CORRECTION DU DÉBIT EN PRÉSENCE DE GLACE ET ESTIMATION DE L'ÉPAISSEUR DE LA GLACE DE RIVIÈRE, APPLICATION À QUATRE RIVIÈRES DU CANADA

Rapport de recherche No R-886 *Septembre 2006*

CORRECTION DU DÉBIT EN PRÉSENCE DE GLACE ET ESTIMATION DE L'ÉPAISSEUR DE LA GLACE DE RIVIÈRE, APPLICATION À QUATRE RIVIÈRES DU CANADA

Rapport préparé à l'attention de:

Raymond Bourdages

Service des Relevés hydrologiques

Environnement Canada

373 Sussex Drive, Block E-101

Ottawa, Ontario, Canada, K1A 0H3

par:

Karem Chokmani

Taha B.M.J. Ouarda

Chaire Hydro-Québec/CRSNG/Alcan en Hydrologie Statistique

Chaire du Canada en estimation des variables hydrologiques

Institut National de la Recherche Scientifique, INRS-ETE

490, rue de la Couronne, Québec (Québec) G1K 9A9

Rapport de recherche N° R-886

Septembre 2006

ÉQUIPE DE RECHERCHE

Ont participé à la réalisation de cette étude:

Institut National de la Recherche Scientifique, INRS-ETE

Karem Chokmani

Taha B.M.J. Ouarda

Zeljka Ristic-Rudolf

Environnement Canada

Raymond Bourdages

AVANT PROPOS

L'équipe de recherche de la chaire industrielle en Hydrologie statistique / Chaire du Canada en estimation des variables hydrologiques de l'INRS-ETE a été mandatée par les services d'Environnement Canada, Service des Relevés hydrologiques, afin de développer une méthodologie de correction du débit de rivières durant la période hivernale, en présence de glace, et pour l'estimation de l'épaisseur de la glace de rivière. Le mandat a porté aussi sur le développement d'un outil informatique pour la visualisation et la calibration des données hydrométriques et météorologiques disponibles ainsi que l'estimation en temps réel du débit corrigé et de l'épaisseur de glace.

Nous tenons à remercier M. Raymond Bourdages pour avoir fourni les données utilisées dans cette étude.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

1 INTRODUCTION

Une proportion importante des rivières canadiennes est affectée par l'effet de glace. Les séries de débits de ces rivières correspondant à la période hivernale sont souvent de qualité inférieure à celle correspondant au reste de l'année ; i.e. le débit estimé par la courbe de tarage ne correspond pas au débit réel dans la rivière à cause de la présence de glace dans la rivière (glace de surface, glace de fond, glace en aiguilles, etc.).

En général, la courbe de tarage est construite dans une section stable de la rivière à partir de plusieurs observations niveau-débit durant la période de débit libre (non affectée par la glace). Cependant, cette courbe de tarage ne peut pas être représentative de la période hivernale à cause du changement de la section d'écoulement même et des conditions très variables qui peuvent exister quand l'écoulement est affecté par la présence de glace. Pour remédier à cette situation, les services ayant la gestion du réseau de stations hydrométriques effectuent des jaugeages durant la période de présence de glace pour estimer le débit réel qui s'écoule dans les rivières. Ensuite, les débits pendant le reste de la période hivernale sont corrigés par interpolation tout en tenant compte des événements pluvieux ou des réchauffements de température qui peuvent avoir eu lieu ainsi que du comportement hydrologique des autres rivières de la région. Cette approche mène généralement à des résultats satisfaisants mais risque d'introduire des erreurs assez importantes lors de fonte hivernale, embâcles de glace, etc.

En effet, pour chaque rivière, quelques jaugeages sont effectués durant la période de présence de glace. Ces jaugeages représentent la base de l'interpolation des débits pour tout le reste de l'hiver (une période qui excède souvent les 4 mois). D'autre part, la méthodologie n'est pas reproductible étant donné qu'elle contient un degré de subjectivité assez important. En effet, l'application successive de la technique par différents individus aux mêmes données de débits donne souvent des résultats assez différents. Il a aussi été observé que les hypothèses de correction du débit évoluent à travers le temps pour un même individu. Finalement, la méthodologie adoptée présentement ne permet pas l'obtention des estimations des débits de

rivière d'une façon continue durant la saison hivernale. En effet, ces débits ne sont disponibles qu'à la fin de l'hiver, après correction globale de toutes les séries des rivières affectées.

Par ailleurs, la présence de glace affecte l'exploitation hydroélectrique des cours d'eau en réduisant leur capacité d'écoulement, bloquant les prises d'eau, par exemple, ou en réduisant la hauteur de chute. Également, la glace de rivière est souvent à l'origine d'inondations dues aux embâcles. Elle perturbe le trafic sur les voies navigables et affecte la sécurité de certaines activités récréotouristiques telle que la motoneige ou la pêche. Par conséquent, il est important de prédire l'évolution de l'épaisseur de glace de rivière afin de pouvoir prévoir l'impact de la présence de glace sur de tels usages.

Ouarda *et al.* (2000) ont effectué une étude critique de l'approche adoptée présentement pour l'estimation des débits en présence de glace et ont proposé le développement d'une approche efficace, objective et reproductible. L'étude a présenté une revue de littérature complète des différentes méthodes utilisées pour la correction du débit en présence des glaces. Il en ressort que trois méthodes analytiques à savoir la régression multiple, le filtre de Kalman et le réseau de neurones artificiels sont parmi les plus prometteuses pour la correction du débit en présence de glace. Cependant, ces méthodes nécessitent l'accès à des données de bonne qualité pour leur assurer une bonne calibration.

Également, Seidou *et al.* (2005), lors d'une étude effectuée à l'INRS-ETE pour la modélisation de la croissance de glace de lac, ont démontré la supériorité des réseaux de neurones artificiels (RNA) pour le suivi de l'épaisseur de glace. Ils ont comparé les RNA à deux modèles thermodynamiques (basés sur la loi de Stefan). Il s'est avéré que les RNA sont plus flexibles et peuvent efficacement remplacer ces modèles, surtout en présence de données en quantité et en qualité suffisante.

Dans la présente étude, il a été question de développer des algorithmes robustes pour la correction du débit sous glace utilisant les méthodes les plus performantes identifiées par Ouarda *et al.* (2000). Il a été également question de développer une méthodologie pour l'estimation de l'épaisseur de glace de rivière en se basant sur les travaux de Seidou *et al.* (2005). Ainsi, un modèle de réseaux de neurones artificiels (RNA) et deux modèles régressifs (régression multiple

et régression stepwise) ont été développés pour la correction du débit en présence de glace d'une part et pour l'estimation de l'épaisseur de la glace de l'autre part. Les différentes approches étudiées ont été calibrées à l'aide d'une combinaison de variables hydrométriques et météorologiques facilement disponibles en quantité et qualité adéquate.

Afin d'atteindre ces objectifs, les différentes méthodes d'estimation ont été implantées dans un outil informatique développé dans l'environnement Matlab (The MathWorks, 2000). Cet outil, baptisé *UNICCO* (Under Ice Correction), permet entre autres de visualiser les données utilisées dans la calibration, de calibrer les différents modèles et de calculer les débits corrigés pour l'effet de glace. Il s'agit de la version 2.0 de l'outil, la version 1.0 ayant été développée lors d'une étude réalisée par l'équipe de la chaire pour le compte du Environnement Canada, Colombie Britannique (Chokmani *et al.*, 2003). Lors de cette étude il a été souligné l'intérêt de l'utilisation des RNA pour la correction du débit de rivière en présence de glace.

Par ailleurs, les approches étudiées pour la correction des débits sous glace et l'estimation de l'épaisseur de glace ont été appliquées aux données de quatre rivières canadiennes à savoir :

- la rivière Athabasca à Fort McMurray (07DA001);
- la rivière Pembina à Entwistle (07BB002);
- la rivière Clearwater à Draper (07CD001);
- la rivière North Saskatchewan à Edmonton (05DF001).

Le premier chapitre de présent rapport a été consacré à exposer les fondements théoriques des méthodes d'estimation implantées dans le logiciel *UNICCO 2.0.* Le deuxième chapitre est dédié à la comparaison entre les différentes approches à travers un cas pratique d'application portant sur des données recueillies sur les quatre rivières sélectionnées.

2 MÉTHODES D'ESTIMATION

2.1 RÉSEAUX NEURONAUX ARTIFICIELS

2.1.1 Introduction

L'évolution phénoménale des outils informatiques a largement contribué au développement des réseaux de neurones. Les réseaux de neurones font actuellement l'objet de beaucoup de recherches, en raison de leurs propriétés intéressantes d'apprentissage de modèles non linéaires et leurs possibilités d'application à des problèmes de classification, de diagnostic, de prédiction et de contrôle de procédés. En plus, un réseau de neurones permet d'optimiser la meilleure approximation non linéaire basée sur la structure complexe du réseau, et ceci sans aucune contrainte sur la linéarité ou sur la non linéarité spécifiée a priori comme dans les méthodes usuelles de régression.

Il existe plusieurs types de réseaux de neurones tels que les "perceptrons", les réseaux à fonctions de base radiales et les réseaux récurrents. Parmi eux, les perceptrons à alimentation directe (*feed-forward*) et entraînés par rétropropagation (*backpropagation*) ont eu un succès important dans plusieurs applications. Leur intérêt provient de la simplicité de leur utilisation ainsi que de la rapidité et de l'efficacité de leur algorithme de rétropropagation. Cet algorithme d'apprentissage a été proposé par Werbos (Werbos, 1974) et diffusé par Rumelhart *et al.* (Rumelhart *et al.*, 1986).

