RAPPORT

CORRECTION DU DÉBIT EN PRÉSENCE D'UN EFFET DE GLACE : DÉVELOPPEMENT DU LOGICIEL UNICCO

Rapport de recherche No R-687 *Avril 2003*

CORRECTION DU DÉBIT EN PRÉSENCE D'UN EFFET DE GLACE : DÉVELOPPEMENT DU LOGICIEL *UNICCO*

Rapport préparé à l'attention de:

Environnement Canada
Région Pacifique et Yukon,
Suite 120, 1200 West 73rd Ave.
Vancouver, B.C., V6P 6H9

par:

Karem Chokmani

Hosni Ghedira

Hugo Gingras

Taha B.M.J. Ouarda

Stuart Hamilton

Bernard Bobée

Chaire Hydro-Québec/CRSNG/Alcan en Hydrologie Statistique
Institut National de la Recherche Scientifique, INRS-ETE
2800, rue Einstein, C.P. 7500, Sainte-Foy (Québec) G1V 4C7

Rapport de recherche No R-687

ISBN : 2-89146-503-2

Avril 2003

ÉQUIPE DE RECHERCHE

Ont participé à la réalisation de cette étude:

Chaire en Hydrologie Statistique

Institut National de la Recherche Scientifique, INRS-ETE

Karem Chokmani

Hosni Ghedira

Hugo Gingras

Taha B.M.J. Ouarda

Bernard Bobée

Environnement Canada

Stuart Hamilton

AVANT PROPOS

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

Nous tenons à remercier M. Stuart Hamilton pour sa précieuse collaboration en particulier pour ses commentaires et remarques judicieux ainsi que pour avoir fourni les données utilisées dans cette étude.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

1 INTRODUCTION

Pour assurer une bonne gestion des ressources hydriques, et un développement durable dans les secteurs associés à l'eau, il est nécessaire de posséder une base de données hydrométriques fiable et de bonne qualité. Cependant, les séries de débits de rivières correspondant à la période hivernale sont souvent de qualité inférieure à celle correspondant au reste de l'année. En effet, une proportion importante des rivières canadiennes sont affectées par l'effet de glace ; 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.). La courbe de tarage définit la relation entre le niveau d'eau dans une section transversale et le débit correspondant. En général, la courbe est construite dans une section stable de la rivière à partir de plusieurs observations niveau-débit. Idéalement, ces observations doivent inclure des valeurs extrêmes pour assurer la bonne extrapolation de la courbe. Cependant, cette courbe de tarage ne peut pas être représentative de la période hivernale à cause du changement de la section d'écoulement 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 Environnement Canada (EC) effectuent généralement des jaugeages durant la période de présence de glace pour estimer le débit réel qui s'écoule dans les rivières. Ensuite, pendant le reste de la période hivernale, les débits 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. En effet, il s'agit d'interpoler les deux ou trois jaugeages d'hiver pour chaque station, afin de reconstituer les débits de toute la période hivernale. Cette approche mène généralement à des résultats satisfaisants mais risque d'introduire des erreurs assez importantes lors de fonte hivernale, embâcles de glace, etc. De surcroît, l'approche est caractérisée par sa subjectivité et sa non-reproductibilité. En plus, elle ne peut être appliquée qu'a posteriori, après la fin de la période hivernale, et donc ne permet pas d'estimer les débits de rivières sur une base journalière durant la période hivernale.

Par ailleurs, le débit durant la période hivernale est de plusieurs ordres de grandeur inférieur à celui de la période printanière et estivale. Une erreur dans son estimation risque alors d'être critique et d'avoir des conséquences fâcheuses pour la faune aquatique qui devient extrêmement vulnérable durant cette période. De plus, la dilution et la dispersion des effluents acquièrent une importance encore plus grande durant cette période à cause des faibles débits dans les rivières. D'autre part, les embâcles et les débâcles de glace sont directement reliés à la formation et à la fonte de la glace dans les rivières, et à la valeur du débit durant les périodes critiques. Les crues causées par les embâcles de glace causent plus de 60 millions de dollars en dommages chaque année au Canada. D'où l'importance d'effectuer une bonne correction du débit en présence d'un effet de glace.

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. Celles-ci ont été classées essentiellement en deux catégories, soit les méthodes subjectives, qui font appel à une certaine forme de jugement de la part de l'utilisateur, et les méthodes analytiques, indépendantes du jugement de l'utilisateur. D'autre part, les auteurs ont étudié la faisabilité du développement d'un logiciel qui permettra d'effectuer la tâche de correction du débit en présence de glace d'une façon automatique et en temps réel. Ils ont conclu de la faisabilité d'un tel système informatique. Il en ressort également 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 souvent des quantités de données qui ne sont pas toujours disponibles pour leur assurer une bonne calibration.

Dans la présente étude, il a été question d'explorer les performances des réseaux de neurones artificiels (RNA) pour la correction du débit en présence de glace et ce, à l'aide d'une combinaison de variables hydrométriques et météorologiques facilement disponibles en quantité et qualité adéquate. Les performances des RNA ont été comparées à celles du modèle régressif utilisant la même combinaison de variables explicatives. Les résultats des

deux méthodes d'estimation ont été également confrontés aux débits corrigés par les services d'Environnement Canada à l'aide de leur propre méthode. 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. Lors d'une étude similaire (Gingras *et al.*, 2002) réalisée par l'équipe de la chair pour le compte du Ministère de l'Environnement du Québec, il a été souligné l'intérêt de l'utilisation de la méthode «stepwise » afin d'identifier le modèle de régression optimal qui fait intervenir les variables explicatives les plus significatives pour expliquer les fluctuations de la variable dépendante (débit). D'autre part, les variables explicatives utilisées pour la calibration de la fonction régressive peuvent ne pas être complètement indépendantes (multicollinéarité). Ce qui peut conduire à produire des paramètres du modèle de régression imprécis. La régression « ridge » peut être utilisée afin de tenir compte du problème de multicollinéarité. Par conséquent, les deux méthodes régressives stepwise et ridge ont été également implantées dans l'outil informatique *UNICCO* et leurs résultats ont été comparés aux deux autres techniques.

Le premier chapitre a été consacré à exposer les fondements théoriques des méthodes d'estimation implantées dans le logiciel *UNICCO.* 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 la rivière Fraser en Colombie Britannique. La description détaillée du logiciel *UNICCO* est présentée en annexe du présent rapport.

2 MÉTHODES D'ESTIMATION

2.1 RÉGRESSION LINÉAIRE

2.1.1 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).

La régression linéaire est une méthode statistique utilisée pour étudier la nature et la signification de la relation linéaire qui peut exister entre une variable réponse et une ou plusieurs variables explicatives. Elle permet ainsi de construire un modèle décrivant cette relation. Celui-ci peut être alors utilisé simplement pour décrire cette relation ou dans une optique prédictive, c'est-à-dire qu'il permet, pour une autre période que celle qui a servi à calibrer le modèle, d'utiliser l'information mesurée des variables explicatives pour estimer la valeur de la variable réponse.

2.1.2 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 :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon \qquad (1)$$

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 :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \qquad (2)$$

ou encore

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i \qquad (3)$$

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;
- ε_i, $i = 1, 2, ..., n$, sont des variables aléatoires indépendantes et identiquement distribuées selon une loi normale centrée de variance σ^2.

Le modèle (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. Trois hypothèses de base concernant les termes ε_i sont associées au modèle de régression multiple:

1. *Normalité des erreurs* : ε_i est une variable aléatoire distribuée selon une loi normale centrée de variance σ^2;
2. *Absence de corrélation des erreurs* : les termes ε_i et ε_j relatifs à deux observations i et j n'ont aucune corrélation entre eux, $Cov\{\varepsilon_i, \varepsilon_j\} = 0, i \neq j$;
3. *Homoscédasticité* : la variance des ε_i est constante quelles que soient les valeurs des variables explicatives $x_{i1}, x_{i2}, ..., x_{ip}$, $Var\{\varepsilon_i\} = \sigma^2, \forall i$.

Lorsqu'une seule variable explicative est utilisée dans le modèle de régression (régression simple), cette fonction est une droite. Les paramètres β_0 et β_1 sont alors respectivement l'ordonnée à l'origine et la pente de cette droite. Si le modèle possède deux variables explicatives, la fonction de régression est un plan. Enfin, si plus de deux variables sont utilisées, la fonction est un hyperplan.

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

$$\underset{n\times 1}{\mathbf{Y}}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \underset{n\times(p+1)}{\mathbf{X}}=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots x_{1p} \\ 1 & x_{21} & x_{22} & \cdots x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots x_{np} \end{bmatrix}$$

$$\underset{(p+1)\times 1}{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \underset{n\times 1}{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times(p+1)}{\mathbf{X}}\ \underset{(p+1)\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \tag{4}$$

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 e_i définis comme suit :

$$e_i = y_i - \left(\hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k x_{ik} \right) \tag{5}$$

$\hat{\beta}_0$ 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 :

$$Q=(\mathrm{Y}-\mathrm{Xb})'(\mathrm{Y}-\mathrm{Xb})=\mathrm{Y}'\mathrm{Y}-\mathrm{bX}'\mathrm{Y} \tag{6}$$

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

$$(X'X)b=X'Y \tag{7}$$

et on déduit :

$$b=(X'X)^{-1}X'Y \tag{8}$$

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

Une fois les paramètres de la régression sont estimés, il est possible de calculer le vecteur $\hat{\mathbf{Y}}$ des valeurs prédites $\hat{y}_i$, et le vecteur correspondant **e** des écarts résiduels $e_i = y_i - \hat{y}_i$. Sous forme matricielle, les valeurs prédites de la variable dépendante ainsi que les résidus s'expriment respectivement comme suit :

$$\hat{\mathbf{Y}} = \mathbf{Xb} \text{ et } \mathbf{e} = \mathbf{Y} - \mathbf{Xb}$$

A l'aide de ces valeurs il est possible d'étudier les différentes sources de variation associées à la variable aléatoire dépendante *Y*. Les différentes sources de variation nous amènent à définir une mesure permettant de quantifier l'effet de l'introduction des variables explicatives $X_1, X_2, ..., X_p$ dans le modèle sur la réduction de la variation totale de *Y*. Cette statistique est le coefficient de détermination multiple R^2 qui s'exprime comme suit:

$$R^2 = \frac{SCT - SCE}{SCT} = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} \tag{9}$$

* *SCT* est la somme des carrés des écarts entre les données et leur valeur moyenne. Elle représente en fait une mesure de la variation totale :

$$SCT=\sum_{i=1}^{n}(y_i-\bar{y})^2=Y'Y-\frac{1}{n}Y'll'Y \tag{10}$$

où **1** est un vecteur colonne de dimension n dont tous les éléments sont égaux à 1.

* *SCE* est la somme des carrés des écarts entre les données et le modèle de régression et elle exprime la variabilité non expliquée par le modèle de régression :

$$SCE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\mathrm{Y'Y-b'X'Y} \tag{11}$$

* *SCR* est la somme des carrées de la régression et elle exprime la variabilité expliquée par la régression :

$$SCR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\mathrm{b'X'Y}-\frac{1}{n}\mathrm{Y'll'Y} \tag{12}$$

R^2 peut être interprété comme le pourcentage de variation totale expliquée par le modèle. Plus R^2 est grand, plus l'apport des variables explicatives est important pour expliquer la variabilité totale de *Y*. Cette mesure permet de juger de l'adéquation du modèle. Toutefois, il est important de noter que R^2 augmente à mesure que l'on ajoute des variables explicatives au modèle, et ce même si ces variables ne sont pas significativement reliées à la variable dépendante. C'est pourquoi il est plus judicieux d'utiliser le R^2 ajusté, noté $R_a^{\,2}$, que l'on définit comme suit :

$$R_a^{\,2}=1-\frac{n-1}{n-p}\times(1-R^2) \tag{13}$$

où *n* est la taille de l'échantillon et *p* est le nombre de variables explicatives

2.1.3 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 X_k les plus significatives pour expliquer les fluctuations aléatoires de la variable dépendante *Y*. 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 obtenus 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 multicollinéarité 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 que l'on note F^*, 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 lorsque aucune variable explicative ne peut être ajoutée ou retranchée de l'équation de régression.

Ainsi, à la première étape, on ajoute la variable explicative X_k ayant la valeur de $F^*(k)$ la plus élevée supérieure à la valeur théorique d'entrée F_a. Si la valeur de la statistique est inférieure, la régression « stepwise » est terminée et aucun lien n'existe entre Y et les autres variables explicatives. Pour chacun des modèles de régression simple, la statistique $F^*(k)$ est calculée à l'aide la relation suivante:

$$F^*(k)=(n-2)\frac{SCR(X_k)}{SCE(X_k)}, k=1,\ldots,p \qquad (14)$$

où $SCR(X_k)$ et $SCE(X_k)$ sont respectivement les sommes des carrés de la régression et des résidus correspondant au modèle incluant X_k.

À l'étape suivante, on ajuste alors tous les modèles à deux variables dont l'une d'elles est la variable X_i retenue à l'étape précédente et l'autre est l'une des variables explicatives inutilisée. Pour chacun de ces modèles de régression, la statistique suivante est calculée :

$$F^*(i,k)=(n-3)\frac{SCR(X_i,X_k)-SCR(X_i)}{SCE(X_i,X_k)},k=1,\ldots,i-1,i+1,\ldots,p \qquad (15)$$

où $SCR(X_i,X_k)$ et $SCE(X_i,X_k)$ sont respectivement la somme des carrées de la régression et des résidus du modèle à deux variables, et $SCR(X_i)$ la somme des carrés de la régression du modèle à une variable obtenue à l'étape précédente. Comme dans la première étape, la variable explicative X_k correspondant à la plus grande valeur de $F^*(i,k)$ est retenue. Cette valeur est ensuite comparée à F_a. Si elle est supérieure, la variable correspondante est ajoutée au modèle de l'étape précédente, sinon, la sélection est terminée et une régression simple avec X_i est jugée satisfaisante.

Par la suite, la procédure « stepwise » examine si l'une des deux variables explicatives du modèle doit être retranchée. Une statistique F^* est alors calculée pour chaque variable comme si chacune d'elle était la dernière variable introduite dans le modèle. Dans notre exemple, il n'y a qu'un seul calcul à effectuer puisqu'il s'agit d'un modèle à deux variables explicatives X_i et X_j. Il est question alors de calculer la statistique de retrait de la variable X_i entrée à la première étape, $F^*(i,j)$:

$$F^*(i,j)=(n-3)\frac{SCR(X_i,X_j)-SCR(X_i)}{SCE(X_i,X_j)} \qquad (16)$$

Si $F^*(i,j)$ est inférieure à la valeur théorique critique de retrait F_r, la variable X_i est retirée du modèle. Dans le cas où on est en présence d'un modèle à plusieurs variables, la

variable présentant la plus petite statistique et dont la valeur est inférieure à F_r est retirée du modèle. Sinon, elle demeure dans le modèle

Les valeurs critiques d'entrée et de sortie des variables explicatives, F_a et F_r , sont établies en fonction du risque α (niveau de signification des tests), du nombre d'observations n et du nombre potentiel de variables explicatives p. Habituellement, on suggère souvent d'utiliser $F_a = F_r = F_{1,n-p-1}(1-\alpha)$, où $F_{1,n-p-1}(1-\alpha)$ est le quantile de probabilité au non-dépassement 1 - α de la loi de Fisher à *1* et (*n-p-1*) degrés de liberté.

2.1.4 Régression « Ridge »

L'un des postulats de la régression linéaire multiple est l'indépendance des variables explicatives utilisées pour bâtir le modèle régressif. En pratique, nous sommes souvent confrontés au phénomène de multicollinéarité où les variables explicatives sont plus ou moins corrélées entre elles. Ce phénomène se traduit par des problèmes typiques tel :

- Le retrait ou l'ajout d'une variable explicative au modèle peut changer radicalement la valeur des paramètres estimés des autres variables;
- Les estimations des paramètres peuvent avoir une variance très élevée et donc être imprécises, et parfois même erronées;
- Les paramètres peuvent ne pas être statistiquement significatifs alors qu'une relation entre la variable dépendante et les variables explicatives existe.

En effet, en présence de multicollinéarité, la méthode des moindres carrés utilisée pour l'estimation paramètres de la régression perd de son efficacité. Il faut rappeler que l'estimation du vecteur des paramètres β fait intervenir l'inversion de la matrice $\mathbf{X'X}$. Or, cette matrice devient de plus en plus difficile à inverser à mesure que la corrélation entre les variables explicatives augmente. Ainsi, si ces corrélations croisées sont grandes, les éléments de la matrice inversée peuvent être erronés et engendrer une instabilité dans l'estimation des paramètres. À la limite, s'il existe une corrélation parfaite entre deux

variables explicatives, la matrice **X'X** sera singulière, donc non-inversible, et les paramètres ne pourront être déterminés.

Une solution à ce problème consiste à bien choisir les variables explicatives à incorporer dans le modèle. Il s'agit de limiter le plus possible le nombre de variables explicatives, d'éviter l'utilisation de variables explicatives fortement corrélées entre elles, d'utiliser les variables les plus pertinentes pour le problème à traiter. Dans nombreux cas, on n'a pas la possibilité de pouvoir effectuer ce genre de choix, soit les variables explicatives sont disponibles en nombre insuffisant soit par manque de connaissance du phénomène à modéliser. Dans ce cas, on peut faire appel à d'autres approches moins subjectives telle que la régression ridge.

La régression ridge a été proposée pour la première fois par Hoerl et Kennard (1970a, 1970b). Cette approche intervient directement sur l'estimation des paramètres de la régression en modifiant la méthode des moindres carrés. Les paramètres obtenus par la régression ridge sont biaisés mais néanmoins plus stables que ceux issus de la régression multiple en présence de multicollinéarité. Ainsi, on troque un faible biais sur les paramètres contre une plus grande précision de leurs valeurs augmentant du coup la probabilité d'être proches des valeurs cible inconnues.

Lors de la régression ridge, nous procédons tout d'abord à la standardisation des observations. Les nouvelles observations standardisées sont données par :

$$y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{y_i - \bar{y}}{s_Y}\right) \quad \text{et} \quad x_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{x_{ik} - \bar{x}_k}{s_k}\right), \quad \text{pour } k = 1, 2, ..., p$$

où S_y et S_k sont les écart-types de la variable dépendante et des variables explicatives, respectivement. Le modèle de régression en fonction des variables standardisées peut s'écrire alors comme suit:

$$y_i^* = \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + ... + \beta_p^* x_{ip}^* + \varepsilon_i \qquad (17)$$

Il possible de démontrer alors que les paramètres β_k du modèle original peuvent être obtenus en fonction des paramètres standardisés β_k^* :

$$\beta_k = \left(\frac{s_Y}{s_k}\right)\beta_k^* \text{ , pour } k = 1, 2, ..., p \quad \text{et} \quad \beta_0 = \bar{y} - \sum_{k=1}^{p}\beta_k \bar{x}_k \qquad (18)$$

Il est possible aussi de démontrer que les matrices **X'X** et **X'Y** qui interviennent dans le calcul des estimateurs (éq. 8) peuvent être respectivement remplacées par la matrice des corrélations entre les variables explicatives, $\mathbf{r}_{XX}$, et la matrice des corrélations entre la variable dépendante et les variables explicatives, $\mathbf{r}_{YX}$. Les estimateurs obtenus par la méthode des moindres carrés s'écrivent alors sous forme matricielle comme suit :

$$\mathbf{b}^* = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX} \qquad (19)$$

où $\mathbf{r}_{XX}^{-1}$ est l'inverse de la matrice des corrélations entre les variables explicatives.

