Freshwater Biology (2002) **47**, 1976–1995

APPLIED ISSUES

A comparison of statistical approaches for modelling fish species distributions

JULIAN D. OLDEN and DONALD A. JACKSON
Department of Zoology, University of Toronto, Toronto, Ontario, Canada

SUMMARY

1. The prediction of species distributions is of primary importance in ecology and conservation biology. Statistical models play an important role in this regard; however, researchers have little guidance when choosing between competing methodologies because few comparative studies have been conducted.
2. We provide a comprehensive comparison of traditional and alternative techniques for predicting species distributions using logistic regression analysis, linear discriminant analysis, classification trees and artificial neural networks to model: (1) the presence/absence of 27 fish species as a function of habitat conditions in 286 temperate lakes located in south-central Ontario, Canada and (2) simulated data sets exhibiting deterministic, linear and non-linear species response curves.
3. Detailed evaluation of model predictive power showed that approaches produced species models that differed in overall correct classification, specificity (i.e. ability to correctly predict species absence) and sensitivity (i.e. ability to correctly predict species presence) and in terms of which of the study lakes they correctly classified. On average, neural networks outperformed the other modelling approaches, although all approaches predicted species presence/absence with moderate to excellent success.
4. Based on simulated non-linear data, classification trees and neural networks greatly outperformed traditional approaches, whereas all approaches exhibited similar correct classification rates when modelling simulated linear data.
5. Detailed evaluation of model explanatory insight showed that the relative importance of the habitat variables in the species models varied among the approaches, where habitat variable importance was similar among approaches for some species and very different for others.
6. In general, differences in predictive power (both correct classification rate and identity of the lakes correctly classified) among the approaches corresponded with differences in habitat variable importance, suggesting that non-linear modelling approaches (i.e. classification trees and neural networks) are better able to capture and model complex, non-linear patterns found in ecological data. The results from the comparisons using simulated data further support this notion.
7. By employing parallel modelling approaches with the same set of data and focusing on comparing multiple metrics of predictive performance, researchers can begin to choose

Correspondence: Julian D. Olden, Department of Biology and Graduate Degree Program in Ecology, Colorado State University, Fort Collins, CO 80523-1878, U.S.A. E-mail: olden@lamar.colostate.edu

predictive models that not only provide the greatest predictive power, but also best fit the proposed application.

Keywords: artificial neural networks, classification trees, discriminant analysis, logistic regression, species presence/absence

Introduction

Ecologists have long been interested in understanding and predicting the distributions of species across landscapes (Orians, 1980; Buckland & Elston, 1993; Pickett, Kolasa & Jones, 1994; Lawton, 1996; Gaston & Blackburn, 1999). However, the relative emphasis placed on the explanatory and predictive components of this ecological research varies substantially across disciplines and taxa (Keddy, 1992). For example, plant ecologists have traditionally spent more effort on developing models to predict species distributions (e.g. Hill & Keddy, 1992; Toner & Keddy, 1997; Wiser, Peet & White, 1998), as have stream ecologists in predicting the occurrence of invertebrates (e.g. Bailey *et al*., 1998; Chessman, 1999; Moss *et al*., 1999) and fish (e.g. Scheller *et al*., 1999; Oberdorff *et al*., 2001). In contrast, lake ecologists have generally focused on understanding species-environment processes rather than attempting to formulate this knowledge into testable, predictive models. Moreover, models that have been developed have primarily focused on making predictions at small spatial scales (stream reaches or lakes within a single watershed), as opposed to landscape or regional scales.

Our understanding of fish-environment associations in lakes has emerged primarily from comparative studies that describe statistical relationships between sets of environmental variables and species occurrence or abundance (see Jackson, Peres-Neto & Olden, 2001 for a review). Such studies (e.g. Jackson & Harvey, 1989; Tonn *et al*., 1990; Rodriguez & Lewis, 1997; Magnuson *et al*., 1998) identify the influence of abiotic conditions (lake morphology, water chemistry), biotic interactions (predation, competition), habitat isolation and human-related factors (e.g. land-use practices) in structuring fish populations at local, landscape and regional spatial scales. The next and often missing essential step is to place this understanding in a quantitative framework where species distributions can be readily and accurately predicted from these environmental factors. More than ever, predictive models are urgently needed as the modification and loss of aquatic habitat is now recognised as the primary factor threatening the conservation of fish populations and communities throughout many parts of the world (Williams *et al*., 1989; Richter *et al*., 1997; Harig & Bain, 1998; Ricciardi & Rasmussen, 1999).

Predictive models have a number of important applications for the conservation and management of fish populations. Predictive fish-habitat models can play an important role in prioritising surveys and monitoring programmes for fish populations because limitations to resources often preclude exhaustive and continual sampling of sites and that extensive sampling is needed to accurately sample lake fish communities (Jackson & Harvey, 1997). Applications of model predictions include: (1) forecasting or measuring the effects of habitat alteration and changing land-use patterns (Oberdorff *et al*., 2001); (2) providing first-order estimates of habitat suitability to establish potential locations for re-introduction (Evans & Oliver, 1995); (3) predicting the likelihood of local establishment and spread of exotic species (Peterson & Vieglais, 2001) that may help set conservation priorities for preserving vulnerable species and populations that might be lost locally; (4) predicting 'hotspots' of species persistence for the conservation of biodiversity (Williams & Araujo, 2000); and (5) revealing additional populations of threatened species, or alternatively revealing unexpected gaps in their range.

Although there are obvious conservation implications from being able to quantify the predictability of species distributions, the development of these models is a difficult task because patterns of fish occurrence and abundance commonly exhibit complex, non-linear relationships to habitat heterogeneity and biotic interactions. Of the statistical approaches, logistic regression and linear discriminant analysis are most commonly used, although our confidence in their results is often limited by the inability to meet a number of assumptions, such as statistical distributions of variables, independence of variables and

model linearity (James & McCulloch, 1990). Consequently, researchers are now recognising the potential utility of non-linear statistical approaches such as classification and regression trees (e.g. Magnuson *et al*., 1998; Rathert *et al*., 1999; Rejwan *et al*., 1999; De'ath & Fabricius, 2000), artificial neural networks (e.g. Lek *et al*., 1996; Mastrorillo *et al*., 1997; Brosse & Lek, 2000; Olden & Jackson, 2001) and genetic algorithms (e.g. D'Angelo *et al*., 1995) for modelling ecological data. It is believed that these alternative approaches can provide researchers with more flexible tools for modelling complex ecological relationships. Although it is encouraging that a broad array of quantitative approaches is currently available to model species distributions, we are now faced with the difficulty of choosing among a large number of competing statistical methodologies. In a recent synthesis, Guisan & Zimmermann (2000) highlighted the need for comparative studies where more than two statistical methods are applied to the same data set to help address this problem, as these comparisons are lacking in the literature.

The primary objective of our study is to compare the predictive power and explanatory insight provided by traditional, linear approaches (i.e. logistic regression analysis and linear discriminant analysis) and alternative, non-linear approaches (i.e. classification trees and artificial neural networks) for modelling species presence/absence. We address this objective by developing fish-habitat models for 27 fish species in north-temperate lakes of Canada and providing a detailed evaluation and comparison among species and among the four modelling approaches. Comparisons involve both the predictability of species based on a number of performance metrics and the relative importance of the habitat variables for predicting the occurrence of the species. To strengthen the methodological comparisons, we test the performance of the four modelling approaches in predicting simulated patterns of species presence/absence across an environmental gradient. The results from this comparison demonstrate the predictive ability of these various approaches under known conditions, thus providing a robust comparison of methodologies that can be generalised with other data. Taken together, these comparative analyses are advancements over more conventional model evaluations and provide important information regarding the comparison of modelling approaches and insight into the predictability of fish species occurrence across large spatial scales.

Methods

Study site and ecological data

The study system consisted of 286 freshwater lakes from five drainage basins located in Algonquin Provincial Park, south-central Ontario, Canada (Fig. 1). Algonquin Provincial Park (7630 km^2) is situated on Precambrian Canadian Shield bedrock and is located in the transition zone between the northern boreal and the southern deciduous hardwood forests. Aquatic communities in this region represent relatively natural ecosystems because these lakes are located in a provincial park and are subject to minimal perturbations from development and species introductions, although there were some species introductions (e.g. smallmouth bass) during the early 1900s, which have subsequently colonised some adjacent waters. We developed fish-habitat models for 27 fish species (Table 1) by modelling species presence/absence as a function of 12 or 13 whole-lake habitat characteristics (Table 2). These predictor variables were chosen to include factors related to habitat requirements of fish in this region (e.g. Matuszek & Beggs, 1988; Minns, 1989) and included: lake surface area; lake volume; total shoreline perimeter (sum of lake and island perimeters); maximum depth; surface measurements (taken at depths ≤ 2.0 m) of total dissolved solids and pH; lake altitude; growing degree-days (the average daily temperature above 5 °C, summed across all days); occurrence of summer thermal stratification (calculated as a function of thermocline depth and maximum depth; see Hanna, 1990); occurrence of a large littoral-zone piscivore (i.e. northern pike, smallmouth bass or largemouth bass) when modelling small-bodied fish; and three binary variables used in combination to delineate the five drainage basins, i.e. Amable du Fond, Bonnechere, Madawaska, Oxtongue and Petawawa Rivers, to account for the potential influence of biogeography on fish distributions. All data were obtained from the Algonquin Park Fish Inventory Data Base; a data base that involved a combination of extensive sampling of lakes in the park during 1989–91 using multiple gear types (multipanel survey gill nets, plastic traps, seines, Gee minnow traps) and