2.1.2 Caractéristiques d'un réseau de neurones

Le nombre de niveaux cachés et le nombre de neurones par niveau représentent les paramètres de l'architecture d'un réseau de neurones (Figure 1). La valeur de ces paramètres dépend principalement de la quantité et de la complexité des données. Cependant, une architecture qui donne de bons résultats pour une application donnée ne peut être déterminée que d'une façon expérimentale. En outre, une architecture optimale trouvée pour une application spécifique ne

garantit pas des résultats similaires dans d'autres applications. Toutefois, un nombre élevé de neurones dans les niveaux intermédiaires augmente le temps de calcul et diminue la généralisation du réseau, d'où la nécessité de trouver le meilleur compromis possible entre le nombre de niveaux et de neurones cachés.

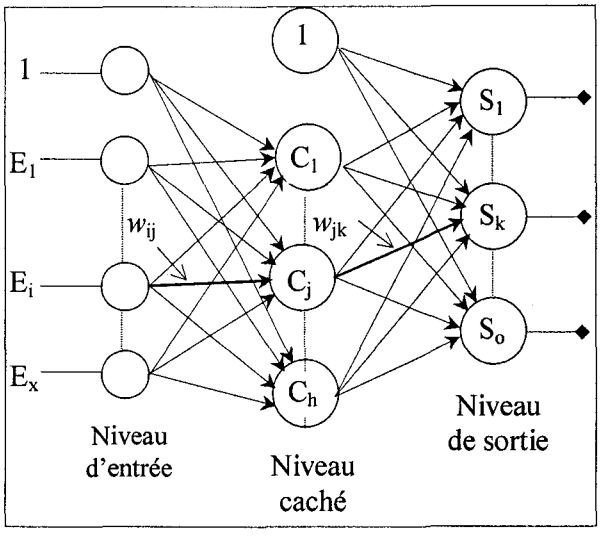

Figure 1 : Architecture d'un réseau multicouche

Les nœuds sont considérés comme éléments processeurs d'un réseau de neurones, chaque nœud permettant la transformation de l'information contenue dans les entrées (E_j) par une fonction non linéaire dite d'activation (Figure 2). La valeur de la sortie d'un neurone quelconque (j) est calculée à partir des entrées qu'il reçoit, ces entrées correspondant aux sorties de la couche précédente.

La réponse d'un neurone dépend des entrées qu'il reçoit, les entrées d'un neurone étant données par les sorties des neurones des couches précédentes pondérées par un facteur de poids (w) qui caractérise le lien entre deux neurones. La configuration et le fonctionnement de base pour chaque neurone intermédiaire sont présentés à la Figure 2.

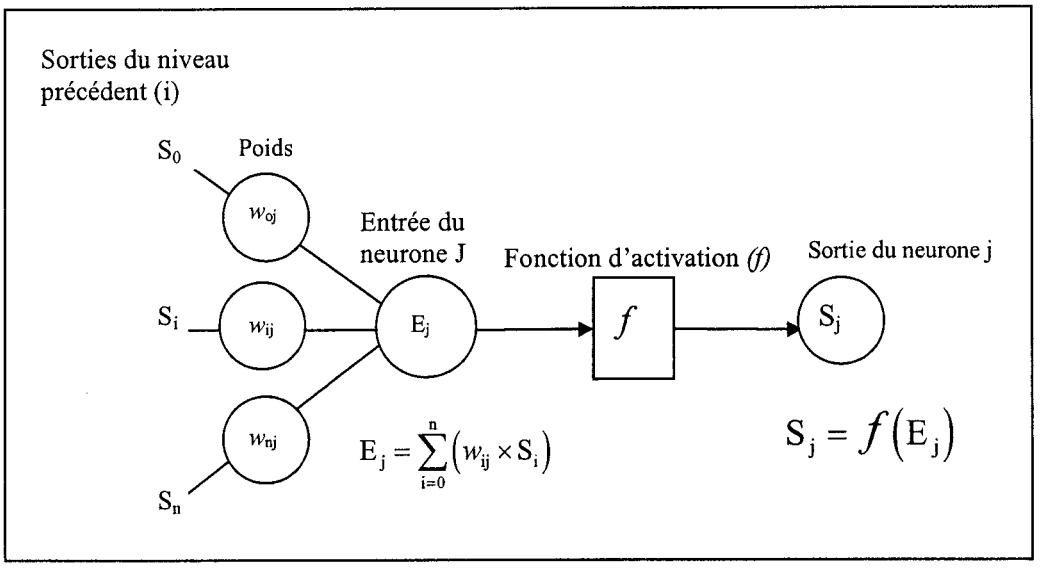

Figure 2 : Connections d'un élément processeur (nœud j)

La méthode utilisée pour le transfert de l'information entre deux neurones i et j appartenant à deux couches successives est basée sur les trois équations suivantes:

$$S_j = f(E_j) \tag{1}$$

$$E_j = \sum_{i=0}^{n} (w_{ij} \times S_i) \tag{2}$$

$$f(x) = \tanh(x) \approx \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3}$$

Où

S_i : La valeur à la sortie du neurone *i*
S_j : La valeur à la sortie du neurone *j*
f : La fonction d'activation (Exemple : tangente hyperbolique)
w_{ij} : Le coefficient de pondération (poids) entre les neurones *i* et *j*
E_j : La valeur à l'entrée du neurone *j*

Dans la plupart des applications, les fonctions d'activation utilisées sont soit la sigmoïde soit la tangente hyperbolique. Ces deux fonctions sont non linéaires et ont une forme asymptotique (Figure 3). Elles travaillent comme des amplificateurs non linéaires du signal. Les fonctions d'activation permettent de compresser la sortie d'un neurone dans un intervalle [0,1] pour la fonction sigmoïde et dans un intervalle [-1,1] pour la tangente hyperbolique afin d'éviter la saturation du signal.

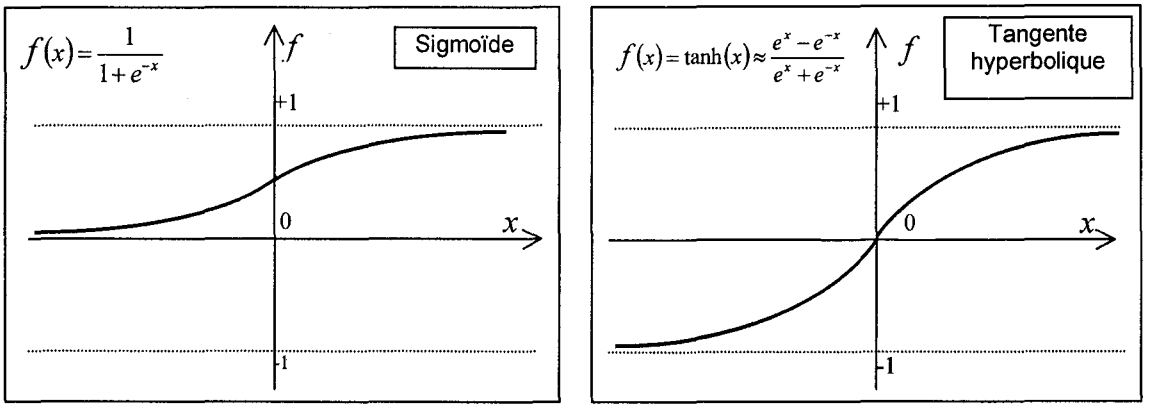

Figure 3 : Fonctions d'activation

2.1.3 Apprentissage du réseau

Les réseaux de neurones sont des outils de modélisation numérique qui tentent de prédire les sorties d'un système à partir de la connaissance des entrées. Cette prédiction est réalisée en construisant au cours d'une phase d'apprentissage (ou d'entraînement) un modèle non linéaire entre des couples entrées-sorties. Les poids (w) précisent le lien entre deux neurones appartenant à deux niveaux successifs (Figure 2). Leurs valeurs sont ajustées et affinées continuellement tout au long de la phase d'apprentissage. Pendant cette phase, un certain nombre de couples entrées-sorties sont fournis au réseau. Ces données représentent le groupe d'apprentissage et elles sont constituées de l'information disponible.

Dans un premier temps, les poids sont fixés aléatoirement pour permettre au réseau de calculer ses propres estimations à partir des entrées déjà fournies. Les poids sont alors corrigés de manière à minimiser la différence entre les sorties ainsi calculées et les sorties réelles. Cette phase de

minimisation correspond à l'apprentissage ; elle est primordiale à l'efficacité du réseau. L'ensemble des données utilisées pour cette étape doit donc être représentatif des situations qui seront rencontrées ultérieurement, lors de l'utilisation réelle. En effet, le réseau ne peut fournir de réponses correctes si les valeurs présentées lui paraissent inconnues.

Afin d'assurer un bon fonctionnement du réseau, les données présentées à l'entrée doivent être normalisées. Cette opération garantit une réponse significative de la fonction d'activation. C'est à dire que, pendant l'ajustement des poids, la sortie ajustée de chaque neurone doit refléter les ajustements initiaux. Ceci nous permet d'éviter que de petits changements dans l'entrée du réseau génèrent des grands changements à la sortie en entraînant la saturation du réseau.

Contrairement au nombre de neurones des niveaux cachés (qui doivent être déterminés d'une manière expérimentale), le nombre de neurones du niveau d'entrée et du niveau de sortie est directement lié aux informations disponibles et aux résultats attendus du réseau.