Les paramètres standardisés estimés par la régression ridge sont calculés en introduisant une constante $k \geq 0$ dans les équations de la méthode des moindres carrés :

$$\mathbf{b}^R = (\mathbf{r}_{XX} - k\mathbf{I}_p)^{-1}\mathbf{r}_{YX} \qquad (20)$$

où $\mathbf{b}^R = (b_1^R, b_2^R, ..., b_p^R)$ est le vecteur des paramètres standardisés estimés par la régression ridge et $\mathbf{I}_p$ est la matrice identité de dimension $p \times p$. L'addition d'une constante k à chacun des éléments situés sur la diagonale de la matrice $\mathbf{r}_{XX}$ a pour effet de faciliter l'inversion de celle-ci. Une fois paramètres standardisés sont calculés, il est possible retrouver les paramètres non-standardisés en appliquant l'équation (18).

La constante k représente le biais commis sur les paramètres. Si $k = 0$, l'équation (20) est équivalente à (19) et les paramètres obtenus par la régression ridge coïncident avec ceux non-biaisés des moindres carrés ordinaires. Lorsque $k > 0$, les paramètres sont biaisés mais plus précis que ceux de la régression multiple. À mesure que la constante k augmente, le biais de $\mathbf{b}^R$ augmente mais sa variance diminue. On peut montrer qu'il existe toujours une valeur de k telle que $\mathbf{b}^R$ possède un écart quadratique moyen inférieur à celui **b** (Hoerl et

Kennard, 1970a,b). Toutefois, il est difficile de choisir la constante k puisque la valeur optimale varie d'un ensemble de données à l'autre.

Une méthode visuelle utilisée couramment pour déterminer la constante k repose sur l'examen du graphique des traces. Ces derniers représentent simultanément des estimations des p paramètres standardisés $b_1^R, b_2^R, ..., b_p^R$ obtenus pour différentes valeurs de la constante k, généralement comprises entre 0 et 1. Il s'agit de choisir graphiquement la plus petite valeur de k qui correspond à une zone de stabilité des courbes associées à chacun des paramètres (Hoerl et Kennard, 1970b). Toutefois, Vinod (1976) a montré que cette procédure peut amener à surestimer la valeur de k. Cet auteur a donc proposé, comme méthode complémentaire, une procédure automatique permettant de déterminer la valeur du paramètre k et que nous avons adopté dans la présente étude. Cette procédure utilise l'indice *ISRM* défini par :

$$ISRM = \sum_{i=1}^{p} \left[\frac{(\lambda_i - k)^2}{\sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + k}} - 1 \right]^2 \qquad (21)$$

où $\lambda_1, \lambda_2, ..., \lambda_p$ sont les valeurs propres de la matrice $\mathbf{r}_{XX}$. Cet indice est nul si les variables explicatives sont non-corrélées. Vinod (1976) suggère d'utiliser le k qui correspond à la valeur minimale de l'indice *ISRM*. Pour un ensemble d'observations donné, la valeur optimale de k est obtenue en minimisant la fonction de l'*ISRM* (éq. 21).

2.2 RÉSEAUX NEURONAUX

2.2.1 Introduction

Les réseaux de neurones sont constitués d'un ensemble de neurones artificiels ou nœuds qui sont analogues aux neurones biologiques. Ils sont issus d'une tentative de conception d'un modèle mathématique très simplifié du cerveau humain en se basant sur notre façon d'apprendre et de corriger nos erreurs. Les premiers travaux sur les réseaux de neurones ont été réalisés en 1943 par Mc Culloch et Pitts. Ces deux chercheurs sont les premiers à

montrer théoriquement que des réseaux de neurones formels et simples peuvent réaliser des fonctions logiques, arithmétiques et symboliques. Ils ont ainsi présenté un modèle assez simple pour les neurones permettant d'explorer différentes possibilités d'applications.

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 (1974) et diffusé par Rumelhart *et al.* (1986).

2.2.2 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.

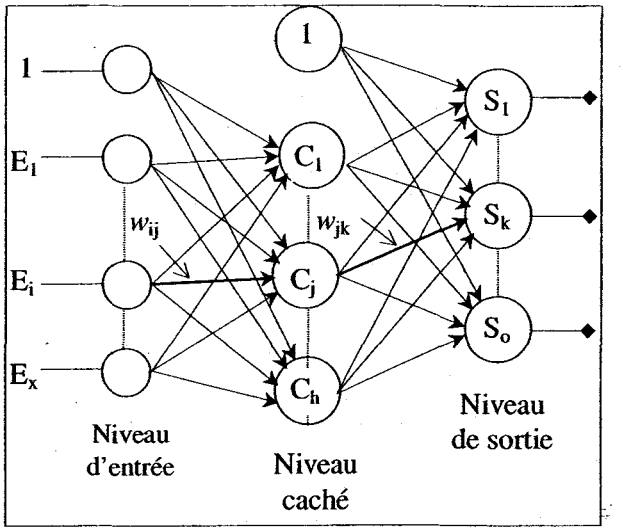


Figure 1 Architecture d'un réseau multicouches

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.

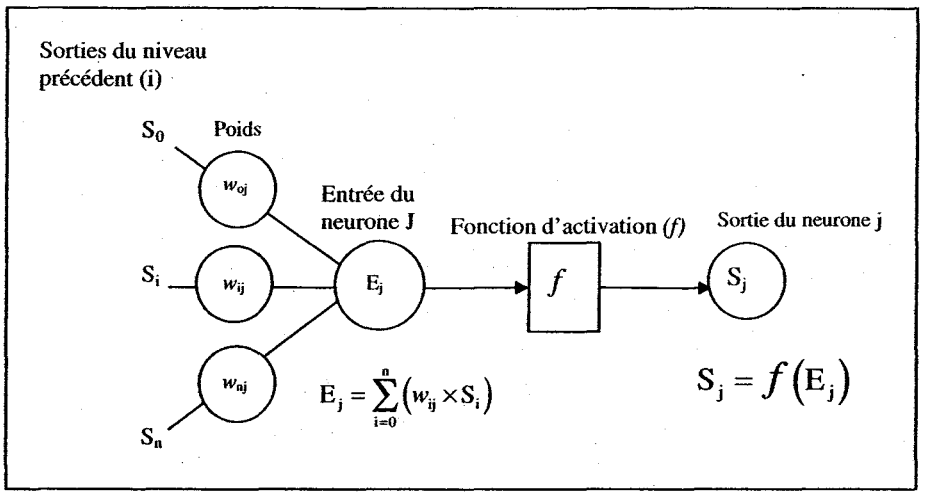


Figure 2 Connections d'un élément processeur (nœud j)

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

$$S_j = f(E_j)$$

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

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

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.

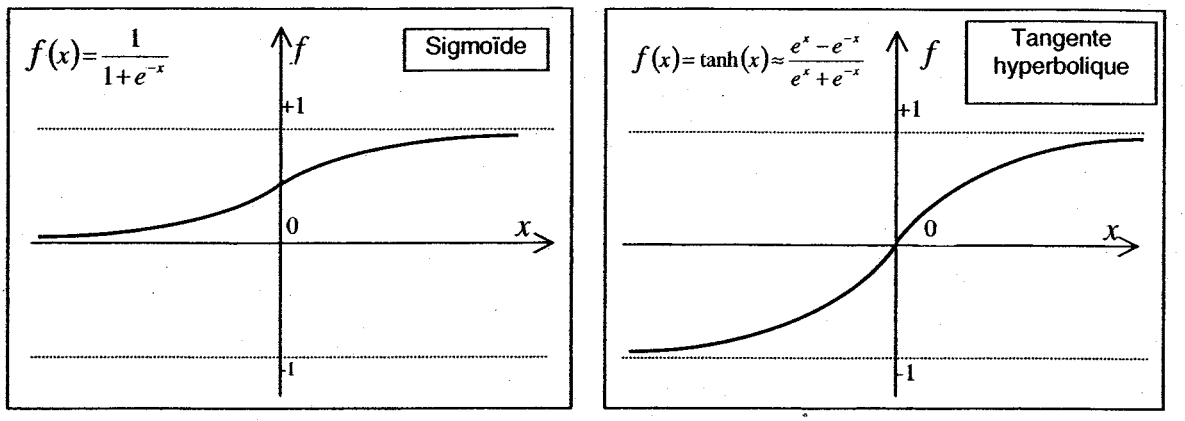


Figure 3 Fonctions d'activation

2.2.3 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 sorties 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 expérimentalement), 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 (**Erreur ! Source du renvoi introuvable.**). 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 appartenants 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.

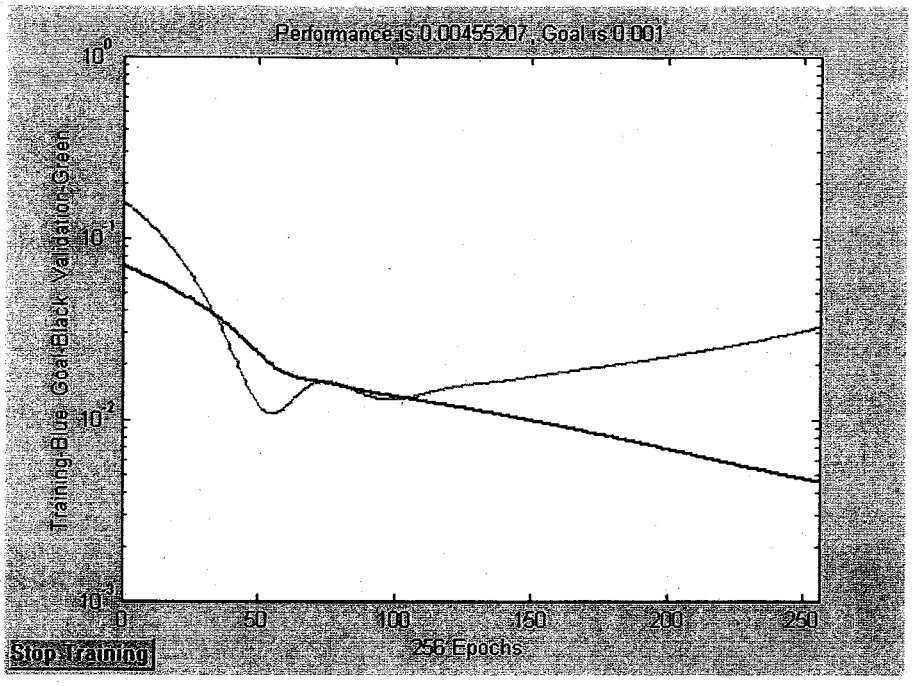


Figure 4 Évolution de l'erreur au cours de la phase d'apprentissage

2.2.4 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.

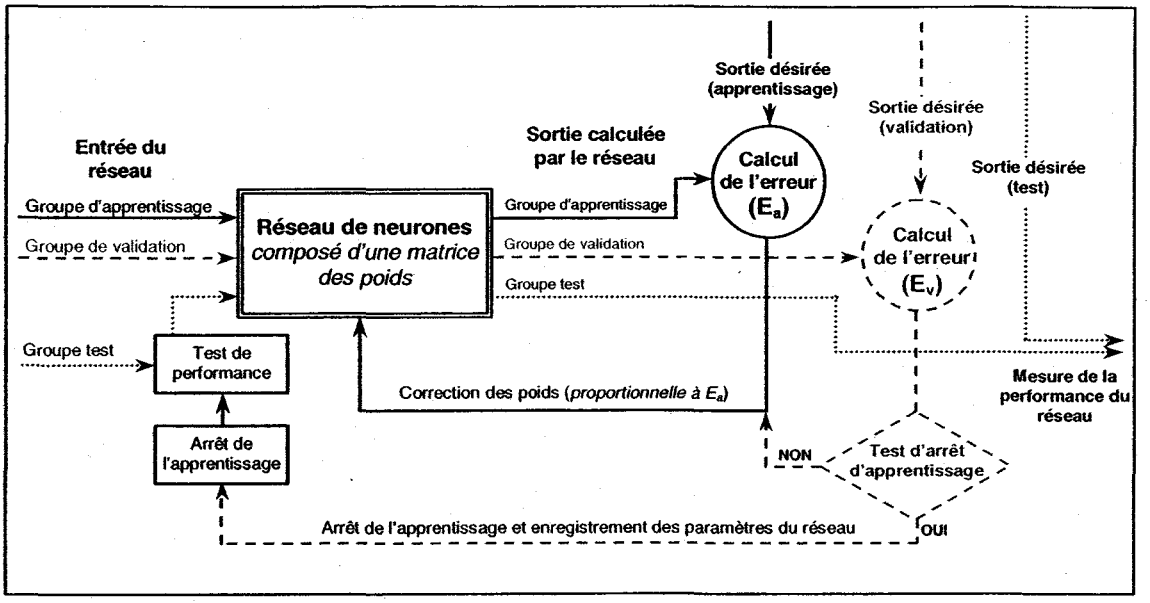


Figure 5 Description schématique de la phase d'apprentissage

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.

3 APPLICATION

3.1 MÉTHODOLOGIE

L'objectif de la présente étude est d'explorer les performances des réseaux de neurones artificiels ainsi que des techniques régressives pour l'estimation du débit en présence de glace et ce, à l'aide d'une combinaison de variables hydrométriques et météorologiques facilement disponibles. Les résultats des deux méthodes d'estimation ont été également confrontés aux débits corrigés par les services d'Environnement Canada.

Les variables explicatives utilisées ont été recueillies sur la rivière Fraser en Colombie Britannique et nous ont été fournies par les services d'Environnement Canada. Elles comprennent essentiellement : le niveau d'eau dans la section de la rivière, la température journalière moyenne de l'air, la température moyenne décadaire, les précipitations liquides journalières, le cumul de précipitations liquides décadaires, la hauteur de la neige au sol, le nombre de jour depuis le gel de la surface de l'eau, les degrés-jours depuis le gel de la surface de l'eau, le dernier débit enregistré avant le gel de la surface de l'eau et la radiation solaire.

En commun accord avec les services d'Environnement Canada, il a été question aussi d'explorer les performances de ces méthodes en faisant abstraction du niveau d'eau. Il a été question également d'utiliser une sélection réduite de variables explicatives. Celles-ci ont été choisies selon leur pertinence pour l'estimation du débit et leur degré de corrélation observée avec celui-ci.

En premier lieu, nous avons procédé à la calibration des modèles des différentes techniques d'estimation à l'aide des débits jaugés en période hivernale comme variable réponse et le niveau automatique et les variables météorologiques, enregistrées à la station météorologique la plus proche de la station de jaugeage, comme variables explicatives. Il faut noter, que seuls les jaugeages pour lesquels la présence de glace en surface est confirmée ont été utilisés dans la calibration. Pour se faire, nous nous sommes basés sur les observations sur l'état de surface de l'eau recueillies par les services d'Environnement

Canada. En cas d'absence du niveau automatique, nous nous sommes rabattus sur le niveau jaugé mesuré à la station à la même heure, quand il est disponible,.

Afin de juger de la qualité de la calibration des modèles, nous nous sommes basés sur quatre critères. Le premier est le coefficient de détermination entre les données originales et les valeurs estimées(décrit plus haut). Le deuxième est l'erreur quadratique relative moyenne (*RMSEr*) qui représente la racine carrée de la moyenne des écarts au carré entre les observations originales et les estimées rapportés aux valeurs des observations originales. Elle peut être calculée comme suit :

$$RMSEr = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \tag{22}$$

où y et $\hat{y}$ étant, respectivement, les valeurs originales et estimées de la variable réponse et n le nombre d'observations. Le troisième est le biais relatif moyen (*BIAISr*). Ce dernier représente la moyenne des écarts relativisés par rapport aux observations originales :

$$BIAISr = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \tag{23}$$

Quant au critère de *Nash*, il consiste en le rapport entre la somme des carrées et la variance de la variable observée :

$$Nash = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{24}$$

où y, $\bar{y}$ et $\hat{y}$ représentent, respectivement, les valeurs originales, leur moyenne et les valeurs estimées par le modèle. Si le critère de *Nash* est négatif, le modèle est moins bon que l'utilisation de la moyenne des valeurs observées. Pour un modèle s'ajustant parfaitement aux données observées le critère de *Nash* vaut 1. En général, on s'attend d'un modèle satisfaisant que son critère de *Nash* soit supérieur à 0,8.

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. C'est l'objet de la validation. Afin de comparer les performances des modèles régressives à ceux du modèle neuronal, nous avons conduit une validation classique. Nous nous sommes servie du groupe d'observations « test » du modèle neuronal comme groupe de validation après avoir ajusté tous les modèles à l'aide du groupe de calibration (regroupant les groupes « apprentissage » et « validation » du réseau nueronal). Il va sans dire que dans certains cas où le jeu de données est très limité les résultats de cette validation ne sont présentés qu'à titre indicatif. Cependant, dans de nombreux cas, nous ne disposons pas toujours d'un jeu de données assez exhaustif pour pouvoir consacrer une partie significative des données pour des fins de validation. C'est pour cette raison que nous pouvons faire appel à la validation croisée. Au cours de cette dernière, la valeur d'une observation donnée de la variable réponse est temporairement retirée de l'ensemble des données. La valeur pour cette observation est alors estimée à l'aide du modèle calibré avec les observations restantes. Cette opération est reprise pour l'ensemble des observations, un à un. Ensuite, les valeurs estimées sont comparées aux vraies valeurs à l'aide des différents critères d'appréciation définis plus haut. Toutefois, cette procédure s'avère très lourde à appliquer dans le cas du modèle neuronal. Par conséquent, nous avons limité l'application de la validation croisée, dans le cas du modèle neuronal, à deux combinaisons de variables explicatives à savoir l'ensemble de variables et une sélection réduite de variable.

Enfin, nous avons conduit, pour chaque station retenue, des simulations du débit à l'aide des modèles ainsi calibrés. Les simulations ont été réalisées sur trois périodes hivernales distinctes choisies parmi la période d'observation de la station. Ainsi, les débits prédits par les différentes méthodes d'estimation ont été comparés aux débits correspondants estimés par les services d'Environnement Canada avec leur propre méthode.

3.2 INVENTAIRE DES DONNÉES

Pour mener à bien la présente étude, les services d'Environnement Canada ont mis à notre disposition les données de 6 stations météorologiques (Tableau 1) ainsi que les données de 8 stations hydrométrique situés sur la rivière Frazer en Colombie Britannique (Tableau 2). La Figure 6 présente la localisation des différentes stations sur le territoire de l'étude.

