Université du Québec
Institut National de la Recherche Scientifique
Centre Eau Terre Environnement

LA RÉGRESSION FONCTIONNELLE POUR MODÉLISER LA TEMPÉRATURE DE L'EAU ET L'HABITAT DU SAUMON ATLANTIQUE JUVÉNILE

Par

Jérémie Boudreault

Mémoire présenté pour l'obtention du grade de
Maître ès sciences (M.Sc.)
en Sciences de l'eau

Jury d'évaluation

Examinateur externe	Thierry Duchesne Université Laval
Examinateur interne	Taha B. M. J. Ouarda INRS – Eau-Terre-Environnement
Directeur de recherche	Fateh Chebana INRS – Eau-Terre-Environnement
Codirecteurs de recherche	André St-Hilaire INRS – Eau-Terre-Environnement Canadian River Institute
	Normand Bergeron INRS – Eau-Terre-Environnement

REMERCIEMENTS

Premièrement, merci à mon directeur de recherche, Fateh Chebana, d'avoir été disponible et à l'écoute avec moi qui en étais à mes débuts en recherche. Il a été un grand plaisir de mener ces travaux de maîtrise sous ta supervision et tes critiques constructives ont permis de m'améliorer constamment durant ces deux dernières années.

Deuxièmement, merci à mes deux codirecteurs qui ont eu une valeur ajoutée significative (*je ne nommerai pas le test statistique effectué*) à mes travaux. Merci André St-Hilaire d'avoir été présent, rigoureux, encourageant et disponible. Ta rigueur et tes connaissances si diverses seront deux choses dont je me rappellerai toujours. Merci Normand Bergeron d'avoir cru en moi dès le début de ce projet et de m'avoir donné tout l'encadrement possible pour que je puisse être autonome et mener à terme ce projet. Les nombreuses conversations que nous avons eues m'ont fait me questionner longuement sur mes analyses pour en arriver avec un produit final plutôt réussi.

Troisièmement, merci aux deux examinateurs de ce mémoire, M. Taha B. M. J. Ouarda et M. Thierry Duchesne, d'avoir gentiment accepté d'évaluer mes travaux de recherche ce qui m'a permis de grandement les bonifier.

Quatrièmement, je tiens à remercier mes deux professeurs d'actuariat de l'Université Laval, Étienne Marceau et Hélène Cossette, qui ont cru en moi et qui m'ont aidé à réorienter mes idées de carrière après mon baccalauréat (idées qui allaient un peu n'importe où). C'est grâce à vous que j'ai atterri à l'INRS où j'ai eu la chance de travailler sur ce projet plus que stimulant.

Finalement, merci à toutes les personnes qui ont été là au cours des deux dernières années. Mes parents (merci notamment à mon père d'être mon correcteur par excellence), mon frère (qui est et sera toujours mon modèle de réussite) et mes amis qui ont su être là pour me sortir la tête de mes travaux durant ces années et d'avoir fait en sorte que « ça se passe ». Merci à la plus importante, ma copine, qui a été ma plus grande *fan*, de m'avoir supporté pendant mes *down*, mais aussi d'avoir célébré mes victoires et de s'être intéressée autant que moi à mes modèles statistiques de petits poissons. Merci.

RÉSUMÉ

Alors que les modèles de régression classiques et d'apprentissage artificiel sont basés sur des variables ponctuelles (scalaires ou vecteurs), l'approche de régression fonctionnelle l'utilisation de fonctions ou courbes tant pour les variables réponses qu'explicatives. Dans le domaine de l'hydrologie, les phénomènes peuvent souvent être mieux représentés par des courbes continues que des quantités scalaires. Par exemple, la température de l'eau observée pendant la saison estivale peut être vue de manière continue comme une seule observation : une courbe. Les variables d'habitat du saumon peuvent aussi être traitées comme des distributions de fréquence représentant mieux son habitat. Ainsi, la régression fonctionnelle semble mieux adaptée à ces problématiques que les approches classiques, permettant de reproduire plus naturellement les phénomènes observés. Dans ce mémoire, cette approche statistique (la régression fonctionnelle), qui connait de nombreux développements dans les dernières années, est utilisée pour deux problématiques distinctes, mais fortement liées. La première étude concerne la température de l'eau où la régression fonctionnelle est utilisée pour modéliser la courbe complète des températures de l'eau pour la saison estivale à partir de la température de l'air. Son application sur trois cours d'eau des États-Unis a montré des températures prédites plus précises que deux modèles classiques comparés : le modèle logistique et le modèle additif généralisé. De plus, la régression fonctionnelle indique facilement et clairement les périodes où la température de l'air influence celle de l'eau. Dans la deuxième étude, l'habitat du saumon atlantique juvénile est considéré, alors que les modèles actuels n'arrivent que peu à prédire la productivité d'une rivière et manquent de transférabilité (c.-à-d. un modèle développé sur une rivière ne fonctionne pas sur d'autres rivières). L'utilisation de la régression fonctionnelle nous permet d'inclure toute la variabilité des variables de l'habitat du poisson sélectionnées comme prédicteurs en utilisant des histogrammes lissés (des fonctions de densité) pour chaque variable typiquement utilisée dans la modélisation de l'habitat : la profondeur de l'eau, la vitesse du courant et la taille du substrat. De plus, l'ajout d'une nouvelle variable en modélisation de l'habitat aquatique, soit la température de l'eau, est une innovation importante car cette variable est souvent absente des modèles d'habitat. Pour valider cette approche, le

travail de terrain a été effectué durant l'été 2017 alors que deux rivières à saumon ont été échantillonnées par pêche électrique à plusieurs sites : la rivière Sainte-Marguerite et la rivière Petite–Cascapédia. Le modèle de régression fonctionnelle a été utilisé pour prédire l'abondance ou la présence-absence de saumon juvénile pour les trois classes (0+, 1+ et 2+) à chacun des sites étudiés. À l'aide d'une validation croisée, le modèle fonctionnel a montré les meilleurs résultats pour les trois classes comparé à un modèle linéaire généralisé et un modèle additif généralisé, deux approches fréquemment utilisées en modélisation d'habitat. De plus, le modèle fonctionnel est celui qui a fourni les meilleures informations quant aux valeurs préférées des variables d'habitat par le saumon juvénile. Pour ce qui est de la transférabilité, le potentiel de transférabilité n'a pu être démontré que pour le modèle de régression fonctionnel de présence-absence. Les deux sujets étudiés démontrent que la régression fonctionnelle est un outil clé pour la modélisation et la prédiction dans le domaine de l'hydrologie et l'écologie en rivière.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

Abréviations des termes statistiques :

ADF (*FDA*)	Analyse de données fonctionnelles (*Functional Data Analysis*)
MRF (*FRM*)	Modèle de régression fonctionnelle (*Functional Regression Model*)
MFLC (*FFLM*)	Modèle fonctionnel linéaire complet (*Fully Functionnal Linear Model*)
MFLH (*HFLM*)	Modèle fonctionnel linéaire historique (*Historical Functional Linear Model*)
MFLS (*FLMS*)	Modèle fonctionnel linéaire pour réponse scalaire (*Functional Linear Model for Scalar response*)
MLG (*GLM*)	Modèle linéaire généralisé (*Generalised Linear Model*)
MAG (*GAM*)	Modèle additif généralisé (*Generalised Additive Model*)
ML (*LM*)	Modèle logistique (*Logistic Model*)
FDP (*PDF*)	Fonction de densité de probabilité (*Probability Density Function*)
EDN (*KDE*)	Estimation de la densité par noyau (*Kernel Density Estimate*)
EXA (*ACU*)	Exactitude (*Accuracy*)
TPP (*TPR*)	Taux de présences correctement prédites (*True Presence Rate*)
TAP (*TAR*)	Taux d'absence correctement prédites (*True Absence Rate*)
NSC (*NSC*)	Coefficient de Nash-Sutcliffe (*Nash-Sutcliffe Coefficient of efficiency*)
RMSE (*RMSE*)	Racine de l'erreur quadratique moyenne (*Root Mean Square Error*)

Abréviations à propos de l'habitat du saumon :

APU (*WUA*)	Aire pondérée utile (*Weighted Usable Area*)
IQH (*HSI*)	Indice de qualité de l'habitat (*Habitat Sustainability Index*)
IPH (*HPI*)	Indice probabiliste de qualité de l'habitat (*Habitat Probabilistic Index*)

Abréviations des rivières à l'étude :

RSM (*SMR*)	Rivière Sainte-Marguerite (*Sainte-Marguerite River*)
RPC (*PCR*)	Rivière Petite-Cascapédia (*Petite-Cascapedia River)*

1 INTRODUCTION

Après avoir décrit le contexte général des travaux de ce mémoire, le saumon atlantique et ses stades de vie seront brièvement abordés dans le but d'introduire les notions clés à la compréhension du document. Par la suite, les impacts touchant son habitat et la variable de la température de l'eau seront traités. Les divers modèles de la littérature et leurs limites seront décrits puis l'approche fonctionnelle sera introduite avec ses avantages. Finalement, les objectifs spécifiques du mémoire seront présentés.

1.1 Mise en contexte

La température de l'eau est l'une des variables les plus importantes pour les écosystèmes aquatiques (Beschta *et al.*, 1987; Caissie, 2006). En effet, les invertébrés et les poissons sont grandement affectés par des températures inadéquates (Bjornn & Reiser, 1991; Hinz & Wiley, 1998; Lessard & Hayes, 2003; Handeland *et al.*, 2008). Ainsi, les changements climatiques, par l'augmentation de la température de l'air, et par conséquent, celle de l'eau, auront de forts impacts sur la faune ichtyenne dont le saumon atlantique (p. ex. Webb, 1996; Morrison *et al.*, 2002; Ferrari *et al.*, 2007; Webb & Nobilis, 2007; Van Vliet *et al.*, 2011; Daigle *et al.*, 2015; Sundt-Hansen *et al.*, 2018). Cependant, les modifications aux régimes thermiques des rivières ne sont pas le seul problème auquel les saumons sont confrontés. Des changements dans le régime hydrologique, par des périodes d'étiages plus longues et des débits printaniers plus élevés par exemple, occasionneront aussi leur lot de conséquences chez les juvéniles comme des conditions hydrauliques sous-optimales ou des pertes d'habitat (Heggenes, 1990; Armstrong *et al.*, 2003; Tetzlaff *et al.*, 2005). Avec les activités anthropiques qui se multiplient autour des cours d'eau (barrage, déforestation, routes, activités industrielles), les facteurs de stress augmentent (p. ex. la température de l'eau) et les surfaces habitables diminuent (Heggenes, 1990; Gibson, 1993; Bardonnet & Baglinière, 2000; Prévost *et al.*, 2002; Nyqvist *et al.*, 2017). Étant donné tous ces changements dans l'habitat du saumon, des baisses de population se sont déjà fait sentir (Noakes *et al.*, 2000; Lackey, 2003), et ce, malgré des mesures telles la réduction des quotas de pêche et l'interdiction de pêche commerciale dans

certaines zones, notamment au Canada (Klemetsen *et al.*, 2003). Selon Heggenes *et al.* (1995), la perte d'habitats d'eau douce serait l'une des principales causes de déclin du saumon atlantique. En 2011, une étude commissionnée par la Fédération du Saumon atlantique avait estimé que les activités liées à la pêche sportive du saumon avaient des retombés économiques de 166 millions de dollars au Québec et dans les provinces atlantiques (Pinfold, 2011).

Ainsi, vu l'intérêt économique de la conservation du saumon atlantique, mais aussi celui de la préservation des écosystèmes lotiques, il devient primordial de (**1**) s'intéresser aux variables affectant ces écosystèmes, particulièrement celle de la **température de l'eau** et de (**2**) comprendre les variables affectant l'**habitat du saumon atlantique juvénile** pour mieux le conserver.

1.2 Le cycle de vie du saumon atlantique

Dans cette section, le cycle de vie du saumon est brièvement décrit. L'accent sera mis sur le stade juvénile ainsi que sur la définition des termes nécessaires généraux à la compréhension des parties subséquentes du mémoire. Des informations supplémentaires peuvent être trouvées sur le site de Saumon Québec (2018).

L'adulte et la reproduction : Le saumon adulte passe l'hiver en mer (eau salée). Il peut y passer entre un et trois ans avant de revenir en rivière (eau douce) durant l'été pour s'y reproduire, dans exactement la même rivière où il est né. Lors de la reproduction se déroulant à la fin de l'été, la femelle creusera d'abord un nid dans le lit de la rivière. Un habitat pouvant inclure plusieurs nids se nomme frayère (*Redd* en anglais). La femelle y déposera ensuite ses œufs qui seront fécondés par un saumon adulte ou même parfois par un saumon juvénile précoce (Beall *et al.*, 1994). Tous les hivers, les saumons adultes quitteront la rivière (eau douce) pour rejoindre l'océan (eau salée) après la reproduction. Ils pourront revenir plusieurs étés consécutifs pour s'y reproduire. Certains adultes qui passent l'hiver en rivière sont nommés saumons noirs. Les œufs fécondés passent tout l'hiver dans l'eau froide des rivières et sont recouverts de petits graviers pour les protéger, mais permettre leur oxygénation. Ceux-ci écloront au printemps suivant pour donner naissance à de jeunes saumons appelés alevins.

L'alevin : Le stade alevin ou jeune de l'année (*Fry* ou *Young-of-the-year* en anglais) débute à l'émergence en mars-avril alors que le saumon juvénile quitte les graviers où ont été pondus les œufs (la frayère). Celui-ci devra alors faire face au courant, ce qui pourra l'amener à dériver de son lieu de naissance et à se trouver un nouvel habitat (Beall *et al.*, 1994). L'alevin commencera donc à s'alimenter par lui-même, en captant de petits insectes qui dérivent dans le courant (Héland *et al.*, 1995). Ce dernier choisira les habitats qui lui sont les plus favorables pour s'alimenter, mais sans avoir à trop dépenser d'énergie pour nager à contre-courant (Gibson, 1993). À ce stade, le saumon juvénile est noté 0+, car il est âgé entre zéro et un an et mesure généralement entre 3 et 9 cm.

Le tacon : Après avoir passé un premier hiver en rivière, l'alevin porte désormais le nom de tacon (*Parr* en anglais) et on utilisera la notation 1+ et 2+ pour différencier les tacons âgés d'un et de deux ans qui utiliseront des habitats différents de par leur mobilité et leur taille. Ce stade de croissance, qui durera généralement d'un à deux ans, peut s'échelonner jusqu'à huit ans (Heland & Dumas, 1994). Les tacons 1+ et 2+ mesurent généralement entre 9 et 15 cm. C'est habituellement après trois étés passés en rivière que les tacons seront prêts à migrer de la rivière (eau douce) à la mer (eau salée) au printemps suivant. Les tacons qui s'apprêtent à quitter la rivière porteront le nom de saumoneau (*Smolt* en anglais) et des changements physionomiques s'opèreront pour se préparer à leur migration en eau salée : la smoltification (Boeuf *et al.*, 1994). Les saumoneaux quitteront la rivière en même temps que la crue printanière, processus nommé la dévalaison des smolts. Ces saumons seront désormais prêts à devenir des adultes et à revenir en rivière pour s'y reproduire.

1.3 **Impacts**

Les différents impacts humains sur la température de l'eau et sur l'habitat du saumon sont divisés en deux grandes catégories : ceux découlant des changements climatiques et ceux des activités anthropiques sur et près des cours d'eau. Finalement, l'impact de la température de l'eau sur les organismes aquatiques et le saumon atlantique est traité dans une section distincte.

1.3.1 Les changements climatiques

Selon le rapport de 2013 du groupe d'experts intergouvernemental sur l'évolution du climat (GIEC), le climat se modifie à l'échelle planétaire et il y a un consensus des scientifiques et des décideurs quant à sa cause anthropique (GIEC, 2013). Par exemple, la température moyenne de l'air au niveau mondial a subi une augmentation de 0.85°C depuis les cent-vingt dernières années (GIEC, 2013). De plus, il est démontré que le taux de cette augmentation est en croissance lui aussi (Jones *et al.*, 2012; GIEC, 2013; Rohde *et al.*, 2013). Outre une température de l'air plus élevée, une augmentation des débits de pointe hivernaux et une diminution à la fois des précipitations estivales et du ruissellement sont d'autres conséquences du changement climatique ayant des impacts sur la température de l'eau des rivières en Amérique du Nord (GIEC, 2013). Dans les scénarios futurs, la température de l'eau augmentera avec un taux d'augmentation plus grand que celui actuel (Webb, 1996; Morrison *et al.*, 2002; Ferrari *et al.*, 2007; Van Vliet *et al.*, 2011). Les rivières du nord de l'Europe se sont déjà réchauffées de 1°C depuis 1900 (Webb, 1996; Webb & Nobilis, 2007) et même s'il manque de données à long terme quant aux rivières de l'Amérique du Nord, on peut supposer les mêmes conclusions (Kaushal *et al.*, 2010). Au niveau mondial, on prévoit une augmentation de 1°C à 3°C de la température moyenne des rivières, en réponse à l'augmentation des températures de l'air (Morrill *et al.*, 2005; Van Vliet *et al.*, 2011).

Les changements climatiques attendus dans l'habitat du saumon atlantique autres que l'augmentation de la température de l'eau sont : des hivers plus doux avec plus de précipitations tombant sous forme d'eau que de neige, une diminution dans la période de couvert de glace et une augmentation des événements météorologiques extrêmes (Jonsson & Jonsson, 2009). L'hiver, le tacon a un co t métabolique énergétique plus élevé lorsque le couvert de glace est absent, ce qui peut augmenter la mortalité hivernale dans le cas d'une diminution du couvert de glace (Finstad *et al.*, 2004a; Finstad *et al.*, 2004b). Lors des périodes de très faibles débits, l'alevin est particulièrement vulnérable en raison de sa faible mobilité et sa capacité réduite à s'évader (Elliott, 1985; Elliott & Elliott, 2006). Le tacon, lui, a une meilleure capacité à se déplacer des seuils (section de rivière peu profonde souvent utilisée par les juvéniles) vers les mouilles (section plus

profonde) lors des faibles débits (Armstrong *et al.*, 1998). Cependant, les tacons qui demeurent dans les seuils en période de faible débit ont un risque de mortalité plus élevé (Berland *et al.*, 2004). Par exemple, en réponse à des débits hivernaux plus faibles dans la rivière Orkla en Norvège, l'augmentation de la relâche d'eau en période hivernale (ce qui a augmenté les débits hivernaux) a permis une hausse de la productivité des tacons (Hvidsten *et al.*, 2015). L'été, une des conséquences des étiages sur les saumons est la perte d'habitat d'alevins et de fraie, par une diminution de la superficie mouillée du chenal, amenant une possible augmentation des échouages et des mortalités de saumons juvéniles (Tetzlaff *et al.*, 2005; Bradford & Heinonen, 2008; Graham & Harrod, 2009; Jonsson & Jonsson, 2009). Aussi, les périodes de débits élevés sont risquées pour les saumons juvéniles, qui n'utilisent qu'une gamme de vitesses et de profondeurs restreintes (Heggenes, 1990; Armstrong *et al.*, 2003), ce qui pourrait réduire les habitats utilisés lors des crues plus intenses par les juvéniles (Tetzlaff *et al.*, 2005). Les précipitations hivernales et les crues printanières accrues peuvent aussi détruire les nids, causant des densités d'alevins plus faibles (Clark *et al.*, 2001; Jonsson & Jonsson, 2009; Mantua *et al.*, 2010). Par exemple, les périodes de fortes crues ont été associées à de hauts taux de mortalité d'alevins dans la rivière Saltdalselva, au nord de la Norvège (Jensen & Johnsen, 1999). La productivité des saumoneaux est d'ailleurs affectée par les débits connus par les alevins au cours de leur premier été en rivière (Jonsson & Jonsson, 2017). Cela dit, en modifiant à la fois le régime thermique et le cycle hydrologique, les changements auront des effets multiples chez le saumon atlantique.

1.3.2 Les activités anthropiques

Les activités anthropiques majeures pour la température de l'eau et l'habitat du saumon sont d'abord traitées, à savoir les barrages et la déforestation. Puis, les autres activités sont brièvement décrites avec leurs impacts respectifs.

Les barrages : Que ce soit pour l'hydroélectricité, la demande en eau potable, la régularisation des crues, l'irrigation pour l'agriculture ou les activités récréatives, les rivières sont de plus en plus régulées (Dynesius & Nilsson, 1994; Murchie *et al.*, 2008). La présence d'un barrage influence le régime thermique d'une rivière de diverses manières selon le type d'ouvrage, le mode d'opération, la prise d'eau, la position dans le

bassin, la présence d'un réservoir, etc. (Webb, 1996; Webb & Walling, 1997; Bartholow *et al.*, 2004; Olden & Naiman, 2010). Pour les barrages possédant un réservoir, leur effet sur la température dépendra de la superficie du réservoir, du temps de résidence de l'eau, de la stratification thermique et de la profondeur de la prise d'eau (Lessard & Hayes, 2003). Par exemple, des études sur des rivières régulées d'Angleterre ont noté que la présence d'un barrage causait une hausse de la température estivale, une diminution des variations diurnes et une diminution des températures maximales estivales (Webb & Walling, 1993; Webb & Walling, 1997). Ces effets sont résumés par Liu *et al.* (2005a) comme un lissage du cycle journalier et saisonnier des températures de l'eau, puisque le réservoir joue un rôle tampon en diminuant les variations positives et négatives et de la température l'eau (Olden & Naiman, 2010). Récemment, Maheu *et al.* (2016) ont trouvé que la présence de barrages sur les rivières de l'est du Canada causait une augmentation dans les températures mensuelles moyennes de septembre. L'augmentation de la température de l'eau en aval des réservoirs est soutenue par les études de Singer and Gangloff (2011) et de Poirel *et al.* (2010), qui portent respectivement sur une rivière régulée en Alabama et sur la simulation des températures et des débits sur la rivière Ain en France. Quant à eux, Preece and Jones (2002) ont trouvé que la température maximale annuelle avait diminué de 5°C et qu'elle se produisait 3 semaines plus tôt après la régulation pour une rivière en Australie. En modifiant le régime thermique, un barrage peut causer chez le saumon atlantique une émergence des alevins différente de celle optimale, et ainsi, diminuer le taux de survie de ceux-ci (Angilletta *et al.*, 2008).

Outre les effets indirects d'une modification de la température de l'eau du cours d'eau, les barrages ont aussi d'autres effets sur le saumon. Par exemple, la présence d'un barrage peut réduire les environnements propices à la fraie du saumon en diminuant les environnements constitués de graviers en aval de ce dernier (Kondolf, 1997). Les séquences seuils-mouilles peuvent aussi être fortement réduites, diminuant la superficie d'habitats utilisables par les saumons juvéniles (Assani & Petit, 2004). Une opération déficiente des barrages peut causer un assèchement des nids, l'échouage et même le piégeage des juvéniles (Harnish *et al.*, 2014). Un barrage peut diminuer ou même supprimer l'accès à des habitats aquatiques potentiels (Kondolf, 1997), ce qui peut affecter ultimement la persistance d'une espèce migratoire comme le saumon atlantique

(Lawrence et al., 2016). Pour ce qui est de leur croissance, Puffer *et al.* (2017) ont montré à l'aide d'un canal expérimental que les débits de pointe d'hydroélectricité n'avaient qu'une faible influence sur celle-ci.

La déforestation : Bien que les barrages influencent sans l'ombre d'un doute la température de l'eau, les forêts ont aussi un rôle important à jouer quant à la régulation thermique des cours d'eau (Beschta *et al.*, 1987). La coupe forestière riveraine influence directement la température de l'eau (Hannah *et al.*, 2008). En effet, lorsque la bande riveraine est coupée et récoltée, le rayonnement solaire incident augmente, ce qui peut causer une hausse de la température de l'eau allant jusqu'à 6-7°C (Moore *et al.*, 2005). Dans le Maine par exemple, l'élimination de la bande riveraine par l'exploitation forestière a créé une augmentation de 1.4°C à 4.4°C de la température maximale hebdomadaire de l'eau (Wilkerson *et al.*, 2006). Une étude de plusieurs cours d'eau au Minnesota a montré une différence de 2.5°C de la température hebdomadaire moyenne entre les cours d'eau ombragés ou non (Blann *et al.*, 2002). En ce qui concerne la température journalière de l'eau, St-Hilaire *et al.* (2000) ont noté qu'une coupe forestière supérieure à 50% pouvait faire augmenter la température journalière moyenne de 1°C, ce qui est assez élevé pour faire migrer les poissons ectothermes vers des refuges thermiques plus froids.

Les autres impacts anthropiques : Les activités anthropiques comme les rejets industriels, urbains ou agricoles, qui contaminent les eaux disponibles pour les saumons, et la mise en place d'obstacles infranchissables comme les ponceaux ou les barrages diminuent la quantité d'habitats potentiels disponibles pour les saumons (Gibson, 1993; Bardonnet & Baglinière, 2000; Nyqvist *et al.*, 2017). Par exemple, la présence de ponceaux amène des sédiments fins dans la rivière et modifie les conditions hydrauliques naturelles de celle-ci, ce qui impacte négativement les habitats lotiques (en rivière) (Prévost *et al.*, 2002). La construction de routes et de ponceaux mal adaptés peut limiter la connectivité des habitats pour les poissons et ainsi les forcer à utiliser des habitats moins favorables et ainsi diminuer leur productivité (Gibson *et al.*, 2005). La construction de route augmente la concentration de métaux lourds dans les cours d'eau alors que l'épandage de sel déglaçant augmente la quantité de chlorure et de sodium dans les rivières, ce qui diminue la qualité de l'eau (Trombulak & Frissell, 2000). Finalement, en

remplaçant le couvert forestier par des routes, le ruissèlement augmente et les sédiments sont amenés plus facilement vers les cours d'eau, ce qui augmente les sédiments en suspension dans l'eau et peut diminuer la productivité des organismes qui vivent dans l'eau ou même les tuer (Newcombe & Jensen, 1996; Wood & Armitage, 1997).

1.3.3 La température de l'eau

Comme la biomasse est fortement corrélée avec la température de l'eau, des modifications au régime thermique amènent leur lot de conséquences sur les organismes aquatiques, du plancton jusqu'au poisson (Hinz & Wiley, 1998). En effet, la température de l'eau est un facteur clé dans le taux de décomposition de la matière organique et dans la concentration en oxygène dissout, ce qui influence la production primaire et la biomasse produite (Nemerow, 1991). D'ailleurs, Lessard and Hayes (2003) ont noté une diminution de la densité d'invertébrés en réponse à une modification du régime thermique en aval des barrages avec réservoir tel que décrit plus haut. Dans certains cas, la perte du signal physiologique environnemental donné par le cycle diurne de la température de l'eau peut impacter les espèces ayant une période de diapause (Lehmkuhl, 1972). Évidemment, ces perturbations sur la biomasse et les invertébrés ébranleront toute la chaîne trophique, dont les populations de poissons.

Étant donné que les poissons sont typiquement des espèces poïkilothermes (c.-à-d. que leur température varie avec celle du milieu), ces derniers sont adaptés au régime thermique de la rivière dans laquelle la population a évolué (Verspoor & Jordan, 1989). Les poissons sont sensibles à la température de l'eau, spécialement lors des périodes de croissance et de reproduction (Handeland *et al.*, 2008). Ils vont choisir les habitats de température optimale (thermorégulation) pour leur métabolisme énergétique et leurs fonctions cardio-respiratoires puisque leurs fonctions biologiques peuvent être fortement modifiées si la température est inadéquate (Bjornn & Reiser, 1991; Farrell, 2002; Breau *et al.*, 2011). Cependant, les poissons peuvent tolérer des températures trop faibles ou trop élevées durant une certaine période de temps, appelée zone de tolérance thermique, où leur survie n'est pas menacée (Wehrly *et al.*, 2007). Lors de cette période, les poissons cessent de s'alimenter et leurs processus s'effectuent en anaérobie, c'est-à-dire sans oxygène (Breau *et al.*, 2011). En été, des températures plus froides à l'aval des barrages

peuvent par exemple favoriser les salmonidés qui préfèrent les eaux fraîches. D'un autre côté, une température plus froide est associée à un plus faible taux de croissance chez les saumons atlantiques juvéniles (Olden & Naiman, 2010). De plus, la fécondité et la longévité du saumon atlantique sont impactées négativement par une hausse des températures en rivière (Jonsson & Jonsson, 2009).

L'œuf de saumon atlantique nécessite une température minimale de 6°C et sa mortalité augmente significativement dans les 12°C (Bley, 1987). La juvénile a une croissance optimale à 16.6°C, alors que la croissance normale se situe entre 15°C et 19°C. Ce dernier peut tolérer des températures allant jusqu'à 27°C avant de rechercher des refuges thermiques (Bley, 1987). Des températures trop basses pour les salmonidés les rendent incapables d'accomplir leurs fonctions biologiques (Sigholt & Finstad, 1990), alors que des températures trop hautes (23°C-25°C) peuvent causer leur mort (Lee & Rinne, 1980; Bjornn & Reiser, 1991). L'expérience de Bley (1987) montre qu'un saumon adapté à 13°C subissait une mortalité de 50% dans les 6 heures s'il était en présence d'eau à 26.7°C. L'article de Daigle *et al.* (2015) donne des bornes maximales de températures pour l'alimentation, le comportement et la survie du saumon atlantique à partir de données colligées de Breau *et al.* (2011), Elliott and Elliott (2010) et Jonsson and Jonsson (2009). Les métriques de températures ont été reproduites au tableau 1.1.

Tableau 1.1 : Températures critiques pour le saumon atlantique *Salmo Salar* (ºC)

Description des métriques	Températures (ºC)
Croissance optimale	16
Début des conditions létales	27.8
Limite supérieure pour l'alimentation	22–28
Limite ultime létale	30–33
Changement de comportement observé	22–24

Vu les changements dans le régime thermique des rivières, certaines espèces devront migrer pour trouver des températures qui leur sont plus adaptées (Parmesan & Yohe, 2003). Par exemple, une augmentation des températures de l'eau dans la rivière Veidness au nord de la Norvège a été associée avec une hausse de la productivité de saumon atlantique entre 1998 et 2010, mais aussi à une baisse, celle de l'omble de fontaine (*Salvelinus fontinalis*), une espèce de truite partageant des habitats parfois similaires à ceux des saumons atlantiques (Svenning *et al.*, 2016). Jonsson and Jonsson (2009) tirent les mêmes conclusions quant à une migration au nord des populations de saumon atlantique, et une diminution, voire une extinction au sud de son aire de répartition. Des simulations pour les années 2071-2100 de Hedger *et al.* (2013) arrivent eux aussi aux mêmes résultats quant à la modification de la répartition géographique du saumon atlantique.

1.4 Modélisation de la température de l'eau en rivière

Comme on l'a vu ci-dessus, la température de l'eau en rivière est une variable clé tant pour le saumon atlantique que les autres organismes vivant dans l'eau. Ainsi, la modélisation de cette variable est d'une grande importante et les modèles se divisent en deux grandes catégories: déterministes et statistiques (Benyahya *et al.*, 2007a).

1.4.1 Modèles déterministes

Dans cette catégorie, des relations mathématiques caractérisent les processus physiques de transfert d'énergie et font le lien entre la température de l'eau (*sortie du modèle*) et les autres variables environnementales (*entrées du modèle*) comme la température de l'air, le couvert forestier, le débit, la radiation solaire, la vitesse du vent, etc. Ces modèles sont basées sur une approche par bilan d'énergie (Morin *et al.*, 1987). Quelques exemples de modèles déterministes sont SNTEMP (Bartholow, 1995), SHADE (Chen *et al.*, 1998) et CEQUEAU (Morin *et al.*, 1981; St-Hilaire *et al.*, 2015). Ces modèles donnent souvent de bonnes estimations de la température de l'eau lorsque l'on veut simuler des changements dans les variables d'entrée (St-Hilaire *et al.*, 2000). Cependant,

ils nécessitent de nombreuses variables d'entrée et sont souvent co teux sur le plan du temps de calcul (Benyahya *et al.*, 2007a).

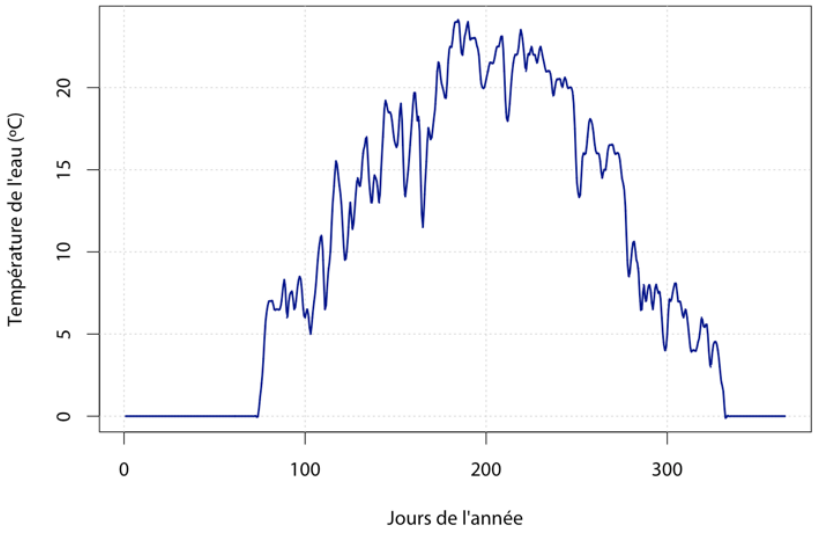
1.4.2 Modèles statistiques

Les modèles statistiques requièrent habituellement moins de variables explicatives et moins de temps de calcul que les approches déterministes, menant à une plus grande simplicité. Les modèles statistiques comprennent les approches non paramétriques comme les réseaux de neurones (Bélanger *et al.*, 2005; Chenard & Caissie, 2008; DeWeber & Wagner, 2014; Hebert *et al.*, 2014; Piotrowski *et al.*, 2015) et l'approche des *k* voisins les plus proches (Benyahya *et al.*, 2008; St-Hilaire *et al.*, 2012). Les modèles non paramétriques sont souvent critiqués pour être des approches de type « boîte noire » et les modèles paramétriques sont donc parfois préférés. Ces derniers sont divisés en deux catégories : les modèles stochastiques et les modèles de régression. La première catégorie inclut les modèles de séries chronologiques comme le processus de Markov de second ordre (Cluis, 1972; Caissie *et al.*, 1998b; Caissie *et al.*, 2001), le modèle autorégressif périodique (Benyahya *et al.*, 2007b) et le modèle autorégressif non linéaire avec variables exogènes (Kwak *et al.*, 2017). Dans les approches de régression, la température de l'eau est modélisée en fonction d'un ou de plusieurs prédicteurs. La relation est souvent présumée linéaire (Crisp & Howson, 1982; Jeppesen & Iversen, 1987; Mackey & Berrie, 1991; Jourdonnais *et al.*, 1992; Stefan & Preud'homme, 1993), mais parfois non linéaire comme le modèle logistique (Mohseni *et al.*, 1998), le processus gaussien (Grbić *et al.*, 2013) et le modèle additif généralisé (Wehrly *et al.*, 2009; Laanaya *et al.*, 2017). Caissie (2006), Benyahya *et al.* (2007a) et Webb *et al.* (2008) ont produit des revues de littérature complètes sur la modélisation statistique de la température de l'eau.

1.4.3 Problématique en modélisation de la température de l'eau

Pour tous ces modèles décrits plus haut et ceux de la littérature, la variable de température de l'eau est traitée de manière ponctuelle telle que mesurée, alors qu'il s'agit plutôt d'un phénomène continu (voir figure 1.1).

Figure 1.1 Exemple de la variation intra-annuelle de la température de l'eau en rivière



La variable de la température de l'eau affiche une forte saisonnalité, ce qui témoigne de l'importance du moment *t* où elle est observée, alors que certains modèles de régression actuellement utilisés ne permettent pas de le prendre en compte (Li *et al.*, 2014). Pour modéliser la température de l'eau, de nombreux modèles sont basés sur la température de l'air. Or, comme la relation entre les températures de l'air et de l'eau est plus dispersée lorsque des moyennes journalières ou hebdomadaires sont utilisées plutôt que des moyennes mensuelles et annuelles, la modélisation des moyennes annuelles ou mensuelles a parfois été préférée dans les modèles de régression (Pilgrim *et al.*, 1998). Cependant, en travaillant avec ces moyennes (agrégation de l'information), une sérieuse perte d'information est causée et rend l'utilisation des résultats peu pratiques d'un point de vue de gestion des rivières. Aussi, on peut parler du problème de « régression fallacieuse » dans le cas où on utilise dans une régression des séries chronologiques non stationnaires (comme c'est le cas avec les séries de températures de l'air et de l'eau), ce qui mène à des résultats plus optimistes que réels (Hoover, 2003). Un remède classique est de retirer la tendance/saisonnalité dans les séries chronologiques comme étape préliminaire avant de modéliser les résidus résultants (Cluis, 1972; Caissie *et al.*, 1998a). Finalement, des modèles basés sur la variable de température de l'air à différents pas de temps, par exemple à *t*-1 et *t*-2 (Kothandaraman, 1971), contiennent de l'autocorrélation (de la corrélation entre les variables explicatives). Cela viole l'hypothèse

d'indépendance des résidus et ainsi, les estimateurs peuvent être biaisés et leur variance sous-estimée (Masselot, 2017).

1.5 Modélisation de l'habitat du saumon atlantique juvénile

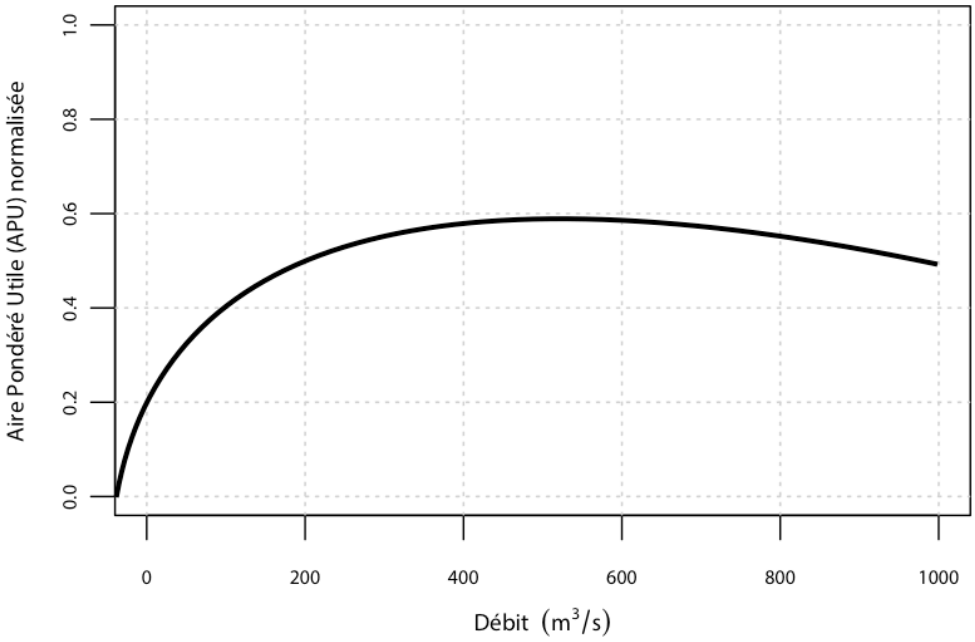
Deux approches sont couramment utilisées dans la littérature pour caractériser l'habitat du saumon atlantique juvénile. La première repose sur la création d'un indice de qualité d'habitat tandis que la seconde repose sur la régression. Ahmadi-Nedushan *et al.* (2006) et Yi *et al.* (2017) donnent des revues de littérature complètes sur la modélisation des habitats aquatiques de poisson.

1.5.1 Les approches basées sur l'indice de qualité d'habitat

L'approche de la définition d'un indice de qualité d'habitat (IQH) à partir de courbes de préférence d'habitat est la technique la plus largement utilisée pour caractériser la qualité de l'habitat (Bovee, 1978; DeGraaf & Bain, 1986; Morantz *et al.*, 1987). Pour chaque variable physique (typiquement la profondeur de l'eau, la vitesse de courant et la taille du substrat), une courbe de préférence est définie comme la proportion de la variable utilisée (pourcentage des poissons qui utilisent cette gamme de valeurs) sur la proportion disponible (pourcentage de la surface représentée par cette gamme de valeurs) pour un tronçon ou pour la rivière en entier (Guay *et al.*, 2000). Une valeur de la courbe de 1 correspond à une utilisation totale de cette gamme de valeurs alors qu'une valeur de 0 correspond à une inutilisation. Une mesure de qualité globale de l'habitat, un IQH, est finalement obtenue par une moyenne arithmétique, géométrique ou géométrique pondérée des valeurs prises par chacune des courbes de préférence dans l'habitat étudié (Ahmadi-Nedushan *et al.*, 2006).

Pour valider les modèles calculant un IQH, la pondération de chaque parcelle de la rivière selon leur superficie et leur IQH donne une mesure appelée l'aire pondérée utile (APU). L'APU est par la suite divisée par la superficie de la rivière et est donc exprimée comme une mesure normalisée entre 0 et 1 représentant la superficie utilisable de la rivière par une espèce de poisson donnée. La figure 1.2 montre un exemple de la variation de l'APU (normalisée par la superficie de la rivière) selon différents débits observés sur une rivière.