Pour le niveau d'entrée, on affecte généralement un neurone pour chaque information fournie au réseau. L'ordre de présentation des données d'entrée n'est pas important. Par contre, le format de valeur présentée au réseau a un effet primordial sur les phases d'entraînement et de classification. Les informations présentées à l'entrée seront filtrées par le réseau en donnant des poids différents pour chaque information. Comme cela, seules les données utiles seront prises en considération pour calculer la sortie.

L'apprentissage a été effectué par un algorithme de rétropropagation avec un taux d'apprentissage variable et une fonction d'activation sigmoïde.

Au début de la phase d'apprentissage, les groupes apprentissage et validation sont présentés au réseau avec les valeurs de sortie correspondantes. Les poids sont ajustés et affinés continuellement tout au long de la phase d'apprentissage. La correction des poids au cours de l'entraînement ne tient compte que des données appartenant au groupe d'apprentissage. Au cours de cette phase, les poids du réseau sont corrigés de manière à minimiser l'erreur au carré entre la réponse calculée par le réseau et la réponse attendue.

Généralement, l'erreur calculée sur le groupe d'apprentissage diminue continuellement au cours de l'entraînement. Toutefois, une longue phase d'entraînement diminue la capacité de généralisation du réseau en l'adaptant uniquement aux données de l'apprentissage (Figure 4). Ce phénomène est appelé le surentraînement ou «*overfitting*» en anglais. À cet effet, nous avons ajouté un autre groupe de données (groupe de validation) pour déterminer à quel moment l'apprentissage doit être arrêté. Les données appartenant à ce groupe servent uniquement à vérifier le comportement du réseau au cours de l'entraînement face à des données qui lui sont étrangères. Contrairement à l'erreur calculée sur le groupe d'apprentissage qui diminue continuellement au cours de l'entraînement, celle calculée sur le groupe de validation diminue dans la première phase d'entraînement en suivant une allure semblable à celle du groupe d'apprentissage avant de commencer à s'accroître (Figure 4). Ceci s'explique par le fait que le réseau commence à perdre son pouvoir de généralisation en adaptant ces neurones uniquement au groupe d'apprentissage.

L'entraînement du réseau est donc arrêté dès que cette erreur commence son ascension. Toutefois, afin d'éviter un arrêt prématuré de l'apprentissage causé par une augmentation ponctuelle de l'erreur du groupe de validation, nous avons introduit un seuil de décision qui tolère de légères ascensions successives de l'erreur. Si cette erreur continue son ascension au-delà de ce seuil, on arrête l'apprentissage du réseau et on conserve les valeurs des poids qui correspondent à l'itération qui précède cette ascension. Après plusieurs tests, nous avons trouvé qu'un seuil de 50 itérations est largement suffisant pour contourner ce genre de situation.

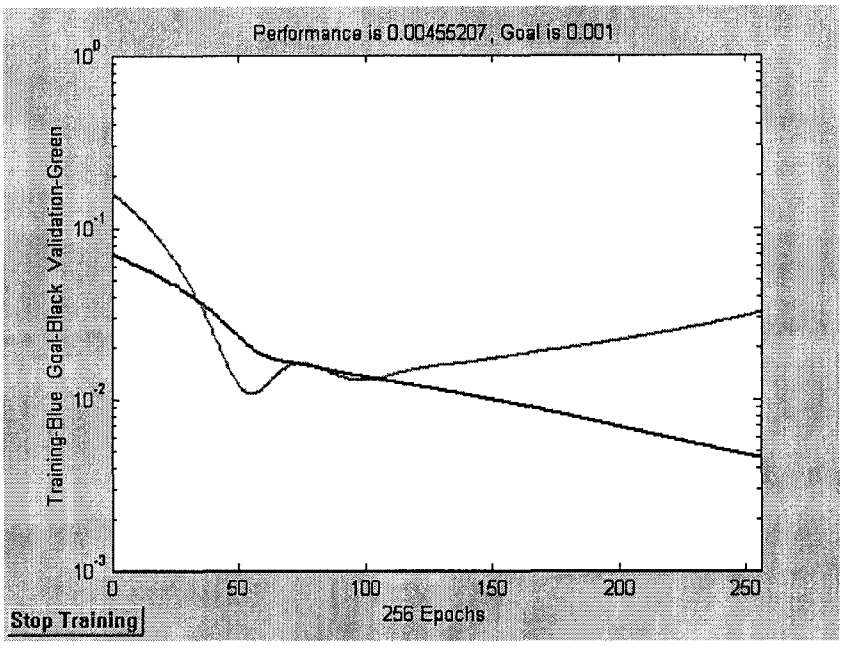

Figure 4 : Évolution de l'erreur au cours de la phase d'apprentissage

2.1.4 Validation du réseau

Après l'arrêt de l'apprentissage, il est toujours préférable de vérifier la performance du réseau avec un troisième groupe de données (groupe test). Ce groupe doit être constitué d'un ensemble de données qui n'ont pas servi à l'apprentissage et qui n'ont joué aucun rôle dans le choix du moment de l'arrêt de l'apprentissage. Le groupe test est utilisé uniquement pour mesurer la performance du réseau après l'arrêt de l'apprentissage. Si le réseau arrive à prédire correctement les débits contenus dans ce groupe de données avec une précision « acceptable », on peut dire que le réseau est opérationnel. Dans le cas contraire, il faut réviser les intrants du réseau et recommencer l'apprentissage.

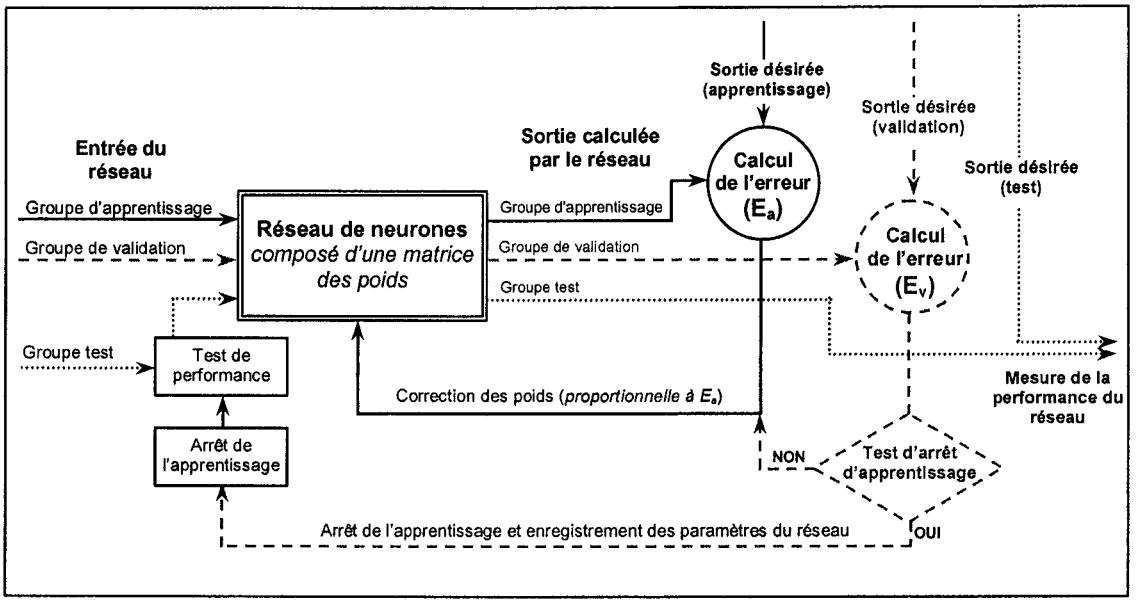

Figure 5 : Description schématique de la phase d'apprentissage

L'organigramme illustré à la Figure 5 résume les principales étapes suivies pendant l'apprentissage et la validation du réseau de neurones et montre le cheminement de chaque groupe de données.

2.2 MODÈLE RÉGRESSIF

2.2.1 Introduction

La régression linéaire est largement utilisée dans différents domaines et elle a été abondamment documentée. Pour cette raison nous nous contenterons ici d'en présenter qu'une brève description. Pour plus de détails sur le sujet, il est possible de consulter Draper et Smith (1966), Weisberg (1985) ou Neter *et al.* (1985).

2.2.2 Régression multiple

La régression linéaire multiple utilise p variables explicatives, supposées indépendantes entre elles, $X_1, X_2, ..., X_p$, pour modéliser une variable réponse Y (ou variable dépendante). Alors, le modèle général de régression multiple est de la forme :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon \tag{4}$$

Où ε est une variable aléatoire distribuée selon une loi normale.

Les réalisations des variables explicatives $X_1, X_2, ..., X_p$, sont notées $x_1, x_2, ..., x_p$. On note aussi la réalisation de la variable dépendante Y par y. Ainsi, pour une réalisation donnée de l'ensemble des variables, le modèle s'écrit :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \tag{5}$$

Où, l'indice $i = 1, 2, ..., n$ réfère à la réalisation de l'ensemble des $p+1$ variables; n désigne la taille d'échantillon; $\beta_0, \beta_1, ..., \beta_p$ sont les paramètres de la régression multiple; et ε_i, $i = 1, 2, ..., n$, sont des variables aléatoires indépendantes et identiquement distribuées selon une loi normale centrée.

Le modèle (éq.1) est dit *multiple* puisqu'il fait intervenir plus d'une variable explicative, et *linéaire* parce que celles-ci apparaissent dans le modèle à la puissance 1.