Tableau 1 : Liste des stations météorologiques

Numéro	Nom	Latitude en °	Longitude en °
1160899	Blue River A	52.13	-119.30
1096450	Prince George A	53.88	-122.67
1098940	Williams Lake A	52.18	-122.07
1163780	Kamloops A	50.70	-120.45
1100119	Agassiz CS	49.25	-121.76
109QJ3G	Tete Jaune	53.00	-119.53

Tableau 2 : Liste des stations hydrométriques

Numéro	Nom	Latitude en ° ' "	Longitude en ° ' "
08KA007	Fraser River at Red Pass	52 58 55 N	119 0 15 W
08KA005	Fraser River at McBride	53 17 10 N	120 6 46 W
08KA004	Fraser River at Hansard	54 4 43 N	121 50 52 W
08KB001	Fraser River at Shelley	54 0 40 N	122 37 0 W
08MC018	Fraser River at Marguerite	52 31 48 N	122 26 32 W
08MF040	Fraser River at Texas Creek	50 36 50 N	121 51 10 W
08LF022	Thompson River at Spences Bridge	50 21 25 N	121 23 38 W
08MF005	Fraser River at Hope	49 22 50 N	121 27 5 W

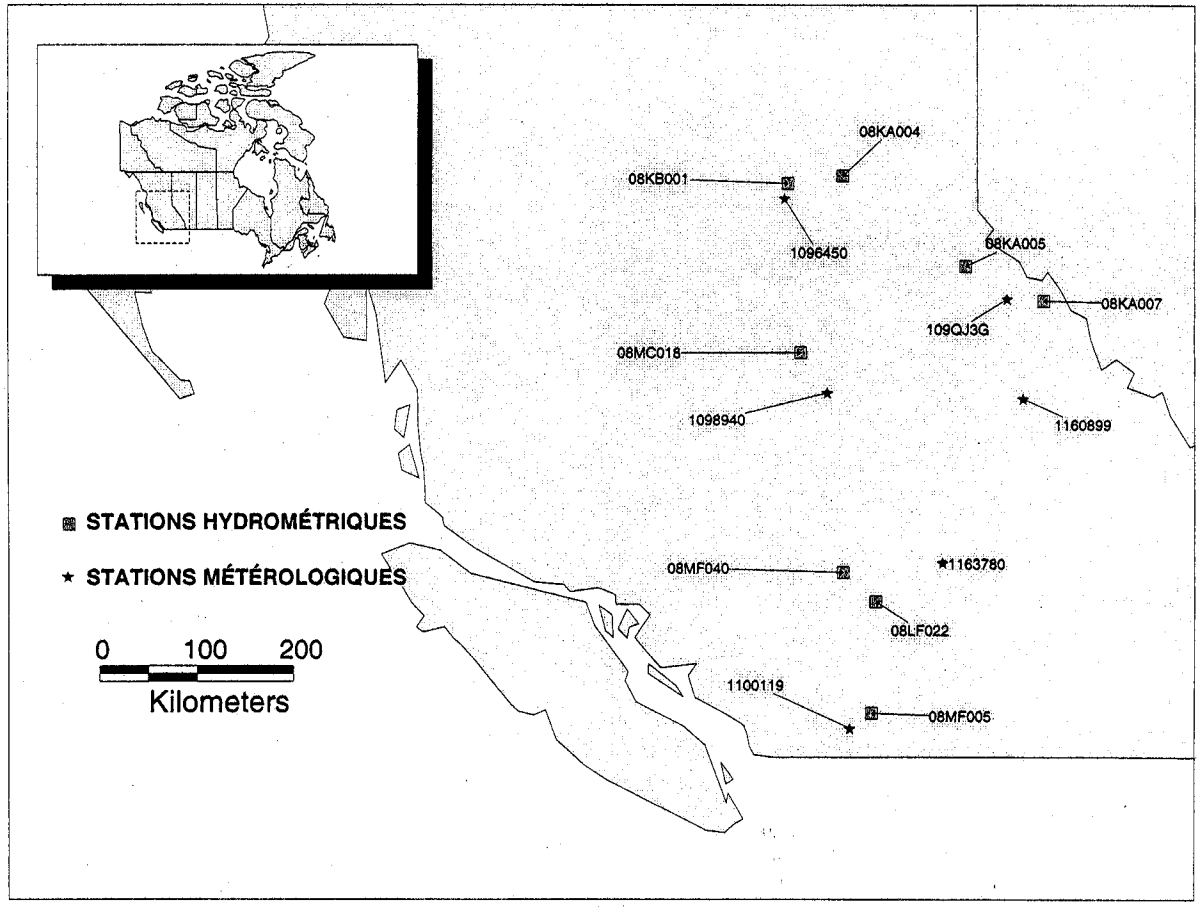


Figure 6 Localisation des stations météorologiques et des stations hydrométriques

Les fichiers de données météorologiques contiennent les éléments météorologiques journaliers suivants : 001 (température maximale), 002 (température minimale), 010 (précipitations liquides), 011 (précipitations solides), 012 (précipitations totales) et 013 (hauteur de la neige au sol). Nous avons obtenu aussi l'élément horaire 061 relatif à la radiation solaire globale et ce, pour la station Prince George (1096450) uniquement. Ces éléments météorologiques ont servi à calculer les variables météorologiques explicatives mentionnées ci-haut. Il faut signaler que les fichiers météorologiques présentaient de nombreuses plages de données manquantes. Un décompte de la disponibilité des données météorologiques et des plages de données manquantes est présenté à l'ANNEXE 1.

Les données hydrométriques, quant à elles, se répartissent en trois catégories : les données de jaugeage, les données sur les débits journaliers estimés par les services d'Environnement Canada et les mesures automatiques horaires du niveau d'eau dans la section de la rivière. Les données de jaugeage comprennent en plus de la date et heure du jaugeage les mesures de température de l'air, l'aire de la section mouillée, la largeur de celle-ci, la température de l'eau, la vitesse de l'eau et le débit. Les données de jaugeage mises à notre disposition commencent, pour la plus part des stations, dans les années 1950 pour se terminer en 2001. Le nombre de jaugeages hivernaux est variable (entre 0 et 4) d'une année à une autre et d'une station à une autre. En revanche, en plus de présenter, pour certaines stations, de nombreuses plages de données manquantes les données sur le débit estimé se terminent 1999 (ANNEXE 2). Il est à noter que l'information sur l'état de la surface nous a été fournie dans les fichiers de débits estimés. Par ailleurs, la période disponible de mesures du niveau automatiques est variable d'une station à une autre (Tableau 3).

Afin de construire les fichiers de calibration, contenant la variable réponse (débit jaugé) et les variables explicatives correspondantes, nous avons associé à chaque station hydrométrique la station météorologique la plus proche (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.

Vu la présence de données manquantes dans les différents types de données, la disparité dans l'occurrence et l'étendue la période d'observation ainsi que le faible nombre de

jaugeages hivernaux en présence de glace confirmée[1], le nombre de stations et la taille du jeu de données pouvant être utilisés s'est trouvé grandement limité. Ainsi, seules quatre stations hydrométriques sur huit ont été retenues, à savoir : les stations 08KA004, 08KB001, 08KA005 et 08MC018. De plus, le nombre d'observations par station est variable et ce, selon les variables explicatives pouvant être considérées. L'ANNEXE 3 présente la taille des échantillons pour chaque station selon les variables retenues.

Tableau 3 : Correspondance entre stations hydrométriques et stations météorologiques

HYDROMÉTRIQUE				MÉTÉOROLOGIQUE	
Numéro	PO*			Numéro	PO
	Débit jaugé	Débit estimé	Niveau automatique		
08KA004	1952-2001	1952-1999	1971-1987 1989-2001	1096450	1960-2001
08KA005	1953-2001	1953-1999	1987-2001	109QJ3G	1989-2001
08KA007	1955-2001	1955-1999	1986-2001	109QJ3G	1989-2001
08KB001	1950-2001	1950-1999	1970-2001	1096450	1960-2001
08LF022 (08LF051)**	1921-1951 (1988-1992)	1921-1951 (1988-1992)	2001	1163780	1960-2001
08MC018	1950-2001	1950-1999	1985-2001	1098940	1961-2001
08MF005	1912-2001	1912-1999	1969-1979 1980-2001	1100119	1988-2001
08MF040	1916-2001	1916-1999	1969-2000	1163780	1960-2001

* Période d'Observation
** La station 08LF051 a remplacé la 08LF022

[1] Pour certaines stations, notamment celles situées au sud du territoire, les jaugeages hivernaux en présence de glace confirmée sont rares voir inexistants.

3.3 RÉSULTATS

3.3.1 Calibration

Tout d'abord, nous avons procédé à l'examen des données de calibration afin d'en vérifier la qualité et détecter la présence éventuelle de données aberrantes et ce, à l'aide de graphiques et de statistiques descriptives. Ces statistiques sont présentées à l'ANNEXE 4.

Généralement, les jaugeages retenus ont eu lieu entre 60 et 80 jours en moyenne après la formation de la glace sous une température moyenne qui varie entre –6 et -2°C et un couvert neigeux variant entre 20 et 40 cm d'épaisseur selon la station. D'autre part, 4 observations ont été retirées des données de la station 08KA004 dont les valeurs de niveaux sont soupçonnées d'être aberrantes. S'agissant de niveaux jaugés, ces observations proviennent du même hiver et présentent des valeurs de niveau trois fois plus importantes que celles du reste des données. Par ailleurs, nous avons développé un algorithme neuronal pour la détection des observations aberrantes dans les valeurs du débit jaugé. Cet algorithme se base sur l'erreur d'estimation du débit par le modèle neuronal. Ainsi, une observation supplémentaire dans les données de la station 08KA004, deux observations dans celles de la station 08KB001 et trois dans celles de la station 08MC018 ont été identifiées comme aberrantes par l'algorithme et par conséquent retirées du jeu de données.

Dans le but d'explorer la nature des relations qui pourraient exister entre la variable réponse et les variables explicatives, nous avons produit les matrices de corrélation entre les différentes variables (ANNEXE 4). 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 08MC018, où le niveau présente une faible corrélation négative avec les données du débit. Ce qui nous amène à se questionner sur la qualité des données de cette station. Outre le niveau, ce sont les variables de température et l'épaisseur de la neige qui présentent les plus importants niveaux de corrélation avec les données du débit jaugé.

Par ailleurs, les services d'EC ont exprimé le souhait d'évaluer l'effet de la radiation solaire sur la correction du débit en présence de glace. Toutefois, cette variable n'est généralement

observée qu'au niveau des stations météorologiques principales du pays (stations synoptiques). Dans notre cas, nous ne disposons de mesures de radiation solaire qu'au niveau de la station Prince-Goerge (1096450). Malgré la faible corrélation de la radiation solaire avec le débit jaugé (ANNEXE 4) et du faible nombre d'observations disponibles, nous avons effectué des tests de calibration de la régression multiple et du réseau de neurones à l'aide d'une sélection de variables, identifiées par la technique stepwise, et la radiation solaire. Cet exercice a été conduit pour les deux stations hydrométriques (associées à la station météorologique 1096450) pour lesquelles des données de radiation sont disponibles. Les résultats sont présentés au Tableau 4.

Il s'est avéré que l'ajout de la radiation solaire n'améliore pas d'une manière significative la calibration de la régression multiple. De plus, il semble que le modèle neuronal est légèrement mieux calibré sans la radiation solaire qu'il l'est en son présence. Vu ces résultats peu concluants et le fait que tenir compte de la variable radiation solaire limite significativement la taille des échantillons disponibles, nous avons opté pour écarter la radiation solaire du reste de l'étude. Cependant, nous ne concluons pas que l'apport de cette variables est nécessairement négligeable pour la correction du débit sous l'effet de la glace. Il serait intéressant de refaire cet exercice avec un jeu de données plus substantiel afin d'être en mesure de se prononcer avec plus d'exactitude quant à la pertinence de cette variable.

Tableau 4 : Résultats de calibration du réseau de neurones et de la régression multiple en présence et en absence de la radiation solaire

Station	Modèle†	Taille	Réseau Neuronal				Régression Multiple			
			R^2	RMSEr (%)	BIAISr (%)	Nash	R^2	RMSEr (%)	BIAISr (%)	Nash
08KA004	1 2 10	26	0.82	23.1	10.6	0.79	0.71	22.0	-2.8	0.71
	1 2 10 + radiation		0.79	18.7	-6.1	0.74	0.74	20.1	3.0	0.74
08KB001	1 2 4 5 9 10	28	0.89	11.6	0.16	0.89	0.82	14.1	1.6	0.82
	1 2 4 5 9 10 + radiation		0.87	12.2	1.6	0.87	0.84	13.6	1.4	0.84

† : 1-Constante; 2-Niveau; 4-Température journalière moyenne; 5-Température décadaire moyenne; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige. Le modèle sans la radiation a été déterminé à l'aide la technique stepwise.

Le Tableau 5 présente les résultats de la calibration des modèles des différentes méthodes d'estimation à l'aide de toutes les variables explicatives disponibles. Le Tableau 6, quant à lui, présente les résultats obtenus à l'aide d'une sélection réduite de variables explicatives. Il va sans dire que dans ce cas la régression stepwise perd de son intérêt. Par conséquent, la technique stepwise n'a pas été appliquée à la sélection réduite de variables. La sélection de variables explicatives a été choisie selon le degré de corrélation de ces dernières avec le débit jaugé. Cette sélection varie d'une station à une autre. Toutefois, dans le cas de la station 08MC018, le niveau faisait partie de cette sélection même s'il affiche une faible corrélation avec le débit. Par ailleurs, la partie supérieure de chacun des deux tableaux présente les résultats obtenus en tenant compte du niveau, tandis que dans la partie inférieure, nous présentons les résultats obtenus sans la variable niveau.

Tableau 5 : Résultats de calibration des différentes méthodes d'estimation à l'aide du groupe de d'apprentissage du modèle neuronal et de toutes les variables explicatives

	Station	Taille	Régression Multiple				Régression Stepwise					Régression Ridge				Réseau Neuronal			
			R^2 ajusté	RMSEr (%)	BIAISr (%)	Nash	Modèle†	R^2 ajusté	RMSEr (%)	BIAISr (%)	Nash	R^2	RMSEr (%)	BIAISr (%)	Nash	R^2	RMSEr (%)	BIAISr (%)	Nash
Avec Niveau	08KA004	35	0.62	20.3	2.7	0.71	1 2 10	0.68	21.5	3.1	0.69	0.29	20.2	4.1	0.67	0.82	15.6	-0.5	0.82
	08KB001	59	0.54	19.1	3.1	0.60	1 2 5 10	0.57	19.6	3.2	0.58	0.21	19.8	4.6	0.55	0.56	19.0	3.1	0.56
	08KA005	17	0.60	9.7	0.9	0.80	1 2	0.66	11.7	1.4	0.66	0.28	10.2	1.2	0.78	0.88	7.7	1.7	0.87
	08MC018	22	0.58	9.5	0.8	0.74	1 7 4	0.48	14.1	1.7	0.50	0.22	10.3	1.3	0.68	0.79	8.5	0.9	0.79
Sans Niveau	08KA004	65	0.48	29.2	6.2	0.54	1 9 6	0.41	29.4	8.1	0.42	0.28	28.4	7.5	0.51	0.79	19.1	2.7	0.80
	08KB001	68	0.18	24.8	5.5	0.27	1 10 7 9	0.23	25.9	5.8	0.25	0.08	25.1	6.0	0.26	0.25	25.9	2.0	0.22
	08KA005	17	0.59	11.4	1.1	0.77	1 3 10	0.62	14.0	1.7	0.64	0.24	11.7	1.6	0.75	0.77	19.9	1.3	0.77
	08MC018	38	0.01	18.8	3.4	0.20	1 5	0.14	19.6	3.6	0.14	0.00	19.0	3.5	0.18	0.40	15.8	1.9	0.39

† : 1-Constante; 2-Niveau; 3-Nombre de jours depuis le gel; 4-Température journalière moyenne; 5-Température décadaire moyenne; 6-Degrés-jours depuis le gel; 7-Débit d'avant le gel; 8-Précipitations liquides journalières; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige.

Tableau 6 : Résultats de calibration des différentes méthodes d'estimation à l'aide du groupe de d'apprentissage du modèle neuronal et d'une sélection de variables explicatives

	Station	Taille	Modèle†	Régression Multiple				Régression Ridge				Réseau Neuronal			
				R^2 ajusté	RMSEr (%)	BIAISr (%)	Nash	R^2 ajusté	RMSEr (%)	BIAISr (%)	Nash	R^2	RMSEr (%)	BIAISr (%)	Nash
Avec Niveau	08KA004	35	1 2 4 5 6 9 10	0.66	20.8	2.9	0.72	0.37	20.5	4.2	0.68	0.89	14.6	-1.3	0.89
	08KB001	59	1 2 4 5 9 10	0.57	20.5	3.5	0.55	0.27	20.0	4.5	0.54	0.74	15.6	2.3	0.74
	08KA005	17	1 2 3 6 10	0.75	11.2	1.2	0.70	0.56	11.6	1.4	0.70	0.61	14.4	2.6	0.61
	08MC018	22	1 2 5 7	0.71	13.4	1.6	0.51	0.47	13.6	1.9	0.49	0.09	20.6	6.1	-0.27
Sans Niveau	08KA004	65	1 4 5 6 9 10	0.51	29.9	7.2	0.46	0.32	29.7	8.0	0.45	0.44	28.8	3.8	0.43
	08KB001	68	1 4 5 9 10	0.24	28.2	6.5	0.18	0.14	28.0	6.7	0.18	0.12	28.8	4.8	0.08
	08KA005	17	1 3 6 10	0.74	13.7	1.7	0.65	0.51	13.7	1.9	0.65	0.65	17.4	-8.7	0.11
	08MC018	38	1 5 7	0.18	19.5	3.5	0.15	0.11	19.5	3.7	0.15	0.04	25.0	6.6	-0.23

† : 1-Constante; 2-Niveau; 3-Nombre de jours depuis le gel; 4-Température journalière moyenne; 5-Température décadaire moyenne; 6-Degrés-jours depuis le gel; 7-Débit d'avant le gel; 8-Précipitations liquides journalières; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige.

De prime abord, il en ressort que le modèle neuronal est le modèle qui réussi le mieux à s'ajuster aux données de calibration et ce, selon tous les critères d'évaluation. Dans la plupart des cas, il réussit à expliquer près de 80% de la variance observée dans le débit jaugé. Ceci est vrai aussi bien pour le cas où nous avons utilisé toutes les variables explicatives que pour le cas de la sélection réduite de variables. Il est à remarquer ici que les résultats de la station 08MC018 ont été les plus médiocres, peu importe la technique d'estimation employée. Ce qui confirme nos soupçons quant à la qualité des données de cette station. D'autre part, les calibrations conduites sans le niveau ont été moins probantes, quoique, là aussi, le modèle neuronal s'est mieux comporté, en général. Ceci est attribuable à la place prépondérante qu'occupe le niveau dans l'explication de la variance du débit jaugé.

Dans le cas de la régression multiple et ridge, les calculs effectués à l'aide d'une sélection réduite de variables explicatives ont abouti à des performances légèrement supérieures à ceux effectués avec l'ensemble de variables explicatives et ce, si on se fie aux coefficients

de déterminations ajustés. Ce qui représente un résultat prévisible. En effet, l'utilisation d'une sélection de variables a pour but de réduire l'effet de la corrélation croisée qui pourrait existée entre les variables explicatives. En revanche, l'erreur sur l'estimation a été légèrement plus faible dans le cas de l'emploi de toutes les variables explicatives. Quant au modèle neuronal, il semble, lui aussi, mieux s'ajuster aux données en présence de la sélection réduite de variables explicatives, notamment dans les cas des deux stations 08KA004 et 08KB001 ayant plus d'observations. Ceci serait attribuable au fait qu'un nombre réduit de variables explicatives nécessite une architecture neuronale (nombre de connections) plus réduite augmentant ainsi les chances du réseaux de converger et s'adapter aux données d'entrées plus rapidement. En fait, l'ajout d'une variable supplémentaire aux p variables explicatives initiales, augmenterait le nombre de connections du réseau de neurones de $2(1+p)$ augmentant, par conséquent, de la même proportion le nombre de paramètres à optimiser par le système et ce, pour un nombre constant d'observations.

Si on s'intéresse aux résultats des modèles régressifs uniquement (Tableau 7, Tableau 8)[2], il s'en dégage que la régression stepwise produit des performances aussi bonnes que la régression multiple. Ce qui démontre qu'il est possible d'identifier d'une manière rigoureuse la sélection optimale de variables explicatives permettant de saisir l'essentiel de la variation de la variable à modéliser. Ainsi, pour toutes les stations, le niveau d'eau est ressorti comme la variable la plus significative puisqu'elle a été sélectionnée en premier lieu par l'algorithme de la stepwise, sauf pour la station 08MC018. Pour cette dernière c'est le débit d'avant le gel qui a occupé cette place. Dans deux cas, c'est la neige au sol qui est arrivée en deuxième lieu. Dans les deux autres cas c'est les variables de température qui sont ressorties : la température moyenne décadaire pour la station 08KB001 et la température moyenne journalière pour la station 08MC018. Il faut signaler que ces deux dernières stations sont situées, respectivement, plus en aval sur le cours d'eau par rapport

[2] Les modèles régressifs ont été calibrés à l'aide de toutes les observations disponibles.

aux deux autres. De plus, la dernière station est la plus méridionale du groupe, où probablement la température journalière aurait un effet sur le régime d'écoulement plus prononcé. En ce qui concerne la régression ridge, mis à part les plus faibles valeurs du coefficient de détermination, les résultats obtenus sont comparables à ceux de la régression multiple voir légèrement supérieurs à la régression stepwise. Néanmoins, à l'aide de la régression ridge, nous sommes assurés d'obtenir des paramètres plus précis. Il est à noter que R^2 baisse quant le constante k de la régression ridge croît. En effet, la somme des carrés des erreurs est minimum lorsque k est nul, mais elle augment proportionnellement à la constante k alors que la somme des carrés totale reste fixe. Ainsi, la baisse de la part de la variance expliquée en fonction de l'augmentation de k est compensée par la précision et la stabilité dans le calcul des coefficients de la régression ridge.

