Freshwater Biology (2005) **50,** 539–551 doi:10.1111/j.1365-2427.2005.01340.x

Predicting microhabitat selection in juvenile Atlantic salmon *Salmo salar* by the use of logistic regression and classification trees

KATRINE TURGEON AND MARCO A. RODRÍGUEZ
Département de chimie-biologie, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada

SUMMARY

1. We compared the capacity of logistic regression (LR) and classification tree (CT) models to predict microhabitat use and the summer distribution of juvenile Atlantic salmon, *Salmo salar*, in two reaches of a small stream in eastern Quebec.
2. The models predicted the presence or absence of salmon at a location on the basis of habitat features (depth, current velocity, presence of instream and overhead cover, substratum particle size, and distance to stream bank) measured at that location. Models were validated by means of crossover field tests evaluating the performance of models developed for one reach (calibration trials) when applied to the other reach (validation trials). Model performance was evaluated with regard to accuracy, generality and ease of use and interpretation. Prediction maps based on habitat features were also built to compare the observed position of fish with those predicted by LR and CT models.
3. The spatial distribution of active fish differed markedly from that of resting fish, apparently as a result of the selection for water greater than about 30 cm depth by active fish and for the presence of rocky cover by resting fish.
4. All models made accurate predictions, validated by crossover trials. For both LR and CT models, the prediction maps reflected well the actual fish distributions. However, CT models were easier to build and interpret than LR models. CT models also had less variable performance and a smaller decline in predictive capability in crossover trials (for fish at rest), suggesting that they may be more transferable than LR models.

Keywords: classification methods, fish behaviour, habitat modelling, *Salmo salar*, stream habitats

Introduction

Habitat selection by juvenile salmonids during their stream-dwelling phase reflects variable tradeoffs between net energy gain and avoidance of various risks, such as predation, stranding, and entrapment or injury by ice. As such, habitat selection is a dynamic and flexible process, which responds to temporal and spatial variations in habitat conditions (Heggenes *et al*., 2002). Previous studies have shown that habitat selection in stream-dwelling fish is influenced by net energetic gain from foraging on invertebrate drift (Fausch, 1984; Hughes & Dill, 1990; Hill & Grossman, 1993), swimming costs associated with current velocity (Fausch, 1984), predation risk (Metcalfe, Huntingford & Thorpe, 1987; Gotceitas & Godin, 1993; Gregory & Griffith, 1996), agonistic interactions (Kalleberg, 1958; Fausch & White, 1981), and the availability of instream (Cunjak, 1988; Gries & Juanes, 1998) and overhead cover (Shirvell, 1990; Grand & Dill, 1997). Instream structures, such as large unembedded rocks, can provide refuge from predators and fast flow (Fausch, 1984) and reduce agonistic interactions by preventing individuals seeing each other (Kalleberg, 1958; Fausch & White, 1981).

Correspondence: Marco A. Rodríguez, Département de chimie-biologie, Université du Québec à Trois-Rivières, C.P. 500, Trois-Rivières, Québec, G9A 5H7, Canada.
E-mail: marco_rodriguez@uqtr.ca

Many models of microhabitat selection do not distinguish between active behaviour, such as foraging, and more passive behaviour, such as resting and sheltering. However, this distinction may be informative in some cases. For example, use of cover can vary in space and time as a function of predation risk, hydrological fluctuations (Gotceitas & Godin, 1993; Giannico & Healey, 1999), or ontogenetic development (Cunjak, 1988; Gries & Juanes, 1998). Also, daytime sheltering by juvenile Atlantic salmon, *Salmo salar* L., in summer appears to be more common than previously thought and may be a key factor affecting production (Gries & Juanes, 1998). Therefore, distinguishing between active and resting behaviour in microhabitat models may enhance our understanding of the habitat needs of stream salmonids and provide more accurate predictions of their spatial distribution.

To be useful as conservation and management tools, habitat models should be accurate (correctly predict presence and absence), general (transferable to new sites), and easily applied (parsimonious, readily interpretable) (Lek *et al*., 1996; Guisan & Zimmermann, 2000). Logistic regression (LR) (Hosmer & Lemeshow, 2000) and classification trees (CT) (Breiman *et al*., 1984) are powerful tools for modelling ecological data (Manel, Dias & Ormerod, 1999; De'ath & Fabricius, 2000; Olden & Jackson, 2002). CTs offer several advantages over conventional linear models: they can readily detect complex interactions among predictors, are relatively easy to conceptualise and represent graphically, and have no distributional assumptions (Breiman *et al*., 1984; Rejwan *et al*., 1999; De'ath & Fabricius, 2000). Although both techniques have been used to model habitat selection in salmonids (LR: Rieman & McIntyre, 1995; Knapp & Preisler, 1999; Torgersen *et al*., 1999; Guay *et al*., 2000; CT: Stoneman & Jones, 2000), we know of no studies that directly compare the two techniques in this context.

Validation and assessment of performance are critical steps in developing useful models (Fielding & Bell, 1997; Manel, Williams & Ormerod, 2001; Olden, Jackson & Peres-Neto, 2002). Data-partitioning techniques (Olden *et al*., 2002) are often used to validate 'internally' a model based on statistical properties of a single data set whenever independent data are not available. However, examining the predictive performance of models when applied to new or independent data is a more rigorous, and thus preferable, method of 'external' validation (Verbyla & Litaitis, 1989; 'prospective sampling' *sensu* Fielding & Bell, 1997). To assess model performance, many studies of habitat selection rely solely on the percentage of correctly predicted presences and absences, or accuracy, a measure calculated from the confusion matrix (cross-tabulated values for observed versus predicted presence and absence). However, accuracy may be artificially inflated when the prevalence (frequency of occurrence) is low (Fielding & Bell, 1997). Other measures of model performance, such as Cohen's kappa (κ), Matthews correlation (MC), normalised mutual information (NMI), and odds ratio (OR) or log-odds ratio (LOR), use the information in the confusion matrix more effectively and allow for assessment of the extent to which models correctly predict occurrence at rates better than expected by chance (Fielding & Bell, 1997; Baldi *et al*., 2000; Manel *et al*., 2001). Two of these measures, κ and NMI, have been shown to be relatively insensitive to variation in prevalence (Manel *et al*., 2001).

