UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE DE LA MAÎTRISE
EN SCIENCES DE L'ENVIRONNEMENT

PAR
KATRINE TURGEON

PRÉDICTION DE LA SÉLECTION DU MICROHABITAT CHEZ LES JUVÉNILES DU SAUMON DE L'ATLANTIQUE (*SALMO SALAR*) À L'AIDE DE LA RÉGRESSION LOGISTIQUE ET DES ARBRES DE CLASSIFICATION

AOÛT 2004

Résumé

Nous avons comparé la capacité de la régression logistique (RL) et des arbres de classification (AC) à prédire l'utilisation du microhabitat et la distribution estivale des juvéniles du saumon de l'Atlantique, *Salmo salar*, dans deux tronçons d'un petit ruisseau de l'est du Québec. Les modèles développés prédisent la présence vs. l'absence des saumons à un site en fonction des caractéristiques de l'habitat (profondeur, vitesse de courant, présence de couvert submergé et du couvert émergent, taille du substrat et distance à la berge) mesurées au site. Les modèles ont été développés séparément en fonction du comportement adopté par les individus (actif ou passif). Les modèles ont été validés à l'aide de tests croisés sur le terrain (angl. : «*crossover field tests*») qui ont permis d'évaluer la performance des modèles développés sur un tronçon (essais en calibration) lorsqu'ils étaient appliqués sur l'autre tronçon (essais en validation). La performance des modèles a été évaluée en fonction de leur efficacité, de leur généralité et de leur facilité d'utilisation et d'interprétation. Des cartes prévisionnelles, construites à partir des caractéristiques environnementales, ont servi à comparer les positions des poissons prédites par les modèles RL et AC avec les positions observées des poissons. Il semblerait qu'une sélection de l'habitat différentielle en fonction du comportement menait à des distributions spatiales différentes pour les poissons en activité et ceux au repos. Les saumons en activité sélectionnaient des positions en fonction des profondeurs supérieures à 30 cm, tandis que les saumons au repos étaient associés à la présence d'un couvert rocheux. Tous les modèles avaient un fort pouvoir prévisionnel et étaient transférables lors des tests de validation croisée sur le terrain. Pour les deux types de modèles, les cartes prévisionnelles représentaient de façon précise la distribution spatiale des poissons. Cependant, les modèles AC étaient plus faciles à bâtir et à interpréter que les modèles RL. Les modèles AC avaient également une performance moins variable et un déclin de performance moins important en validation croisée (notamment pour les saumons au repos), suggérant ainsi qu'ils pourraient être plus transférables que les modèles RL.

Remerciements

Premièrement, je tiens à remercier mon directeur de maîtrise, Dr Marco A. Rodríguez, pour m'avoir donné l'opportunité de réaliser ce projet et pour m'avoir transmis un intérêt marqué pour la biologie quantitative. J'aimerais également remercier les membres de mon comité de lecture, Drs Pierre Magnan et Gilbert Cabana pour les critiques et commentaires constructifs durant l'évaluation de ce mémoire, ainsi que pour les deux années passées en leur compagnie.

J'aimerais également remercier : Marie-Noëlle Rivard, Myriam Chénier-Soulière, Nicolas Martel, Sébastien Rouleau, Geneviève Turgeon et Jean-François Therrien pour leur assistance très appréciée sur le terrain et pour leur amitié. Remerciements sincères à Julie Deschênes et Pedro Peres-Neto pour les suggestions durant la rédaction des articles, ainsi qu'aux membres du GRÉA (Groupe de recherche sur les écosystèmes aquatiques). Merci également à mes parents, Réjean et Linda, qui me soutiennent depuis toujours dans mon cheminement académique et personnel.

J'aimerais également exprimer ma reconnaissance à la Société Cascapédia (Marc Gauthier et Marc-André Bernard) pour le support financier et logistique. Cette étude était financée par le Fond québécois de la recherche sur la nature et les technologies (FQRNT), le Conseil de Recherches en sciences naturelles et en génie du Canada (CRSNG) et le Centre interuniversitaire de recherche sur le saumon Atlantique (CIRSA).

Avant-propos

Ce mémoire comprend deux chapitres. Le premier chapitre est une synthèse en français du projet de maîtrise. Le second chapitre est l'article soumis pour publication dans le périodique *Freshwater Biology*. Cet article compare la capacité de deux approches quantitatives, la régression logistique et les arbres de classification, à prédire l'utilisation du microhabitat et la distribution estivale des juvéniles du saumon de l'Atlantique, *Salmo salar*.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1. Prédiction de la sélection du microhabitat chez les juvéniles du saumon de l'Atlantique, *Salmo salar*, à l'aide de la régression logistique et des arbres de classification

Introduction

La sélection de l'habitat chez les poissons salmonidés reflète différents compromis entre les gains énergétiques nets et l'évitement des risques (ex. : prédation, séquestration et blessures causés par la glace). Les études précédentes ont démontré que la sélection de l'habitat chez les poissons est influencée par les gains énergétiques provenant de la dérive d'invertébrés (Fausch 1984; Hughes & Dill, 1990; Hill & Grossman, 1993), par les coûts énergétiques reliés à la nage, qui varient en fonction de la vitesse du courant (Fausch 1984), par les risques de prédation (Metcalfe, Huntingford & Thorpe, 1987; Gotceitas & Godin, 1993; Gregory & Griffith, 1996), par les interactions agonistiques (Kalleberg 1958; Fausch & White 1981), et par la disponibilité de couvert submergé (Cunjak 1988; Gries & Juanes, 1998) et émergent (Shirvell, 1990; Grand & Dill, 1997).

Généralement, les modèles de sélection du microhabitat ne font pas la distinction entre les comportements actifs, tels l'alimentation, et les comportements plus passifs, tels le repos sur le substrat et l'utilisation d'abris ou du couvert. Or, cette distinction peut s'avérer instructive puisque l'utilisation du couvert varie dans le temps et dans l'espace en fonction des risques (prédation et compétition), des fluctuations hydrologiques (Gotceitas & Godin, 1993; Giannico & Healey, 1999) ou du développement ontogenétique (Cunjak 1988; Gries & Juanes, 1998). De plus, chez les juvéniles du saumon de l'Atlantique en période estivale, l'utilisation du couvert serait peut-être un phénomène plus répandu que ce qu'on suggère dans la littérature

et pourrait avoir un effet positif sur la production de saumons (Gries & Juanes, 1998). Faire la distinction entre les comportements actifs et passifs dans les modèles de sélection du microhabitat pourrait améliorer notre compréhension des besoins des salmonidés en rivière et nous permettre également d'obtenir des prédictions plus exactes de leur distribution spatiale.

Objectifs

Cette étude vise à : (1) comparer la capacité de la régression logistique (RL) et des arbres de classification (AC) à prédire l'utilisation du microhabitat des juvéniles du saumon de l'Atlantique, *Salmo salar*, dans deux tronçons d'un petit ruisseau de l'est du Québec; (2) valider les modèles produits par des tests de validation croisée sur le terrain (angl. « *crossover field tests* »), où les modèles développés sur un tronçon (essais en calibration) sont appliqués sur l'autre tronçon (essais en validation); (3) évaluer les modèles à l'aide de plusieurs mesures de performance et (4) bâtir des cartes prévisionnelles, à partir des caractéristiques environnementales, pour vérifier la concordance entre les positions prédites par les modèles avec les positions observées des poissons afin de prédire la distribution estivale des juvéniles du saumon de l'Atlantique.