Tableau 7 : Résultats de calibration des trois modèles régressifs à l'aide de toutes les variables explicatives

	Station	Taille	Régression Multiple				Régression Stepwise					Régression Ridge				
			R^2 ajusté	RMSEr (%)	BIAISr (%)	Nash	Modèle†	R^2 ajusté	RMSEr (%)	BIAISr (%)	Nash	k‡	R^2 ajusté	RMSEr (%)	BIAISr (%)	Nash
Avec Niveau	08KA004	45	0.69	19.6	2.7	0.74	1 2 10	0.73	20.1	-2.7	0.73	0.35	0.34	19.9	4.3	0.69
	08KB001	77	0.59	18.7	3.0	0.63	1 2 5	0.55	20.5	-3.5	0.56	0.35	0.32	19.3	4.2	0.59
	08KA005	20	0.69	9.1	0.8	0.82	1 2 10	0.74	10.1	-1.1	0.76	0.21	0.41	9.6	1.2	0.80
	08MC018	26	0.49	12.0	1.3	0.65	1 7 4	0.46	15.6	-2.1	0.48	0.26	0.09	12.9	2.1	0.58
Sans Niveau	08KA004	84	0.49	28.2	5.8	0.54	1 9 5 3	0.50	28.7	-6.2	0.51	0.28	0.28	27.3	7.3	0.50
	08KB001	89	0.28	23.6	5.1	0.33	1 10 9 7	0.25	25.9	-4.8	0.27	0.34	0.13	24.3	5.8	0.31
	08KA005	21	0.63	10.8	1.0	0.76	1 3 10	0.61	12.8	-1.5	0.63	0.25	0.28	11.0	1.5	0.73
	08MC018	48	0.10	17.6	3.0	0.23	1 5	0.16	18	-3.2	0.16	0.27	0.00	17.9	3.2	0.21

† : 1-Constante; 2-Niveau; 3-Nombre de jours depuis le gel; 4-Température journalière moyenne; 5-Température décadaire moyenne; 6-Degrés-jours depuis le gel; 7-Débit d'avant le gel; 8-Précipitations liquides journalières; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige.

‡ : Constante de la régression ridge estimée en minimisant la fonction de l'*ISRM*

Tableau 8 : Résultats de calibration des modèles régressifs à l'aide d'une sélection de variables explicatives

	Station	Taille	Modèle†	Régression Multiple				Régression Ridge				
				R^2 ajusté	RMSEr (%)	BIAISr (%)	Nash	k‡	R^2 ajusté	RMSEr (%)	BIAISr (%)	Nash
Avec Niveau	08KA004	45	1 2 4 5 6 9 10	0.70	20.0	2.7	0.74	0.27	0.43	20.2	4.1	0.70
	08KB001	77	1 2 4 5 9 10	0.59	19.0	3.1	0.61	0.26	0.39	19.5	4.1	0.59
	08KA005	20	1 2 3 6 10	0.75	9.7	1.0	0.79	0.10	0.65	9.9	1.1	0.78
	08MC018	26	1 2 5 7	0.45	14.8	2.1	0.49	0.28	0.25	15.1	2.5	0.47
Sans Niveau	08KA004	84	1 4 5 6 9 10	0.44	29.4	6.9	0.47	0.22	0.30	28.9	7.8	0.45
	08KB001	89	1 4 5 9 10	0.20	27.4	6.3	0.22	0.20	0.12	27.4	6.5	0.22
	08KA005	21	1 3 6 10	0.61	12.6	1.4	0.65	0.14	0.49	12.8	1.7	0.64
	08MC018	48	1 5 7	0.16	18.3	3.2	0.17	0.18	0.10	18.4	3.3	0.17

† : 1-Constante; 2-Niveau; 3-Nombre de jours depuis le gel; 4-Température journalière moyenne; 5-Température décadaire moyenne; 6-Degrés-jours depuis le gel; 7-Débit d'avant le gel; 8-Précipitations liquides journalières; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige.
‡ : Constante de la régression ridge estimée en minimisant la fonction de l'*ISRM*

3.3.2 Validation

Il faut rappeler que la validation simultanée des modèles régressifs et du modèle neuronal (Tableau 9, Tableau 10) n'a pu être effectuée que sur les groupes de test du modèle neuronal. C'était le seul moyen pour pouvoir juger, sur les mêmes bases, de la qualité d'extrapolation des différents modèles. Cependant, les résultats des stations dont la taille du groupe de validation inférieure à 5 observations ne sont présentés ici qu'à titre indicatif.

Là encore, le modèle neuronal affiche les meilleurs résultats et ceci en comptant le niveau parmi les variables explicatives ou non, ou encore en considérant l'ensemble des variables explicatives ou juste une sélection réduite, alors que les trois modèles régressifs se valent entre eux. Si on s'attarde sur les résultats de l'estimation du débit sous la glace produits par le modèle neuronal au niveau des stations 08KA004 et 08KB001 (Tableau 9, Tableau 10) on remarque que ce dernier, contrairement aux deux modèles régressifs (régression multiple et ridge), réussi à produire des estimations de meilleure qualité à l'aide de toutes les variables explicatives. Ce qui signifie que la méthode d'estimation neuronale serait à

mieux de tirer profit de toute l'information disponible et semble peu sensible à la redondance d'information due à la multicollinéarité au niveau des variables explicatives. Par ailleurs, là aussi, se confirme la place prépondérante du niveau d'eau dans l'estimation du débit sous la glace. Peu importe la méthode, les modèles calibrés sans la variable niveau d'eau réussissent à modéliser à peine 50% de la variance observée au niveau du débit jaugé.

Tableau 9 : Validation des différentes méthodes d'estimation à l'aide du groupe test du modèle neuronal et de toutes les variables explicatives

	Station	Taille	Régression Multiple				Régression Stepwise					Régression Ridge				Réseau Neuronal			
			R^2	RMSEr (%)	BIAISr (%)	Nash	Modèle†	R^2	RMSEr (%)	BIAISr (%)	Nash	R^2	RMSEr (%)	BIAISr (%)	Nash	R^2	RMSEr (%)	BIAISr (%)	Nash
Avec Niveau	08KA004	10	0.81	22.8	-10.3	0.72	1 2 10	0.87	15.5	-3.7	0.86	0.73	21.8	-8.3	0.65	0.93	15.9	8.9	0.80
	08KB001	18	0.68	20.7	3.6	0.60	1 2 5 10	0.60	22.4	4.2	0.54	0.70	17.8	3.5	0.69	0.85	13.1	-0.5	0.84
	08KA005	3	0.17	11.2	4.9	-0.26	1 2	0.94	17.0	16.6	-2.3	0.60	8.5	5.5	0.27	1.00	23.5	14.3	-4.92
	08MC018	4	0.41	22.8	-11.8	0.10	1 7 4	0.77	20.1	-17.1	0.26	0.39	21.4	-12.7	0.02	0.99	20.3	-5.8	0.31
Sans Niveau	08KA004	19	0.03	29.3	11.5	-0.07	1 9 6	0.39	26.9	8.8	0.38	0.51	24.4	6.7	0.45	0.84	18.5	1.1	0.83
	08KB001	21	0.48	25.2	9.3	0.46	1 10 7 9	0.35	28.9	11.4	0.30	0.48	25.6	9.7	0.44	0.59	22.8	9.5	0.55
	08KA005	4	0.39	7.2	1.6	0.13	1 3 10	0.33	6.2	-0.3	0.29	0.35	7.2	3.4	0.14	0.99	9.7	-8.3	-0.32
	08MC018	10	0.38	11.8	0.4	0.36	1 5	0.33	12.2	0.67	0.32	0.45	12.7	1.6	0.36	0.87	6.8	9.7	0.83

† : 1-Constante; 2-Niveau; 3-Nombre de jours depuis le gel; 4-Température journalière moyenne; 5-Température décadaire moyenne; 6-Degrés-jours depuis le gel; 7-Débit d'avant le gel; 8-Précipitations liquides journalières; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige.

Tableau 10 : Validation des différentes méthodes d'estimation à l'aide du groupe test du modèle neuronal d'une sélection de variables explicatives

	Station	Taille	Modèle†	Régression Multiple				Régression Ridge				Réseau Neuronal			
				R^2	RMSEr (%)	BIAISr (%)	Nash	R^2	RMSEr (%)	BIAISr (%)	Nash	R^2	RMSEr (%)	BIAISr (%)	Nash
Avec Niveau	08KA004	10	1 2 4 5 6 9 10	0.83	14.4	-6.9	0.68	0.80	17.4	-1.7	0.63	0.92	19.4	-12.9	0.68
	08KB001	18	1 2 4 5 9 10	0.66	20.7	4.8	0.61	0.70	18.4	4.5	0.69	0.88	17.3	5.0	0.55
	08KA005	3	1 2 3 6 10	1.00	16.1	15.9	-1.93	0.92	17.3	16.8	-2.32	1.00	19.3	19.2	-3.34
	08MC018	4	1 2 5 7	0.60	18.4	-7.2	0.31	0.60	19.7	-6.7	0.24	0.94	12.5	-3.1	0.71
Sans Niveau	08KA004	19	1 4 5 6 9 10	0.52	24.7	5.9	0.48	0.51	24.4	6.7	0.45	0.69	21.2	-2.0	0.60
	08KB001	21	1 4 5 9 10	0.41	28.7	13.5	0.35	0.41	28.9	12.9	0.35	0.64	22.4	7.7	0.61
	08KA005	4	1 3 6 10	0.42	6.1	-2.6	0.29	0.45	5.8	-2.3	0.35	1.00	8.0	-8.0	-0.04
	08MC018	10	1 5 7	0.27	13.7	1.9	0.27	0.27	14.0	2.2	0.26	0.51	21.2	5.9	-0.32

† : 1-Constante; 2-Niveau; 3-Nombre de jours depuis le gel; 4-Température journalière moyenne; 5-Température décadaire moyenne; 6-Degrés-jours depuis le gel; 7-Débit d'avant le gel; 8-Précipitations liquides journalières; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige.

En se basant sur les résultats de la validation croisée (Tableau 11, Tableau 12) relatifs aux modèles régressifs pour lesquelles la taille des échantillons est plus significative, il s'avère que la régression stepwise a mieux réussi à expliquer la variation dans le débit jaugé. En effet, les résultats de la stepwise affichent, à quelques exceptions près, les valeurs d'erreur les plus faibles et les coefficients de détermination les plus élevés. Ceci plaide en faveur du fait qu'avec un choix judicieux des variables à inclure dans le modèle de régression, on est en mesure de bien modéliser la relation entre variable réponse et variables explicatives ce qui se traduit par une amélioration de la qualité de l'estimation. Ainsi, en considérant deux variables uniquement (le niveau d'eau et la neige au sol ou la température de l'air selon les stations) il est possible d'expliquer plus 60% de la variance totale.

Tableau 11 : Validation croisée des trois modèles régressifs calibrés à l'aide de toutes les variables explicatives

	Station	Taille	Régression Multiple				Stepwise					Régression Ridge			
			R^2	RMSEr (%)	BIAISr (%)	Nash	Modèle†	R^2	RMSEr (%)	BIAISr (%)	Nash	R^2	RMSEr (%)	BIAISr (%)	Nash
Avec Niveau	08KA004	45	0.46	31.4	-0.15	0.35	1 2 10	0.68	21.4	2.8	0.68	0.41	30.2	1.4	0.37
	08KB001	77	0.53	21.1	3.4	0.53	1 2 5	0.52	21.4	3.5	0.52	0.53	21.1	4.5	0.52
	08KA005	20	0.27	26.3	6.6	-0.32	1 2 10	0.60	13.0	1.8	0.52	0.38	20.1	4.6	0.25
	08MC018	26	0.06	34.3	6.0	-2.0	1 7 4	0.03	21.5	4.3	-0.03	0.18	18.1	2.8	0.10
Sans Niveau	08KA004	84	0.21	32.7	5.6	0.17	1 9 5 3	0.39	30.6	6.1	0.39	0.22	30.4	7.3	0.22
	08KB001	89	0.21	26.0	5.6	0.20	1 10 9 7	0.21	27.1	6.1	0.20	0.21	26.0	6.2	0.21
	08KA005	21	0.19	26.9	6.2	-0.25	1 3 10	0.51	14.8	1.4	0.50	0.31	19.9	4.4	0.25
	08MC018	48	0.01	28.9	7.2	-1.1	1 5	0.10	19.2	3.3	0.09	0.01	24.7	6.1	-0.52

† : 1-Constante; 2-Niveau; 3-Nombre de jours depuis le gel; 4-Température journalière moyenne; 5-Température décadaire moyenne; 6-Degrés-jours depuis le gel; 7-Débit d'avant le gel; 8-Précipitations liquides journalières; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige.

Tableau 12 : Validation croisée des modèles régressifs calibrés à l'aide d'une sélection de variables explicatives

	Station	Taille	Modèle†	Régression Multiple				Régression Ridge			
				R^2	RMSEr (%)	BIAISr (%)	Nash	R^2	RMSEr (%)	BIAISr (%)	Nash
Avec Niveau	08KA004	45	1 2 4 5 6 9 10	0.61	23.3	2.65	0.60	0.60	22.7	4.1	0.59
	08KB001	77	1 2 4 5 9 10	0.54	20.7	3.3	0.54	0.54	20.7	4.2	0.53
	08KA005	20	1 2 3 6 10	0.59	13.7	1.8	0.48	0.61	13.1	1.8	0.56
	08MC018	26	1 2 5 7	0.31	17.8	2.8	0.28	0.31	17.3	3.0	0.31
Sans Niveau	08KA004	84	1 4 5 6 9 10	0.33	31.9	7.1	0.32	0.33	30.8	8.0	0.33
	08KB001	89	1 4 5 9 10	0.15	28.7	6.6	0.15	0.16	28.5	6.7	0.15
	08KA005	21	1 3 6 10	0.50	15.0	1.4	0.49	0.51	14.8	1.6	0.51
	08MC018	48	1 5 7	0.07	19.9	3.8	0.05	0.07	19.8	3.9	0.06

† : 1-Constante; 2-Niveau; 3-Nombre de jours depuis le gel; 4-Température journalière moyenne; 5-Température décadaire moyenne; 6-Degrés-jours depuis le gel; 7-Débit d'avant le gel; 8-Précipitations liquides journalières; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige.

Les résultats de la validation croisée conduite sur le modèle neuronal (Tableau 13) désavouent les bonnes performances de ce modèle par rapport aux modèles régressifs, constatées plus haut. En effet, dans le cas où toutes les variables explicatives sont employées, tous les critères d'évaluation sont plus médiocres que ceux de la régression.

Toutefois, dans le cas des validations conduites à l'aide d'une sélection de variables, on observe une nette amélioration des performances du modèle neuronal. Ceci est attribuable au fait qu'avec un nombre plus petit de variables explicatives, le modèle neuronal serait plus stable spécialement dans un contexte où la taille de l'échantillon est problématique. Dans certains cas, les performances de ce dernier sont comparables à celles de la régression particulièrement au niveau de l'erreur quadratique. Cette dernière, en raison du biais d'estimation plus important du modèle neuronal (Tableau 11, Tableau 12, Tableau 13), est moins entachée de variance sur l'estimation que les pour les modèles régressifs.

L'origine de cette contre-performance est multiple. En plus la taille limitée de l'échantillon, la technique de la validation croisée est par nature à l'avantage des modèle paramétriques comme la régression où, à chaque itération, la forme du modèle évalué ne change pas. En revanche, dans le cas du modèle neuronal, à chaque étape de la validation croisée, on est en présence d'un modèle complètement différents des autres obtenus dans les autres itérations. À ceci s'ajout la faible attitude des modèles neuronaux par comparaison aux modèles régressifs à l'extrapolation. À cet égard, il faut rappeler que le groupe validation du modèle neuronal n'est pas complètement indépendant du processus de calibration. En effet, en raison du nombre d'observations limité, nous avons conduit plusieurs calibrations (une centaine pour chaque station et pur chaque combinaison de variables) et celle qui donne les meilleurs résultats pour le groupe de validation a été retenue. Cette procédure s'est avérée indispensable dans un contexte de rareté de données de calibration afin d'assurer la convergence du réseau neuronal vers une solution optimale.

Tableau 13 : Validation croisée du modèle neuronal

Station	Taille	Modèle†	Réseau Neuronal			
			R^2	RMSEr (%)	BIAISr (%)	Nash
08KA004	45	Tout	0.32	36.7	8.0	0.07
		2 4 5 6 9 10	0.59	25.4	8.3	0.56
08KB001	77	Tout	0.32	26.2	6.4	0.26
		2 4 5 9 10	0.44	22.3	2.5	0.36
08KA005	20	Tout	0.36	20.3	-0.2	0.24
		2 3 6 10	0.48	16.2	-0.2	0.41
08MC018	26	Tout	0.11	22.1	1.0	-0.14
		2 5 7	0.33	17.6	4.4	0.24

† : 1-Constante; 2-Niveau; 3-Nombre de jours depuis le gel; 4-Température journalière moyenne; 5-Température décadaire moyenne; 6-Degrés-jours depuis le gel; 7-Débit d'avant le gel; 8-Précipitations liquides journalières; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige.

Dans le but d'identifier les variables explicatives les plus significatives pour l'estimation du débits sous la glace, nous avons calibré le modèle neuronal en considérant différentes combinaisons de variables explicatives. Cet exercice a été réalisé à l'aide des données de la station 08KB001 ayant le nombre d'observations le plus important (77 observations en considérant toutes les variables explicatives). Pour chaque combinaison de variables, la calibration a été effectuée sur le groupe d'apprentissage du modèle neuronal (59 observations). Les modèles ainsi calibrés ont été validés sur le groupe « test » (18 observations). Le Tableau 14 en présente le résultats.

Tout d'abord, nous avons considéré toutes le variables explicatives disponibles. Ce modèle arrive à expliquer 85% de la variance dans le débit jaugé du groupe test avec une erreur quadratique moyenne relative de l'ordre de 14% et un *Nash* près de 0.80. Ensuite, nous avons procédé à retirer toutes les variables du modèle une à la fois. Il s'est avéré que le niveau d'eau est la variable indispensable pour l'estimation du débit sous la glace. Puisque, en son absence, les performances du modèle chute à une variance expliquée d'environ 0.17%, une erreur quadratique moyenne relative plus de 25% et un *Nash* négatif. Ceci est en accord avec les résultats de la régression stepwise pour cette station qui avait identifié cette celle-ci comme la première variable significative(Tableau 5). D'ailleurs, si on ne considère que la variable niveau d'eau uniquement, le modèle arrive à expliquer près de

60% de la variance. Par conséquent, nous avons retenu cette variable et nous avons examiné l'effet de lui ajouter une autre variable sur les performances du modèle. Il en ressort, en accord avec la régression stepwise, que la température décadaire moyenne en association avec le niveau d'eau constitue le modèle qui explique le mieux la variation du débit jaugé. En procédant comme précédemment, nous avons retenu ces dernières et nous avons testé les modèles augmentés d'une variables parmi les restantes, une à la fois. Il en résulte que les variables degrés-jours depuis le gel et la neige au sol semble être les variables qui contribuent le plus à améliorer les performances du modèles neuronal. Là aussi, ces résultats sont en accord avec ceux obtenus par la régression stepwise puisque cette dernière a identifié la neige au sol comme étant la troisième variable en terme d'importance pour cette station. Toutefois, le processus de sélection des variables les plus significative à l'aide du modèle neuronal a identifié une variable supplémentaire ignorée jusqu'à là par la stepwise à savoir les degrés-jours depuis le gel.