In this paper we develop and test quantitative models for predicting the spatial distribution of active and resting juvenile Atlantic salmon. For the two types of behaviour, we: (i) compare LR and CT models for predicting summer distributions at the microhabitat scale, (ii) validate the models based on crossover field tests in which models developed for one reach (calibration trials) are applied to the other reach (validation trials), (iii) use multiple measures of prediction capability to assess model performance, and (iv) build prediction maps based on instream habitat features and compare observed fish positions with those predicted by LR and CT models.

Methods

Study site and sampling schedule

Field work was conducted in Big Jonathan Brook (drainage area: 98 km^2), a third-order tributary of the Grande Cascapedia River in eastern Quebec, Canada (48°27′20″N, 66°01′70″W). Two reaches were studied, one (R1) located approximately 100 m from the brook mouth, 75 m a.s.l., and the other (R2) 100 m upstream of R1. Both reaches were 75 m long and 15–20 m wide, and encompassed sequences of riffle, run and pool habitats (Fig. 1). Atlantic salmon, brook trout (*Salvelinus fontinalis* Mitchill) and slimy sculpin (*Cottus cognatus* Richardson) were present at the site.

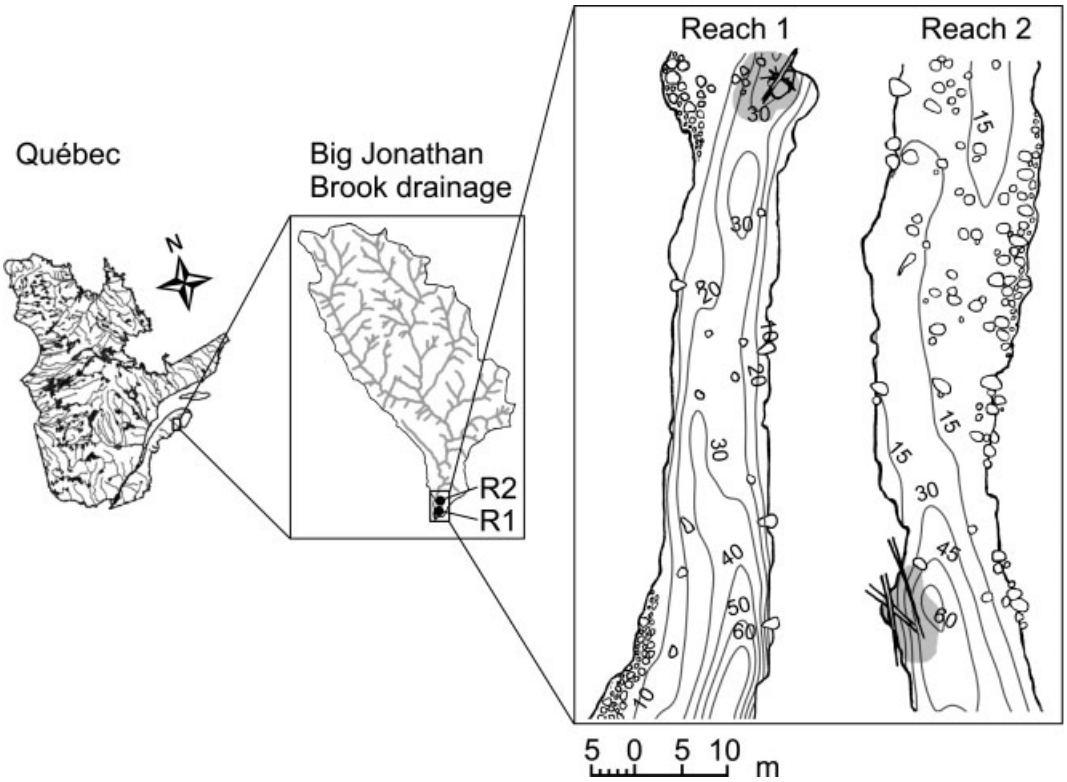

Fig. 1 Study reaches in Big Jonathan Brook, a tributary of the Grande Cascapedia River, Quebec. Both reaches are approximately 75 m long and 15–20 m wide. Contour lines within the study reaches represent water depth (cm). Woody debris and submerged rocks >30 cm are also shown.

The sampling schedule covered a 63-day period between 28 June (flow: 1.58 $m^3 s^{-1}$) and 29 August 2002 (flow: 0.55 $m^3 s^{-1}$). Water temperature (Vemco thermograph recorder placed midway between the two reaches) varied between 5.8 °C and 16.3 °C (mean ± SD: 10.1 ± 2.1 °C) over the sampling period. The two reaches were divided into adjacent sections 5 m in length, which were sampled in a fixed sequence, one section at a time and alternating between reaches, from the downstream end to the upstream end of both reaches. For each 5-m section, all sampling was carried out on two consecutive days: fish observations and microhabitat measurements were made on first day, and habitat characterisations used to build prediction maps were made on the second day.

Underwater fish observation and microhabitat measurement

Atlantic salmon parr (1+ and older; 5–16 cm total length) were observed by snorkelling, following the protocol in Heggenes *et al.* (2002). Underwater visibility exceeded 5 m during dives, which were always carried out between 11:00 and 14:00 hours. Each diving session covered one 5-m section of the reach and lasted 60–120 min depending on the number of fish encountered. To avoid startling fish, the diver entered the stream 10–15 m downstream of the target section. Within the section, the diver moved slowly upstream in a 'zigzag' pattern until a fish was encountered. The fish was then observed for 3–5 min to ensure that it was holding a position and was not disturbed by the diver. Species identity, total length (nearest centimetre), distance from bottom (nearest centimetre), and behaviour (activity or at rest) were noted for each fish. Active fish held a position in the water column and were observed foraging or engaging in agonistic interactions with other fish. Fish at rest lay on the substratum and were largely immobile. An assistant on the shore recorded data called by the diver, and the location of the fish was then marked by placing a numbered rock on the streambed.

For each reach, a subset of locations (90 locations in R1 and 106 locations in R2, to approximately match the number of locations with fish observations) were randomly selected from a uniform *xy* grid (1 × 1 m cells) covering the entire surface of the reach. A selected location was marked as an 'absence' location if it was not used by fish at the time of observation (no fish seen within a radius of 50 cm of the location over a period of at least 3 min); otherwise, it was discarded and replaced by another randomly chosen location not used by fish. A random subset was used because including all of the absence locations in the reaches would have greatly reduced prevalence, possibly leading to an artificial increase in accuracy (Fielding & Bell, 1997).