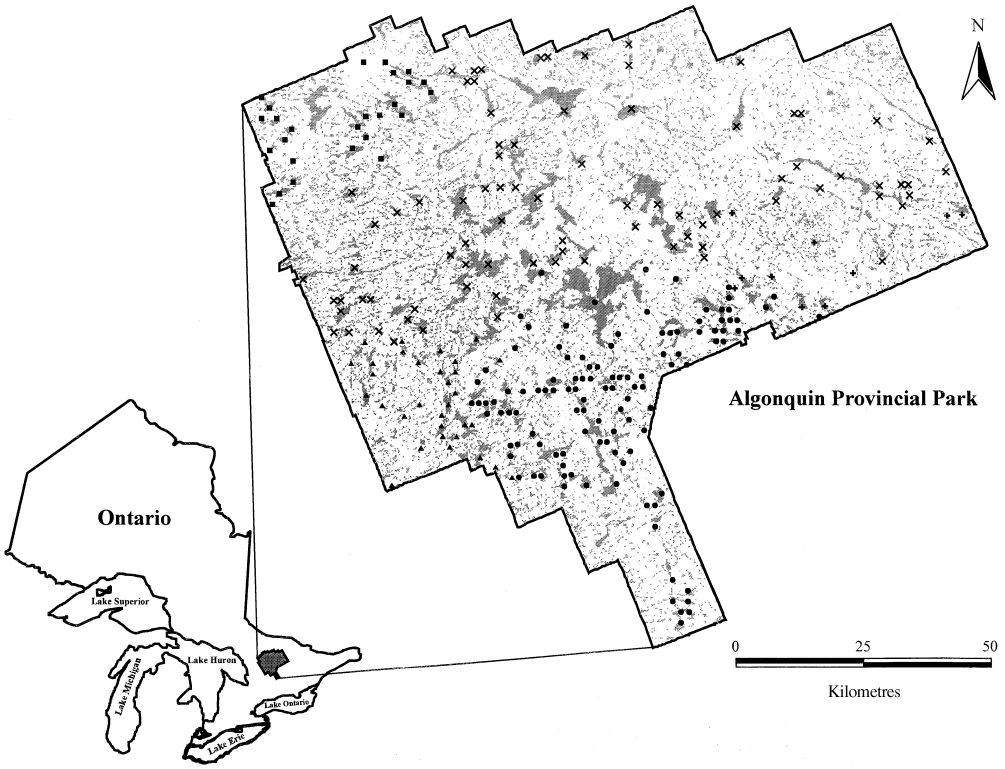

Fig. 1 Map of 286 study lakes located in Amable du Fond River (■), Bonnechere River (+), Madawaska River (●), Oxtongue River (▲) and Petawawa River (×) basins of Algonquin Provincial Park, Ontario, Canada (45°50′N, 78°20′W).

Table 1 List of fish species (organised by family), including species abbreviation (Code) and frequency of occurrence (%) in the 286 study lakes

Code	Common name	Scientific name	%
Catostomidae			
LS	Longnose sucker	*Catostomus catostomus* (Forster, 1773)	18.5
WS	White sucker	*Catostomus commersoni* (Lacepède, 1803)	83.6
Centrarchidae			
PKS	Pumpkinseed	*Lepomis gibbosus* (Linnaeus, 1758)	65.0
RB	Rock bass	*Ambloplites rupestris* (Rafinesque, 1817)	12.2
SMB	Smallmouth bass	*Micropterus dolomieu* Lacepède, 1802	21.3
Cyprinidae			
BCS	Blackchin shiner	*Notropis heterodon* (Cope, 1865)	5.6
BNS	Blacknose shiner	*Notropis heterolepis* Eigenmann & Eigenmann, 1893	40.6
CC	Creek chub	*Semotilus atromaculatus* (Mitchill, 1818)	68.2
CS	Common shiner	*Luxilus cornutus* (Mitchill, 1817)	54.2
F	Fallfish	*Semotilus corporalis* (Mitchill, 1817)	11.2
FSD	Finescale dace	*Phoxinus neogaeus* Cope, 1868	15.0
GS	Golden shiner	*Notemigonus crysoleucas* (Mitchill, 1814)	39.9
LC	Lake chub	*Couesius plumbeus* (Agassiz, 1850)	16.4
NRD	Northern redbelly dace	*Phoxinus eos* (Cope, 1862)	54.5
PD	Pearl dace	*Margariscus margarita* Cope, 1868	41.3
Gadidae			
B	Burbot	*Lota lota* (Linnaeus, 1758)	30.8

Table 1 (*Continued*)

Code	Common name	Scientific name	%
Gasterosteidae			
BSB	Brook stickleback	*Culaea inconstans* (Kirtland, 1841)	27.3
Ictaluridae			
BB	Brown bullhead	*Ameiurus nebulosus* (Lesueur, 1819)	47.2
Percidae			
ID	Iowa darter	*Etheostoma exile* (Girard, 1859)	20.3
YP	Yellow perch	*Perca flavescens* (Mitchill, 1814)	71.0
Percopsidae			
T-P	Trout-perch	*Percopsis omiscomaycus* (Walbaum, 1792)	9.4
Salmonidae			
BT	Brook trout	*Salvelinus fontinalis* (Mitchill, 1814)	76.9
C	Cisco	*Coregonus artedi* Lesueur, 1818	21.3
LT	Lake trout	*Salvelinus namaycush* (Walbaum, 1792)	52.8
LW	Lake whitefish	*Coregonus clupeaformis* (Mitchill, 1818)	13.6
RW	Round whitefish	*Prosopium cylindraceum* (Pallas, 1784)	10.5
SL	Splake	*S. fontinalis* × *S. namaycush*	7.7

Table 2 Summary statistics for the habitat variables used in model development to predict species presence/absence in the 286 study lakes

Predictor variable	Minimum	First quartile	Median	Third quartile	Maximum
Surface area (ha)	1.6	24.3	54.5	151.7	5154.2
Volume ($\times 10^4$ m^3)	1.0	8.3	29.0	91.4	8560.0
Total shoreline perimeter (km)	1.0	3.0	6.0	12.0	171.0
Maximum depth (m)	1.5	9.5	15.1	24.4	107.4
Altitude (m)	165	390	415	442	488
pH	5.5	6.5	7.0	7.0	8.0
Total dissolved solids (mg L^{-1})	2.0	22.2	27.0	32.0	210.0
Growing degree days	1613	1624	1625	1638	1797
Occurrence of summer stratification			0 – Absence, 1 – Presence		
Occurrence of littoral-zone piscivore			0 – Absence, 1 – Presence		
Three drainage basin dummy variables			0, 1		

records obtained from the Ontario Ministry of Natural Resources Lake Inventory and the Royal Ontario Museum collection (see Crossman & Mandrak, 1992). The standardised sampling methodology for this inventory is described in Dodge *et al*. (1985).

Simulated data

We compared the predictive performance of logistic regression analysis, linear discriminant analysis, classification trees and neural networks (more details are provided in the section *Statistical approaches to modelling species presence/*absence) using a Monte Carlo simulation experiment with data exhibiting defined numerical relationships between species occurrence and a habitat gradient. A number of species response curves are seen in nature including linear, logarithmic, logistic and Gaussian (Fig. 2a: also see Jongman, ter Braak & van Tongeren, 1995, pp. 31–32). For example, patterns in the probability of occurrence of various fish species commonly exhibit linear, logarithm, logistic and Gaussian responses to patterns in shoreline complexity, lake area, lake maximum depth and pH, respectively. However, because species are observed as either present or absent at a site, the linear, logarithmic and logistic response curves are seen as a step function (Fig. 2b) and the Gaussian response curve as a block function (Fig. 2c). Consequently, we generated two statistical populations following Fig. 2b,c (the Gaussian response curve

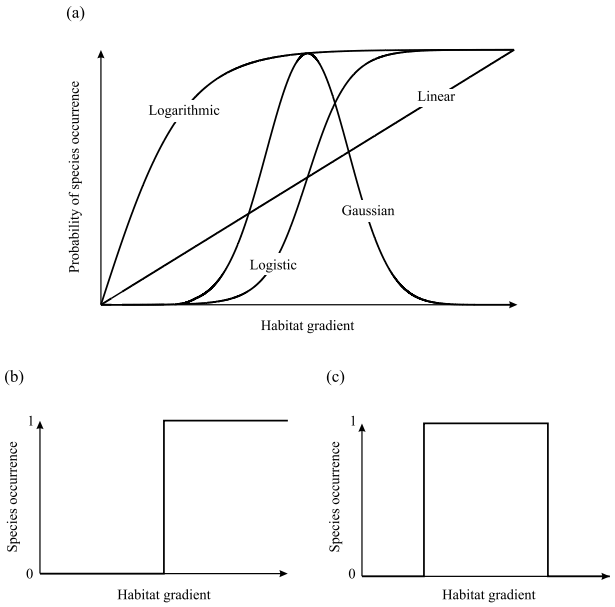

Fig. 2 (a) Species response curves where the probability of species occurrence can have either linear or non-linear relationships with an environmental gradient; (b) simulated data where species occurrence has a linear response to a single environmental variable [represented as a step-function when the linear, logarithmic or logistic response curves in (a) are converted to a presence/absence response]; (c) simulated data where species occurrence has a non-linear response to a single environmental variable [represented as a block-function when the Gaussian response curve in (a) is converted to a presence/absence response]. See text for more details regarding the simulation of data sets.

characterised by a mean of five and a variance of one), each containing 10 000 observations (where an observation is the presence or absence of a species at a particular site) and with values of a single habitat variable ranging from zero to 10 (data points were generated at uniform distances along the variable). Note that the response variables were simulated to have equal numbers of presence and absence values. The Monte Carlo experiment consisted of randomly sampling 30 observations from the statistical population, constructing a logistic regression model, discriminant function, classification tree and neural network and recording overall per cent correct classification (more details are provided in the section *Model construction, validation and performance metrics*). This procedure was repeated 500 times to ensure that meaningful conclusions emerged and summary statistics were calculated. A sample size of 30 was chosen as this corresponded to the mean sample size (29.3) based on a review of 98 statistical models reported in the literature (Tables 1–5 in Fausch, Hawkes & Parsons, 1988) and thus represents a sample size that provides a reasonable degree of generality. The step-function response curve (which is generated from the linear, logarithmic or logistic curves) represents 'optimal' data types for the traditional linear approaches in terms of distributional characteristics, whereas the Gaussian response curve represents a non-linear relationship where the probability of species occurrence or abundance is maximised at intermediate values of a habitat variable.