Le modèle neuronal bâti en considérant les variables ainsi identifiées : le niveau d'eau, la température décadaire moyenne, la neige au sol et les degrés-jours depuis le gel, réussi à produire, au niveau de la variance expliquée, des performances comparables à ceux obtenus par le modèle complet. En revanche, le modèle réduit arrive à des estimations plus précises et moins biaisées que celles produites par le modèle complet. Ceci renforce l'idée que nous avons déjà avancée à savoir que un nombre plus réduit de variables explicatives favoriserait la convergence de l'apprentissage du réseau de neurones et améliorait par conséquent la qualité de l'ajustement aux données et celle de l'estimation.

Tableau 14 : Résultats de validation du modèle neuronale à l'aide du groupe test (n=18) d e la station 08KB001 et de différentes combinaisons de variables explicatives[3]

Modèle	R^2	RMSEr (%)	BIAISr (%)	Nash
Tout	0.85	13.8	-6.7	0.79
Tout-2	0.17	25.5	-2.7	-0.03
Tout-3	0.71	16.6	-1.3	0.61
Tout-4	0.73	18.4	-0.6	0.68
Tout-5	0.78	14.8	-7.5	0.63
Tout-6	0.79	13.0	2.5	0.75
Tout-7	0.80	16.5	-2.9	0.60
Tout-8	0.82	15.1	0.11	0.76
Tout-9	0.77	13.2	-0.3	0.75
Tout-10	0.72	17.2	-10.1	0.54
2	0.58	17.2	-4.4	0.50
2+3	0.75	16.3	1.2	0.68
2+4	0.67	14.1	-2.9	0.63
2+5	0.80	12.3	-2.1	0.78
2+6	0.76	13.1	0.1	0.76
2+7	0.62	16.9	-5.0	0.54
2+8	0.59	17.9	-4.2	0.52
2+9	0.65	15.5	-3.1	0.61
2+10	0.78	14.7	-1.4	0.74
2+5+3	0.82	13.9	-0.7	0.73
2+5+4	0.80	11.6	-3.9	0.77
2+5+6	0.79	13.1	-2.0	0.77
2+5+7	0.78	13.5	1.6	0.76
2+5+8	0.78	12.2	-1.7	0.77
2+5+9	0.74	14.2	-5.8	0.66
2+5+10	0.80	12.7	-1.2	0.78
2+5+6+10	0.85	12.1	-1.3	0.80

† : 1-Constante; 2-Niveau; 3-Nombre de jours depuis le gel; 4-Température journalière moyenne; 5-Température décadaire moyenne; 6-Degrés-jours depuis le gel; 7-Débit d'avant le gel; 8-Précipitations liquides journalières; 9-Précipitations liquides décadaires et 10-Épaisseur de la neige.

[3] Les différentes combinaisons ont été calibrées avec le même groupe d'apprentissage (n=59)

3.3.3 Estimation

Une fois les quatre modèles calibrés et validés, nous avons conduit des simulations du débit à l'aide du niveau enregistré automatiquement dans stations et des variables météorologiques correspondantes. Pour chaque station, nous avons choisi trois périodes hivernales continues pour conduire les simulations. Ne sont présentés ici que les résultats de la station 08KA004 pour l'hiver 1971-1972. Le reste des résultats est présenté en annexe. Nous avons comparé par la suite les résultats de simulation avec les débits estimés et fournis par les services d'Environnement Canada (EC).

La Figure 7 présente le débit estimé pour l'hiver 1971-1972 à la station 08KA004. Nous y avons rapporté aussi les variables explicatives les plus significatives observées à la même période, présentant les plus fortes corrélations avec le débit jaugé. Par soucis d'en améliorer la lisibilité, toutes les variables ont été standardisées par rapport à leurs valeurs maximales et minimales.

Généralement, le débit estimé épouse le patron de variation du niveau de l'eau, sauf pour le mois de décembre 1971 et pour la fin du mois de mars et le mois d'avril 1972. Pour la première période, le débit évolue en sens inverse que l'évolution du niveau. Au moment où, nous assistons à une baisse du débit estimé, une baisse de la température et une croissance de l'épaisseur de la neige au sol, le niveau continue de croître. Cette incohérence nous amène à se questionner sur la qualité de ce dernier! Quant à la deuxième période, le pic du débit estimé est en retard par rapport à celui du niveau. Or ce dernier coïncide exactement avec un pic de précipitations liquides et réchauffement des températures et qui devrait se traduire en une augmentation proportionnelle du niveau observé.

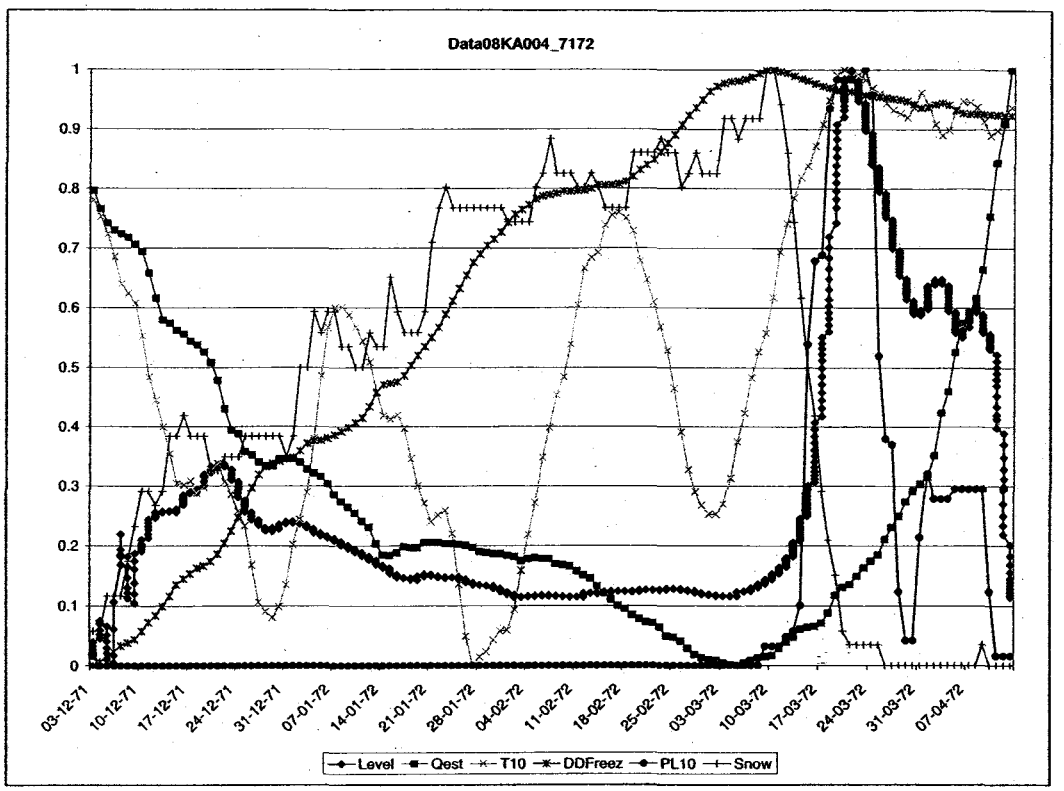


Figure 7 Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC (station 08KA004 pendant l'hiver 1971-1974)[4]

La Figure 8 montre l'évolution des débits calculés pour la station 08KA004 à l'aide des quatre modèles étudiés. Dans la même figure, nous avons rapporté les valeurs du débit estimé par EC correspondantes. Ainsi, pour la période allant de la fin de décembre 1971 au début de mars 1972, l'allure générale de l'évolution des débits estimés par nos modèles concordent , grosso modo, avec celle du débit estimé par EC. Cette période, correspondrait à la période pour laquelle la glace dans la rivière est complètement formée et stable. Donc, il s'agit d'une période pour laquelle les différents modèles étudiés ont été calibrés. Puisque, nous n'avons retenu dans leur calibration que les données observées en présence confirmée de glace. Les périodes de discordance, en début et en fin de période, correspondraient alors à la formation et puis la fonte de la glace, pour lesquelles nous ne nous attendons pas à de

[4] Level :Niveau; Qest :Débit estimé; T10 :Température décadaire moyenne; DDFreez :Degrés-jours depuis le gel; PL10 :Précipitations liquides décadaires; Snow : Neige au sol.

bonnes performances de la part de nos modèles. Cette tendance est confirmée par l'ensemble des stations pour lesquelles nous avons produit des estimations pendant la période de formation et/ou fonte des glaces.

Comme nous l'avons déjà mentionné, le modèle neuronal produit, par rapport aux modèles régressifs, les résultats se rapprochant le plus des valeurs jaugées. Par ailleurs, les débits estimés par les modèles étudiés, contrairement au débit d'EC qui est manifestement lissé, se font l'écho des fluctuations dans les variables explicatives météorologiques. Par exemple, un pic de chaleur en fin du mois de février 1972 et un autre en début du mois de mars jumelé avec un pic de précipitations liquides se sont traduits par des pics de débits dans l'estimation des modèles (Figure 7, Figure 8). Ce phénomène est plus accentué dans le cas des simulations sans la variable niveau d'eau (Figure 8B). Cette sensibilité aux fluctuations des variables explicatives météorologiques représente-t-elle un artefact de la modélisation, ou l'évolution réelle que devrait afficher le débit de la rivière? Le seul moyen pour le vérifier est de disposer d'une série indépendante de débits jaugés et la comparer aux résultats qu'auraient produit la méthode d'EC et nos quatre modèles.

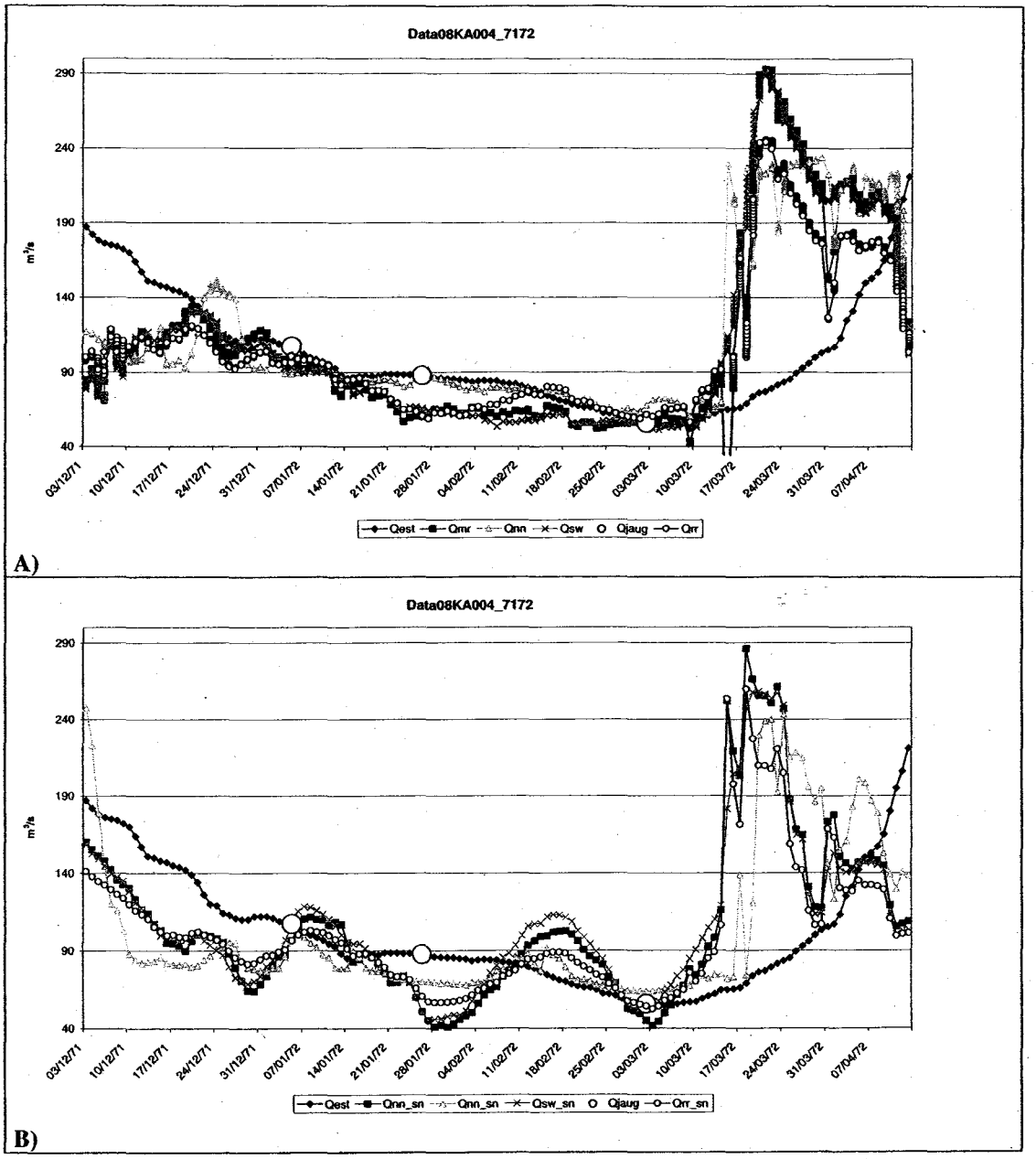


Figure 8 Débits simulés à l'aide des quatre modèles en fonction du débit EC pour la station 08KA004 pendant l'hiver 1971-1972 : A) simulations incluant le niveau; B) simulations sans le niveau[5]

[5] Qest : Débit estimé; Qmr : Débit simulé par la régression multiple; Qnn : Débit simulé par le modèle neuronal; Qsw : Débit simulé par la régression stepwise; Qjaug : Débit jaugé; Qrr : Débit simulé par la régression ridge.

4 CONCLUSIONS

L'objectif visé dans cette étude était d'explorer les performances de deux types d'approches pour la correction du débit en présence de glace, la première étant l'approche régressive (régression multiple, régression stepwise et régression ridge) et la seconde est basée sur les réseaux de neurones artificiels. Ces approches sont caractérisées par leur objectivité et leur reproductibilité par opposition aux méthodes présentement appliquées à EC qui sont plutôt caractérisées par leur subjectivité et non-reproductibilité. Les deux approches ont été calibrées à l'aide de différentes combinaisons de différents variables explicatives faisant intervenir l'ensemble de variables ou juste une sélection réduite et ce, en tenant compte non du niveau d'eau.

Les résultats de calibration, de validation et de simulation montre la supériorité de l'approche neuronale par rapport à l'approche régressive pour l'estimation du débit en présence de glace. Toutefois, certaines limitations en terme d'expérience de l'usager (choix de l'architecture optimale et de la technique d'apprentissage approprié du réseau de neurones), de disponibilité des données (données pour l'apprentissage et d'autres pour l'arrêt de celui-ci) et de temps de calcul viennent réduire quelque peu le potentiel d'application l'approche neuronale. Dans le cas où quelqu'un choisie de se rabattre sur l'approche régressive pour leur simplicité, transparence et rapidité, la régression stepwise représente une alternative intéressante. En plus de permettre d'effectuer un choix automatique mais rigoureux du modèle de régression, elle permet d'obtenir des résultats satisfaisants.

Par ailleurs, la disponibilité de données en quantité en qualité adéquates a représenté une limitation majeure pour la calibration et la validation des différentes techniques étudiées. Au départ, on disposait d'un jeu de données à première vue exhaustif (une cinquantaine d'années d'observations). Mais, les multiples plages de données manquantes et la disparité entre les périodes de disponibilité des données selon leur nature (données de jaugeage, mesures automatiques du niveau d'eau ou observations météorologiques) ont grandement

réduit le jeu de données disponibles pour la calibration (quelques 80 observations pour le meilleur cas et une vingtaine dans le cas le plus défavorable).

D'autre part, un de nos objectifs était de comparer les performances des méthodes étudiées dans le présent travail pour l'estimation du débit en présence de glace à ceux obtenus par la méthode employée par EC. Cependant, nous nous sommes contentés que d'une comparaison visuelle dans la mesure où nous ne disposons pas d'une série de jaugeage indépendante qui aurait pu être utilisée pour valider les performances réelles et de la technique conventionnelle utilisée par EC et celles étudiées ici.

Par ailleurs, les différentes méthodes d'estimation ont été calibrées avec des débits jaugés qui correspondent à la phase de présence stable de la glace en rivière. Ce qui explique en partie les divergences observées en début et en fin de période hivernale avec les débits produits par les services EC. Il serait donc intéressant de construire des modèles utilisables dans toutes les phases de présence de la glace (formation, phase stable et cassure). Une façon pour répondre à ce besoin serait d'identifier les jaugeages effectués pendant les phases de formation et cassure de la glace et pour les inclure dans la calibration. Cependant, cette information n'est pour le moment pas disponible.

Nous avons parvenu, ainsi, à développer un outil informatique de correction à temps réel des débits en présence de 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 recalibrer 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 ici, 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

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

Hoerl, A.E., Kennard, R.W. (1970a). Ridge regression : biased estimation for nonorthogonal problems. Technometrics, 12, pp. 55-67.

Hoerl, A.E., Kennard, R.W. (1970b). Ridge regression : biased application to nonorthogonal problems. Technometrics, 12, pp. 69-82.

Mc Culloch, W.S., Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Mathem. Biophys.*, **5**, 115-133.

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 de recherche No. R-559, INRS-Eau, Québec, 75 pages.

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

The MathWorks, Inc. (2000). *Using Matlab, Version 6.*

Weisberg, S. (1985). *Applied linear regression (Second Edition).* Wiley & Sons, New York, 324 pages.

Vinod, H.D., 1976, Application of new ridge regression methods to a study of Bell system scale economies, Journal of the American Statistical Association, vol. 71, pp. 835-841.