At each marked location, we recorded water depth, current velocity at 15 and 40% depth (from bottom) (pygmy-type meter; Scientific Instruments 1205, Milwaukee, WI, U.S.A.), substratum particle size (Wentworth scale; DeGraaf & Bain, 1986), presence of instream cover within a 15 cm radius of the location (unembedded rock >20 cm along the major axis, submerged vegetation or woody debris), presence of overhead cover (broken water surface, undercut bank, or overhanging vegetation), and distance to the stream bank.

Model development

All models were fit to aggregate data collected over the 63-day study period (28 June to 29 August). LR and CT models were developed separately for each study reach and behaviour in calibration trials (a total of eight models: two model types × two behaviours × two study reaches). The models aimed at predicting presence or absence of salmon at a location, either active or at rest, on the basis of habitat features at the location. An alternative approach, in which activity and rest were integrated in a single outcome variable, would also have been feasible (i.e. one polytomous instead of two binary LR, and one three-group CT instead of two two-group CT). However, models obtained by the latter approach, although more synthetic, would also be less specific and more difficult to interpret than the models with simpler outcome (dependent) variables (Hosmer & Lemeshow, 2000).

Logistic regression represents the probability of occurrence, P, as a function of a linear combination of habitat predictors, which can include single variables as well as higher-order (quadratic and interaction) terms:

$$P = \frac{e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{k} \beta_i x_i}}$$

where the x_i are single-variable or higher-order habitat predictors, β_0 is a constant, the β_i are regression coefficients associated with the k predictors, and e is the base of natural logarithms. Higher-order terms were included among the potential predictors in the variable selection procedure to allow for significant nonlinear effects in addition to linear ones.

The program SYSTAT, v. 10.2, was used to build LR models. Squared variables and all pairwise interactions between single variables were included as potential predictors. All variables were z-standardised prior to calculating products of variables, to remove non-essential collinearity in quadratic and interaction terms and facilitate comparisons among predictors. A stepwise selection procedure, with nominal cut-off at $P = 0.05$, was used to determine which variables should be retained in the final model. The tolerance (a measure of the amount of variation unique to each predictor; Tabachnick & Fidell, 2000) was >0.68 for all predictor variables in final models, indicating only mild collinearity among predictors. Because model performance can be highly sensitive to the choice of prediction threshold (Fielding & Bell, 1997; Manel *et al*., 1999; Hosmer & Lemeshow, 2000), an optimal decision threshold (ODT) was used, in addition to the normal threshold of $P = 0.5$ used in applications of LR models, to predict presence or absence. Receiver-operating characteristic plots were drawn to evaluate predictive ability over all decision thresholds in the calibration trials (Pearce & Ferrier, 2000), and the ODT was chosen to equalise the costs of misclassifying species as present (sensitivity) or absent (specificity) (Fielding & Bell, 1997). McFadden's ρ^2 was used as a measure of association between P and the predictor variables in the final models. ρ^2 tends to be much lower than R^2 for multiple regression, with values in the range 0.2–0.4 considered highly satisfactory (Tabachnick & Fidell, 2000).

The RPART3 software library (Atkinson & Therneau, 2000) was used to develop CT models (S-PLUS program, v. 6.2). RPART3 uses the binary recursive partitioning algorithm developed by Breiman *et al.*

(1984), which is the best known, most dependable and most thoroughly tested available (Lim, Loh & Shih, 2000). Beginning with the entire data set (the 'root node' at the top of the tree), the algorithm examines all possible splits for each possible value of the predictor variables, and selects the candidate split (the 'splitting value') that maximises the homogeneity within the two resulting subgroups, or nodes, with respect to the response variable. We used 'pruning' and cross-validation to select optimal trees (Atkinson & Therneau, 2000; De'ath & Fabricius, 2000). First, we generated a sequence of trees of increasing size, using a cost-complexity parameter, CP, to eliminate ('prune off') splits that were obviously not worthwhile, i.e. that did not improve the fit by at least the value of CP (=0.01 for all trials) (Atkinson & Therneau, 2000). Then, 10-fold cross-validation was used to estimate prediction error, and final tree size was determined by the 1-SE rule, which favours the largest tree for which the cross-validated error falls within 1 SE of the minimum relative error determined by cross-validation (Atkinson & Therneau, 2000; Feldesman, 2002). Given that the selected tree size will vary under repeated cross-validation, 50 sets of 10-fold cross-validation were run and the most frequently occurring tree size was chosen (De'ath & Fabricius, 2000). The influence of individual predictor variables was gauged by the proportional reduction in error (PRE, a measure of the variability accounted for by the splits associated with each predictor in the tree), an approach similar to the use of partial R^2 to assess the contribution of individual predictors in multiple regression.

Model validation and assessment

Crossover field tests were used to validate models and assess transferability. Models developed and calibrated with data from R1 were used to predict presence or absence on the basis of habitat data from R2, and *vice versa*, yielding a total of eight validation trials.

To evaluate model accuracy, the following measures were obtained from the confusion matrices: correct classification rate (CCR; percentage of all cases correctly predicted), sensitivity (percentage of true positives correctly predicted), and specificity (percentage of true negatives correctly predicted).

Four additional measures were calculated from the confusion matrices to assess whether model performance differed from expectations based on chance alone: κ (proportion of specific agreement; range: −1 to 1), MC (range: −1 to 1), NMI (range: 0–1), and the LOR (range: $-\infty$ to ∞). For all measures, a value of zero indicates no difference from random prediction. We used the following formulae:

$$\kappa = \frac{(a+d) - \frac{(a+c)(a+b)+(b+d)(c+d)}{N}}{N - \frac{(a+c)(a+b)+(b+d)(c+d)}{N}}$$

$$\mathrm{MC} = \frac{ad - cb}{\sqrt{(a+c)(a+b)(b+d)(c+d)}}$$

$$\mathrm{NMI} = 1 - \frac{-a\ln(a) - b\ln(b) - c\ln(c) - d\ln(d) + (a+b)\ln(a+b) + (c+d)\ln(c+d)}{N\ln N - (a+c)\ln(a+c) - (b+d)\ln(b+d)}$$

$$\mathrm{LOR} = \ln\left(\frac{ad}{cb}\right),$$

where a (true positives), b (false positives), c (false negatives), and d (true negatives) are the four entries in a 2×2 confusion matrix, and $N = a + b + c + d$ is the total number of cases.

