illustrates the relative importance of the environmental variables for predicting community membership (Fig. 4). Six environmental variables were found to be statistically significant predictors of fish community composition (averaged across all species values; see Appendix B), in decreasing order of importance: reach-scale descriptors of upstream elevation and stream order; catchment area; mean annual air temperature; land use principal component 1 (PC1; see Appendix C for the results from the principal-components analysis); and mean annual precipitation. Predicted occurrence of all species (i.e., species richness) was positively related to increasing

stream size (reflected in stream order and catchment area) and negatively related to site elevation. Sites experiencing greater annual air temperatures were predicted to exhibit greater species richness, as were those located in catchments containing greater proportions of native vegetation vs. pastoral farming (land use PC1).

DISCUSSION

Growing threats to biological diversity and ecosystem integrity call for innovative approaches for meeting current conservation challenges. Recent years have seen a major shift in the focus of conservation and restoration efforts from single species to entire communities and even ecosystems (Franklin 1993, Young et al. 2005). Perhaps the greatest hurdle in the development of community models for species occurrence is that biological communities are composed of both co-occurring, repeatable groups of species and aggregations of species arranged independently of one another in "continua" along environmental gradients (Shipley and Keddy 1987). Therefore, any particular community is expected to have both individualistic and discrete elements, and ideally we would like to be able to model both components simultaneously.

Consistent with this methodological need, our study presents a novel neural network approach for modeling entire communities in an integrative and species-specific manner, thus embracing the complexity by which communities assemble in nature. A primary strength of this approach is that it provides a single statistical model that predicts complete species membership from a common set of environmental variables while respecting

TABLE 3. Comparison of methodologies for predicting fish community composition across the study sites.

Methodology	Simple matching		Jaccard's similarity	
	Mean (sd)	Correct (%)	Mean (sd)	Correct (%)
Logistic regression models (LOG)	83.7 (9.9)	48.5	46.7 (23.9)	45.6
Assemblage-type (MDA)†	85.3 (8.5)	54.1	45.8 (24.1)	41.7
Multiresponse artificial neural network (MANN)	91.0 (7.4)	82.1	65.5 (25.8)	76.0

Note: Reported values include mean and standard deviation of the similarity metrics comparing predicted and actual species composition and the percentage of sites whose communities were correctly predicted.

† Also referred to as multiple discriminant analysis.

that species are expected to exhibit different functional relationships with the environment (Olden 2003). This allows the simultaneously assessment of the relationship between environmental change and both species occurrence and community structure, thereby having the flexibility to concurrently model species that exhibit similar habitat affinities (and therefore tend to co-occur or form groups) and species showing more individualistic responses to the environment. Compared to other community-level approaches that have been used extensively in ecology (see review by Ferrier et al. [2002]), our evaluation showed that the predictive ability of the neural network outperformed the combined logistic regression models and a classification-then-modeling approach based on TWINSpan and MDA.

The neural network exhibited greater accuracy and precision for predicting fish community membership and richness at each site and, on average, correctly predicting community composition in nearly twice the number of sites compared to the other approaches.

Community-level modeling approaches based on assemblage classifications continue to be popular in conservation planning and decision making as they are seen to alleviate the need to consider a large number of individual species by instead focusing on a reduced number of higher-level assemblage types (Ferrier 2002). While this approach has played an important role in planning activities, at the same time it is recognized that “the principle disadvantage of modelling communities or assemblages instead of species is that the approach