Méthodes

Sites à l'étude

L'échantillonnage s'est déroulé du 28 juin au 29 août 2002 dans le ruisseau Big Jonathan, un tributaire de la rivière Grande Cascapédia, en Gaspésie. Deux tronçons ont été échantillonnés, le tronçon 1 (T1) situé à 100 m de l'embouchure du ruisseau et le tronçon 2 (T2) situé environ 100 m en amont de T1. Les deux tronçons mesuraient 75 m de longueur par 15 à 20 m de largeur (Fig. 2.1).

Localisation des poissons et échantillonnage du microhabitat

Un total de 28 plongées ont été effectuées durant la période d'échantillonnage, entre 11h00 et 14h00, couvrant des sections de 5 m de long. Les deux tronçons étaient échantillonnés en parallèle dans le temps, une section à la fois et alternant entre les tronçons.

Lorsque localisé, le poisson était observé durant 3-5 min. Le poisson était identifié à l'espèce, sa longueur et sa position dans la colonne d'eau étaient estimées (± 1 cm), son comportement était noté et sa position était marquée à l'aide d'une roche numérotée. Pour chaque tronçon, des positions non occupées (absences) par les individus ont été choisies aléatoirement dans la grille XY (90 positions dans T1; 106 positions dans T2). À chaque site marqué, nous avons mesuré la profondeur (cm), la vitesse de courant à 15% et 40% de la colonne d'eau a partir du fond ($cm \cdot s^{-1}$), la taille moyenne des particules du substrat (Tableau 2.1), la présence du couvert submergé et du couvert émergent et la distance à la berge (m).

Développement des modèles et analyses statistiques

Développement des modèles, validation et évaluation

Les modèles RL représentent la probabilité d'occurrence des saumons en fonction d'une combinaison linéaire des variables environnementales (variables simples et termes d'interaction). Une procédure de sélection pas à pas progressive (avec un seuil de $p = 0.05$) a été utilisée pour sélectionner les variables retenues dans les modèles finaux (logiciel SYSTAT, v.10.2). Un graphique de la fonction d'efficacité de l'observateur (angl. :« *Receiver-operating characteristic* ») a été produit pour évaluer la capacité prédictive des modèles RL. Ce graphique permet de représenter la sensibilité du modèle (pourcentage des présences correctement prédites) sur l'axe des X et la spécificité (pourcentage des absences correctement

prédites) sur l'axe des Y en fonction de l'étendue des seuils de décision possibles dans les essais en calibration (Pearce & Ferrier, 2000). Le seuil optimal de décision (SDO) a été choisi pour égaliser les coûts associés au classement erroné des présences (sensibilité) et des absences (spécificité) (Fielding & Bell, 1997).

Le module RPART3 (Atkinson & Therneau 2000) a été utilisé pour développer les modèles AC (logiciel S-PLUS, v.6.1). Les arbres ont été élagués par une validation croisée à dix partitions (estimation de l'erreur associée à la prédiction par la règle du 1-SE) (Atkinson and Therneau 2000; De'ath and Fabricius 2000; Feldesman 2002).

Des tests de validation croisée sur le terrain «angl. : *crossover field tests* » ont été utilisés pour valider les modèles, c'est-à-dire que les modèles développés et calibrés avec les données de T1 ont été utilisés pour prédire la présence et l'absence de saumons en fonction des caractéristiques environnementales de T2, et vice versa. Pour évaluer l'exactitude des modèles, les mesures suivantes ont été obtenues à partir de la matrice de confusion (tableau de contingence confrontant les classes prédites et les classes observées) : le PCP (pourcentage des présences et des absences observées qui sont correctement prédites), la sensibilité (pourcentage des présences observées qui sont correctement prédites) et la spécificité (pourcentage des absences observées qui sont correctement prédites). Quatre mesures additionnelles ont permis d'estimer si la performance du modèle différait de ce qui est attendu par hasard seulement (Fielding & Bell 1997, Manel et al. 2001; Baldi et al. 2000). Ces mesures sont : le kappa de Cohen (κ), la correlation de Matthews (CM), l'information mutuelle normalisée (IMN) et le logarithme du rapport des cotes «angl. : *log odds-ratio* » (LOR).

Caractérisation de l'habitat et cartes prévisionnelles

Les cartes prévisionnelles bâties en appliquant les modèles RL et AC dans chaque tronçon représentent la distribution spatiale des probabilités d'occurrence des saumons, en activité ou au repos, en fonction des caractéristiques environnementales des tronçons. La profondeur, les vitesses de courant, la taille du substrat, le couvert submergé et émergent et la distance à la berge ont été échantillonnées (tel que décrit dans la section *Localisation des poissons et échantillonnage du microhabitat*) au centre de chaque cellule de 1 x 1 m de la grille XY. Les modèles RL (avec l'utilisation du SOD) et AC ont été utilisés pour prédire la présence (1) ou l'absence (0) de saumons dans chaque cellule de la grille XY en fonction du comportement adopté par les poissons (activité ou repos). Les valeurs binaires produites par les modèles ont été lissés pour obtenir des valeurs de probabilités continues sur la surface des deux tronçons (moindres carrés pondérés en fonction de la distance, « angl : *distance-weighted least-squares* »; logiciel SYSTAT, v. 10.2). Les probabilités d'occurrence sont représentées sous forme d'intervalles associés avec des couleurs différentes sur les cartes prévisionnelles. La position réelle des poissons est également superposée sur les cartes pour évaluer la concordance entre les prédictions des modèles et la distribution observée des saumons.

Résultats

Les caractéristiques environnementales (moyenne ± É-.T.) des deux tronçons d'étude étaient assez semblables (Tableau 2.1); cependant, les patrons d'hétérogénéité spatiale, illustrés par les isolignes bathymétriques, différaient entre les tronçons (Fig. 2.1). En effet, la bathymétrie du T1 était plus variable que celle du T2. La température de l'eau a variée entre 5.8 °C et 16.3 °C (moyenne ± SD : 10.1 ± 2.1 °C) durant la période d'échantillonnage.

Modèles de sélection de l'habitat

Les mêmes variables prévisionnelles dominantes ont été retenues par les modèles RL et AC pour les deux comportements, soit l'activité et le repos (Tableau 2.2 et Fig. 2.2). Globalement, les saumons en activité sélectionnaient une position en fonction de la profondeur de l'eau, tandis que les saumons au repos sélectionnent une position sous ou derrière une roche non-imbriquée > 20 cm de diamètre. Les modèles RL retenaient toujours plus de termes (variables simples et interactions) que les modèles AC. Les modèles AC finaux pour les saumons en activité étaient presque identiques dans les deux tronçons. Pour les saumons au repos, les modèles AC finaux différaient légèrement entre les tronçons, mais la variable prédominante était la présence d'une roche non-imbriquée > 20 cm.

L'exactitude des modèles (PCP) pour les essais en calibration variait légèrement entre les modèles RL et AC (Fig. 2.3). Pour les essais en validation, le PCP demeurait élevé pour les deux types de modèles. Les modèles RL et AC avaient généralement des valeurs de spécificité supérieures à celles de la sensibilité autant en calibration qu'en validation et la variabilité de ces deux mesures était plus importante dans les essais en validation. Les quatre mesures qui tiennent compte des fluctuations dues au hasard étaient fortement corrélées et démontraient que la performance des modèles diffère de ce qui est espéré par chance seulement (aucun modèle inclus le zéro à l'intérieur de son intervalle de confiance) (Fig. 2.4). Également, la performance des modèles déclinait généralement entre les essais en calibration et les essais en validation pour les deux comportements et les deux types de modèles. La performance (valeurs absolues et les patrons de déclin entre les essais de calibration et validation) des modèles RL et AC était semblable pour les saumons en activité. Cependant, la performance des modèles AC déclinait moins que celle des modèles RL pour les essais en calibration chez les saumons au repos.