Le modèle de régression multiple à p variables explicatives (éq. 2) peut-être exprimé sous forme matricielle. Pour se faire, définissons les matrices suivantes :

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \underset{n\times(p+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots x_{1p} \\ 1 & x_{21} & x_{22} & \cdots x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots x_{np} \end{bmatrix}; \quad \underset{(p+1)\times 1}{\mathrm{B}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \text{ et } \underset{n\times 1}{\mathrm{E}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Le vecteur colonne **Y** contient les n valeurs observées de la variable dépendante, la matrice **X** les valeurs correspondantes des p variables explicatives et une colonne de 1, le vecteur **B** contient les

$p+1$ paramètres de la régression, et enfin le vecteur **E** les n termes d'erreur aléatoire. Donc, sous forme matricielle, le modèle de régression linéaire multiple (éq. 2) s'exprime de la façon suivante :

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times(p+1)}{\mathbf{X}} \quad \underset{(p+1)\times 1}{\mathrm{B}} + \underset{n\times 1}{\mathrm{E}} \tag{6}$$

En pratique, la matrice **X** des observations des variables explicatives et la matrice **Y** des observations de la variable dépendante sont connues. Pour déterminer la fonction de régression, il suffit alors d'estimer les paramètres $\beta_0, \beta_1, \ldots, \beta_p$. Afin d'obtenir de bons estimateurs, on emploie la méthode des moindres carrés qui consiste à minimiser la somme des carrés des résidus ε_i définis comme suit :

$$\varepsilon_i = y_i - \left(\hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k x_{ik} \right) \tag{7}$$

$\hat{\beta}_0$ étant l'estimateur de β_0 et $\hat{\beta}_k$ l'estimateur de β_k. La fonction à minimiser, sous forme matricielle, s'exprime alors de la façon suivante :

$$Q = (\mathrm{Y} - \mathrm{X}b)' (\mathrm{Y} - \mathrm{X}b) = \mathrm{Y}'\mathrm{Y} - b\mathrm{X}'\mathrm{Y} \tag{8}$$

Où $\mathrm{b} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$, $\mathbf{X}'$ désignant la transposée d'une matrice **X**. Ceci revient à résoudre un système de p équations à p inconnus pour obtenir le vecteur des paramètres estimés **b**. Ce système d'équations s'écrit :

$$(\mathrm{X}'\mathrm{X})b = \mathrm{X}'\mathrm{Y} \tag{9}$$

et on déduit :

$$b = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y} \tag{10}$$

Où $\mathbf{X}'$ est la matrice transposée de **X** et $(\mathrm{X}'\mathrm{X})^{-1}$ est la matrice inverse de $(\mathrm{X}'\mathrm{X})$.

2.2.3 Régression "Stepwise"

La régression multiple vise à obtenir, à partir d'un ensemble de variables explicatives, un modèle de régression faisant intervenir les variables explicatives les plus significatives pour expliquer les fluctuations aléatoires de la variable dépendante. Ceci permet d'obtenir une équation de régression possédant un coefficient de détermination R^2 élevé. Le moyen le plus sûr pour y arrivé est d'effectuer la régression en entrant progressivement les variables explicatives une à une dans le modèle. À chaque fois, on trace les résidus du modèle en fonction des variables explicatives inutilisées. Si aucun de ces graphiques ne révèle une relation entre les résidus et les variables inutilisées, ceci signifie qu'aucune de ces dernières n'est utile. En revanche, si l'une d'elles affiche une relation avec les résidus on peut conclure qu'elle est probablement pertinente. Il est aussi important de s'assurer que la valeur du paramètre de la première variable reste stable suite à l'inclusion de la nouvelle variable explicative. Si par exemple la première variable devient non significative, il y a probablement un problème de multicolinéarité entre les variables explicatives.

Cependant cette technique est lourde et nécessite une bonne expertise. Il existe néanmoins d'autres méthodes automatiques permettant de sélectionner un modèle optimal. L'une d'entre elles est la procédure stepwise (Neter *et al.*, 1985; Weisberg, 1985). Pour en faire un bref résumé, cette procédure consiste à ajouter ou retrancher une variable explicative du modèle selon un critère de sélection. Le critère d'entrée ou de sortie d'une variable explicative est un rapport de sommes des carrés, et que l'on compare à une valeur théorique critique établie a priori. Ce critère permet d'évaluer l'effet de l'ajout d'une nouvelle variable explicative sur la contribution de la ou des variables explicatives déjà contenues dans le modèle. Si cet effet est significatif, la nouvelle variable est gardée dans le modèle. Si cet apport n'est pas significatif, la variable correspondante est alors retranchée de l'équation de régression. La sélection se termine lorsqu'aucune variable explicative ne peut être ajoutée ou retranchée de l'équation de régression.

3 APPLICATION

3.1 MÉTHODOLOGIE

3.1.1 Données

En commun accord avec les Relevés hydrologiques du Canada (Environnement Canada), il a été décidé de tester les performances des différentes méthodes étudiées en les appliquant au niveau de quatre stations hydrométriques de l'Alberta (Figure 6). Il s'agit de la rivière Athabasca à Fort McMurray (07DA001), de la rivière Clearwater à Draper (07CD001), de la rivière Pembina à Entwistle (07BB002) et de la rivière North Saskatchewan à Edmonton (05DF001). Afin d'obtenir les conditions climatiques correspondant aux relevés hydrométriques, ces stations ont été jumelées avec les stations météorologiques les plus proches (Figure 6). Les caractéristiques des stations hydrométriques et météorologiques sont présentées au Tableau 1 et Tableau 2, respectivement. Ces données été mises à notre disposition par les Relevés hydrologiques du Canada.

Tableau 1 : Caractéristiques des stations hydrométriques étudiées

Nom	Identifiant	Long.	Lat.	Superficie du bassin (km^2)	En service depuis...
Athabasca à Fort McMurray	07DA001	-111°24'00"	56°46'50"	131000	1957
Clearwater à Draper	07CD001	-111°15'15"	56°41'07"	30800	1930
Pembina à Entwistle	07BB002	-115°00'14"	53°36'18"	4350	1914
North Saskatchewan à Edmonton	05DF001	-113°29'04"	53°32'15"	27300	1911

Tableau 2 : Caractéristiques des stations météorologiques utilisées

Nom	Identifiant	Long. (°)	Lat.	Altitude (m)
Fort McMurray A	3062693	-111,22	56,65	369
Edmonton International Airport A	3012205	-111,58	53,32	723
Magnolia	3064157	-114,88	53,58	755
Wildwood Newbery	3067NF0	-113,32	53,57	853
Entwistle	3062451	-114,98	53,60	780

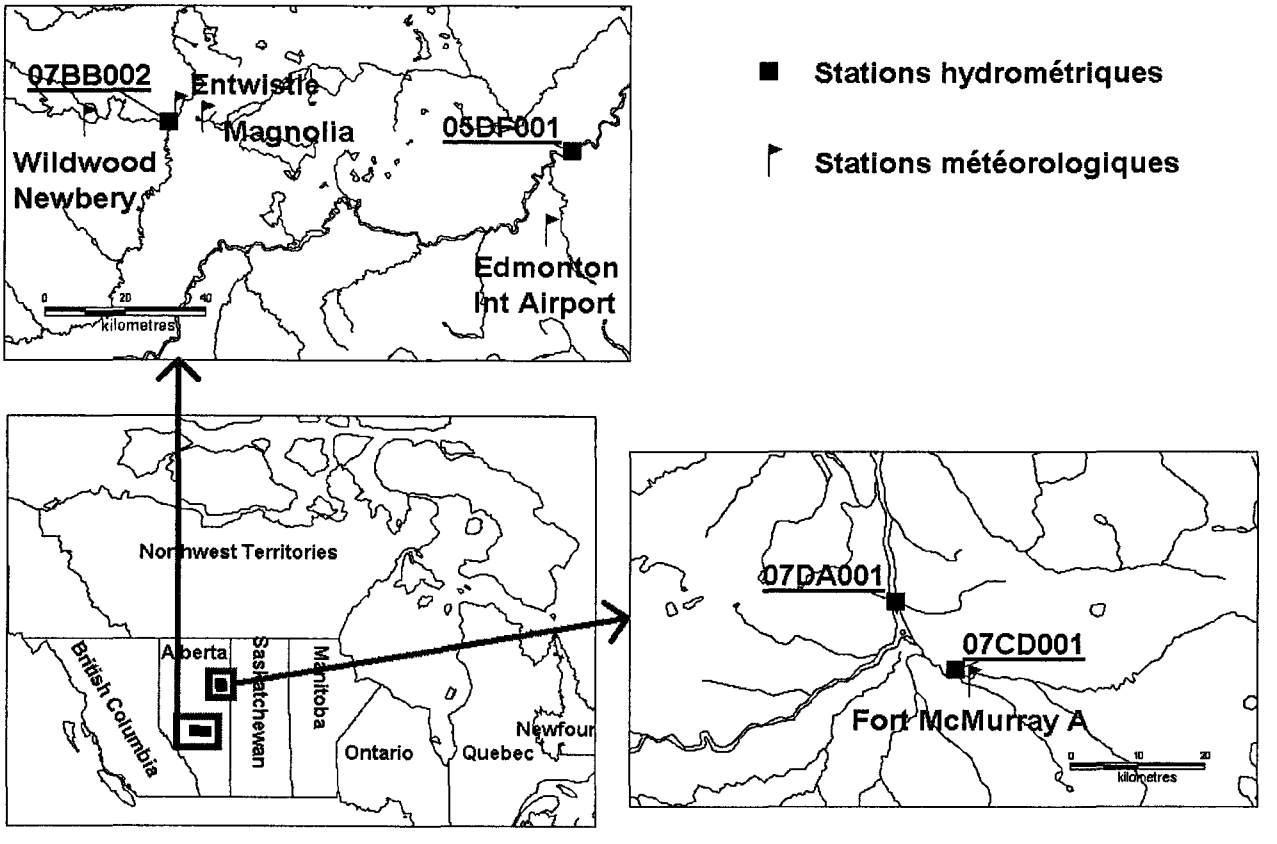

Figure 6 Localisation des stations hydrométriques et des stations météorologiques

Les fichiers de données météorologiques contiennent les éléments météorologiques journaliers suivants :

- Température maximale (Tmax);
- Température minimale (Tmin);
- Température moyenne (Tmoy);
- Précipitations liquides (PLT);
- Précipitations solides (PST);
- Précipitations totales (PT);
- Hauteur de la neige au sol (NS).