Figure 1.2 : Exemple de la variation de l'aire pondérée utile en fonction du débit



La première approche de modélisation d'habitat (Bovee & Milhous, 1978) a mené à la création du système de simulation d'habitat physique (PHABISM pour *physical habitat simulation system*) (Bovee, 1982). Depuis, la modélisation d'habitat n'a que peu évolué et près d'un demi-siècle plus tard, PHABISM est encore largement utilisé aux États-Unis et ailleurs dans le monde même si souvent critiqué (Railsback, 2016). Premièrement, l'utilisation de la métrique de l'APU n'a pas de signification biologique pour quantifier l'habitat (Railsback, 2016). Des métriques plus directes seraient par exemple d'utiliser simplement la densité ou l'abondance de poissons (Manly *et al.*, 2007). Deuxièmement, les hypothèses que les variables d'habitat agissent indépendamment et ont des effets égaux sur la sélection d'habitats (les courbes de préférences prennent des valeurs entre 0 et 1) ont de grandes chances d'introduire des erreurs considérables dans l'estimation de l'IQH (Orth & Maughan, 1982). Troisièmement, les courbes de préférence utilisées dans les modèles comme PHABISM peuvent avoir des effets considérables sur l'estimation résultante (Ayllón *et al.*, 2012). Finalement les mesures au nez du poisson introduisent de fausses valeurs dans l'approche par courbes de préférence, ce qui peut biaiser l'estimation des débits requis pour les poissons avec l'approche PHABISM (Beecher *et al.*, 2010). Par exemple, un poisson observé derrière une roche alors qu'il se

nourrit dans les courants plus rapides voisins introduit une sous-représentation des vitesses élevées dans la courbe de préférence des vitesses. Alors que quelques études ont démontré un lien entre l'APU, obtenue via les modèles basés sur les courbes de préférence, et la densité de poissons (Orth & Maughan, 1982; Boudreau *et al.*, 1996; Bovee *et al.*, 1998), d'autres ont trouvé des résultats opposés (Scott & Shirvell, 1987; Bourgeois *et al.*, 1996). Quant à la transférabilité de ces modèles d'habitat, les résultats sont souvent non concluants ou manquent parfois de rigueur statistique (Scott & Shirvell, 1987; Freeman *et al.*, 1997; Mäki-Petäys *et al.*, 2002; Hedger *et al.*, 2004).

Ainsi, une autre méthode pour définir l'IQH a émergé et est basée sur l'avis d'experts et la logique floue (Ahmadi-Nedushan *et al.*, 2006). Elle a été utilisée dans le cas de la truite brune *Salmon trutta* (Jorde *et al.*, 2001). Cette méthode a obtenu de bons résultats pour le lien entre l'IQH et l'abondance de truites observée, sauf dans les classes de valeurs d'IQH 0 – 0.1 et 0.8 – 0.9. Ahmadi-Nedushan *et al.* (2008) ont utilisé la logique floue pour calculer les débits réservés pour le saumon atlantique. Mocq *et al.* (2013) ont modélisé l'habitat du tacon de saumon atlantique avec la logique floue et ont montré de bons résultats de ce modèle par son application sur la Rivière Romaine au Québec, mais sa transférabilité sur plusieurs rivières demeure encore à prouver. Les limites de l'application des modèles basés sur la logique floue sont que le nombre de règles augmente exponentiellement avec le nombre de variables explicatives et qu'ils nécessitent la codification de l'expertise d'un nombre suffisamment élevé d'experts (Ahmadi-Nedushan *et al.*, 2006).

1.5.2 Les approches de type régression

Les modèles de régression ont aussi émergé associés à la problématique de l'habitat aquatique permettant de faire le lien direct entre les variables de l'habitat et l'abondance/densité/présence-absence de poissons. Ils se veulent plus simples et directs que l'approche par courbe de préférence ou la modélisation floue. On y retrouve la régression multiple, le modèle linéaire généralisé (MLG), le modèle additif généralisé (MAG) et les réseaux de neurones (Ahmadi-Nedushan *et al.*, 2006). En ce qui concerne les modèles utilisés pour l'habitat du saumon atlantique juvénile, on y trouve

premièrement la régression logistique, un cas spécial du MLG avec une fonction de lien *logit* (Legendre & Legendre, 2012). Elle a été utilisée pour prédire la présence-absence de tacon du saumon atlantique juvénile et comparée à l'approche classique de l'IQH, avec de meilleurs résultats pour la régression logistique (Guay *et al.*, 2000). Cette même approche a été utilisée pour caractériser la transférabilité du modèle logistique comparé à l'approche IQH avec des résultats plus prometteurs pour le modèle logistique (Guay *et al.*, 2003). Le MLG n'a été que faiblement utilisé dans la modélisation de l'habitat du poisson (Labonne *et al.*, 2003; Ahmadi-Nedushan *et al.*, 2006). Plus récemment, Beakes *et al.* (2014) l'ont utilisé pour modéliser la présence-absence d'alevins du saumon chinook (*Oncorhynchus tshawytscha*). Comme extension du MLG, le MAG a été utilisé pour la densité d'alevins et de tacons de saumon atlantique (Hedger *et al.*, 2005). Dans Millidine *et al.* (2016), le MAG a aussi été utilisé pour tester la transférabilité d'un modèle d'abondance d'alevins de saumon atlantique à différentes sections de la rivière. Finalement, pour ce qui est des réseaux de neurones appliqués à l'habitat du saumon juvénile, aucune étude n'a fait l'application de cette méthode sur le saumon atlantique juvénile bien qu'elle a été quelque peu utilisée avec d'autres poissons (Olden & Jackson, 2001; Olden & Jackson, 2002a; Ibarra *et al.*, 2003). Le désavantage de travailler avec les réseaux de neurones réside dans l'interprétation des résultats, particulièrement dans la contribution de chaque variable (Ahmadi-Nedushan *et al.*, 2006), bien que certaines études s'adressent spécifiquement à ce problème (Olden & Jackson, 2002b).

1.5.3 Problématique en modélisation de l'habitat du saumon

Parmi les limitations ou inconvénients des modèles cités plus haut, ceux-ci sont majoritairement basés sur la profondeur d'eau, la vitesse du courant et la taille du substrat. La température de l'eau n'est que rarement incluse comme variable d'entrée, alors qu'on sait que les refuges thermiques sont d'une grande importance pour les habitats des saumons (Dugdale *et al.*, 2013). De plus, alors que Railsback (2016) a proposé récemment de développer de nouveaux modèles d'habitat basés sur les données, peu d'efforts ont été faits dans ce sens dans les dernières années. Une critique que nous faisons des modèles vus dans la littérature est que ceux-ci reposent toujours sur une seule valeur des variables physiques pour représenter l'habitat du saumon à un

site ou pour une parcelle. Par exemple, Hedger *et al.* (2005) ont utilisé une moyenne sur 3 à 10 mesures des variables physiques (profondeur, vitesse et substrat) par site pour caractériser l'habitat du saumon et le modéliser à l'aide d'un modèle additif généralisé. En procédant ainsi, une perte importante d'information sur l'habitat naturel du saumon est encourue, ce qui pourrait expliquer les faibles mesures d'adéquation obtenues (R^2 de 28% pour la densité d'alevins et de 47% pour le tacon). Plus encore, Hedger *et al.* (2006) ont montré qu'une mesure moyenne par site du substrat était meilleure qu'une mesure au nez du poisson pour caractériser sa préférence de substrat. Même si l'étude montre que la mesure au site est meilleure que celle au nez du poisson, il n'en demeure pas moins qu'une perte d'information sérieuse est causée sur la description du substrat en ne tenant compte que de la valeur moyenne. À ce jour, l'habitat aquatique a toujours été décrit en fonction de mesures ponctuelles (au nez du poisson ou comme une moyenne à un site), ce qui limite grandement sa description. Cette représentation incomplète de l'habitat par une valeur moyenne pourrait donc expliquer le faible pouvoir explicatif des modèles d'habitat à même la rivière échantillonnée ou sur d'autres rivières.

1.6 **Approche proposée : modélisation par régression fonctionnelle**

Tel qu'il a été vu dans les deux sections précédentes, les variables d'habitat du saumon juvénile ou celles de la température de l'eau et de l'air sont décrites par des valeurs ponctuelles dans les modèles classiques, causant de sérieuses pertes d'information et limitant parfois l'utilisation pratique de ces modèles et leurs résultats. Par exemple, pour la variable de la température de l'eau, l'agrégation sur une certaine période de temps peut rendre les prévisions peu utilisables et l'inclusion de plusieurs variables explicatives temporellement corrélées entre elles introduit de l'autocorrélation et de la dépendance temporelle dans les modèles de régression. Pour le saumon juvénile, les variables ponctuelles d'habitat (au nez du poisson ou comme une moyenne par site) ne semblent pas décrire son habitat adéquatement, vu les faibles résultats des modèles classiques obtenus en termes de prédiction et de transférabilité. Étant donné ces divers problèmes, il semble donc plus naturel de considérer le **cadre fonctionnel** pour traiter ces deux sujets, permettant de traiter ces variables comme des courbes/fonctions. En effet,

l'utilisation de courbes pour représenter ces phénomènes semble plus naturelle et plus adaptée que les approches classiques utilisées dans la littérature basée sur des valeurs scalaires. Cette approche est susceptible de fournir de nouvelles informations quant au lien entre les variables explicatives et la variable réponse, et aussi, d'améliorer le pouvoir prédictif de ces modèles.

Dans ce contexte, il est donc d'intérêt d'introduire le cadre statistique de **l'analyse de données fonctionnelle** (ADF) pour ces deux problématiques. Ce dernier permet de travailler avec des courbes ou fonctions continues $x(s)$ plutôt que des scalaires ou vecteurs x_1, x_2, etc. L'ADF a été introduite par Ramsay (1982) et est devenue très populaire au cours des dernières années avec la publication de nombreux livres (Ferraty & Vieu, 2006; Ramsay, 2006; Dabo-Niang & Ferraty, 2008; Ramsay *et al.*, 2009; Bosq, 2012; Horváth & Kokoszka, 2012) et diverses applications en écologie (Bel *et al.*, 2011; Stewart-Koster *et al.*, 2014; McDonald *et al.*, 2015), transport (Chiou, 2012), énergie (Goia *et al.*, 2010; Chaouch, 2014; Brockhaus *et al.*, 2015), finance (Wang *et al.*, 2008), gestion des déchets (Bernardi *et al.*, 2017), biotechnologie (Brockhaus *et al.*, 2017a), médecine (Ciarleglio *et al.*, 2016) et neuroscience (McLean *et al.*, 2014; Ivanescu *et al.*, 2015; Meyer *et al.*, 2015). Le développement de nouvelles librairies facilitant l'utilisation de l'ADF, notamment en R (R Core Team, 2017), a grandement contribué à sa popularité : *fda* (Ramsay *et al.*, 2013), *far* (Damon & Guillas, 2015)*, refund* (Goldsmith *et al.*, 2016), *fdANOVA* (Gorecki & Smaga, 2017) et *FDboost* (Brockhaus et al., 2017b). Chebana *et al.* (2012) ont été les premiers à introduire l'ADF en hydrologie, en voyant la série annuelle des débits, l'hydrogramme, comme une courbe c.-à-d. une observation fonctionnelle. Depuis, plusieurs travaux ont porté sur l'utilisation de l'ADF en hydrologie : classification des hydrogrammes selon leur forme (Ternynck *et al.*, 2016), modélisation de l'hydrogramme à partir des précipitations (Masselot *et al.*, 2016), étude de la variabilité temporelle et spatiale des précipitations (Suhaila & Yusop, 2017), calibration plus réaliste des modèles hydrologiques (Larabi *et al.*, 2017), estimation de la courbe des débits classés (Requena *et al.*, 2018). Bien que de nombreuses études aient porté sur l'application du cadre fonctionnel sur des données de débits et de précipitations, l'utilisation de l'ADF en hydrologie peut aller bien au-delà de ces deux variables.

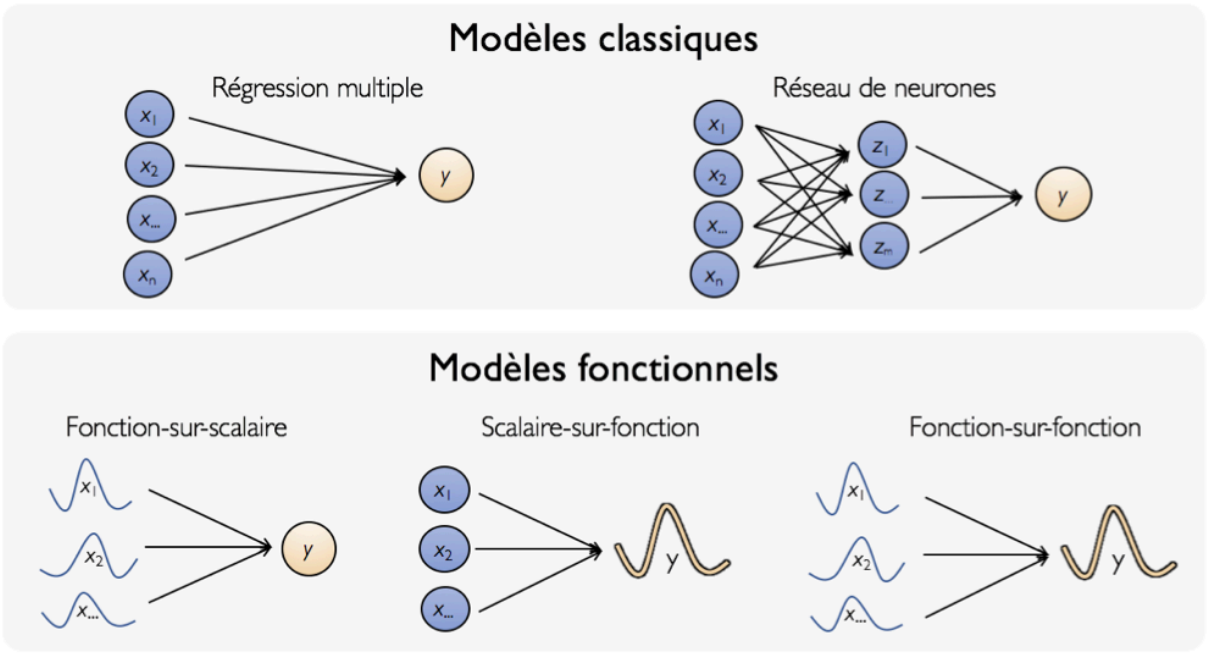
Ainsi, dans le contexte de la modélisation, on s'intéressera particulièrement aux **modèles de régression fonctionnelle** (MRF) permettant de faire le lien entre des variables qui peuvent être sous forme de courbes ou non. Ces modèles seront introduits plus en détail dans les sections suivantes du mémoire selon le type de modèle nécessaire à la réalisation de chacun des objectifs spécifiques de recherche. Notons cependant que leur but est de permettre que les variables explicatives ($x_1, x_2, \ldots$) et/ou réponse (y) soient des courbes/fonctions plutôt que des valeurs scalaires, donnant lieu à trois types de modèles de régression fonctionnelle (Ramsay, 2006) :

1. Fonction-sur-scalaire : Au moins une des variables explicatives est une courbe/fonction et la variable réponse est un scalaire ;
2. Scalaire-sur-fonction : Les variables explicatives sont toutes des scalaires et la variable réponse est une courbe/fonction ;
3. Fonction-sur-fonction : Au moins une des variables explicatives est une courbe/fonction et la variable réponse est aussi une courbe/fonction.

Le cas scalaire-sur-scalaire représente la régression classique non fonctionnelle.

La figure 1.3 illustre la comparaison entre les modèles de régression dits « classiques » (non-fonctionnels) dans le présent mémoire et les différents types de modèles fonctionnels.

Figure 1.3 : Illustration des modèles classiques et des modèles fonctionnels



1.7 Objectifs spécifiques

Pour les raisons évoquées précédemment (notamment dans les problématiques), les MRF semblent bien adaptés aux deux variables d'intérêt de ce mémoire (qui sont la température de l'eau et l'habitat du saumon juvénile). Ces modèles sont novateurs par leur capacité à reproduire les phénomènes plus naturellement que les modèles classiques, en plus d'être faciles à interpréter et de fournir plus d'information. Finalement, le lien entre la variable de température de l'eau et l'habitat saumon atlantique étant plus qu'évident, l'étude dans ce mémoire de ces deux sujets est plus que pertinente. Les objectifs spécifiques se déclineront ainsi en deux parties : une première touchant la modélisation de la température de l'eau et une seconde touchant l'habitat du saumon atlantique juvénile.

1.7.1 Température de l'eau en rivière

Dans le cas de la température de l'eau, comme c'était pour le débit qui a été beaucoup étudié avec l'ADF, la série annuelle ou pour une saison des températures de l'eau peut

être vue comme une courbe, et ce, de manière encore plus évidente que le débit étant donné sa variation saisonnière très prononcée (revoir la figure 1.1). Cette courbe peut ensuite être modélisée à l'aide d'un MRF par des variables explicatives, qui sont aussi fonctionnelles, comme la température de l'air, le débit, la précipitation, le vent, etc. Un des avantages des MRF est de permettre la modélisation directe de la *courbe* comme une seule observation, plutôt que de multiples valeurs modélisées dans le cas des autres modèles classiques. De plus, les MRF permettent l'utilisation des *courbes* complètes des variables explicatives comme entrées plutôt que plusieurs observations à différents pas de temps dans la régression classique (p. ex. la température de l'air à *t*-1 et à *t*-2), pouvant causer de l'autocorrélation, ce qui est évité ici (Cuevas *et al.*, 2002). Aussi, les problèmes de régression fallacieuse sont aussi corrigés en modélisant directement les séries chronologiques comme des *courbes*.

Comme il s'agit d'une première tentative de modéliser la *courbe* de la température de l'eau avec la régression fonctionnelle, une seule variable explicative sera utilisée, à savoir la *courbe* des températures de l'air. Deux modèles fonctionnels seront utilisés : le modèle fonctionnel linéaire complet (MFLC), précédemment utilisé par Masselot *et al.* (2016) pour modéliser la *courbe* des débits en fonction de celle des précipitations (Masselot *et al.*, 2016) et le modèle fonctionnel linéaire historique (MFLH). Le MFLH, qui sera introduit plus en détail au chapitre 2, a comme avantage de pouvoir restreindre le domaine utilisé de la *courbe* de la variable explicative pour modéliser la *courbe* de la variable réponse (par exemple en n'utilisant que les valeurs historiques). Ce dernier a récemment été mis en place en R dans le package *FDboost*, facilitant son utilisation pratique (Brockhaus *et al.*, 2017a). Ces deux modèles fonctionnels seront comparés à deux modèles tirés de la littérature : le modèle logistique (ML) (Mohseni *et al.*, 1998), le plus largement utilisé dans la littérature, et le modèle additif généralisé (MAG) (Laanaya *et al.*, 2017), ayant montré les meilleurs résultats récemment pour la modélisation des températures de l'eau journalières en fonction de la température de l'air et du débit.

Ainsi, les objectifs spécifiques de ce premier projet sont :

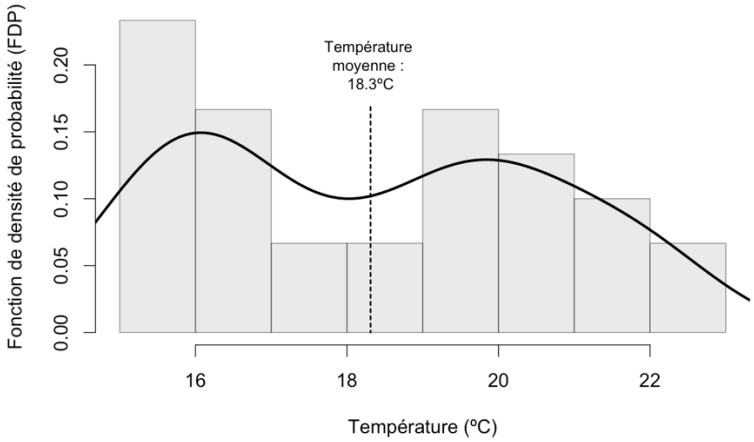
1. Considérer la régression fonctionnelle pour modéliser la *courbe* des températures de l'eau en fonction de la *courbe* des températures de l'air;

2. Évaluer deux modèles de régression fonctionnelle, le modèle fonctionnel linéaire complet (MFLC) et celui historique (MFLH);
3. Comparer ces deux modèles avec deux modèles non-fonctionnels tirés de la littérature : le modèle logistique (ML) et le modèle additif généralisé (MAG).

1.7.2 Habitat du saumon atlantique juvénile

En ce qui concerne l'habitat du poisson, l'utilisation des MRF dans l'habitat du poisson permettra d'employer la distribution complète de chacune des variables d'habitat du poisson dans le processus de modélisation. En effet, les variables d'entrée seront plutôt des histogrammes lissés, c'est-à-dire des fonctions de densité de probabilité (FDP) représentant la disponibilité de ces variables dans l'environnement du poisson plutôt que de grossières moyennes peu représentatives et causant de sérieuses pertes d'information. Ainsi, le fait qu'un poisson a différents besoins sera totalement pris en compte dans le MRF (p. ex. il a besoin de vitesses faibles pour se réfugier, mais aussi de vitesses plus fortes pour se nourrir) comparé aux approches classiques basées sur des moyennes. Il en sera de même pour l'apport de refuges thermiques par les tributaires, dont l'effet refroidissant sur le cours d'eau serait masqué par l'utilisation d'une moyenne. La figure 1.4 illustre la distribution (FDP) de la variable de la température de l'eau dans un habitat influencé par un tributaire froid. On y voit comment la FDP représente mieux la variabilité des températures de l'habitat comparée à la moyenne.

Figure 1.4 : Exemple des températures de l'eau mesurées pour une section de rivière



L'idée de considérer les variables d'habitat comme des *courbes* (des FDP) via la régression fonctionnelle est nouvelle, et non seulement en modélisation de l'habitat aquatique, mais aussi en statistique. En effet, dans le domaine de l'ADF, les courbes sont souvent vues comme des fonctions variant dans le temps ou dans l'espace. Ici, l'idée proposée de considérer les FDP comme des courbes pouvant être utilisées dans un MRF peut être vue comme une innovation en soit. Ainsi, comme il s'agit d'une première utilisation des MRF avec des FDP, un modèle simple de régression fonctionnelle sera utilisé pour faire le lien entre la présence-absence et/ou l'abondance de poissons et les FDP : le modèle fonctionnel linéaire pour réponse scalaire (MFLS). Deux autres modèles non-fonctionnels précédemment utilisés pour modéliser l'habitat du saumon juvénile seront comparés. Le premier est le modèle linéaire généralisé (MLG) (Guay *et al.*, 2000; Guay *et al.*, 2003; Beakes *et al.*, 2014) et le second est le modèle additif généralisé (MAG) (Hedger *et al.*, 2005; Millidine *et al.*, 2016), ce dernier permettant une plus grande flexibilité que le MLG quant à la forme de la relation entre les variables explicatives et celle réponse. Finalement, la variable de la température de l'eau sera mesurée sur le terrain et, pour une des premières fois, ajoutée aux modèles d'habitat aquatique. La valeur ajoutée de cette variable sera aussi quantifiée.

Les objectifs spécifiques de ce second projet sont donc :

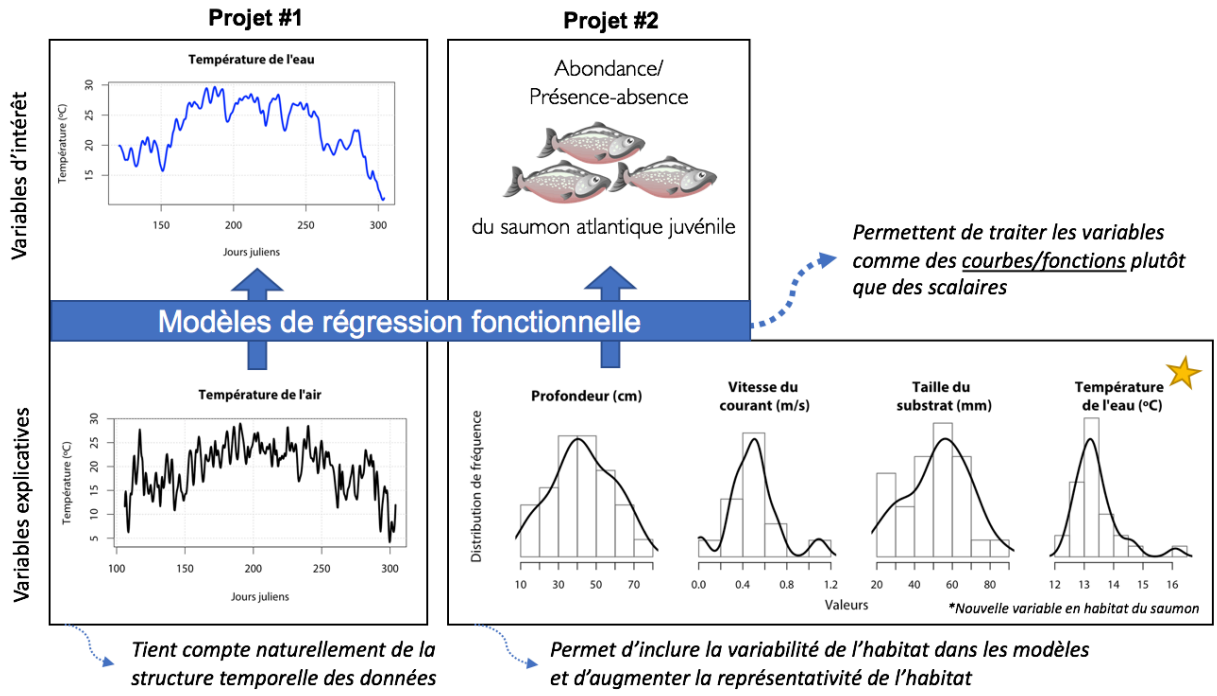
1. Considérer les variables d'habitat du saumon atlantique juvénile comme des *courbes* (des FDP) via un modèle de régression fonctionnel (MFLS);
2. Comparer le modèle fonctionnel à un modèle linéaire généralisé (MLG) et un modèle additif généralisé (MAG);
3. Ajouter la température de l'eau comme variable d'entrée aux modèles d'habitat du saumon atlantique et quantifier la valeur ajoutée de cette variable.

1.8 Présentation du mémoire

Les travaux de recherche de cette maîtrise sont présentés sous la forme d'un mémoire par articles. Après ce premier chapitre d'introduction, le chapitre 2 présente la synthèse des travaux de recherche de l'application des modèles de régression fonctionnelle aux

deux problématiques. Les deux chapitres suivants présentent les deux articles dans leur forme intégrale tels que soumis (ou prêt à soumettre) pour publication : le chapitre 3 présente l'article touchant la température de l'eau en rivière alors que le chapitre 4 contient l'article portant sur l'habitat du saumon atlantique juvénile. Finalement, le chapitre 5 contient une conclusion générale ainsi que quelques recommandations et avenues à explorer. La figure 1.5 résume les principaux travaux de ce mémoire.

Figure 1.5 : Synthèse des travaux du présent mémoire



Les deux projets ont été réalisés sous la direction du professeur Fateh Chebana et la codirection des professeurs André St-Hilaire et Normand Bergeron. Le rôle de l'étudiant a été la collecte des données (via Internet et sur le terrain), les analyses statistiques, le développement de code en R ainsi que la rédaction du mémoire et des articles en tant que premier auteur. L'étudiant a planifié, participé et supervisé la collecte des données sur le terrain lors de l'été 2017. Le travail de terrain s'est fait avec l'aide des personnes suivantes : André Boivin, Marc-André Pouliot, Killian Dolais, Michael Deetjens, Andrée-Sylvie Carbonneau et Antoine Boudry. Les trois professeurs supervisant les travaux de maîtrise ont donné leur soutien tout au long du projet, de la collecte, l'analyse et le

traitement des données jusqu'à la rédaction des articles en émettant tout au long du processus des idées, commentaires, suggestions et corrections.

2 SYNTHÈSE DES TRAVAUX DE RECHERCHE

Les résultats sont divisés en deux parties : une première touchant la température de l'eau et une seconde sur l'habitat du saumon atlantique juvénile.

2.1 Modélisation fonctionnelle de la température de l'eau en rivière

Tel que mentionné précédemment, la modélisation fonctionnelle sera utilisée pour l'estimation de la *courbe* de la variation intra-annuelle des températures de l'eau (appelée dans les sections suivantes « *courbe* des températures de l'eau ») en fonction de celle des températures de l'air. La méthodologie, les résultats obtenus et la discussion sont synthétisés dans les paragraphes qui suivent. Les détails complets ainsi que tous les tableaux/figures/équations sont disponibles dans l'article complet reproduit au chapitre 3.

2.1.1 Méthodologie et cas d'étude

Pour modéliser la *courbe* des températures de l'eau en fonction de celle des températures de l'air, deux modèles fonctionnels sont étudiés : le modèle fonctionnel linéaire complet (MFLC, équation 3.3) et historique (MFLH, équation 3.4). Le MFLC fait l'usage de toutes les valeurs de la *courbe* de la variable explicative (ici la température de l'air) pour prédire la *courbe* de la variable réponse, alors que le MFLH n'utilise que les valeurs historiques de la courbe. Autrement dit, lorsque le MFLH prédit la valeur de la courbe des températures de l'eau à t, seulement les valeurs avant t de la courbe des températures de l'air sont utilisées, alors que dans le MFLC, toutes les valeurs sont utilisées, qu'elles soient observées avant ou après t. Dans Masselot *et al.* (2016), le MFLC était utilisé pour modéliser la *courbe* des débits en fonction de celle des précipitations, ce qui faisait en sorte que les précipitations qui survenaient après t étaient utilisées pour prédire le débit à t, ce qui n'a pas de sens d'un point de vue météorologique. Dans notre étude, nous avons aussi utilisé le modèle historique MFLH qui permet de n'utiliser que les informations de la *courbe* explicative avant t pour prédire la valeur de la *courbe* réponse à t. Cela a plus de sens puisque les températures de l'air futures (après t) ne peuvent influencer la température de l'eau aujourd'hui (à t) d'un point de vue

météorologique. De nouveaux développements informatiques (Brockhaus *et al.*, 2015; Brockhaus *et al.*, 2017a) permettent désormais l'application en R du modèle MFLH, ce qui n'était pas possible lors de l'étude de Masselot *et al.* (2016).

À des fins de comparaison, deux modèles classiques basés seulement sur les températures journalières de l'air et de l'eau sont comparés aux modèles fonctionnels. Le premier est le modèle logistique (ML, équation 3.8) (Mohseni *et al.*, 1998), permettant une relation non linéaire en forme de S de la température de l'air sur la température de l'eau (à ne pas confondre avec la régression logistique). Le ML a souvent été utilisé sur des données hebdomadaires, mais récemment utilisé avec des températures journalières (Kelleher *et al.*, 2012; St-Hilaire *et al.*, 2012; Grbić *et al.*, 2013; Bustillo *et al.*, 2014; Laanaya *et al.*, 2017), comme ce sera fait dans la présente étude. Les quatre paramètres du modèle sont calibrés pour les parties montante (réchauffement) et descendante (refroidissement) des données pour tenir compte de l'hystérèse dans les données (c'est-à-dire que le lien entre la température de l'air et de l'eau diffère lors du réchauffement au printemps et du refroidissement en automne) (Mohseni *et al.*, 1998). Le second modèle est le modèle additif généralisé (MAG, équation 3.6), permettant une plus grande flexibilité quant à la forme de la relation entre la variable explicative et la variable réponse. Le MAG a montré les meilleurs résultats récemment pour modéliser la température journalière de l'eau en fonction de la température de l'air et du débit, comparé à trois autres modèles statistiques (Laanaya *et al.*, 2017).

Pour valider l'approche de la régression fonctionnelle avec la température de l'eau, les données provenant de trois cours d'eau avec plusieurs années disponibles (~25 années) des États-Unis ont été utilisées : les rivières Potomac, Missouri et Delaware (USGS, 2017). La figure 3.1 montre l'emplacement de chacune des rivières alors que le tableau 3.1 énonce les caractéristiques des diverses rivières à leur station de mesure. Les températures de l'air de la station météorologique la plus rapprochée ont été utilisées (NOAA, 2017). Le choix des stations est reporté au tableau 3.2. Les quatre modèles (MFLC, MFLH, ML et MAG) sont calibrés pour toutes les n-1 années disponibles, où n est le nombre d'années totales disponibles pour chaque rivière et j est l'année retirée. Par la suite, ces modèles sont utilisés pour prédire, à partir des températures de l'air, les

températures de l'eau de l'année *j* qui a été retirée de la calibration. Cette méthode de validation croisée est nommée « Leave-one-out » (Quenouille, 1949) et a été souvent utilisée en modélisation de la température de l'eau pour tester l'adéquation du modèle à de nouvelles données (Benyahya *et al.*, 2007b; Benyahya *et al.*, 2008; Laanaya *et al.*, 2017)

Par la suite, trois critères de performance sont calculés à partir des estimations obtenues: la racine de l'erreur quadratique moyenne (RMSE, équation 3.8), le biais (équation 3.9) et le coefficient de Nash-Sutcliffe (NSC, équation 3.10). Pour les modèles fonctionnels, comme des courbes sont prédites plutôt que des valeurs journalières, des mesures d'adéquation des modèles fonctionnels sont aussi calculées, comme le R^2 fonctionnel (équation 3.11) et les homologues fonctionnels des trois critères discutés plus haut : RMSE, biais et NSC fonctionnels (équations 3.12, 3.13 et 3.14 respectivement).

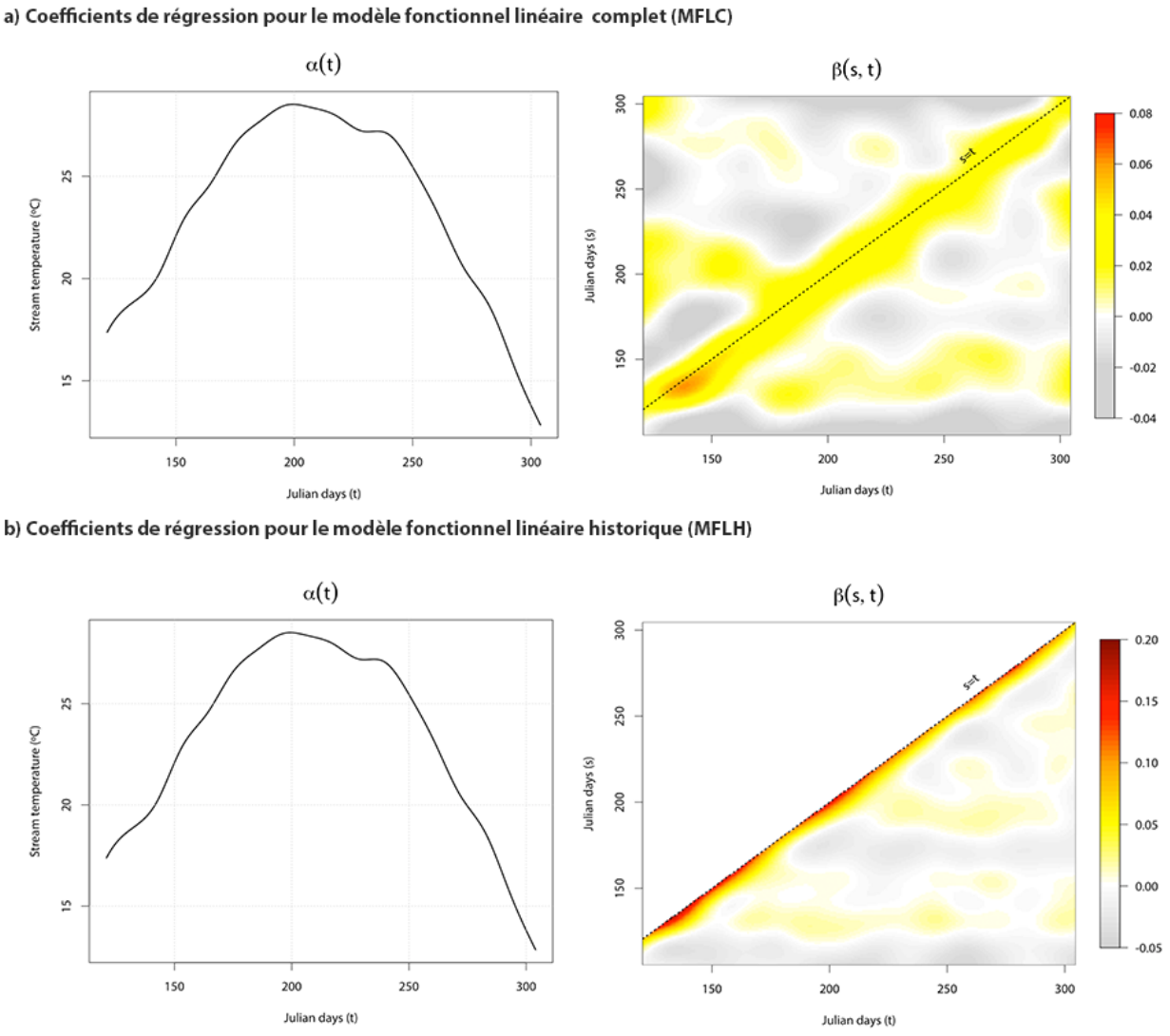
2.1.2 Résultats

Premièrement, les coefficients des modèles fonctionnels sont calculés et illustrés pour l'exemple de la rivière Potomac lorsque la période d'estimation exclut l'année 2015. La figure 2.1 illustre les surfaces de régression $\beta(s, t)$ pour les modèles fonctionnels MFLC (a) et MFLH (b). L'effet de la température de l'air sur la température de l'eau est très prononcé dans les 15 jours suivant/précédant la valeur prédite par le MFLC, alors qu'elle se concentre dans les 10 jours précédant la valeur prédite pour le MFLH. Les autres effets sont près de 0 ailleurs. Les coefficients du MAG sont représentés à la figure 3.4, montrant l'effet non linéaire de la température de l'air sur la température de l'eau particulièrement pour les valeurs élevées. Pour le ML, l'équation obtenue pour la partie montante et pour la partie descendante des données est retranscrite à l'équation 3.15.

Par la suite, les quatre modèles sont comparés selon leur capacité à prédire les moyennes journalières de la température de l'eau comme celles-ci sont d'un certain intérêt pour les gestionnaires des rivières (comparées à des moyennes hebdomadaires ou mensuelles par exemple). Des exemples de températures de l'eau prédites selon la méthode « Leave-one-out » sont illustrés à la figure 3.6 pour trois années de la rivière Potomac. Alors que les deux modèles fonctionnels prédisent directement une courbe

continue des températures de l'eau, les modèles classiques (MAG et LM) ont été utilisés pour faire plusieurs prédictions des températures journalières de l'eau, ce qui ne reproduit pas le comportement continu naturel de la température de l'eau.

Figure 2.1 : Surfaces de régression pour les modèles fonctionnels pour la rivière Potomac



Surfaces de régression pour a) le modèle fonctionnel linéaire complet (MFLC) et b) le modèle fonctionnel linéaire historique (MFLH). L'échelle de couleur est la même pour les deux graphiques.

Pour ce qui est des mesures de performance, les résultats complets sont disponibles au tableau 3.3 pour les températures journalières et au tableau 3.4 pour les critères fonctionnels. Pour les rivières Potomac et Missouri, le MFLH obtient les plus faibles

RMSE et biais, et les plus forts NSC. Le MFLH est suivi par le MAG, puis le MFLC et finalement par le ML. Pour ce qui est de la rivière Delaware, elle obtient les meilleurs résultats avec le modèle MAG, suivi par le MFLH, le MFLC et par le ML. En comparant les modèles fonctionnels entre eux, les critères fonctionnels du R^2, RMSE, biais et NSC indiquent que le modèle MFLH est le meilleur. En affichant graphiquement le R^2 *fonctionnel* (figure 3.8), les résultats demeurent en faveur du MFLH et ce, pour les trois rivières.

2.1.3 Discussion

Les résultats des modèles obtenus sur la base de la validation croisée « Leave-one-out » se comparent bien avec ceux de la littérature. Par exemple, le modèle MFLH obtient des RMSE de 1.62°C et 1.36°C pour les rivières Potomac et Missouri, ce qui est comparable avec le MAG de Laanaya *et al.* (2017) appliqué à la rivière Sainte-Marguerite (RMSE de 1.36°C). En se comparant à un modèle non paramétrique, le modèle des *k*-voisins les plus proches de St-Hilaire *et al.* (2012) a obtenu un RMSE de 1.57°C sur la rivière Moisie. Cela étant dit, le modèle fonctionnel historique (MFLH) obtient des résultats comparables aux modèles de la littérature et n'est basé que sur la température de l'air comme variable d'entrée versus la température de l'air et le débit dans Laanaya *et al.* (2017) et à la température de l'eau à différents pas de temps dans St-Hilaire *et al.* (2012).

Pour ce qui est de la rivière Delaware, le MAG est le modèle qui a obtenu les meilleurs résultats. Cependant, il faut noter que pour cette troisième rivière, les résultats du RMSE sont supérieurs à 2°C pour tous les modèles utilisés, ce qui est un indicateur d'une mauvaise estimation de la température de l'eau par la température de l'air. En effet, la littérature citée précédemment ainsi que les résultats obtenus sur les deux autres rivières étudiées montrent des RMSE avoisinants les 1.5°C, alors que dans le cas de la rivière Delaware, le 2°C d'erreur d'estimation semble trop élevé. Ainsi d'autres variables explicatives devraient s'ajouter à la modélisation de la température de l'eau pour cette rivière comme le débit ou les précipitations.