Les cartes prévisionnelles

Les cartes prévisionnelles illustrent clairement que la distribution spatiale des saumons en activité (Fig. 2.5 a-h) diffère considérablement de celle des saumons au repos (Fig. 2.5 i-p). Pour les saumons en activité, les probabilités de présence prédites étaient spatialement hétérogènes (Fig. 2.5). Dans T1, les modèles RL et AC donnaient des cartes prévisionnelles très semblables pour les essais en calibration (Fig. 2.5 a,b). Dans T2, les modèles RL et AC donnaient également des cartes de prédiction très semblables; cependant, le modèle AC tendait à surestimer la probabilité de présence près d'une aire où les saumons étaient localement abondants comparé au modèle RL (aire rouge à gauche, Fig. 2.5 d). Dans les essais en validation, les modèles RL sous-estimait la probabilité de présence des saumons en activité dans T1 (aire bleu à gauche, Fig. 2.5 e), ce qui contraste avec les modèles AC, qui donnaient une carte prévisionnelle plus exacte (Fig. 2.5 f). Les modèles RL et les modèles AC surestiment la probabilité de présence (aire verte à droite et aire rouge à gauche, Fig. 2.5 g et aire rouge à gauche, Fig. 2.5 h) dans T2.

Pour les saumons au repos, les probabilités de présence étaient spatialement plus homogènes que pour les saumons en activité (Fig. 2.5 i-p). Dans les essais en calibration, les modèles RL surestimaient l'abondance locale (aires rouges à gauche de la Fig. 2.5 i et à droite de Fig. 2.5,k), alors que les modèles AC sous-estimaient légèrement l'abondance dans T2 (aires rouge et verte à gauche dans la Fig. 2.5 l). Dans les essais en validation, le modèle RL donnait de bonnes prédictions dans T1 (Fig. 2.5 m) mais surestimait la probabilité de présence dans T2 (aire rouge à gauche de la Fig. 2.5 o), alors que les modèles AC sous-estimaient la probabilité de présence dans T2 (aire verte à gauche et aires rouges au bas de la Fig. 2.5 p). Globalement, les deux types de modèles (RL et AC) produisent des cartes prévisionnelles pour

les essais en calibration et en validation qui représentent de façon assez précise la distribution actuelle des saumons en activité et au repos.

Discussion

Sélection de l'habitat et comportement

Les modèles finaux pour les saumons en activité et au repos incorporaient différentes variables prévisionnelles, suggérant ainsi que des modèles de sélection de l'habitat basés seulement sur des comportements actifs peuvent mener à une vue d'ensemble incomplète de la sélection de l'habitat pour les juvéniles du saumon de l'Atlantique. La compétition (Kalleberg 1958; Fausch & White 1981) et les risques associés à la prédation (Metcalfe et al. 1987; Gotceitas & Godin 1993) sont des interactions biotiques qui peuvent inciter les poissons à rechercher du couvert et des refuges (Gries & Juanes 1998). Les roches non-imbriquées dans le lit de la rivière procurent une isolation visuelle contre les interactions agonistiques ainsi qu'une protection contre les prédateurs et les forts courants. Par ailleurs, la présence de couvert pourrait permettre de soutenir des densités plus élevées de juvéniles des poissons salmonidés (Kalleberg 1958; Fausch & White 1981; Gries & Juanes 1998).

Comparaisons des modèles

Pour être utile comme outils de gestion et de conservation, les modèles de sélection de l'habitat devraient être exactes (haut pouvoir prédictif), généraux (bonne capacité de transfert, peu de changement dans les prédictions entre les sites) et faciles à utiliser et à appliquer (parcimonie, facilité d'interprétation) (Lek et al. 1996; Guisan & Zimmermann 2000). Nos résultats suggèrent que la RL et les AC sont des outils adéquats, mais non équivalents, pour modéliser la distribution des juvéniles du saumon de l'Atlantique.

Pour les deux types de modèles (RL et AC), il y avait une bonne concordance entre les prédictions et les observations autant en calibration qu'en validation (Fig. 2.3). Dans les essais en calibration et validation, la performance de la RL et des AC était très semblable (Fig. 2.3 et Fig. 2.4). Cependant, les deux mesures de performance (Kappa et le log des odds-ratio) pour les AC dans les essais en validation étaient moins variables que dans le cas des RL. De plus, ces deux mesures de performance entre les deux tronçons déclinaient moins (particulièrement pour les saumons au repos) pour les AC que pour les RL (Fig. 2.4). Le déclin de performance entre les essais en calibration et en validation donne une mesure utile de la généralité d'un modèle et illustre l'importance d'effectuer des tests de validation croisée sur le terrain. Étant donné que la validation externe est moins sujette à des spécificités statistiques ou écologiques particulières à un site d'étude que la validation interne (ex. : méthodes d'amorce « angl. : *bootstrap* » ou de partition en portefeuille « angl. : *jacknife* »), elle permet une évaluation plus rigoureuse et plus réaliste du potentiel de transfert des modèles (Fielding & Bell 1997).

Les modèles RL et AC différaient dans leur facilité d'utilisation et d'interprétation. L'élaboration des modèles RL peut être plus contraignante et peut demander plus d'étapes (vérification des suppositions associées à la méthode, standardisation des variables, examen de la tolérance du modèle et détermination du SOD) que celle des modèles AC. De plus, les modèles et les interactions entre les variables associées aux modèles AC étaient plus faciles à interpréter que dans le cas des modèles RL, les AC étaient plus parcimonieux, retenaient moins de variables prévisionnelles, étaient plus comparables entre eux (retenaient les mêmes termes dans les modèles) que les modèles RL. Également, les AC généraient des résultats graphiques simples.

Cette étude souligne l'importance d'examiner les comportements passifs dans les

modèles de sélection du microhabitat, afin de raffiner la description de la distribution spatiale des juvéniles du saumon de l'Atlantique et d'identifier les besoins de l'habitat associés aux différents comportements. Plus spécifiquement, la distribution spatiale des saumons en activité différait de celle des saumons au repos, apparemment comme résultat d'une association différentielle avec les caractéristiques de l'habitat. En effet, les saumons en activité sélectionnaient un habitat en fonction des profondeurs (> 30 cm) tandis que les saumons au repos étaient associés avec la présence de couvert rocheux. Étant donné que les cartes prévisionnelles sont habituellement développées à des grandes échelles spatiales, en utilisant des procédures quantitatives assez complexes (cf. Guay *et al.* 2000; Guensch, Hardy & Addley, 2001), il est remarquable de constater que des modèles RL et AC relativement simples généraient des cartes précises de la distribution des saumons dans un petit ruisseau tributaire (Fig. 2.5).

Chapitre 2. Predicting microhabitat selection in juvenile Atlantic salmon *Salmo salar* by 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

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

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

SUMMARY

1. We compared the capacity of logistic regression (LR) and classification tree (CT) models for predicting microhabitat use and summer distributions of juvenile Atlantic salmon, *Salmo salar*, in two reaches of a small stream in eastern Quebec.