En plus de ces variables météorologiques brutes disponibles au départ, nous avons calculé des variables dérivées telle que :

- La température maximale durant la décade précédant la date du jaugeage (Tmax10)
- La température minimale durant la décade précédant la date du jaugeage (Tmin10)
- La température moyenne durant la décade précédant la date du jaugeage (Tmoy10)
- La variation journalière de la température moyenne (DTmoy)
- La moyenne de la variation journalière de la température durant la décade précédant la date du jaugeage (DTmoy 10)
- Les degrés-jours cumulés en bas de zéro (DJNeg)
- La précipitation liquide cumulée durant la décade précédant la date du jaugeage (PLT10)
- La précipitation liquide cumulée durant la décade précédant la date du jaugeage (PST10)
- Les degrés-jours cumulés entre le 1er novembre et le 30 avril (DJH)
- Les précipitations solides cumulées entre le 1er novembre et le 30 avril (PSCH).

Il est à noter que les fichiers météorologiques présentaient de nombreuses plages de données manquantes. La neige au sol représente une variable problématique pour la station météorologique de Fort McMurray A, puisqu'elle est manquante pour les années 2000 à 2005.

Les données hydrométriques, quant à elles, se répartissent en deux catégories : les données de jaugeage et les mesures automatiques du niveau d'eau moyen journalier dans la section de la rivière. Les données de jaugeage comprennent en plus de la date du jaugeage l'épaisseur moyenne de la glace dans la section de la rivière, le niveau d'eau moyen, la vitesse moyenne de l'eau et le débit.

Pour bâtir la base de données à partir de laquelle nous avons construit les fichiers de calibration et de validation, nous avons en premier temps établi la correspondance entre les jaugeages hivernaux, les données hydrométriques et les données météorologiques correspondantes (Figure 6). Le Tableau 3 présente la correspondance entre stations hydrométriques et stations météorologiques ainsi que les périodes d'observations de chaque type de données.

Tableau 3 : Correspondance entre stations hydrométriques et stations météorologiques

Stations hydrométriques					Stations météorologiques	
Nom	**Identifiant**	**Jaugeages hivernaux†**	**Niveau d'eau automatique**	**Distance à la station météo. (km)**	**Nom**	**Période d'observation**
Athabasca à Fort McMurray	07DA001	1979-2005 **89** [2; 6; 3,5]	1979-2005	18	**Fort McMurray A**	1978-2005
Clearwater à Draper	07CD001	1979-2005 **61** [1; 7; 3,4]	1979-2005	4,5		
Pembina à Entwistle	07BB002	1979-2005 **160** [3; 10; 6,2]	1979-2004	8,6	**Magnolia**	1978-1979
				21	**Wildwood Newbery**	1980-1986
				1,7	**Entwistle**	1987-2005
North Saskatchewan à Edmonton	05DF001	1982-2005 **69** [1; 8; 3,6]	1979-2005	25	**Edmonton International Airport A**	1981-2005

† La période d'observation, le nombre total de jaugeages hivernaux et, entre crochets, respectivement, le nombre minimal de jaugeages par hiver, le nombre maxima de jaugeages par hiver et le nombre moyen de jaugeages par hiver.

Généralement, nous disposons d'environ 25 ans de données. Ceci correspond à environ 3 jaugeages hivernaux en moyenne. Sauf pour la station 07BB002 pour laquelle nous disposons de plus 6 jaugeages/hiver, pour un total de 160 jaugeages. Pour les trois autres stations, le nombre total de jaugeages varie entre 61 et 89. Par ailleurs, les données hydrométriques de la station 07BB002 sont associées à des observations météorologiques recueillies pour la plupart à proximité immédiate de la station (station Entwistle entre 1987-2005). En revanche, la station 05DF001 se trouve dans la situation contraire (la station météorologique correspondante est à 25 km de distance). Vu la présence de données manquantes dans les données météorologiques, la disparité dans l'occurrence et l'étendue la période d'observation ainsi que le nombre relativement faible de jaugeages hivernaux, la taille du jeu de données pouvant être utilisés s'est trouvé réduit par rapport aux jaugeages initialement disponibles (dont le nombre est déjà faible pour les stations 07CD001 et 05DF001, par exemple.

3.1.2 Modélisation

En premier lieu, nous avons procédé à l'analyse de la corrélation entre les deux variables réponse (variables expliquées) à savoir le débit jaugé et l'épaisseur de la glace, et l'ensemble des variables explicatives disponibles correspondantes (le niveau de l'eau et les variables météorologiques brutes et dérivées).

Pour calibrer les différentes méthodes d'estimation étudiées, nous nous sommes servi des données recueillies avant l'année 2000. Les données après l'année 2000 ont servi à la validation.

Pour chaque station hydrométrique et pour chacune des deux variables réponse, le modèle de RNA et le modèle de la régression multiple ont été calibrés à l'aide des variables explicatives présentant une corrélation statistiquement significative. Le modèle RNA a été également calibré à l'aide d'une sélection de variables explicatives identifiées par la régression stepwise parmi les variables retenues initialement.

En raison la taille limitée du jeu de données disponible, la qualité de la calibration du modèle RNA i.e. les chances du modèle de converger vers un entraînement optimal est tributaire, d'une part de la répartition de l'échantillon en différents groupes de données (groupe d'apprentissage,

groupe de validation et groupe test), d'autre part, de la valeur initiale des poids du réseau de neurones. Par conséquent, nous avons opté pour un processus de calibration du RNA en deux étapes itératives. La première étape consiste à trouver la répartition optimale des groupes de données. Les poids initiaux du réseau étant fixés à une série de valeurs choisie d'une manière aléatoire. Pour cela, on départage d'une manière aléatoire le jeu de calibration en trois groupes : la moitié de l'échantillon étant consacrée au groupe d'apprentissage et les deux autres groupes sont constitués d'un quart de l'échantillon chacun, tout en s'assurant que les deux valeurs extrêmes de l'échantillon (la valeur minimale et maximale de la variable réponse) sont incluses dans le groupe d'apprentissage. Cette opération est répétée un grand nombre de fois. À chaque fois, le modèle obtenu est appliqué à trois groupes de données pour en estimer les valeurs de la variable réponse. Ces valeurs sont par la suite comparées aux valeurs observées et le coefficient de détermination (R^2) pour chaque groupe de données est calculé. Enfin, la répartition du jeu de données correspondant au modèle qui produit la valeur du R^2 du groupe test la plus élevée est retenues. Par ailleurs, ce modèle doit avoir produit une valeur du R^2 des groupes apprentissage et validation combinés supérieure ou égale 0,60. Cette condition aurait pour but d'obtenir une répartition des données correspondant à un modèle neuronal ayant une capacité de généralisation la plus optimale. La deuxième étape consiste à optimiser le choix initial des poids du réseau en appliquant la même procédure itérative, avec la même condition pour le choix du modèle optimum et avec la répartition du jeu en différents groupes identifiée à l'étape précédente. Ce modèle optimum est retenu pour être utilisé pour l'estimation de la variable réponse en question.

Afin de juger de la qualité de la calibration des modèles, nous nous sommes basés, en plus du coefficient de détermination. sur les paramètres suivant : l'erreur quadratique moyenne (*RMSE*), l'erreur quadratique relative moyenne (*RMSEr*), le biais moyen (*B*) et le biais relatif moyen (*Br*). Ces critères calculés à partir des résultats de la calibration nous ne permettent uniquement que d'apprécier les capacités d'interpolation d'un modèle. En effet, ils ne permettent pas de juger de ses capacités d'extrapolation, c'est-à-dire de ses performances lorsqu'il est utilisé avec des données qui n'ont pas servi à sa calibration. Par conséquent, nous avons conduit une validation classique sur les données recueillies entre les années 2000 et 2005. Ces critères peuvent être calculés comme suit :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

$$RMSEr = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{y_i})^2} \tag{12}$$

$$B = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{13}$$

$$Br = \frac{1}{n}\sum_{i=1}^{n}(\frac{\hat{y}_i - y_i}{y_i}) \tag{14}$$

Où y et $\hat{y}$ étant, respectivement, les valeurs originales et estimées de la variable réponse et n le nombre d'observations.

3.2 RÉSULTATS

Tout d'abord, nous avons procédé à l'examen des données de jauges afin d'en vérifier la qualité. Leurs statistiques descriptives sont présentées au Tableau 4. En moyenne, les débits et les épaisseurs de glace les plus élevés ont été enregistrés à la station 07DA001. Les stations 05DF001, 07CD001et 07BB002 arrivent au 2^{e}, 3^{e} et 4^{e} rang, respectivement. En ce qui concerne l'épaisseur de la glace, la station 05DF001 affichent une grande variabilité avec une valeur maximale dépassant les 2 m. Ces valeurs correspondent à l'hiver exceptionnel de 1994.