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

ANNEXE 1: DISPONIBILITÉ DES DONNÉES MÉTÉOROLOGIQUES ET PLAGES DES DONNÉES MANQUANTES

STATIONS		1096450	1098940	109QJ3G	1100119	1160899	1163780
PÉRIODE D'OBSETRVATION		01/1960-12/2001	01/1961-12/2001	11/1989-04/2001	11/1988-04/2001	01/1970-04/2001	01/60-12/2001
DONNÉES MAQUNATES	**001** Tmax			23,25,27/02/1997 01-12/02/1999 06/2000	**01-05/1989** **02/1990** 07-08/1992 **01/1997**	**04/1979**	
	002 Tmin			23,25,27/02/1997 07-08/1998 01-12/02/1999 06/2000	**01-05/1989** **02/1990** 07-08/1992 **01/1997**	**04/1979**	
	010 Précipitations liquides			23,25,27/02/1997 01-12/02/1999 06/2000	**02-05/1989** **02/1990** **01-02;07-12/1991** **01/1992-06/1997** 08/1997-**12/1998** **02-06/1999** 08/1999 **11/1999-04/2001**	**04/1979**	
	011 Précipitations solides			23,25,27/02/1997 **11/1998** 01-12/02/1999 06/2000	**02-05/1989** **02/1990** **01/1991-06/1997** 08/1997-**12/1998** **02-06/1999** 08/1999 **11/1999-04/2001**	**04/1979**	
	012 Précipitations totales			23,25,27/02/1997 **11/1998** 01-12/02/1999 06/2000 **04/2001**	**02-05/1989** **02/1990** 07-08/1992 **11/1992** **10-12/1994-01/1995** 10/1996 **01/1997** **04/2001**	**04/1979** **04/2001**	
	013 Neige au sol			**06/2000**	ND*	04/1979-071980 09/1980-10/1980 02/1984	
DISPONIBILITÉ	**061** (horaire) Radiation globale	01/10/1973- 30/01/1995	ND*	ND	ND	ND	ND

- Non disponible

ANNEXE 2: DISPONIBILITÉ DES DONNÉES SUR LE DÉBIT ESTIMÉ ET PLAGES DES DONNÉES MANQUANTES

STATIONS	08KA004	08KA005	08KA007	08KB001	08MC018	08MF005	08MF040
PÉRIODE D'OBSETRVATION	01/1952-12/1999	01/1953-12/1999	01/1955-12/1999	01/1950-12/1999	01/1950-12/1999	01/1912-12/1999	01/1951-12/1999
DONNÉES MAQUNATES	01/01/1952-30/09/152 01/12/1952-28/02/1953	01/01/1953-28/04/1953 01/01/1954-20/04/1954 01/01/1955-30/04/1955 01/02/1957-27/04/1957 01/12/1957-27/03/1958	01/01/1955-31/05/1955	01/01/1950-30/04/1950	01//01/1950-18/04/1950 01/12/1956-31/01/1957 01/12/1958-31/03/1959 01/01/1960-31/03/1960 01/12/1961-31/03/1962 11/12/1963-05/03/1964	01/01/1912-29/02/1912	01/01/1951-11/08/1951

ANNEXE 3: TAILLE DES ÉCHANTILLONS EN FONCTION DES VARIABLES EXPLICATIVES RETENUES

08KA004-1096450

TAILLE DE L'ÉCHANTILLON	VARIABLES EXPLICATIVES									
	Niveau	Tmoy	Tmoy10	Jour depuis Gel	DJ depuis Gel	Débit avant Gel	Précipitation liquide	Précipitation liquide 10 jours	Neige au Sol	Radiation
89										
48										
50										
34										

08MC018-1098940

TAILLE DE L'ÉCHANTILLON	VARIABLES EXPLICATIVES								
	Niveau	Tmoy	Tmoy10	Jour depuis Gel	DJ depuis Gel	Débit avant Gel	Précipitation liquide	Précipitation liquide 10 jours	Neige au Sol
51									
29									

08KB001-1096450

TAILLE DE L'ÉCHANTILLON	VARIABLES EXPLICATIVES									
	Niveau	Tmoy	Tmoy10	Jour depuis Gel	DJ depuis Gel	Débit avant Gel	Précipitation liquide	Précipitation liquide 10 jours	Neige au Sol	Radiation
91										
36										
79										
34										

08KA005_109QJ3G

TAILLE DE L'ÉCHANTILLON	VARIABLES EXPLICATIVES								
	Niveau	Tmoy	Tmoy10	Jour depuis Gel	DJ depuis Gel	Débit avant Gel	Précipitation liquide	Précipitation liquide 10 jours	Neige au Sol
21									
22									
20									
21									

ANNEXE 4: STATISTIQUES DESCRIPTIVES ET MATRICES DE CORRÉLATION

08KA004-1096450

DESCRITIVE STATISTICS (n=45)

	Mean	Median	Max	Min	Std-Dev	Skewness	Kurtosis
Q	109.81	98.00	331.00	56.00	47.91	2.22	8.79
Area	375.95	368.79	690.00	147.70	105.96	0.46	3.47
Width	265.38	225.77	2050.00	20.12	204.82	7.97	70.33
Level	2.55	2.52	3.25	2.12	0.27	0.71	3.13
DayFreez	84.05	84.50	148.00	6.00	34.69	-0.24	2.25
T	-4.97	-2.65	8.65	-29.20	7.93	-1.19	4.00
T10	-5.84	-4.38	3.59	-24.00	6.19	-0.70	2.90
DDFreez	-660.41	-619.92	-74.45	-1417.15	317.77	-0.21	2.32
Qfreez	227.45	191.00	476.00	105.00	94.29	1.04	3.07
LP	0.12	0.00	4.60	0.00	0.59	6.07	43.41
LP10	2.77	0.30	37.10	0.00	5.54	3.65	19.97
Snow	20.35	15.00	76.00	0.00	19.75	1.06	3.35

CORRELATION MATRIX

	Q	Area	Width	Level	DayFreez	T	T10	DDFreez	Qfreez	LP	LP10	Snow	Radiation
Q	1.00	0.71	-0.07	0.76	-0.08	0.29	0.42	0.30	-0.02	-0.10	0.25	-0.43	-0.20
Area	0.71	1.00	-0.01	0.73	-0.20	0.23	0.37	0.34	0.10	-0.27	0.24	-0.31	-0.38
Width	-0.07	-0.01	1.00	0.03	0.25	0.03	0.17	-0.13	0.21	0.00	-0.06	-0.14	0.27
Level	0.76	0.73	0.03	1.00	-0.17	0.18	0.29	0.23	0.10	-0.10	0.26	-0.05	-0.22
DayFreez	-0.08	-0.20	0.25	-0.17	1.00	0.49	0.57	-0.61	0.08	0.13	0.13	-0.16	0.77
T	0.29	0.23	0.03	0.18	0.49	1.00	0.69	-0.11	-0.02	0.13	0.29	-0.27	0.35
T10	0.42	0.37	0.17	0.29	0.57	0.69	1.00	0.04	-0.17	0.13	0.20	-0.56	0.42
DDFreez	0.30	0.34	-0.13	0.23	-0.61	-0.11	0.04	1.00	-0.22	-0.09	0.08	-0.34	-0.52
Qfreez	-0.02	0.10	0.21	0.10	0.08	-0.02	-0.17	-0.22	1.00	-0.14	0.21	0.19	0.02
LP	-0.10	-0.27	0.00	-0.10	0.13	0.13	0.13	-0.09	-0.14	1.00	0.06	-0.07	-0.02
LP10	0.25	0.24	-0.06	0.26	0.13	0.29	0.20	0.08	0.21	0.06	1.00	-0.18	012
Snow	-0.43	-0.31	-0.14	-0.05	-0.16	-0.27	-0.56	-0.34	0.19	-0.07	-0.18	1.00	-0.25
Radiation(n=33)	-0.20	-0.38	0.27	-0.22	0.77	0.35	0.42	-0.52	0.02	-0.02	0.12	-0.25	1.00

08MC018-1098940

DESCRITIVE STATISTICS (n=26)

	Mean	Median	Max	Min	Std-Dev	Skewness	Kurtosis
Q	389.00	378.00	552.00	258.00	78.27	0.42	2.42
Area	451.00	441.00	752.51	305.65	98.61	1.13	4.43
Width	179.29	182.88	208.79	118.87	15.92	-1.44	6.49
Level	2.25	2.00	4.85	0.83	0.91	1.64	5.48
DayFreez	63.50	71.50	143.00	4.00	33.75	0.21	2.32
T	-4.28	-2.70	6.10	-24.50	6.87	-0.98	3.51
T10	-5.92	-4.25	2.10	-25.00	6.34	-0.90	3.21
DDFreez	-534.29	-515.05	-4.30	-1251.40	316.31	-0.51	2.54
Qfreez	618.58	592.00	1800.00	292.00	291.66	2.73	11.91
LP	0.16	0.00	5.60	0.00	0.82	6.29	42.11
LP10	1.02	0.00	22.10	0.00	3.44	5.13	31.02
Snow	33.13	30.50	89.00	0.00	24.74	0.49	2.44

CORRELATION MATRIX

	Q	Area	Width	Level	DayFreez	T	T10	DDFreez	Qfreez	LP	LP10	Snow
Q	1.00	0.31	0.52	-0.16	0.25	0.32	0.45	-0.25	0.57	0.01	0.17	0.20
Area	0.31	1.00	0.60	0.41	-0.24	-0.17	-0.10	0.23	0.10	0.06	0.08	0.02
Width	0.52	0.60	1.00	0.21	0.10	-0.36	-0.08	-0.19	0.51	0.03	0.05	0.38
Level	-0.16	0.41	0.21	1.00	0.09	-0.33	-0.05	-0.25	0.17	-0.17	-0.15	0.10
DayFreez	0.25	-0.24	0.10	0.09	1.00	0.32	0.55	-0.80	0.43	-0.27	-0.14	-0.15
T	0.32	-0.17	-0.36	-0.33	0.32	1.00	0.48	-0.06	-0.12	0.04	0.11	-0.46
T10	0.45	-0.10	-0.08	-0.05	0.55	0.48	1.00	-0.28	0.22	-0.05	0.13	-0.28
DDFreez	-0.25	0.23	-0.19	-0.25	-0.80	-0.06	-0.28	1.00	-0.67	0.24	0.38	-0.27
Qfreez	0.57	0.10	0.51	0.17	0.43	-0.12	0.22	-0.67	1.00	-0.10	0.02	0.41
LP	0.01	0.06	0.03	-0.17	-0.27	0.04	-0.05	0.24	-0.10	1.00	0.63	-0.01
LP10	0.17	0.08	0.05	-0.15	-0.14	0.11	0.13	0.38	0.02	0.63	1.00	-0.17
Snow	0.20	0.02	0.38	0.10	-0.15	-0.46	-0.28	-0.27	0.41	-0.01	-0.17	1.00

08KB001-1096450

DESCRITIVE STATISTICS (n=77)

	Mean	Median	Max	Min	Std-Dev	Skewness	Kurtosis
Q	195.74	183.00	425.00	104.00	59.05	1.23	4.78
Area	334.90	318.00	604.74	168.00	93.78	0.63	3.14
Width	194.56	204.21	274.31	1.81	37.27	-2.81	14.39
Level	2.69	2.60	3.90	1.80	0.47	0.34	2.58
DayFreez	79.06	78.00	147.00	11.00	32.95	-0.11	2.32
T	-6.00	-4.55	6.65	-35.00	7.95	-1.03	4.26
T10	-6.16	-5.20	4.10	-25.30	6.28	-0.62	2.87
DDFreez	-682.84	-648.70	-42.00	-1416.50	315.01	-0.26	2.32
Qfreez	382.85	334.00	729.00	198.00	132.27	1.00	3.08
LP	0.27	0.00	6.40	0.00	1.04	4.55	24.09
LP10	2.10	0.30	24.80	0.00	3.89	3.12	15.55
Snow	25.30	20.00	98.00	0.00	20.63	1.14	4.23

CORRELATION MATRIX

	Q	Area	Width	Level	DayFreez	T	T10	DDFreez	Qfreez	LP	LP10	Snow	Radiation
Q	1.00	0.63	-0.09	0.68	0.02	0.29	0.34	0.08	0.16	0.02	0.34	-0.33	0.09
Area	0.63	1.00	0.07	0.75	-0.25	0.06	0.01	0.17	0.32	0.14	0.18	-0.02	-0.02
Width	-0.09	0.07	1.00	0.09	-0.13	0.00	-0.09	0.17	0.10	0.06	0.10	0.09	-0.17
Level	0.68	0.75	0.09	1.00	-0.28	0.02	0.04	0.33	0.26	0.05	0.16	-0.12	-0.13
DayFreez	0.02	-0.25	-0.13	-0.28	1.00	0.45	0.55	-0.66	0.02	0.17	0.21	-0.19	0.74
T	0.29	0.06	0.00	0.02	0.45	1.00	0.57	-0.26	0.07	0.29	0.33	-0.25	0.21
T10	0.34	0.01	-0.09	0.04	0.55	0.57	1.00	-0.29	0.01	0.16	0.32	-0.38	0.55
DDFreez	0.08	0.17	0.17	0.33	-0.66	-0.26	-0.29	1.00	0.02	-0.04	-0.18	-0.20	-0.54
Qfreez	0.16	0.32	0.10	0.26	0.02	0.07	0.01	0.02	1.00	0.28	-0.06	0.18	-0.20
LP	0.02	0.14	0.06	0.05	0.17	0.29	0.16	-0.04	0.28	1.00	0.27	0.07	0.03
LP10	0.34	0.18	0.10	0.16	0.21	0.33	0.32	-0.18	-0.06	0.27	1.00	-0.17	0.19
Snow	-0.83	-0.02	0.09	-0.12	-0.19	-0.25	-0.38	-0.20	0.18	0.07	-0.17	1.00	-0.16
Radiation(n=34)	0.09	-0.02	-0.17	-0.13	0.74	0.21	0.55	-0.54	-0.20	0.03	0.19	-0.16	1.00

08KA005_109QJ3G

DESCRITIVE STATISTICS (n=20)

	Mean	Median	Max	Min	Std-Dev	Skewness	Kurtosis
Q	34.27	33.00	46.00	25.00	6.86	0.41	2.04
Area	115.67	108.22	161.00	80.60	23.62	0.39	1.88
Width	100.61	98.00	115.00	82.00	7.75	-0.14	2.83
Level	3.53	3.48	4.17	3.25	0.21	1.45	5.13
DayFreez	76.36	83.50	138.00	24.00	34.57	0.07	2.00
T	-2.12	-0.55	5.80	-22.50	6.62	-1.56	5.37
T10	-6.24	-4.05	1.90	-22.70	5.85	-1.06	3.90
DDFreez	-480.89	-338.50	-109.30	-1006.50	309.68	-0.51	1.72
Qfreez	77.07	74.80	127.00	47.60	23.49	0.65	2.71
LP	1.25	0.00	16.20	0.00	3.73	3.27	13.08
LP10	3.75	0.00	21.40	0.00	6.36	1.56	4.20
Snow	39.95	38.00	79.00	10.00	18.99	0.31	2.53

CORRELATION MATRIX

	Q	Area	Width	Level	DayFreez	T	T10	DDFreez	Qfreez	LP	LP10	Snow
Q	1.00	0.89	0.24	0.82	-0.68	-0.17	-0.19	0.56	0.06	-0.26	-0.06	-0.53
Area	0.89	1.00	0.18	0.86	-0.80	-0.16	-0.18	0.80	-0.06	-0.37	-0.20	-0.28
Width	0.24	0.18	1.00	0.32	-0.16	-0.08	-0.03	-0.15	0.18	-0.05	-0.15	-0.37
Level	0.82	0.86	0.32	1.00	-0.71	-0.26	-0.27	0.59	0.11	-0.33	-0.25	-0.31
DayFreez	-0.68	-0.80	-0.16	-0.71	1.00	0.34	0.27	-0.81	0.34	0.37	0.16	0.18
T	-0.17	-0.16	-0.08	-0.26	0.34	1.00	0.61	-0.21	0.18	0.26	0.43	-0.17
T10	-0.19	-0.18	-0.03	-0.27	0.27	0.61	1.00	-0.11	0.24	0.08	0.26	-0.24
DDFreez	0.56	0.80	-0.15	0.59	-0.81	-0.21	-0.11	1.00	-0.24	-0.26	-0.18	-0.02
Qfreez	0.06	-0.06	0.18	0.11	0.34	0.18	0.24	-0.24	1.00	-0.09	-0.25	-0.18
LP	-0.26	-0.37	-0.05	-0.33	0.37	0.26	0.08	-0.26	-0.09	1.00	0.70	-0.31
LP10	-0.06	-0.20	-0.15	-0.25	0.16	0.43	0.26	-0.18	-0.25	0.70	1.00	-0.48
Snow	0.53	-0.28	-0.37	-0.31	0.18	-0.17	-0.24	-0.02	-0.18	-0.31	-0.48	1.00

ANNEXE 5: RÉSULTATS D'ESTIMATION

08KA004

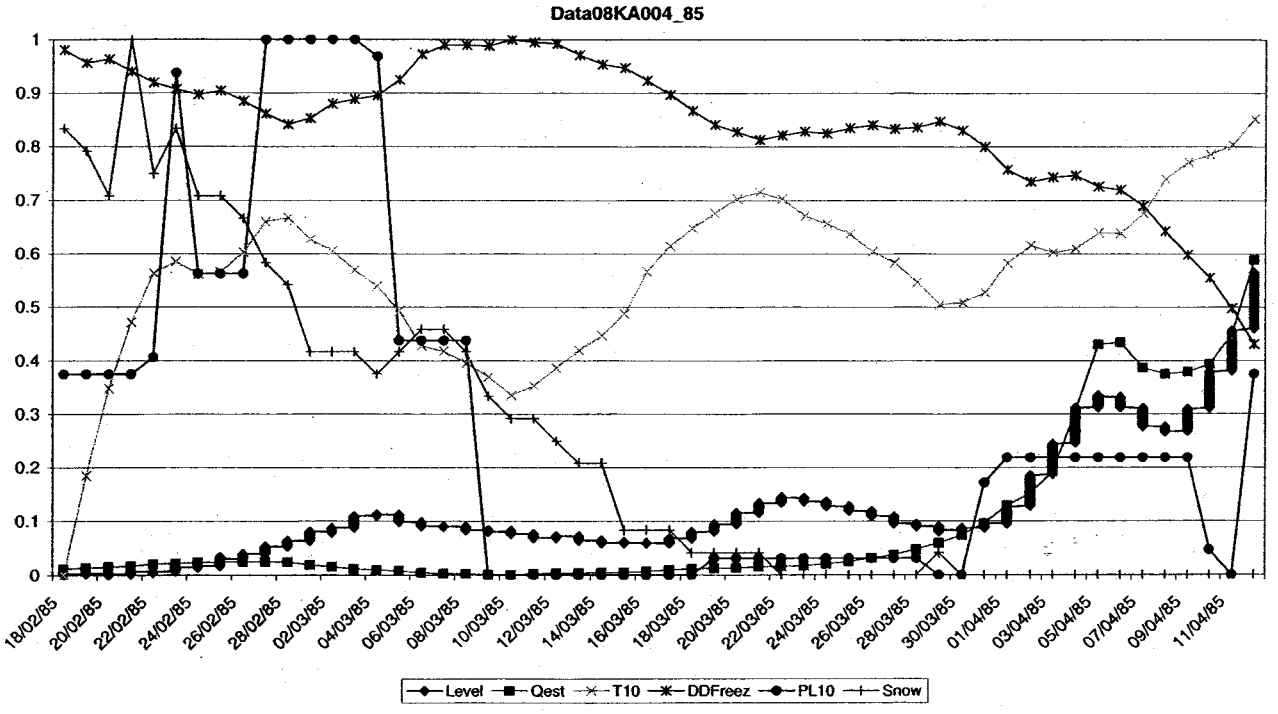


Fig. A.5. 1 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08KA004 pendant l'hiver 1985.