Malgré la bonne adéquation du modèle MFLH comparé à ceux de la littérature et aux autres modèles évalués dans la présente étude, le modèle MFLC a obtenu de moins

bons résultats (il demeure meilleur que le LM, mais légèrement moins performant que le MAG et le MFLH pour les trois rivières étudiées). Cela s'explique par le fait que le modèle MFLC a une surface $\beta(s, t)$ beaucoup plus grande à estimer (le double) que celle du modèle MFLH (Revoir la figure 2.1). En effet, la surface $\beta(s, t)$ n'est définie que pour la région $s < t$ dans le modèle MFLH, ce qui facilite l'estimation de la surface avec un petit nombre de données (~ 25 années = 25 observations fonctionnelles). Le fait de limiter à priori la surface $\beta(s, t)$ du modèle augmente grandement le pouvoir prédictif et diminue les erreurs dans le MFLH, ce qui se solde en un modèle plus parcimonieux. De plus, le modèle MFLH a une meilleure signification du point de vue climatique que le MFLC, ce dernier considérant que la température de l'air après t peut influencer la température de l'eau à t. Outre les performances, les coefficients des modèles fonctionnels nous indiquent les régions où la température de l'air a un impact sur la température de l'eau, ce qui nous permettrait de simplifier encore davantage le modèle pour ne garder que les périodes d'intérêts (où la surface est différente de 0). Comme l'étude était avant tout exploratoire, toute l'information a été utilisée dans un premier temps (MFLC) puis seulement l'information passée (MFLH). À la lueur des résultats obtenus, les quelques 10 à 15 jours passés (revoir la figure 2.1 b) seraient suffisants pour l'estimation de la température de l'eau, ce qui pourrait rendre le modèle fonctionnel encore plus parcimonieux.

Un frein majeur à l'utilisation des modèles fonctionnels est la nécessité d'avoir un nombre suffisamment grand de données de températures pour ajuster un modèle fonctionnel qui s'adaptera bien aux données. Par exemple, 25 années étaient insuffisantes pour le modèle MFLC alors que c'était suffisant pour le MFLH qui a obtenu des résultats supérieurs aux autres modèles considérés sur deux des trois rivières étudiées. Néanmoins, avec les efforts déployés pour la collecte de données en rivière (p. ex. le réseau RivTemp.ca), nous sommes d'avis que les modèles fonctionnels pourront être déployés sur davantage de rivières et ultimement utilisés dans un mode opérationnel. Pour ce faire, une prévision initiale des températures saisonnières de l'eau pourrait être fournie aux gestionnaires des rivières à partir de la climatologie passée ou d'un modèle climatique. Cette prévision pourrait être ajustée au fur et à mesure que la température de l'air est observée et enregistrée, ce qui viendrait préciser la prévision future. Il serait utile

pour les gestionnaires de rivières d'avoir une idée générale du comportement thermique de la rivière avant la saison à venir, pour pouvoir mieux cibler les périodes de stress thermiques pour les poissons, puis d'avoir des prévisions automatiquement mises à jour selon la température de l'air observée chaque jour, par exemple.

2.2 Modélisation fonctionnelle de l'habitat du saumon

Comme il a été vu précédemment, la modélisation fonctionnelle permet de voir les variables d'habitat du saumon atlantique juvénile comme des histogrammes lissés représentés par des fonctions de densité de probabilité (FDP). Ces prédicteurs, sous forme de *courbes*, représentent plus adéquatement et naturellement l'habitat du saumon juvénile que le feraient des valeurs moyennes. Les sections qui suivent résument la méthodologie utilisée, les résultats obtenus ainsi que la discussion alors que la chapitre 4 présente les travaux complets ainsi que tous les tableaux/figures/équations.

2.2.1 Méthodologie et cas d'étude

Pour valider l'approche proposée, le modèle fonctionnel linéaire pour réponse scalaire (MFLS, équation 4.4) est considéré, permettant d'expliquer une variable réponse scalaire avec des fonctions / courbes qui seront des FDP dans le présent. Les FDP représentent mieux l'habitat du poisson que des moyennes utilisées dans certaines approches classiques et seront définies à l'aide d'une estimation de la densité par noyau (EDN, équation 4.2) calculée à l'aide de R (fonction *density*). Les courbes seront inspectées visuellement pour s'assurer d'une bonne adéquation avec les histogrammes et s'assurer que l'EDN capture bien la variabilité dans les variables d'habitat. Pour sa part, la variable réponse peut être soit l'abondance de poissons relevée ou une variable de présence-absence prenant les valeurs 0 ou 1.

Outre le MFLS, deux autres modèles classiques de régression utilisés précédemment en habitat du saumon juvénile seront utilisés. Le premier est le modèle linéaire généralisé (MLG, équation 4.7) (Guay *et al.*, 2000; Guay *et al.*, 2003; Beakes *et al.*, 2014), qui a montré de meilleurs résultats que l'approche du calcul de l'indice de qualité d'habitat (IQH) via des courbes de préférence (Guay *et al.*, 2000; Guay *et al.*, 2003). Le second

modèle est le modèle additif généralisé (MAG, équation 4.8) (Hedger *et al.*, 2005; Millidine *et al.*, 2016). Ce dernier, comparé au MLG, n'impose pas une relation linéaire entre les prédicteurs et la variable réponse. En effet, une fonction de transformation appropriée $f_j(.)$ est estimée pour chaque prédicteur *j*, ce qui pourrait être mieux adapté que de supposer un effet linéaire, même si aucune étude n'a comparé spécifiquement les modèles MLG et MAG en habitat aquatique comme nous le ferons. Des fonctions *g*(.) de lien *logit* (équation 4.5) pour la présence-absence ou de lien *logarithme* (équation 4.6) pour l'abondance seront utilisées (Guay *et al.*, 2000; Feyrer *et al.*, 2007; Millidine *et al.*, 2016) pour lier la variable réponse (présence-absence ou abondance) aux prédicteurs dans les trois modèles considérés.

Pour valider ces modèles, le coefficient de Nash-Sutcliffe (*NSC*, équation 4.9) pour les modèles d'abondance sera calculé. Lorsque ce coefficient est calculé avec les mêmes données de calibration, il est l'équivalent du traditionnel R^2. Pour les modèles de présence-absence (données binaires), le critère du *PseudoR*^{2} (équation 4.10) sera plutôt évalué. Des mesures de performance seront aussi calculées pour ces modèles de présence-absence (Beakes *et al.*, 2014) : l'exactitude (*EXA*, équation 4.13), le taux de présences correctement prédites (*TPP*, équation 4.14) et le taux d'absences correctement prédites (*TAP*, équation 4.15). La valeur seuil pour la conversion des probabilités prédites par chacun des modèles en variables dichotomiques 0-1 est choisie pour maximiser les bonnes classifications en validation croisée (Liu *et al.*, 2005b). Aussi, l'aire sous la courbe ROC (*Receiver Operating Characteristic*) sera calculée comme mesure de performance considérant toutes les valeurs seuils possibles. Finalement, un critère R^2 *ajusté* (équation 4.12) sera calculé pour faire la comparaison des modèles avec et sans la variable de la température de l'eau et tenir compte de la parcimonie des modèles.

Dans le but d'obtenir un jeu de données adapté à la modélisation à effectuer, deux rivières à saumon du Québec (Canada) ont été échantillonnées au cours de l'été 2017 : les rivières Sainte-Marguerite (RSM) et Petite-Cascapédia (RPC) telles qu'illustrées à la carte de la figure 4.2. Plusieurs sites ont été échantillonnés sur ces deux rivières pour couvrir une grande gamme d'habitats du saumon atlantique juvénile. Qui plus est, le choix des rivières a été motivé par diverses études passées sur leur structure en liens

sédimentaires (Davey, 2005; Davey & Lapointe, 2007; Bouchard & Boisclair, 2008; Kim, 2009; Johnston & Bergeron, 2010; Kim & Lapointe, 2011; Lanthier *et al.*, 2014). Un lien sédimentaire un segment longitudinal d'une rivière caractérisé par la présence de sédiments de grandes tailles en amont, un affinement granulométrique et une diminution de la pente du cours d'eau (Rice & Church, 1998; Rice *et al.*, 2001). Cette structure semble donc être idéale pour la récolte de données couvrant une grande variabilité des habitats disponibles pour les saumons. À chaque site, 30 parcelles de $4m^2$ ont été étudiées (voir la figure 4.3 pour la disposition des 30 parcelles). À chacune de ces parcelles, la pêche électrique a été effectuée (récolte et mesure des saumons atlantiques juvéniles) et les mesures d'habitat ont été prises : vitesse du courant, profondeur et température de l'eau et diamètre moyen du substrat (axe *b* de la roche médiane de la parcelle). La RSM a été utilisée comme rivière de calibration de par sa plus grande variabilité des sites échantillonnés alors que la RPC a été utilisée comme rivière de validation.

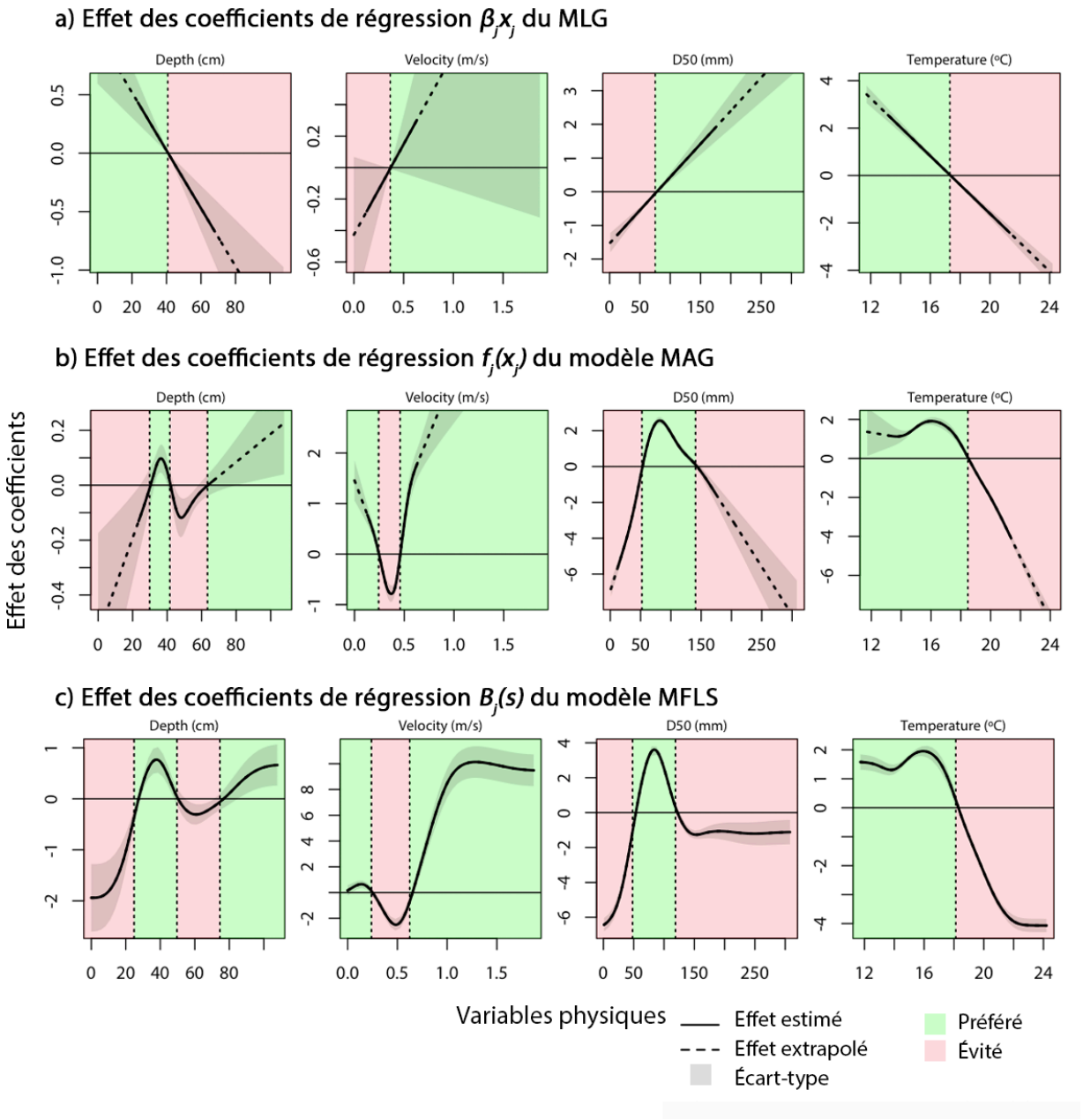
Selon les abondances relevées sur le terrain, des modèles d'abondance ou de présence-absence ont été développés pour les trois âges de saumons atlantiques juvéniles (0+, 1+ et 2+). Comme aucune absence n'a été trouvée pour les 0+ et 1+ au niveau des sites échantillonnés, des modèles d'abondance ont été développés. Par contre, un grand nombre de sites avec absence de 2+ ont été mesurés (50% de sites avec absence), ce qui a motivé la modélisation de la présence-absence plutôt que de l'abondance directement (Feyrer *et al.*, 2007). Le tableau 4.2 résume quelle variable a été modélisée pour chaque âge ainsi que la justification.

2.2.2 Résultats

À des fins d'illustration, les résultats du modèle de présence-absence des 2+ sont d'abord présentés, suivis de ceux des modèles d'abondances de 0+ et 1+. Finalement, la valeur ajoutée de la variable de la température de l'eau est quantifiée.

Modèle de présence-absence du 2+ : La figure 2.2 montre l'effet de chaque prédicteur sur la probabilité d'observer un saumon juvénile 2+ à chacun des sites.

Figure 2.2 : Effet des variables sur la sélection d'habitat du 2+ pour les trois modèles



Dans le cas du MLG (a), le modèle impose un effet linéaire alors que dans le cas du MAG (b) et du MFLS (c), les effets des prédicteurs sont non linéaires. Le MLG semble inadapté pour identifier les valeurs des variables d'habitat préférées par le saumon puisque les effets semblent plutôt être non linéaires si l'on regarde les coefficients obtenus par le MAG et le MFLS. En comparant ces deux modèles, on constate que les coefficients du MFLS sont définis sur un domaine plus grand que ceux du MAG. Cela s'explique par le

fait que les effets du MAG sont estimés à partir des valeurs moyennes des prédicteurs à chaque site, alors que dans le MFLS, ceux-ci sont estimés à partir des FDP pour chaque variable d'habitat. Cela permet au MFLS de mieux estimer les valeurs des coefficients où peu de valeurs ont été échantillonnées (notamment dans les valeurs extrêmes). Les estimations dans les bornes du domaine des prédicteurs semblent donc erronées pour les modèles MAG, en particulier pour les vitesses élevées, le substrat grossier et les hautes températures, ce qui sera discuté plus amplement dans la section 2.2.3.

En ce qui concerne les résultats de modélisation, le *PseudoR*^{2} obtenu par les trois modèles est respectivement de 0.08, 0.60 et 0.62 pour le MLG, MAG et MFLS. Le MLG semble inadapté (confirmant la discussion ci-dessus) à la modélisation de l'habitat du poisson de par son très faible *PseudoR*^{2} obtenu. Les mesures *EXA*, *TPP* et *TAP* sont calculées selon les seuils maximisant les bonnes classifications en validation croisée pour chacun des modèles et reportées au tableau 4.3 de même que l'aire sous la courbe ROC. Les résultats selon la validation croisée « *5-fold* » sont d'abord regardés. Le MFLS a une exactitude de 84.5% alors les modèles MLG et MAG obtiennent respectivement 76.9% et 73.1%. Les MAG et MFLS prédisent le mieux les présences en validation croisée (*TPP* de 92.3% pour les deux modèles) alors que les MLG et MFLS prédisent le mieux les absences (*TAP* de 76.9% pour les deux modèles). Ces résultats indiquent une plus grande capacité pour le modèle fonctionnel MFLS à décrire à la fois les sites d'absences et de présences sur la RSM via la validation croisée que les modèles MAG et MLG. Ces conclusions sont les mêmes lorsque le critère de l'aire sous la courbe ROC est regardé. Le MLG, le MAG et le MFLS obtiennent respectivement des valeurs de 0.67, 0.77 et 0.82.

Finalement, ces mêmes métriques sont calculées lorsque les modèles développés sur la RSM sont utilisés avec les données de la RPC pour étudier la transférabilité des modèles d'habitat. Les mesures *EXA*, *TPP* et *TAP* sont calculées avec les mêmes seuils que ceux trouvés avec la validation croisée pour chacun des modèles, puisqu'en pratique, les observations réelles de présence-absence sur une nouvelle rivière ne seront pas disponibles. Les modèles MLG et MAG semblent être les meilleurs pour détecter les sites où il y a présence de saumon (*TPP* respectifs de 100% et de 83.3%), mais ne détectent correctement aucun site avec absence (*TAP* de 0% pour les deux modèles). Cela dit, de

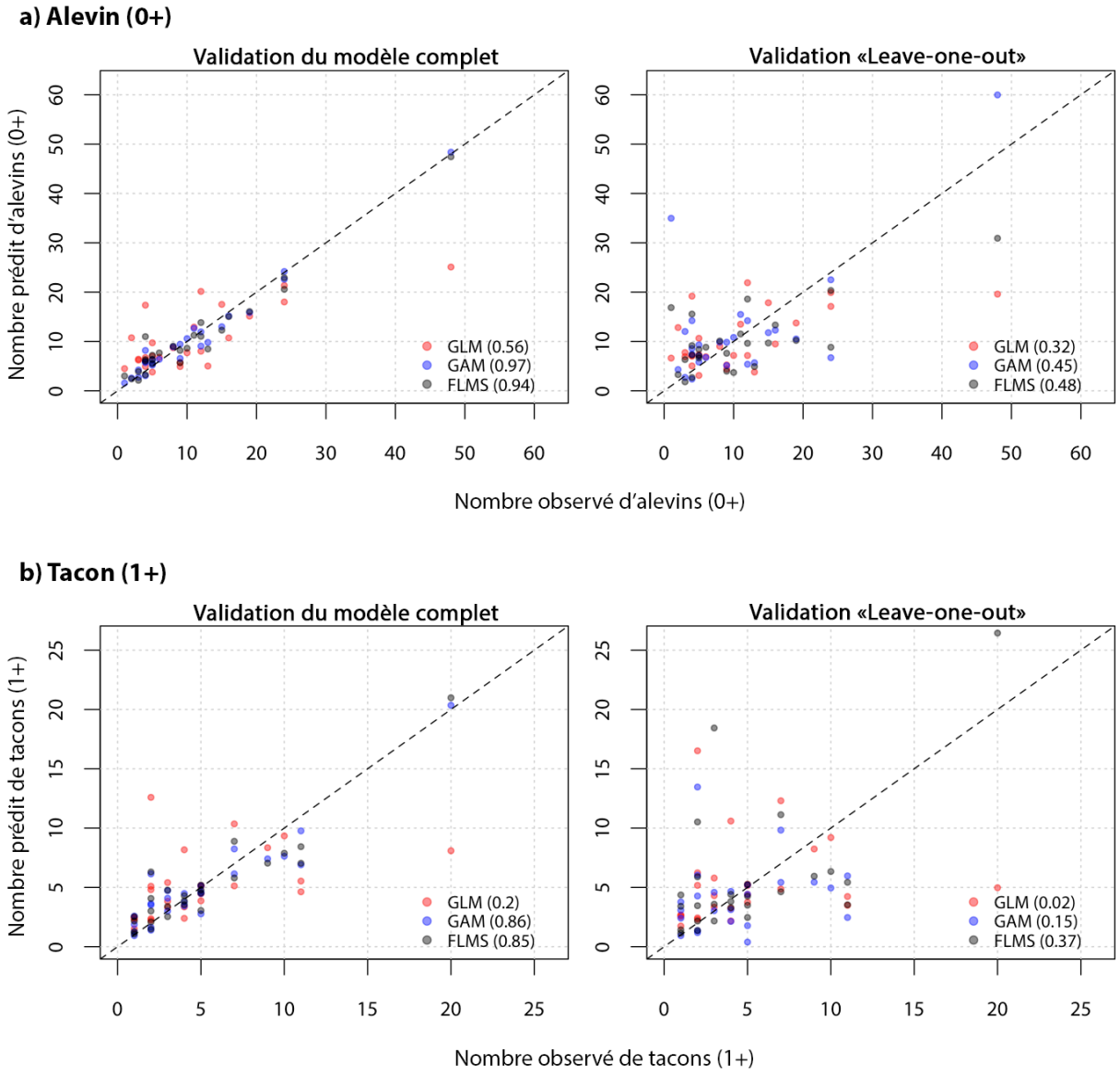
par leur incapacité à prédire les absences, ces deux modèles ne semblent pas transférables à la RPC. Pour ce qui est du MFLS, il détecte correctement trois quarts des présences (*TPP* de 75%) et un tiers des absences (*TAP* de 33.3%). Somme toute, le MFLS est le modèle qui semble le mieux adapté à prédire à la fois les absences et les présences sur la RPC lorsque les valeurs seuils de la validation croisée sont utilisées avec les données de la PCR. En regardant les aires sous le courbe ROC pour la validation de transférabilité, le GAM semble être le plus performant (0.58), suivi du MFLS (0.54) et du MLG (0.39). Ces résultats diffèrent de ceux obtenus avec les seuils optimaux trouvés en validation croisée. Il faut noter la métrique de l'aire sous la courbe ROC considère tous les seuils possibles alors les métriques *EXA*, *TPP* et *TAR* calculées ci-dessus ont utilisé le meilleur seuil obtenu en validation croisée, ce qui est plus susceptible d'être utilisé en pratique puisque les présences-absences observées ne seront pas disponibles.

Modèles d'abondances du 0+ et du 1+ : Pour le cas des modèles d'abondances, les résultats sont affichés comme les valeurs prédites versus les valeurs observées à la figure 4.6. Le *NSC* est affiché et permet d'évaluer le pouvoir explicatif des modèles. La validation de transférabilité (avec les données de la RPC) n'est pas affichée puisque les résultats ont donné des *NSC* inférieurs à 0.10 pour les trois modèles et les deux âges. Cela sera abordé plus en détail dans la partie discussion.

Pour ce qui est du 0+, les modèles MAG et FLMS semblent avoir une très bonne adéquation, avec des *NSC* supérieurs à 0.90. Pour le MLG, celui-ci diminue à 0.56, ce qui montre encore une fois que l'effet non linéaire des prédicteurs n'est pas pris en compte par ce modèle. Pour la validation « Leave-one-out », celle-ci donne des valeurs de *NSC* plus élevées pour le MAG (0.45) et le FLMS (0.48) que pour le MLG (0.32).

Concernant les saumons 1+, les résultats sont encore une fois meilleurs pour le MAG et le MFLS (*NSC* d'environ 0.85) alors qu'ils sont médiocres pour le MLG (*NSC* de 0.20). En validation « Leave-one-out », le FLMS a le *NSC* le plus élevé (0.37), suivi par le MAG (0.15) et le MLG avec un très faible *NSC* (0.02). Ainsi, on voit que dans les modèles d'abondances, c'est le MFLS qui obtient les *NSC* les plus élevés en validation « Leave-one-out », ce qui veut dire que c'est lui qui performe le mieux avec des données non utilisées dans la calibration.

Figure 2.3 : Résultat de la modélisation des abondances de 0+ et de 1+ avec les trois modèles



Note : Les *NSC* ont été reproduits entre parenthèses. Pour la validation du modèle complet, le *NSC* est l'équivalent du traditionnel R^2.

Valeur ajoutée de la variable de la température de l'eau : Des valeurs du R^2*ajusté* ont été calculées pour tous les âges (0+, 1+, 2+) pour des modèles avec et sans la variable de la température de l'eau. Les différences de R^2*ajusté* obtenues ont été reproduites au tableau 4.4. Dans tous les cas, le R^2*ajusté* augmente lorsque la température est incluse, en dépit du fait qu'un prédicteur additionnel est utilisé (c.-à-d. que le critère tient compte de la parcimonie des modèles). Cette augmentation est entre +0.03 et +0.43 pour le MLG,

alors qu'elle se chiffre plutôt entre +0.31 et +0.57 pour les MAG et MFLS. Dans tous les cas, une augmentation du critère du R^2*ajusté* est notée. Ainsi, la variable de température de l'eau est bel et bien importante en modélisation de l'habitat du saumon atlantique juvénile et son effet semble non linéaire avec de plus fortes augmentations du R^2*ajusté* avec les modèles non linéaires (MAG et MFLS) versus le modèle linéaire (MLG). Par exemple, comme le montre le modèle MFLS à la figure 2.2, les températures de l'eau inférieures à 18ºC semblent être préférées par le tacon 2+, avec une préférence optimale pour une température de 16ºC, alors que les températures plus chaudes (> 18ºC) sont évitées, plus particulièrement les températures supérieures à 22ºC.

2.2.3 Discussion

Comme il a été vu, le modèle MLG utilisé est peu performant de par les effets des prédicteurs supposés linéaires. Dans les deux autres modèles (MAG et MFLS), ceux-ci sont beaucoup plus performants de par leur capacité à tenir compte des effets non linéaires. Cela dit, le MFLS est un modèle linéaire, mais le fait que les coefficients $\beta_j(.)$ varient (possiblement non linéairement) selon la valeur prise par $x_j(.)$ permet de tenir en compte de ces effets non linéaires. Qui plus est, les validations croisées et de transférabilité avec les données de la RPC ont permis de montrer que l'utilisation des FDP pour chacun des prédicteurs dans le MFLS améliorait les résultats obtenus, comparés au MAG qui était basé sur des valeurs moyennes à chacun des sites. Aussi, tel que vu à la figure 2.2, les coefficients du MFLS étaient définis sur une grande partie du domaine des prédicteurs que le MAG, ce qui peut expliquer la plus grande performance obtenue. En effet, les extrapolations des coefficients pour le MAG sont susceptibles de fausser les estimations de ce modèle d'habitat dans les valeurs où peu de sites sont échantillonnés. Ces extrapolations ne semblent pas être représentatives biologiquement, alors que les coefficients du MFLS le sont beaucoup plus. Par exemple, des vitesses très élevées pour le saumon juvénile de 1 m/s et de 2 m/s ont le même effet dans le MFLS alors que dans le MAG, la vitesse de 2 m/s semble grandement préférée, ce qui n'est pas justifié d'un point de vue biologique. D'ailleurs, Jowett and Davey (2007) ont comparé les résultats d'extrapolation du MAG et de l'approche des courbes de préférence et ont souligné que les extrapolations au-delà des valeurs calibrées

représentaient un défi pour les deux approches testées. Ainsi, la régression fonctionnelle peut être vue comme une solution prometteuse à ces problèmes d'extrapolation.

Dans la littérature, le MAG de Hedger *et al.* (2005) obtient des valeurs de *NSC* de 0.28 pour la densité d'alevin (0+) et de 0.47 pour le tacon (1+ et 2+ combinés), lorsque le *NSC* est calculé avec les mêmes données que celles utilisées dans la calibration. Cela ressemble aux résultats obtenus avec le FLMS utilisé dans notre étude : *NSC* de 0.45 pour le 0+ et de 0.37 pour le 1+. Cependant, nos *NSC* sont le résultat d'une validation « Leave-one-out » alors que ceux de Hedger *et al.* (2005) ont été calculés à même les données utilisées pour la calibration. Ainsi, ces résultats auraient été fort probablement inférieurs si la validation « Leave-one-out » avait aussi été effectuée. En fait, dans la littérature, la majorité des résultats reportés n'incluent pas de validation croisée (Guay *et al.*, 2000; Mäki-Petäys *et al.*, 2002; Guay *et al.*, 2003; Hedger *et al.*, 2004; Hedger *et al.*, 2005; Millidine *et al.*, 2012; Beakes *et al.*, 2014). Cependant, notons que nos résultats incluent une variable supplémentaire, soit celle de la température de l'eau, qui s'est montrée hautement importante pour ce modèle.

Pour ce qui est des résultats des modèles de présence-absence développés pour le 2+, nous les avons comparés d'une manière similaire à la méthode développée par Guay *et al.* (2000). Il s'agit de tracer le lien entre les abondances observées de poissons et les probabilités prédites par les modèles de régression, dénommées indices probabilistes de qualité d'habitat (IPH). Ensuite, l'adéquation de la relation entre l'abondance et l'IPH est quantifiée à l'aide d'un R^2. Ces résultats sont comparés à un R^2 calculé en mettant plutôt en relation les IQH (calculés via l'approche par courbe de préférence de Leclerc *et al.* (1994)) et les abondances de poissons. L'adéquation est reportée à la figure 4.7. Les MAG et MFLH ont des R^2 supérieurs à 0.90 alors que l'approche de l'IQH obtient un faible R^2 de 0.41. Ce résultat est comparable aux autres résultats obtenus dans la littérature pour l'approche de l'IQH : R^2 de 0.39 pour le tacon dans Guay *et al.* (2000) et des R^2 variant de 0.30 à 0.46 pour l'alevin et de 0.35 à 0.69 pour le tacon (Hedger *et al.*, 2004). Ainsi, les modèles de régression considérant les effets non linéaires des prédicteurs (comme le MAG et le FLMS) semblent mieux adaptés à l'habitat aquatique que l'approche de l'IQH et ont un plus fort R^2 pour expliquer l'abondance de poissons.

En ce qui concerne la transférabilité, les modèles d'abondance n'ont pas montré de potentiel pour être utilisés sur une autre rivière. Cela peut d'abord s'expliquer par les différences d'abondance observées sur les deux rivières : abondance moyenne d'alevin (0+) de 10.6 sur la RSM contre 3.1 sur la RPC. De plus, plusieurs sites n'avaient que 0 alevin (4 sites) ou 1 alevin (4 sites) sur la RPC alors qu'il n'y avait sur la RSM aucun site avec 0 alevin et qu'un seul site avec 1 alevin. Malgré ces différences au niveau des alevins, les modèles d'abondance du 1+, où les abondances moyennes semblaient davantage similaires, n'ont tout de même pas été en mesure d'être transférés. Le fait que la mobilité des 0+ et du 1+ soit inférieure à celle du 2+ peut expliquer que des habitats d'une moins bonne qualité aient aussi été utilisés sur la RPC, ce qui est décrit comme la plasticité comportementale des poissons et pourrait expliquer la faible transférabilité (Mocq *et al.*, 2013). Aussi, bien que cette étude était l'une des premières à inclure la variable de la température de l'eau, d'autres prédicteurs comme l'intensité lumineuse (Heggenes & Gunnar Dokk, 2001; Girard *et al.*, 2003), la distance entre la frayère et l'habitat du saumon (Klemetsen *et al.*, 2003), la disponibilité en nourriture (Vehanen, 2003), la compétition (Gabler & Amundsen, 1999) et la prédation potentielle (Dionne & Dodson, 2002; Vehanen, 2003) demeurent à être étudiées. Finalement, d'autres tests ont aussi été effectués pour tenter de valider la transférabilité des modèles d'abondance, comme la standardisation des données d'abondance par rivière, mais ces essais n'ont pas montré de meilleurs résultats et n'ont pas donc pas été explicités.

Outre les avantages discutés plus haut de la régression fonctionnelle, il faut souligner qu'elle a nécessité un travail de terrain plus ardu pour obtenir les courbes (les FDP) basées sur 30 mesures par site. Bien que nous ayons essayé de définir ces courbes avec moins de valeurs (~15), cette faible quantité d'information était insuffisante et ne représentait pas bien l'hétérogénéité des mesures d'habitat à chaque site. Néanmoins, avec les avantages de performance et d'explication du modèle fonctionnel, son intérêt en habitat aquatique est tout de même démontré. Avec le développement de techniques de prise de mesures automatiques par l'analyse de photo/vidéo et la capture d'images aériennes, ces caractéristiques de l'habitat seront plus facilement récoltables (taille du substrat (Carbonneau *et al.*, 2005; Hedger *et al.*, 2006), mesures thermiques (Dugdale,

2016), vitesses de surface (Smith *et al.*, 2005)) et les modèles fonctionnels deviendront plus facilement applicables.

Finalement, il faut noter que plusieurs modèles d'habitat sont basés sur le microhabitat (caractérisation de l'habitat au nez du poisson). Cependant, ces mesures ne décrivent pas bien l'habitat du poisson et l'utilisation potentielle des habitats qui sont à proximité (Shirvell, 1994; Beecher *et al.*, 2010). D'ailleurs, l'inclusion de mesures intermédiaires dans les modèles d'habitat pourrait améliorer leur pouvoir prédictif (Mocq *et al.*, 2018). Ainsi, le cadre proposé dans le présent projet permettrait aussi d'utiliser dans les modèles de microhabitat des FDP prenant en compte la variabilité dans l'habitat immédiat du poisson comparé à des mesures seulement à son nez dans les modèles classiques.

3 ARTICLE 1 : MODÉLISATION DE LA TEMPÉRATURE DE L'EAU EN RIVIÈRE AVEC DES MODÈLES DE RÉGRESSION FONCTIONNELLE

Stream Temperature Modelling using Functional Regression Models

J. Boudreault[a*], N. E. Bergeron[a], A. St-Hilaire[a,b] and F. Chebana[a]

[a] Institut National de la recherche scientifique – Centre Eau Terre Environnement, Québec, Canada ;

[b] Canadian River Institute, University of New Brunswick, Fredericton, Canada.

Manuscript submitted

March 16th 2018

*Corresponding author: Jeremie Boudreault (jeremie.boudreault@ete.inrs.ca)

Abstract

Stream temperature is one of the most important river variable in lotic habitats as it affects the ecosystem profoundly. Given the continuous nature of the stream temperature variable, the aim of this paper is to introduce functional regression for the air-stream temperature relation, being capable to model an entire seasonal or annual curve of temperatures as one entity, rather than multiple daily or weekly values in the classical models. Two functional models were explored, the fully functional linear model (FFLM) and the historical functional linear model (HFLM), and compared with two classical models. Three rivers from the US were considered and the HFLM had the best fit for two of them based on three performance criteria. When comparing the functional models, the HFLM clearly outperforms the FFLM for all three rivers. Functional regression, especially the HFLM, leads to encouraging results to model the complete stream temperature curve compared to other classical approaches.

Keywords: Stream temperature modelling; functional regression; fully functional linear model; historical functional linear model; generalized additive model; logistic model

Résumé

La température de l'eau en rivière est l'une des variables les plus importantes pour les habitats lotiques comme elle affecte profondément l'écosystème aquatique. Étant donné la nature continue de la variable de la température de l'eau, le but de cette étude est d'introduire le modèle de régression fonctionnelle pour modéliser la relation entre la température de l'air et de l'eau. Le modèle de régression fonctionnelle est capable de modéliser une année ou saison entière de températures de l'eau comme une seule entité, plutôt que plusieurs valeurs quotidiennes ou hebdomadaires dans les autres modèles classiques. Deux modèles fonctionnels ont été explorés, le modèle fonctionnel linéaire complet (MFLC) et le modèle fonctionnel linéaire historique (MFLH), et comparés à deux modèles classiques. Trois rivières des États-Unis ont été considérées et le MFLH avait la meilleure performance pour deux des trois rivières basée sur trois critères de validation. Lorsque les modèles fonctionnels étaient comparés entre eux, le MFLH surpassait largement le MFLC pour les trois rivières. La régression fonctionnelle, particulièrement le MFLH, a donné des résultats encourageant pour modéliser la température de l'eau en rivière sous forme d'une courbe comparée aux approches classiques.

Mots-clés: modélisation de la température de l'eau en rivière ; régression fonctionnelle ; modèle fonctionnel linéaire complet ; modèle fonctionnel linéaire historique ; modèle additif généralisé ; modèle logistique.

3.1 Introduction

Stream temperature is one of the most important variables when studying aquatic ecosystems (Beschta et al., 1987, Caissie, 2006). Indeed, many fishes are sensitive to water temperature, especially during growth and spawning periods (Handeland et al., 2008). Their ability to achieve their biological functions correctly can be severely modified if the water temperature is inadequate (Bjornn and Reiser, 1991, Lee and Rinne, 1980, Sigholt and Finstad, 1990). Furthermore, the thermal regime of a river can be modified through human activities. For example, forestry, and more precisely deforestation, influences the canopy cover, which causes changes to the heat input into the river (Beschta et al., 1987). Webb and Walling (1993) showed that flow regulation increases the mean value of the stream temperature. More recently, Maheu et al. (2016) found that the presence of dams on medium sized rivers of Eastern Canada cause an increase in September monthly mean temperature. Finally, global warming has an impact on increasing stream temperature (Kaushal et al., 2010) and this increase is more likely to continue in the future and with a higher amplitude (Morrison et al., 2002).

For these reasons, models that predict and simulate stream temperature are needed. They can be divided into two categories: deterministic and statistical models (Benyahya et al., 2007a). The former focus on mathematical relations that characterize the physical processes of heat transfer and link the stream temperature to other hydro-meteorological and physical variables, based on an energy budget approach (Morin et al., 1981). However, these models require several input variables and are often costly in computation time (Benyahya et al., 2007a). Statistical models typically use fewer predictors and computational time than deterministic models, often leading to more simplicity than deterministic models. Statistical models include non-parametric approaches like the artificial neural-network (DeWeber and Wagner, 2014, Piotrowski et al., 2015, Bélanger et al., 2005, Chenard and Caissie, 2008) and the k-nearest neighbors model (St-Hilaire et al., 2012, Benyahya et al., 2008). Parametric models can be divided into two categories: stochastic and regression models. The first category includes times series models like the second-order Markov process (Caissie et al., 1998, Cluis, 1972, Caissie et al., 2001), the periodic autoregressive model (Benyahya et al., 2007b) and the

non-linear autoregressive process with exogenous variable (Kwak et al., 2017). In the second category, i.e. regression models, the stream temperature is modelled as a function of one or multiples predictors. This relation is often assumed to be linear (Crisp and Howson, 1982, Jeppesen and Iversen, 1987, Jourdonnais et al., 1992, Mackey and Berrie, 1991, Stefan and Preud'homme, 1993) or non-linear like the logistic model of Mohseni et al. (1998), the Gaussian process of Grbić et al. (2013) and the generalized additive model (Laanaya et al., 2017, Wehrly et al., 2009). However, regression models can suffer from collinearity when significant correlation exists between predictors and thus, a Ridge regression can be used to overcome this problem (Ahmadi-Nedushan et al., 2007). See (Caissie, 2006, Benyahya et al., 2007a, Webb et al., 2008) for reviews on stream temperature modelling.

For all the aforementioned models and others described in the literature, stream temperature data is considered in a punctual or discrete manner, as it is recorded. However, stream temperature is a continuous variable and shows a strong seasonal pattern. Hence, it can be better represented by a curve over a year or a season rather than using aggregate values such as a mean over a certain time period. Pilgrim et al. (1998) highlighted the fact that the air-water temperature relation is less scattered when using long aggregation period (annual, monthly) instead of short ones (weekly, daily), motivating the use of such long aggregation to model stream temperature. For some current methods, a serious loss of information is sustained, which is not the case when working with an entire curve of stream temperature. Moreover, extracted stream temperature descriptive statistics (e.g. maximum temperature, the number of days with a temperature over a threshold, the day at which the maximum temperature occurs) are fully described or incorporated by a curve of stream temperature. For these reasons, it is of interest to introduce the **functional data analysis** (FDA) in this context. FDA is a framework that uses curves or functions as observations, in contrast with scalar or vectors in other classical contexts. Hence, working with a curve instead of daily, weekly or monthly metrics naturally leads to a better understanding of the whole phenomenon and captures more of its variability. Another interesting fact about using FDA in this context is to avoid collinearity in the predictors used in regression models. For example, models using air temperature at different time scales, e.g. at t-1 and t-2, to model stream temperature at t,

will suffer from possible redundancy as the air temperature series is highly autocorrelated. However, in the FDA case, this problem is avoided as the whole curve of predictor values (in the present study, air temperature) can be used as a single input (Cuevas et al., 2002). Finally, in the functional case, the prediction errors remain constant over the modelled period because it is a single predicted curve while this error would increase when using times series approaches (Box et al., 2015).

FDA was introduced by Ramsay (1982) and became very popular and present in the literature, with multiple textbooks (Bosq, 2012, Dabo-Niang and Ferraty, 2008, Ferraty and Vieu, 2006, Ramsay, 2006, Ramsay et al., 2009) and recent applications in various fields including ecology (Bel et al., 2011, McDonald et al., 2015), transportation (Chiou, 2012), energy (Chaouch, 2014, Brockhaus et al., 2015), waste management (Bernardi et al., 2017), biotechnology (Brockhaus et al., 2017a), medecine (Ciarleglio et al., 2016), neuroscience (Ivanescu et al., 2014, McLean et al., 2014, Meyer et al., 2015). Chebana et al. (2012) were the first to introduce FDA in the hydrological context, seeing the annual series of streamflow, the hydrograph, as a curve i.e. a functional datum. Ternynck et al. (2016) classified the hydrographs using FDA. Masselot et al. (2016) modelled the streamflow curve using the precipitation curve with functional regression. Larabi et al. (2017) used FDA to improve hydrological model realism. Even though work has been done in applying FDA on streamflow data, this work is the first considering FDA on stream temperature. Similarly to the streamflow, where the hydrograph was seen as a curve over a year (Chebana et al., 2012), such a representation is also suitable for stream temperature, perhaps even more so, given its pronounced seasonal trend. As this work aims at introducing functional regression for stream temperature data, only one predictor, the air temperature, is used. Functional regression models are compared to two other regression models used with the same predictor, i.e. the Logistic Model (LM), widely used in the literature (Mohseni and Stefan, 1999, Caissie et al., 2001, Webb et al., 2003, St-Hilaire et al., 2012, Laanaya et al., 2017, Grbić et al., 2013, Segura et al., 2015, Morrill et al., 2005) and the Generalized Additive Model (GAM) (Wehrly et al., 2009, Laanaya et al., 2017). As a generalization of the LM, it allows for more flexibility. In one case study, it was recently shown to be the best model for daily mean stream temperature, compared to three other models (Laanaya et al., 2017).