Tableau 4 : Statistiques descriptives du débit et de l'épaisseur de glace jaugés au niveau des quatre stations hydrométriques étudiées

05DF001					
	Moyenne	**Max**	**Min**	**Écart-type**	**CV†**
Débit (m^3/s)	111.41	164	52	26.42	23.71%
Épaisseur de la glace (m)	0.72	2.38	0.29	0.31	43.06%
07BB002					
	Moyenne	**Max**	**Min**	**Écart-type**	**CV**
Débit (m^3/s)	5.17	23.4	1.11	4.12	79.69%
Épaisseur de la glace (m)	0.48	0.88	0.13	0.17	35.42%
07CD001					
	Moyenne	**Max**	**Min**	**Écart-type**	**CV**
Débit (m^3/s)	57.19	105	29.2	17.02	29.76%
Épaisseur de la glace (m)	0.57	1.84	0.25	0.2	35.09%
07DA001					
	Moyenne	**Max**	**Min**	**Écart-type**	**CV**
Débit (m^3/s)	198.81	498	104	81.05	40.77%
Épaisseur de la glace (m)	0.71	1.09	0.31	0.16	22.54%

†: Coefficient de variation

Dans le but de sélectionner la les variables explicatives les plus pertinentes à inclure dans les modèles d'estimation, nous avons produit les matrices de corrélation entre ces variables et les deux variables réponse (Tableau 5). Comme on peut s'attendre, le niveau de l'eau est la variable la plus importante pour expliquer la variance dans le débit jaugé. Sauf pour la station 05DF001, où le niveau présente une faible corrélation avec les données du débit par comparaison aux variables météorologique. De plus, l'épaisseur de glace mesurée à cette station n'est significativement corrélée à aucune autre variable explicative autre que le niveau de l'eau et ce, à un degré plus élevé que l'est le débit jaugé ou l'épaisseur de glace au niveau de trois autres stations. Ceci nous amène à se questionner sur la qualité des données de cette station de même que la pertinence de l'associer à la station météorologique de l'aéroport d'Edmonton (une station en milieu urbain et à 25 km du site de jaugeage). Outre le niveau, ce sont les variables de température et la hauteur de la neige au sol qui présentent les plus importants niveaux de corrélation avec les données du débit jaugé. Toutefois, étant donné que la neige au sol est manquante du jeu de données de validation pour les stations 07CD001 et 07DA001, cette variable n'a pas été retenue pour l'étalonnage des différentes méthodes d'estimation du débit pour ces deux stations. Également, la neige au sol présente plusieurs plages de données manquantes dans le jeu de calibration pour le débit de la station 07BB002. L'utilisation de cette

variable dans l'étalonnage des différentes méthodes d'estimation réduirait la taille du jeu de données de calibration de 50%. Par conséquent, nous avons décidé également d'écarter cette variable du processus d'estimation du débit pour cette station.

En ce qui concerne l'épaisseur de la glace mesurée aux stations 07BB002, 07CD001 et 07DA001, cette variable est en générale corrélée aux variables de température en particulier les degrés-jours cumulés pour l'hiver (DJH) calculés à partir du 1er novembre et aux précipitations solides cumulées à partir du 1er novembre de chaque année. Ceci est prévisible puisque la formation de la glace est un processus cumulatif et elle serait a priori corrélée à de telles variables cumulatives qui sont reliées au bilan d'énergie et de matière du couvert de glace.

Tableau 5 : Corrélation entre les deux variables réponse (débit et épaisseur) et les variables explicatives disponibles (les valeurs de corrélation en trame grise sont statistiquement significatives)

	05DF001		07BB002		07CD001		07DA001	
	Débit	Épaisseur de glace	Débit	Épaisseur de glace	Débit	Épaisseur de glace	Débit	Épaisseur de glace
Niveau	0.31	0.52	0.70	0.33	0.75	0.34	0.87	0.36
Tmax	0.33	0.01	0.31	0.35	0.11	0.29	0.29	0.36
Tmin	0.37	-0.06	0.31	0.28	0.07	0.21	0.25	0.27
Tmoy	0.35	-0.02	0.33	0.33	0.09	0.26	0.27	0.32
Tmax10	0.24	-0.08	0.50	0.39	0.15	0.20	0.35	0.31
Tmin10	0.54	0.02	0.43	0.43	0.23	0.29	0.42	0.32
Tmoy10	0.53	-0.02	0.43	0.51	0.20	0.27	0.36	0.34
DTmoy	0.02	0.08	0.04	-0.09	-0.02	0.09	-0.03	0.05
DTmoy10	0.19	0.02	-0.03	0.18	-0.09	0.06	-0.12	0.12
DJNeg	0.10	0.01	0.20	0.32	0.21	0.04	0.30	0.14
PLT	-0.09	-0.01	0.13	0.31	0.03	0.05	-0.14	0.18
PLT10	-0.13	-0.14	0.50	0.14	0.09	0.03	0.18	0.26
PST	0.05	-0.08	-0.13	-0.04	-0.21	-0.05	-0.21	-0.22
PST10	-0.27	0.13	-0.20	-0.19	-0.14	-0.09	-0.15	-0.19
PT	0.05	-0.07	-0.11	0.02	-0.23	-0.05	-0.21	-0.25
NS	-0.16	0.11	-0.43	0.03	-0.27	0.07	-0.40	-0.12
DJH	-0.13	0.02	-0.04	-0.69	0.24	-0.28	-0.14	-0.53
PSHC	0.05	0.02	0.12	0.52	-0.20	0.15	0.24	0.40

Dans un premier temps, nous avons procédé à la calibration du modèle RNA et du modèle de régression pour l'estimation des deux variables réponse à l'aide de leurs variables explicatives

correspondantes identifiées au Tableau 5. Les résultats pour le débit sous glace sont présentés au Tableau 6. Ceux de l'épaisseur de la glace figurent au Tableau 7.

Pour chaque station, la sélection de variables explicatives utilisées est listée en haut du tableau. Les résultats pour le groupe dit de calibration sont présentés au milieu du tableau. Ce groupe comprend le groupe d'apprentissage et le groupe de l'arrêt de l'apprentissage (dit groupe de validation). Il est à noter que dans certains cas les valeurs de R^2 du groupe de calibration sont inférieures à 0,60. Ceci est en raison des faibles R^2 du groupe de l'arrêt de l'apprentissage pour lequel aucun critère d'optimisation n'est appliqué. Il est à noter également que les résultats du groupe test sont présentés comme des résultats de calibration (en bas du tableau). En effet, étant donné que les résultats de ce groupe sont utilisés pour la sélection du schéma de répartition des données pour le réseaux de neurones et dans la sélection du meilleur modèle neuronal, ce groupe ne peut être considéré comme complètement indépendant du processus de calibration et par conséquent ne peut être utilisé comme jeu de validation indépendant. Par souci de transparence, nous présentons séparément les résultats de calibration du groupe test.

Tableau 6 : Résultats de la calibration des deux méthodes d'estimation du débit sous glace à l'aide des variables explicatives identifiées par l'analyse de la corrélation

		05DF001		07BB002		07CD001		07DA001	
Variables explicatives		Niveau Tmax Tmin Tmoy Tmin10 Tmoy10		Niveau Tmax Tmin Tmoy Tmax10 Tmin10 Tmoy10 PLT10		Niveau Tmin10 DJH		Niveau Tmax Tmin Tmoy Tmax10 Tmin10 Tmoy10 DJNeg PSHC	
Groupes calibration (apprentissage + validation)	**Taille de l'échantillon**	46		82		58		56	
		RNA	**RM†**	**RNA**	**RM**	**RNA**	**RM**	**RNA**	**RM**
	R^2	0.50	0.45	0.77	0.58	0.60	0.62	0.76	0.80
	RMSE (m^3/s)	17.53	18.15	1.92	2.61	10.62	10.30	38.46	34.42
	RMSEr (%)	18.26	21.56	43.57	55.96	18.18	16.28	15.53	20.43
	B (m^3/s)	-0.95	0.00	-0.14	0.00	0.34	0.00	-1.67	0.00
	Br (%)	1.69	3.34	10.51	13.38	3.71	2.78	1.16	2.19
Groupe test	**Taille de l'échantillon**	13		25		16		17	
		RNA	**RM**	**RNA**	**RM**	**RNA**	**RM**	**RNA**	**RM**
	R^2	0.59	0.06	0.93	0.70	0.94	0.80	0.98	0.89
	RMSE (m^3/s)	19.14	29.21	1.32	2.78	4.83	9.72	17.80	31.40
	RMSEr (%)	18.75	28.98	28.28	55.87	8.64	12.78	10.03	16.65
	B (m^3/s)	2.48	2.45	-0.09	-0.72	2.15	3.23	-11.99	-2.31
	Br (%)	4.39	8.79	2.73	-8.64	4.49	5.25	-6.94	-1.20

† : Régression multiple

De prime abord, il en ressort que le modèle neuronal est le modèle qui réussi le mieux à s'ajuster aux données de calibration pour le débit sous glace, surtout en ce qui concerne l'erreur quadratique et le biais. Ceci est encore plus évident avec le groupe test. Dans la plupart des cas, il réussit à expliquer près de 80% de la variance observée dans le débit jaugé. Même pour la station 05DF001, le modèle neuronal s'ajuste assez bien aux données. Toutefois, les bons résultats de la station 05DF001 ont été obtenus à l'aide de deux cycles de 2000 itérations (répartition des groupes et initialisation des poids du réseau de neurones) contre 1000 pour les trois autres stations et la réduction du critère sur le R^2 de 0,60 à 0,50 (sinon le modèle neuronal ne convergeait pas). Ce qui confirme nos soupçons quant à la qualité des données de cette station.