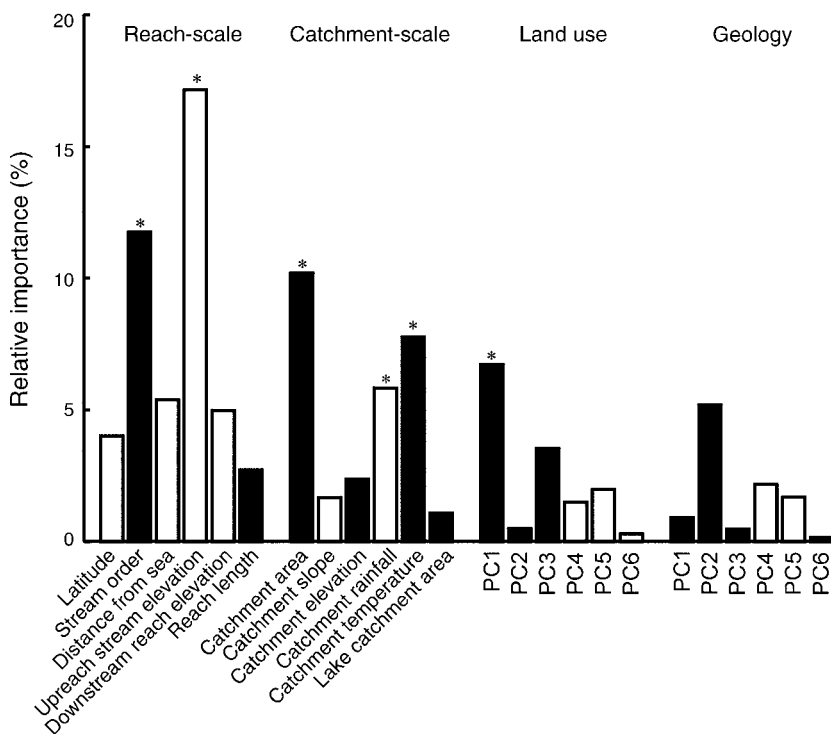


FIG. 4. Relative importance of the 24 environmental variables in the community neural network (PC, principal component). An overall positive connection weight indicates a positive association between the particular environmental variable and predicted probabilities of species (averaged across all species) and therefore greater predicted species richness (indicated by solid bars). The converse is true for negative connection weights (indicated by open bars). Asterisks denote variables that significantly ($P < 0.05$) contribute to network predictions based on the randomization test. Variable contributions for each species are presented in Appendix B.

may not allow planning to give consideration to the needs of individual species of particular conservation concern" (Ferrier et al. 2002:2310–2311). Our finding that the classification-then-modeling technique that relies on the selection of assemblage "types" did not perform as well as the neural network is not entirely surprising. This method makes the critical assumption that communities exist as discrete entities and that all the species in an assemblage type respond identically to changes in the environment. In freshwater systems (and many others) this assumption is typically violated. Nearly a century ago, Forbes (1907) showed that fish do not readily fall into discrete groups but rather occupy positions along an environmental continuum, and more generally, recent studies continue to illustrate that establishing the existence of discrete assemblages is a difficult task for a number of aquatic taxa, specifically fish (Angermeier and Winston 1999, Pusey et al. 2000), macroinvertebrates (Hawkins and Vinson 2000, Heino et al. 2003), riparian plants (Nilsson 1986), and lake shoreline vegetation (Keddy 1983).

Clearly, there is no simply answer to the Clementsian-Gleasonian question: "Do assemblages consist of discrete types?" However, if discrete groups of taxa do not occur in nature it seems unlikely that a modeling technique that imposes boundaries on groups of taxa is going to perform particularly well. Our results are particularly concerning because community classification approaches (e.g., TWINSPLAN followed by MDA) represent the core analytical technique in biomonitoring programs that are used to assess aquatic ecosystem "health" (Hawkins et al. 2000). In fact, studies recognize the limitations of using community classifications in bioassessment (Chessman 1999, Hawkins and Vinson 2000), yet efforts continue to try to enhance the classification component of the analysis by proposing new methodologies for forcing macroinvertebrate assemblages into discrete "types" (e.g., Linke et al. 2005). We reiterate Heino et al. (2003) in emphasizing that macroinvertebrate and fish assemblages do not readily form discrete species groups, and therefore predictive models for bioassessment should not rely too heavily on classification-then-modeling approaches. Such efforts may be futile, and instead, alternative approaches that model entire communities, such as the neural network approach presented here, should be considered.

Despite the apparent advantages of the community neural network over classification-then-modeling, we are not arguing that the use of assemblage types in ecology should be completely abandoned. In fact, such approaches have proven useful in the past (e.g., Ferrier et al. 2002), and predictions from a neural network could be used to test and refine the definition of assemblage types, thus enhancing their use in conservation biology. Research in this area warrants further investigation. We do suggest, however, that neural networks provide greater utility (and we have shown also predictive and explanatory power) for modeling communities. In fact,

work is currently in progress that attempts to modify network architecture to allow for species interactions to be explicitly incorporated into community neural networks, thus respecting the nonadditive manner in which species assemble into communities. This would provide the opportunity to account for species' interactions, both positive and negative, which are seen to play an important role in shaping community structure (Callaway 1997).