Prediction maps

Prediction maps for active and resting salmon were used to represent the spatial distribution of probabilities of occurrence, predicted on the basis of habitat features in the two reaches. As with model development, prediction maps were built from data aggregated over the study period. Depth, current velocity, substratum particle size, instream and overhead cover, and distance to the stream bank were measured (as described above, *Underwater fish observation and microhabitat measurement*) at the centre of each 1×1 m cell of the *xy* grids. For each reach, stream flow was measured approximately twice per week. The LR and CT models were used to predict a binary value reflecting either presence (1) or absence (0) for each cell of the *xy* grids. For the LR models, predictions for individual cells

were made by comparing the *P*-value obtained based on the habitat features of that cell to the ODT threshold. For the CT models, predictions were made by 'dropping down' the cells along tree branches so that assignment of cells to terminal nodes was determined by the habitat features of individual cells (Feldesman, 2002). Then, these binary values were smoothed to obtain continuous values for probability of occurrence over the whole surface of the reaches (distance-weighted least-squares; program SYSTAT, v. 10.2).

Results

Microhabitat models

Average habitat conditions were broadly comparable in the two study reaches ('Absence' columns in Table 1), although the pattern of spatial heterogeneity varied between reaches, as illustrated by the differences in depth contours (Fig. 1). In both reaches, available cover was mostly in the form of embedded rocks (Table 1). LR and CT models identified water depth and the presence of an unembedded rock >20 cm as key predictor variables (Table 2; Fig. 2). Both types of model indicated that active salmon selected positions based mostly on water depth and avoided shallow sites, whereas salmon at rest selected positions behind or alongside an unembedded rock >20 cm. However, the LR models always retained additional predictors beyond those retained by the CT models.

Logistic regression models differed between reaches for a given behaviour, both in the number and identity of predictor variables (Table 2). The final model for active salmon included four single variables and one quadratic term in R1, but only one variable and one quadratic term in R2. The final model for resting salmon included five variables and one interaction term in R1, but only three variables and one quadratic term in R2. When simpler linear models excluding the significant quadratic or interaction terms were considered, McFadden's ρ^2 showed declines ranging from 4.7% (salmon at rest in R2) to 38.1% (active salmon in R2).

Only one or two variables were useful predictors in the final CT models (Fig. 2). Final models for active salmon were almost identical in the two reaches, including the same single predictor (water depth) and very similar splitting values, indicating that active salmon were most likely to be present where water depth exceeded 26.5–31.5 cm. For resting salmon, final models differed slightly between reaches, but in both reaches the most influential variable (as reflected by PRE values) was the presence of an unembedded rock >20 cm. Water depth also contributed secondarily to prediction in R2: the second split in the CT model indicates that even in the presence of rocky cover, salmon were unlikely to be present if water depth was below 19.5 cm.

Model accuracy (CCR) was similar across model types and was slightly higher for resting than for active salmon (Fig. 3). In calibration trials, CCR

Table 1 Fish length and habitat variables (mean ± SD) at locations used by active and resting juvenile Atlantic salmon, and unused locations, by reach

	Reach 1			Reach 2		
Fish length and habitat variables	Active ($N = 92$)	At rest ($N = 37$)	Absence ($N = 90$)	Active ($N = 46$)	At rest ($N = 25$)	Absence ($N = 106$)
Fish length (cm)	10.4 ± 1.7	10.3 ± 1.3	–	9.7 ± 1.7	11.9 ± 1.6	–
Water depth (cm)	41.9 ± 14.9	28.5 ± 12.5	21.7 ± 8.8	34.3 ± 9.2	27.0 ± 12.8	18.8 ± 12.4
Current velocity at 15% depth from bottom (cm s^{-1})	36.7 ± 12.2	32.1 ± 20.5	32.1 ± 20.1	51.1 ± 19.5	41.8 ± 25.7	40.3 ± 27.3
Current velocity at 40% depth from bottom (cm s^{-1})	43.3 ± 14.6	40.2 ± 22.4	35.7 ± 21.8	62.4 ± 25.4	46.9 ± 23.6	43.8 ± 30.3
Substratum particle size (Wentworth scale)	8.5 ± 1.4	7.9 ± 2.4	8.2 ± 1.6	8.7 ± 2.0	9.8 ± 2.3	8.5 ± 1.6
Presence of cover (number of locations)						
Rocky cover	15	28	13	5	20	11
Broken water surface	0	0	1	0	0	2
Overhanging bank	0	1	0	0	0	0
Overhanging vegetation	0	0	1	0	0	0
Submerged vegetation or wood	0	0	0	0	0	0
Distance to stream bank (m)	3.7 ± 1.2	3.3 ± 1.3	2.6 ± 1.5	3.1 ± 1.6	3.9 ± 1.8	3.8 ± 1.8

Table 2 Coefficients of logistic regression models for activity and resting behaviour, by reach. Coefficients are given only for terms retained by the stepwise selection procedure ($P < 0.05$).

	Active		At rest	
Model term	Reach 1 ($N = 182$)	Reach 2 ($N = 152$)	Reach 1 ($N = 127$)	Reach 2 ($N = 131$)
Constant	0.807	−0.673	−1.656	−1.701
Depth	3.544	3.959	0.638	–
Velocity at 40%	−0.075	–	–	1.155
Distance to bank	0.854	–	0.673	–
Substratum particle size	–	–	0.622	1.008
Rock > 20 cm	0.610	–	1.816	1.721
Depth^2	–	−1.756	–	–
(Velocity at 40%)^2	−0.497	–	–	−0.862
Substratum × Depth	–	–	−0.722	–
McFadden's ρ^2	0.52	0.45	0.47	0.50

All models were globally significant at $P < 0.0001$. McFadden's ρ^2 is reported for each model also.

generally exceeded 80% and variation between reaches was low. In validation trials, CCR declined but remained high (>70%) and variation in CCR between reaches remained low. LR and CT models generally had higher specificity than sensitivity in calibration and in validation trials (Fig. 3). Variation in specificity and sensitivity in validation trials was generally higher than in calibration trials.