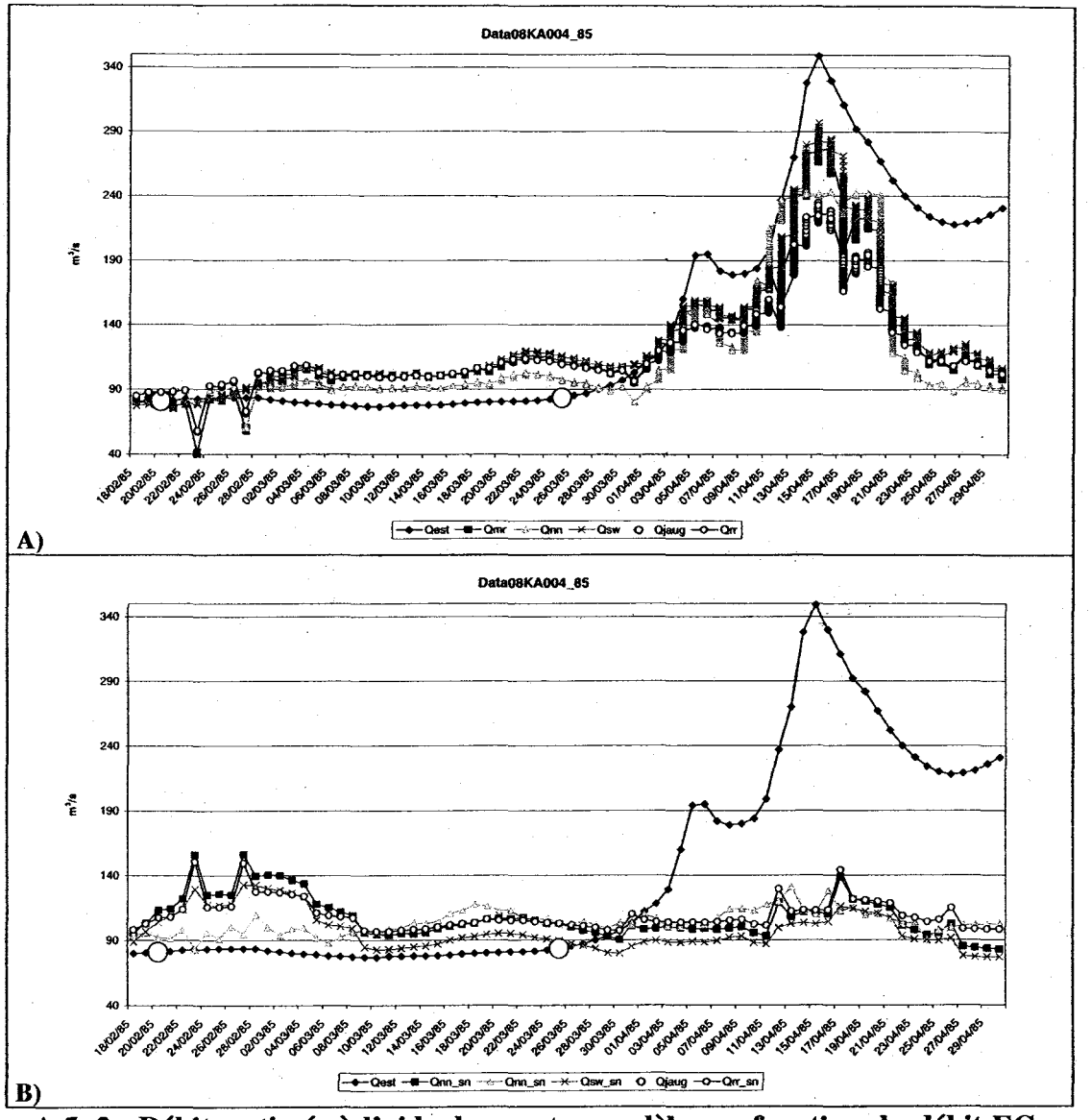


Fig. A.5. 2 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08KA004 pendant l'hiver 1985 : A) simulations incluant le niveau; B) simulations sans le niveau[6]

[6] Qest : Débit estimé; Qmr : Débit simulé par la régression multiple; Qnn : Débit simulé par le modèle neuronal; Qsw : Débit simulé par la régression stepwise; Qjaug : Débit jaugé; Qrr : Débit simulé par la régression ridge.

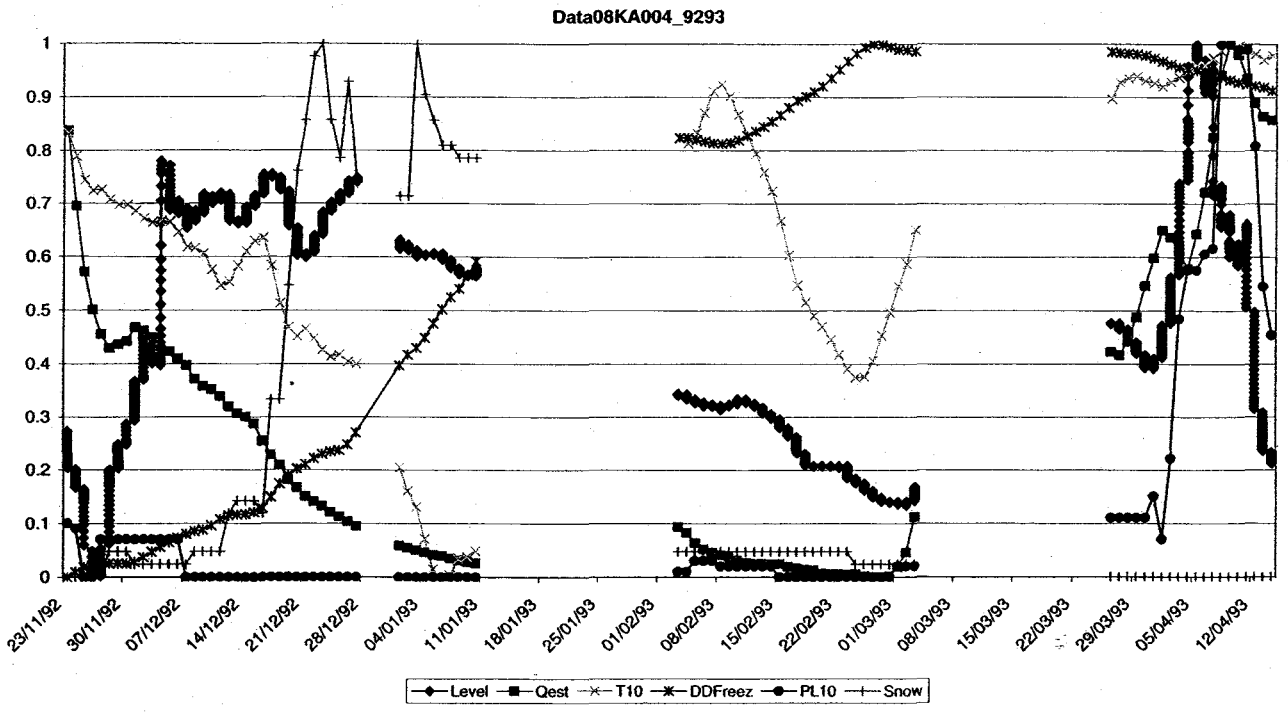


Fig. A.5. 3 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08KA004 pendant l'hiver 1992-1993.

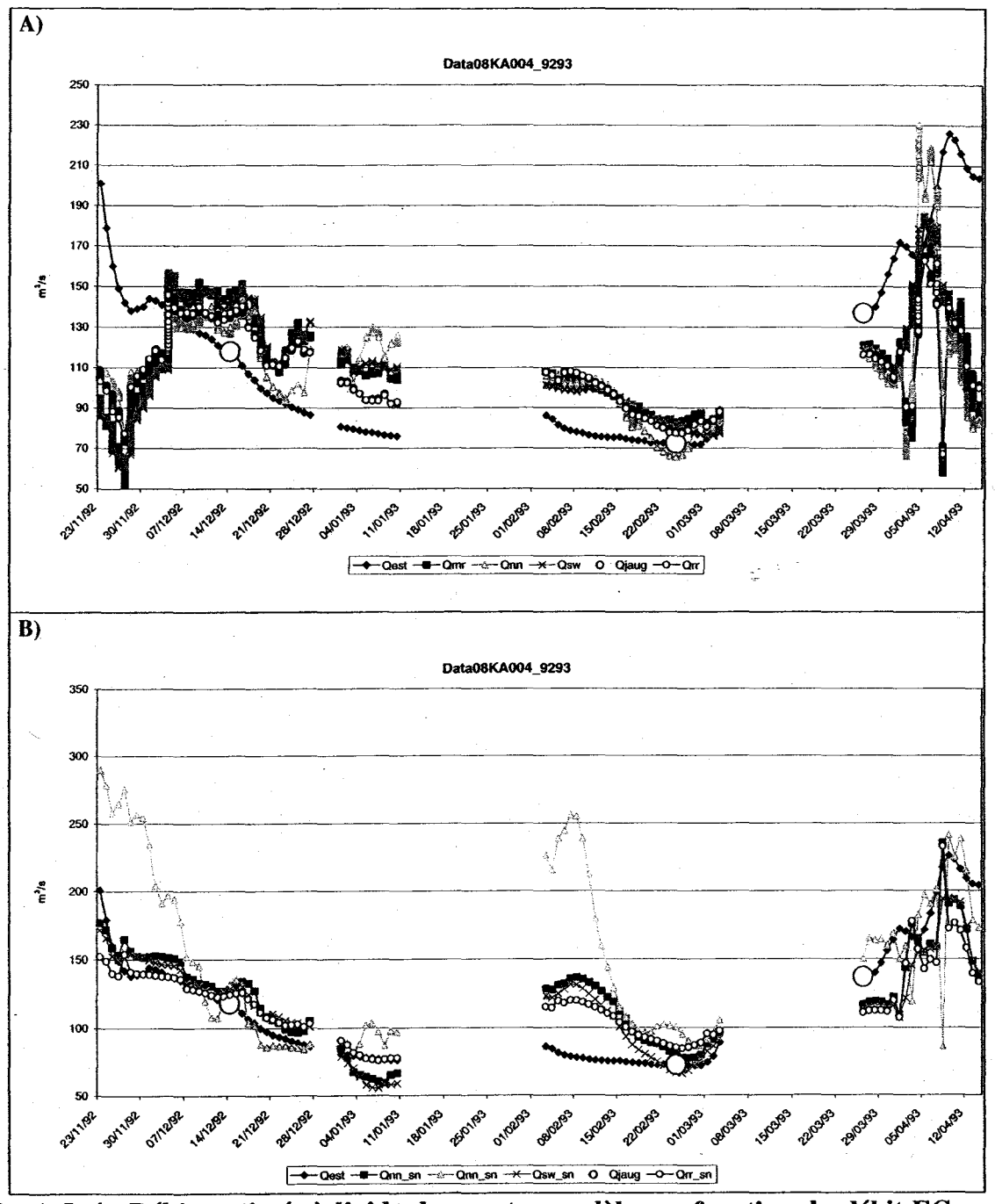


Fig. A.5. 4 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08KA004 pendant l'hiver 1992-1993 : A) simulations incluant le niveau; B) simulations sans le niveau

08KB001

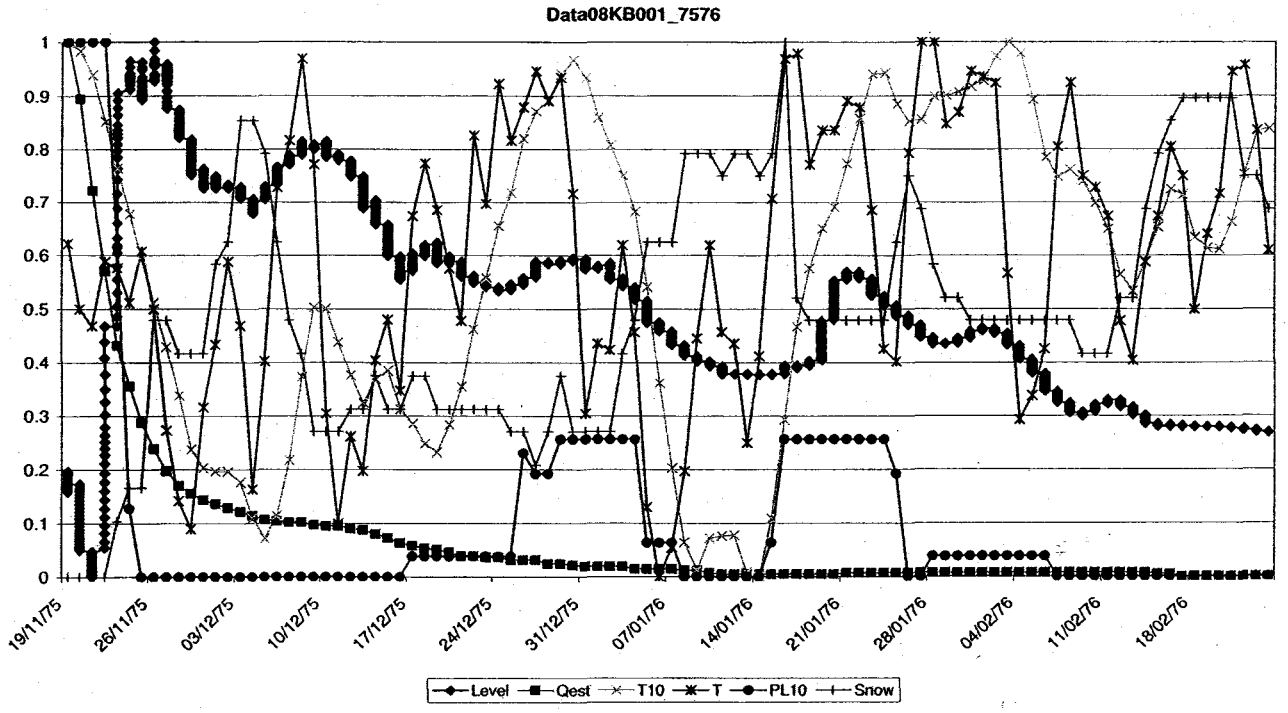


Fig. A.5. 5 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08KB001 pendant l'hiver 1975-1976.

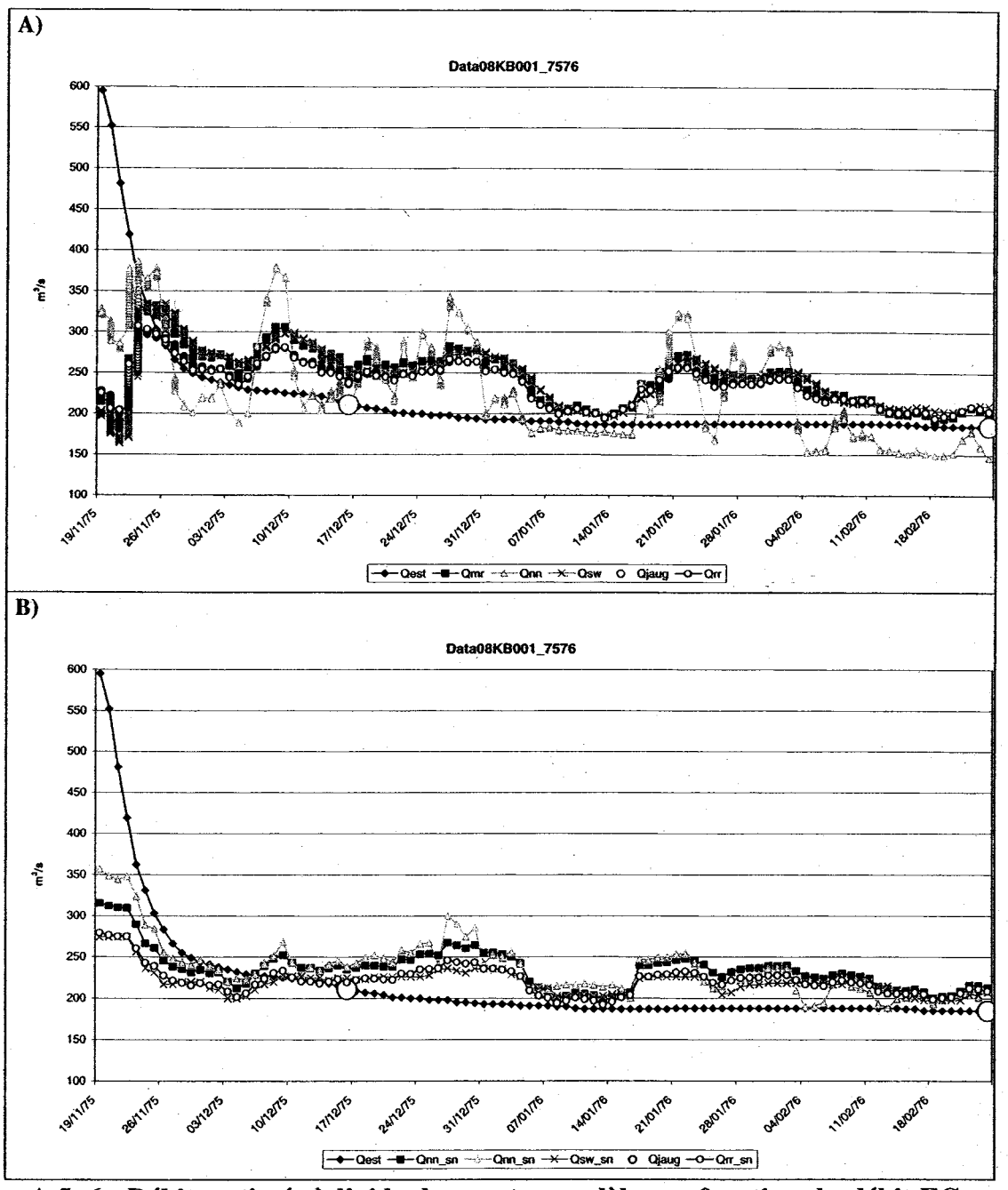


Fig. A.5. 6 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08KB001 pendant l'hiver 1975-1976 : A) simulations incluant le niveau; B) simulations sans le niveau

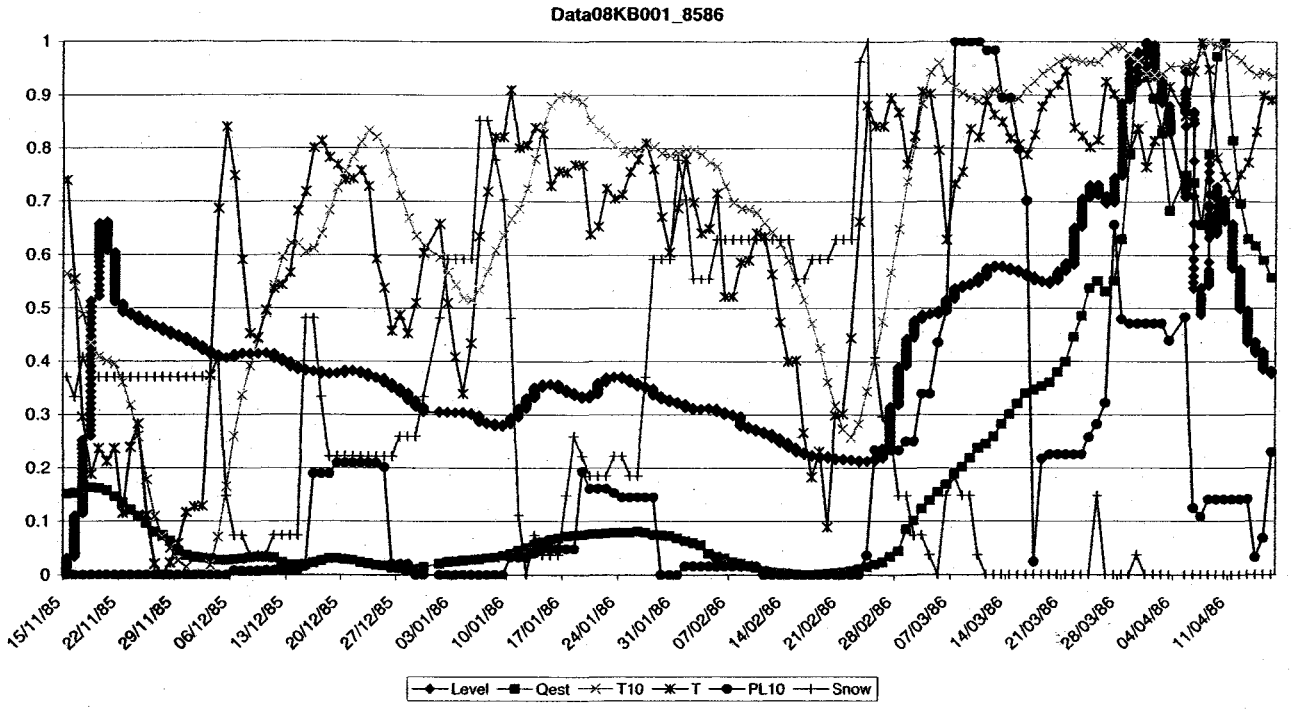


Fig. A.5. 7 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08KB001 pendant l'hiver 1985-1986.