Statistical approaches to modelling species presence/absence

We applied logistic regression analysis (LRA), linear discriminant analysis (LDA), classification trees (CFT) and artificial neural networks (ANN) to develop predictive models for fish species presence/absence. Because of their well-documented use in the ecological literature, we refrain from detailing LRA and LDA methodologies, providing only a brief description (referring the reader to Hand, 1997 for a comprehensive coverage). The LRA is a class of linear models that are parameterised using a maximum likelihood principle and are based on a logistic transformation of the response variable with a linear combination of the independent variables. The LDA is a standard multivariate method that seeks a linear combination of the independent variables to maximally separate between-class means (two classes: presence and absence) relative to the within-class variance. In contrast, researchers are generally less familiar with CFT and ANN and therefore we discuss these methodologies at greater length.

Classification trees. The use of automatic construction of classification or decision trees dates from the pioneering work of Morgan & Sonquist (1963) in the social sciences, but was rekindled in the statistical literature by the seminal monograph of Breiman *et al.* (1984). Classification and regression trees have been used extensively in the social and medical sciences, but only recently recognised as potentially powerful tools for modelling ecological data (De'ath & Fabricius, 2000). Classification and regression trees are nonparametric, classification techniques that are most commonly implemented using a recursive-partitioning algorithm. This algorithm repeatedly partitions the data set into a

nested series of mutually exclusive groups, each of them as homogeneous as possible with respect to the response variable. When modelling species presence/absence, the procedure begins with the entire data set, also called the root node, and formulates split-defining conditions for each possible value of the explanatory variables to create candidate splits. Next, the algorithm selects the candidate split that minimises the misclassification rate and uses it to partition the data set into two subgroups. The algorithm continues recursively with each of the new subgroups until no split yields a significant decrease in the misclassification rate, or until the subgroup contains a small number of observations. A terminal subgroup or 'leaf' is a node that the algorithm cannot partition any further because of group size or because it is a relatively homogeneous group in terms of the values of the response variable. The response class (in this case the presence or absence of a species) for each terminal node is assigned by minimising the resubstitution estimate of the probability of misclassification for the observations of that node. For more detailed accounts on classification trees, see Breiman *et al*. (1984) and De'ath & Fabricius (2000).

In this study, we employed a discriminant-based, univariate split-selection method based on the algorithms used in QUEST (Quick, Unbiased, Efficient Statistical Trees: Loh & Shih, 1997). The QUEST provides an unbiased method for variable selection during tree construction, where bias is commonly associated with predictor variables containing few or many levels and can consequently skew the interpretation of variable importance in the tree (Breiman *et al*., 1984). Furthermore, QUEST has been shown to exhibit one of the best combinations of error rate and speed compared with 22 other decision-tree algorithms using 32 data sets (Lim, Loh & Shih, 2000). To determine the optimal size of each tree (i.e. the number of terminal nodes), the mode of 50 repeated cross-validations using the one standard-error rule (Breiman *et al*., 1984) was used and splitting was stopped when nodes contained less than five observations. The relative importance of each predictor variable in each CST was estimated by summing the changes in misclassification (also called impurity) for each surrogate split across all nodes and was expressed on a 0–100 scale (see Breiman *et al*., 1984, p. 147 for details).

Artificial neural networks. Although ANNs were originally developed to better understand how the mammalian brain functions, researchers have become more interested in the potential statistical utility of neural network algorithms (Cheng & Titterington, 1994; Bishop, 1995). In this study, we used one-hidden-layer feedforward neural network trained by the backpropagation algorithm (Rumelhart, Hinton & Williams, 1986). We used this type of network because it is considered to be a universal approximator of any continuous function (Hornik, Stinchcombe & White, 1989) and we used a single hidden layer because this is generally satisfactory for statistical applications (Bishop, 1995), it greatly reduces computational time and generally produces similar results compared with multiple hidden layers (Kurková, 1992).

The one-hidden-layer feedforward network consists of single input, hidden and output layers, with each layer containing one or more neurones. The input layer contains p neurones, each of which represents one of the p predictor variables, i.e. in our case 12 input neurones for each species, except for small-bodied species where the input layer contained 13 neurones (including the littoral-zone piscivore variable). The optimal number (optimal referring to minimising the trade-off between network bias and variance) of hidden neurones in the neural network is determined empirically by choosing the number of hidden neurones that produces the lowest misclassification rate (Bishop, 1995). The output layer contains one neurone representing the probability of species occurrence. An additional bias neurone with a constant output (equal to one) is added to the hidden and output layers, and plays a similar role to that of the constant term in multiple regression analysis. Each neurone (excluding the bias neurones) is connected to all neurones from adjacent layers by axons, and the axon connection between neurones is assigned a weight that dictates the intensity of the signal transmitted by the axon. In feedforward networks, axon signals are transmitted in a unidirectional path, from input layer to output layer through the hidden layer. The 'state' or 'activity level' of each neurone is determined by the input received from the other neurones connected to it. For example, the state of each input neurone is defined by the incoming signal (i.e. values) of the predictor variables and the states of the other neurones are evaluated locally by calculating the weighted sum of the incoming signals from the neurones of the previous layer. The entire process can be written mathematically as:

$$y_k = \phi_o\left\{\beta_k + \sum_j w_{jk}\phi_h\left(\beta_j + \sum_i w_{ij}x_i\right)\right\} \quad (1)$$

where x_i are the input signals, y_k are the output signals, w_{ij} are the weights between input neurone i and hidden neurone j, w_{jk} are the weights between hidden neurone j and output neurone k, β_j and β_k are the bias associated with the hidden and output layers, and ϕ_h and ϕ_o are activation functions for the hidden and output layers. There are several activation functions (see Bishop, 1995), but the logistic (or sigmoid) function was employed, as it is the most commonly used.

Training the neural network involves the backpropagation algorithm where the goal is to find a set of connection weights that minimises an error function. The cross-entropy criterion, similar to log-likelihood criterion, is minimised during network training for a dichotomous response variable:

$$E = -\sum_n \{t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\} \quad (2)$$

where t_n is the observed output value and y_n is the predicted output value for observation n. Observations are sequentially presented to the network and weights are adjusted after each output is calculated depending on the magnitude and direction of the error. This iterative technique of minimising the error is known as gradient descent, where weights are modified in the direction of greatest descent, travelling 'downhill' in the direction of the minimum.

The explanatory importance of each environmental variable was quantified by calculating the product of the connection weights (i.e. input-hidden × hidden-output weights) between its input neurone and the output neurone and then summing the products across all hidden neurones. This procedure is repeated for each environmental variable and the relative contributions of the variables were calculated by dividing the absolute value of each variable contribution by the grand sum of all absolute variable contributions. This method provides a measure of the explanatory importance of each environmental variable, which were subsequently assessed for their statistical significance using a randomisation test. This test randomises the response variable, then constructs a neural network using the randomised response variable and the original predictor variables and records the relative explanatory importance of each environmental variable. This process is repeated 9999 times to generate a null distribution for the relative importance of each variable, which is then compared with the observed values to calculate the significance level. By using the product of the connection weights (rather than merely summing the absolute value of each input-hidden and hidden-output connection weight separately), we account for the fact that the direction of the connection weights (i.e. positive or negative) can switch between different networks optimised with the same data (i.e. referred to as symmetrical interchanges of weights: Ripley, 1994) without having any effect on the relationship between the input and output neurone. Note that this phenomenon may result in incorrect estimates of variable importance using the commonly employed Garson's algorithm (Garson, 1991: see Olden & Jackson, 2002). We refer the reader to Olden & Jackson (2002) for more details on calculating variable contributions and testing their significance using the randomisation approach.

For all analyses, the optimal number of neurones in the hidden layer was determined empirically by comparing the performances of different cross-validated networks with one to 25 hidden neurones (one to five hidden neurones for the simulation experiment) and choosing the number that produced the greatest predictive performance. Learning rate (η) and momentum (α) parameters (varying as a function of network error) were included during network training to ensure a high probability of global network convergence (see Bishop, 1995 for details) and a maximum of 1000 iterations were used for the backpropagation algorithm to determine the optimal axon weights. Furthermore, to minimise the potential for network overfitting, we used the simplest network architecture (i.e. smallest number of hidden neurones) where equivalent network configurations exhibit identical predictive performance. Prior to training the network, the independent variables were converted to z-scores to standardise the measurement scales of the inputs into the network and thereby ensure that same percentage change in the weighted sum of the inputs caused a similar percentage change in the unit output.