The four measures of model performance that account for chance variation were strongly correlated (Pearson correlation: mean: 0.95, range: 0.89–1.00 for LR; mean: 0.95, range: 0.90–1.00 for CT); therefore, graphical results are presented only for κ and the LOR (Fig. 4). Model performance was better than random for all cases (none of the 95% confidence intervals includes zero). Performance generally declined be-

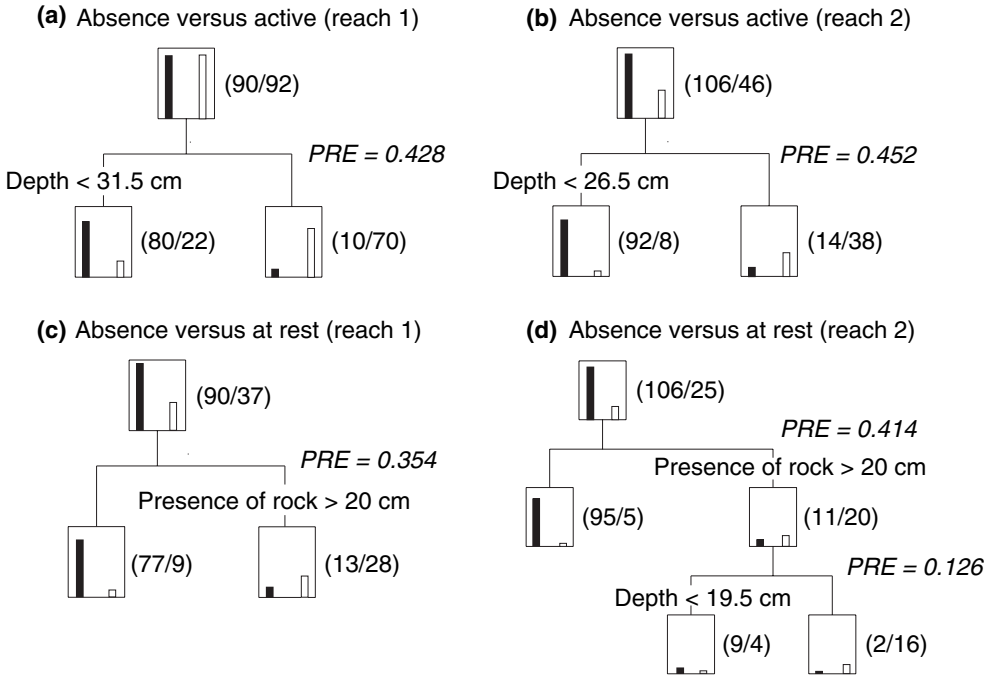

Fig. 2 Classification tree models for predicting activity versus absence (a: reach 1; b: reach 2) and resting versus absence (c: reach 1; d: reach 2). Vertical bars represent the frequency of absence (black) and presence (white) at each node. Splitting values and proportional reduction in error (PRE) values are given on the branches of the trees. Absence/presence numbers for each node are given in parentheses.

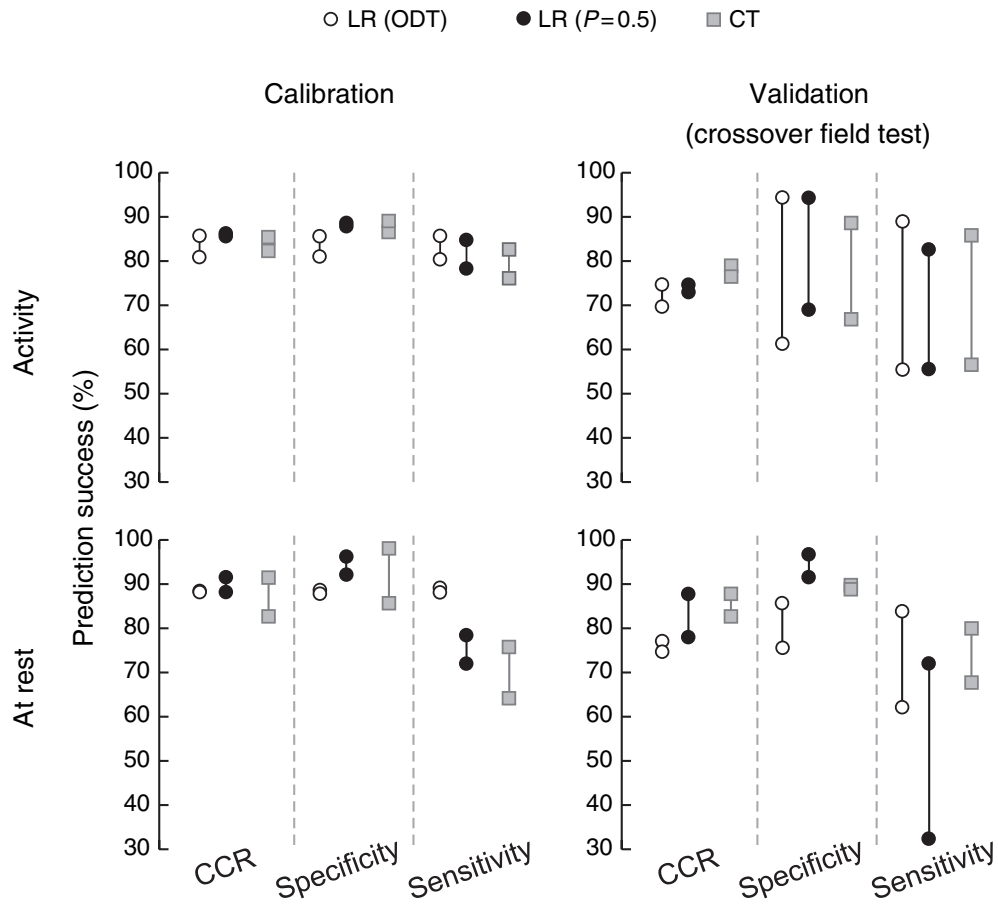

Fig. 3 Correct classification rate (CCR), specificity, and sensitivity of logistic regression (LR) and classification tree (CT) models for activity and resting behaviour, in calibration and validation (crossover field test) trials. LR results are presented both for the optimal decision (ODT) and $P = 0.5$ thresholds. The ODT, determined in calibration trials for the two reaches and also applied in validation trials, were: activity, R1: 0.47, R2: 0.40; resting, R1: 0.23, R2: 0.15. Symbols representing the values for reaches R1 and R2 are connected by a vertical line.

tween calibration and validation trials, for all behaviours and model types. Performance (absolute values and pattern of decline between calibration and validation trials) was similar for LR and CT models for active salmon. However, performance of CT models declined less than that of LR models in validation trials for salmon at rest.