Tableau 7 : Résultats de la calibration des deux méthodes d'estimation de l'épaisseur de la glace à l'aide des variables explicatives identifiées par l'analyse de la corrélation

		05DF001		07BB002		07CD001		07DA001	
Variables explicatives		Niveau		Niveau Tmax Tmin Tmoy Tmax10 Tmin10 Tmoy10 DJNeg DJH PSHC		Niveau Tmax Tmoy Tmin10 Tmoy10 DJH		Niveau Tmax Tmin Tmoy Tmax10 Tmin10 Tmoy10 PLT10 PT DJH PSHC	
Groupes calibration (apprentissage + validation)	**Taille de l'échantillon**	46		82		58		56	
		RNA	**RM†**	**RNA**	**RM**	**RNA**	**RM**	**RNA**	**RM**
	R^2	0.14	0.24	0.29	0.29	0.54	0.26	0.57	0.39
	RMSE (m)	0.31	0.28	0.15	0.15	0.15	0.19	0.10	0.12
	RMSEr (%)	36.35	31.54	65.95	54.52	22.04	27.27	15.06	17.29
	B (m)	-0.04	0.00	-0.01	0.00	-0.03	0.00	-0.01	0.00
	Br (%)	40.75	80.88	12.40	15.21	-0.05	60.89	0.50	20.96
Groupe test	**Taille de l'échantillon**	13		25		16		17	
		RNA	**RM**	**RNA**	**RM**	**RNA**	**RM**	**RNA**	**RM**
	R^2	0.63	0.59	0.53	0.36	0.81	0.50	0.94	0.53
	RMSE (m)	0.14	0.14	0.10	0.12	0.06	0.11	0.07	0.14
	RMSEr (%)	14.22	15.25	26.21	32.49	11.37	20.87	11.71	25.96
	B (m)	-0.05	-0.03	-0.02	-0.02	0.02	0.04	0.01	0.05
	Br (%)	-4.75	-2.40	20.84	40.48	40.08	70.78	40.45	12.78

† : Régression multiple

Les résultats de la calibration du modèle neuronal pour l'estimation de l'épaisseur de la glace Tableau 7) viennent confirmer le même constat quant à la suprématie des RNA par rapport à la régression. Toutefois, les niveaux de la variance de l'épaisseur de la glace expliquée par les deux types de modèles sont plus faibles que dans le cas du débit sous glace. Ceci signifie que les variables explicatives utilisées ne sont pas suffisantes et l'ajout d'autres variables explicatives en relation avec la formation de la glace telle que la radiation solaire ou la température de l'eau serait avantageux pour une meilleure estimation de cette variable réponse.

La calibration du modèle neuronal et de la régression semble bénéficier de la réduction des variables explicatives introduites dans le modèle d'estimation des deux variables réponse

(Tableau 8, Tableau 9). Dans le cas de la régression, ceci réduit le risque de multicolinéarité. Dans le cas du modèle neuronal, la réduction du nombre de variables explicatives contribue à alléger la structure du réseau de neurones et par conséquent à réduire le nombre de poids à estimer et favoriser ainsi les chances du réseau à converger vers la solution optimale. L'amélioration notable dans les résultats se situe au niveau de la réduction du biais moyen relatif et ce, pour l'estimation des deux variables réponse. Dans le cas du débit sous glace, en plus du niveau d'eau, la deuxième variable en importance identifiée par la régression stepwise est la température extrême durant la décade précédant le jaugeage (Tmin10 ou Tmax10 selon les stations). Quant à l'épaisseur de la glace, la première variable en importance est la DJH et le niveau de l'eau arrive en second lieu.

Il est à noter que dans le cas du débit pour la station 07CD001 et de l'épaisseur de la glace pour la station 05DF001, les variables explicatives identifiées par la régression stepwise sont identiques à celles introduites dans le modèle complément. Par conséquent, les résultats de la calibration des deux versions des modèles demeurent identiques.

Tableau 8 : Résultats de la calibration des méthodes d'estimation du débit sous glace à l'aide des variables explicatives identifiées par la régression stepwise

		05DF001		07BB002		07CD001		07DA001	
Variables explicatives		Niveau Tmin10		Niveau Tmax10 PLT10		Niveau Tmin10 DJH		Niveau Tmax10	
Groupes calibration (apprentissage + validation)	**Taille de l'échantillon**	46		82		58		56	
		RNA	**RM†**	**RNA**	**RM**	**RNA**	**RM**	**RNA**	**RM**
	R^2	0.59	0.31	0.62	0.60	0.60	0.62	0.77	0.79
	RMSE (m^3/s)	16.93	21.55	2.67	2.73	10.62	10.30	39.23	35.40
	RMSEr (%)	17.53	25.14	53.16	58.73	18.18	16.28	22.81	21.31
	B (m^3/s)	-2.82	0.00	0.03	0.00	0.34	0.00	0.22	0.00
	Br (%)	-0.54	4.58	20.06	14.73	3.71	2.78	5.31	2.45
Groupe test	**Taille de l'échantillon**	13		25		16		17	
		RNA	**RM**	**RNA**	**RM**	**RNA**	**RM**	**RNA**	**RM**
	R^2	0.61	0.52	0.95	0.52	0.94	0.80	0.99	0.92
	RMSE (m^3/s)	15.85	17.44	2.05	2.80	4.83	9.72	20.33	41.40
	RMSEr (%)	13.65	15.58	31.34	74.71	8.64	12.78	9.45	27.36
	B (m^3/s)	1.58	2.23	-0.72	-0.50	2.15	3.23	-2.12	-18.41
	Br (%)	2.58	3.88	1.03	4.79	4.49	5.25	2.10	-1.20

† : Régression multiple

Tableau 9 : Résultats de la calibration des méthodes d'estimation de l'épaisseur de la glace à l'aide des variables explicatives identifiées par la régression stepwise

		05DF001		07BB002		07CD001		07DA001	
Variables explicatives		Niveau		Niveau DJH		Niveau DJH		Niveau DJH	
Groupes calibration (apprentissage + validation)	**Taille de l'échantillon**	46		82		58		56	
		RNA	**RM†**	**RNA**	**RM**	**RNA**	**RM**	**RNA**	**RM**
	R^2	0.14	0.24	0.36	0.28	0.80	0.18	0.48	0.32
	RMSE (m)	0.31	0.28	0.14	0.15	0.09	0.19	0.12	0.14
	RMSEr (%)	36.35	31.54	45.59	54.66	17.63	27.96	17.77	21.96
	B (m)	-0.04	0.00	-0.02	0.00	0.00	0.00	0.00	0.00
	Br (%)	40.75	80.88	60.39	15.24	2.29	7.06	3.58	4.28
Groupe test	**Taille de l'échantillon**	13		25		16		17	
		RNA	**RM**	**RNA**	**RM**	**RNA**	**RM**	**RNA**	**RM**
	R^2	0.63	0.59	0.35	0.40	0.84	0.56	0.75	0.63
	RMSE (m)	0.14	0.14	0.16	0.13	0.06	0.12	0.07	0.08
	RMSEr (%)	14.22	15.25	36.25	35.15	10.75	31.94	11.91	13.39
	B (m)	-0.05	-0.03	-0.09	-0.05	0.00	0.08	0.03	0.02
	Br (%)	-4.75	-2.40	-11.85	-1.86	1.90	20.48	5.40	4.41

† : Régression multiple

Les différentes méthodes d'estimation ainsi calibrées ont été appliquées au jeu de données consacrées pour la validation et qui correspondent à la période de mesure allant de 2000 à 2005. Ceci permet d'analyser les capacités d'extrapolation des différentes méthodes d'estimation étudiées.

Le Tableau 10 présente les résultats de validation pour l'estimation du débit sous glace. Le premier constat qui s'en dégage est que la limitation du nombre de variables explicatives incluses dans le modèle aux variables les plus significatives pour le débit contribue à améliorer les performances des deux catégories de modèle d'estimation.

Le deuxième constat est que le modèle neuronal est plus performant pour l'estimation du débit que le modèle régressif dans le cas où nous disposons d'un échantillon de calibration assez exhaustif comme dans le cas de la station 07BB002. Dans ces conditions, le réseau de neurones dispose de plus d'information afin de caractériser l'espace de l'erreur et par conséquent dispose de plus de chance pour localiser le minimum absolu de cet espace et converger vers la configuration optimale qui lui confère les capacités de généralisation requises d'un modèle d'estimation performant. Dans le cas contraire ou en présence de données de qualité douteuses, le RNA, malgré qu'il réussisse à s'ajuster aux données de calibration, perd ses capacités d'extrapolation et produit des résultats plus médiocres que le modèle régressif comme dans le cas de la station 05DF001.

Le troisième constat qui se dégage de ces résultats est que, par comparaison au modèle régressif et pour des niveaux d'erreur quadratique comparables, le RNA produit des biais d'estimation plus importants. Sachant que le RMSE est en fait la résultante de la variance et du biais de l'estimation, le RNA devient alors plus intéressant pour l'estimation du débit sous glace puisqu'il produit des estimations dont l'erreur résulte davantage d'un biais systématique plus facile à corriger (à l'aide d'une simple translation des valeurs estimées dans un sens ou dans un autre dépendant du sens du biais) qu'une erreur aléatoire plus difficile à corriger.