Model construction, validation and performance metrics

To evaluate predictive performance, all models (including fish-habitat and simulation models) were

validated using a 'leave-one-out' cross-validation method. This method excludes one observation, constructs the model with the remaining $n-1$ observations and then predicts the response of the excluded observation using this model. This procedure is repeated n times so that each observation, in turn, is excluded in model construction and its response is predicted. Cross-validation was used as it has been shown to produce nearly unbiased estimates of prediction error compared with the commonly used resubstitution approach where the same data are used in model construction and validation (Olden & Jackson, 2000). This analysis provided an opportunity to accurately assess the transferability or generalisation of the models to other lakes in the same geographical region.

The output value from the LRA, LDA, CFT and ANN models range from 0 to 1, representing the probability of species occurrence in a particular lake. Rather than simply following the conventional decision threshold of 0.5 to classify a species as present or absent, we constructed Receiver-Operating Characteristic (ROC) plots for each species to estimate the predictive ability of the models over all decision thresholds (Fielding & Bell, 1997; Hand, 1997). An ROC graph is a plot of the sensitivity/specificity pairs (defined below) resulting from continuously varying the decision threshold over the entire range of results observed. The optimal decision threshold was chosen to maximise overall classification performance of the model, given equal costs of misclassifying the species as present or absent. Using the optimal decision threshold, we partitioned the overall classification success of each species model by deriving 'confusion matrices' following Fielding & Bell (1997). Using these matrices we examined three metrics of model performance. First, we quantified the overall classification performance of the model as the percentage of lakes where the model correctly predicted the presence or absence of the species (CC). Secondly, we examined the ability of the model to correctly predict species presence, termed model sensitivity (SE). Thirdly, we examined the ability of the model to correctly predict species absence, termed model specificity (SP). Cohen's κ statistic (Titus, Mosher & Williams, 1984) was used to assess whether the performance of the model differed from expectations based on chance alone as this measure is relatively independent of species prevalence or frequency of occurrence (Manel, Williams & Ormerod, 2001). Data characteristics of the habitat variables were screened prior to analyses, which resulted in using $\ln(x)$ transformed values of all continuous variables (except pH) for the LRA and LDA models, whereas the raw data were used in the CFT and ANN models. McNemar's test (with Yates correction for continuity; Zar, 1999) was used to compare patterns of lake misclassifications among LRA, LDA, CFT and ANN for each species. Spearman rank-correlation coefficients between absolute values of standardised regression coefficients (LRA), standardised canonical coefficients (LDA), relative percentage importance (CFT) and variable contributions (ANN) were used to compare patterns in the importance of the habitat variables for predicting the occurrence of each species.

All analysis for LRA, LDA and ANN were conducted using macros in the MatLab programming language (written by the authors) and CFT using Statistica software (StatSoft Inc., 1998).

Results

Comparison of modelling approaches: fish distributions in Algonquin Provincial Park

Predictability of species distributions. Predictability of species presence/absence, on average, was high across the 27 species based on the four modelling approaches with correct classification ranging between 80 and 85%, specificity ranging between 75 and 85% and sensitivity ranging between 35 and 70% (Fig. 3). Comparison among the approaches showed no significant differences in correct classification (Kruskal–Wallis: $H = 4.167$, $P = 0.244$) and specificity ($H = 4.281$, $P = 0.233$), but significant differences in sensitivity ($H = 15.984$, $P = 0.001$), which were attributed to significantly higher sensitivity of ANNs, compared with all other approaches (Mann–Whitney: $U = 170.0$–209.5, $P = 0.0007$–0.007).

Rates of correct classification, specificity and sensitivity varied among species and among modelling approaches (Table 3). Accounting for both the magnitude and directionality (i.e. specificity, sensitivity) of species predictability, Cohen's κ test showed that LRA, LDA, CFT and ANN produced 20, 20, 13 and 25 species-habitat models, respectively, whose numbers of correct predictions are greater than expectations based on chance. Because model performances are

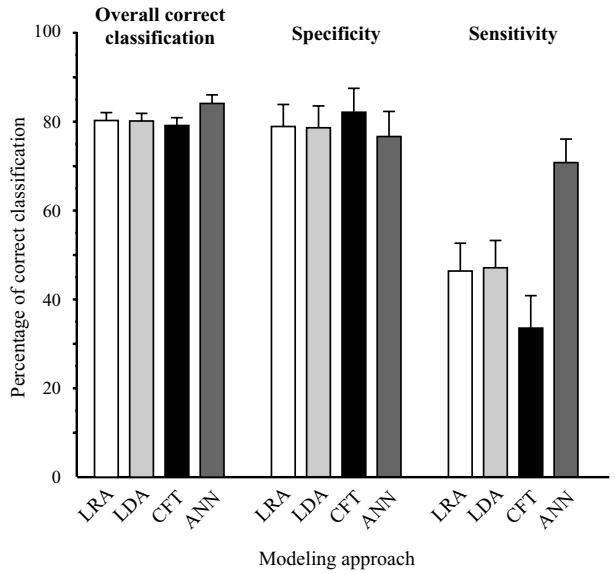

Fig. 3 Mean (and one standard deviation) for overall correct classification, specificity and sensitivity based on all 27 fish-habitat models developed using logistic regression (white), discriminant analysis (light grey), classification trees (black) and artificial neural networks (dark grey).

relatively independent from the influence of species prevalence in the data set, the interpretation of model predictions (i.e. species predictability in the study lakes) is acceptable. Comparisons among species show that for some species their absence was better predicted (e.g. cisco, smallmouth bass), for some species their presence was better predicted (e.g. brook trout, yellow perch) and others exhibited similar levels of specificity and sensitivity (e.g. brown bullhead, lake trout). Furthermore, although many species were correctly classified with equal success, directional strengths in their predictions often differed (e.g. smallmouth bass, brook trout). Comparisons among modelling approaches showed that the magnitude of differences in correct-classification success varied greatly from almost no difference (e.g. brown bullhead, pearl dace) to larger differences between ANN and the other three approaches (e.g. cisco, smallmouth bass) (Table 3).

McNemar's test indicated a number of differences among approaches in the identity of the lakes that were misclassified (synonymous with differences in the identity of the lakes correctly classified). Fig. 4 shows statistically significant pair-wise differences between modelling approaches, highlighting which approaches differed in patterns of lake misclassification for each fish species. Of the 45 significant pair-wise differences between approaches (note that no correction of α was made to maintain family wide error rate but only six pair-wise comparisons were made for each species), 30 (67%) were between linear and non-linear techniques (i.e. LRA or LDA and CFT or ANN), corresponding exactly with the expected proportion of pair-wise differences between linear and non-linear approaches (four of six comparisons = 67%). However, 38 of 45 pair-wise differences existed between ANN and the other three approaches. It is important to note that because all four approaches predicted species presence/absence with moderate to high degrees of success, the results from the McNemar's test generally represent differences in the overall classification success among the modelling approaches. For example, when all four approaches produced significant predictions of presence/absence (13 species: Table 3), minimal differences were observed in patterns in lake misclassification among the approaches (Fig. 4). The results showed that when the modelling approaches did exhibit differences in overall correct classification, the specific lakes that were misclassified also differed. For example, lakes where cisco, smallmouth bass and finescale dace were misclassified differed between ANN and the other three approaches and this result was related to the higher classification success of ANN (Table 3). However, patterns in lake misclassification also differed among approaches that had similar correct classification rates (e.g. longnose sucker, rock bass: Fig. 4), illustrating that, although the approaches misclassified the same number of lakes, different lakes were misclassified by the various methods.

Variable importance in predicting species distributions. Variable contributions in the fish-habitat models varied among species and among modelling approaches. Averaged across the modelling approaches, measures of overall lake size (i.e. surface area, lake volume, shoreline perimeter, maximum depth) were consistently important in model predictions, whereas the contribution of the other habitat variables varied among the species (Table 4). For example, lake altitude, growing-degree days and catchment identity were the most important predictors of smallmouth bass occurrence and lake size and the presence of a piscivore contributed the

Table 3 Summary of performance metrics of species-habitat models. Reported values are percentage correctly classified (CC), specificity (SP: ability to correctly predict species absence) and sensitivity (SN: ability to correctly predict species presence). Predictions significantly different from random (based on κ statistic) are indicated in bold ($P < 0.05$). Species codes are defined in Table 1