Regional conservation planning is a top priority in conservation biology (Abell 2002), and recent decades have seen increasing interest in the development and application of systematic approaches to this process (Margules and Pressey 2000). In a recent review, Ferrier (2002) proposed that information on compositional similarity between different areas may serve as a basis for conservation planning by facilitating the selection of sites that maximize the representation of regional biodiversity. Given that empirical data is typically in short supply, one of the primary challenges is to develop models that predict species compositional similarity based on patterns of environmental similarity. While matrix regression models have been shown to be useful for deriving predictions of compositional similarity (Steinitz et al. 2005), this approach does not provide information regarding the identity of the species that contribute to the predicted similarity. The neural network approach described in this paper provides an analytical solution to this problem by addressing taxonomic diversity as a whole through dual predictions of individual species and entire communities from which compositional similarity can be calculated. Because it allows the simultaneous assessment of the effects of environmental change at the individual species and entire community levels, it may be particularly suitable for the conservation of biodiversity at regional scales. For example, aquatic resource management in New Zealand is at the research forefront and predictive models are seen as playing an essential role in the areas of bioassessment, risk-assessment for biological invasions, and conservation planning for biodiversity (Woods and Howard-Williams 2004). The MANN model developed here is currently used by the Wellington Regional Council; an interactive map called "point-click-fish" (interfaced in a geographic information system) has been developed to allow users to "click" on any stream reach in a region and simultaneously obtain species-specific probabilities of fish occurrence and community-level composition and richness. Similar models have been completed for three other New Zealand Regional Councils, and we are currently developing an interactive map for the entire country that can be used in regional conservation planning.

CONCLUSION

Ecological forecasting has been described as an emerging imperative and thus the quality of predictive models is also critical. Our study demonstrates that the

multispecies neural network possesses all the qualities of an ideal predictive modeling tool, combining the advantages of single-species modeling (focusing on species of interest) with the advantages of community modeling (focusing on collective properties of species groups), which makes it especially useful for both community assessment and conservation planning. Until recently, the ability of researchers to use neural networks was limited to those with computer programming experience, but this is no longer the case. There now exist a number of off-the-shelf, Windows-based programs, modules for commonly used software packages, and libraries for a number of programming languages, and most of these provide the sophistication needed to develop multiresponse neural networks (see Appendix D). In the future, the availability of these techniques, when combined with geographic information systems, will allow managers to integrate the complexity of biological communities and make multispecies decisions about different management strategies. In conclusion, we hope our study will spark interest in the applied potential of artificial neural networks in modeling biological communities for conservation planning and decision making. In combination with multiple-species monitoring efforts (Manley et al. 2004) we see the community neural network as a valuable tool in conservation planning by helping derive environmental quality metrics that ensure both the persistence of rare and special-interest species and community-level attributes that are important for ecosystem functioning.

ACKNOWLEDGMENTS

We thank Mark Kennard and three anonymous reviewers for insightful comments on this paper. J. D. Olden was supported by the David H. Smith Post-doctoral Conservation Research Fellowship Program. M. K. Joy was supported by a grant from the sustainable management fund (New Zealand Ministry for the Environment grant No. 5099 "Nga Ika Waioara": stream health evaluation tool) and the Wellington Regional Council. J. D. Olden thanks the Massey University Center for Ecosystem Modeling and Management for graciously providing travel funds for his visit.