2. The models predicted the presence vs. absence of salmon at a site, either in activity or at rest, on the basis of habitat features (depth, current velocity, presence of instream and overhead cover, substratum size, and distance to stream bank) measured at the site. 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 observed fish positions with those predicted by LR and CT models.

3. The spatial distributions of fish in activity differed markedly from those of fish at rest, apparently as a result of differential selection for depths greater than about 30 cm by fish in activity and for presence of rocky cover by fish at rest.

4. All models had high prediction accuracy, and transferability in crossover validation. For both LR and CT models, the prediction maps provided a remarkably accurate picture of actual fish distributions. However, CT models were easier to build and interpret than LR models. CT models also had less variable performance and smaller decline in performance (for fish at rest) in validation, suggesting that they may be more transferable than LR models.

Introduction

Habitat selection by juvenile salmonid fish 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 should be studied within the context of temporal and spatial variations in habitat conditions (Heggenes *et al.* 2002). Previous studies have shown that habitat selection in 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 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 increasing the visual isolation among individuals (Kalleberg 1958; Fausch & White 1981).

Many models of microhabitat selection do not distinguish between active behaviours, such as foraging, and more passive behaviours, 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 behaviours 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). CT 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 "internally" validate 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 vs. 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), normalized mutual information (NMI), and odds ratio (OR) or log-odds ratio (LOR), use more effectively the information in the confusion matrix and allow for assessment of the extent to which models correctly predict occurrence at rates better than chance expectation (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 juvenile Atlantic salmon in activity and at rest. For the two types of behaviour, we: (1) compare logistic regression (LR) and classification tree (CT) models for predicting summer distributions at the microhabitat scale, (2) 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), (3) use multiple measures of prediction capability to assess model performance, and (4) 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^o 27' 20" N, 66^o 01' 70"

W). Two reaches were studied, one (R1) located approximately 100 m from the brook mouth, 75 m above sea level, and the other (R2) 100 m upstream of R1 (Fig. 2.1). Both reaches were 75 m long and 15 to 20 m wide, and encompassed sequences of riffle, run, and pool habitats. Atlantic salmon, brook trout (*Salvelinus fontinalis* Mitchill), and slimy sculpin (*Cottus cognatus* Richardson) were present at the site.

The sampling schedule covered a 63-d period between 28 June (flow: 1.58 $m^3 \cdot s^{-1}$) and 29 August 2002 (flow: 0.55 $m^3 \cdot s^{-1}$). 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 done on two consecutive days: fish observations and microhabitat measurements were made on first day, and habitat characterizations used to build prediction maps were made on the second day.

Underwater fish observation and microhabitat measurement

Fish were observed by snorkelling. Underwater visibility exceeded 5 m during dives, which were always done between 11:00 and 14:00. Each diving session covered one 5-m section of the reach and lasted 60 to 120 min depending on the number of fish encountered. To avoid startling fish, the diver entered the stream 10 to 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 cm), distance from bottom (nearest cm), 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 provided by the diver, and fish position was then marked using a numbered rock.

For each reach, a subset of sites (90 sites in R1 and 106 sites in R2, to approximately match the number of sites with fish observations) were randomly selected from a uniform XY grid (1 x 1 m cells) covering the entire surface of the reach, and marked as "absence" sites if they were not used by fish at the time of observation (over a period of at least 3 min). A random subset was used because including all absence sites in the reaches would have greatly reduced prevalence, possibly leading to an artificial increase in accuracy (Fielding & Bell, 1997). Any absence site within a radius of 50 cm of a "presence" site was discarded and replaced by another randomly chosen site.

At each marked site, we recorded water depth, current velocities at 15% and 40% depth (from bottom) (pygmy–type meter; Scientific Instruments 1205), substratum size (modified Wentworth scale; Table 2.1), presence of instream cover within a 15 cm radius of the fish (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-d study period (28 June - 29 August). LR and CT models were developed separately for each study reach and behaviour in calibration trials (a total of eight models: 2 model types x 2 behaviours x 2 study reaches). The models aimed at predicting presence vs. absence of salmon at a site, either in activity or at rest,

on the basis of habitat features at the site. An alternative approach, in which activity and resting 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).

LR 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 that can account for non-linear effects:

$$p = \frac{e^{b_0 + \sum_{i=1}^{k} b_i x_i}}{1 + e^{b_0 + \sum_{i=1}^{k} b_i x_i}}$$

where the x_i are single-variable or higher-order habitat predictors, b_0 is a constant, the b_i are regression coefficients associated with the *k* predictors, and *e* is the base of natural logarithms.

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-standardized prior to calculating products of variables, to remove non-essential collinearity in quadratic and interaction terms and facilitate comparisons among predictors. A stepwise procedure, forward selection with a nominal cut-off $p = 0.05$, was used to determine which variables should be retained in the final models. The tolerance (a measure of the amount of variation unique to each predictor; Tabachnick & Fidell, 2000) was greater than 0.68 for all predictor variables in final models, indicating only mild collinearity among predictors. Because model performance can be highly sensitive to 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 threshold of $p = 0.5$ often 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 equalize the costs of misclassifying species as present (sensitivity) or absent (specificity) (Fielding & Bell, 1997).

The RPART3 software library (Atkinson & Therneau, 2000) was used to develop CT models (S-PLUS program, v. 6.1). RPART3 uses the binary recursive partitioning algorithm developed by Breiman *et al.* (1984), which is the best-known, most dependable, and most thoroughly tested one 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 that maximizes the homogeneity within the two resulting subgroups (nodes) with respect to the response variable. We penalized models by use of a cost-complexity parameter to obtain an optimal tree size balancing the number of terminal nodes with the misclassification error rate (Atkinson & Therneau, 2000). A 10-fold cross-validation was used to estimate prediction error. Final tree size was determined by the 1-SE rule, which favours the largest tree for which the cross-validated error falls within one standard error 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).

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 confusion matrices: correct classification rate (CCR; percentage of all cases correctly predicted), sensitivity (percentage of true presences correctly predicted), and specificity (percentage of true absences correctly predicted).

Four additional measures were calculated from the confusion matrices to assess whether model performance differed from expectations based on chance alone: Cohen's kappa (κ, proportion of specific agreement; range: -1 to 1), Matthews correlation (MC; range: -1 to 1), normalized mutual information (NMI; range: 0 to 1), and the log odds-ratio (LOR; range: $-\infty$ to ∞). For all measures, a value of zero indicates no difference from random prediction. We used formulae in Baldi *et al.* (2000) to calculate NMI (there appears to be a typographical error in Fielding & Bell (1997) and Manel *et al.* (2001), who give instead 1 – NMI).

Prediction maps

Prediction maps for salmon in activity and at rest 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 velocities, substratum 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 x 1 m cell of the XY grids. For each

reach, water temperature was measured at 15-min intervals over the whole study period (Vemco thermograph), and stream flow was measured approximately twice a 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 probability values over the whole surface of the reaches (distance-weighted least-squares; program SYSTAT, v. 10.2). The resulting contours of probabilities of occurrence were represented as colour-coded intervals in the maps. Actual fish locations were also included in the maps to evaluate the agreement of model predictions with observed fish distribution.

Results

Microhabitat models

Average habitat conditions were broadly comparable in the two study reaches (Table 2.1), although patterns of spatial heterogeneity varied between reaches, as illustrated by the differences in depth contours (Fig. 2.1). Water temperature varied between 5.8 °C and 16.3 °C (mean ± SD: 10.1 ± 2.1 °C) over the sampling period. The same key predictor variables were retained in LR and CT models, both for active and resting salmon (Table 2.2, Fig. 2.2). Salmon in activity selected positions based mostly on water depth and avoided shallow sites, whereas

salmon at rest selected positions behind or edging an unembedded rock >20 cm. However, the LR models always retained more predictors than the CT models.