Prediction maps

The prediction maps clearly show that the overall spatial distribution for active salmon (Fig. 5a–h) differed markedly from that of salmon at rest (Fig. 5i–p). In addition, the probability of presence was more spatially heterogeneous for active than for resting salmon. LR and CT often yielded similar prediction maps which closely matched the observed distributions (e.g. Fig. 5a,b), but differences between model predictions arose in several trials; e.g. LR underestimated the probability of presence of active salmon in R1 (blue area to the left of Fig. 5e), in contrast with CT, which provided accurate predictions (Fig. 5f). Overall, however, for both types of model the prediction maps for calibration and validation trials reflected well the actual distributions of fish in activity and at rest.

Discussion

Microhabitat selection in active and resting fish

Final models for active and resting fish incorporated substantially different predictor variables and yielded different prediction maps, suggesting that models

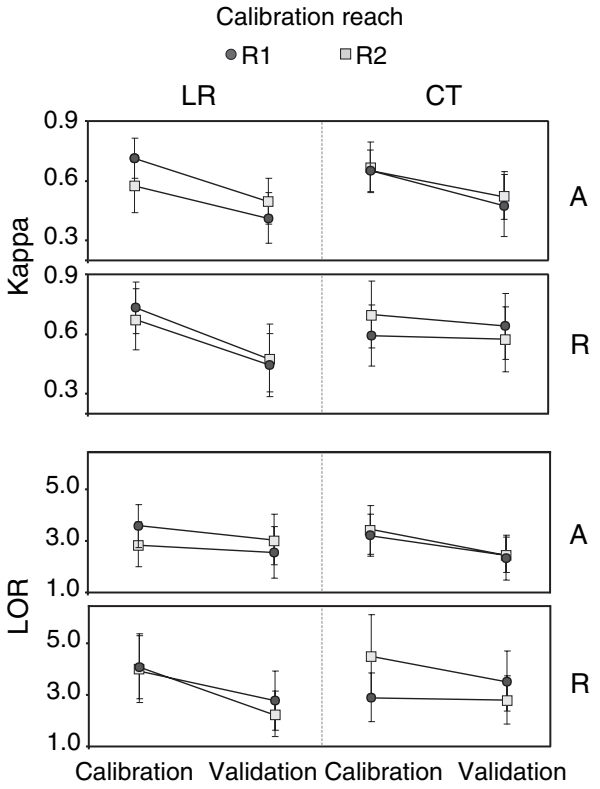

Fig. 4 Performance measures for logistic regression (LR) and classification tree (CT) models in calibration and validation (crossover field tests) trials, for activity (A) and resting (R) behaviours. Results for LR are based on the optimal decision threshold. Reported values are Cohen's kappa (κ) and LOR, with 95% confidence intervals.

based solely on active behaviour such as foraging may yield an incomplete picture of microhabitat selection in juvenile Atlantic salmon. For example, resting salmon can be abundant in areas predicted to have low probability of occurrence by a model focusing on activity (cf. Fig. 5a–p). Because of the close association of fish at rest with rocky cover, it seems likely that these fish were sheltering. Competitive (Kalleberg, 1958; Fausch & White, 1981) and predatory (Metcalfe *et al.*, 1987; Gotceitas & Godin, 1993) interactions may drive fish to seek refuge, thereby increasing the frequency of sheltering behaviour (Gries & Juanes, 1998). The availability of shelter may affect salmon populations because individuals that fail to find shelter may either be forced to emigrate or, more likely, are eaten by predators (Metcalfe *et al.*, 1987; Gotceitas & Godin, 1993).

Juvenile Atlantic salmon can adapt rapidly to a changing environment; their habitat selection behaviour is flexible and stream-specific, which may therefore limit model transferability (Heggenes *et al.*, 2002). However, LR and CT models based on data collected over a summer period (28 June to 29 August) during which flow ranged between 0.55 and 1.58 $m^3 s^{-1}$ provided accurate predictions of habitat selection by both active and resting fish. Fitting habitat models to data collected over an extended time period expands the range of environmental and behavioural variation that must be accounted for by the models. Possible losses in precision arising from increased temporal variability in longer studies must therefore be weighed against potential gains in transferability, relative to models developed from shorter 'snapshot' studies.

Comparison of models

The results suggest that LR and CT are suitable but not equivalent tools for modelling distribution of juvenile Atlantic salmon. For both types of model, accuracy (predictive power) was high. Values of performance measures were high in calibration trials and declined in validation trials for both methods (Fig. 4). In both calibration and validation trials, the performance of LR and CT was broadly similar (Figs 3 & 4) with the exception of the LR ($P = 0.5$) model for resting salmon, which had lower sensitivity than the CT model in the validation trial (Fig. 3). However, performance of CT in validation trials was less variable between reaches and generally declined less (particularly for salmon at rest) than that of LR (Fig. 4). LR models based on the ODT had higher sensitivity and less variable performance between reaches than those based on the $P = 0.5$ threshold. Use of the ODT may thus reduce costs associated with misclassification of true presence, i.e. those incurred when the model incorrectly classifies as poor (absence) a location at which a fish is actually present.

The decline in performance between calibration and validation trials illustrates the value of crossover field tests in model assessment. Because it is less subject to statistical or ecological quirks peculiar to a specific study site, external validation provides a more rigorous and realistic test of model generality than do internal validation or, clearly, no validation at all. Following rigorous validation by means of crossover field tests and assessment of model performance by