The next section presents the models used in this study and the performance measures. Section 3.3 presents the case study and the results. Discussion is in section 3.4 and section 3.5 presents the conclusion.

3.2 **Methodology**

This section first introduces the functional regression models. Then, the generalized additive model and the logistic model are briefly presented. Finally, the model performance criteria are shown.

3.2.1 Functional regression models

As mentioned in the introduction, FDA aims at working with functions (curves) instead of discrete observations. Because data are not recorded continuously, a first step in FDA is to find a function $x(t)$ that adequately represents the given discrete observations. This is achieved by defining $x(t)$ as a weighted sum of M known basis functions (Ramsay, 2006):

$$x(t) = \sum_{m=1}^{M} c_m \phi_m(t) \quad , \quad t \in \Omega_1 \tag{3.1}$$

where c_m and $\phi_m(t)$ are respectively the coefficients to be estimated and the basis functions. For periodic data, a Fourier series expansion is often used as basis functions and for non-periodic data, the most popular basis functions are B-splines (Ramsay, 2006). Once the data curves have been constructed, the functional data (i.e. the curves) are ready to be used in a functional regression model.

In the standard linear regression model, all values are scalar (Neter et al., 1996). In the functional case, three possible scenarios occur (Morris, 2015). The scalar-on-function models, where at least one of the predictors is functional but the predictand is a scalar. On the opposite, the function-on- scalar models, where the predictand is functional but all the predictors are with scalar values. The last one is the function-on-function models, where the predictand and at least one of the predictors are functional (Ramsay, 2006).

In order to model the stream temperature *curve* using all available information from air temperature, i.e. the air temperature *curve*, it is appropriate to consider the function-on-function models. First, the most simple model in this class is called the concurrent model (Ramsay, 2006). It is the analogue to the linear model with functional observations:

$$y_i(t) = \alpha(t) + \beta(t)x_i(t) + \varepsilon_i(t) \quad , \quad t \in \Omega_1 \tag{3.2}$$

where $y_i(.)$ and $x_i(.)$ represent respectively the curves of stream and air temperatures for the year *i* and day *t*. Note that in this model, the intercept $\alpha(t)$ and the regression coefficient $\beta(t)$ are also functions. Basically, their expressions are of the same nature as of $y(t)$ and $x(t)$ (equation 3.1) as linear combination of basis functions. Note that model (3.2) only allows the predictor $x_i(.)$ to influence $y_i(.)$ at the same time *t* or at a fixed lag *t-l*. However, the predictor at different time can affect the predictand at time *t*. In the case of stream temperature, it can take several days before a rise or fall in air temperature induces a change in stream temperature. Hence, the concurrent model in (3.2) is not suitable to incorporate this heat exchange process and the **fully functional linear model** (FFLM) should be introduced:

$$y_i(t) = \alpha(t) + \int_{\Omega_2} \beta(s,t)x_i(s)ds + \varepsilon_i(t) \quad , \quad t \in \Omega_1 \tag{3.3}$$

where $\beta(s,t)$ is a surface representing the effect of $x_i(.)$ at time *s* on $y_i(.)$ at different time *t*. The domain Ω_2 of *s* can differ from the domain Ω_1 of *t*. In this model, all the values of $x_i(.)$ are used to model the value of $y_i(.)$ at time *t*. In Masselot et al. (2016), the FFLM (3.3) was used to model streamflow curve from the precipitation curve, allowing backward effect of the precipitation on the streamflow i.e. precipitations occurring at $s > t$ were used to model streamflow at *t*. Even if from a meteorological point of view the precipitations occurring at $s > t$ cannot affect the streamflow at time *t*, the $\beta(s,t)$ surface was close to 0 for $s > t$, motivating the use of the FFLM with those variables. In the case considered here, it is more appropriate to introduce the **historical functional linear model** (HFLM) as it allows to restrict the domain of influence of the predictor on the predictand (Malfait and Ramsay, 2003, Gervini, 2015, Harezlak et al., 2007). The HFLM has known recent developments

regarding the implementation (Brockhaus et al., 2017a, Brockhaus et al., 2015) and is defined as:

$$y_i(t) = \alpha(t) + \int_{l(t)}^{u(t)} \beta(t,s)x_i(s)ds + \varepsilon_i(t) \quad , \quad t \in \Omega_1 \tag{3.4}$$

where l(t) and u(t) are respectively the lower and the upper bounds where the predictor x_i(.) can influence y_i(.) at time t. This model allows an important flexibility with particular cases of interest. For instance, if one wants to let only the past values of x_i(.) to effect y_i(.) at t, then l(t) = 0 and u(t) = t. If one wants to let only the past values of length d to effect x_i(.) on y_i(.) at t, then l(t) = max(0, t - d) and u(t) = t (Malfait and Ramsay, 2003, Kim et al., 2011). For comparison and appropriateness purposes, models (3.3) and (3.4) will be used.

In order to fit functional models, the functional linear array model (FLAM) implemented in R (R Core Team, 2017) by Brockhaus et al. (2015) in the package *FDboost* (Brockhaus et al., 2017b) was used as it can fit both model (3.3) and (3.4). The FLAM is fit through a component-wise gradient boosting algorithm, a machine learning procedure that estimates the model parameters by minimizing an empirical loss function (Freund et al., 1999, Bühlmann and Hothorn, 2007, Brockhaus et al., 2017a). In our case, the minimized loss function is the mean, but the package can also handle the median or any quantile (Brockhaus et al., 2015):

$$\sum_{i=1}^{n} \int_{\Omega_1} [y_i(t) - \hat{y}_i(t)]^2 dt \tag{3.5}$$

where $\hat{y}_i(.)$ is the predicted stream temperature curve by the functional model. The minimization procedure is iterative and each iteration leads to a rougher β(s, t) surface (Brockhaus et al., 2015). The minimization procedure can be stopped when the error stops decreasing (called *early stop*), leading to more regularized effects (i.e. smoother β(s, t) surface) and more stable predictions (Brockhaus et al., 2017c). More information about the FLAM, the *FDboost* package and the boosting algorithm can be found in the above references.

3.2.2 Generalized additive model

This model was introduced by Hastie and Tibshirani (1990) as a natural extension of the generalized linear model, without the assumption of linearity in the relationship between the predictor and the predictand, as well as relaxing the normality assumption. Even if only few applications of the GAM have been done in stream temperature (Wehrly et al., 2009, Laanaya et al., 2017), it has been widely used in hydrology (Rahman et al., 2018, Falah et al., 2017, Iddrisu et al., 2017, Zhang et al., 2015, Chebana et al., 2014). The generalized additive model (GAM) is defined as:

$$g(E(y_i)) = f_1(x_{1,i}) + f_2(x_{2,i}) + \cdots + f_p(x_{p,i}) \quad (3.6)$$

where g is the link function, $E(y_i)$ is the expected value of the predictand (in our case, the daily mean stream temperature), $x_{j,i}$ is the jth predictor and f_j is a smooth nonlinear function (often combination of cubic splines) applied to the predictor $x_{j,i}$ and i denotes the i^{th} observed stream temperature. Here, we use the identity function as link function $g(.)$ and two covariates, x_1, the Julian day of year to account for the seasonality in stream temperature data and x_2 the daily mean air temperature as implemented by Laanaya et al. (2017).

To avoid overfitting, a penalized GAM is used and fitted by minimizing an error term calculated by cross-validation with the package *mgcv* (minimized generalized cross-validation) (Wood, 2015, Wood, 2006) in R (R Core Team, 2017).

3.2.3 Logistic model

As highlighted in Mohseni et al. (1998), the relation between air and stream temperatures can be represented with a continuous S-shaped function better than assuming a linear relation. A common function to model this relation is the logistic model (LM) defined as:

$$T_w = \mu + \frac{\alpha - \mu}{1 + e^{\gamma(\beta - T_a)}} \quad (3.7)$$

where T_w and T_a are respectively stream and air mean temperatures. The parameters to be adjusted are α, the estimated maximum stream temperature, γ, the measure of the

steepest slope, β, the air temperature inflection point and μ, the minimum temperature. Mohseni et al. (1998) first used the LM (3.7) to estimate the whole year of weekly stream temperatures data for different watercourses in the USA. They highlighted the fact that the air-water relation was different for high and low values and suggested to separate the rising and falling limbs of the data prior to modeling the temperatures, considering the hysteresis in the series. Hence, model (3.7) is fitted separately for the rising and falling limbs of the data. The model is applied on daily mean temperature for comparison purpose, as done in recent applications (Laanaya et al., 2017, Bustillo et al., 2014, Grbić et al., 2013, St-Hilaire et al., 2012, Kelleher et al., 2012).

3.2.4 Performance criteria

To assess model performances, criteria commonly used in the context of stream temperature are calculated for each model as well as new functional criteria. A cross-validation technique, called the leave-one-out cross-validation year method or the jack-knife technique is used (Quenouille, 1949). Basically, one year of observed data is removed prior to fit each model. Then, the fitted model is used to predict the stream temperature for the year that was removed. The performance criteria are then calculated based on the predicted values for each year when it was removed from the calibration period.

Classical criteria: Those criteria are valid for daily predicted stream temperature values. As functional regression models a curve of values, the daily predicted temperatures are extracted from the curve by taking the mean temperature values on every [*t*, *t*+1[period as follow: $\hat{T}_w(t_i) = \int_{t_i}^{t_i+1} \hat{y}(t)dt$ where $\hat{y}(t)$ is the predicted stream temperature curve.

The first criterion is the root mean square error, denoted RMSE (Janssen and Heuberger, 1995) and is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(T_w(t_i) - \hat{T}_w(t_i)\right)^2} \tag{3.8}$$

where $T_w(t_i)$ and $\hat{T}_w(t_i)$ are respectively the observed and predicted daily mean stream temperature at time t_i (in days), and n is the number of data points.

The second criterion is the bias error (Benyahya et al., 2007a). It indicates if the model is overestimating (bias > 0) or underestimating (bias < 0) the observed values. It is defined by:

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{T}_w(t_i) - T_w(t_i)\right) \tag{3.9}$$

Finally, the third criterion is the Nash-Sutcliffe coefficient of efficiency, denoted NSC (Nash and Sutcliffe, 1970). It is less than 1 and a value of 1 indicates a perfect matching of the model to the observed data. It is defined as:

$$NSC = 1 - \frac{\sum_{i=1}^{n}\left(T_w(t_i) - \hat{T}_w(t_i)\right)^2}{\sum_{i=1}^{n}(T_w(t_i) - \bar{T}_w)^2} \tag{3.10}$$

where $\bar{T}_w$ is the mean stream temperature for the observed period of year *i*.

Functional criteria: The functional criteria are used to compare the two functional models as they model a curve of values for each year instead of multiple daily values. Not much effort has been done in the recent literature to develop such criteria in the FDA framework.

The first criterion is the analogue of the coefficient of determination R^2 in the standard regression (Neter et al., 1996) and was first defined in Ramsay (2006) and recently used by McDonald et al. (2015). It is called the squared correlation function (denoted $funR^2(.)$) and is defined as:

$$funR^2(t) = 1 - \frac{\sum_{i=1}^{n}\left(y_i(t) - \hat{y}_i(t)\right)^2}{\sum_{i=1}^{n}\left(y_i(t) - \bar{y}(t)\right)^2} \quad , \quad t \in \Omega_1 \tag{3.11}$$

where $y_i(t)$ and $\hat{y}_i(t)$ are respectively the observed and predicted curve of stream temperature for year *i*, *n* is the number of functional observations (i.e. number of years)

and $\bar{y}(t) = n^{-1}\sum_{i=1}^{n} y_i(t)$ is the mean functional observation. The *fun*R^2(.) is a function of *t* and shows how the model performs at every moment *t* compared to $\bar{y}(t)$. The *fun*R^2(.) can be integrated over its domain, to be a real valued coefficient denoted *fun*R^2, to give an indication of how the functional model performs for the observed period.

Furthermore, a functional version of the classical RMSE (equation 3.8) has been defined in (Brockhaus et al., 2017a, Brockhaus et al., 2015) and given by:

$$funRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\int_{\Omega_1}\left(y_i(t) - \hat{y}_i(t)\right)^2\, dt} \tag{3.12}$$

The *fun*RMSE can be compared to the classical RMSE (equation 3.8), except that discrete observations are replaced by curves integrated over their domain Ω_1. In a similar way, we can define the functional versions of the bias (*fun*Bias) and of the NSC (*fun*NSC) respectively as follows:

$$funBias = \frac{1}{n}\sum_{i=1}^{n}\int_{\Omega_1}\left(\hat{y}_i(t) - y_i(t)\right)dt \tag{3.13}$$

$$funNSC = 1 - \frac{\sum_{i=1}^{n}\int_{\Omega_1}\left(y_i(t) - \hat{y}_i(t)\right)^2 dt}{\sum_{i=1}^{n}\int_{\Omega_1}(y_i(t) - \bar{y}_i)^2\, dt} \tag{3.14}$$

3.3 Case study and results

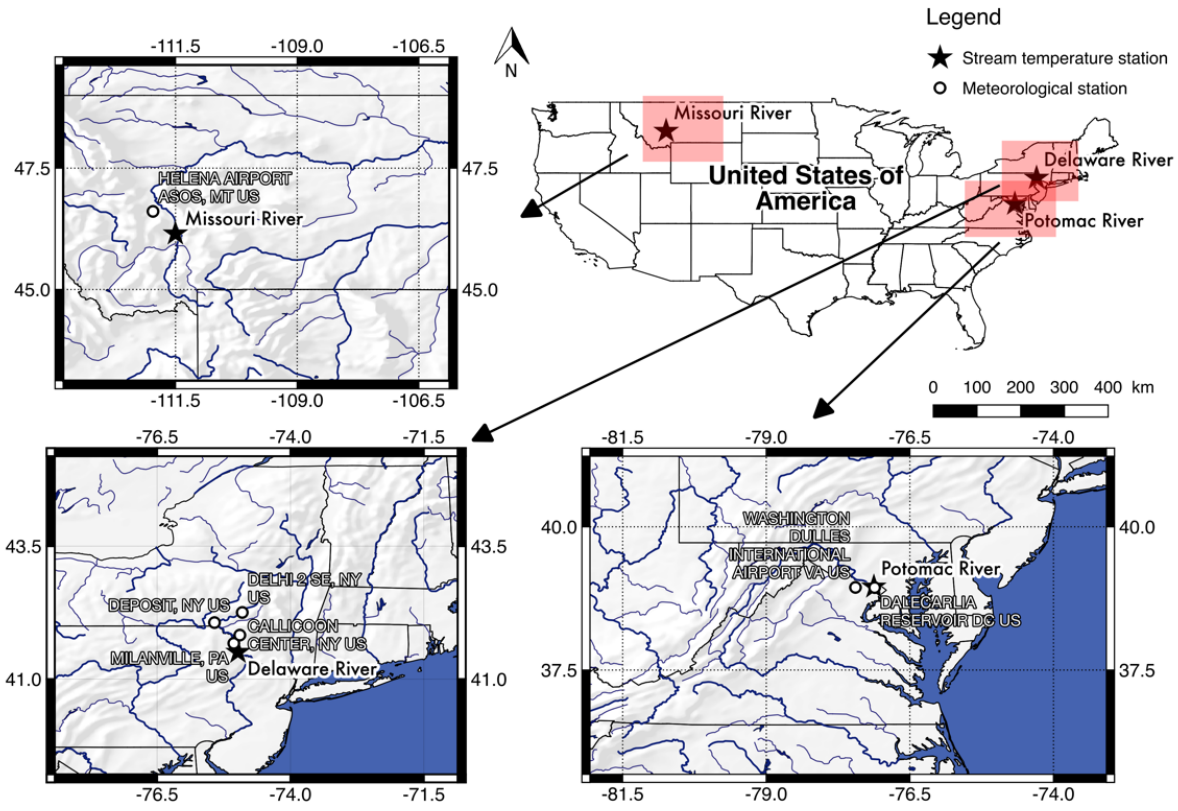
This section first introduces the data used in the study. Then, the models fitting procedure is briefly explained and results are shown.

3.3.1 Data description

Stream temperature regimes can vary widely from a small watercourse to a large river. For instance, Maheu et al. (2014) provided an example of the impact of river size on evaporative cooling, showing the differences that can exist in the relation between air and

stream temperature, according to river size. Thus, the present article proposes a case study based on three different rivers from the United States with long series of observations to see how functional regression models can be applied for different sized rivers and with different thermal regimes. Recall that FDA uses curves instead of scalars or vectors, meaning that a year of daily records corresponds to n=1 functional observation in the functional case while it is n=number of observations as number of days in the classical models. Hence, a considerable number of observed years is needed to apply functional regression to stream temperature, but it predicts a whole year of stream temperature instead of multiple daily values. The map of Figure 3.1 shows the location of the rivers and the meteorological stations used.

Figure 3.1 : Map of the stations



Mean daily stream temperatures were obtained from the United States Geological Survey (USGS, 2017). Three rivers with long enough observed period were found: Potomac River (USGS code: 01646500), Missouri River (06054500) and Delaware River (01428500),

respectively with 25, 25 and 26 usable years. The Potomac River is southernmost and the warmest one. The Missouri River goes under 0ºC and freezes during winter. The river station is situated near Toston, Oregon. The Delaware River is situated in the state of New York and does not experience freezing. The selected period for stream temperature is May 1st to October 31st as it is the period with less missing values for all three rivers. All years with less than 20 missing values in the selected period were kept for the analysis, to ensure a high enough number of observations was used in the functional regression models. Information about the three rivers is summarized in Table 3.1. For air temperatures, the closest meteorological station to the stream temperature station with less than 5 missing values in the observed period was used. Air temperatures were obtained from the *National Oceanic and Atmospheric Administration* (NOAA, 2017). Table 3.2 indicates which stations were used for each river and their respective distances. The selected period was the same than for stream temperature, but the starting date was chosen 15 days before the stream temperature series, leading to the period of April 16th to October 31. This choice was motivated by the fact that there is a lag between an increase in air temperature and the resulting effect on stream temperature, especially for larger systems (Preud'homme and Stefan, 1992). For both series, a linear interpolation was made for missing values to ensure complete data were used in the analysis.

Tableau 3.1 : Rivers data used in the study

River	River temperature station location	Drainage area at the station (km^2)	Mean streamflow in the observed period (m^3/s)	Temperature (ºC)			Number of years available
				Min	Mean	Max	
Potomac	Washington, District of Columbia	29 940	262.1	9.3	23.4	33.3	25
Missouri	Toston, Montana	37 920	153.4	0.5	15.5	25.5	25
Delaware	Lackawaxen, New York	5 232	84.4	5	19.2	30.1	26

Tableau 3.2 : Meteorological station used for each river

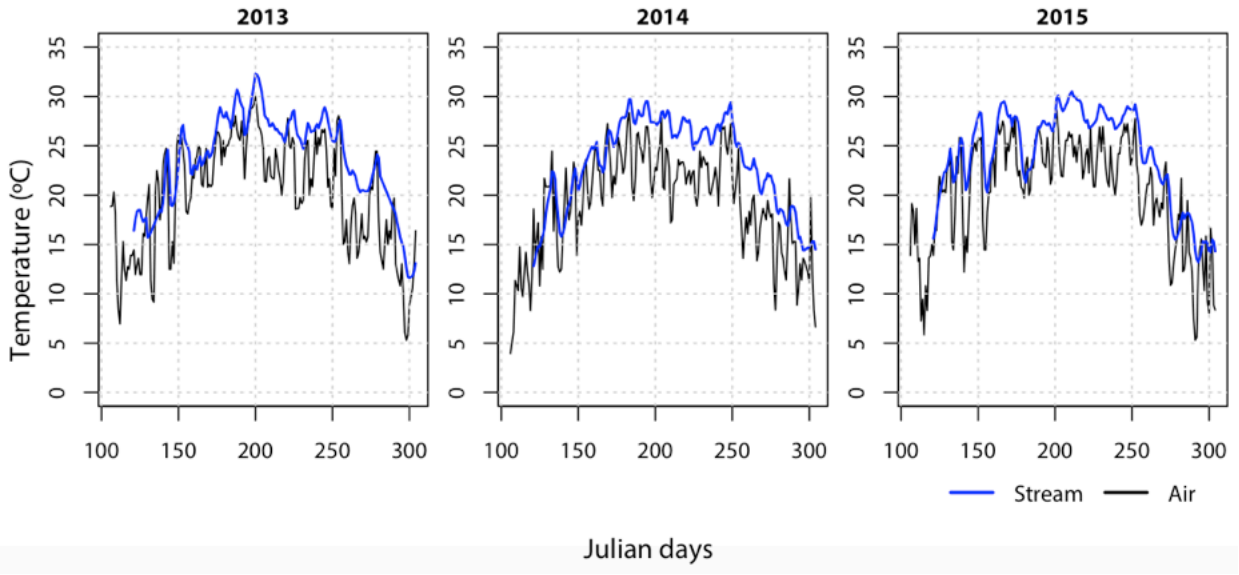
River	Meteorological station	Distance (km)	Years
Potomac	DALECARLIA RESERVOIR DC US	1.87	1989-1993, 1996-1998, 2001-2002, 2005-2012
	WASHINGTON DULLES INTERNATIONAL AIRPORT VA US	14.38	1995, 2000, 2003-2004, 2013-2015
Missouri	HELENA AIRPORT ASOS, MT US	62.1	1978-1979, 1981-1985, 1987-1988, 1990-1993, 1995-1996, 1998-1999, 2001-2008
Delaware	DELHI 2 SE, NY US	14.82	1983, 1988, 1992-1996, 1998, 2000-2010
	DEPOSIT, NY US	17.65	1989, 1991, 2011
	MILANVILLE, PA US	23.27	2013-2014
	CALLICOON CENTER, NY US	37.54	2012

3.3.2 Models fittings and regression coefficients

This section will briefly explain the fitting procedure and the resulting four models with the example of the Potomac River.

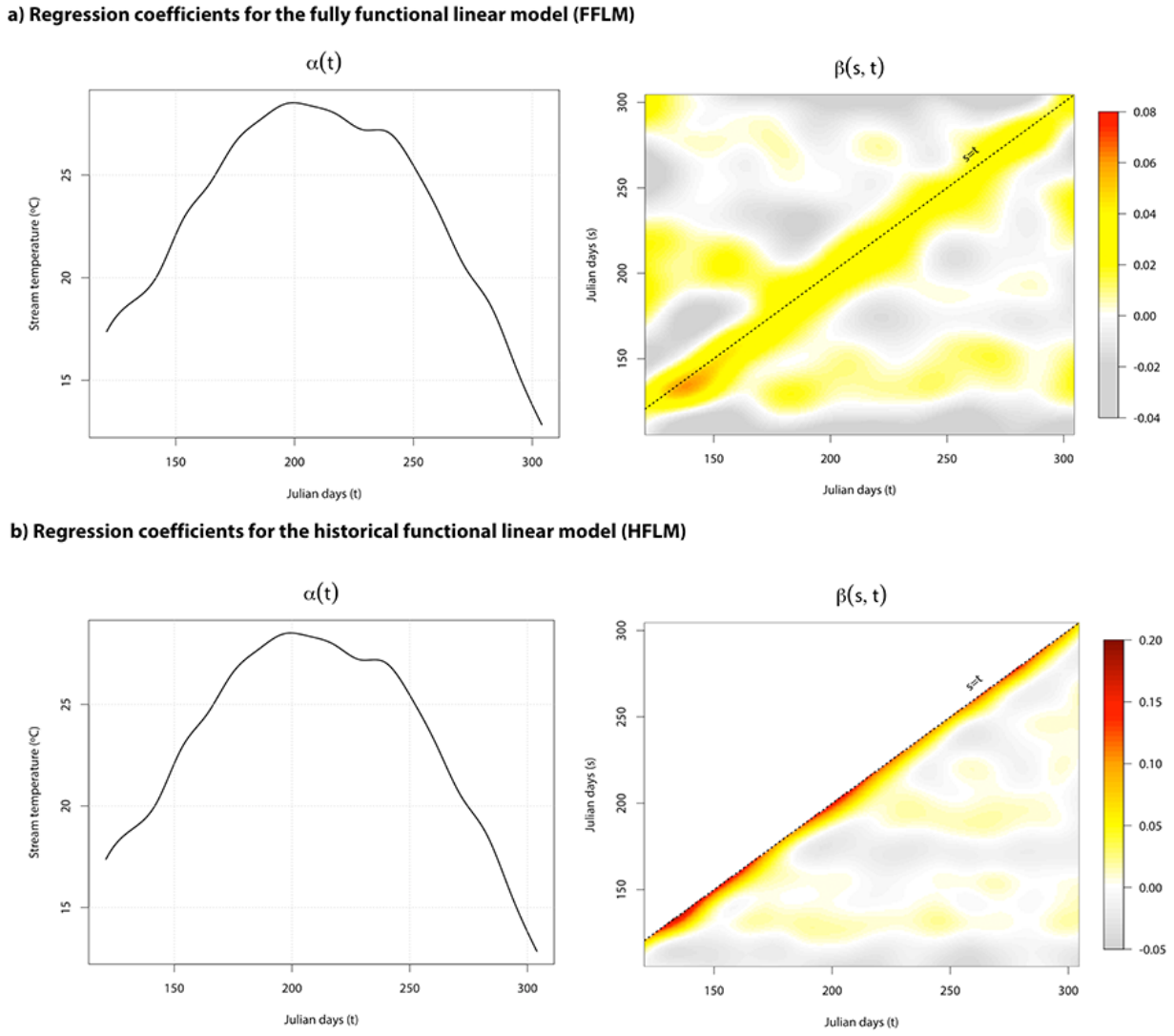
Functional regression models: Potomac River has 25 available years of daily mean temperatures, which means 25 functional observations. The values of stream and air temperature are illustrated in Figure 3.2 for the last three years of this river. The domain of $y(.)$ is Ω_1 = [121; 305] days and the domain of $x(.)$ is Ω_2 = [106; 305] days, both measures in Julian days of year. For the HFLM (model 4), the bounds of the integral are set to $l(t) = 0$ and $u(t) = t$, allowing only the values of air temperature at $s < t$ to influence the stream temperature at time t. The iterative procedure to fit the two functional models (described in section 2.1) is stopped when the computed RMSE from the leave-one-out procedure stops decreasing by more than 0.01ºC or increases. The resulting intercept $\alpha(.)$ and regression coefficient $\beta(., .)$ for the two models are illustrated in Figure 3.3.

Figure 3.2 : Stream and air temperature (Potomac River, Years 2013-2014-2015)



The intercept is the same for both models as it represents the mean functional observation $\bar{y}(.)$. The β(s,t) surface for the FFLM (Figure 3.3a) shows that air temperature influences stream temperature particularly in the region $s = t \pm 15$. Also, the surface highlights a strong effect (in red) of air temperature around May 20th (Julian day (t) = 140). For the the HFLM (Figure 3.3b), the β(s,t) surface is restricted to the region $s < t$ and the effect is concentrated in the 10 days before $s = t$ and the effect gets stronger as the line $s = t$ is approached. The surface also highlights the particular effect of air temperature in late May, but also in late July (Julian day (t) = 210) and beginning of October (Julian day (t) = 280). The β(s,t) surfaces for the two models are close to 0 elsewhere.

Figure 3.3 : Regression coefficients for the functional models for the Potomac river

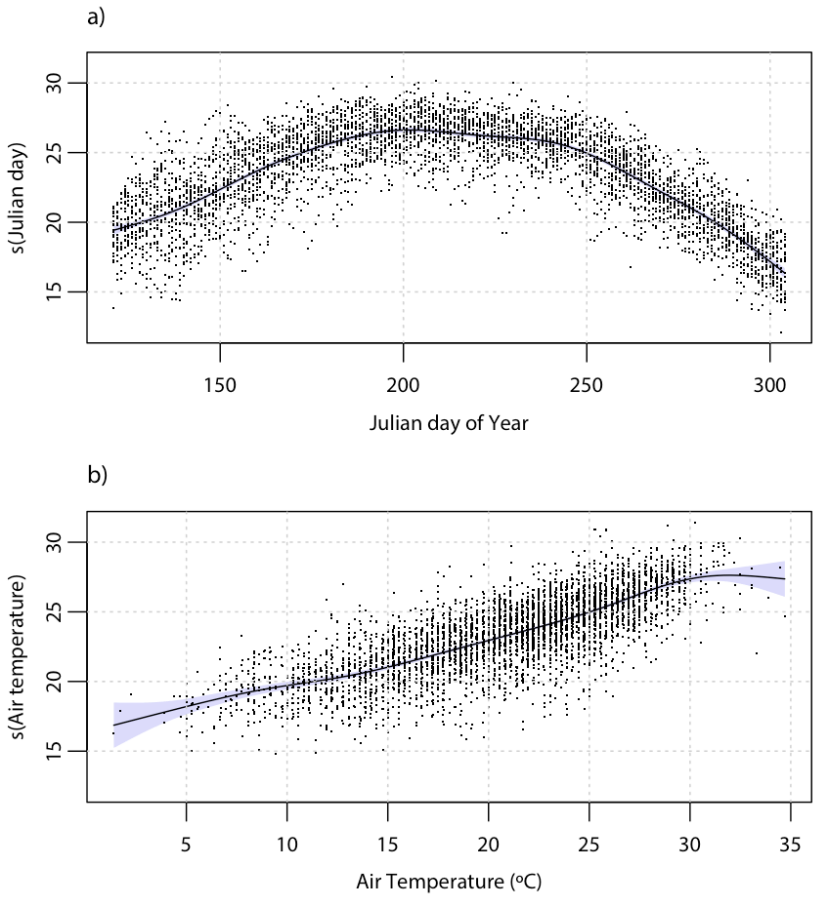


Regression coefficients for a) the fully functional linear model (FFLM) and b) the historical functional linear model (HFLM) (Potomac River). The color scale is the same for both.

Generalized additive model: The smooth effect functions obtained from the minimized generalized cross-validation are illustrated in Figure 3.4. In Figure 3.4a, the smooth function of the Julian day of year captures the seasonal variation in stream temperature, increasing from May to mid-July, and then decreasing. In Figure 3.4b, the effect of air temperature on stream temperature shows the non-linear effect for high and low temperature values as mentioned in the introduction. For high air temperature (>30ºC)

the effect smooth function stops increasing. Also, a small change of slope is observed in the 0ºC to 12.5ºC and 12.5ºC to 30ºC ranges. The confidence intervals are also shown. It is very narrow for any Julian day of year (not seeable on the figure), and it tends to be relatively large in the extremities of the air temperature effects (<5ºC and >30ºC).

Figure 3.4 : Estimated smoothing function for the generalized additive model for the Potomac River

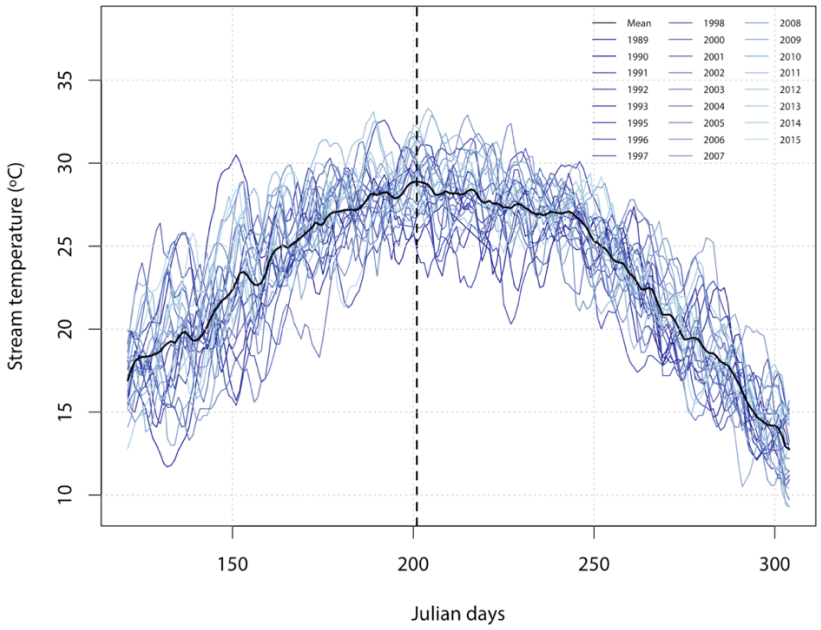


(a) Julian day of Year; (b) Air temperature

Logistic model: Optimal parameters are found by minimizing the sum of square errors for both the rising and falling limbs of the data. The day at which the maximum mean stream temperature occurs was used as a separation between the rising and the falling limbs. This day corresponds to the Julian day 201 (July 21st) in the case of the Potomac River, as illustrated in Figure 3.5. The resulting model equation is:

$$T_w(t) = \begin{cases} 32.48 + \dfrac{18.448}{1 + e^{0.188(21.09 - T_{a(t)})}} & , \quad t \leq 201 \\ 31.91 + \dfrac{20.400}{1 + e^{0.187(18.52 - T_a(t))}} & , \quad t > 201 \end{cases} \tag{3.15}$$

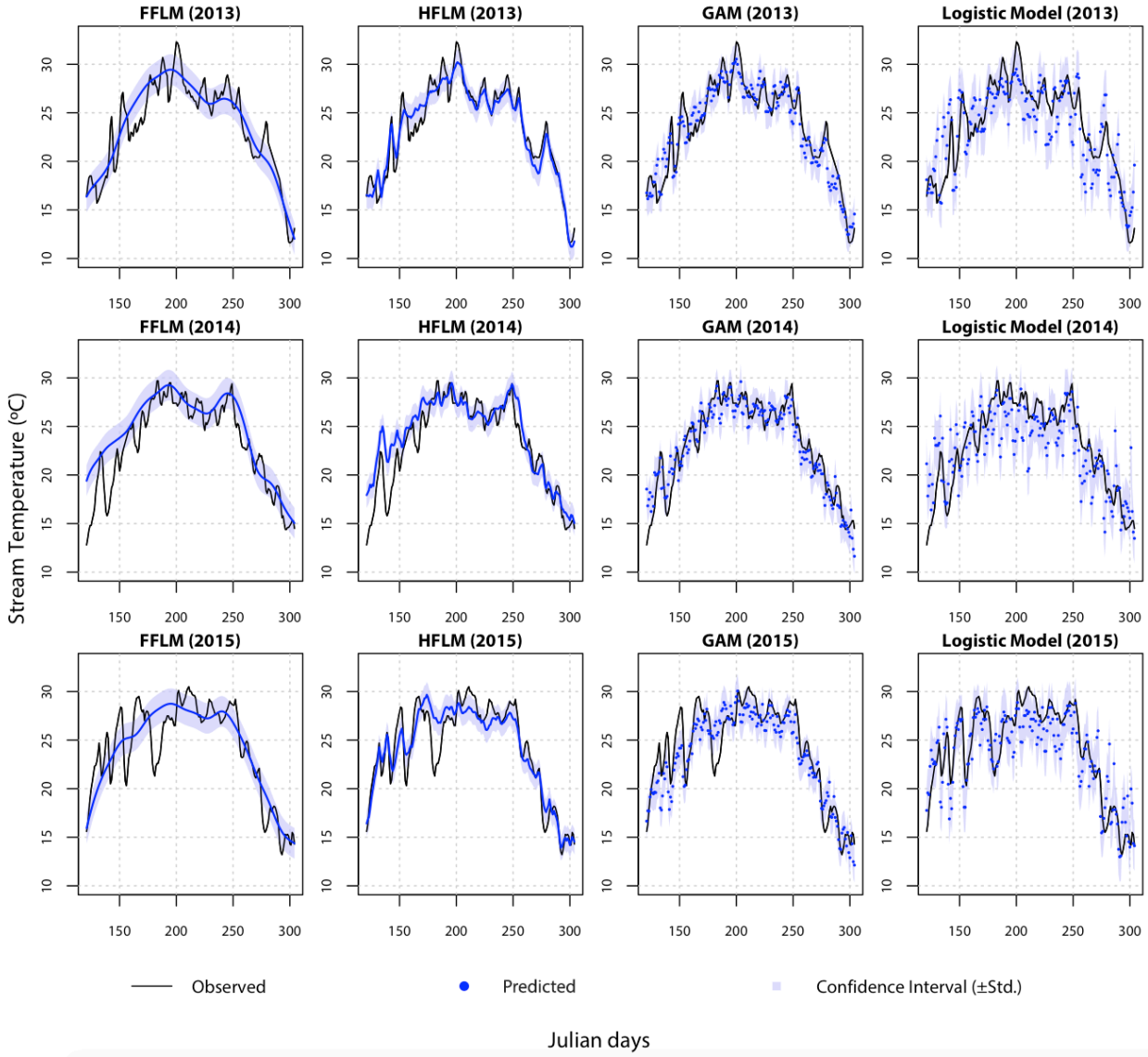
Figure 3.5 : Definition of the rising and falling limb for the logistic model for the Potomac River



3.3.3 Performance criteria

Predicted stream temperatures from the leave-one-out procedure and estimated standard deviation are illustrated for the last three years of the Potomac River in Figure 3.6 for the four models. For the FFLM and HFLM, a continuous curve of stream temperature is predicted for the entire season. The FFLM predicted well the stream temperature trend, but underestimated peak temperatures, while the HFLM appears to be more adapted to model maximum stream temperature for the three illustrated years. However, predictions from the GAM and the LM are plotted as dots as they are represented by multiple daily stream temperatures for the complete season. For these two models, the predicted values vary highly from day to day and tend to have higher standard deviations, especially for the LM. The functional models seem to reproduce more adequately the natural phenomenon of stream temperature heating and cooling than the two traditional models.

Figure 3.6 : Predicted stream temperatures from the Jack-Knife procedure for the four models (Potomac River, Years 2013-2014-2015)



The performance criteria calculated for daily predicted values are shown in Table 3.3. For the Potomac river, the RMSE calculated for the FFLM is 2.05ºC, while it decreases to 1.62ºC for the HFLM. For the GAM, an RMSE of 1.78ºC was found while the LM led to the highest RMSE (2.54ºC). Bias is generally 0ºC for the four models, but its range is the lowest for the HFLM with a standard deviation of ±0.75ºC. Finally, the NSC were respectively 0.70, 0,85, 0.81, and 0.88 for LM, GAM, FFLM and HFLM. Results for the

Missouri River leads to the same conclusions than for the Potomac. For the Delaware river, GAM was slightly better than the HFLM (RMSE of 2.21ºC compared to 2.06ºC).

Figure 3.7 summarizes the RMSE calculated for the four models for the three rivers with standard deviation. The LM provided the highest RMSEs (worst fit) followed by the FFLM for the three rivers whereas the HFLM is the best for the Potomac and the Missouri rivers, and the GAM for the Delaware. Same conclusions are derived when considering the NSC and the standard deviation of the Bias (Table 3.3).

Figure 3.7 : RMSE (with plus and minus one standard deviation) comparison for the four models and the three rivers. The dot indicates the mean values and the red indicates the model with the lower RMSE for each river.