	Logistic regression			Discriminant analysis			Classification tree			Neural network		
Species	CC	SP	SN	CC	SP	SN	CC	SP	SN	CC	SP	SN
BCS	94	100	0	94	99	6	94	100	0	**99**	**100**	**75**
BNS	**73**	**84**	**59**	**72**	**81**	**59**	**69**	**95**	**31**	**78**	**71**	**88**
BSB	**74**	**93**	**23**	**75**	**94**	**24**	73	100	0	**80**	**80**	**79**
BT	**79**	**24**	**95**	**79**	**24**	**96**	**79**	**23**	**96**	**83**	**27**	**100**
BB	**66**	**70**	**61**	**66**	**71**	**61**	**66**	**79**	**52**	**64**	**98**	**25**
B	**79**	**91**	**53**	**78**	**89**	**55**	**76**	**96**	**30**	**87**	**86**	**88**
C	**84**	**94**	**46**	**84**	**95**	**44**	79	100	0	**91**	**97**	**66**
CC	**76**	**44**	**90**	**76**	**43**	**91**	**74**	**47**	**87**	**76**	**31**	**97**
CS	**70**	**63**	**75**	**72**	**66**	**76**	**76**	**76**	**75**	**77**	**63**	**89**
F	90	99	16	89	98	16	89	100	0	**98**	**100**	**78**
FSD	85	100	2	86	100	7	85	100	0	**93**	**99**	**56**
GS	**66**	**83**	**41**	**66**	**83**	**40**	**66**	**89**	**32**	**71**	**61**	**85**
ID	**85**	**96**	**43**	**84**	**95**	**40**	80	100	0	**87**	**93**	**67**
LC	**86**	**99**	**17**	85	99	13	84	100	0	84	100	0
LT	**80**	**76**	**83**	**81**	**77**	**84**	**76**	**86**	**67**	**83**	**69**	**95**
LW	86	99	5	86	98	5	86	100	0	**93**	**97**	**67**
LS	**85**	**97**	**28**	**84**	**96**	**28**	82	100	0	**88**	**100**	**40**
NRD	**63**	**50**	**74**	**63**	**48**	**74**	**63**	**51**	**73**	**69**	**42**	**92**
PD	**68**	**79**	**52**	**67**	**79**	**52**	**66**	**88**	**35**	**67**	**98**	**22**
PKS	**71**	**36**	**89**	**71**	**37**	**90**	**72**	**43**	**87**	**72**	**21**	**99**
RB	**94**	**100**	**51**	**91**	**96**	**51**	**92**	**99**	**43**	**97**	**100**	**77**
RW	89	98	7	90	99	10	90	100	0	**93**	**95**	**73**
SMB	**83**	**96**	**33**	**82**	**96**	**31**	79	100	0	**91**	**96**	**74**
SL	92	100	0	92	100	0	92	100	0	94	100	18
T-P	91	99	19	**92**	**99**	**30**	91	100	0	**95**	**98**	**63**
WS	**85**	**30**	**96**	**86**	**34**	**96**	**89**	**47**	**97**	**88**	**32**	**100**
YP	**74**	**30**	**93**	**74**	**28**	**93**	71	0	100	**76**	**16**	**100**

most to predictions of northern redbelly dace occurrence.

Based on ranked importance of the habitat variables in the models of the 27 species, there were a number of similarities and differences in the relative predictive importance of the habitat variables among modelling approaches. The results based on all species showed that the modelling approaches showed close agreement in the importance of particular variables (e.g. shoreline perimeter, altitude), but disagreed in the importance of others (e.g. total dissolved solids, pH) for predicting species occurrence (Fig. 5). Overall differences in the mean ranked importance of habitat variables existed among all approaches, but were most notable between linear and non-linear approaches. For example, the importance of summer stratification and presence of a piscivore (both dichotomous variables) were higher for both CFT and ANN compared with LRA and LDA.

Correlation analysis of ranked variables indicated many differences in the importance of the habitat variables for predicting the occurrence of each species. Significant pair-wise differences (again not correcting the Type I error rate for multiple comparisons) between modelling approaches highlight which methods differed in their ranked importance of the habitat variables in each species model (Fig. 6). For example, the approaches exhibited marked differences in the importance of the habitat variables for predicting the occurrence of creek chub, northern redbelly dace and pearl dace, whereas they were in agreement for other species, including lake whitefish, round whitefish and smallmouth bass. Of the 61 pair-wise differences between approaches, 49 (80%) were between linear and non-linear techniques, a value substantially greater than the expected 67%. Similar to the results from the McNemar's tests, ANN participated in the greatest number of pair-wise differences (41) com-

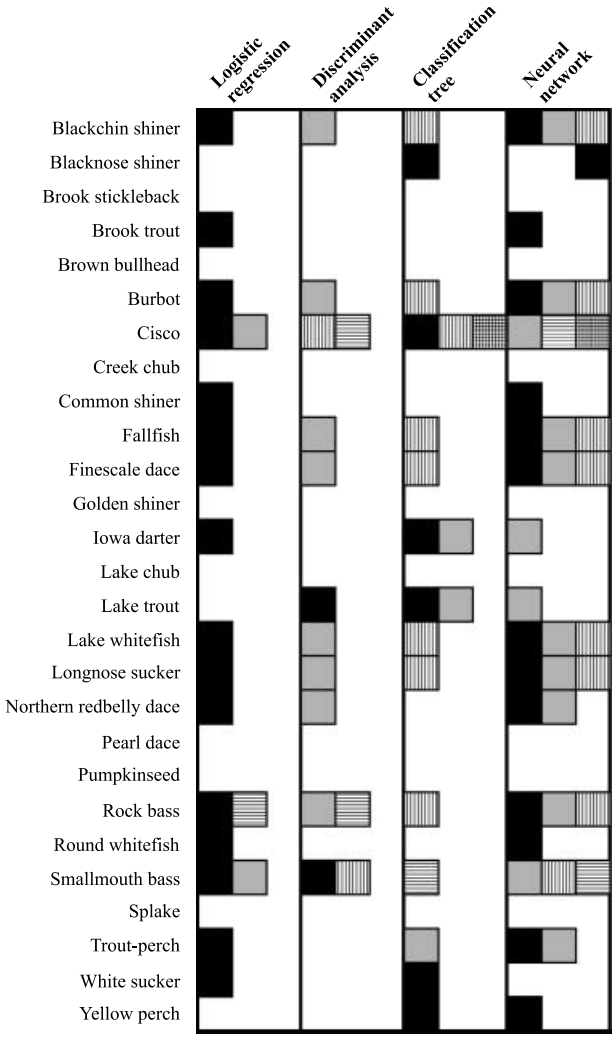

Fig. 4 Results from McNemar's test assessing differences among patterns of lake misclassifications using logistic regression, discriminant analysis, classification trees and artificial neural networks. Shared shading for a species represents significant differences in the sets of lakes in which species occurrence was incorrectly predicted ($P < 0.05$). For example, blackchin shiner was misclassified in different sets of lakes based on the logistic regression model and the neural network.

pared with the other approaches (LRA: 27; LDA: 33; and CFT: 21).

Concordance between patterns in species predictability and habitat variable importance. Differences in species predictability and variable importance among the modelling approaches resulted in four general relationships (Figs 4 & 6). First, several species exhibited minimal differences in patterns of species predictability and variable importance among the approaches (e.g. brown bullhead, round whitefish, yellow perch). Secondly, a number of species (e.g. burbot, rock bass) exhibited significant differences both in patterns of misclassification and in the importance of the variables for making model predictions. Thirdly, the predictability of some species (e.g. cisco, creek chub, longnose sucker) did not differ among the approaches, yet the importance of the habitat variables differed significantly. Fourthly, there were also a number of species where the modelling approaches exhibited different correct classification rates, but the importance of the variables in the models were similar (e.g. finescale dace, smallmouth bass).

Comparison of modelling approaches: simulated data

Results from the Monte Carlo simulations show that for linear species response curves, all modelling approaches exhibited high correct classification rates (Fig. 7). On average, CFT and ANN correctly classified 100% of the observations in all 500 simulated data sets, whereas LRA and LDA correctly classified 97 and 92% of the cases, respectively. In contrast, the two non-linear modelling approaches greatly out-performed the linear approaches for the Gaussian or non-linear species response curve, where CFT and ANN had mean correct-classification rates of 89 and 98%, and LRA and LDA had mean correct-classification rates of 52 and 49%, respectively (corresponding to completely random predictions because the data sets were generated to exhibit species prevalence of 50%) (Fig. 7).

Discussion

Linear and non-linear approaches to developing predictive models should be viewed as both competitive and complementary methodologies for establishing quantitative linkages between species and their environment. Although taking such a comparative approach is favourable, the number of studies examining multiple statistical techniques for modelling species-environment relationships is surprisingly small (Guisan & Zimmermann, 2000). To the best of our knowledge, our study is the first comprehensive comparison of traditional and alternative modelling approaches using both empirical and simulated data. This treatment is especially timely because alternative statistical methodologies, especially classification trees and neural networks, have recently been

Table 4 Mean rank importance of the habitat variables for predicting species presence/absence. Mean rank importance was calculated by averaging the rank of each variable across the four modelling approaches. Species codes are defined in Table 1 and variable codes are defined in Table 2. The three dummy variables coding the drainage basins (DB Code) were averaged to produce the single value shown

	Habitat variables										
Species	S. Area	Volume	Sh. Per.	Max. Dep.	TDS	pH	Altitude	GDD	SS	DB Code	Piscivore
BCS	2.0	5.0	7.8	8.3	5.5	4.0	7.5	4.0	7.3	5.8	3.0
BNS	4.5	5.3	2.5	4.0	9.8	11.8	10.8	10.3	9.8	5.4	6.0
BSB	1.8	2.5	4.5	7.3	6.3	7.8	5.5	2.5	7.5	5.8	8.5
BT	3.0	3.8	5.8	5.0	12.0	9.0	10.5	5.3	3.5	6.7	–
BB	3.0	6.8	2.8	4.3	10.0	9.0	9.0	11.3	9.0	4.3	–
B	3.8	3.5	3.5	5.3	6.3	9.0	8.8	5.8	8.8	7.7	–
C	1.0	2.0	5.3	4.8	7.5	6.8	4.3	4.8	7.8	5.8	–
CC	3.0	3.3	3.5	7.0	10.5	10.0	6.0	9.0	10.3	6.1	10.0
CS	2.8	4.8	2.0	7.8	5.0	8.3	10.3	10.8	11.5	7.3	5.5
F	4.0	6.5	1.5	8.8	8.0	7.8	4.5	6.0	8.3	3.2	6.8
FSD	2.0	3.3	2.5	8.0	6.0	3.8	6.3	3.5	9.3	6.6	7.3
GS	2.3	2.3	5.3	5.3	9.3	6.5	6.8	8.0	7.8	10.4	6.3
ID	2.0	5.3	4.0	5.3	4.5	7.5	8.8	6.5	6.3	4.0	9.5
LC	5.3	1.0	4.5	3.8	7.0	7.5	3.3	7.3	8.5	6.7	3.5
LT	2.3	2.8	2.3	3.3	8.8	7.8	10.3	6.0	8.8	8.5	–
LW	1.0	2.3	3.3	6.3	5.8	7.0	3.8	8.0	6.5	5.9	–
LS	1.3	1.8	4.5	6.8	4.0	8.0	7.0	8.3	6.0	4.7	–
NRD	3.0	6.3	3.8	3.0	9.8	7.3	7.5	9.8	10.0	8.8	4.0
PD	1.8	5.3	1.5	9.3	10.3	10.8	7.0	7.3	9.3	7.4	5.8
PKS	2.3	5.8	5.3	6.0	6.0	7.0	2.3	5.5	7.0	10.3	–
RB	7.5	8.0	7.3	10.5	6.3	4.0	4.0	8.0	8.0	4.5	–
RW	3.8	4.5	5.8	7.5	6.8	5.5	7.5	8.5	4.0	2.6	–
SMB	4.5	4.8	4.8	7.8	7.0	9.0	2.0	3.3	8.3	3.4	–
SL	4.8	2.3	5.0	5.8	6.5	7.0	2.3	3.8	7.3	5.7	–
T-P	2.8	3.8	5.3	6.8	5.8	8.5	4.5	8.5	4.5	4.5	7.8
WS	2.0	6.8	4.0	5.3	5.5	7.3	8.3	10.5	7.8	6.9	–
YP	1.8	7.0	3.3	5.0	1.5	4.5	5.5	6.0	5.0	7.3	–