LR models differed between reaches for a given behaviour (Table 2.2). The final model for salmon in activity included four variables (depth, velocity at 40% depth, distance to bank, and presence of an unembedded rock > 20 cm) and one quadratic term (velocity at 40% · velocity at 40%) in R1, but only one variable (depth) and one quadratic term (depth · depth) in R2. The final model for resting salmon included five variables (depth, distance to bank, substratum size, and presence of an unembedded rock > 20 cm) and one interaction term (substratum size · depth) in R1, but only three variables (velocity at 40% depth, substratum size, and presence of an unembedded rock > 20 cm) and one quadratic term (velocity at 40% · velocity at 40%) in R2.

Only one or two variables were useful predictors in the final CT models (Fig. 2.2). Final models for salmon in activity were almost identical in the two reaches, including the same single predictor (water depth) and very similar splitting values. For salmon at rest, final models differed slightly between reaches, but the most influential variable in both reaches was the presence of an unembedded rock > 20 cm.

Model accuracy (CCR) for calibration trials varied slightly between LR and CT models (Fig. 2.3). For salmon in activity, the CCR was 83.3% for LR and 84.0% for CT. For salmon at rest, the CCR was 88.4 % for LR and 87.2% for CT. Variation in CCR between reaches was low for all calibration models.

CCR remained high for both model types in validation trials. For salmon in activity, the CCR was 72.2% for LR and 77.7% for CT (Fig. 2.3). For salmon at rest, the CCR was 77.9% for LR and 85.3% for CT. Variation in CCR between reaches remained low for both model

types in the validation trials. LR and CT models generally had higher specificity than sensitivity in calibration and in validation trials. Variation in specificity and sensitivity in validation trials was usually 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 Cohen's kappa and the log-odds ratio (Fig. 2.4). Model performance was better than random for all cases (none of the 95% confidence intervals includes zero). Performance generally declined between 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 salmon in activity. 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 salmon in activity (Fig. 2.5 a-h) differed markedly from that of salmon at rest (Fig. 2.5 i-p). For salmon in activity, predicted probabilities of presence were spatially heterogeneous. In calibration trials, LR and CT yielded similar prediction maps in R1, both of which closely matched the observed distributions (Fig. 2.5 a,b). LR and CT also yielded similar prediction maps for salmon in activity in R2, but the CT overestimated the probability of presence near an area where salmon were locally abundant (red area to the left of Fig. 2.5 d). In validation trials, LR underestimated the probability of presence of salmon in activity in R1 (blue area to the left of Fig. 2.5 e), in contrast with CT, which provided accurate predictions (Fig. 2.5 f). LR overestimated the

probability of presence (green area to the right, and red area to the left, of Fig. 2.5 g) in R2, as did CT (red area to the left of Fig. 2.5 h).

For salmon at rest, predicted probabilities of presence were spatially less heterogeneous than for salmon in activity (Fig. 2.5 i-p). In calibration trials, LR tended to overestimate local abundance (red areas to the left of Fig. 2.5 i and the right of Fig.2.5 k), whereas CT slightly underestimated local abundance in R2 (red and green areas to the left of Fig. 2.5 l). In validation trials, LR provided accurate prediction in R1 (Fig. 2.5 m) but overestimated the probability of presence in R2 (red area to the left of Fig. 2.5 o), whereas CT again underestimated the probability of presence in R2 (green area to the left and red areas at the bottom of Fig. 2.5 p). Overall, however, for both types of model the prediction maps for calibration and validation trials provided an accurate picture of the actual distributions of fish in activity and at rest.

Discussion

Microhabitat selection in active and resting fish

Final models for fish in activity and at rest incorporated substantially different predictor variables and yielded different prediction maps, suggesting that models based solely on active behaviours such as foraging may yield an incomplete picture of microhabitat selection in juvenile Atlantic salmon. For example, salmon at rest can be abundant in areas predicted to have low probability of occurrence by a model focusing on activity (cf. Figs. 2.5 a-h; 2.5 i-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). Availability of shelter may affect salmon populations because individuals that fail to find shelter may either be forced to emigrate or, more likely, removed by predators (Metcalfe *et al.* 1987; Gotceitas & Godin, 1993). Consequently, the availability of instream structures such as large unembedded rocks, which provide visual isolation and protection from predators and flow, may affect population density of juvenile salmonids.

Juvenile Atlantic salmon can adapt rapidly to changing environmental conditions; 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 - 29 August) during which flow ranged between 0.55 and 1.58 $m^3 \cdot s^{-1}$ provided accurate prediction of habitat selection for fish both in activity and at rest. 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, which may improve their 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 (CCR > 76.4%). Values of performance measures were high in calibration trials and declined in validation trials for both methods (Fig. 2.4). In both calibration and validation trials, the performance of LR and CT was broadly similar (Figs. 2.3 and 2.4) with the exception of the LR ($p = 0.5$) model for salmon at rest, which had lower sensitivity than the CT model in the validation trial (Fig. 2.3). 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. 2.4). LR models based on 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 presences, i.e. those incurred when the model incorrectly classifies as poor (absence) a site 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, when feasible, 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 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.

LR 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 optimal decision thresholds (although not all these steps will be needed in all instances). In contrast, CT did not require transformation or standardization 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. In comparison, CT models generated simpler graphical interpretations (Fig. 2.2) 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 behaviours in habitat selection models, as a means for refining the description of spatial distribution of juvenile Atlantic salmon and identifying habitat needs in relation to these behaviours. Specifically, the spatial distributions of fish in activity differed markedly from those of fish at rest, 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. Remarkably, relatively simple LR and CT models for the distribution of fish in small stream reaches sufficed to generate accurate prediction maps (Fig. 2.5), 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).

Acknowledgments

We thank M.-N. Rivard and M. Chénier-Soulière for assistance in the field, J. Deschênes and P. Peres-Neto 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 program 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 #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. & 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.

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.

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 Gibson, R.J. & 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.

Grant, J.W.A. & Kramer, D.L. (2000) Territory size as a predictor of the upper limit to population density of juvenile salmonids in streams, *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 1724-1737.

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 south-west 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 ed. 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.

Noakes, D.L.G. & McNicol, Richard E. (1982) Geometry for the eccentric territory. *Canadian Journal of Zoology*, **60**, 1776-1779.

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.

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

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 ed. 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.

Table 2.1. Habitat variables (mean ± SD) characterizing sites used by juvenile Atlantic salmon in activity and at rest, and unused sites, by reach.