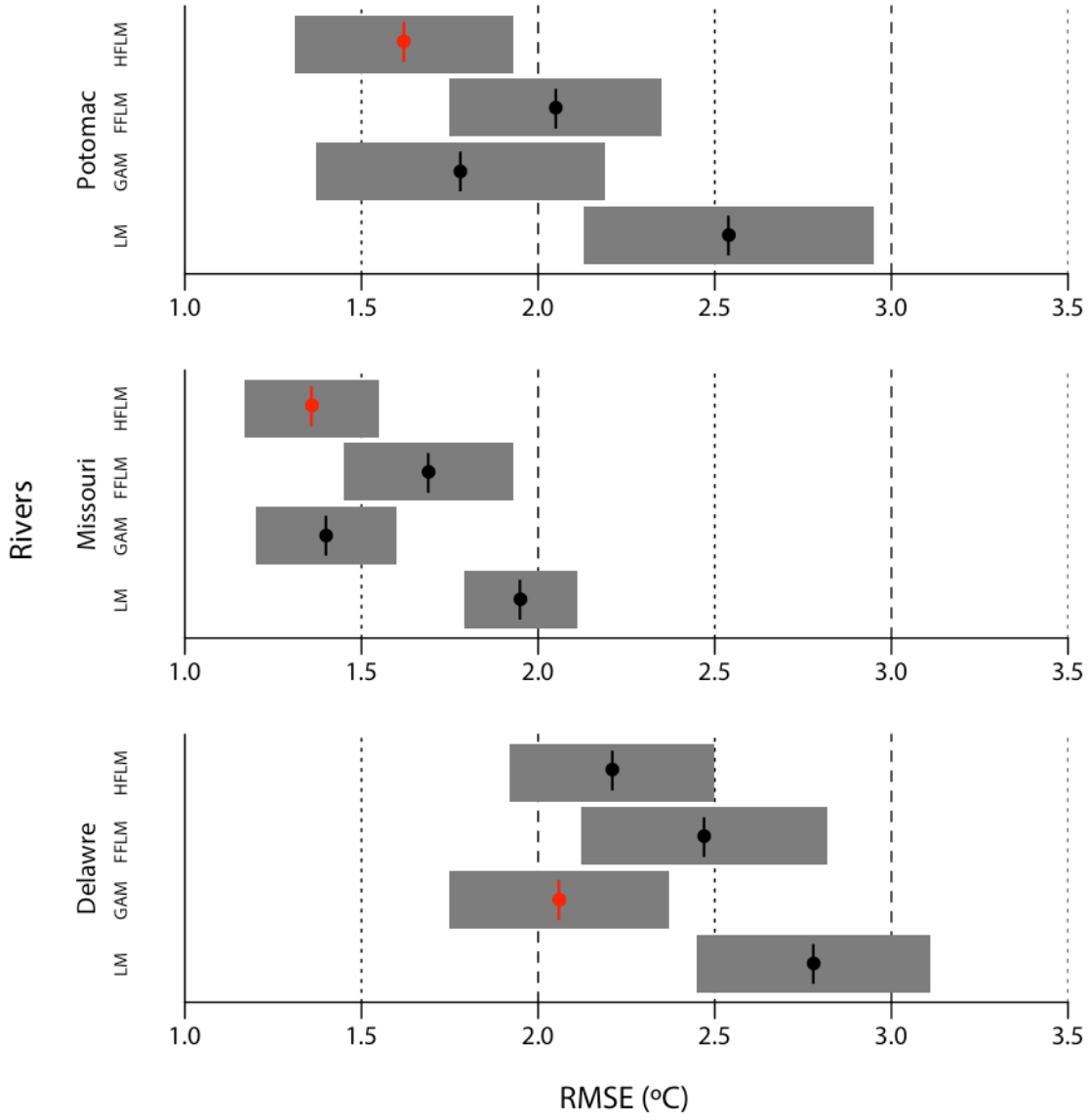


Tableau 3.3 : Classical performance measures

Rivers		Models											
		LM			GAM			FFLM			HFLM		
		RMSE (ºC)	Bias (ºC)	NSC	RMSE (ºC)	Bias (ºC)	NSC	RMSE (ºC)	Bias (ºC)	NSC	RMSE (ºC)	Bias (ºC)	NSC
Potomac	Mean	2.54	0	0.7	1.78	0	0.85	2.05	0	0.81	**1.62**	0	**0.88**
	(± Std)	(±0.41)	(±0.88)	(±0.14)	(±0.41)	(±0.89)	(±0.1)	(±0.3)	(±1.03)	(±0.07)	(±0.31)	**(±0.75)**	(±0.06)
	Range	1.81 - 3.83	-1.36 - 2.59	0.2 - 0.88	1.39 - 3.12	-1.50 - 2.61	0.47 - 0.93	1.5 - 2.78	-1.55 - 1.98	0.66 - 0.88	1.11 - 2.33	-1.25 - 1.81	0.71 - 0.94
Missouri	Mean	1.95	0	0.82	1.4	0	0.9	1.69	-0.03	0.86	**1.36**	-0.01	**0.91**
	(± Std)	(±0.16)	**(±0.49)**	(±0.05)	(±0.2)	(±0.53)	(±0.04)	(±0.24)	(±0.57)	(±0.06)	(±0.19)	(±0.52)	(±0.04)
	Range	1.62 - 2.34	-0.84 - 1.09	0.7 - 0.88	1.14 - 1.89	-0.76 - 1.02	0.79 - 0.95	1.27 - 2.23	-1.03 - 1.22	0.71 - 0.93	0.97 – 1.83	-0.95 - 1.07	0.81 - 0.96
Delaware	Mean	2.78	0	0.65	**2.06**	0	**0.8**	2.47	-0.04	0.71	2.21	-0.02	0.77
	(± Std)	(±0.33)	(±1.06)	(±0.1)	(±0.31)	**(±0.95)**	(±0.09)	(±0.35)	(±1.13)	(±0.12)	(±0.29)	(±1.04)	(±0.1)
	Range	2.18 – 3.44	-1.77 - 2.61	0.37 - 0.79	1.52 – 2.63	-1.80 - 1.95	0.54 - 0.91	1.82 – 3.11	-2.13 - 2.09	0.45 - 0.89	1.62 - 2.78	-1.76 - 2.05	0.52 - 0.9

The bold indicates the lower RMSE and the higher NSC among the four models for each river. For the Bias, the smaller standard deviation was put in bold.

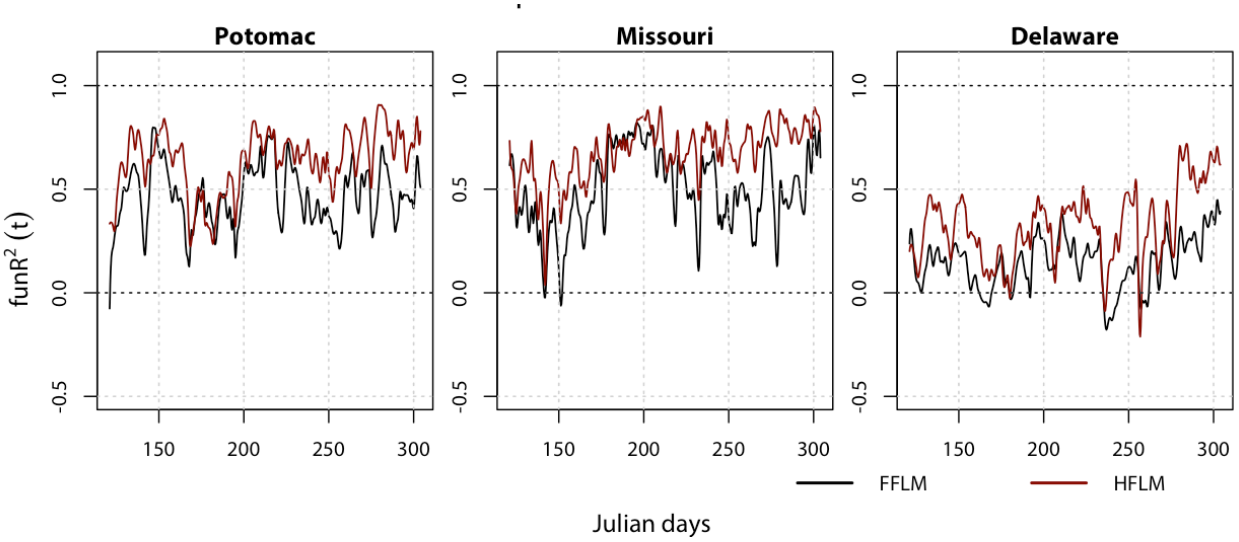
The functional performance criteria (Table 3.4) are then used to compare the functional regression models in between and access goodness-of-fit of these models. The four functional criteria all lead to the conclusion that the HFLM is more suitable than the FFLM for the three rivers. For the squared correlation function $funR^2(t)$ (equation 11), it is illustrated in Figure 3.8 and shows how the two functional models perform across time. The Potomac and the Missouri rivers had higher values of $funR^2(t)$ than the Delaware river showing a better fit with the functional models with these two rivers. The plot also shows that the HFLM has a better explicative power than the FFLM for the three rivers, as the red curve (HFLM) is almost always higher than the black curve (FFLM). This can be explained by the fact that the FFLM may be over parametrized. By comparing the $\beta(s,t)$ surface of the FFLM (Figure 3.3a) and the HFLM (Figure 3.3b), the surface is only defined for $s < t$ for the HFLM, reducing the overfitting that can result from using the FFLM with only $n \sim 25$ functional observations. In the case considered here, the HFLM would be preferred from a meteorological point of view, as the air temperature at s can only affect the stream temperature if $s < t$, which is the case for the HFLM but not for the FFLM.

Tableau 3.4 : Functional performance criteria

Rivers		Models							
		FFLM				HFLM			
		funRMSE (ºC)	funBias (ºC)	funNSC	$funR^2$	funRMSE (ºC)	funBias (ºC)	funNSC	$funR^2$
Potomac	Mean	2.02	0	0.81	0.47	**1.66**	0	**0.87**	**0.54**
	(± Std)	(±0.3)	(±1.03)	(±0.07)		(±0.3)	**(±0.75)**	(±0.06)	
	Range	1.47 - 2.76	-1.54 - 1.97	0.66 - 0.89		1.13 - 2.38	-1.12 - 1.84	0.69 – 0.94	
Missouri	Mean	1.64	-0.03	0.87	0.49	**1.4**	-0.01	**0.9**	**0.68**
	(± Std)	(±0.25)	(±0.57)	(±0.06)		(±0.19)	**(±0.54)**	(±0.04)	
	Range	1.22 – 2.19	-1.03 - 1.21	0.72 – 0.93		0.97 – 1.87	-0.93 - 1.08	0.79 - 0.96	
Delaware	Mean	2.45	-0.04	0.71	0.14	**2.23**	-0.02	**0.76**	**0.32**
	(± Std)	(±0.36)	(±1.13)	(±0.12)		(±0.29)	**(±1.04)**	(±0.1)	
	Range	1.79 - 3.09	-2.15 - 2.08	0.44 - 0.89		1.65 - 2.74	-1.78 - 1.97	0.5 - 0.89	

The bold indicates the lower *fun*RMSE and the higher *fun*NSC and *fun*R^2. For the *fun*Bias, the smaller standard deviation was put in bold.

Figure 3.8 : Squared correlation function (*funR*^{2}(t)) for the FFLM (black line) and the HFLM (red line) for the three rivers



3.4 Discussion

Among the four models tested, the Missouri River had the the best performance, while the Delaware River had the worst ones as illustrated by RMSE values (Figure 3.7). The distance between the stream and the meteorological station (Table 3.2) does not appear to be a factor that influences the modelling results since the Missouri River has the farthest station (62.1 km) and had the best modelling results. The general information about the three rivers (Table 3.1) may explain why the air to stream temperature models are less efficient with the Delaware River. This station has the smallest drainage area, while the Missouri River and Potomac have much higher values (up to 6 times a larger drainage area). Even if the Delaware drainage area is still a relatively large (5 232 km^2 at the gauging station), the stream temperature at this station is probably more likely to be influenced by variables other than the one used in this study (air temperature) like the thermal impact of tributaries, canopy cover or water abstraction (e.g. for the city of New York). The GAM achieves the best modelling results for this river (RMSE of 2.06ºC), but still not as good as the performance by the HLFM obtained on the two other rivers (RMSE of 1.62ºC and 1.36ºC). GAM performance in our study is comparable to that obtained elsewhere on the Sainte-Marguerite river (e.g. RMSE of 1.36ºC and NSC of 0.91, Laanaya et al., 2017).

Non-parametric methods, such as the *k*-nearest neighbors model (KNN) based on water temperature at *t*-1 and *t*-2 of St-Hilaire et al. (2012) applied to the Moisie River (RMSE of 1.57ºC), showed similar performance to the HFLM model results of our study (RMSE of 1.62ºC and 1.36ºC on the Potomac and the Missouri rivers respectively). One should note here that the KNN is based on past stream temperatures while the functional model only relies on air temperature to model stream temperature. Even though, those results are better than the ones we achieved for the Delaware river. Also, we use rivers with very long time series compared to the ones of the Sainte- Marguerite river or the Moisie river.

In addition to better performance, functional models show clear advantages of modelling a whole curve of temperatures instead of multiple daily predicted values in the other classical models. First, the confidence interval for a complete season of predicted values is constant with respect to time, while it typically increases when using time series approach (Box et al., 2015). Second, functional regression leads to a parametric model easier to understand and to interpret compared to non-parametric approaches, often called “black-box models” such as k-nearest neighbours or artificial neural networks (not compared in this article). Third, the use of air temperature curve as predictor corrects the problem of collinearity when the air temperature at different time steps is used in other regression models. However, a limitation of the functional models is that they require several years of data (~25) to be calibrated. If such data are available, and they will likely become more and more available in the actual context of big data and given the increased interest for the thermal regime of rivers, the functional regression model will surely become a very useful tool and this explains the growing interest in the last years for functional regression.

Finally, we conclude on how this model could be used for forecasting a complete season of future stream temperature. Given the current emphasis in meteorology on short to medium term forecasts, an initial implementation of functional regression in forecast mode would have to use the historical climatology or an output from a global/regional climate models as a predictor. Using this as input in a fitted functional regression model will give us a first rough forecast of the stream temperature for the next year (season). Then, as the air temperature is observed and forecasted for shorter timesteps, we would replace our first guess by the real observed or operationally forecasted air temperature. This will

then adjust the predictions that are made for different time-scales, especially for the short-term period of 10-15 days, as shown by the functional regression surface. This could be done for a particularly interesting period for commercial or recreational fisheries thermally sensitive species, for example. This approach can be repeated operationally to provide daily updated predictions of stream temperature curves to managers as the air temperature observation/forecast database is being built.

3.5 Conclusion

Two functional models were compared with two classical approaches to model the stream temperature. The functional models were able to model a complete season of stream temperature as a continuous curve compared to the classical model with multiple daily predicted values. To valid the proposed approach, three rivers were modeled to test the performance of functional models in different scenarios. The fully functional linear model (FFLM) and the historical functional linear model (HFLM) were tested and the latter was found to have better fit based on classical and functional performance criteria. The two functional models clearly outperform the LM while the GAM outperformed the FFLM. By comparing the GAM to the HFLM, the latter was found to be better on two of the three rivers. The GAM outperformed HFLM on the last river but still had a high RMSE value (>2ºC) in the case of daily stream temperature modelling. Hence, functional regression, especially the HFLM, seems to be a promising way to model stream temperature with only one predictor. Besides their conceptual advantages, they show good results on daily mean stream temperature modelling compared to classical models. Hence, they can give good indicator to stream manager of stream temperatures over a complete season and ultimately, in forecasting stream temperature for the future year.

Acknowledgement

The authors wish to acknowledge the financial contribution of Natural Sciences and Engineering Research Council of Canada (scholarship for the lead author) and to the Fonds de recherche du Québec - Nature et technologies (FRQNT). All data used in this study are freely available and can be downloaded directly from the NOOA (for air

temperatures) and USGS (for stream temperatures) websites (links available in the following references). No conflict of interest is reported.

References

AHMADI-NEDUSHAN, B., ST-HILAIRE, A., OUARDA, T. B., BILODEAU, L., ROBICHAUD, E., THIEMONGE, N. & BOBEE, B. 2007. Predicting river water temperatures using stochastic models: case study of the Moisie River (Quebec, Canada). *Hydrological Processes,* 21**,** 21- 34.

BEL, L., BAR-HEN, A., PETIT, R. & CHEDDADI, R. 2011. Spatio-temporal functional regression on paleoecological data. *Journal of Applied Statistics,* 38**,** 695-704.

BÉLANGER, M., EL-JABI, N., CAISSIE, D., ASHKAR, F. & RIBI, J. 2005. Estimation de la température de l'eau de rivière en utilisant les réseaux de neurones et la régression linéaire multiple. *Revue des sciences de l'eau/Journal of Water Science,* 18**,** 403-421.

BENYAHYA, L., CAISSIE, D., ST-HILAIRE, A., OUARDA, T. B. & BOBÉE, B. 2007a. A review of statistical water temperature models. *Canadian Water Resources Journal,* 32**,** 179-192.

BENYAHYA, L., ST-HILAIRE, A., OUARDA, T. B., BOBÉE, B. & DUMAS, J. 2008. Comparison of non- parametric and parametric water temperature models on the Nivelle River, France. *Hydrological sciences journal,* 53**,** 640-655.

BENYAHYA, L., ST-HILAIRE, A., QUARDA, T. B., BOBÉE, B. & AHMADI-NEDUSHAN, B. 2007b. Modeling of water temperatures based on stochastic approaches: case study of the Deschutes River. *Journal of Environmental Engineering and Science,* 6**,** 437-448.

BERNARDI, M. S., SANGALLI, L. M., MAZZA, G. & RAMSAY, J. O. 2017. A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. *Stochastic Environmental Research and Risk Assessment,* 31**,** 23-38.

BESCHTA, R., BILBY, R., BROWN, G. & HOLTBY, L. 1987. Stream temperature and aquatic habitat: pp. 191-232. *Fishery and forestry interactions. Streamside management: forestry and fishery interactions. University of Washington, Institute of Forest Resources. Contr,* 57.

BJORNN, T. & REISER, D. 1991. Habitat requirements of salmonids in streams. *American Fisheries Society Special Publication,* 19**,** 138.

BOSQ, D. 2012. *Linear processes in function spaces: theory and applications*, Springer Science & Business Media.

BOX, G. E., JENKINS, G. M., REINSEL, G. C. & LJUNG, G. M. 2015. *Time series analysis: forecasting and control*, John Wiley & Sons.

BROCKHAUS, S., MELCHER, M., LEISCH, F. & GREVEN, S. 2017a. Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing,* 27**,** 913-926.

BROCKHAUS, S., RUEGAMER, D., HOTHORN, T. & BROCKHAUS, M. S. 2017b. Package 'FDboost'.

BROCKHAUS, S., RÜGAMER, D. & GREVEN, S. 2017c. Boosting Functional Regression Models with FDboost. *arXiv preprint arXiv:1705.10662*.

BROCKHAUS, S., SCHEIPL, F., HOTHORN, T. & GREVEN, S. 2015. The functional linear array model. *Statistical Modelling,* 15**,** 279-300.

BÜHLMANN, P. & HOTHORN, T. 2007. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science***,** 477-505.

BUSTILLO, V., MOATAR, F., DUCHARNE, A., THIÉRY, D. & POIREL, A. 2014. A multimodel comparison for assessing water temperatures under changing climate conditions via the equilibrium temperature concept: case study of the Middle Loire River, France. *Hydrological Processes,* 28**,** 1507-1524.

CAISSIE, D. 2006. The thermal regime of rivers: a review. *Freshwater Biology,* 51**,** 1389-1406.

CAISSIE, D., EL-JABI, N. & SATISH, M. G. 2001. Modelling of maximum daily water temperatures in a small stream using air temperatures. *Journal of Hydrology,* 251**,** 14-28.

CAISSIE, D., EL-JABI, N. & ST-HILAIRE, A. 1998. Stochastic modelling of water temperatures in a small stream using air to water relations. *Canadian Journal of Civil Engineering,* 25**,** 250-260.

CHAOUCH, M. 2014. Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves. *IEEE Transactions on Smart Grid,* 5**,** 411-419.

CHEBANA, F., CHARRON, C., OUARDA, T. B. & MARTEL, B. 2014. Regional frequency analysis at ungauged sites with the generalized additive model. *Journal of Hydrometeorology,* 15**,** 2418-2428.

CHEBANA, F., DABO-NIANG, S. & OUARDA, T. B. 2012. Exploratory functional flood frequency analysis and outlier detection. *Water Resources Research,* 48.

CHENARD, J. F. & CAISSIE, D. 2008. Stream temperature modelling using artificial neural networks: application on Catamaran Brook, New Brunswick, Canada. *Hydrological Processes,* 22**,** 3361-3372.

CHIOU, J.-M. 2012. Dynamical functional prediction and classification, with application to traffic flow prediction. *The Annals of Applied Statistics***,** 1588-1614.

CIARLEGLIO, A., PETKOVA, E., TARPEY, T. & OGDEN, R. T. 2016. Flexible functional regression methods for estimating individualized treatment rules. *Stat,* 5**,** 185-199.

CLUIS, D. A. 1972. Relationship between stream water temperature and ambient air temperature. *Hydrology Research,* 3**,** 65-71.

CRISP, D. & HOWSON, G. 1982. Effect of air temperature upon mean water temperature in streams in the north Pennines and English Lake District. *Freshwater Biology,* 12**,** 359-367.

CUEVAS, A., FEBRERO, M. & FRAIMAN, R. 2002. Linear functional regression: the case of fixed design and functional response. *Canadian Journal of Statistics,* 30**,** 285-300.

DABO-NIANG, S. & FERRATY, F. 2008. *Functional and operatorial statistics*, Springer Science & Business Media.

DEWEBER, J. T. & WAGNER, T. 2014. A regional neural network ensemble for predicting mean daily river water temperature. *Journal of Hydrology,* 517**,** 187-200.

FALAH, F., GHORBANI NEJAD, S., RAHMATI, O., DANESHFAR, M. & ZEINIVAND, H. 2017. Applicability of generalized additive model in groundwater potential modelling and comparison its performance by bivariate statistical methods. *Geocarto International,* 32**,** 1069-1089.

FERRATY, F. & VIEU, P. 2006. *Nonparametric functional data analysis: theory and practice*, Springer Science & Business Media.

FREUND, Y., SCHAPIRE, R. & ABE, N. 1999. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence,* 14**,** 1612.

GERVINI, D. 2015. Dynamic retrospective regression for functional data. *Technometrics,* 57**,** 26-34.

GRBIĆ, R., KURTAGIĆ, D. & SLIŠKOVIĆ, D. 2013. Stream water temperature prediction based on Gaussian process regression. *Expert systems with applications,* 40**,** 7407-7414.

HANDELAND, S. O., IMSLAND, A. K. & STEFANSSON, S. O. 2008. The effect of temperature and fish size on growth, feed intake, food conversion efficiency and stomach evacuation rate of Atlantic salmon post-smolts. *Aquaculture,* 283**,** 36-42.

HAREZLAK, J., COULL, B. A., LAIRD, N. M., MAGARI, S. R. & CHRISTIANI, D. C. 2007. Penalized solutions to functional regression problems. *Computational statistics & data analysis,* 51**,** 4911-4925.

HASTIE, T. & TIBSHIRANI, R. 1990. *Generalized additive models*, Wiley Online Library.

IDDRISU, W. A., NOKOE, K. S., LUGUTERAH, A. & ANTWI, E. O. 2017. Generalized Additive Mixed Modelling of River Discharge in the Black Volta River. *Open Journal of Statistics,* 7**,** 621.

IVANESCU, A. E., STAICU, A.-M., SCHEIPL, F. & GREVEN, S. 2014. Penalized function-on-function regression. JANSSEN, P. & HEUBERGER, P. 1995. Calibration of process-oriented models. *Ecological Modelling,* 83**,** 55-66.

JEPPESEN, E. & IVERSEN, T. M. 1987. Two simple models for estimating daily mean water temperatures and diel variations in a Danish low gradient stream. *Oikos***,** 149-155.

JOURDONNAIS, J., WALSH, R., PICKETT, F. & GOODMAN, D. 1992. Structure and calibration strategy for a water temperature model of the lower Madison River, Montana. *Rivers,* 3**,** 153-169.

KAUSHAL, S. S., LIKENS, G. E., JAWORSKI, N. A., PACE, M. L., SIDES, A. M., SEEKELL, D., BELT, K. T., SECOR, D. H. & WINGATE, R. L. 2010. Rising stream and river temperatures in the United States. *Frontiers in Ecology and the Environment,* 8**,** 461-466.

KELLEHER, C., WAGENER, T., GOOSEFF, M., MCGLYNN, B., MCGUIRE, K. & MARSHALL, L. 2012. Investigating controls on the thermal sensitivity of Pennsylvania streams. *Hydrological Processes,* 2*6***,** 771-785.

KIM, K., ŞENTÜRK, D. & LI, R. 2011. Recent history functional linear models for sparse longitudinal data. *Journal of statistical planning and inference,* 141**,** 1554-1566.

KWAK, J., ST-HILAIRE, A. & CHEBANA, F. 2017. A comparative study for water temperature modelling in a small basin, the Fourchue River, Quebec, Canada. *Hydrological Sciences Journal,* 62**,** 64-75.

LAANAYA, F., ST-HILAIRE, A. & GLOAGUEN, E. 2017. Water temperature modelling: comparison between the generalized additive model, logistic, residuals regression and linear regression models. *Hydrological Sciences Journal,* 62**,** 1078-1093.

LARABI, S., ST-HILAIRE, A., CHEBANA, F. & LATRAVERSE, M. 2017. Multi-Criteria Process-Based Calibration Using Functional Data Analysis to Improve Hydrological Model Realism. *Water Resources Management***,** 1-17.

LEE, R. M. & RINNE, J. N. 1980. Critical thermal maxima of five trout species in the southwestern United States. *Transactions of the American Fisheries Society,* 109**,** 632-635.

MACKEY, A. & BERRIE, A. 1991. The prediction of watertemperatures in chalk streams from air temperatures. *Hydrobiologia,* 210**,** 183-189.

MAHEU, A., CAISSIE, D., ST-HILAIRE, A. & EL-JABI, N. 2014. River evaporation and corresponding heat fluxes in forested catchments. *Hydrological Processes,* 28**,** 5725-5738.

MAHEU, A., ST-HILAIRE, A., CAISSIE, D., EL-JABI, N., BOURQUE, G. & BOISCLAIR, D. 2016. A regional analysis of the impact of dams on water temperature in medium-size rivers in eastern Canada. *Canadian Journal of Fisheries and Aquatic Sciences,* 73**,** 1885-1897.

MALFAIT, N. & RAMSAY, J. O. 2003. The historical functional linear model. *Canadian Journal of Statistics,* 31**,** 115-128.

MASSELOT, P., DABO-NIANG, S., CHEBANA, F. & OUARDA, T. B. 2016. Streamflow forecasting using functional regression. *Journal of Hydrology,* 538**,** 754-766.

MCDONALD, S., KOULIS, T., EHN, J., CAMPBELL, K., GOSSELIN, M. & MUNDY, C. 2015. A functional regression model for predicting optical depth and estimating attenuation coefficients in sea-ice covers near Resolute Passage, Canada. *Annals of Glaciology,* 56**,** 147-154.

MCLEAN, M. W., HOOKER, G., STAICU, A.-M., SCHEIPL, F. & RUPPERT, D. 2014. Functional generalized additive models. *Journal of Computational and Graphical Statistics,* 23**,** 249- 269.

MEYER, M. J., COULL, B. A., VERSACE, F., CINCIRIPINI, P. & MORRIS, J. S. 2015. Bayesian function- on-function regression for multilevel functional data. *Biometrics,* 71**,** 563-574.

MOHSENI, O. & STEFAN, H. 1999. Stream temperature/air temperature relationship: a physical interpretation. *Journal of hydrology,* 218**,** 128-141.

MOHSENI, O., STEFAN, H. G. & ERICKSON, T. R. 1998. A nonlinear regression model for weekly stream temperatures. *Water Resources Research,* 34**,** 2685-2692.

MORIN, G., FORTIN, J.-P., LARDEAU, J.-P., SOCHANSKA, W. & PAQUETTE, S. 1981. *Modèle CEQUEAU: manuel d'utilisation*, INRS-eau.

MORRILL, J. C., BALES, R. C. & CONKLIN, M. H. 2005. Estimating stream temperature from air temperature: implications for future water quality. *Journal of Environmental Engineering,* 131**,** 139-146.

MORRIS, J. S. 2015. Functional regression. *Annual Review of Statistics and Its Application,* 2**,** 321- 359.

MORRISON, J., QUICK, M. C. & FOREMAN, M. G. 2002. Climate change in the Fraser River watershed: flow and temperature projections. *Journal of Hydrology,* 263**,** 230-244.

NASH, J. E. & SUTCLIFFE, J. V. 1970. River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology,* 10**,** 282-290.

NETER, J., KUTNER, M. H., NACHTSHEIM, C. J. & WASSERMAN, W. 1996. *Applied linear statistical models*, Irwin Chicago.

NOAA. 2017. *National centers for environmental information* [Online]. https://www.ncdc.noaa.gov/cdo-web/. [Accessed].

PILGRIM, J. M., FANG, X. & STEFAN, H. G. 1998. Stream temperature correlations with air temperatures in Minnesota: implications for climate warming. *JAWRA Journal of the American Water Resources Association,* 34**,** 1109-1121.

PIOTROWSKI, A. P., NAPIORKOWSKI, M. J., NAPIORKOWSKI, J. J. & OSUCH, M. 2015. Comparing various artificial neural network types for water temperature prediction in rivers. *Journal of Hydrology,* 529**,** 302-315.

PREUD'HOMME, E. B. & STEFAN, H. G. 1992. Relationship between water temperatures and air temperatures for Central US Streams.

QUENOUILLE, M. H. 1949. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological),* 11**,** 68-84.

R CORE TEAM 2017. R language definition. *Vienna, Austria: R foundation for statistical computing*.

RAHMAN, A., CHARRON, C., OUARDA, T. B. & CHEBANA, F. 2018. Development of regional flood frequency analysis techniques using generalized additive models for Australia. *Stochastic Environmental Research and Risk Assessment,* 32**,** 123-139.

RAMSAY, J. 1982. When the data are functions. *Psychometrika,* 47**,** 379-396.

RAMSAY, J. O. 2006. *Functional data analysis*, Wiley Online Library. RAMSAY, J. O., HOOKER, G. & GRAVES, S. 2009. *Functional data analysis with R and MATLAB*, Springer Science & Business Media.

SEGURA, C., CALDWELL, P., SUN, G., MCNULTY, S. & ZHANG, Y. 2015. A model to predict stream water temperature across the conterminous USA. *Hydrological processes,* 29**,** 2178-2195.

SIGHOLT, T. & FINSTAD, B. 1990. Effect of low temperature on seawater tolerance in Atlantic salmon (Salmo salar) smolts. *Aquaculture,* 84**,** 167-172.

ST-HILAIRE, A., OUARDA, T. B., BARGAOUI, Z., DAIGLE, A. & BILODEAU, L. 2012. Daily river water temperature forecast model with a k-nearest neighbour approach. *Hydrological Processes,* 26**,** 1302-1310.

STEFAN, H. G. & PREUD'HOMME, E. B. 1993. Stream temperature estimation from air temperature. *JAWRA Journal of the American Water Resources Association,* 29**,** 27-45.

TERNYNCK, C., BEN ALAYA, M. A., CHEBANA, F., DABO-NIANG, S. & OUARDA, T. B. 2016. Streamflow hydrograph classification using functional data analysis. *Journal of Hydrometeorology,* 17**,** 327-344.

USGS. 2017. *USGS Daily Values Web Service* [Online]. https://waterservices.usgs.gov/rest/DV-Test-Tool.html. [Accessed].

WEBB, B., CLACK, P. & WALLING, D. 2003. Water–air temperature relationships in a Devon river system and the role of flow. *Hydrological processes,* 17**,** 3069-3084.

WEBB, B. & WALLING, D. 1993. Temporal variability in the impact of river regulation on thermal regime and some biological implications. *Freshwater Biology,* 29**,** 167-182.

WEBB, B. W., HANNAH, D. M., MOORE, R. D., BROWN, L. E. & NOBILIS, F. 2008. Recent advances in stream and river temperature research. *Hydrological processes,* 22**,** 902-918.

WEHRLY, K. E., BRENDEN, T. O. & WANG, L. 2009. A comparison of statistical approaches for predicting stream temperatures across heterogeneous landscapes. *JAWRA Journal of the American Water Resources Association,* 45**,** 986-997.

WOOD, S. 2006. *Generalized additive models: an introduction with R*, CRC press.

WOOD, S. 2015. Package 'mgcv'. *R package version*, 1-7.

ZHANG, Q., GU, X., SINGH, V. P., XIAO, M. & CHEN, X. 2015. Evaluation of flood frequency under non-stationarity resulting from climate indices and reservoir indices in the East River basin, China. *Journal of Hydrology,* 527, 565-575.

4 ARTICLE 2 : MODÉLISATION DE LA SÉLECTION D'HABITAT PAR LE SAUMON ATLANTIQUE JUVÉNILE EN UTILISANT LA RÉGRESSION FONCTIONNELLE

Modelling Habitat Selection by Juvenile Atlantic Salmon using Functional Regression

J. Boudreault[a,b*], N. E. Bergeron[a,b], A. St-Hilaire[a,b,c] and F. Chebana[a,b]

[a] Institut National de la recherche scientifique – Centre Eau Terre Environnement, Québec, Canada ;

[b] Laboratoire d'Analyse et de Modélisation de l'Habitat Aquatique (LAMHA), Québec, Canada ;

[c] Canadian River Institute, University of New Brunswick, Fredericton, Canada.

Manuscript ready to be submitted

August 8th 2018

*Corresponding author: Jeremie Boudreault (jeremie.boudreault ete.inrs.ca)

Abstract

Fish habitat is impacted by increased human activities on watercourses and climate changes. Hence, models to assess habitat quality are developed. In classical habitat models, the habitat variables, often depth, velocity and substrate size, are discrete measurements leading to serious loss of information on fish habitat. In this study, complete frequency distributions (curves) are used for each habitat variable as measured in fish nearby habitat as well as including the water temperature as a new predictor. The functional regression model (FRM) allows to model fish presence-absence or abundance by these curves. Hence, this may overcome current challenges in habitat modelling such as low predictive power or lack of transferability by representing more naturally fish habitat. To consider the proposed approach, the Sainte-Marguerite River (Quebec, Canada) was surveyed at 26 sites for juvenile Atlantic salmon (*Salmo Salar*). FRMs are developed for the three ages of juvenile salmon (0+, 1+, 2+) and compared to a generalized linear model and a generalized additive model. The FRM gave more insights on habitat selection (preferred range of each habitat variable) and had the greatest accuracy based on a cross-validation. Furthermore, a transferability checking was performed with data from the Petite Cascapedia River (Quebec, Canada) and the functional presence-absence model looks the most promising for transferability between rivers. The use of the FRM in habitat modelling is innovative as it represents more thoroughly the fish habitat with curves of availability and promising as it gives interpretable coefficients and performs better than other regression models.

Keywords: Juvenile Atlantic salmon, river ecology, fish habitat modelling, functional regression model, generalized additive model, generalized linear model

Résumé

L'habitat du poisson est impacté par des activités anthropiques plus fréquentes sur les cours d'eau et par les changements climatiques. Ainsi, des modèles pour évaluer la qualité de l'habitat sont développés. Dans les modèles classiques, les variables d'habitat, souvent la profondeur de l'eau, la vitesse du courant et la taille du substrat, sont des mesures discrètes menant à une sérieuse perte d'information sur l'habitat du poisson. Dans cette étude, des distributions complètes de fréquences (des courbes) sont utilisées pour chaque variable d'habitat telles que mesurées dans l'habitat immédiat du poisson, en plus d'ajouter la variable de la température de l'eau comme nouveau prédicteur. Le modèle de régression fonctionnelle (MRF) permet de modéliser la présence-absence ou l'abondance de poissons via ces courbes. Ainsi, cela pourrait régler certains problèmes connus en modélisation de l'habitat comme le faible pouvoir prédictif ou le manque de transférabilité en représentant plus naturellement l'habitat du poisson. Pour considérer l'approche proposée, la rivière Sainte-Marguerite (Québec, Canada) a été étudiée à 26 sites pour les saumons atlantiques juvéniles (*Salmo Salar*). Des MRFs ont été développés pour les trois âges du saumon juvénile (0+, 1+, 2+) et comparés à un modèle linéaire généralisé et un modèle additif généralisé. Le MRF donne plus d'information sur l'habitat sélectionné (gamme de valeurs choisie par le poisson) et est le plus performant basé sur une validation croisée. De plus, la transférabilité du modèle a été étudiée avec des données provenant de la rivière Petite-Cascapédia (Québec, Canada) et le modèle fonctionnel de présence-absence avait le meilleur potentiel de transférabilité entre les rivières. L'utilisation du MRF en modélisation de l'habitat est innovante comme le MRF représente plus adéquatement l'habitat du poisson avec des courbes de disponibilité et prometteuse comme il donne des coefficients interprétables et performe mieux que les approches classiques.

Mots-clés: Saumon atlantique juvénile ; écologie en rivière ; modélisation de l'habitat du poisson ; modèle de régression fonctionnelle ; modèle additif généralisé ; modèle linéaire généralisé

4.1 Introduction

Fish habitat in many rivers has been subjected to several alterations in last decades due to multiple human activities in or close to many watercourses such as flow regulation, road construction, deforestation and industrial activities (Heggenes, 1990; Gibson, 1993; Heggenes *et al.*, 1995; Bardonnet & Baglinière, 2000; Prévost *et al.*, 2002; Nyqvist *et al.*, 2017; Sundt-Hansen *et al.*, 2018). In addition, climate change will likely impact both hydrological and thermal regimes of rivers, which could result in a decrease of fish population, especially salmonids (Armstrong *et al.*, 2003; Tetzlaff *et al.*, 2005; Jonsson & Jonsson, 2009; Elliott & Elliott, 2010; Daigle *et al.*, 2015). Given all these changes in its habitat, its population has suffered from significant decrease in the last decades (Noakes *et al.*, 2000; Lackey, 2003). Hence, to quantify these changes, models that assess fish habitat quality have been developed and are continuously being improved (Ahmadi-Nedushan *et al.*, 2006; Yi *et al.*, 2017).

Juvenile salmon habitat is often studied through preference curves that calculate a habitat suitability index (HSI) (Leclerc *et al.*, 1994; Guay *et al.*, 2000; Mäki-Petäys *et al.*, 2002; Guay *et al.*, 2003; Hedger *et al.*, 2004; Hedger *et al.*, 2006). This approach consists in determining fish preference over the range of available physical habitat characteristics, usually including depth, flow velocity and substrate size using an index varying from 0 to 1. HSI are subsequently used to calculate a weighted usable area (WUA) that can be linked to fish presence/absence, abundance or density. While some studies have found a significant correlation between WUA and fish density (Orth & Maughan, 1982; Boudreau *et al.*, 1996; Bovee *et al.*, 1998), others have found no significant relationship (Scott & Shirvell, 1987; Bourgeois *et al.*, 1996). The physical habitat simulation model, PHABISM, based on the preference curves approach, was strongly criticized in the last decade (Railsback, 2016). Firstly, the calculated WUA is outdated as it has no biological meaning and simpler quality/quantity metrics such as density or abundance should be used (Manly *et al.*, 2007). Also, the hypothesis that physical variables act independently and have equal effects on habitat suitability (ranging from 0 to 1) may introduce errors in HSI estimation (Orth & Maughan, 1982). Moreover, focal measures of physical habitat variables (e.g. at the “fish nose”) may not be fully representative of fish habitat and thus

could introduce error in estimation of the preference curves (Beecher *et al.*, 2010), which can greatly impact predicted HSI (Ayllón *et al.*, 2012). Finally, the preference curves developed on a certain river have low transferability to other rivers (Scott & Shirvell, 1987; Freeman *et al.*, 1997; Mäki-Petäys *et al.*, 2002; Guay *et al.*, 2003; Hedger *et al.*, 2004).

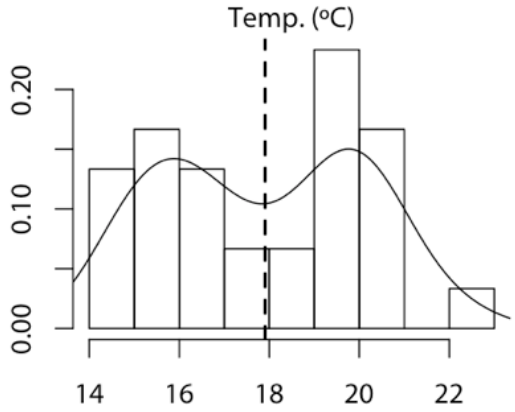
Hence, regression based habitat models have emerged, linking directly habitat variables (i.e. predictors) to fish presence-absence and/or abundance/density (Ahmadi-Nedushan *et al.*, 2006). The logistic regression, a special case of the generalized linear model (GLM) with a logit link function, was used for juvenile salmon habitat modelling with a greater performance (Guay *et al.*, 2000) and transferability (Guay *et al.*, 2003) than the traditional HSI approach. The GLM was used in a few fish habitat modelling studies (Labonne *et al.*, 2003; Ahmadi-Nedushan *et al.*, 2006), and recently applied for presence-absence of Chinook salmon fry (Beakes *et al.*, 2014). In Hedger *et al.* (2005), the juvenile salmon density was modelled using a generalized additive model (GAM), with the advantage of allowing smooth non-linear effects of the habitat variables on fish density. Millidine *et al.* (2016) also used a GAM to model fry abundance. Besides regression models, the fuzzy logic based on experts' knowledge was also used in fish habitat modelling (Jorde *et al.*, 2001; Ahmadi-Nedushan *et al.*, 2008; Mocq *et al.*, 2013). The limiting factor to the application of fuzzy logic is that the number of fuzzy rules increases exponentially with the number of predictors (Ahmadi-Nedushan *et al.*, 2006).

While Railsback (2016) has suggested developing new data-driven approaches to assess fish habitat quality and quantity, not much effort has been made in recent years to develop such approaches. Considering all models mentioned above and the ones used in the literature, the fish habitat is always described by single values of depth, substrate size and velocity. Using only one value of each predictor per habitat patch causes serious loss of information on fish habitat. In addition, other important habitat variables such as water temperature are lacking. For example, Hedger *et al.* (2005) used a mean value based on three to ten measurements for the predictors (depth, velocity and substrate size) at each site to describe juvenile salmon habitat. Using the mean, of course, implies that there is no indication on variability of habitat conditions. In doing so, information on fish habitat is lost, which may explain the relatively low explicative power of the GAM tested in this study ($R2$ of 28% for fry density and 47% for parr density). Moreover, Hedger *et al.* (2006)

compared the mean substrate size calculated at each site to spot measurements taken at the fish nose and found significantly better result for mean values. Even if this study suggests that the mean value at the site better represents juvenile salmon habitat than a spot (e.g. “at the nose”) measurement, the use of the mean substrate size may not be a representative measure of the site substrate composition therefore limiting habitat description. Hence, the proposed idea is to include more information in fish habitat models by using probability density functions (PDF) for each predictor that better represent the natural habitat of fish than an aggregated value of these measurements.