LITERATURE CITED

- Abell, R. 2002. Conservation biology for the biodiversity crisis: a freshwater follow-up. *Conservation Biology* **16**:1435–1437.
- Anderson, M. J., and A. Clements. 2000. Resolving environmental disputes: a statistical method for choosing among competing cluster models. *Ecological Applications* **10**:1341–1355.
- Angermeier, P. L., and M. R. Winston. 1999. Characterizing fish community diversity across Virginia landscapes: prerequisite for conservation. *Ecological Applications* **9**:335–349.
- Belbin, L., and C. McDonald. 1993. Comparing three classification strategies for use in ecology. *Journal of Vegetation Science* **4**:341–348.
- Bishop, C. M. 1995. *Neural networks for pattern recognition*. Oxford University Press, New York, New York, USA.
- Callaway, R. M. 1997. Positive interactions in plant communities and the individualistic continuum concept. *Oecologia* **112**:143–149.
- Cao, Y., D. P. Larsen, R. M. Hughes, P. L. Angermeier, and T. M. Patton. 2002. Sampling effort affects multivariate comparisons of stream assemblages. *Journal of the North American Benthological Society* **21**:701–714.
- Caro, T. M., and G. O'Doherty. 1999. On the use of surrogate species in conservation biology. *Conservation Biology* **13**:805–814.
- Caughley, G. 1994. Directions in conservation biology. *Journal of Animal Ecology* **62**:215–235.
- Chessman, B. C. 1999. Predicting the macroinvertebrate faunas of rivers by multiple regression of biological and environmental differences. *Freshwater Biology* **41**:747–757.
- Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? *Systematic Biology* **51**:331–363.
- Ferrier, S., M. Drielsma, G. Manion, and G. Watson. 2002. Extending statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and Conservation* **11**:2309–2338.
- Forbes, S. A. 1907. On the local distribution of certain Illinois fishes: an essay in statistical ecology. *Bulletin of the Illinois Natural History Survey* **7**:273–303.
- Franklin, J. F. 1993. Preserving biodiversity: Species, ecosystems, or landscapes? *Ecological Applications* **3**:202–205.
- Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* **135**:147–186.
- Hawkins, C. P., R. H. Norris, J. N. Hogue, and J. W. Feminella. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications* **10**:1456–1477.
- Hawkins, C. P., and M. R. Vinson. 2000. Weak correspondence between landscape classifications and stream invertebrate assemblages: implications for bioassessment. *Journal of the North American Benthological Society* **19**:501–517.
- Heino, J., T. Muotka, H. Mykra, R. Paavola, H. Hamalainen, and E. Koskenniemi. 2003. Defining macroinvertebrate assemblage types of headwater streams: implications for bioassessment and conservation. *Ecological Applications* **13**:842–852.
- Hill, M. O. 1979. DECORANA—a FORTRAN program for detrended correspondence analysis and reciprocal averaging. *Ecology and Systematics*, Cornell University, Ithaca, New York, USA.
- Jackson, D. A. 1993. Stopping rules in principal components analysis: a comparison of heuristic and statistical approaches. *Ecology* **74**:2204–2214.
- Jackson, D. A., K. M. Somers, and H. H. Harvey. 1989. Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence? *American Naturalist* **133**:436–453.
- Jowett, I. G., and J. Richardson. 2003. Fish communities in New Zealand rivers and their relationship to environmental variables. *New Zealand Journal of Marine and Freshwater Research* **37**:347–366.
- Joy, M. K., and R. G. Death. 2004. Predictive modelling and spatial mapping of freshwater fish and decapod assemblages: an integrated GIS and neural network approach. *Freshwater Biology* **49**:1036–1052.
- Keddy, P. A. 1983. Shoreline vegetation in Axe Lake, Ontario: effects of exposure on zonation patterns. *Ecology* **64**:331–344.
- Krebs, C. J. 1999. *Ecological methodology*. Benjamin/Cummings, Menlo Park, California, USA.
- Lek, S., and J. F. Guégan, editors. 2000. *Artificial neuronal networks: applications to ecology and evolution*. Springer-Verlag, New York, New York, USA.
- Linke, S., R. H. Norris, D. P. Faith, and D. Stockwell. 2005. ANNA: a new prediction method for bioassessment programs. *Freshwater Biology* **50**:147–158.

- Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* **38**:921–931.
- Manley, P. N., W. J. Zielinski, M. D. Schlesinger, and S. R. Mori. 2004. Evaluation of a multiple-species approach to monitoring species at the ecoregional scale. *Ecological Applications* **14**:296–310.
- Margules, C. R., and R. L. Pressey. 2000. Systematic conservation planning. *Nature* **405**:243–253.
- Matthews, W. J. 1998. Patterns in freshwater fish ecology. Kluwer Academic, Boston, Massachusetts, USA.
- McDowall, R. M., and J. Richardson. 1983. The New Zealand freshwater fish survey: a guide to input and output. Ministry of Agriculture and Fisheries, Wellington, New Zealand.
- McIntosh, R. P. 1995. H. A. Gleason's "individualistic community concept" and theory of animal communities: a continuing controversy. *Biological Reviews* **70**:317–357.
- Newsome, P. F. J. 1990. New Zealand land resources inventory ARC/INFO data manual. Landcare Research, Lincoln, New Zealand.
- Nilsson, C. 1986. Change in riparian plant community composition along two rivers in northern Sweden. *Canadian Journal of Botany* **64**:589–592.
- 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., and D. A. Jackson. 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* **154**:135–150.
- Olden, J. D., M. K. Joy, and R. G. Death. 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* **178**:389–397.
- Pielke, R. A., and R. T. Conant. 2003. Best practices in prediction for decision-making: lessons from the atmospheric and earth sciences. *Ecology* **84**:1351–1358.
- Pusey, B. J., M. J. Kennard, and A. H. Arthington. 2000. Discharge variability and the development of predictive models relating stream fish assemblage structure to habitat in northeastern Australia. *Ecology of Freshwater Fish* **9**:30–50.
- Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. Learning representations by back-propagating errors. *Nature* **323**:533–536.
- Rushton, S. P., S. J. Ormerod, and G. Kerby. 2004. New paradigms for modelling species distributions? *Journal of Applied Ecology* **41**:193–200.
- Shipley, B., and P. A. Keddy. 1987. The individualistic and community-unit concepts as falsifiable hypotheses. *Vegetatio* **69**:47–55.
- Simberloff, D. 1998. Flagships, umbrellas, and keystones: Is single-species management passé in the landscape era? *Biological Conservation* **83**:247–257.
- Smith, M. J., et al. 1999. AusRivAS: using macroinvertebrates to assess ecological condition of rivers in Western Australia. *Freshwater Biology* **41**:269–282.
- Snelder, T., and B. J. F. Biggs. 2002. Multi-scale river environment classification for water resources management. *Journal of the American Water Resources Association* **38**:1225–1240.
- Soulé, M., and G. Orians, editors. 2001. Conservation biology: research priorities for the next decade. Island Press, Washington, D.C., USA.
- Steinitz, O., J. Heller, A. Tsoar, D. Rotem, and R. Kadmon. 2005. Predicting regional patterns of similarity in species composition for conservation planning. *Conservation Biology* **19**:1978–1988.
- Whittaker, R. H., editor. 1978. Classification of plant communities. Junk, The Hague, The Netherlands.
- Williams, P. H., and M. B. Araújo. 2000. Using probability of persistence to identify important areas for biodiversity conservation. *Proceedings of the Royal Society of London B* **267**:1959–1966.
- Woods, R., and C. Howard-Williams. 2004. Advances in freshwater sciences and management in New Zealand. Pages 1.1–1.20 in J. S. Harding, M. P. Mosley, C. Pearson, and B. Sorrell, editors. *Freshwaters of New Zealand*. Caxton Press, Christchurch, New Zealand.
- Wright, J. F. 1995. Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. *Australian Journal of Ecology* **20**:181–197.
- Wright, J. F., D. W. Sutcliffe, and M. T. Furse. 2000. Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK.
- Young, T. P., D. A. Peterson, and J. J. Clary. 2005. The ecology of restoration: historical links, emerging issues and unexplored realms. *Ecology Letters* **8**:662–673.

APPENDIX A

Sample calculation illustrating the manner in which community composition is predicted from multiple discriminant analysis (MDA) classifications of TWINSPAN fish assemblage types (*Ecological Archives* A016-048-A1).

APPENDIX B

Relative contributions of the environmental variables in the multiresponse artificial neural network (MANN) for predicting individual species occurrence (*Ecological Archives* A016-048-A2).

APPENDIX C

Results from the principal-components analysis on land use and geology variables (*Ecological Archives* A016-048-A3).

APPENDIX D

A partial list of Window-based programs, modules for commonly used software packages, and libraries for a number of programming languages that conduct artificial neural network analysis (*Ecological Archives* A016-048-A4).