Reach	Site type	Fish length (cm)	Depth (cm)	Mean current velocity ($cm \cdot s^{-1}$)	Substratum size*	Distance to stream bank (m)
1	Activity (N=92)	10.4 ± 1.7	41.9 ± 14.9	43.3 ± 14.6	8.5 ± 1.4	3.7 ± 1.2
	At rest (N=37)	10.3 ± 1.3	28.5 ± 12.5	40.2 ± 22.4	7.9 ± 2.4	3.3 ± 1.3
	Absence (N=90)	-	21.7 ± 8.8	35.7 ± 21.8	8.2 ± 1.6	2.6 ± 1.5
2	Activity (N=46)	9.7 ± 1.7	34.3 ± 9.2	62.4 ± 25.4	8.7 ± 2.0	3.1 ± 1.6
	At rest (N=25)	11.9 ± 1.6	27.0 ± 12.8	46.9 ± 23.6	9.8 ± 2.3	3.9 ± 1.8
	Absence (N=106)	-	18.8 ± 12.4	43.8 ± 30.3	8.5 ± 1.6	3.8 ± 1.8

Wentworth scale, adapted from DeGraaf and Bain (1986): 1: clay (<0.004 mm); 2: silt (0.004-0.062 mm); 3: sand (0.062-2.0 mm); 4: fine gravel (21.-8.0 mm) ; 5: gravel (8.1-16 mm); 6: pebble (16.1-32 mm); 7: coarse pebble (32.1-64 mm); 8: cobble (64.1-128 mm); 9: rubble (128.1-256 mm); 10: small boulder (256.1-384 mm); 11: medium boulder (384.1-512 mm); 12: large boulder (>512 mm); 13: bedrock.

Table 2.2. Coefficients of logistic regression models for activity and resting behaviours, by reach. Coefficients are given only for terms retained by the stepwise selection procedure ($p < 0.05$). All models were globally significant at $p < 0.0001$.

Model term	Activity		At rest	
	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 size	-	-	0.622	1.008
Rock > 20 cm	0.610	-	1.816	1.721
Depth · Depth	-	-1.756	-	-
Velocity 40% · Velocity 40%	-0.497	-	-	-0.862
Substratum · Depth	-	-	-0.722	-

FIGURE CAPTIONS

Figure 2.1. Study reaches in Big Jonathan Brook, a tributary of the Grande Cascapedia River, Quebec. Both reaches are approximately 75 m long and 15 to 20 m wide. Contour lines within the study reaches represent water depth (cm). Woody debris and submerged rocks > 30 cm are also shown.

Figure 2.2. Classification tree models for predicting activity vs. absence (a: Reach 1; b: Reach 2) and resting vs. absence (c: Reach 1; d: Reach 2). Vertical bars represent the frequency of absence (black) and presence (white) at each node. Splitting rules 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.

Figure 2.3. Correct classification rate (CCR), specificity, and sensitivity of logistic regression (LR) and classification tree (CT) models for activity and resting behaviours, in calibration and validation (crossover field test) trials. Logistic regression 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 applied as well in validation trials, were: activity, R1: 0.47, R2: 0.40; resting, R1: 0.23, R2: 0.15. Symbols represent the mean of the two reaches; vertical lines represent the range between reaches.

Figure 2.4. 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 logistic regression are based on the optimal decision threshold. Reported values are Cohen's kappa (κ) and log-odds ratio (LOR), with 95% confidence intervals.

Figure 2.5. 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.