Quant aux résultats de validation des méthodes d'estimation de l'épaisseur de la glace, ils sont présentés au Tableau 11. Les constats relevés dans le cas de l'estimation du débit sous glace se trouvent confirmés avec l'épaisseur de la glace. Toutefois, les niveaux de variance expliqués par

les deux types de modèles demeurent plus faibles que dans le cas de l'estimation du débit, soulignant ainsi la nécessité de faire appel à des variables explicatives supplémentaires facilement disponible. Par ailleurs, mise à part la station 05DF001, le modèle neuronal en particulier celui utilisant la sélection réduite de variables explicatives a été plus performant que le modèle de régression dans l'estimation de l'épaisseur de glace.

Tableau 10 : Résultats de la validation des différentes méthodes étudiées pour l'estimation du débit sous glace

05DF001				
n=10	**RNA**	**RNASW†**	**RM‡**	**RSW£**
R^2	0.06	0.10	0.27	0.39
RMSE (m^3/s)	27.24	31.58	22.48	21.31
RMSEr (%)	24.80	41.87	20.64	19.35
B (m^3/s)	-0.86	-4.62	-3.35	-1.72
Br (%)	4.00	-1.01	1.08	3.04
07BB002				
n=24	**RNA**	**RNASW**	**RM**	**RSW**
R^2	0.74	0.75	0.54	0.56
RMSE (m^3/s)	1.40	1.51	2.56	2.32
RMSEr (%)	37.34	39.38	70.46	70.59
B (m^3/s)	1.01	1.22	1.27	1.27
Br (%)	56.24	70.78	58.21	59.34
07CD001				
n=13	**RNA**	**RNASW**	**RM**	**RSW**
R^2	0.65	0.65	0.71	0.71
RMSE (m^3/s)	6.93	6.93	6.96	6.96
RMSEr (%)	14.30	14.30	14.20	14.20
B (m^3/s)	0.63	0.63	-0.77	-0.77
Br (%)	4.06	4.06	1.48	1.48
07DA001				
n=16	**RNA**	**RNASW**	**RM**	**RSW**
R^2	0.71	0.81	0.80	0.86
RMSE (m^3/s)	56.08	46.64	51.63	57.30
RMSEr (%)	25.34	24.61	23.16	36.90
B (m^3/s)	38.30	40.78	26.50	23.50
Br (%)	27.07	31.09	15.50	10.91

† : Modèle de RNA utilisant la sélection de variables explicatives identifiées par la régression stepwise
‡ : Régression multiple utilisant la sélection originale de variables explicatives
£ : Régression multiple utilisant la sélection de variables explicatives identifiées par la régression stepwise

Tableau 11 : Résultats de validation des différentes méthodes étudiées pour l'estimation de l'épaisseur de la glace

05DF001				
n=10	**RNA**	**RNASW†**	**RM‡**	**RSW£**
R^2	0.00	0.00	0.21	0.21
RMSE (m)	0.33	0.33	0.29	0.29
RMSEr (%)	30.98	30.98	31.40	31.40
B (m)	0.21	0.21	0.21	0.21
Br (%)	46.76	46.76	43.33	43.33
07BB002				
n=23	**RNA**	**RNASW**	**RM**	**RSW**
R^2	0.19	0.42	0.33	0.33
RMSE (m)	0.18	0.17	0.16	0.17
RMSEr (%)	42.65	44.44	36.40	39.08
B (m)	-0.08	-0.12	-0.08	-0.10
Br (%)	-11.57	-18.74	-10.08	-13.65
07CD001				
n=13	**RNA**	**RNASW**	**RM**	**RSW**
R^2	0.20	0.67	0.17	0.60
RMSE (m)	0.23	0.17	0.24	0.19
RMSEr (%)	44.10	33.68	52.51	33.42
B (m)	-0.14	-0.12	-0.17	-0.10
Br (%)	-15.13	-15.32	-18.67	-6.81
07DA001				
n=15	**RNA**	**RNASW**	**RM**	**RSW**
R^2	0.46	0.77	0.53	0.58
RMSE (m)	0.14	0.12	0.15	0.12
RMSEr (%)	19.77	18.42	22.20	19.25
B (m)	0.10	0.10	0.12	0.08
Br (%)	19.55	20.16	25.39	18.02

† : Modèle de RNA utilisant la sélection de variables explicatives identifiées par la régression stepwise

‡ : Régression multiple utilisant la sélection originale de variables explicatives

£ : Régression multiple utilisant la sélection de variables explicatives identifiées par la régression stepwise

4 CONCLUSIONS

L'objectif visé dans cette étude était de développer une méthodologie pour la correction du débit de rivières en présence de glace et pour l'estimation de l'épaisseur de la glace. Pour ce faire nous avons exploré les performances de deux types d'approches d'estimation la première étant les réseaux de neurones artificiels et la seconde l'approche régressive (régression multiple, régression stepwise). En ce qui concerne la correction du débit, ces approches sont caractérisées par leur objectivité et leur reproductibilité par opposition aux méthodes présentement appliquées qui sont plutôt caractérisées par leur subjectivité et non reproductibilité. Pour l'estimation de l'épaisseur de la glace, les méthodes étudiées sont caractérisées par leur simplicité et leur flexibilité puisqu'elles ne requièrent que les données hydrométriques et météorologiques standards facilement accessible. Les deux approches ont été appliquées sur quatre rivières de l'Alberta présentant différents régimes hydrologiques et différentes conditions et disponibilité et qualité de données.

Les résultats ont démontré que l'approche neuronale est supérieure par rapport à l'approche régressive pour l'estimation du débit en présence de glace surtout dans le cas de la disponibilité d'un grand jeu de données de calibration. Pour l'estimation de l'épaisseur de la glace l'approche neuronale est également plus appropriée et ce, dans tous les cas. Toutefois, les résultats obtenus pour cette variable laissent croire qu'il y a place à l'amélioration en utilisant d'autres variables explicatives supplémentaires. Les résultats ont démontré également que l'emploi d'une sélection de variables explicatives améliore les performances des deux approches. Dans ce sens, la régression stepwise constitue un outil efficace pour l'identification automatique mais rigoureuse des variables explicatives les plus pertinentes à inclure dans le modèle d'estimation. Par ailleurs, en termes d'erreur, l'approche neuronale serait plus appropriée pour l'estimation du débit sous glace ainsi que pour l'épaisseur de glace puisque l'erreur d'estimation du modèle est constituée pour la plupart d'erreur systématique pouvant être facilement corrigée.

Nous avons démontré dans cette étude le potentiel des deux approches étudiées pour l'estimation du débit et de l'épaisseur de la glace. Cependant, il serait intéressant de comparer les résultats obtenus ici avec les résultats d'autres approches appliquées sur les mêmes rivières, surtout en ce qui concerne l'épaisseur de glace où il serait intéressant de comparer les performances des RNA à celles des modèles thermodynamiques et des modèles plus déterministes de croissance de la glace.

Par ailleurs, nous avons développé un outil informatique d'estimation à temps réel des débits en présence de glace et de l'épaisseur de la glace. Cependant, les méthodes d'estimation implantées dans cet outil ne sont pas adaptatives. En effet, les techniques d'estimation étant basées sur des données historiques, les nouveaux jaugeages hivernaux acquis au cours de la saison ne peuvent être pris en compte automatiquement pour la saison en cours. Pour se faire, il faut soit calibrer de nouveau les différents modèles chaque fois que les données d'un nouveau jaugeage sont disponibles soit attendre la fin de la saison hivernale pour les prendre en compte. Grâce à la flexibilité de l'outil développé, la première option ne présente pas vraiment un obstacle car les modèles mis à jour peuvent être disponibles dans un délai raisonnable : moins d'une heure suivant l'acquisition des nouvelles données.

RÉFÉRENCES

Chokmani K., Ghedira M. H., Gingras H., Ouarda T. B. M. J., Bobée B. 2003. Correction du débit en présence d'un effet de glace : développement du logiciel UNICCO. Rapport scientifique No. R-687, INRS-ETE, Québec, 125 p.

Draper N. R., Smith H. 1966. Applied regression analysis. Wiley, New York.

Neter J., Wasserman W., Kutner M. H. 1985. Applied linear Statistical models. Irwin, Homewood, Illinois.

Ouarda T. B. M. J., Faucher D., Coulibaly P., Bobée B. 2000. Correction du débit en présence d'un effet de glace; étude de faisabilité pour le développement d'un logiciel. Rapport scientifique No. R-559, INRS-Eau, Québec, 70 p.

Rumelhart D. E., Hinton G. E., Williams R. J. 1986. Learning internal representation by error propagation. Parallel distributed processing : Exploration in the microstructure of cognition. D. E. Rumelhart, J. L. McClelland, MIT Press. 1: 318-364.

Seidou O., Hessami M., Ouarda T. B. M. J., St-Hilaire A., Bilodeau L., Bruneau P., Bobée B. 2005. Modélisation de la croissance de glace de lac par réseaux de neurones artificiels et estimation du volume de la glace abandonnée sur les berges des réservoirs hydroélectriques pendant les opérations d'hiver. Rapport scientifique No. R-788, INRS-ETE, Québec, 72 p.

Weisberg S. 1985. Applied linear regression (Second Edition). Wiley & Sons, New York.

Werbos P. J. 1974. Beyond regression : New tools for prediction and analysis in the behavioral sciences. Thèse de doctorat non publiée, Université Harvard.