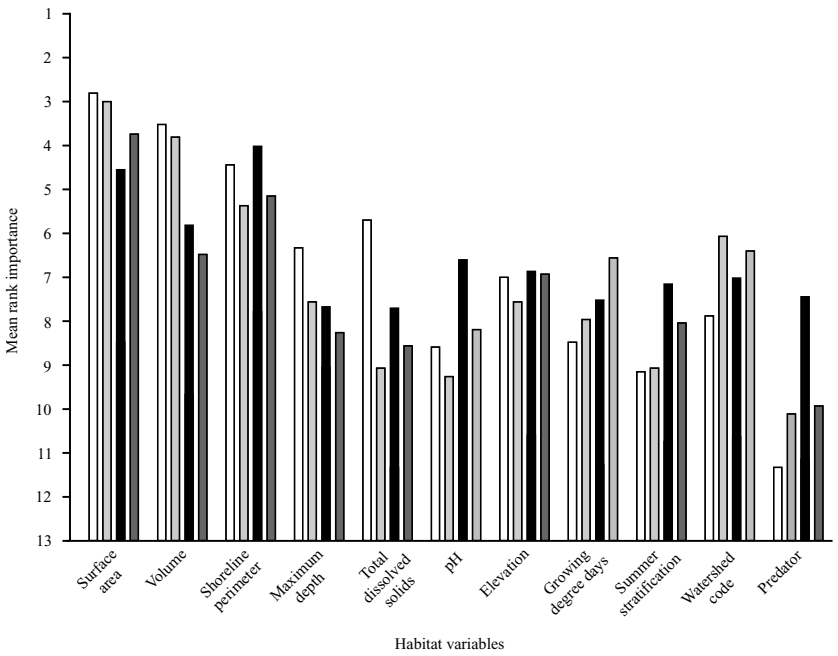

Fig. 5 Mean rank importance (average importance of each habitat variable based on all 27 species-habitat models) of the habitat variables for predicting species presence/absence using logistic regression (white), discriminant analysis (light grey), classification trees (black) and artificial neural networks (dark grey). Note that standard deviations were omitted for clarity and that the three dummy variables coding the drainage basins were averaged to produce the single value shown.

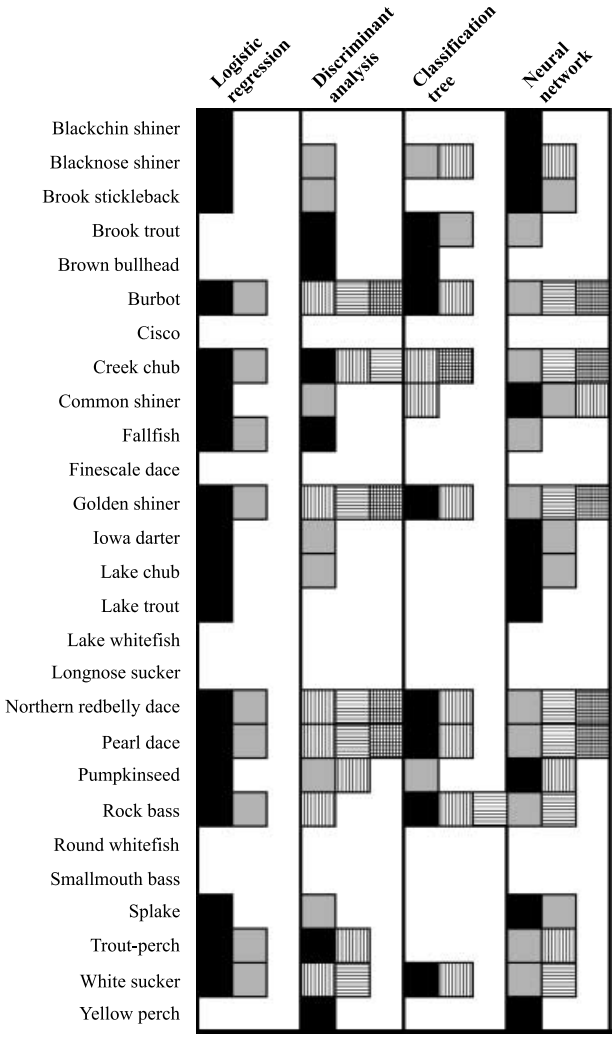

Fig. 6 Results from Spearman rank correlation test assessing differences among relative importance of the habitat variables for predicting species presence/absence using logistic regression, discriminant analysis, classification trees and artificial neural networks. Shared shading for a species represents significant differences between models in the importance of the habitat variables for predicting its occurrence ($P < 0.05$). For example, the ranked importance of the habitat variables in the blackchin shiner model differed between logistic regression and the neural network.

introduced and advocated as attractive modelling approaches in the ecological literature (Lek *et al*., 1996; De'ath & Fabricius, 2000).

Comparison of modelling approaches: predictive power

We found that average predictive performances of LRA, LDA, CFT and ANN were similar across all species, although differences in the directionality of

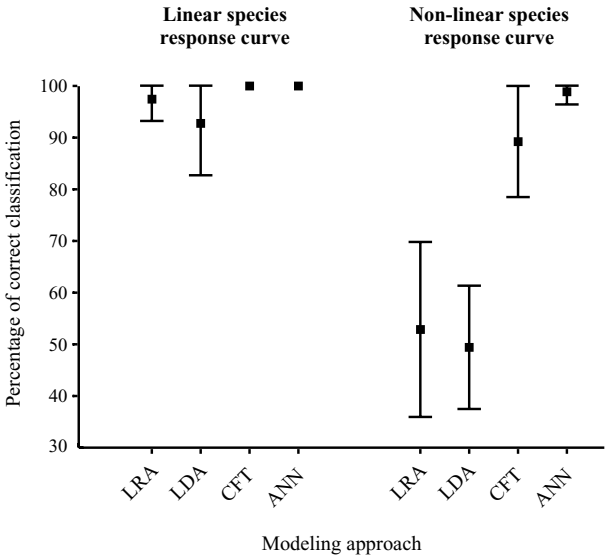

Fig. 7 Whisker plots comparing the correct classification rates for logistic regression (LRA), discriminant analysis (LDA), classification trees (CFT) and artificial neural networks (ANN) based on simulated, linear and non-linear species response curves. Mean values of 500 simulated data sets are shown and the whiskers represent one standard deviation.

correct prediction were evident as CFT exhibited the highest specificity and neural networks exhibited the highest sensitivity. Furthermore, the predictability for individual species varied greatly, emphasising that no single linear or non-linear approach was optimal for all species (although neural networks produced the greatest number of statistically significant models). Recent studies modelling species presence/absence have shown the predictive advantages of LRA (e.g. Manel *et al*., 1999; Özesmi & Özesmi, 1999), LDA (e.g. Reichard & Hamilton, 1997; Scheller *et al*., 1999), CFT (e.g. Rejwan *et al*., 1999) and ANN (e.g. Mastrorillo *et al*., 1997; Olden & Jackson, 2001) relative to their linear or non-linear counterparts.

Where the underlying data structure and assumptions are met for a particular statistical method, there is no reason to expect major differences in the suitability between traditional and alternative techniques. For example, one might expect LRA and LDA to perform as well where linear relationships exist, whereas CFT and ANN should prove better in non-linear situations. Indeed, the results from our simulation experiment support this expectation. The CFTs and ANNs were shown to be superior to linear approaches for nonlinearly distributed data (i.e. Gaussian species response curves), whereas all four

approaches showed nearly identical predictive power for the linear species response curves. The use of simulated data to compare statistical techniques is preferable because the properties of the data are known (properties which cannot be determined, but only estimated from field data) and thus true differences among techniques can be accurately assessed.