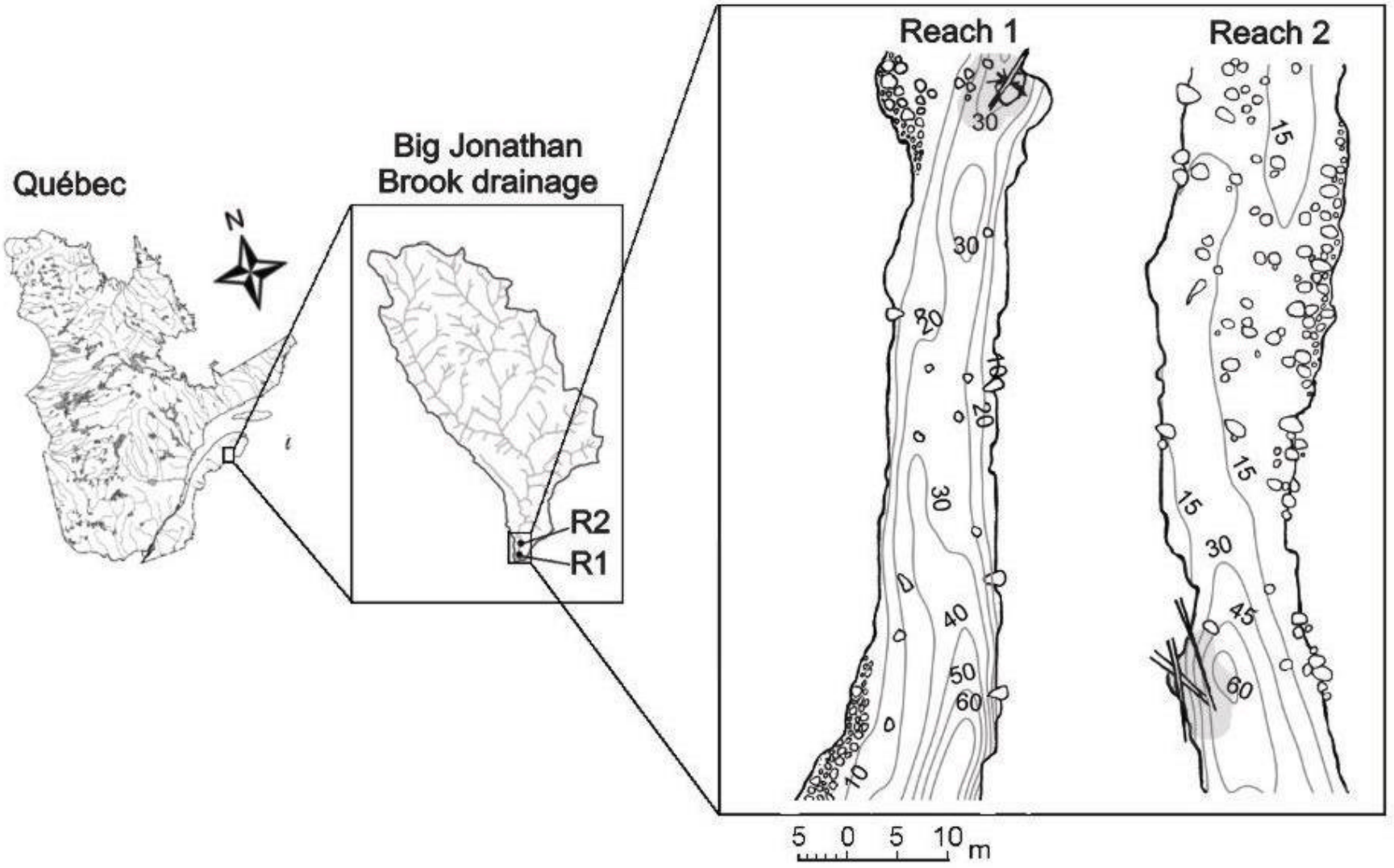

Figure 2.1
Turgeon & Rodríguez 2004

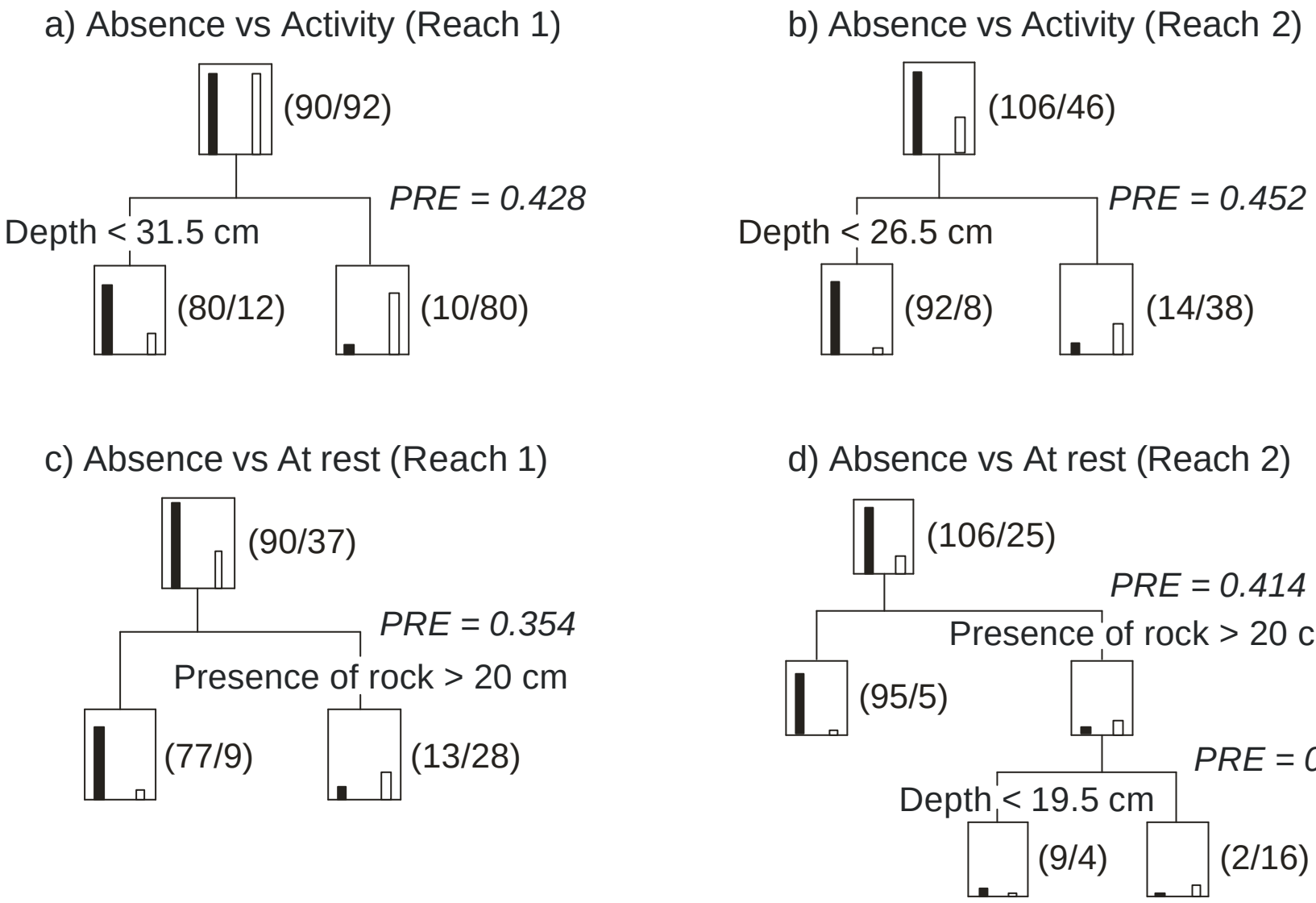

Figure 2.2
Turgeon & Rodríguez 2004

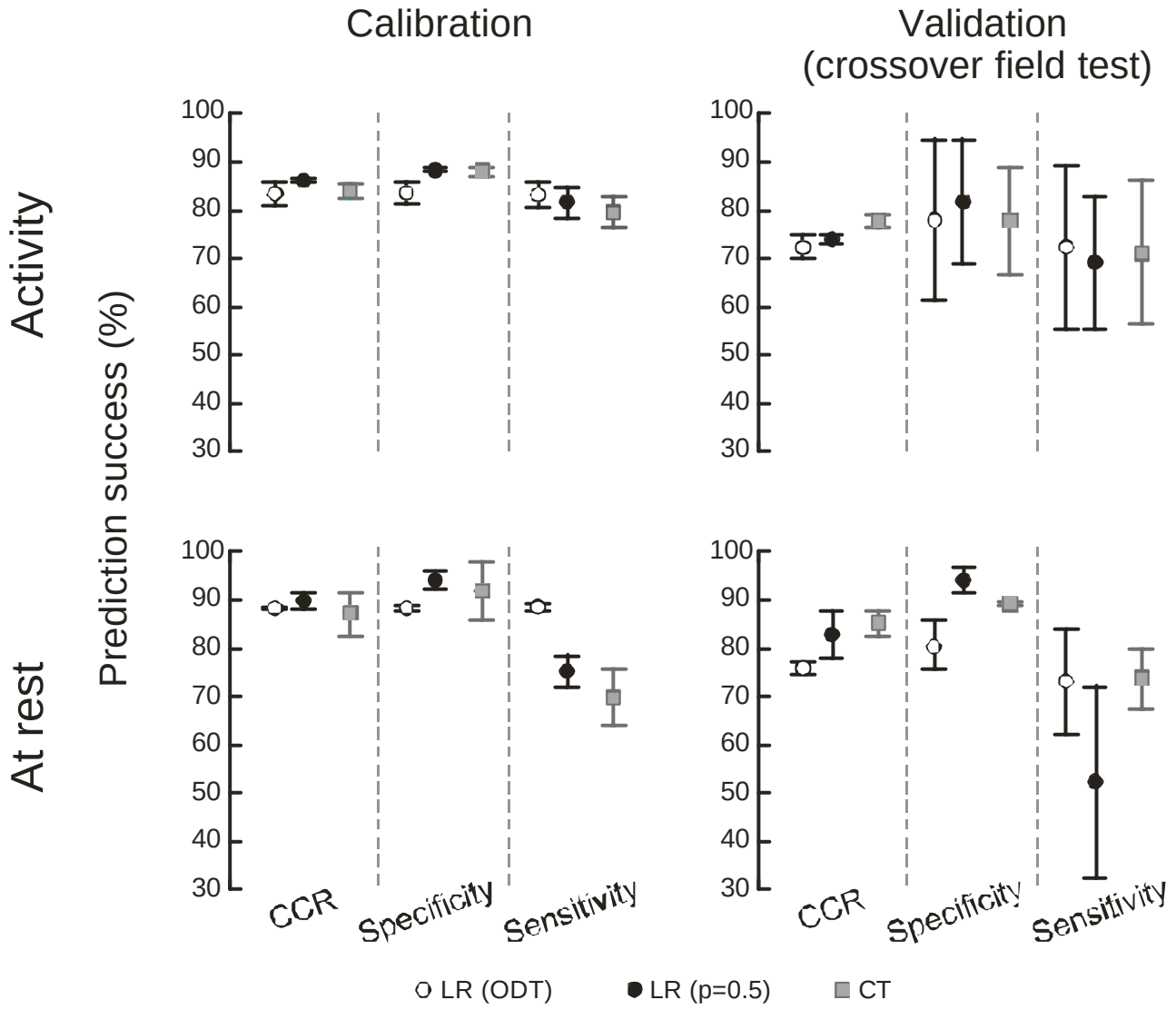