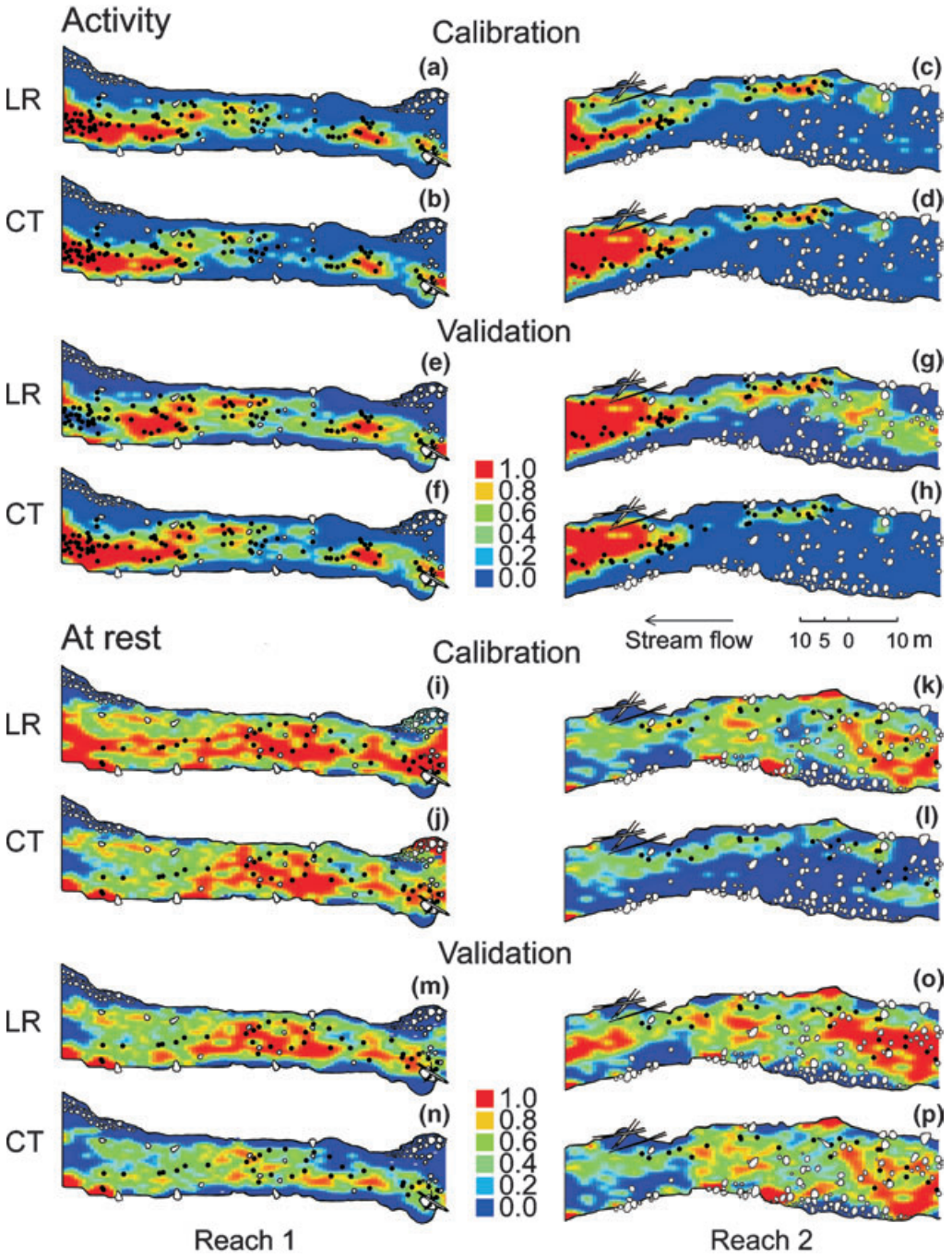

Fig. 5 (a–p): Prediction maps based on output of logistic regression (LR) and classification tree (CT) models for active and resting fish in validation and calibration trials, by reach. Probabilities of occurrence predicted as a function of habitat features are coded as colour hues (six intervals). Active and resting fish are represented by black dots. Woody debris and submerged rocks >30 cm are also shown.

use of chance-corrected measures, we found that both LR and CT had high prediction accuracy in calibration, and generality, as indicated by their accuracy in crossover validation. CT models had less variable performance and smaller decline in performance (for salmon at rest) in validation trials, suggesting that they may be more transferable than LR models. In evaluating potential model transferability, however, it must be noted that the crossover field validation in the present study involved relatively minor changes

between reaches within a single stream. Clearly, more stringent tests comparing the transferability of LR and CT models across rivers (Mäki-Petäys *et al.*, 2002) would be desirable.

Logistic regression and CT models differed in ease of use and interpretation. Building LR models required verification of statistical assumptions, transformation of variables, tolerance checks, and determination of ODTs (although not all these steps will be needed in all instances). In contrast, CT did not require transformation or standardisation because they use the rank-order of a variable to determine a split (De'ath & Fabricius, 2000). LR models were generally more difficult to interpret than CT models because the former retained more predictors, including quadratic and interaction terms, and the predictors retained for a given behaviour differed substantially between reaches (Table 2). In comparison, CT models generated simpler graphical interpretations (Fig. 2), were more consistent in predictor selection between reaches (Fig. 2), and were more parsimonious, requiring only one or two variables to generate predictions comparable in accuracy to those of more complex LR models.

This study highlights the value of examining passive behaviour in habitat selection models, as a means for refining the description of the spatial distribution of juvenile Atlantic salmon and identifying habitat needs in relation to this behaviour. Specifically, the spatial distribution of active fish differed markedly from that of resting fish, apparently as a result of differential association with habitat features, primarily depths greater than about 30 cm for fish in activity and rocky cover for fish at rest. Few previous studies have found summer sheltering in Atlantic salmon parr (e.g. Gries & Juanes, 1998), and none seems to have examined the association between summer spatial distribution and activity and sheltering behaviour. Remarkably, relatively simple LR and CT models for the distribution of fish in small stream reaches sufficed to generate accurate prediction maps (Fig. 5), which have traditionally been developed at a larger spatial scale, by use of more complex quantitative procedures (cf. Guay *et al.*, 2000; Guensch, Hardy & Addley, 2001). Our results indicate that, in addition to identifying areas of favourable habitat, prediction maps can clearly depict associations between reach-wide horizontal patterns of spatial distribution and specific behaviour.

Acknowledgments

We thank M.-N. Rivard and M. Chénier-Soulière for assistance in the field, J. Deschênes, P. Peres-Neto and A.G. Hildrew for constructive comments, and the Société Cascapédia for logistic and financial support. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and le Fond québécois de la Recherche sur la Nature et les Technologies (FQRNT). This paper is a contribution to the programme of the Centre Interuniversitaire de Recherche sur le Saumon Atlantique (CIRSA).

References

Atkinson E.J. & Therneau T.M. (2000) *An Introduction to Recursive Partitioning Using the RPART Routines*. Technical Report number 61. Mayo Foundation, Rochester, 52 pp.