In this context, it is useful to introduce **functional data analysis (FDA)**, a statistical framework able to manipulate curves or functions in contrast with scalar or vector (discrete measures) in classical contexts. FDA was first suggested by Ramsay (1982) and has become very popular in the last decade (Ferraty & Vieu, 2006; Ramsay, 2006; Dabo-Niang & Ferraty, 2008; Ramsay *et al.*, 2009; Bosq, 2012; Horváth & Kokoszka, 2012) with various applications in neuroscience (McLean *et al.*, 2014; Ivanescu *et al.*, 2015; Meyer *et al.*, 2015; Ciarleglio *et al.*, 2016), energy (Goia *et al.*, 2010; Chaouch, 2014; Brockhaus *et al.*, 2015), ecology (Bel *et al.*, 2011; Stewart- Koster *et al.*, 2014; McDonald *et al.*, 2015) and hydrology (Chebana *et al.*, 2012; Masselot *et al.*, 2016; Ternynck *et al.*, 2016; Larabi *et al.*, 2017; Requena *et al.*, 2018). One main advantage of the FDA framework is the possibility to use functions/curves in regression models called **functional regression models (FRM)**. In the case presented here, a FRM can model the presence-absence/abundance of juvenile Atlantic salmon using PDFs representing available depths, velocities and substrate sizes that fully describe the salmon habitat as well as including water temperature as an additional predictor. FRM, in addition to representing more adequately fish habitat, will handle naturally heterogeneous habitat compared to classical approaches based on a mean value. For example, Figure 4.1 shows how a PDF will capture a heterogeneous habitat with a bimodal distribution of temperatures with warmer temperatures in the main river stem and a colder temperature indicating a potential cold thermal refuge for fish. A classical model will lose this information by only taking the mean temperature value for this site.

Figure 4.1 : Example of a site on the river with a bimodal distribution of temperatures



This site shows peaks at 16ºC (e.g. cold tributary) and 20ºC (e.g. main channel), whereas the mean is approximately at 18ºC (dashed line)

To assess the added value and performance of the proposed approach, the FRM is evaluated and compared to two regression models based on mean values previously used in salmon habitat modelling: GLM (Guay *et al.*, 2000; Guay *et al.*, 2003; Beakes *et al.*, 2014) and GAM, as a generalization of the GLM with smooth non-linear transformations of the predictors (Hedger *et al.*, 2005; Millidine *et al.*, 2016).

The next section introduces the methodology and the case study. The results are in section 3. Section 4 provides a discussion, while the conclusion is in section 5.

4.2 Material and methods

The next section presents the statistical models, their fitting and performance measures. Then, the case study is presented.

4.2.1 Statistical models

Functional data analysis and the functional regression model: A first step to use FRM is to recover the probability density functions from the histograms of observed data. To simplify this problem, the function $x(.)$ to be found is often expressed as linear combination of M basis functions as below (Ramsay, 2006):

$$x(s) = \sum_{m=1}^{M} c_m \, \phi_m(s) \tag{4.1}$$

where c_m are the coefficients to be estimated and $\phi_m(s)$ are known basis function. Basis function can be either b-splines for non-periodic data or Fourier series for periodic data (Ramsay, 2006). Equation (4.1) is the classical way to get a continuous function from observed data, but the FDA can handle any kind of functions. In our case, the functions will be PDFs estimated through empirical kernel density estimates (KDE) to obtain smooth non-parametric PDF, thereby allowing the function to have many shapes (Tukey, 1977). The KDE is defined as:

$$\hat{f}_h(w) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{w - w_i}{h}\right) \tag{4.2}$$

where w_1, …, w_n are univariate independent and identically distributed random variables with unknown probability density function $f(.)$, K is the kernel (often the standard normal density function) and h is a smoothing parameter called the bandwidth that can be adjusted for smoothness (Tukey, 1977). The KDE can be obtained easily with the function *density* of R (R Core Team, 2017).

Once the KDE have been constructed from the measured habitat variables at each site, they can be readily used in a FRM. Many variants of the FRM exist and a literature review is available in (Morris, 2015). For the case of modelling fish habitat using the PDFs of each predictor, the **functional linear model for scalar response** (denoted FLMS) will be considered (Ramsay, 2006). It aims at explaining a scalar response variable (e.g. fish presence-absence or abundance) with functions/curves as predictors (e.g. PDFs for each habitat variable at the site). The FLMS is defined as:

$$g(E(y_i)) = \alpha + \int_{\Omega} \beta(s) \, x_i(s) ds \tag{4.3}$$

where $E(y_i)$ is the expected value of the response variable at site i, g is the link function, $x_i(s)$ is the predictor curve (a PDF observed on the domain Ω) and α is the intercept. The $\beta(s)$ is the regression coefficient giving the impact of $x(.)$ at s on $g(E(y_i))$ and it is defined

in the same manner as $x(s)$ in equation (4.1) (Ramsay, 2006). Model (4.3) can be easily adapted for multiple predictors as follows:

$$g(E(y_i)) = \alpha + \sum_{j=1}^{p} \int_{\Omega_j} \beta_j(s)\, x_{i,j}(s) ds \tag{4.4}$$

where p is the number of predictors.

To model the presence-absence of juvenile salmons at each site, the logit function for dichotomous variables (0 or 1) is used as done in past studies (Guay *et al.*, 2000; Feyrer *et al.*, 2007):

$$g(E(y_i)) = \log\left(\frac{p_i}{1 - p_i}\right) \tag{4.5}$$

where $p_i = \Pr[y_i = 1 \mid \boldsymbol{X}]$ is the probability to observe a fish at site *i.* This probability is also called the habitat probabilistic index (HPI) (Guay *et al.*, 2000). As the HSI, the HPI ranges between 0 and 1. It can be understood as a HSI (1 is the preferred habitat while 0 is the avoided habitat), but calculated from a regression model instead of the classic preference curves approach.

Finally, if instead, the abundance of juvenile salmons at each site is modelled, the logarithmic function is used as a link function (Millidine *et al.*, 2016):

$$g(E(y_i)) = \log(E[y_i]) \tag{4.6}$$

where $E[y_i]$ is the expected abundance at site *i*. These link functions (4.5) and (4.6) will also be used in the GLM and the GAM presented in the next two sections.

Generalized linear model : The GLM is a generalization of the multiple regression model allowing for non-normality (Neter et al., 1996). It is defined as :

$$g(E(y_i)) = \alpha + \sum_{j=1}^{p} \beta_j\, \bar{x}_{i,j} \tag{4.7}$$

where $\bar{x}_{i,j}$ is the mean value of predictor *j* for the site *i* and β_j is the regression coefficient for predictor *j*. Here and in the GAM presented in the next section, the use of the complete distribution of frequency (i.e. the PDF) of each predictor is impossible and hence the mean measured value for each site is used as done in past studies (Hedger *et al.*, 2005; Hedger *et al.*, 2006).

Generalized additive model : In the GAM approach (Hastie & Tibshirani, 1990), a smooth transformation of each predictor is linked to the response variable. It is defined as :

$$g(E(y_i)) = \alpha + \sum_{j=1}^{M} f_j(\bar{x}_{i,j}) \tag{4.8}$$

where f_j is the smooth non-linear function (often a combination of cubic splines) applied to $\bar{x}_{i,j}$. For example, Hedger *et al.* (2005) used cubic splines with 4 degrees of freedom for depth, velocity and substrate size to model fry and parr density.

4.2.2 Fitting of Models

To calibrate the different models, all analysis are conducted in R (R Core Team, 2017) For the GLM in (4.7), it is fitted with the function *glm*. For the FLMS (4.4) and the GAM (4.8), they are both fitted with the FDboost package (Brockhaus et al., 2017b). The GAM is calibrated using cubic splines with 4 degrees of freedom for smooth transformation of the predictors as suggested in Hedger et al. (2005). The regression coefficients for the FLMS and the GAM are found by a machine learning iterative procedure using a component-wise gradient boosting algorithm, allowing to fit multiple predictors (Bühlmann & Hothorn, 2007; Brockhaus et al., 2017a). At each iteration, the algorithm adjusts the parameters to minimize a loss function. To fit data with a binomial distribution (i.e. presence-absence data), the loss function is the negative likelihood of the binomial distribution whereas it is the negative likelihood of the negative binomial distribution for abundance data (Brockhaus et al., 2017c). Optimal parameters can be found by stopping the algorithm before full convergence, called *early stop*. In doing so, it avoids overfitting of the regression coefficients and leads to more regularized effects (Brockhaus et al.,

2017c). Hence, the optimal number of iterations is found by a leave-one-out procedure when error stops decreasing or increases (Brockhaus et al., 2017a).

4.2.3 Models performance

To assess performance of the three models tested in this study, the models will be calibrated and validated with three different approaches. For the first approach called "Full model validation", the calibration will use all available sites and validation will be performed on these same sites. For the second validation approach called the "cross-validation", either a leave-one-out cross-validation (for abundance models) or a 5-fold cross-validation (for presence-absence models) is performed. For the leave-one-out validation, the model is calibrated on all available data except that one site is left out. The model is then used to predict abundance on the site removed from the calibration and this is repeated for each site. For the 5-fold cross-validation, the dataset is separated into five equal sized subsamples. One of the five subsamples is removed prior to fit the model. Then, the fitted model is used to predict presence-absence on the sites of the removed subsample. This is repeated for the five subsamples. Cross-validation procedures are often used with statistical models to test whether a model calibrated on a certain dataset will work on data not used for the calibration (Shao, 1993). For the third validation called "Transferability validation", the full model will be used but validated with data from another river. This is to test whether a model calibrated on a certain river will work on another river with possibly different characteristics as it is a well-known problem in fish habitat modelling.

To assess models performance for models based on fish abundance, the Nash-Sutcliffe model efficiency coefficient (*NSC*) is calculated (Nash and Sutcliffe, 1970):

$$NSC = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4.9}$$

where y_i and $\hat{y}_i$ are respectively the observed and predicted abundance of fish at site *i* and *n* is the number of sites. A *NSC* of 1 represents a perfect fit and a *NSC* of 0 or below denotes a model not better than the mean value. It is used to evaluate the efficiency of abundance models from a leave-one-out cross-validation procedure or with data from

another river. Note that the *NSC* is equivalent to the R^2 if it is calculated on the same data used in the calibration.

The Nagelkerke pseudo R^2 (denoted *PseudoR*^{2}) is a homologous measure of R^2 for presence-absence data defined as a ratio of log-likelihood function (Nagelkerke, 1991) :

$$PseudoR^2 = \frac{1 - \left[\frac{-2\ LLF_{reduced}}{-2\ LLF_{full}}\right]^{2/n}}{1 - [-2\ LLF_{reduced}]^{2/n}} \tag{4.10}$$

where $LLF_{reduced}$ is the log-likelihood of a non-informative model (with no predictor), LLF_{full} is the log-likelihood of the tested model. A high value corresponds to a great improvement compared to the non-informative model. The *PseudoR*^{2} is comparable to the R^2 (it also ranges from 0 and 1), but it does not indicate the percentage of variance explained by the model. To calculate the *PseudoR*^{2}*,* the likelihood function (LLF) (Neter *et al.*, 1996) is calculated as follows:

$$LLF = \sum_{i=1}^{n} [\ y_i \ln(\hat{p}_i) + \ (1 - \ y_i) \ln(1 - \hat{p}_i))] \tag{4.11}$$

where y_i is the observed presence-absence and $\hat{p}_i = Pr(\hat{y}_i = 1 \,|\boldsymbol{X})$ is the predicted probability to observe a presence by the model.

Finally, an adjusted measure of R^2 (denoted R^2*adj*) will also be calculated to assess the added value of water temperature as a new predictor in fish habitat models. Based on the traditional Pearson R^2 (or *PseudoR*^{2} in the case of presence-absence data), the R^2*ajd* is calculated as:

$$R^2_{adj} = \ 1 - (1 - R^2)\frac{(n-1)}{(n-k-1)} \tag{4.12}$$

where *n* is the number of observations and *k* is the number of predictors. The R^2*ajd* takes in account the number of predictors and hence can be used to check model parsimony compared to the R^2 that always increase as the number of predictor increases.

Apart from goodness-of-fit, three performance criteria for diagnostic tests (models with binary response) will be calculated, previously used to assess performance of the

presence-absence model for chinook salmon fry by Beakes *et al.* (2014). The first one is the accuracy (denoted *ACU*) which quantifies the proportion of cases for which the model is right on predicting whether juvenile salmon were present or absent at a certain site. The second metric refers to sensibility or true presence rate (denoted *TPR*) and is the proportion of cases where the model predicted presence when there is actual presence of fish. The third metric refers to specificity and is called the true absence rate (denoted *TAR*) and gives the proportion of correctly predicted absence.

$$ACU = \frac{TP + TA}{P + A} \tag{4.13}$$

$$TPR = \frac{TP}{P} \tag{4.14}$$

$$TAR = \frac{TA}{A} \tag{4.15}$$

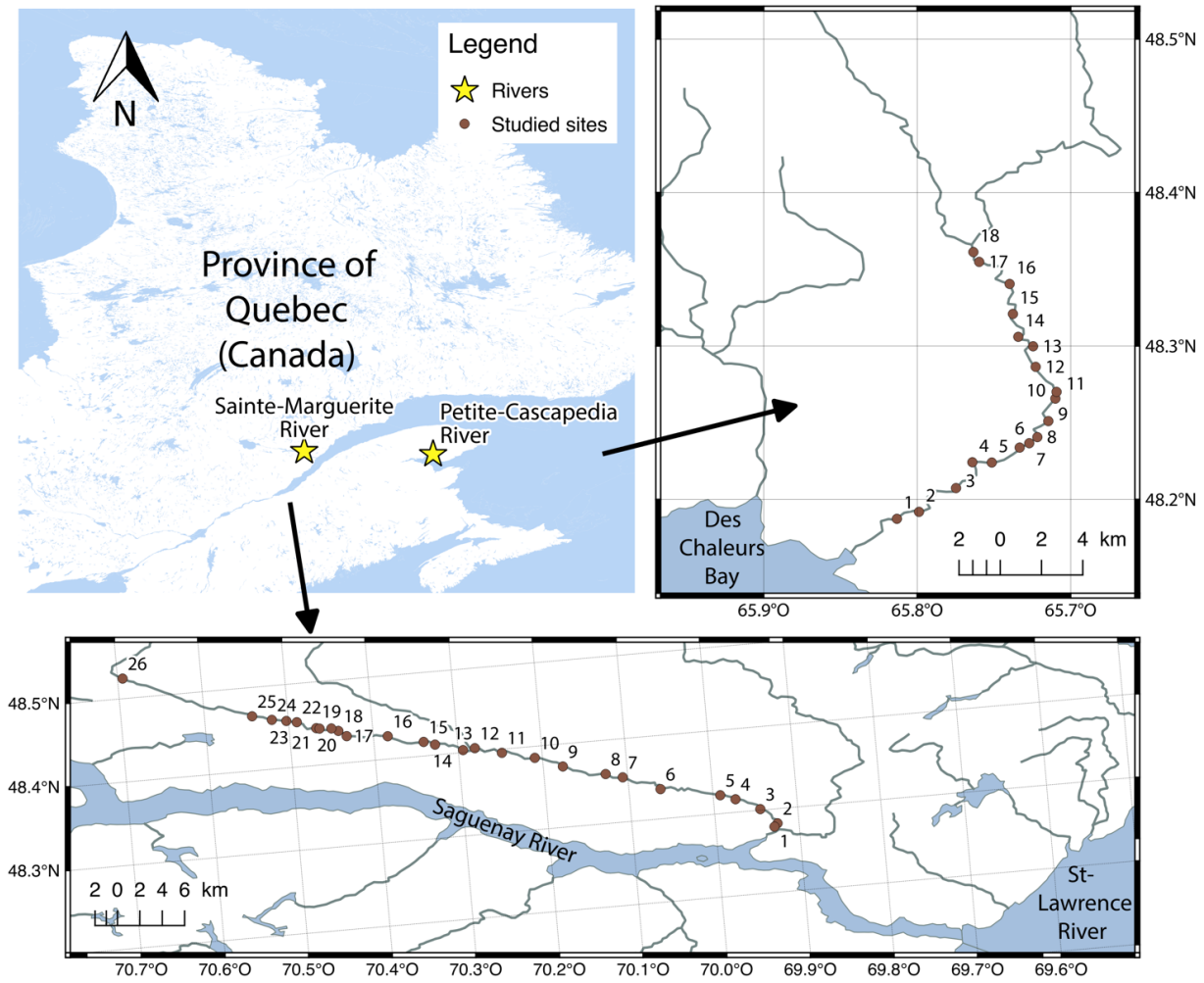
where *TP* is the number of correct presence (true presence) detected by the model, *TA* is the number of correct absence (true absence) detected by the model and *P* and *A* are the respective number of observed presence and absence in the sample. All those metrics are proportion between 0 and 1, with high values associated with high model predictability. To compare the model output to observed presence/absence, a threshold must be selected to convert continuous probabilities of presence to a dichotomous (0-1) series of presence and absence. Hence, this conversion is done in an objective manner by taking a threshold that maximizes the good classifications on a validation dataset (Liu *et al.*, 2005b). The optimal threshold is found for each model on the 5-fold cross-validation when the number of good classifications is maximized using the R package *Epi* (Carstensen et al., 2015). Once these thresholds are obtained, they can also be used to convert predicted probabilities of presence on another river to presence-absence data and to calculate performance measure. Finding optimal thresholds is essential for a future operational application of the model. Without thresholds, models will not be able to classify sites with absence and sites with presence if true observations are not measured. In addition to *ACU*, *TPR* and *TAR*, the area under the receiver operating characteristic

curve (*AUC*) will also be calculated as a measure of performance considering all possible thresholds, computed with the R package *Epi* (Carstensen et al., 2015).

2.4 A case study on juvenile Atlantic salmon habitat

To test and validate the proposed approach, two salmon rivers in the province of Quebec (Canada) were surveyed during summer 2017. The first river is the Sainte-Marguerite River (SMR) and flows into the Bay Sainte-Marguerite in the Saguenay Fjord. The second river is the Petite-Cascapedia River (PCR) located in the region of Gaspésie. It empties into the Bay des Chaleurs. The map of Figure 4.2 illustrates the location of the two rivers and the surveyed sites on these rivers.

Figure 4.2 : Map of the rivers surveyed and studied sites on each river



The choice of these rivers was motivated by the various past studies on their segmentation in sedimentary links (Davey, 2005; Davey & Lapointe, 2007; Bouchard & Boisclair, 2008; Kim, 2009; Johnston & Bergeron, 2010; Kim & Lapointe, 2011; Lanthier *et al.*, 2014), which guaranteed that the sites were representing a large variability of fish habitat, well suited for habitat modelling (Rice & Church, 1998; Rice *et al.*, 2001).

For the SMR (the calibration river), 26 sites were surveyed between July 27th and August 16th. All sites were separated by more than 500 meters along the river to ensure independence between sites. At each site, 30 patches of 4m^2 equally distributed along 5 transects (6 patches per transect) were surveyed (see Figure 4.3 for an illustration of the sampling method). First, each parcel was fished using a Smith-Root LR24 Electrofisher. The electrofisher was automatically calibrated with the "Quick Set-up" option before each site on a small patch outside the surveyed site. Voltage was then increased or decreased (by step of 50V) based on fish reaction until optimal reaction was obtained (no death and an optimal recovery in a bucket filled with stream water). Once the electrofisher was set, each of the 30 patches of 4m^2 was electrofished during a constant amount of time (fishing time per patch was measured at 54.0 seconds with a standard error of 3.6 seconds). The operator of the electrofisher remained the same during all field surveys to ensure that the fishing time was constant. Each collected fish was noted, but only the juvenile salmon were measured to fork length using a handheld ruler. Then, fishes were immediately returned to water in the downstream direction to ensure that the same fish was not caught twice. Also, at each patch, physical measurements of fish habitat were taken. The flow velocity at 0.4 of the depth was taken by an acoustic velocity meter (Sontek Flow Tracker 2). The temperature was recorded by this same instrument and the depth was noted. A visual estimation of the median substrate size of the *b*-axis (denoted D_{50}) for each 4m^2 patch was made by a method comparable to the one developed by Latulippe *et al.* (2001). The three operators calibrated themselves on a selected 4m^2 patch and tried to estimate the D_{50}. After, the real D_{50} was calculated and provided to operators so that they can adjust themselves if they underestimate or overestimate the real value. This was repeated on a new patch until each operator can guess the correct D_{50} with tolerance of 10%. The associated D_{50} for each patch was then obtained when the 3 operators agreed with the given value. Finally, the location was taken for each site by a handheld GPS (Garmin

eTrex 20x). Table 4.1 summarizes the main characteristics of the surveyed sites. For the PCR (the validation river), 30 sites were surveyed between September 1st and September 16th with the same methodology as the one described for the SMR, but only 18 are kept for the validation because the temperature was significantly colder than on the SMR for some of the PCR sites, as sampling on this river occurred later in the fall.

Figure 4.3 : Survey design at each site

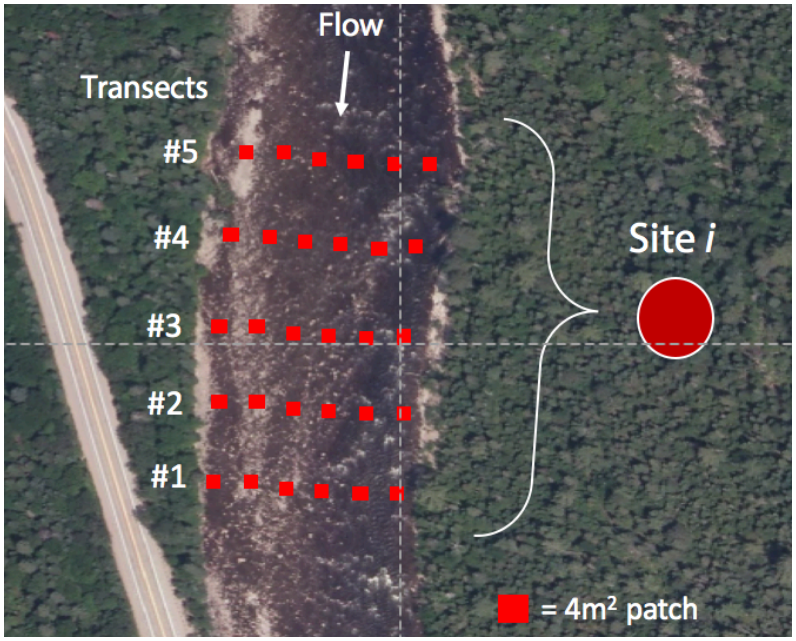


Tableau 4.1 : Main characteristics of the studied sites on the two rivers

Rivers	Location	Number of sites	Habitat variables (Mean value)				Mean abundance of juvenile Atlantic salmon per site		
			Depth (cm)	Velocity (m/s)	D50 (mm)	Tempera -ture (ºC)	Fry (0+)	Parr (1+)	Parr (2+)
Sainte-Marguerite River (SMR)	Sacre-Coeur, Saguenay (Qc) Canada	26	41.0 (±11.1)	0.37 (±0.13)	77.9 (±41.1)	17.4 (±2.5)	10.6 (±9.9)	5.0 (±4.3)	1.4 (±1.8)
Petite Cascapedia River (PCR)	New Richmond, Gaspesie (Qc) Canada	18	35.7 (±7.5)	0.43 (±0.12)	50.6 (±10.43)	13.5 (±0.9)	3.1 (±3.5)	4.9 (±3.7)	1.6 (±1.6)

Measured juvenile salmons were then classified by age using fork length (0+ for fry, 1+ and 2+ for parr). Mean abundance per site for the two rivers are reported in Table 4.1.

For all sites of the SMR, at least one fry and one 1+ parr were captured and models for abundance are hence developed. However, it was noted that 50% of the sites have absence of 2+ parr in the SMR. A model based on abundance will not be suitable for such low density and these abundances were transferred to presence-absence data to minimize effect of high abundance sites as suggested in Feyrer *et al.* (2007). Table 4.2 summarizes which response variable (i.e. presence/absence or abundance) for each age is modelled.

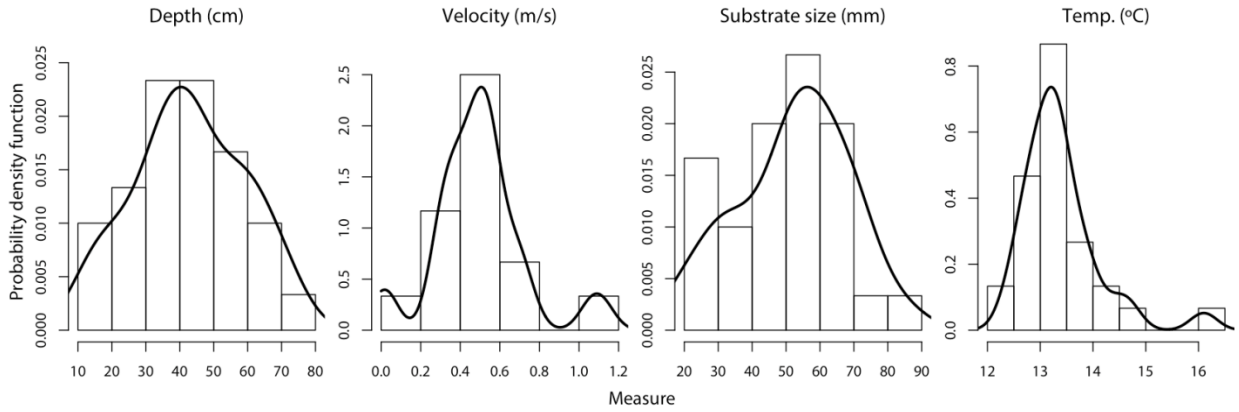
Tableau 4.2 : Motivation of the choice the modelled response variable for the three ages

Age	Proportion of the surveyed sites on the SMR with no fish observed	Abundance range	Modelled response variable (y_i)		Explanation
			Abundance	Presence-Absence	
Fry (0+)	0%	1 – 48	X		No absence noted
Parr (1+)	0%	1 – 20	X		
Parr (2+)	50%	0 – 6		X	50% of the sites with absence and high abundances at some sites (6 fish)

4.3 **Results**

The KDE (estimates of PDFs) are obtained by first plotting the histograms of available depth, velocity, median substrate size (D_{50}) and temperature for each site. Then, kernel density estimates are fitted to the sample histograms. To ensure the coherence of the density estimates with the observed histogram, they were visually inspected. The number of knots was set to 128 in order to capture all the variability by the KDE, especially for heterogeneous sites (i.e. with bimodal distribution). For example, Figure 4.4 shows the resulting KDE for site 1 of the SMR. Once obtained, the KDE curves are ready to be used in the FLMS.

Figure 4.4 : Obtained kernel density estimates (KDE) with 128 knots for the site #1 of the SMR



The black lines illustrate the KDE while the histogram is plotted for fitting check.

In the next sections, results for the 2+ parr presence-absence model are first shown for illustration purpose in terms of regression coefficients effects and performance. Then, results for models of fry and 1+ parr abundances are presented. Finally, the added value of the water temperature variable in habitat models is presented.

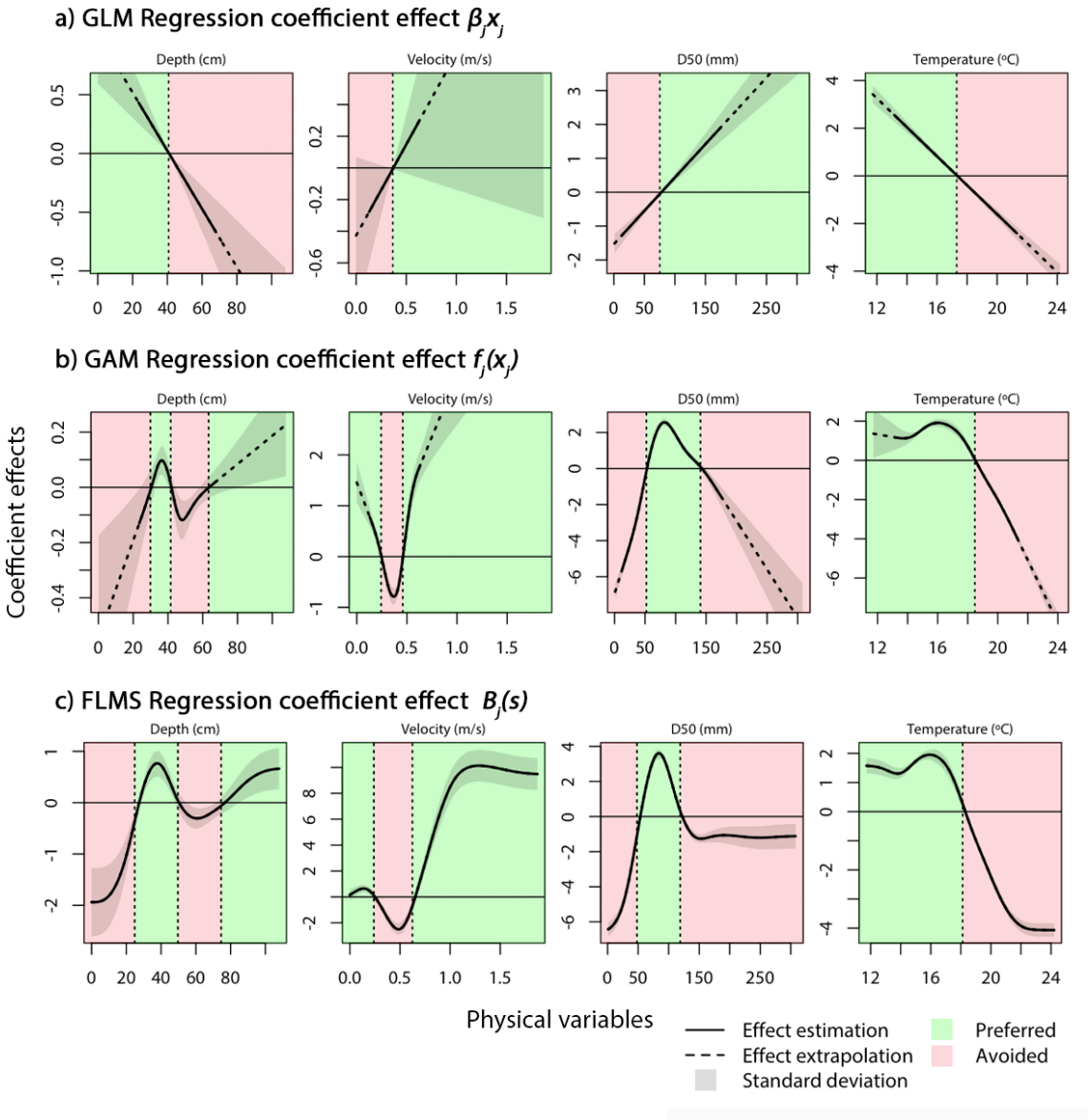
4.3.1 Results for 2+ parr presence-absence model

Resulting regression effects: Fitted effect coefficients/curves of the three considered models are presented in Figure 4.5. This figure shows how each value of the variable affect the logit transformation of the probability to observe a 2+ parr at the site. For the three models, the standard deviation of each effect is also calculated based on the leave-one-out estimates of each effect. In the case of the GLM, the velocity estimate has intervals that merge into the 0 line, probably showing a non-significant effect. Once tested with an F-test, the hypothesis that the most parsimonious model (without the velocity) is better can be rejected, so the velocity was kept in the GLM (tolerance level of 0.10; *p-value* of 0.067). All predictors are significant for the GAM and the FLMS.

The GLM imposes a linear relation between the predictors and the logit probability. Hence, this model gives one preferred (in green) and avoided (in red) region for each predictor. For example, low depths (<40 cm) are preferred compared to higher depths (>40 cm). For the GAM, a more appropriate function, $f_j(.)$, links each of the predictors to the logit probability. This non-linear relationship shows multiple preferred ranges for each

predictor. For example, two ranges of velocities (<0.2 m/s and >0.5 m/s) are more likely selected by 2+ parr. This may be explained because salmon (as other fish species) have multiple biological activities to accomplish and these different activities require different types of habitat (e.g. feeding in higher-velocity habitat vs seeking refuge in slower flowing water). Finally, in the FLMS, the coefficient $\beta_j(.)$ varies over the domain of $x_j(.)$, allowing also to account for non-linear effects and multiple ranges of habitat preference even if the FLMS is a linear model.

Figure 4.5 : Regression coefficient effects for the three models



From Figure 4.5, the coefficient curves from the FLMS are defined on a wider domain than the ones for the GLM and the GAM This is because the coefficients are estimated from PDFs for each predictor instead of a mean value, letting the model capture the effect of each predictor even in the extreme values. One can see that the coefficient effects in the extreme value for the FLMS are more regularized than the ones of the GAM. For example, the GAM coefficients for the velocity indicate that the velocity increases, the probability to observe a fish will also increase. In the case of the FLMS, the effect is stable for high velocities (i.e. not increasing with velocities higher than 1 m/s). Figure 4.5c also shows that this stability of the FLMS effects for high values is also observed when substrate is coarse (>175 mm) and temperature is high (>20ºC). This is an important advantage of the FLMS since it avoids direct extrapolations which are usually inappropriate with the GAM.

Goodness-of-fit and performance: Once the three models are fitted, validation criteria are calculated in Table 4.3. The best *PseudoR*^{2} value is obtained by the FLMS (0.62) followed closely by the GAM (0.60). The GLM obtains 0.08 meaning that it is only slightly better than a non-informative model, while both the GAM and the FLMS are much better with greater values of *PseudoR*^{2}. This may be explained because of the various non-linear effects of the predictors on the logit probability not modelled in the GLM, as stated above.

To assess models performance, thresholds of 0.43, 0.12 and 0.15 are found respectively for GLM, GAM and FLMS to convert probabilities into presence-absence data as described in section 4.2.3. For the “full model validation”, the GLM obtains 76.9% for the *ACU*, *TPR* and *TAR* while both the GAM and the FLMS obtain the same results: 96.2% for the *ACU*, 100% for the *TPR*, 92.3% for the *TAR* and a *AUC* of 1.00. These results suggest that these two last models are maybe over-fitted. Hence, the “5-fold cross-validation” is done to check this possible issue. While the GLM remains at 76.9% for the *ACU*, *TPR* and *TAR*, the performance criteria for the GAM decrease respectively to 73.1%, 92.3% and 53.8% while the FLMS has 84.6%, 92.3% and 76.9%. The latter has the best results for these three criteria. Also, the *AUC* is the highest for the FLMS (0.82), while it is lower for the GAM (0.77) and the GLM (0.67). For the “Transferability validation”, the model calibrated on the SMR is tested with 18 surveyed sites on the PCR with the

same thresholds than the ones found with the cross-validation. Both the GLM and the GAM are unable to detect absence of salmon (they both obtain a *TAR* of 0% on the PCR). The GLM obtains a relatively good *ACU* (66.7%) because it predicted presence of salmon at all sites (in fact, the true proportion of presence was 66.7% of the sites on the PCR). The same conclusions are found for the GAM, except that the GAM predicted correctly 83.3% of the presences versus 100% for the GLM. Over-predicting presence is problematic and likely means that these two models show little potential for transferability as they are unable to classify good and poor sites (i.e. sites with presence and absence). The FLMS obtains a *ACU* of 61.1% and it is the model that looks the most promising to detect both presences (*TPR* of 75%) and some of the absences (*TAR* of 33.3%). Even if these metrics are lower on the PCR that the ones calculated on the SMR where the model is calibrated, they show clearly an advantage of using the FLMS on new rivers. However, the *AUC* for the transferability validation is higher for the GAM (0.58) than the FLMS (0.54). In a real context of using a habitat model on a new river, observed fish presence-absence will not be available and hence, a threshold will have to be selected (as we did to calculate *ACU*, *TPR* and *TAR*). Therefore, the *AUC* as a performance measure for the transferability check may be misleading here.

Tableau 4.3 : Goodness-of-fit and performance criteria for the 2+ parr presence-absence model

Validation type	Models	Performance Criteria				PseudoR^2
		***ACU* (%)**	***TPR* (%)**	***TAR* (%)**	***AUC***	
Full model	GLM	76.9	76.9	76.9	0.83	0.08
	GAM	**96.2**	**100**	**92.3**	**1.00**	0.60
	FLMS	**96.2**	**100**	**92.3**	**1.00**	**0.62**
5-fold cross-validation	GLM	76.9	76.9	**76.9**	0.67	
	GAM	73.1	**92.3**	53.8	0.77	
	FLMS	**84.5**	**92.3**	**76.9**	**0.82**	
Transferability	GLM	**66.7**	**100**	0	0.39	
	GAM	55.6	83.3	0	**0.58**	
	FLMS	61.1	75.0	**33.3**	0.54	

***ACU*, *TPR*, *TAR and AUC* denote respectively accuracy, true positive rate, true absence rate and area under the ROC curve. The highest value for each criteria of each validation procedure was put in bold.**

4.3.2 Results for fry and 1+ parr abundance models

The results for fry and 1+ parr abundance are plotted as observed versus predicted abundances in Figure 4.6 for the three models. The *NSC* is also computed and added to the plot to assess models performance and comparison. The transferability validation is not plotted because all calculated *NSC* were below 0.10 when the model calibrated on the SMR is used with data from the PCR. As shown in Figure 4.6a, for the "Full model validation", fry abundance is well modelled by the GAM and the FLMS (respective *NSC* of 0.97 and 0.94), but the GLM is less able to capture the relation between abundance and the predictors (*NSC* of 0.56) as already noted before (i.e. the effect of the predictors is non-linear and cannot be captured by the GLM). When looking at the "Leave-one-out validation", the *NSC* for the three models decreases to 0.32, 0.45 and 0.48 respectively for the GLM, GAM and FLMS, showing comparable slightly better result for the FLMS than the GAM and the GLM to model fry abundance at sites not used in the calibration. For the 1+ parr results (Figure 4.6b), the model performance for the leave-one-out validation is lower than for fry abundance. The *NSC* are respectively 0.02, 0.15 and 0.37 for GLM, GAM and FLMS. Among the three models, the FLMS had the highest *NSC* (0.37) for the Leave-one-out cross-validation, meaning that this model appears to be more adapted to predict abundance on sites not used for calibration, compared to the GLM (*NSC* of 0.02) and the GAM (*NSC* of 0.15).

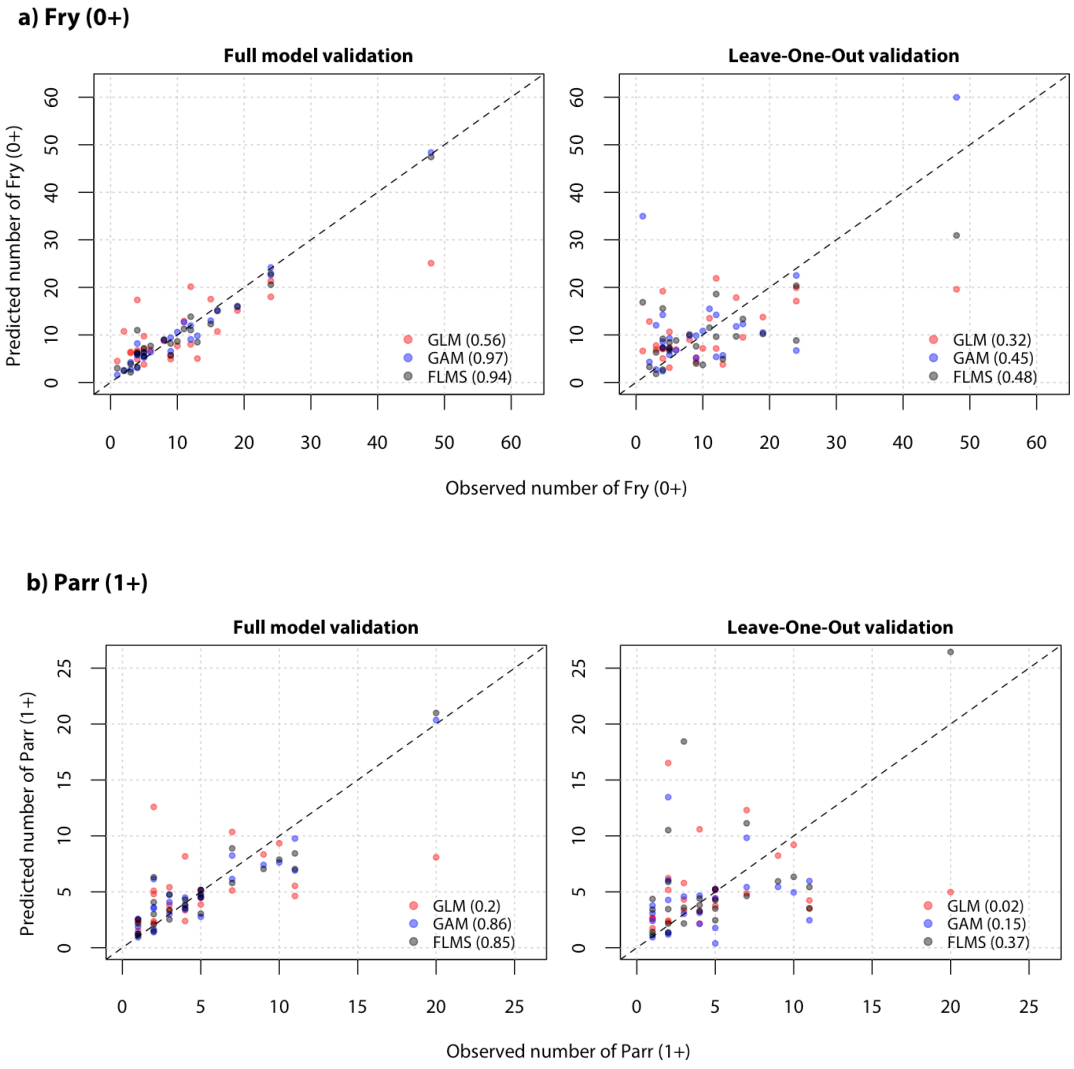
4.3.3 Added value of the water temperature predictor

As this study is one of the first considering water temperature as input for juvenile salmon habitat models, the added value of the water temperature in the three considered habitat models is quantified by the difference between the R^2adj (equation 12) for models with and without water temperature. The differences between the two obtained values are reported in Table 4.4. The R^2adj increases for all models and all ages despite the fact that one predictor is added. The increase varies from 0.03 to 0.61, showing the strong importance of its inclusion in habitat models with great increase of the models R^2adj and for the three ages.