Figure 2.3
Turgeon & Rodríguez 2004

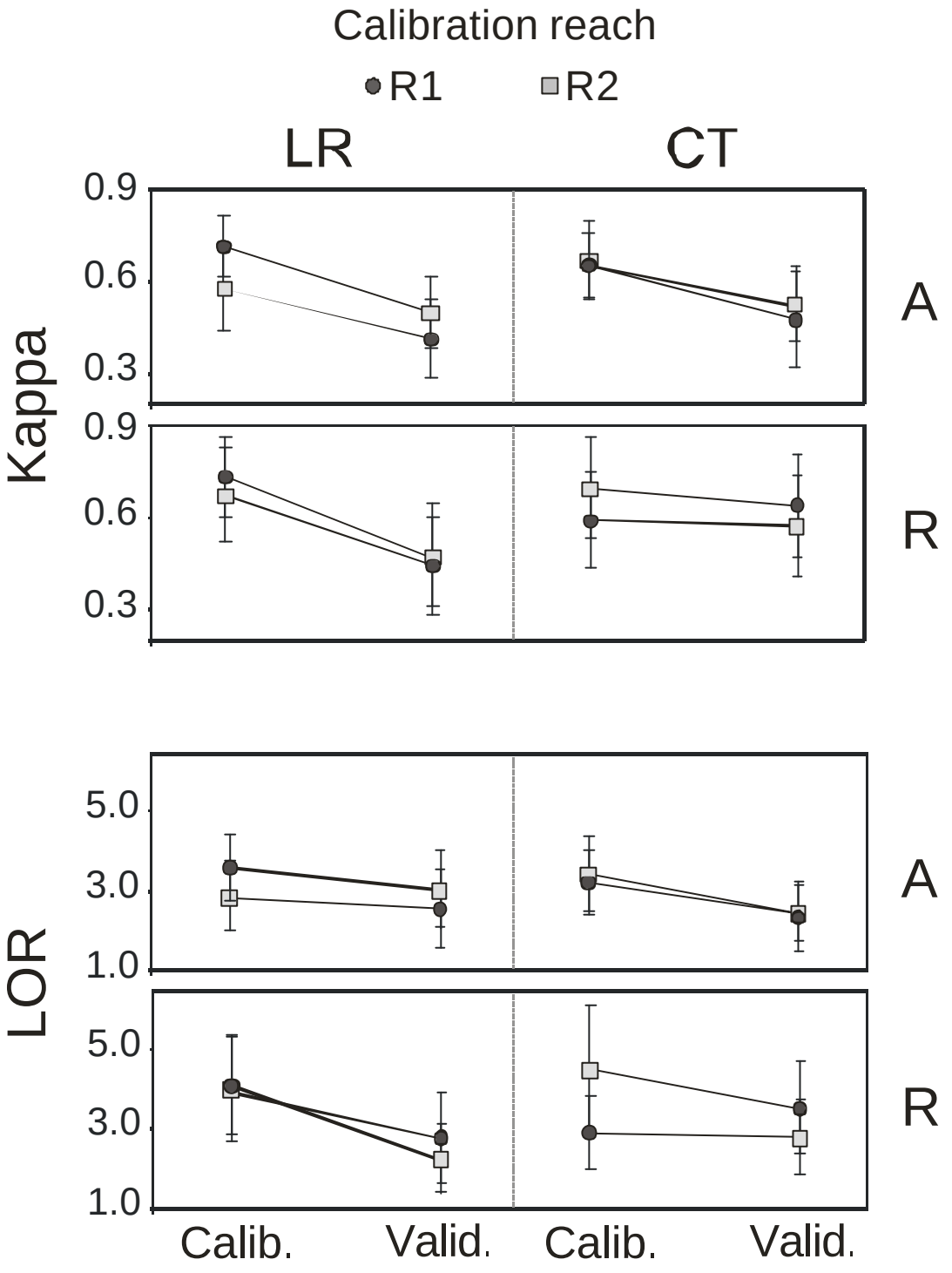

Figure 2.4.
Turgeon & Rodríguez 2004

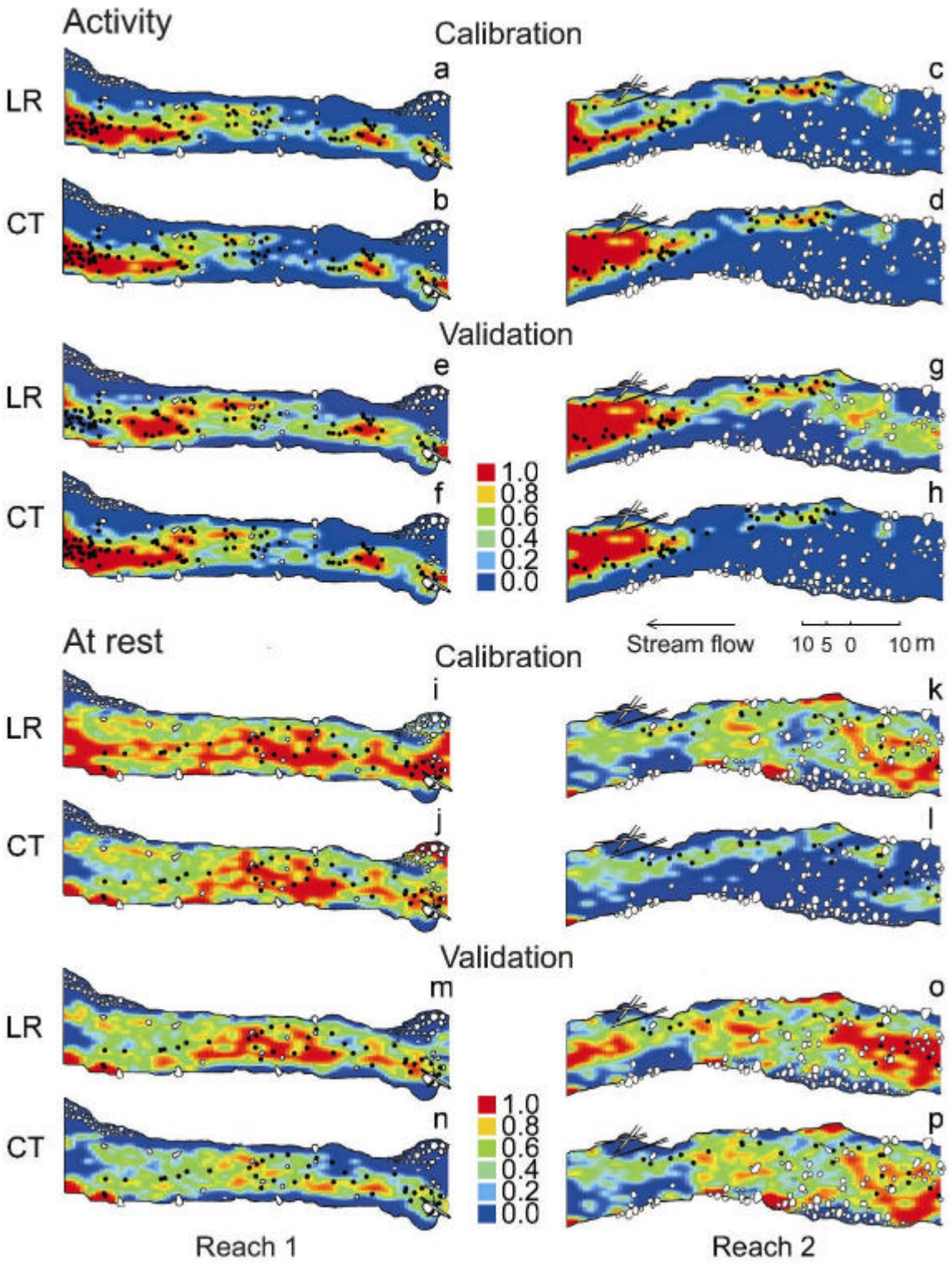

Figure 2.5
Turgeon & Rodríguez 2004