Baldi P., Brunak S., Chauvin Y., Andersen C.A.F. & Nielsen H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Breiman L., Friedman J.H., Olshen R.A. & Stone C.J. (1984) *Classification and Regression Trees*. Chapman and Hall, New York.

Cunjak R.A. (1988) Behaviour and microhabitat of young Atlantic salmon (*Salmo salar*) during winter. *Canadian Journal of Fisheries and Aquatic Sciences*, **45**, 2156–2160.

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

DeGraaf D.A. & Bain L.H. (1986) Habitat use by and preferences of juvenile Atlantic salmon in two Newfoundland rivers. *Transactions of the American Fisheries Society*, **115**, 671–681.

Fausch K.D. (1984) Profitable stream positions for salmonids: relating specific growth rate to net energy gain. *Canadian Journal of Fisheries and Aquatic Sciences*, **62**, 441–451.

Fausch K.D. & White R.J. (1981) Competition between brook trout (*Salvelinus fontinalis*) and brown trout (*Salmo trutta*) for positions in a Michigan stream. *Canadian Journal of Fisheries and Aquatic Sciences*, **38**, 1220–1227.

Feldesman M.R. (2002) Classification trees as an alternative to linear discriminant analysis. *American Journal of Physical Anthropology*, **119**, 257–275.

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**, 39–49.

Giannico G.R. & Healey M.C. (1999) Ideal free distribution theory as a tool to examine juvenile coho salmon (*Oncorhynchus kisutch*) habitat choice under different conditions of food abundance and cover. *Canadian Journal of Fisheries and Aquatic Sciences*, **56**, 2362–2373.

Gotceitas V. & Godin J.G.J. (1993) Effects of aerial and instream threat of predation on foraging by juvenile Atlantic salmon (*Salmo salar*), in natural waters. In: *Production of Juvenile Atlantic Salmon,* Salmo salar, in Natural Waters (Eds R.J. Gibson & R.E. Cutting), pp. 35–41. *Canadian Special Publication on Fisheries and Aquatic Sciences*, 118.

Grand T.C. & Dill L.M. (1997) The energetic equivalence of cover to juvenile coho salmon (*Oncorhynchus kisutch*): ideal free distribution theory applied. *Behavioural Ecology*, **8**, 437–447.

Gregory J.S. & Griffith J.S. (1996) Winter concealment by subyearling rainbow trout: space size selection and reduced concealment under surface ice and in turbid water conditions. *Canadian Journal of Zoology*, **49**, 237–245.

Gries G. & Juanes F. (1998) Microhabitat use by juvenile Atlantic salmon (*Salmo salar*) sheltering during the day in summer. *Canadian Journal of Zoology*, **76**, 1441–1449.

Guay J.C., Boisclair D., Rioux D., Leclerc M., Lapointe M. & Legendre P. (2000) Development and validation of numerical habitat models for juveniles of Atlantic salmon (*Salmo salar*). *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 2065–2075.

Guensch G.R., Hardy T.B. & Addley R.C. (2001) Examining feeding strategies and position choice of drift-feeding salmonids using an individual-based, mechanistic foraging model. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 446–457.

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

Heggenes J., Saltveit S.J., Bird D. & Grew R. (2002) Static habitat partitioning and dynamic selection by sympatric young Atlantic salmon and brown trout in southwest England streams. *Journal of Fish Biology*, **60**, 72–86.

Hill J. & Grossman G.D. (1993) An energetic model of microhabitat use for rainbow trout and rosyside dace. *Ecology*, **74**, 685–698.

Hosmer D.W. & Lemeshow S. (2000) *Applied Logistic Regression*, 2nd edn. Wiley-Interscience, New York.

Hughes N.F. & Dill L.M. (1990) Position choice by drift-feeding salmonids: models and test for arctic grayling (*Thymallus arcticus*) in subarctic mountain streams, interior Alaska. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 2039–2048.

Kalleberg H. (1958) *Observations in a Stream Tank of Territoriality and Competition in Juvenile Salmon and Trout (Salmo salar L. and S. trutta)*. Institute of Freshwater Research, Drottningholm, Report, 39, 55–98.

Knapp R.A. & Preisler H.K. (1999) Is it possible to predict habitat use by spawning salmonids? A test using California golden trout (*Oncorhynchus mykiss aguabonita*). *Canadian Journal of Fisheries and Aquatic Sciences*, **56**, 1576–1584.

Lek S., Delacoste M., Bran P., Dimopoulos I., Lauga J. & Aulagnier S. (1996) Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, **1634**, 1–13.

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

Mäki-Petäys A., Huusko A., Erkinaro J. & Muotka T. (2002) Transferability of habitat suitability criteria of juvenile Atlantic salmon (*Salmo salar*). *Canadian Journal of Fisheries and Aquatic Sciences*, **59**, 218–228.

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

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

Metcalfe N.B., Huntingford F.A. & Thorpe J.E. (1987) The influence of predation risk on the feeding motivation and foraging strategy of juvenile Atlantic salmon. *Animal Behaviour*, **35**, 901–911.

Olden J.D. & Jackson D.A. (2002) A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology*, **47**, 1976–1995.

Olden J.D., Jackson D.A. & Peres-Neto P.R. (2002) Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329–336.

Pearce J. & Ferrier S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.

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

Rieman B.E. & McIntyre J.D. (1995) Occurrence of bull trout in naturally fragmented habitat patches of varied size. *Transactions of the American Fisheries Society*, **124**, 285–296.

Shirvell C.S. (1990) Role of instream rootwads as juvenile coho salmon (*Oncorhynchus kisutch*) and steelhead trout (*O. mykiss*) cover habitat under varying streamflows. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 852–861.

Stoneman C.L. & Jones M.L. (2000) The influence of habitat features on the biomass and distribution of three species of southern Ontario stream salmonines. *Transactions of the American Fisheries Society*, **129**, 639–657.

Tabachnick B. & Fidell L. (2000) *Using Multivariate Statistics*, 4th edn. Pearson Allyn & Bacon, New York.

Torgersen C.E., Price D.M., Li H.W. & McIntosh B.A. (1999) Multiscale thermal refugia and stream habitat associations of chinook salmon in northeastern Oregon. *Ecological Applications*, **9**, 301–319.

Verbyla D.L. & Litaitis J.A. (1989) Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management*, **13**, 783–787.

(Manuscript accepted 5 January 2005)