Tableau 4.4 : Variation in R^2adj from models without to models with water temperature as predictor

Age	GLM	GAM	FLMS
0+	+0.43	+0.61	+0.45
1+	+0.10	+0.57	+0.51
2+	+0.03	+0.31	+0.45

Figure 4.6 : Result for modelling 0+ fry and 1+ parr abundances with the three models



Note: *NSC* are written in parentheses. In the case of the "Full model validation", the *NSC* is equivalent to the R^2.

4.4 Discussion

First, the preferred variables by the 2+ parr are discussed and the performance are compared with the literature. The results on modelling the fry and 1+ parr abundance are also discussed with focus on the importance of the cross validation. Finally, drawbacks and advantages of the proposed methodology are listed with literature comparison as well as future use of the functional regression models.

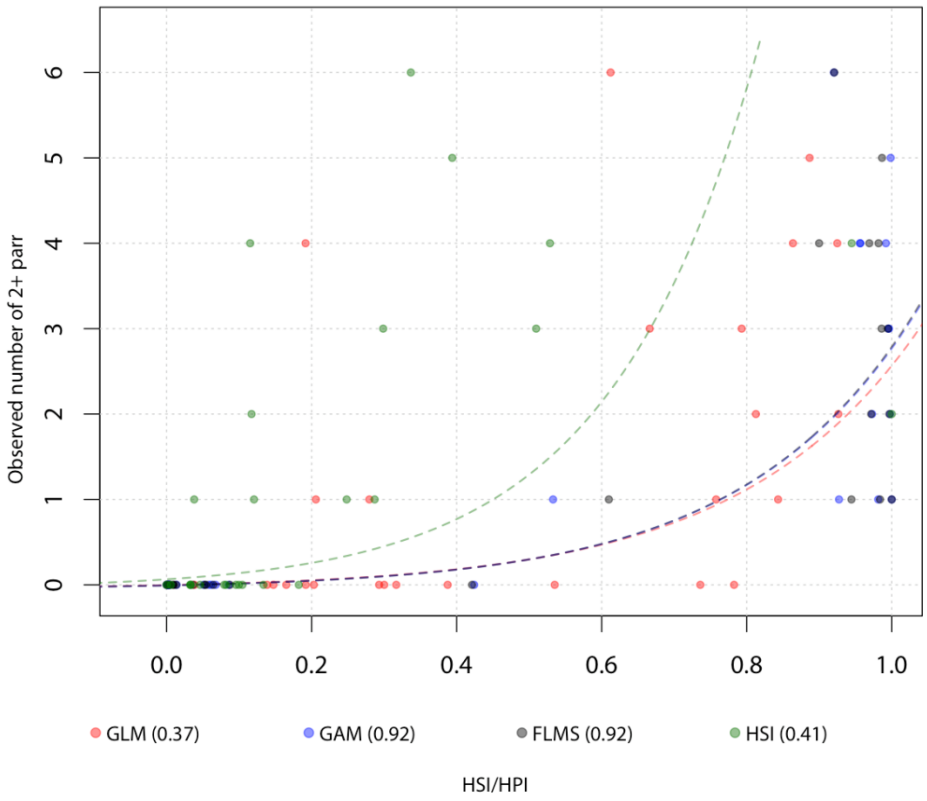
Coefficients effects interpretation: As seen in Figure 4.5, the GLM looks not adapted for fish habitat modelling because of the non-linear relations highlighted by both GAM and FLMS. For GAM, habitat evaluation for high values of velocities, substrate size or temperature looks to be wrong, suggesting high positive effect for velocity and high negative effects for substrate and temperature. This may lead to misinterpretation when using GAM for extrapolation while FLMS have much more biologically interpretable results in these extremes. Hence, this is a first important advantage of the FLMS in fish habitat modelling where it models a wider range of effect of each predictor and allows to potentially model maximum thresholds of suitability for habitat variables.

Evaluation of HPI approach for fish habitat modelling: To evaluate the HPIs (recall that HPI is the probability p_i to observe a fish at site *i* obtained from a regression model), a look is taken at how these indices can explain fish abundance observed at each site. Guay *et al.* (2000) highlighted the fact that HPI from a logistic regression was better to explain fish abundance than the classical HSI based on preference curve approach of Leclerc *et al.* (1994). The HSI values for each site were also calculated with the method described in Leclerc *et al.* (1994). The relationship between HSI/HPI and observed 2+ parr abundance are compared in Figure 4.7.

The HSI approach has a R^2 of 0.41, which is comparable to the results of Guay *et al.* (2000) (R^2 of 0.39) while the GLM has a R^2 of 0.36, which is lower than the R^2 obtained by the logistic regression of Guay *et al.* (2000) (R^2 of 0.86). Both GAM and FLMS have high values of R^2 (0.92) showing that models including non-linear effects of the predictors seem better suited to explain fish abundance from the HPI than the GLM and the HSI approach. The latter has shown low explicative power in previous studies as shown by the results of Guay *et al.* (2000). Also, in Hedger *et al.* (2004), the local HSI calculated for

5 rivers has R^2 values that vary from 0.30 to 0.46 for fry and from 0.35 to 0.69 for parr. This is much lower than the values obtained by the two non-linear models used in our study (*NSC* (or R^2) greater than 0.90 for GAM and FLMS), motivating the use of such non-linear approaches.

Figure 4.7 : Comparison between the HSI approach and HPI produced by the three regression models for 2+ parr presence-absence and relation with the observed fish abundance at each site



In parentheses is noted the R^2 corresponding to the link between HSI/HPI and abundance as modelling by a simple exponential relation.

Modelling abundances for fry and 1+ parr and importance of the cross validation:

In the literature, the GAM of Hedger *et al.* (2005) has a *NSC* of 0.28 for fry density and 0.47 for parr density using depth, substrate size and velocity as predictors and based on a full model validation. With performance being higher when all data are used for both calibration and validation, this would likely mean that their result would have been lower if a leave-one-out validation had been performed. With an extra predictor (water temperature), similar results are achieved on the leave-one-out cross-validation as the

ones of Hedger *et al.* (2005), with a *NSC* of 0.45 for fry and *NSC* of 0.37 for the 1+ parr abundances, whereas *NSC* greater than 0.90 are obtained on the full model validation for GAM and FLMS.

As observed in Figure 4.6, the relation between observed and predicted fish abundance is more scattered when passing from full model validation to leave-one-out cross-validation. In many models employed in the literature, no such validation is done (Guay *et al.*, 2000; Guay *et al.*, 2003; Hedger *et al.*, 2004; Hedger *et al.*, 2005; Beakes *et al.*, 2014; Millidine *et al.*, 2016). If the model is not able to estimate abundance for new data from the same river, it will surely not be able to predict the entire river productivity or to be transferable to other rivers. Hence, our study shows the importance of such cross validation in fish habitat modelling.

As noted before, transferring models for abundance of fry and 1+ parr from one river to the other is not as efficient as transferring the presence-absence model of the 2+ parr with the FLMS. Various reasons may explain why the models for abundances show no potential for transferability for fry and 1+ parr. First, the two rivers have very different fry abundance for the surveyed sites (see Table 4.1). The mean fry abundance per site is 10.6 for the SMR while it is only 3.1 for the PCR. Also, low values were noted on the PCR (there was no fry at 4 sites and only 1 fry at 4 sites) compared to the SMR (no site with 0 fry and all but one site had more than 1 fry). In spring 2017, a huge flood happened just after the egg hatching on the PCR, but not on the SMR. This flood has an estimated return period of 53 years based on a Gumbel distribution with online streamflow data on the PCR (CEHQ, 2018). This could have washed the fry downstream, explaining the low density observed on the PCR and the low transferability for fry abundance model (Jensen & Johnsen, 1999). Secondly, fry and 1+ parr have less mobility than the 2+ parr to select optimal habitat, potentially leading to poorer habitat also used by those younger ages. This is described as plasticity of juvenile salmon which is not taken into account is the modelling process and could explain low transferability (Mocq *et al.*, 2013). Finally, other predictors may be included to model juvenile salmon habitat. Even if this study is one of the first to consider the added value of the water temperature as a predictor for juvenile salmon habitat modelling, more predictors may still be needed to fully understand the variations in fish abundance like light intensity (Heggenes & Gunnar Dokk, 2001; Girard

et al., 2003), the distance between redds and salmon habitat (Klemetsen *et al.*, 2003), food availability (Vehanen, 2003), competition (Gabler & Amundsen, 1999) and potential predation (Dionne & Dodson, 2002; Vehanen, 2003).

Advantages, drawbacks and future use of functional regression: Despite the important improvements associated with functional regression, this new methodology to quantify habitat quality for fish with the use of functional regression model (i.e. the FLMS) shows better performance on the cross-validation for all three ages of juvenile salmon compared to the GAM and the GLM and looks better than the classical HSI method to explain fish density for the 2+ parr. Even if no interaction between the predictors were included in the model because of the small dataset (n=26 sites), the FLMS explained fish abundance or presence-absence more adequately and is not over-fitted, as shown by the cross-validation procedure. A drawback of the FLMS is that it requires curves (e.g. PDFs) of available habitat variable at each site, built from 30 measurements per site in our study. Investigations have been done to check if curves could be constructed with fewer measurements (~15), but so few data could not capture the heterogeneity of the habitat variables at each site. Even if the use of the FLMS requires more data and consequently more field work, the use of those distribution curves as inputs in fish habitat model improves the predictive power compared to the classical models and provides a better representation of real fish habitat than using a mean value per site. With automatic measurement techniques like airborne sensors and photo/video analysis (e.g. for substrate size (Carbonneau *et al.*, 2005; Hedger *et al.*, 2006), thermal measures (Dugdale, 2016), surface flow velocities (Smith *et al.*, 2005)), these curves of habitat variables availability would be easier to obtain and such model will hence be more easily applicable. Once these curves are obtained, the functional model is easy to fit in R. Using the functional model instead of other regression models or the old HSI approach would improve fish habitat modelling in terms of performance, but also in terms of which variables are more likely selected by the fish.

Finally, many habitat models are based on microhabitat measurement (i.e. habitat characterization at the fish nose). However, the fish, in its natural habitat, is not strictly exposed to these unique measured values but rather to a large range of values that it may also use for other biological tasks (Beecher *et al.*, 2010). Measures in microhabitat

model ignore habitat conditions in the vicinity of the fish and are too small to reflect the fish needs in terms of habitat (Shirvell, 1994). Hence, Mocq *et al.* (2018) recently showed the importance of an intermediate scale that can result in improved model predictions. Even if such findings are not directly addressed in this study (fish nose measurements versus PDF were not compared), our framework, using curves instead of discrete measurements, could directly address this problem in a future research.

4.5 Conclusion

A new method is developed to characterize fish habitat by making the use of the complete distribution of each predictor instead of single measurements in other classical regression models. Our method makes the use of a functional regression model (denoted FLMS), being able to work with curves or functions as inputs. Habitat of juvenile Atlantic salmon was modelled from either abundance data (for fry and 1+ parr) or presence-absence data (for 2+ parr) using the PDFs of four predictors: velocity, depth, substrate size and water temperature. The latter has been only seldom used in such habitat models and showed a strong importance in habitat models. A cross validation showed better results for the FLMS compared to a GLM and a GAM for the three ages of juvenile Atlantic salmon. With data from another river than the one used for the model calibration, the best transferability potential was obtained for the FLMS with the presence-absence model for the 2+ parr. Finally, the HPIs calculated by the three regression models for presence-absence of 2+ parr were compared to the classical HSI approach and both the GAM and the FLMS were better to explain abundance than the HSI. A particularly novel advantage of FLMS is the extended range of the predictors since it is based on the PDF instead of mean values. Our new method has shown good results by taking in consideration the variability within the fish habitat as characterized by the complete distribution based on numerous (30) measurements for each habitat variable.

Acknowledgement

The authors wish to acknowledge the financial contribution of Natural Sciences and Engineering Research Council of Canada (scholarship for the lead author) and to the

Fonds de recherche du Québec - Nature et technologies (FRQNT). Also, the authors would like to thank André Boivin, Marc-André Pouliot, Killian Dolais, Andrée-Sylvie Carbonneau, Michael Deetjens and Antoine Boudry for their help in the field work.

References

Ahmadi-Nedushan B, St-Hilaire A, Berube M, Ouarda T & Robichaud E (2008) Instream flow determination using a multiple input fuzzy-based rule system: A case study. *River Research and Applications* 24(3):279-292.

Ahmadi-Nedushan B, St-Hilaire A, Bérubé M, Robichaud É, Thiémonge N & Bobée B (2006) A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. *River Research and Applications* 22(5):503-523.

Armstrong J, Kemp P, Kennedy G, Ladle M & Milner N (2003) Habitat requirements of Atlantic salmon and brown trout in rivers and streams. *Fisheries research* 62(2):143-170.

Ayllón D, Almodóvar A, Nicola G & Elvira B (2012) The influence of variable habitat suitability criteria on PHABSIM habitat index results. *River Research and Applications* 28(8):1179-1188.

Bardonnet A & Baglinière J-L (2000) Freshwater habitat of Atlantic salmon (Salmo salar). *Canadian Journal of Fisheries and Aquatic Sciences* 57(2):497-506.

Beakes M, Moore J, Retford N, Brown R, Merz J & Sogard S (2014) Evaluating statistical approaches to quantifying juvenile Chinook salmon habitat in a regulated California river. *River Research and Applications* 30(2):180-191.

Beecher HA, Caldwell BA, DeMond SB, Seiler D & Boessow SN (2010) An Empirical Assessment of PHABSIM Using Long-Term Monitoring of Coho Salmon Smolt Production in Bingham Creek, Washington. *North American Journal of Fisheries Management* 30(6):1529-1543.

Bel L, Bar-Hen A, Petit R & Cheddadi R (2011) Spatio-temporal functional regression on paleoecological data. *Journal of Applied Statistics* 38(4):695-704.

Bosq D (2012) *Linear processes in function spaces: theory and applications.* Springer Science & Business Media,

Bouchard J & Boisclair D (2008) The relative importance of local, lateral, and longitudinal variables on the development of habitat quality models for a river. *Canadian Journal of Fisheries and Aquatic Sciences* 65(1):61-73.

Boudreau P, Bourgeois G, Leclerc M, Boudreault A & Belzile L (1996) Two-dimensional habitat model validation based on spatial fish distribution: application to juvenile Atlantic salmon of Moisie River (Québec, Canada). *Ecohydraulics 2000: Proceedings of the 2nd International Symposium on Habitat Hydraulics, Québec, Qc.* p B365-B380.

Bourgeois G, Cunjak RA, Caissie D & El-Jabi N (1996) A spatial and temporal evaluation of PHABSIM in relation to measured density of juvenile Atlantic salmon in a small stream. *North American Journal of Fisheries Management* 16(1):154-166.

Bovee KD, Lamb BL, Bartholow JM, Stalnaker CB & Taylor J (1998) Stream habitat analysis using the instream flow incremental methodology. (GEOLOGICAL SURVEY RESTON VA BIOLOGICALRESOURCES DIV).

Brockhaus S, Melcher M, Leisch F & Greven S (2017a) Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing* 27(4):913-926.

Brockhaus S, Ruegamer D, Hothorn T & Brockhaus MS (2017b) Package 'FDboost'.

Brockhaus S, Rügamer D & Greven S (2017c) Boosting Functional Regression Models with FDboost. *arXiv preprint arXiv:1705.10662*.

Brockhaus S, Scheipl F, Hothorn T & Greven S (2015) The functional linear array model. *Statistical Modelling* 15(3):279-300.

Bühlmann P & Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* :477-505.

Carbonneau PE, Bergeron N & Lane SN (2005) Automated grain size measurements from airborne remote sensing for long profile measurements of fluvial grain sizes. *Water Resources Research* 41(11).

CEHQ (2018) Historique des données de différentes stations hydrométriques, Centre d'expertise hydrique du Québec. *https://www.cehq.gouv.qc.ca/hydrometrie/index.htm*.

Chaouch M (2014) Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves. *IEEE Transactions on Smart Grid* 5(1):411-419.

Chebana F, Dabo-Niang S & Ouarda TB (2012) Exploratory functional flood frequency analysis and outlier detection. *Water Resources Research* 48(4).

Ciarleglio A, Petkova E, Tarpey T & Ogden RT (2016) Flexible functional regression methods for estimating individualized treatment rules. *Stat* 5(1):185-199.

Dabo-Niang S & Ferraty F (2008) *Functional and operatorial statistics.* Springer Science & Business Media,

Daigle A, Jeong DI & Lapointe MF (2015) Climate change and resilience of tributary thermal refugia for salmonids in eastern Canadian rivers. *Hydrological Sciences Journal* 60(6):1044-1063.

Davey C & Lapointe M (2007) Sedimentary links and the spatial organization of Atlantic salmon (Salmo salar) spawning habitat in a Canadian Shield river. *Geomorphology* 83(1):82-96.

Davey CE (2005) Longitudinal trends in grain size, shear stress and sediment mobility along sedimentary links of a Canadian Shield River, Saguenay Region: A geomorphic perspective on assessing Atlantic salmon (Salmo salar) productivity in rivers.

Dionne M & Dodson JJ (2002) Impact of exposure to a simulated predator (Mergus merganser) on the activity of juvenile Atlantic salmon (Salmo salar) in a natural environment. *Canadian journal of zoology* 80(11):2006-2013.

Dugdale SJ (2016) A practitioner's guide to thermal infrared remote sensing of rivers and streams: recent advances, precautions and considerations. *Wiley Interdisciplinary Reviews: Water* 3(2):251-268.

Elliott J & Elliott J (2010) Temperature requirements of Atlantic salmon Salmo salar, brown trout Salmo trutta and Arctic charr Salvelinus alpinus: predicting the effects of climate change. *Journal of fish biology* 77(8):1793-1817.

Ferraty F & Vieu P (2006) *Nonparametric functional data analysis: theory and practice.* Springer Science & Business Media,

Feyrer F, Nobriga ML & Sommer TR (2007) Multidecadal trends for three declining fish species: habitat patterns and mechanisms in the San Francisco Estuary, California, USA. *Canadian Journal of Fisheries and Aquatic Sciences* 64(4):723-734.

Freeman MC, Bowen ZH & Crance JH (1997) Transferability of habitat suitability criteria for fishes in warmwater streams. *North American Journal of Fisheries Management* 17(1):20-31.

Gabler HM & Amundsen PA (1999) Resource partitioning between Siberian sculpin (Cottus poecilopus Heckel) and Atlantic salmon parr (Salmo salar L.) in a sub-Arctic river, northern Norway. *Ecology of Freshwater Fish* 8(4):201-208.

Gibson R (1993) The Atlantic salmon in fresh water: spawning, rearing and production. *Reviews in fish biology and fisheries* 3(1):39-73.

Girard P, Boisclair D & Leclerc M (2003) The effect of cloud cover on the development of habitat quality indices for juvenile Atlantic salmon (Salmo salar). *Canadian Journal of Fisheries and Aquatic Sciences* 60(11):1386-1397.

Goia A, May C & Fusai G (2010) Functional clustering and linear regression for peak load forecasting. *International Journal of Forecasting* 26(4):700-711.

Guay J, Boisclair D, Leclerc M & Lapointe M (2003) Assessment of the transferability of biological habitat models for Atlantic salmon parr (Salmo salar). *Canadian Journal of Fisheries and Aquatic Sciences* 60(11):1398-1408.

Guay J, 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(10):2065-2075.

Hastie T & Tibshirani R (1990) *Generalized additive models.* Wiley Online Library,

Hedger R, Dodson J, Bergeron N & Caron F (2004) Quantifying the effectiveness of regional habitat quality index models for predicting densities of juvenile Atlantic salmon (Salmo salar L.). *Ecology of freshwater fish* 13(4):266-275.

Hedger R, Dodson J, Bergeron N & Caron F (2005) Habitat selection by juvenile Atlantic salmon: the interaction between physical habitat and abundance. *Journal of Fish Biology* 67(4):1054-1071.

Hedger R, Dodson J, Bourque J, Bergeron N & Carbonneau P (2006) Improving models of juvenile Atlantic salmon habitat use through high resolution remote sensing. *ecological modelling* 197(3):505-511.

Heggenes J (1990) Habitat utilization and preferences in juvenile Atlantic salmon (Salmo salar) in streams. *River Research and Applications* 5(4):341-354.

Heggenes J, Baglinière J & Cunjak R (1995) Note de synthèse sur la sélection de niche spatiale et la compétition chez le jeune saumon atlantique (Salmo salar) et la truite commune (Salmo trutta) en milieu lotique. *Bulletin Français de la Pêche et de la Pisciculture* (337-338-339):231-239.

Heggenes J & Gunnar Dokk J (2001) Contrasting temperatures, waterflows, and light: seasonal habitat selection by young Atlantic salmon and brown trout in a boreonemoral river. *River Research and Applications* 17(6):623-635.

Horváth L & Kokoszka P (2012) *Inference for functional data with applications.* Springer Science & Business Media,

Ivanescu AE, Staicu A-M, Scheipl F & Greven S (2015) Penalized function-on-function regression. *Computational Statistics* 30(2):539-568.

Jensen A & Johnsen B (1999) The functional relationship between peak spring floods and survival and growth of juvenile Atlantic salmon (Salmo salar) and brown trout (Salmo trutta). *Functional Ecology* 13(6):778-785.

Johnston P & Bergeron N (2010) Variation of juvenile Atlantic salmon (Salmo salar) body composition along sedimentary links. *Ecology of freshwater fish* 19(2):187-196.

Jonsson B & Jonsson N (2009) A review of the likely effects of climate change on anadromous Atlantic salmon Salmo salar and brown trout Salmo trutta, with particular reference to water temperature and flow. *Journal of fish biology* 75(10):2381-2447.

Jorde K, Schneider M, Peter A & Zoellner F (2001) Fuzzy based models for the evaluation of fish habitat quality and instream flow assessment. *Proceedings of the 3rd international symposium on environmental hydraulics.* p 27-28.

Kim M & Lapointe M (2011) Regional variability in Atlantic salmon (Salmo salar) riverscapes: a simple landscape ecology model explaining the large variability in size of salmon runs across Gaspé watersheds, Canada. *Ecology of Freshwater Fish* 20(1):144-156.

Kim MS (2009) *The controls of sedimentary links on the spatial distribution of Atlantic salmon (Salmo salar) juveniles and spawning activity along rivers in the Gaspé Peninsula, Canada*. (McGill University).

Klemetsen A, Amundsen PA, Dempson J, Jonsson B, Jonsson N, O'connell M & Mortensen E (2003) Atlantic salmon Salmo salar L., brown trout Salmo trutta L.

and Arctic charr Salvelinus alpinus (L.): a review of aspects of their life histories. *Ecology of freshwater fish* 12(1):1-59.

Labonne J, Allouche S & Gaudin P (2003) Use of a generalised linear model to test habitat preferences: the example of Zingel asper, an endemic endangered percid of the River Rhone. *Freshwater Biology* 48(4):687-697.

Lackey RT (2003) Pacific Northwest salmon: forecasting their status in 2100. *Reviews in fisheries Science* 11(1):35-88.

Lanthier G, Bédard M-E, Lapointe M & Boisclair D (2014) Assessment of the structural role of sedimentary links on the spatial distribution of periphyton and fish in a Canadian Shield river. *Aquatic sciences* 77(1):141-152.

Larabi S, St-Hilaire A, Chebana F & Latraverse M (2017) Multi-Criteria Process-Based Calibration Using Functional Data Analysis to Improve Hydrological Model Realism. *Water Resources Management* :1-17.

Latulippe C, Lapointe MF & Talbot T (2001) Visual characterization technique for gravel-cobble river bed surface sediments; validation and environmental applications Contribution to the programme of CIRSA(Centre Interuniversitaire de Recherche sur le Saumon Atlantique). *Earth Surface Processes and Landforms* 26(3):307-318.

Leclerc M, Boudreau P, Bechara J, Belzile L & Villeneuve D (1994) Modélisation de la dynamique de l'habitat des jeunes stades de saumon atlantique (Salmo salar) de la rivière Ashuapmushuan (Québec, Canada). *Bulletin Français de la Pêche et de la Pisciculture* (332):11-32.

Liu C, Berry PM, Dawson TP & Pearson RG (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28(3):385-393.

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(2):218-228.

Manly B, McDonald L, Thomas DL, McDonald TL & Erickson WP (2007) *Resource selection by animals: statistical design and analysis for field studies.* Springer Science & Business Media,

Masselot P, Dabo-Niang S, Chebana F & Ouarda TB (2016) Streamflow forecasting using functional regression. *Journal of Hydrology* 538:754-766.

McDonald S, Koulis T, Ehn J, Campbell K, Gosselin M & Mundy C (2015) A functional regression model for predicting optical depth and estimating attenuation coefficients in sea-ice covers near Resolute Passage, Canada. *Annals of Glaciology* 56(69):147-154.

McLean MW, Hooker G, Staicu A-M, Scheipl F & Ruppert D (2014) Functional generalized additive models. *Journal of Computational and Graphical Statistics* 23(1):249-269.

Meyer MJ, Coull BA, Versace F, Cinciripini P & Morris JS (2015) Bayesian function-on-function regression for multilevel functional data. *Biometrics* 71(3):563-574.

Millidine K, Malcolm I & Fryer R (2016) Assessing the transferability of hydraulic habitat models for juvenile Atlantic salmon. *Ecological Indicators* 69:434-445.

Mocq J, St-Hilaire A & Cunjak RA (2013) Assessment of Atlantic salmon (Salmo salar) habitat quality and its uncertainty using a multiple-expert fuzzy model applied to the Romaine River (Canada). *Ecological modelling* 265:14-25.

Mocq J, St-Hilaire A & Cunjak RA (2018) Do habitat measurements in the vicinity of Atlantic salmon (Salmo salar) parr matter? *Fisheries Management and Ecology* 25(1):22-30.

Morris JS (2015) Functional regression. *Annual Review of Statistics and Its Application* 2:321-359.

Nagelkerke NJ (1991) A note on a general definition of the coefficient of determination. *Biometrika* 78(3):691-692.

Neter J, Kutner MH, Nachtsheim CJ & Wasserman W (1996) *Applied linear statistical models.* Irwin Chicago,

Noakes DJ, Beamish RJ & Kent ML (2000) On the decline of Pacific salmon and speculative links to salmon farming in British Columbia. *Aquaculture* 183(3):363-386.

Nyqvist D, Greenberg LA, Goerig E, Calles O, Bergman E, Ardren WR & Castro-Santos T (2017) Migratory delay leads to reduced passage success of Atlantic salmon smolts at a hydroelectric dam. *Ecology of Freshwater Fish* 26(4):707-718.

Orth DJ & Maughan OE (1982) Evaluation of the incremental methodology for recommending instream flows for fishes. *Transactions of the American Fisheries Society* 111(4):413-445.

Prévost L, Levesque D & Plamondon AP (2002) *Méthodologie pour évaluer l'effet de l'installation d'un ponceau sur le substrat des frayères de l'omble de fontaine (Salvelinus fontinalis).* Ministère des ressources naturelles,

R Core Team (2017) R language definition. *Vienna, Austria: R foundation for statistical computing*.

Railsback SF (2016) Why it is time to put PHABSIM out to pasture. *Fisheries* 41(12):720-725.

Ramsay J (1982) When the data are functions. *Psychometrika* 47(4):379-396.

Ramsay JO (2006) *Functional data analysis.* Wiley Online Library,

Ramsay JO, Hooker G & Graves S (2009) *Functional data analysis with R and MATLAB.* Springer Science & Business Media,

Requena AI, Chebana F & Ouarda TB (2018) A functional framework for flow-duration-curve and daily streamflow estimation at ungauged sites. *Advances in Water Resources*.

Rice S & Church M (1998) Grain size along two gravel-bed rivers: statistical variation, spatial pattern and sedimentary links. *Earth Surface Processes and Landforms* 23(4):345-363.

Rice SP, Greenwood MT & Joyce CB (2001) Tributaries, sediment sources, and the longitudinal organisation of macroinvertebrate fauna along river systems. *Canadian Journal of Fisheries and Aquatic Sciences* 58(4):824-840.

Scott D & Shirvell C (1987) A critique of the instream flow incremental methodology and observations on flow determination in New Zealand. *Regulated streams,* Springer. p 27-43.

Shao J (1993) Linear model selection by cross-validation. *Journal of the American statistical Association* 88(422):486-494.

Shirvell C (1994) Effect of changes in streamflow on the microhabitat use and movements of sympatric juvenile coho salmon (Oncorhynchus kisutch) and chinook salmon (O. tshawytscha) in a natural stream. *Canadian Journal of Fisheries and Aquatic Sciences* 51(7):1644-1652.

Smith J, Bérubé F & Bergeron N (2005) A field application of particle image velocimetry (PIV) for the measurement of surface flow velocities in aquatic habitat studies. *26th Canadian Symposium on Remote Sensing.*

Stewart-Koster B, Olden JD & Gido KB (2014) Quantifying flow–ecology relationships with functional linear models. *Hydrological sciences journal* 59(3-4):629-644.

Sundt-Hansen L, Hedger R, Ugedal O, Diserud O, Finstad A, Sauterleute J, Tøfte L, Alfredsen K & Forseth T (2018) Modelling climate change effects on Atlantic salmon: Implications for mitigation in regulated rivers. *Science of The Total Environment* 631:1005-1017.

Ternynck C, Ben Alaya MA, Chebana F, Dabo-Niang S & Ouarda TB (2016) Streamflow hydrograph classification using functional data analysis. *Journal of Hydrometeorology* 17(1):327-344.

Tetzlaff D, Soulsby C, Youngson A, Gibbins C, Bacon P, Malcolm I & Langan S (2005) Variability in stream discharge and temperature: a preliminary assessment of the implications for juvenile and spawning Atlantic salmon. *Hydrology and Earth System Sciences* 9(3):193-208.

Tukey JW (1977) *Exploratory data analysis.* Reading, Mass.,

Vehanen T (2003) Adaptive flexibility in the behaviour of juvenile Atlantic salmon: short-term responses to food availability and threat from predation. *Journal of Fish Biology* 63(4):1034-1045.

Yi Y, Cheng X, Yang Z, Wieprecht S, Zhang S & Wu Y (2017) Evaluating the ecological influence of hydraulic projects: A review of aquatic habitat suitability models. *Renewable and Sustainable Energy Reviews* 68:748-762.

5 CONCLUSION ET RECOMMANDATIONS

L'objectif principal de ce mémoire était d'introduire les approches de régression fonctionnelle à la modélisation de la température de l'eau et de l'habitat du saumon atlantique juvénile. Compte tenu des diverses problématiques soulevées dans les deux cas d'études, la régression fonctionnelle semblait naturellement bien adaptée pour répondre à ces problématiques, ce pour quoi elle a été retenue comme méthode de modélisation. Son avantage principal est de permettre l'inclusion de variables (explicatives et/ou réponse) vues comme des courbes plutôt que seulement des valeurs scalaires ou vectorielles (c.-à-d. des nombres ou des groupes de nombres) dans les autres approches de modélisation classiques. L'introduction de cette méthode permet de traiter (1) les variables de température de l'eau et de l'air comme des fonctions variant dans le temps au cours d'une saison et (2) les variables d'habitat du poisson comme des fonctions de densité de probabilité (FDP) pour mieux représenter son habitat. Dans les deux cas d'étude, des modèles classiques basés sur des valeurs scalaires ont été comparés aux approches fonctionnelles. Finalement, autant des données obtenues via Internet que collectées sur le terrain (durant l'été 2017) ont été utilisées pour accomplir les objectifs spécifiques de ce mémoire.

Pour ce qui est du premier projet, le modèle fonctionnel linéaire historique (MFLH) s'est avéré plus adapté aux données de températures de l'air et de l'eau que le modèle fonctionnel linéaire complet (MFLC) de par sa plus grande parcimonie et signification d'un point de vue météorologique. Notre première recommandation serait donc d'introduire le MFLH pour modéliser la *courbe* du débit, qui pourrait être plus performant que le MFLC précédemment utilisé dans Masselot *et al.* (2016). Deuxièmement, dans un contexte exploratoire, le domaine utilisé de la courbe de la variable explicative dans le MFLH était toutes les valeurs de $x(s)$ avant t (c'est-à-dire $s \in [0, t[$). Or, à partir de la surface $\beta(s, t)$ obtenue (figure 3.3 b), on a vu que le domaine utilisé de la variable explication pourrait se limiter aux 15 jours précédant t (c'est-à-dire $s \in [t\text{-}15, t[$). Cela réduirait encore davantage le nombre de paramètres à estimer et rendrait le modèle encore plus parcimonieux.

Troisièmement, le modèle MFLH a montré des résultats meilleurs que les autres modèles testés et comparables à ceux de la littérature pour deux des trois rivières à l'étude (racine de l'erreur quadratique moyenne supérieure, RMSE, d'environ 1.5Cº). Pour la troisième rivière (Delaware), tous les modèles ont obtenu des RMSE supérieurs à 2ºC. Ainsi, une troisième recommandation serait de considérer l'effet d'autres variables explicatives (p. ex. le débit, les précipitations, le vent, etc.), dans la modélisation de la température de l'eau de cette rivière.

Une quatrième recommandation, vu les bons résultats obtenus par la modélisation fonctionnelle, serait de valider la capacité du modèle fonctionnel à agir en mode « prévisionnel ». Une façon de le valider serait d'utiliser la climatologie passée ou le résultat d'un modèle climatique pour obtenir une série des températures de l'air *temporaires* pour l'année complète à venir et de l'utiliser pour faire une prévision initiale des températures de l'eau. Puis, au fur et à mesure que les températures de l'air seraient observées, celles-ci remplaceraient les valeurs temporaires utilisées, ce qui mettrait à jour la prévision initiale de la courbe des températures de l'eau et rendrait celle-ci de plus en plus précise pour les jours à venir. Cette courbe pourrait être fournie aux gestionnaires quotidiennement, ce qui leur permettrait de connaître la thermie à venir du cours d'eau dans les jours qui suivent, mais aussi d'avoir un portrait global de la thermie de la saison.

Comme cinquième recommandation, un modèle additif généralisé différent de celui de Laanaya *et al.* (2017) pourrait être considéré comme modèle de comparaison. Comme première exemple, l'allure de la figure 3.4 propose une relation quasi linéaire entre la température de l'air et de l'eau pour la rivière Potomac. Ainsi, un MAG avec un prédicteur linéaire pour la température de l'eau pourrait être considéré plutôt qu'un effet non linéaire comme il avait été fait par Laanaya *et al.* (2017). Aussi, un MAG considérant des effets différents de la température de l'air sur celle de l'eau selon le jour de l'année en utilisant $f(x, t)$ comme prédicteur, plutôt que seulement $f(x)$ et $f(t)$ indépendamment, serait une deuxième option. Finalement, l'effet retardé de la température de l'air sur celle de l'eau pourrait être pris en compte, comme c'est le cas avec les modèles fonctionnels utilisés dans cette étude, avec l'utilisation du *Distributed Lag Non-linear Model* (DLNM) (Gasparrini *et al.*, 2010).

Finalement, il serait profitable de mettre en place des modèles de régression fonctionnelle sur des rivières à saumon de l'est du Canada. Cependant, le manque de longues séries de données n'a pas permis de le faire dans la présente étude. De nombreux efforts sont actuellement mis sur pied pour collecter ces informations sur les régimes thermiques des rivières à saumon étant donné l'importance primordiale de la variable de la température de l'eau sur le saumon, notamment pour la gestion de la pêche (RivTemp, 2018). D'ici quelques années, les modèles fonctionnels pourront être calibrés et testés sur celles-ci et pourront servir aux gestionnaires de rivières à saumon grâce au mode opérationnel décrit plus haut.

Pour le second projet, il faut d'abord souligner qu'outre l'utilisation du modèle fonctionnel, l'ajout de la variable de la température de l'eau a permis d'expliquer un plus grand pourcentage de la variation dans la sélection de l'habitat du saumon juvénile, ce qui est une avancée en soit en modélisation des habitats aquatiques. Le modèle fonctionnel MFLS a donné de meilleurs résultats pour les trois âges du saumon juvénile (0+, 1+, 2+) que les modèles classiques utilisés dans la validation croisée. De plus, un pas vers une potentielle transférabilité a pu être démontré pour le modèle fonctionnel de présence-absence du 2+. Comme les résultats de transférabilité étaient assez faibles pour les modèles d'abondance, une première recommandation serait de faire une étude distincte et plus poussée sur la transférabilité des modèles d'habitat pour prédire l'abondance de poissons en utilisant plusieurs rivières avec des caractéristiques similaires et différentes à la fois. Aussi, un modèle incluant des données de plusieurs rivières pourrait être considéré pour analyser la transférabilité comme il a été fait dans Hedger *et al.* (2004) avec l'approche des courbes de préférence, mais en utilisant plutôt des modèles de régression, dont le fonctionnel.

Deuxièmement, une étude comparant un modèle d'habitat basé sur des mesures « spot » (au nez du poisson) à un modèle d'habitat fonctionnel basé sur plusieurs mesures dans l'habitat immédiat où le poisson est observé devrait être réalisée. Le cadre développé ici permet de tester facilement la valeur ajoutée de plusieurs mesures dans l'habitat immédiat du poisson (résumée dans des FDP) à une mesure seulement à son nez comme c'est le cas dans les modèles classiques de microhabitat.

Troisièmement, bien que l'étude vise l'inclusion d'une quatrième variable explicative en modélisation de l'habitat aquatique (la température de l'eau), d'autres variables restent tout de même à être considérées. Ces variables (p. ex. intensité lumineuse, prédation, compétition, nourriture, distance de la frayère), avec la possibilité pour certaines d'être vues comme des courbes, sont aussi susceptibles de donner plus d'indications quant aux habitats propices pour le saumon atlantique juvénile dans l'optique de mieux les conserver.

Finalement, un modèle fonctionnel d'habitat pourrait être utilisé pour calculer des débits réservés et être comparé à des approches classiques comme PHABISM. La valeur ajoutée de l'utilisation d'un modèle fonctionnel pour cette application bien précise pourrait par la suite être quantifiée, ce qui serait un argument de plus parmi les autres soulevés jusqu'à présent quant à l'utilisation du modèle fonctionnel en habitat aquatique.

Somme toute, le présent mémoire démontre clairement l'intérêt d'utiliser l'approche statistique de la régression fonctionnelle en modélisation de la température de l'eau et des habitats aquatiques. La rigueur statistique des analyses effectuées ainsi que la grande variabilité de données utilisées nous a permis de montrer clairement les avantages tant conceptuels qu'en termes de performance des modèles fonctionnels versus les approches classiques. De plus, ces deux cas d'étude ont permis de montrer que la régression fonctionnelle fournissait de nouvelles informations quant au lien entre les variables explicatives et celles réponses (par exemple la période d'intérêt où la température de l'air influence la température de l'eau ou l'effet des valeurs extrêmes des variables d'habitat du saumon sur la sélection de l'habitat). Ceci s'explique notamment par la représentativité et la richesse des modèles fonctionnels, mais aussi en raison de leur facilité d'interprétation, ce qui n'est pas toujours le cas avec des modèles plus complexes comme ceux non paramétriques (p. ex. les réseaux de neurones) de plus en plus utilisés de nos jours. La régression fonctionnelle est donc un outil clé prometteur en modélisation des phénomènes hydrologiques et de l'écologie en rivière.

RÉFÉRENCES

Ahmadi-Nedushan B, St-Hilaire A, Berube M, Ouarda T & Robichaud E (2008) Instream flow determination using a multiple input fuzzy-based rule system: A case study. *River Research and Applications* 24(3):279-292.

Ahmadi-Nedushan B, St-Hilaire A, Bérubé M, Robichaud É, Thiémonge N & Bobée B (2006) A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. *River Research and Applications* 22(5):503-523.

Angilletta MJ, Ashley Steel E, Bartz KK, Kingsolver JG, Scheuerell MD, Beckman BR & Crozier LG (2008) Big dams and salmon evolution: changes in thermal regimes and their potential evolutionary consequences. *Evolutionary Applications* 1(2):286-299.

Armstrong J, Kemp P, Kennedy G, Ladle M & Milner N (2003) Habitat requirements of Atlantic salmon and brown trout in rivers and streams. *Fisheries research* 62(2):143-170.

Armstrong JD, Braithwaite VA & Fox M (1998) The response of wild Atlantic salmon parr to acute reductions in water flow. *Journal of Animal Ecology* 67(2):292-297.

Assani A & Petit F (2004) Impact of hydroelectric power releases on the morphology and sedimentology of the bed of the Warche River (Belgium). *Earth Surface Processes and Landforms* 29(2):133-143.

Ayllón D, Almodóvar A, Nicola G & Elvira B (2012) The influence of variable habitat suitability criteria on PHABSIM habitat index results. *River Research and Applications* 28(8):1179-1188.