However, simulation studies are not without problems. If not designed properly, simulation studies are likely to be biased towards the success of particular methods. For example, if all the data were generated from two multivariate normal classes with equal covariance matrices, LDA would perform better than LRA (Efron, 1975). This is a good example because the above assumption of variable distribution was undoubtedly violated when we sampled from our simulated data sets, which helps explain why LRA tended to outperform LDA (this is also supported by Press & Wilson, 1978). Furthermore, our simulated data sets are idealised species response curves and have no error associated with the species-habitat relationships. Although this perfect relationship will not be found in field data, it was used to highlight the major differences among the methods. However, it is important to state as the error distribution may bias the results in favour of particular methods depending on the individual data set. More detailed simulation studies using a broader variety of known data conditions are needed to more accurately compare the traditional and alternative statistical approaches. The results from our simulation study, however, clearly show that non-linear modelling approaches to modelling species-environment relationships should be favoured because the non-linear approaches should perform as well as linear methods when the data show linear relationships. When the data relationships are non-linear, the non-linear modelling approaches should provide superior results. Consequently, although non-linear approaches (such as the techniques used in this study) have greater costs (e.g. computational time and effort) associated with their use compared with linear approaches, we recommend that such approaches should be considered when modelling ecological data given that: (1) patterns within ecological data are commonly non-linear in nature, (2) different model solutions (i.e. predictive power and model parameters) may arise as a result of specific choices of transformations and (3) achieving linearity is often not possible.

Comparison of modelling approaches: explanatory insight

We found that the degree of similarity in the relative importance of the habitat variables was highly variable among the modelling approaches. For more than half of the species, little difference among ranked variable importance was evident, indicating that all approaches established similar quantitative relationships between the habitat variables and species occurrence (although we did not compare the directionality of the relationships). For the remaining species, there were substantial differences in the relative contribution of each habitat variable to predictions of species occurrence. The majority of these differences were observed between linear and non-linear; a finding which is not surprising given the inherent differences between these two types of approaches. For example, the recursive-partitioning algorithm of CFT and the backpropagation algorithm of ANN have a number of advantages over the training algorithms of LRA and LDA, including their ability to handle mixed data types, model non-linear relationships and capture non-additive behaviour without having to specify *a priori* the form of the interactions (Breiman *et al*., 1984; Bishop, 1995). Therefore, these advantages are most likely responsible for the predictive and explanatory differences between linear and non-linear approaches. For example, for a number of species where either ANN or CFT exhibited greater predictive power compared with the linear approaches (e.g. brook stickleback, Iowa darter, lake chub, pumpkinseed, white sucker), they also exhibited differences in the relative importance of the habitat variables for producing those predictions. These findings follow given that CFT and ANN are better suited to quantify non-linear relationships between species and environmental variables and may result in a better representation of variable importance for predicting species occurrence. Furthermore, the idea that CFT and ANN can more readily model mixed data types is supported by the result that two dichotomous variables (summer stratification and presence of a piscivore) had greater contributions in the CFTs and ANNs compared with the linear approaches.

In summary, similarities or differences in the relative importance of the habitat variables for predicting species occurrence corresponded generally with similarities or differences in the classification success of the models among the statistical approaches. Specifically, when models differed in classification success, so did the ranked importance of the habitat variables in the models. Given the results from the empirical and simulation components of this study and others, we believe that non-linear techniques provide a more flexible set of analytical tools for modelling ecological data compared with traditional, linear techniques, as they can model either linear or non-linear species response curves. Furthermore, the fact that a number of species models showed different patterns of variable importance, even when the approaches correctly predicted the occurrence in a similar number and set of lakes, is an important finding. This result suggests that traditional, linear approaches to modelling species distributions (without the inclusion of interaction terms) are providing a different representation of the importance of habitat factors shaping species distributions relative to that provided by non-linear approaches.

Predictability of fish species distributions: implications for conservation and management

More effective conservation of aquatic biodiversity will require new approaches that recognise the protection of key local- and regional-scale processes that shape fish distributions (Angermeier & Winston, 1999). Developments in these areas require an increased reliance on probabilistic models and will represent an important advancement in both population and community ecology. Our study shows that statistical modelling approaches exhibit considerable promise in providing testable, predictive models for fish ecology. Although the models presented here are correlative, and thus we cannot determine, but only imply causation, the results are consistent with findings from many studies of north-temperate fish populations (e.g. Jackson & Harvey, 1989; Tonn *et al*., 1990; Magnuson *et al*., 1998). The fact that many species were highly predictable from measures of whole-lake habitat features is promising, especially for species such as smallmouth bass and rock bass, which adversely impact littoral prey fish abundance and diversity in north-temperate lakes (Whittier, Halliwell & Paulsen, 1997; Findlay, Bert & Zheng, 2000; MacRae & Jackson, 2001) and can have competitive impacts on populations of native top predators by reducing prey fish populations (Vander Zanden, Casselman & Rasmussen, 1999; Jackson, 2002). Therefore, predictive models can play an important role by forecasting the likelihood of local establishment and spread of non-native species and thus help set proactive conservation priorities for preserving vulnerable populations.

Patterns of species occurrence were found to be related to various aspects of lake habitat. For example, the presence/absence of salmonids and burbot (cold-water species) were best predicted with variables describing overall lake size (i.e. surface area, shoreline perimeter, maximum depth). Surface area and maximum depth are known to influence the occurrence of these species (e.g. Jackson & Harvey, 1989) because these factors influence the mixing characteristics and the thermal regime of lakes. Furthermore, lake area and depth serve as indirect measures of the diversity of habitats available in lakes, which may be important to support small-bodied, forage fish upon which these species feed. Lake altitude and the number of growing-degree days were the most important predictors of smallmouth bass occurrence, a finding consistent with the sensitive thermal requirements of this species (Shuter & Post, 1990). Presence of a littoral-zone piscivore had a strong contribution to predictions of lake chub and northern redbelly dace occurrence. This result is consistent with studies that suggest that the distributions of these species are greatly affected by piscivory (Whittier *et al*., 1997; Findlay *et al*., 2000; Jackson *et al*., 2001).

Although a thorough discussion of the potential applications of predictive models to aquatic conservation is beyond the scope of this paper, we contend that exploring alternative measures of model performance and the use of multiple statistical approaches will play critical roles in determining the potential utility of predictive models of species distributions. Conventionally, the predictive abilities of species distribution models are assessed from overall classification rates alone. However, we show that by partitioning the predictive performance of the models into measures such as sensitivity and

specificity, we can assess more readily the strengths and weaknesses of the models and better evaluate their applicability. For instance, we found that although the overall correct-classification rates for some species were similar, levels of specificity and sensitivity were often quite different. Examining alternative measures of prediction success can provide more accurate comparisons among different modelling approaches (e.g. Manel *et al*., 1999) and among species. Alternative metrics can also provide additional knowledge into the factors shaping species occurrence, as well as provide important insight into the importance and causes of model misclassifications, ultimately leading to the development of more robust predictive models. Although traditionally ignored by ecologists, the consideration of model sensitivity and specificity is important as they may impose limitations on the success of predictive models, particularly when used in applied contexts. For example, low model sensitivity for rare species implies that it will be more difficult to predict the occurrence of organisms whose conservation and management may be the most critical. This finding has great importance in developing models for guiding searches for populations in previously unsampled areas and for indicating site suitability for the reintroduction of rare species (e.g. Hill & Keddy, 1992; Wiser *et al*., 1998) because the predictive ability of the models will be limited. Low model specificity could limit our ability to monitor and predict local extinction events caused by habitat modification, as well as reduce our confidence in drawing inferences from observed absences of species from sites containing suitable habitat conditions (e.g. indirect evidence for dispersal, predation, competition).

Given that most data sets are expensive to collect both in terms of time and money, we believe that more effort should be spent in choosing and comparing different statistical methods that best suit the particular questions of interest and characteristics of the data at hand. By employing parallel modelling approaches with the same set of data and focusing on comparing multiple metrics of predictive performance, researchers can begin to choose predictive models that not only provide the greatest predictive power, but also those models that best fit the proposed application (e.g. maximising sensitivity for predicting potential sites for species re-introduction). Such advances will provide more statistically and biologically powerful predictions for applied aquatic conservation.

Acknowledgments

We would like to thank Nick Mandrak for providing the Algonquin Park Fish Inventory Data Base and Bob Bailey, Nick Collins, Laura Hartt, Pedro Peres-Neto, Brian Shuter, Keith Somers and two anonymous reviewers for their comments. Funding for this research was provided by a Natural Sciences and Engineering Council of Canada (NSERC) Graduate Scholarship and University of Toronto scholarships to J.D. Olden and an NSERC Research Grant to D.A. Jackson.

References

Angermeier P.L. & Winston M.R. (1999) Characterizing fish community diversity across Virginia landscapes: prerequisite for conservation. *Ecological Applications*, **9**, 335–349.

Bailey R.C., Kennedy M.G., Dervish M.Z. & Taylor R.M. (1998) Biological assessment of freshwater ecosystems using a reference condition approach: comparing predicted and actual benthic invertebrate communities in Yukon streams. *Freshwater Biology*, **39**, 765–774.

Bishop C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press Inc, New York, USA.

Breiman L., Friedman J.H., Olshen A. & Stone C.G. (1984) *Classification and Regression Trees.* Wadsworth International Group, Belmont, CA, USA.

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

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

Cheng B. & Titterington D.M. (1994) Neural networks: a review from a statistical perspective (with discussion). *Statistical Science*, **9**, 2–54.

Chessman B.C. (1999) Predicting the macroinvertebrate faunas of rivers by multiple regression of biological and environmental differences. *Freshwater Biology*, **41**, 747–757.

Crossman E.J. & Mandrak N.E. (1992) *An Analysis of Fish Distribution and Community Structure in Algonquin Park: Annual Report for 1991 and Completion Report, 1989–1991*. Ontario Ministry of Natural Resources, Toronto, Ontario, Canada.

D'Angelo D.J., Howard L.M., Meyer J.L., Gregory S.V. & Ashkenas L.R. (1995) Ecological uses for genetic algorithms: predicting fish distributions in complex physical habitats. *Canadian Journal of Fisheries and Aquatic Sciences*, **52**, 1893–1908.

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

Dodge D.P., Goodchild G.A., MacRitchie I., Tily J.C. & Waldriff D.G. (1985) *Manual of Instructions: Aquatic Habitat Inventory Surveys*. Ontario Ministry of Natural Resources, Fisheries Branch, Toronto, Ontario, Canada.

Efron B. (1975) The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, **70**, 892–898.

Evans D.O. & Oliver C.H. (1995) Introduction of lake trout to inland lakes of Ontario, Canada: Factors contributing to successful colonization. *Journal of Great Lakes Research*, **21** (Suppl. 1), 30–53.

Fausch K.D., Hawkes C.L. & Parsons M.G. (1988) *Models that Predict the Standing Crop of Stream Fish from Habitat Variables: 1950–1985.* General Technical Report PNW-GTR-213. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, OR, USA.

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

Findlay C.S., Bert D.G. & Zheng L. (2000) Effect of introduced piscivores on native minnow communities in Adirondack lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 570–580.

Garson G.D. (1991) Interpreting neural network connection weights. *Artificial Intelligence Expert*, **6**, 47–51.

Gaston K.J. & Blackburn T.M. (1999) A critique for macroecology. *Oikos*, **84**, 353–368.

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

Hand D.J. (1997) *Construction and Assessment of Classification Rules*. John Wiley & Sons, Chichester, England.

Hanna M. (1990) Evaluation of models predicting mixing depth. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 940–947.

Harig A.L. & Bain M.B. (1998) Defining and restoring biological integrity in wilderness lakes. *Ecological Applications*, **8**, 71–87.

Hill N.M. & Keddy P.A. (1992) Prediction of rarities from habitat variables: coastal plain plants on Nova Scotian landshores. *Ecology*, **73**, 1852–1857.

Hornik K., Stinchcombe M. & White H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366.

Jackson D.A. (2002) Ecological impacts of *Micropterus* introductions: the dark side of black bass. In: *Black Bass: Ecology, Conservation and Management* (Eds D. Phillip & M. Ridgway), pp. 221–232. American Fisheries Society, Bethesda, MD.

Jackson D.A. & Harvey H.H. (1989) Biogeographic associations in fish assemblages: local versus regional processes. *Ecology*, **70**, 1472–1484.

Jackson D.A. & Harvey H.H. (1997) Qualitative and quantitative sampling of lake fish communities. *Canadian Journal of Fisheries and Aquatic Sciences*, **54**, 2807–2813.

Jackson D.A., Peres-Neto P.R. & Olden J.D. (2001) What controls who is where in freshwater fish communities – the roles of biotic, abiotic, and spatial factors. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 157–170.

James F.C. & McCulloch C.E. (1990) Multivariate analysis in ecology and systematics: panacea or Pandora's box? *Annual Review of Ecology and Systematics*, **21**, 129–166.

Jongman R.H.G., ter Braak C.J.F. & van Tongeren O.F.R. (1995) *Data Analysis in Community and Landscape Ecology*. Cambridge University Press, Cambridge, England.

Keddy P.A. (1992) Assembly and response rules: two goals for predictive community ecology. *Journal of Vegetation Science*, **3**, 157–164.

Kurková V. (1992) Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, **5**, 501–506.

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

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

Lim T.S., Loh W.Y. & Shih Y.S. (2000) A comparison of prediction accuracy, complexity, and training time of thirty-tree old and new classification algorithms. *Machine Learning*, **40**, 203–228.

Loh W.Y. & Shih Y.S. (1997) Split selection methods for classification trees. *Statistica Sinica*, **7**, 815–840.

MacRae P.S.D. & Jackson D.A. (2001) The influence of smallmouth bass (*Micropterus dolomieu*) predation and habitat complexity on the structure of littoral zone fish assemblages. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 342–351.

Magnuson J.J., Tonn W.M., Banerjee A., Toivonen J., Sanchez O. & Rask M. (1998) Isolation vs. extinction in the assembly of fishes in small northern lakes. *Ecology*, **79**, 2941–2956.

Manel S., Dias J.M., Buckton S.T. & Ormerod S.J. (1999) Comparing discriminant analysis, neural networks and logistic regression for predicting species distribution: a case study with a Himalayan river bird. *Ecological Modelling*, **120**, 337–347.

Manel S., Williams H.C. & Ormerod S.J. (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921–931.

Mastrorillo S., Lek S., Dauba F. & Belaud A. (1997) The use of artificial neural networks to predict the presence of small-bodied fish in a river. *Freshwater Biology*, **38**, 237–246.

Matuszek J.E. & Beggs G.L. (1988) Fish species richness in relation to lake area, pH, and other abiotic factors in Ontario lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **45**, 1931–41.

Minns C.K. (1989) Factors affecting fish species richness in Ontario lakes. *Transactions of the American Fisheries Society*, **118**, 533–545.

Morgan J.N. & Sonquist J.A. (1963) Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, **58**, 415–434.

Moss D.M., Wright J.F., Furse M.T. & Clarke R.T. (1999) A comparison of alternative techniques for prediction of the fauna of running-water sites in Great Britain. *Freshwater Biology*, **41**, 167–181.

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

Olden J.D. & Jackson D.A. (2000) Torturing the data for the sake of generality: how valid are our regression models? *Écoscience*, **7**, 501–510.

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

Olden J.D. & Jackson D.A. (2002) Illuminating the 'black box': a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, **154**, 135–150.

Orians G.H. (1980) Micro and macro in ecological theory. *Bioscience*, **30**, 79.

Özesmi S.L. & Özesmi U. (1999) An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling*, **166**, 15–31.

Peterson A.T. & Vieglais D.A. (2001) Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. *Bioscience*, **51**, 363–371.

Pickett S.T.A., Kolasa J. & Jones C.G. (1994) Ecological understanding: the nature of theory and the theory of nature.

Press S.J. & Wilson S. (1978) Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, **73**, 699–705.

Rathert D., White D., Sifneos J.C. & Hughes R.M. (1999) Environmental correlates of species richness for native freshwater fish in Oregon, U.S.A. *Journal of Biogeography*, **26**, 257–273.

Reichard S.H. & Hamilton C.W. (1997) Predicting invasions of woody plants introduced into North America. *Conservation Biology*, **11**, 193–203.

Rejwan C., Collins N.C., Brunner J., Shuter B.J. & Ridgway M.S. (1999) Tree regression analysis on the nesting habitat of smallmouth bass. *Ecology*, **80**, 341–348.

Ricciardi A. & Rasmussen J.B. (1999) Extinction rates of North American freshwater fauna. *Conservation Biology*, **13**, 1220–1222.

Richter B.D., Braun D.P., Mendelson M.A. & Master L.L. (1997) Threats to imperiled freshwater fauna. *Conservation Biology*, **11**, 1081–1093.

Ripley B.D. (1994) Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society, Series B*, **56**, 409–456.

Rodriguez M.A. & Lewis W.M. Jr (1997) Structure of fish assemblages along environmental gradients in floodplain lakes of the Orinoco River. *Ecological Monographs*, **67**, 109–128.

Rumelhart R.E., Hinton R.J. & Williams R.J. (1986) Learning representations by back-propagating error. *Nature*, **323**, 533–536.

Scheller R.M., Snarski V.M., Eaton J.G. & Oehlert G.W. (1999) An analysis of the influence of annual thermal variables on the occurrence of fifteen warmwater fishes. *Transactions of the American Fisheries Society*, **128**, 257–264.

Shuter B.J. & Post J.R. (1990) Climate, population viability, and the zoogeography of temperature fishes. *Transactions of the American Fisheries Society*, **119**, 314–336.

StatSoft Inc. (1998) *STATISTICA for Windows*. Tulsa, OK.

Titus K., Mosher J.A. & Williams B.K. (1984) Chance-corrected classification for use in discriminant analysis: ecological applications. *American Midland Naturalist*, **111**, 1–7.

Toner M. & Keddy P. (1997) River hydrology and riparian wetlands: a predictive model for ecological assembly. *Ecological Applications*, **7**, 236–246.

Tonn W.M., Magnuson J.J., Rask M. & Toivonen J. (1990) Intercontinental comparison of small-lake fish assemblages: the balance between local and regional processes. *American Naturalist*, **136**, 345–375.

Vander Zanden M.J., Casselman J.M. & Rasmussen J.B. (1999) Stable isotope evidence for the food web consequences of species invasions in lakes. *Nature*, **401**, 464–467.

Whittier T.R., Halliwell D.B. & Paulsen S.G. (1997) Cyprinid distributions in Northeast U.S.A. lakes:

evidence of regional-scale minnow biodiversity losses. *Canadian Journal of Fisheries and Aquatic Sciences*, **54**, 1593–1607.

Williams P.H. & Araujo M.B. (2000) Using probability of persistence to identify important areas for biodiversity conservation. *Proceedings of the Royal Society of London B*, **267**, 1959–1966.

Williams J.E., Johnson J.E., Hendrickson D.A., Contreras-Bladeras S., Williams J.D., Navarro-Mendoza M., McAllister D.E. & Deacon J.E. (1989) Fishes of North America: endangered, threatened, or of special concern. *Fisheries*, **14**, 2–20.

Wiser S.K., Peet R.K. & White P.S. (1998) Prediction of rare-plant occurrence: a southern Appalachian example. *Ecological Applications*, **8**, 909–920.

Zar J.H. (1999) *Biostatistical Analysis*, 4th edn. Prentice-Hall, Upper Saddle River, NJ, USA.

(Manuscript accepted 19 April 2002)