Bardonnet A & Baglinière J-L (2000) Freshwater habitat of Atlantic salmon (Salmo salar). *Canadian Journal of Fisheries and Aquatic Sciences* 57(2):497-506.

Bartholow JM (1995) The stream network temperature model (SNTEMP): A decade of results. *Workshop on computer application in water management: Fort Collins, CO. Water Resources Research Institute, Colorado State University, Fort Collins, CO.* p 57-60.

Bartholow JM, Campbell SG & Flug M (2004) Predicting the thermal effects of dam removal on the Klamath River. *Environmental Management* 34(6):856-874.

Beakes M, Moore J, Retford N, Brown R, Merz J & Sogard S (2014) Evaluating statistical approaches to quantifying juvenile Chinook salmon habitat in a regulated California river. *River Research and Applications* 30(2):180-191.

Beall E, Dumas J, Claireaux D, Barriere L & Marty C (1994) Dispersal patterns and survival of Atlantic salmon (Salmo salar L.) juveniles in a nursery stream. *ICES journal of marine science* 51(1):1-9.

Beecher HA, Caldwell BA, DeMond SB, Seiler D & Boessow SN (2010) An Empirical Assessment of PHABSIM Using Long-Term Monitoring of Coho Salmon Smolt

Production in Bingham Creek, Washington. *North American Journal of Fisheries Management* 30(6):1529-1543.

Bel L, Bar-Hen A, Petit R & Cheddadi R (2011) Spatio-temporal functional regression on paleoecological data. *Journal of Applied Statistics* 38(4):695-704.

Bélanger M, El-Jabi N, Caissie D, Ashkar F & Ribi J (2005) Estimation de la température de l'eau de rivière en utilisant les réseaux de neurones et la régression linéaire multiple. *Revue des sciences de l'eau/Journal of Water Science* 18(3):403-421.

Benyahya L, Caissie D, St-Hilaire A, Ouarda TB & Bobée B (2007a) A review of statistical water temperature models. *Canadian Water Resources Journal* 32(3):179-192.

Benyahya L, St-Hilaire A, Ouarda TB, Bobée B & Dumas J (2008) Comparison of non-parametric and parametric water temperature models on the Nivelle River, France. *Hydrological sciences journal* 53(3):640-655.

Benyahya L, St-Hilaire A, Quarda TB, Bobée B & Ahmadi-Nedushan B (2007b) Modeling of water temperatures based on stochastic approaches: case study of the Deschutes River. *Journal of Environmental Engineering and Science* 6(4):437-448.

Berland G, Nickelsen T, Heggenes J, Økland F, Thorstad E & Halleraker J (2004) Movements of wild Atlantic salmon parr in relation to peaking flows below a hydropower station. *River research and Applications* 20(8):957-966.

Bernardi MS, Sangalli LM, Mazza G & Ramsay JO (2017) A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. *Stochastic Environmental Research and Risk Assessment* 31(1):23-38.

Beschta R, Bilby R, Brown G & Holtby L (1987) Stream temperature and aquatic habitat: pp. 191-232. *Fishery and forestry interactions. Streamside management: forestry and fishery interactions. University of Washington, Institute of Forest Resources. Contr* 57.

Bjornn T & Reiser D (1991) Habitat requirements of salmonids in streams. *American Fisheries Society Special Publication* 19(837):138.

Blann K, Frost Nerbonne J & Vondracek B (2002) Relationship of riparian buffer type to water temperature in the driftless area ecoregion of Minnesota. *North American Journal of Fisheries Management* 22(2):441-451.

Bley PW (1987) Age, Growth, and Mortality of Juvenile Atlantic Salmon in Streams: A Review. (MAINE COOPERATIVE FISHERY RESEARCH UNIT ORONO).

Boeuf G, Marc AM, Prunet P, Le Bail PY & Smal J (1994) Stimulation of parr-smolt transformation by hormonal treatment in Atlantic salmon (Salmo salar L.). *Aquaculture* 121(1-3):195-208.

Bosq D (2012) *Linear processes in function spaces: theory and applications.* Springer Science & Business Media,

Bouchard J & Boisclair D (2008) The relative importance of local, lateral, and longitudinal variables on the development of habitat quality models for a river. *Canadian Journal of Fisheries and Aquatic Sciences* 65(1):61-73.

Boudreau P, Bourgeois G, Leclerc M, Boudreault A & Belzile L (1996) Two-dimensional habitat model validation based on spatial fish distribution: application to juvenile Atlantic salmon of Moisie River (Québec, Canada). *Ecohydraulics 2000: Proceedings of the 2nd International Symposium on Habitat Hydraulics, Québec, Qc.* p B365-B380.

Bourgeois G, Cunjak RA, Caissie D & El-Jabi N (1996) A spatial and temporal evaluation of PHABSIM in relation to measured density of juvenile Atlantic salmon in a small stream. *North American Journal of Fisheries Management* 16(1):154-166.

Bovee KD (1978) The incremental method of assessing habitat potential for coolwater species, with management implications. *American Fisheries Society Special Publication* 11:340-343.

Bovee KD (1982) Guide to stream habitat analysis using the instream flow incremental methodology. *Available from the National Technical Information Service, Springfield VA 22161 as PB 83-131052. Report*.

Bovee KD, Lamb BL, Bartholow JM, Stalnaker CB & Taylor J (1998) Stream habitat analysis using the instream flow incremental methodology. (GEOLOGICAL SURVEY RESTON VA BIOLOGICALRESOURCES DIV).

Bovee KD & Milhous R (1978) Hydraulic simulation in instream flow studies: theory and techniques. IFIP No. 5. (US Fish and Wildlife Service).

Bradford MJ & Heinonen JS (2008) Low flows, instream flow needs and fish ecology in small streams. *Canadian water resources Journal* 33(2):165-180.

Breau C, Cunjak RA & Peake SJ (2011) Behaviour during elevated water temperatures: can physiology explain movement of juvenile Atlantic salmon to cool water? *Journal of Animal Ecology* 80(4):844-853.

Brockhaus S, Melcher M, Leisch F & Greven S (2017a) Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing* 27(4):913-926.

Brockhaus S, Ruegamer D, Hothorn T & Brockhaus MS (2017b) Package 'FDboost'.

Brockhaus S, Rügamer D & Greven S (2017c) Boosting Functional Regression Models with FDboost. *arXiv preprint arXiv:1705.10662*.

Brockhaus S, Scheipl F, Hothorn T & Greven S (2015) The functional linear array model. *Statistical Modelling* 15(3):279-300.

Bühlmann P & Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science* :477-505.

Bustillo V, Moatar F, Ducharne A, Thiéry D & Poirel A (2014) A multimodel comparison for assessing water temperatures under changing climate conditions via the

equilibrium temperature concept: case study of the Middle Loire River, France. *Hydrological Processes* 28(3):1507-1524.

Caissie D (2006) The thermal regime of rivers: a review. *Freshwater Biology* 51(8):1389-1406.

Caissie D, El-Jabi N & Bourgeois G (1998a) Évaluation du débit réservé par méthodes hydrologiques et hydrobiologiques. *Revue des sciences de l'eau/Journal of Water Science* 11(3):347-364.

Caissie D, El-Jabi N & Satish MG (2001) Modelling of maximum daily water temperatures in a small stream using air temperatures. *Journal of Hydrology* 251(1):14-28.

Caissie D, El-Jabi N & St-Hilaire A (1998b) Stochastic modelling of water temperatures in a small stream using air to water relations. *Canadian Journal of Civil Engineering* 25(2):250-260.

Carbonneau PE, Bergeron N & Lane SN (2005) Automated grain size measurements from airborne remote sensing for long profile measurements of fluvial grain sizes. *Water Resources Research* 41(11).

Carstensen B, Plummer M, Laara E & Hills M (2015) Epi: a package for statistical analysis in epidemiology. R package version 1.1. 71.).

CEHQ (2018) Historique des données de différentes stations hydrométriques, Centre d'expertise hydrique du Québec. *https://www.cehq.gouv.qc.ca/hydrometrie/index.htm*.

Chaouch M (2014) Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves. *IEEE Transactions on Smart Grid* 5(1):411-419.

Chebana F, Dabo-Niang S & Ouarda TB (2012) Exploratory functional flood frequency analysis and outlier detection. *Water Resources Research* 48(4).

Chen YD, Carsel RF, McCutcheon SC & Nutter WL (1998) Stream temperature simulation of forested riparian areas: I. Watershed-scale model development. *Journal of Environmental Engineering* 124(4):304-315.

Chenard JF & Caissie D (2008) Stream temperature modelling using artificial neural networks: application on Catamaran Brook, New Brunswick, Canada. *Hydrological Processes* 22(17):3361-3372.

Chiou J-M (2012) Dynamical functional prediction and classification, with application to traffic flow prediction. *The Annals of Applied Statistics* :1588-1614.

Ciarleglio A, Petkova E, Tarpey T & Ogden RT (2016) Flexible functional regression methods for estimating individualized treatment rules. *Stat* 5(1):185-199.

Clark ID, Lauriol B, Harwood L & Marschner M (2001) Groundwater contributions to discharge in a permafrost setting, Big Fish River, NWT, Canada. *Arctic, Antarctic, and Alpine Research* :62-69.

Cluis DA (1972) Relationship between stream water temperature and ambient air temperature. *Hydrology Research* 3(2):65-71.

Crisp D & Howson G (1982) Effect of air temperature upon mean water temperature in streams in the north Pennines and English Lake District. *Freshwater Biology* 12(4):359-367.

Cuevas A, Febrero M & Fraiman R (2002) Linear functional regression: the case of fixed design and functional response. *Canadian Journal of Statistics* 30(2):285-300.

Dabo-Niang S & Ferraty F (2008) *Functional and operatorial statistics.* Springer Science & Business Media,

Daigle A, Jeong DI & Lapointe MF (2015) Climate change and resilience of tributary thermal refugia for salmonids in eastern Canadian rivers. *Hydrological Sciences Journal* 60(6):1044-1063.

Damon J & Guillas S (2015) far: Modelization for Functional AutoRegressive Processes. *R Package, URL : https://cran.r-project.org/web/packages/far/index.html*.

Davey C & Lapointe M (2007) Sedimentary links and the spatial organization of Atlantic salmon (Salmo salar) spawning habitat in a Canadian Shield river. *Geomorphology* 83(1):82-96.

Davey CE (2005) Longitudinal trends in grain size, shear stress and sediment mobility along sedimentary links of a Canadian Shield River, Saguenay Region: A geomorphic perspective on assessing Atlantic salmon (Salmo salar) productivity in rivers.

DeGraaf D & Bain L (1986) Habitat use by and preferences of juvenile Atlantic salmon in two Newfoundland rivers. *Transactions of the American Fisheries Society* 115(5):671-681.

DeWeber JT & Wagner T (2014) A regional neural network ensemble for predicting mean daily river water temperature. *Journal of Hydrology* 517:187-200.

Dionne M & Dodson JJ (2002) Impact of exposure to a simulated predator (Mergus merganser) on the activity of juvenile Atlantic salmon (Salmo salar) in a natural environment. *Canadian journal of zoology* 80(11):2006-2013.

Dugdale SJ (2016) A practitioner's guide to thermal infrared remote sensing of rivers and streams: recent advances, precautions and considerations. *Wiley Interdisciplinary Reviews: Water* 3(2):251-268.

Dugdale SJ, Bergeron NE & St-Hilaire A (2013) Temporal variability of thermal refuges and water temperature patterns in an Atlantic salmon river. *Remote Sensing of Environment* 136:358-373.

Dynesius M & Nilsson C (1994) Fragmentation and flow regulation of river systems in the northern third of the world. *Science* 266(5186):753-762.

Elliott J (1985) Population regulation for different life-stages of migratory trout Salmo trutta in a Lake District stream, 1966-83. *The Journal of Animal Ecology* :617-638.

Elliott J & Elliott J (2006) A 35-year study of stock–recruitment relationships in a small population of sea trout: assumptions, implications and limitations for predicting targets. *Sea trout: biology, conservation and management* :257-278.

Elliott J & Elliott J (2010) Temperature requirements of Atlantic salmon Salmo salar, brown trout Salmo trutta and Arctic charr Salvelinus alpinus: predicting the effects of climate change. *Journal of fish biology* 77(8):1793-1817.

Farrell A (2002) Cardiorespiratory performance in salmonids during exercise at high temperature: insights into cardiovascular design limitations in fishes. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* 132(4):797-810.

Ferrari MR, Miller JR & Russell GL (2007) Modeling changes in summer temperature of the Fraser River during the next century. *Journal of Hydrology* 342(3-4):336-346.

Ferraty F & Vieu P (2006) *Nonparametric functional data analysis: theory and practice.* Springer Science & Business Media,

Feyrer F, Nobriga ML & Sommer TR (2007) Multidecadal trends for three declining fish species: habitat patterns and mechanisms in the San Francisco Estuary, California, USA. *Canadian Journal of Fisheries and Aquatic Sciences* 64(4):723-734.

Finstad AG, Forseth T, Næsje TF & Ugedal O (2004a) The importance of ice cover for energy turnover in juvenile Atlantic salmon. *Journal of Animal Ecology* 73(5):959-966.

Finstad AG, Ugedal O, Forseth T & Næsje TF (2004b) Energy-related juvenile winter mortality in a northern population of Atlantic salmon (Salmo salar). *Canadian Journal of Fisheries and Aquatic Sciences* 61(12):2358-2368.

Freeman MC, Bowen ZH & Crance JH (1997) Transferability of habitat suitability criteria for fishes in warmwater streams. *North American Journal of Fisheries Management* 17(1):20-31.

Gabler HM & Amundsen PA (1999) Resource partitioning between Siberian sculpin (Cottus poecilopus Heckel) and Atlantic salmon parr (Salmo salar L.) in a sub-Arctic river, northern Norway. *Ecology of Freshwater Fish* 8(4):201-208.

Gasparrini A, Armstrong B & Kenward MG (2010) Distributed lag non-linear models. *Statistics in medicine* 29(21):2224-2234.

Gibson R (1993) The Atlantic salmon in fresh water: spawning, rearing and production. *Reviews in fish biology and fisheries* 3(1):39-73.

Gibson RJ, Haedrich RL & Wernerheim CM (2005) Loss of fish habitat as a consequence of inappropriately constructed stream crossings. *Fisheries* 30(1):10-17.

GIEC (2013) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovern- mental Panel on Climate Change. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.*

Girard P, Boisclair D & Leclerc M (2003) The effect of cloud cover on the development of habitat quality indices for juvenile Atlantic salmon (Salmo salar). *Canadian Journal of Fisheries and Aquatic Sciences* 60(11):1386-1397.

Goia A, May C & Fusai G (2010) Functional clustering and linear regression for peak load forecasting. *International Journal of Forecasting* 26(4):700-711.

Goldsmith JFS, Lei Huang, Julia Wrobel, Jonathan Gellar, Jaroslaw, Harezlak MWM, Bruce Swihart, Luo Xiao, Ciprian Crainiceanu and Philip T. & Reiss (2016) refund: Regression with Functional Data. R package version 0.1-16.).

Gorecki T & Smaga L (2017) fdANOVA: Analysis of Variance for Univariate and Multivariate Functional Data. *R Package, URL : https://CRAN.R-project.org/package=fdANOVA*.

Graham C & Harrod C (2009) Implications of climate change for the fishes of the British Isles. *Journal of Fish Biology* 74(6):1143-1205.

Grbić R, Kurtagić D & Slišković D (2013) Stream water temperature prediction based on Gaussian process regression. *Expert systems with applications* 40(18):7407-7414.

Guay J, Boisclair D, Leclerc M & Lapointe M (2003) Assessment of the transferability of biological habitat models for Atlantic salmon parr (Salmo salar). *Canadian Journal of Fisheries and Aquatic Sciences* 60(11):1398-1408.

Guay J, 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(10):2065-2075.

Handeland SO, Imsland AK & Stefansson SO (2008) The effect of temperature and fish size on growth, feed intake, food conversion efficiency and stomach evacuation rate of Atlantic salmon post-smolts. *Aquaculture* 283(1):36-42.

Hannah DM, Malcolm IA, Soulsby C & Youngson AF (2008) A comparison of forest and moorland stream microclimate, heat exchanges and thermal dynamics. *Hydrological Processes* 22(7):919-940.

Harnish RA, Sharma R, McMichael GA, Langshaw RB & Pearsons TN (2014) Effect of hydroelectric dam operations on the freshwater productivity of a Columbia River fall Chinook salmon population. *Canadian journal of fisheries and aquatic sciences* 71(4):602-615.

Hastie T & Tibshirani R (1990) *Generalized additive models.* Wiley Online Library,

Hebert C, Caissie D, Satish MG & El-Jabi N (2014) Modeling of hourly river water temperatures using artificial neural networks. *Water Quality Research Journal* 49(2):144-162.

Hedger R, Dodson J, Bergeron N & Caron F (2004) Quantifying the effectiveness of regional habitat quality index models for predicting densities of juvenile Atlantic salmon (Salmo salar L.). *Ecology of freshwater fish* 13(4):266-275.

Hedger R, Dodson J, Bergeron N & Caron F (2005) Habitat selection by juvenile Atlantic salmon: the interaction between physical habitat and abundance. *Journal of Fish Biology* 67(4):1054-1071.

Hedger R, Dodson J, Bourque J, Bergeron N & Carbonneau P (2006) Improving models of juvenile Atlantic salmon habitat use through high resolution remote sensing. *ecological modelling* 197(3):505-511.

Hedger RD, Sundt-Hansen LE, Forseth T, Ugedal O, Diserud OH, Kvambekk ÅS & Finstad AG (2013) Predicting climate change effects on subarctic–Arctic populations of Atlantic salmon (Salmo salar). *Canadian journal of fisheries and aquatic sciences* 70(2):159-168.

Heggenes J (1990) Habitat utilization and preferences in juvenile Atlantic salmon (Salmo salar) in streams. *River Research and Applications* 5(4):341-354.

Heggenes J, Baglinière J & Cunjak R (1995) Note de synthèse sur la sélection de niche spatiale et la compétition chez le jeune saumon atlantique (Salmo salar) et la truite commune (Salmo trutta) en milieu lotique. *Bulletin Français de la Pêche et de la Pisciculture* (337-338-339):231-239.

Heggenes J & Gunnar Dokk J (2001) Contrasting temperatures, waterflows, and light: seasonal habitat selection by young Atlantic salmon and brown trout in a boreonemoral river. *River Research and Applications* 17(6):623-635.

Heland M & Dumas J (1994) Ecologie et comportement des juvéniles. *Le saumon atlantique* :29-46.

Héland M, Gaudin P & Bardonnet A (1995) Mise en place des premiers comportements et utilisation de l'habitat après l'émergence chez les salmonidés d'eau courante. *Bulletin Français de la Pêche et de la Pisciculture* (337-338-339):191-197.

Hinz LC & Wiley MJ (1998) *Growth and production of juvenile trout in Michigan streams: influence of potential ration and temperature.* Michigan Department of Natural Resources, Fisheries Division,

Hoover KD (2003) Nonstationary time series, cointegration, and the principle of the common cause. *The British Journal for the Philosophy of Science* 54(4):527-551.

Horváth L & Kokoszka P (2012) *Inference for functional data with applications.* Springer Science & Business Media,

Hvidsten N, Diserud O, Jensen A, Jensås J, Johnsen B & Ugedal O (2015) Water discharge affects Atlantic salmon Salmo salar smolt production: a 27 year study in the River Orkla, Norway. *Journal of fish biology* 86(1):92-104.

Ibarra AA, Gevrey M, Park Y-S, Lim P & Lek S (2003) Modelling the factors that influence fish guilds composition using a back-propagation network: assessment of metrics for indices of biotic integrity. *Ecological Modelling* 160(3):281-290.

Ivanescu AE, Staicu A-M, Scheipl F & Greven S (2015) Penalized function-on-function regression. *Computational Statistics* 30(2):539-568.

Jensen A & Johnsen B (1999) The functional relationship between peak spring floods and survival and growth of juvenile Atlantic salmon (Salmo salar) and brown trout (Salmo trutta). *Functional Ecology* 13(6):778-785.

Jeppesen E & Iversen TM (1987) Two simple models for estimating daily mean water temperatures and diel variations in a Danish low gradient stream. *Oikos* :149-155.

Johnston P & Bergeron N (2010) Variation of juvenile Atlantic salmon (Salmo salar) body composition along sedimentary links. *Ecology of freshwater fish* 19(2):187-196.

Jones P, Lister D, Osborn T, Harpham C, Salmon M & Morice C (2012) Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research: Atmospheres* 117(D5).

Jonsson B & Jonsson N (2009) A review of the likely effects of climate change on anadromous Atlantic salmon Salmo salar and brown trout Salmo trutta, with particular reference to water temperature and flow. *Journal of fish biology* 75(10):2381-2447.

Jonsson B & Jonsson N (2017) Fecundity and water flow influence the dynamics of Atlantic salmon. *Ecology of Freshwater Fish* 26(3):497-502.

Jorde K, Schneider M, Peter A & Zoellner F (2001) Fuzzy based models for the evaluation of fish habitat quality and instream flow assessment. *Proceedings of the 3rd international symposium on environmental hydraulics.* p 27-28.

Jourdonnais J, Walsh R, Pickett F & Goodman D (1992) Structure and calibration strategy for a water temperature model of the lower Madison River, Montana. *Rivers* 3(3):153-169.

Jowett IG & Davey AJ (2007) A comparison of composite habitat suitability indices and generalized additive models of invertebrate abundance and fish presence–habitat availability. *Transactions of the American Fisheries Society* 136(2):428-444.

Kaushal SS, Likens GE, Jaworski NA, Pace ML, Sides AM, Seekell D, Belt KT, Secor DH & Wingate RL (2010) Rising stream and river temperatures in the United States. *Frontiers in Ecology and the Environment* 8(9):461-466.

Kelleher C, Wagener T, Gooseff M, McGlynn B, McGuire K & Marshall L (2012) Investigating controls on the thermal sensitivity of Pennsylvania streams. *Hydrological Processes* 26(5):771-785.

Kim M & Lapointe M (2011) Regional variability in Atlantic salmon (Salmo salar) riverscapes: a simple landscape ecology model explaining the large variability in size of salmon runs across Gaspé watersheds, Canada. *Ecology of Freshwater Fish* 20(1):144-156.

Kim MS (2009) *The controls of sedimentary links on the spatial distribution of Atlantic salmon (Salmo salar) juveniles and spawning activity along rivers in the Gaspé Peninsula, Canada*. (McGill University).

Klemetsen A, Amundsen PA, Dempson J, Jonsson B, Jonsson N, O'connell M & Mortensen E (2003) Atlantic salmon Salmo salar L., brown trout Salmo trutta L. and Arctic charr Salvelinus alpinus (L.): a review of aspects of their life histories. *Ecology of freshwater fish* 12(1):1-59.

Kondolf G (1997) Hungry water—effects of dams and gravel mining on river channels and floodplains. *Aggregate resources—a global perspective. AA Balkema, Vermont* :113-129.

Kothandaraman V (1971) Analysis of water temperature variations in large river. *Journal of the Sanitary Engineering Division* 97(1):19-31.

Kwak J, St-Hilaire A & Chebana F (2017) A comparative study for water temperature modelling in a small basin, the Fourchue River, Quebec, Canada. *Hydrological Sciences Journal* 62(1):64-75.

Laanaya F, St-Hilaire A & Gloaguen E (2017) Water temperature modelling: comparison between the generalized additive model, logistic, residuals regression and linear regression models. *Hydrological Sciences Journal* 62(7):1078-1093.

Labonne J, Allouche S & Gaudin P (2003) Use of a generalised linear model to test habitat preferences: the example of Zingel asper, an endemic endangered percid of the River Rhone. *Freshwater Biology* 48(4):687-697.

Lackey RT (2003) Pacific Northwest salmon: forecasting their status in 2100. *Reviews in fisheries Science* 11(1):35-88.

Lanthier G, Bédard M-E, Lapointe M & Boisclair D (2014) Assessment of the structural role of sedimentary links on the spatial distribution of periphyton and fish in a Canadian Shield river. *Aquatic sciences* 77(1):141-152.

Larabi S, St-Hilaire A, Chebana F & Latraverse M (2017) Multi-Criteria Process-Based Calibration Using Functional Data Analysis to Improve Hydrological Model Realism. *Water Resources Management* :1-17.

Latulippe C, Lapointe MF & Talbot T (2001) Visual characterization technique for gravel-cobble river bed surface sediments; validation and environmental applications Contribution to the programme of CIRSA(Centre Interuniversitaire de Recherche sur le Saumon Atlantique). *Earth Surface Processes and Landforms* 26(3):307-318.

Leclerc M, Boudreau P, Bechara J, Belzile L & Villeneuve D (1994) Modélisation de la dynamique de l'habitat des jeunes stades de saumon atlantique (Salmo salar) de la rivière Ashuapmushuan (Québec, Canada). *Bulletin Français de la Pêche et de la Pisciculture* (332):11-32.

Lee RM & Rinne JN (1980) Critical thermal maxima of five trout species in the southwestern United States. *Transactions of the American Fisheries Society* 109(6):632-635.

Legendre P & Legendre LF (2012) *Numerical ecology.* Elsevier,

Lehmkuhl D (1972) Change in thermal regime as a cause of reduction of benthic fauna downstream of a reservoir. *Journal of the Fisheries Board of Canada* 29(9):1329-1332.

Lessard JL & Hayes DB (2003) Effects of elevated water temperature on fish and macroinvertebrate communities below small dams. *River research and applications* 19(7):721-732.

Li H, Deng X, Kim DY & Smith EP (2014) Modeling maximum daily temperature using a varying coefficient regression model. *Water Resources Research* 50(4):3073-3087.

Liu B, Yang D, Ye B & Berezovskaya S (2005a) Long-term open-water season stream temperature variations and changes over Lena River Basin in Siberia. *Global and Planetary Change* 48(1-3):96-111.

Liu C, Berry PM, Dawson TP & Pearson RG (2005b) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28(3):385-393.

Mackey A & Berrie A (1991) The prediction of water temperatures in chalk streams from air temperatures. *Hydrobiologia* 210(3):183-189.

Maheu A, St-Hilaire A, Caissie D, El-Jabi N, Bourque G & Boisclair D (2016) A regional analysis of the impact of dams on water temperature in medium-size rivers in eastern Canada. *Canadian Journal of Fisheries and Aquatic Sciences* 73(12):1885-1897.

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(2):218-228.

Manly B, McDonald L, Thomas DL, McDonald TL & Erickson WP (2007) *Resource selection by animals: statistical design and analysis for field studies.* Springer Science & Business Media,

Mantua N, Tohver I & Hamlet A (2010) Climate change impacts on streamflow extremes and summertime stream temperature and their possible consequences for freshwater salmon habitat in Washington State. *Climatic Change* 102(1-2):187-223.

Masselot P (2017) *Approches statistiques avancées pour la modélisation des séries chronologiques en régression, appliquées à l'épidémiologie environnementale*. (Université du Québec, Institut national de la recherche scientifique).

Masselot P, Dabo-Niang S, Chebana F & Ouarda TB (2016) Streamflow forecasting using functional regression. *Journal of Hydrology* 538:754-766.

McDonald S, Koulis T, Ehn J, Campbell K, Gosselin M & Mundy C (2015) A functional regression model for predicting optical depth and estimating attenuation coefficients in sea-ice covers near Resolute Passage, Canada. *Annals of Glaciology* 56(69):147-154.

McLean MW, Hooker G, Staicu A-M, Scheipl F & Ruppert D (2014) Functional generalized additive models. *Journal of Computational and Graphical Statistics* 23(1):249-269.

Meyer MJ, Coull BA, Versace F, Cinciripini P & Morris JS (2015) Bayesian function-on-function regression for multilevel functional data. *Biometrics* 71(3):563-574.

Millidine K, Malcolm I & Fryer R (2016) Assessing the transferability of hydraulic habitat models for juvenile Atlantic salmon. *Ecological Indicators* 69:434-445.

Millidine K, Malcolm I, Gibbins C, Fryer R & Youngson A (2012) The influence of canalisation on juvenile salmonid habitat. *Ecological indicators* 23:262-273.

Mocq J, St-Hilaire A & Cunjak RA (2013) Assessment of Atlantic salmon (Salmo salar) habitat quality and its uncertainty using a multiple-expert fuzzy model applied to the Romaine River (Canada). *Ecological modelling* 265:14-25.

Mocq J, St-Hilaire A & Cunjak RA (2018) Do habitat measurements in the vicinity of Atlantic salmon (Salmo salar) parr matter? *Fisheries Management and Ecology* 25(1):22-30.

Mohseni O, Stefan HG & Erickson TR (1998) A nonlinear regression model for weekly stream temperatures. *Water Resources Research* 34(10):2685-2692.

Moore R, Spittlehouse D & Story A (2005) Riparian microclimate and stream temperature response to forest harvesting: a review. *JAWRA Journal of the American Water Resources Association* 41(4):813-834.

Morantz D, Sweeney R, Shirvell C & Longard D (1987) Selection of microhabitat in summer by juvenile Atlantic salmon (Salmo salar). *Canadian Journal of Fisheries and Aquatic Sciences* 44(1):120-129.

Morin G, Couillard D, Cluis D, Jones HG & Gauthier J-M (1987) Prévision des températures de l'eau en rivière à l'aide d'un modèle conceptual. *Hydrological sciences journal* 32(1):31-41.

Morin G, Fortin J-P, Lardeau J-P, Sochanska W & Paquette S (1981) *Modèle CEQUEAU: manuel d'utilisation.* INRS-eau,

Morrill JC, Bales RC & Conklin MH (2005) Estimating stream temperature from air temperature: implications for future water quality. *Journal of Environmental Engineering* 131(1):139-146.

Morris JS (2015) Functional regression. *Annual Review of Statistics and Its Application* 2:321-359.

Morrison J, Quick MC & Foreman MG (2002) Climate change in the Fraser River watershed: flow and temperature projections. *Journal of Hydrology* 263(1):230-244.

Murchie K, Hair K, Pullen C, Redpath T, Stephens H & Cooke S (2008) Fish response to modified flow regimes in regulated rivers: research methods, effects and opportunities. *River Research and Applications* 24(2):197-217.

Nagelkerke NJ (1991) A note on a general definition of the coefficient of determination. *Biometrika* 78(3):691-692.

Nemerow NL (1991) Stream, lake, estuary, and ocean pollution.

Neter J, Kutner MH, Nachtsheim CJ & Wasserman W (1996) *Applied linear statistical models.* Irwin Chicago,

Newcombe CP & Jensen JO (1996) Channel suspended sediment and fisheries: a synthesis for quantitative assessment of risk and impact. *North American Journal of Fisheries Management* 16(4):693-727.

NOAA (2017) *National centers for environmental information.*https://www.ncdc.noaa.gov/cdo-web/,

Noakes DJ, Beamish RJ & Kent ML (2000) On the decline of Pacific salmon and speculative links to salmon farming in British Columbia. *Aquaculture* 183(3):363-386.

Nyqvist D, Greenberg LA, Goerig E, Calles O, Bergman E, Ardren WR & Castro-Santos T (2017) Migratory delay leads to reduced passage success of Atlantic salmon smolts at a hydroelectric dam. *Ecology of Freshwater Fish* 26(4):707-718.

Olden JD & Jackson DA (2001) Fish–habitat relationships in lakes: gaining predictive and explanatory insight by using artificial neural networks. *Transactions of the American Fisheries Society* 130(5):878-897.

Olden JD & Jackson DA (2002a) A comparison of statistical approaches for modelling fish species distributions. *Freshwater biology* 47(10):1976-1995.

Olden JD & Jackson DA (2002b) Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological modelling* 154(1-2):135-150.

Olden JD & Naiman RJ (2010) Incorporating thermal regimes into environmental flows assessments: modifying dam operations to restore freshwater ecosystem integrity. *Freshwater Biology* 55(1):86-107.

Orth DJ & Maughan OE (1982) Evaluation of the incremental methodology for recommending instream flows for fishes. *Transactions of the American Fisheries Society* 111(4):413-445.

Parmesan C & Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature* 421(6918):37.

Pilgrim JM, Fang X & Stefan HG (1998) Stream temperature correlations with air temperatures in Minnesota: implications for climate warming. *JAWRA Journal of the American Water Resources Association* 34(5):1109-1121.

Pinfold G (2011) Economic Value of Wild Atlantic Salmon. *Prepared by Gardner Pinfold. Accessed online: http://asf.ca/gardner-pinfold-report.html*.

Piotrowski AP, Napiorkowski MJ, Napiorkowski JJ & Osuch M (2015) Comparing various artificial neural network types for water temperature prediction in rivers. *Journal of Hydrology* 529:302-315.

Poirel A, Gailhard J & Capra H (2010) Influence des barrages-réservoirs sur la température de l’eau: exemple d’application au bassin versant de l’Ain. *La Houille Blanche* (4):72-79.

Preece RM & Jones HA (2002) The effect of Keepit Dam on the temperature regime of the Namoi River, Australia. *River Research and Applications* 18(4):397-414.

Prévost L, Levesque D & Plamondon AP (2002) *Méthodologie pour évaluer l'effet de l'installation d'un ponceau sur le substrat des frayères de l'omble de fontaine (Salvelinus fontinalis).* Ministère des ressources naturelles,

Puffer M, Berg OK, Huusko A, Vehanen T & Einum S (2017) Effects of intra-and interspecific competition and hydropeaking on growth of juvenile Atlantic salmon (Salmo salar). *Ecology of freshwater fish* 26(1):99-107.

Quenouille MH (1949) Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological)* 11(1):68-84.

R Core Team (2017) R language definition. *Vienna, Austria: R foundation for statistical computing*.

Railsback SF (2016) Why it is time to put PHABSIM out to pasture. *Fisheries* 41(12):720-725.

Ramsay J (1982) When the data are functions. *Psychometrika* 47(4):379-396.

Ramsay J, Wickham H, Graves S & Hooker G (2013) fda: Functional data analysis. *R Package, URL http://cran. r-project. org/package= fda*.

Ramsay JO (2006) *Functional data analysis.* Wiley Online Library,

Ramsay JO, Hooker G & Graves S (2009) *Functional data analysis with R and MATLAB.* Springer Science & Business Media,

Requena AI, Chebana F & Ouarda TB (2018) A functional framework for flow-duration-curve and daily streamflow estimation at ungauged sites. *Advances in Water Resources*.

Rice S & Church M (1998) Grain size along two gravel-bed rivers: statistical variation, spatial pattern and sedimentary links. *Earth Surface Processes and Landforms* 23(4):345-363.

Rice SP, Greenwood MT & Joyce CB (2001) Tributaries, sediment sources, and the longitudinal organisation of macroinvertebrate fauna along river systems. *Canadian Journal of Fisheries and Aquatic Sciences* 58(4):824-840.

RivTemp (2018) *Réseau de stations de mesure de température des rivières de l'Est du Canada.* http://RivTemp.ca,

Rohde R, Muller R, Jacobsen R, Muller E, Perlmutter S, Rosenfeld A, Wurtele J, Groom D & Wickham C (2013) A new estimate of the average Earth surface land temperature spanning 1753 to 2011. Geoinfor Geostat Overview 1: 1. *of* 7:2.

Saumon Québec (2018) *Le saumon atlantique.* https://www.saumonquebec.com/decouvrir/s-initier/les-especes/le-saumon-atlantique/

Scott D & Shirvell C (1987) A critique of the instream flow incremental methodology and observations on flow determination in New Zealand. *Regulated streams,* Springer. p 27-43.

Shao J (1993) Linear model selection by cross-validation. *Journal of the American statistical Association* 88(422):486-494.

Shirvell C (1994) Effect of changes in streamflow on the microhabitat use and movements of sympatric juvenile coho salmon (Oncorhynchus kisutch) and chinook salmon

(O. tshawytscha) in a natural stream. *Canadian Journal of Fisheries and Aquatic Sciences* 51(7):1644-1652.

Sigholt T & Finstad B (1990) Effect of low temperature on seawater tolerance in Atlantic salmon (Salmo salar) smolts. *Aquaculture* 84(2):167-172.

Singer EE & Gangloff MM (2011) Effects of a small dam on freshwater mussel growth in an Alabama (USA) stream. *Freshwater Biology* 56(9):1904-1915.

Smith J, Bérubé F & Bergeron N (2005) A field application of particle image velocimetry (PIV) for the measurement of surface flow velocities in aquatic habitat studies. *26th Canadian Symposium on Remote Sensing.*

St-Hilaire A, Boucher M-A, Chebana F, Ouellet-Proulx S, Zhou QX, Larabi S, Dugdale S & Latraverse M (2015) Breathing a new life to an older model: the CEQUEAU tool for flow and water temperature simulations and forecasting. *Proceedings of the 22nd Canadian Hydrotechnical Conference, Montreal, QC, Canada.*

St-Hilaire A, Morin G, El-Jabi N & Caissie D (2000) Water temperature modelling in a small forested stream: implication of forest canopy and soil temperature. *Canadian Journal of Civil Engineering* 27(6):1095-1108.

St-Hilaire A, Ouarda TB, Bargaoui Z, Daigle A & Bilodeau L (2012) Daily river water temperature forecast model with a k-nearest neighbour approach. *Hydrological Processes* 26(9):1302-1310.

Stefan HG & Preud'homme EB (1993) Stream temperature estimation from air temperature. *JAWRA Journal of the American Water Resources Association* 29(1):27-45.

Stewart-Koster B, Olden JD & Gido KB (2014) Quantifying flow–ecology relationships with functional linear models. *Hydrological sciences journal* 59(3-4):629-644.

Suhaila J & Yusop Z (2017) Spatial and temporal variabilities of rainfall data using functional data analysis. *Theoretical and Applied Climatology* 129(1-2):229-242.

Sundt-Hansen L, Hedger R, Ugedal O, Diserud O, Finstad A, Sauterleute J, Tøfte L, Alfredsen K & Forseth T (2018) Modelling climate change effects on Atlantic salmon: Implications for mitigation in regulated rivers. *Science of The Total Environment* 631:1005-1017.

Svenning M-A, Sandem K, Halvorsen M, Kanstad-Hanssen Ø, Falkegård M & Borgstrøm R (2016) Change in relative abundance of Atlantic salmon and Arctic charr in Veidnes River, Northern Norway: a possible effect of climate change? *Hydrobiologia* 783(1):145-158.

Ternynck C, Ben Alaya MA, Chebana F, Dabo-Niang S & Ouarda TB (2016) Streamflow hydrograph classification using functional data analysis. *Journal of Hydrometeorology* 17(1):327-344.

Tetzlaff D, Soulsby C, Youngson A, Gibbins C, Bacon P, Malcolm I & Langan S (2005) Variability in stream discharge and temperature: a preliminary assessment of the implications for juvenile and spawning Atlantic salmon. *Hydrology and Earth System Sciences* 9(3):193-208.

Trombulak SC & Frissell CA (2000) Review of ecological effects of roads on terrestrial and aquatic communities. *Conservation biology* 14(1):18-30.

Tukey JW (1977) *Exploratory data analysis.* Reading, Mass.,

USGS (2017) *USGS Daily Values Web Service.*https://waterservices.usgs.gov/rest/DV-Test-Tool.html,

Van Vliet M, Ludwig F, Zwolsman J, Weedon G & Kabat P (2011) Global river temperatures and sensitivity to atmospheric warming and changes in river flow. *Water Resources Research* 47(2).

Vehanen T (2003) Adaptive flexibility in the behaviour of juvenile Atlantic salmon: short-term responses to food availability and threat from predation. *Journal of Fish Biology* 63(4):1034-1045.

Verspoor E & Jordan W (1989) Genetic variation at the Me-2 locus in the Atlantic salmon within and between rivers: evidence for its selective maintenance. *Journal of Fish Biology* 35(sA):205-213.

Wang S, Jank W & Shmueli G (2008) Explaining and forecasting online auction prices and their dynamics using functional data analysis. *Journal of Business & Economic Statistics* 26(2):144-160.

Webb B (1996) Trends in stream and river temperature. *Hydrological processes* 10(2):205-226.

Webb B & Walling D (1993) Temporal variability in the impact of river regulation on thermal regime and some biological implications. *Freshwater Biology* 29(1):167-182.

Webb B & Walling D (1997) Complex summer water temperature behaviour below a UK regulating reservoir. *River Research and Applications* 13(5):463-477.

Webb BW, Hannah DM, Moore RD, Brown LE & Nobilis F (2008) Recent advances in stream and river temperature research. *Hydrological processes* 22(7):902-918.

Webb BW & Nobilis F (2007) Long-term changes in river temperature and the influence of climatic and hydrological factors. *Hydrological Sciences Journal* 52(1):74-85.

Wehrly KE, Brenden TO & Wang L (2009) A comparison of statistical approaches for predicting stream temperatures across heterogeneous landscapes. *JAWRA Journal of the American Water Resources Association* 45(4):986-997.

Wehrly KE, Wang L & Mitro M (2007) Field-based estimates of thermal tolerance limits for trout: Incorporating exposure time and temperature fluctuation. *Transactions of the American Fisheries Society* 136(2):365-374.

Wilkerson E, Hagan JM, Siegel D & Whitman AA (2006) The effectiveness of different buffer widths for protecting headwater stream temperature in Maine. *Forest Science* 52(3):221-231.

Wood PJ & Armitage PD (1997) Biological effects of fine sediment in the lotic environment. *Environmental management* 21(2):203-217.

Yi Y, Cheng X, Yang Z, Wieprecht S, Zhang S & Wu Y (2017) Evaluating the ecological influence of hydraulic projects: A review of aquatic habitat suitability models. *Renewable and Sustainable Energy Reviews* 68:748-762.