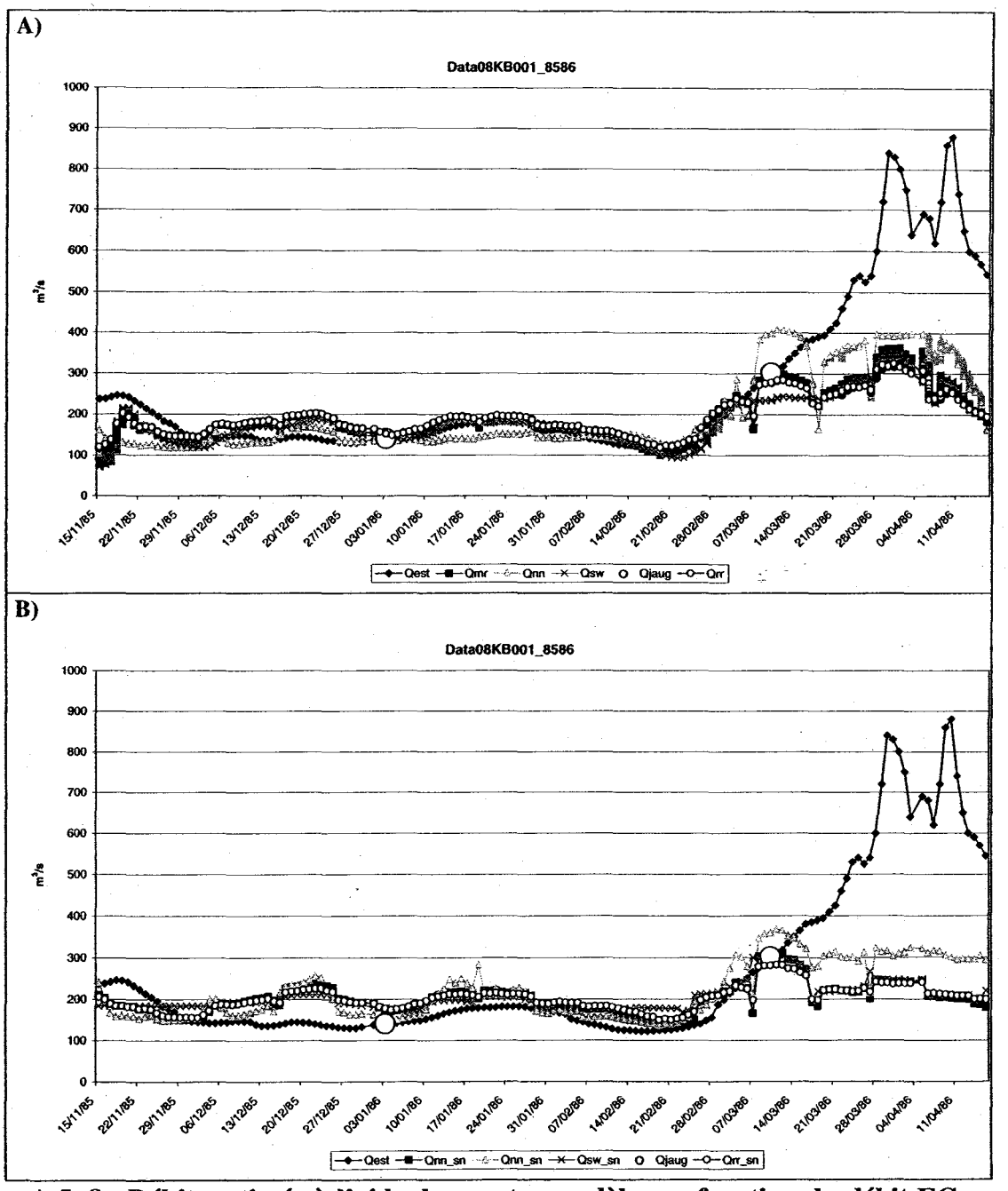


Fig. A.5. 8 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08KB001 pendant l'hiver 1985-1986 : A) simulations incluant le niveau; B) simulations sans le niveau

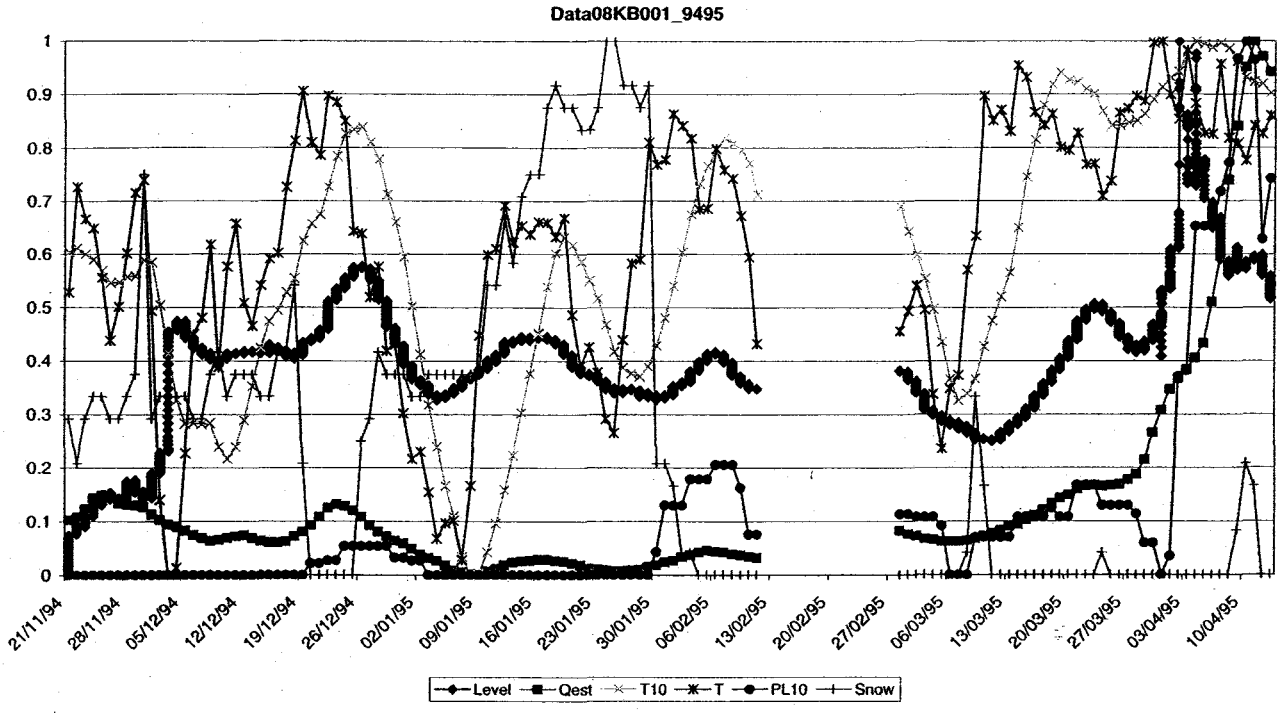


Fig. A.5. 9 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08KB001 pendant l'hiver 1994-1995.

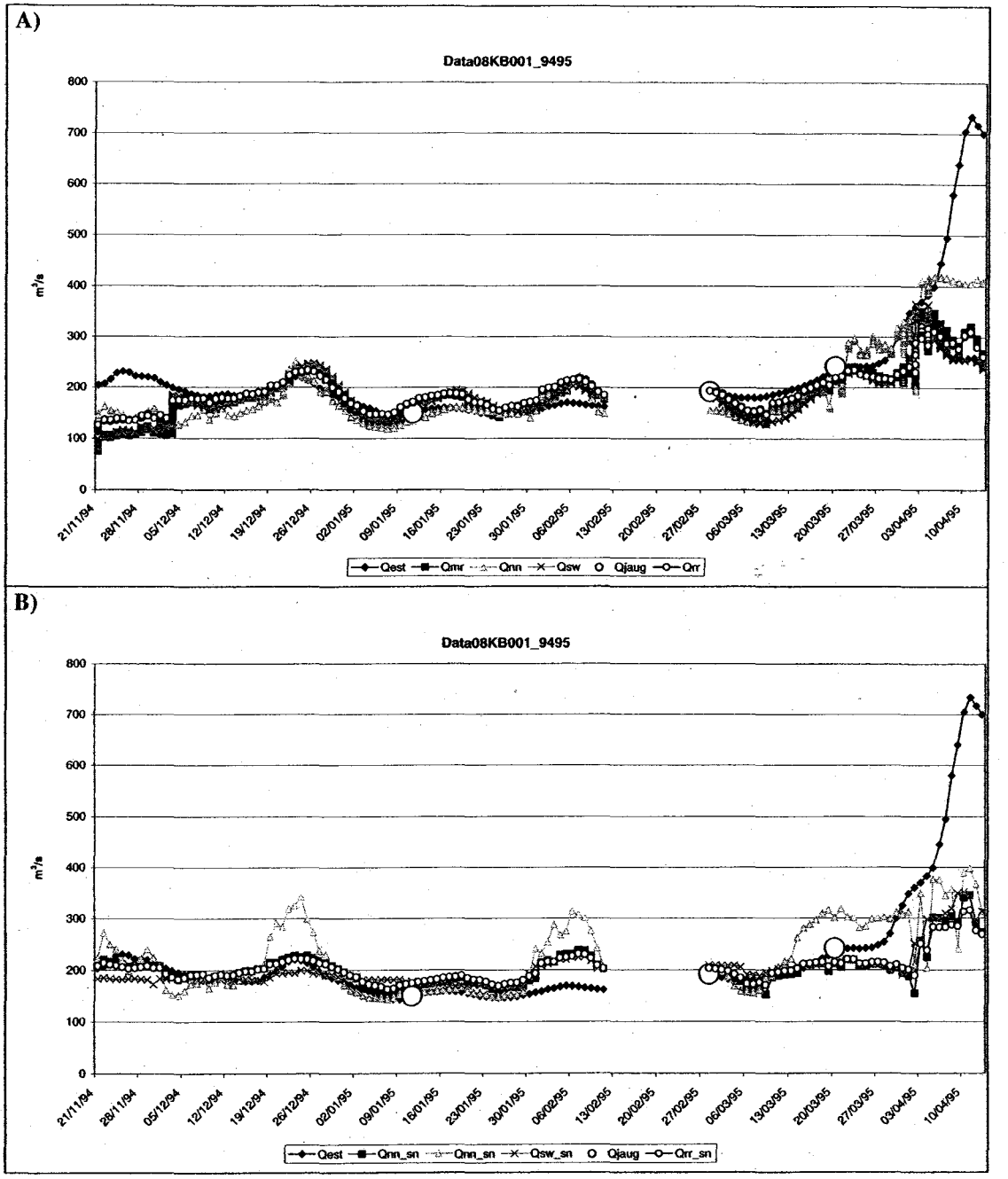


Fig. A.5. 10 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08KB001 pendant l'hiver 1994-1995 : A) simulations incluant le niveau; B) simulations sans le niveau

08KA005

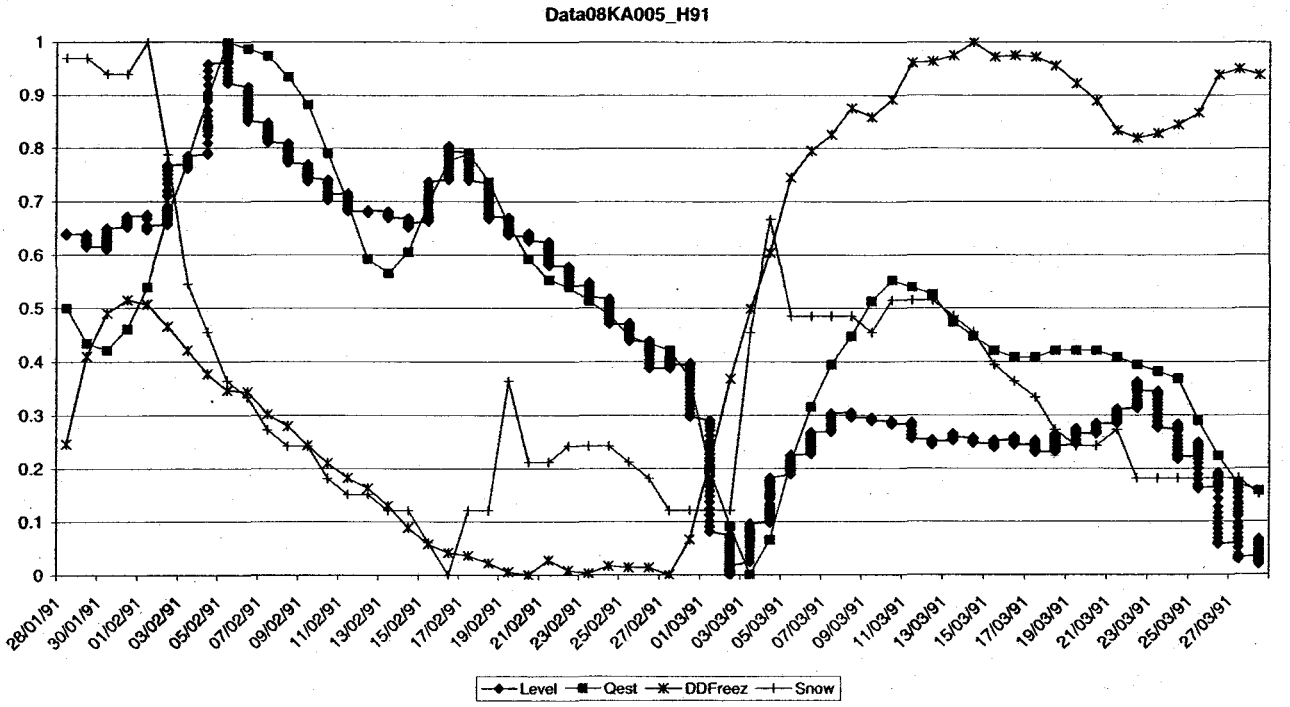


Fig. A.5. 11 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08KA005 pendant l'hiver 1991.

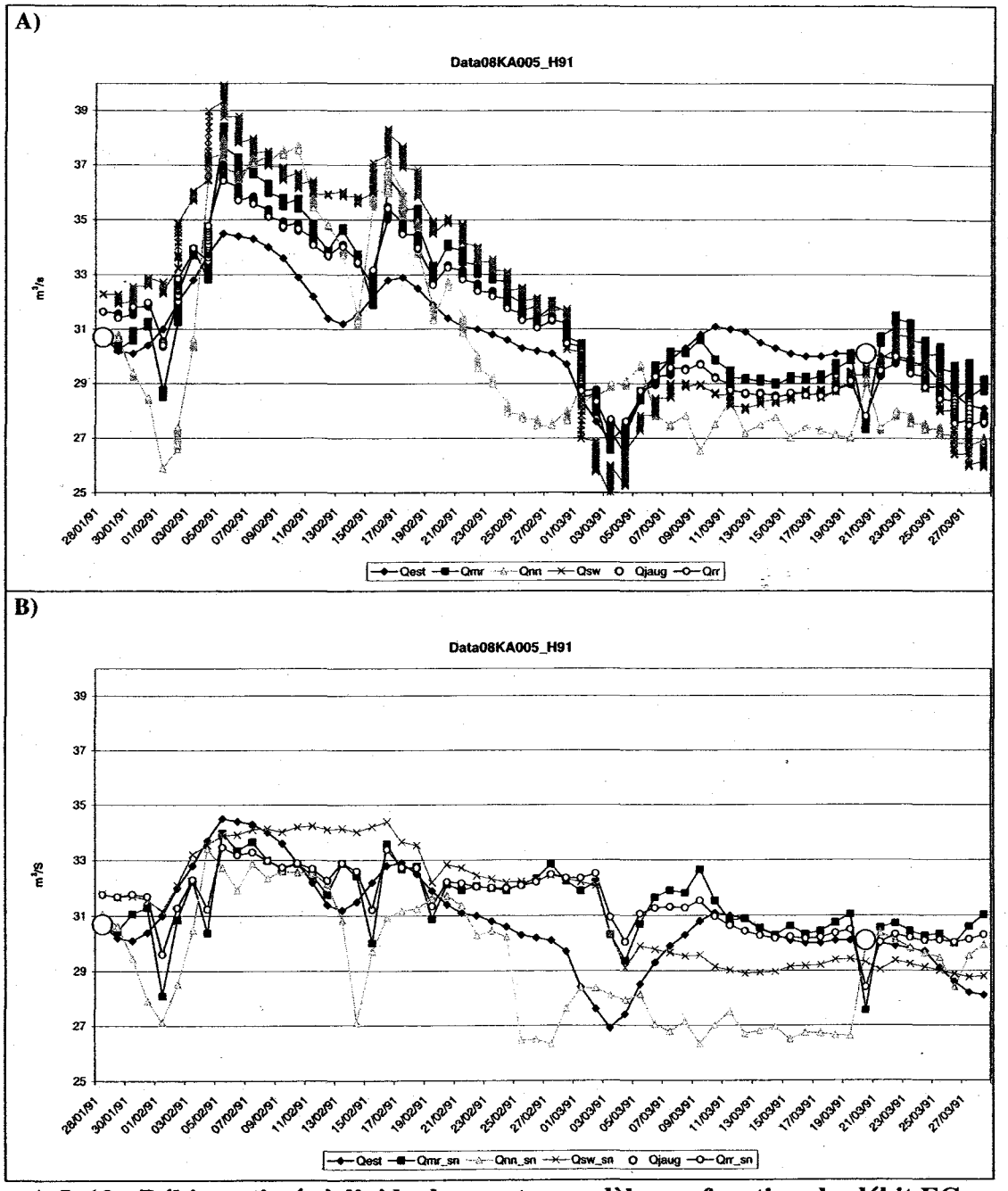


Fig. A.5. 12 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08KA005 pendant l'hiver 1991 : A) simulations incluant le niveau; B) simulations sans le niveau

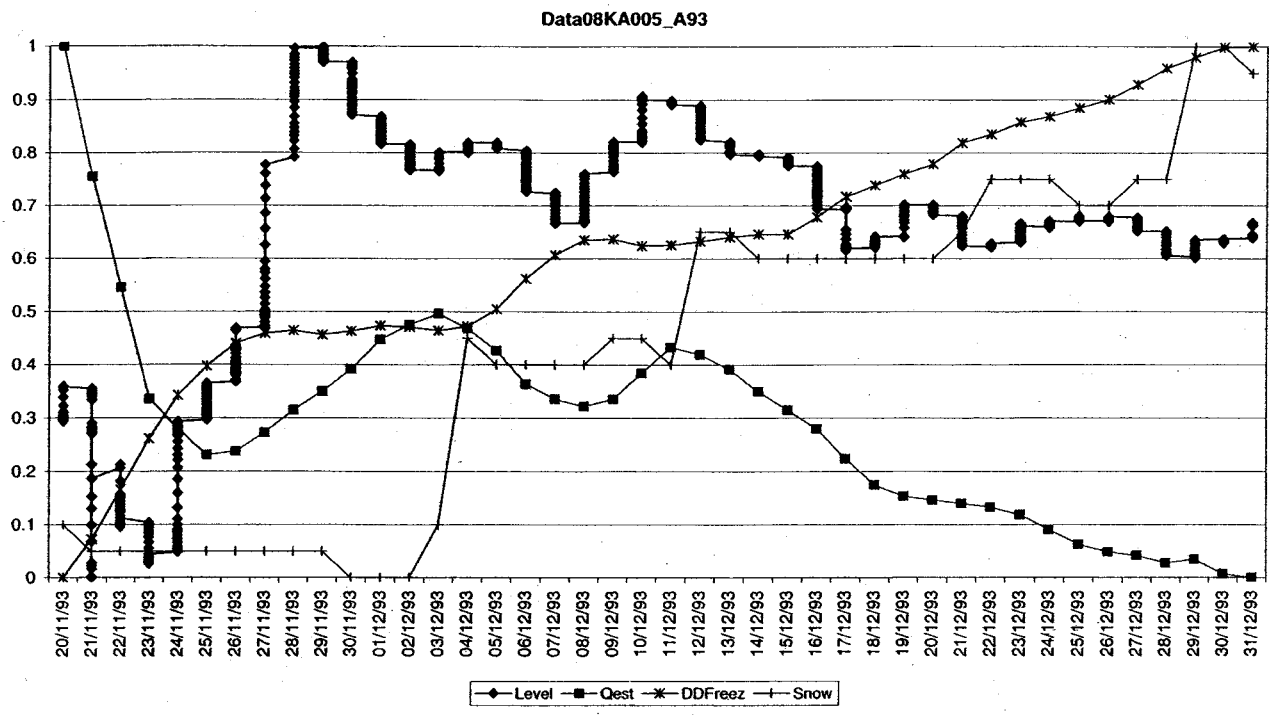


Fig. A.5. 13 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08KA005 pendant l'automne 1993.

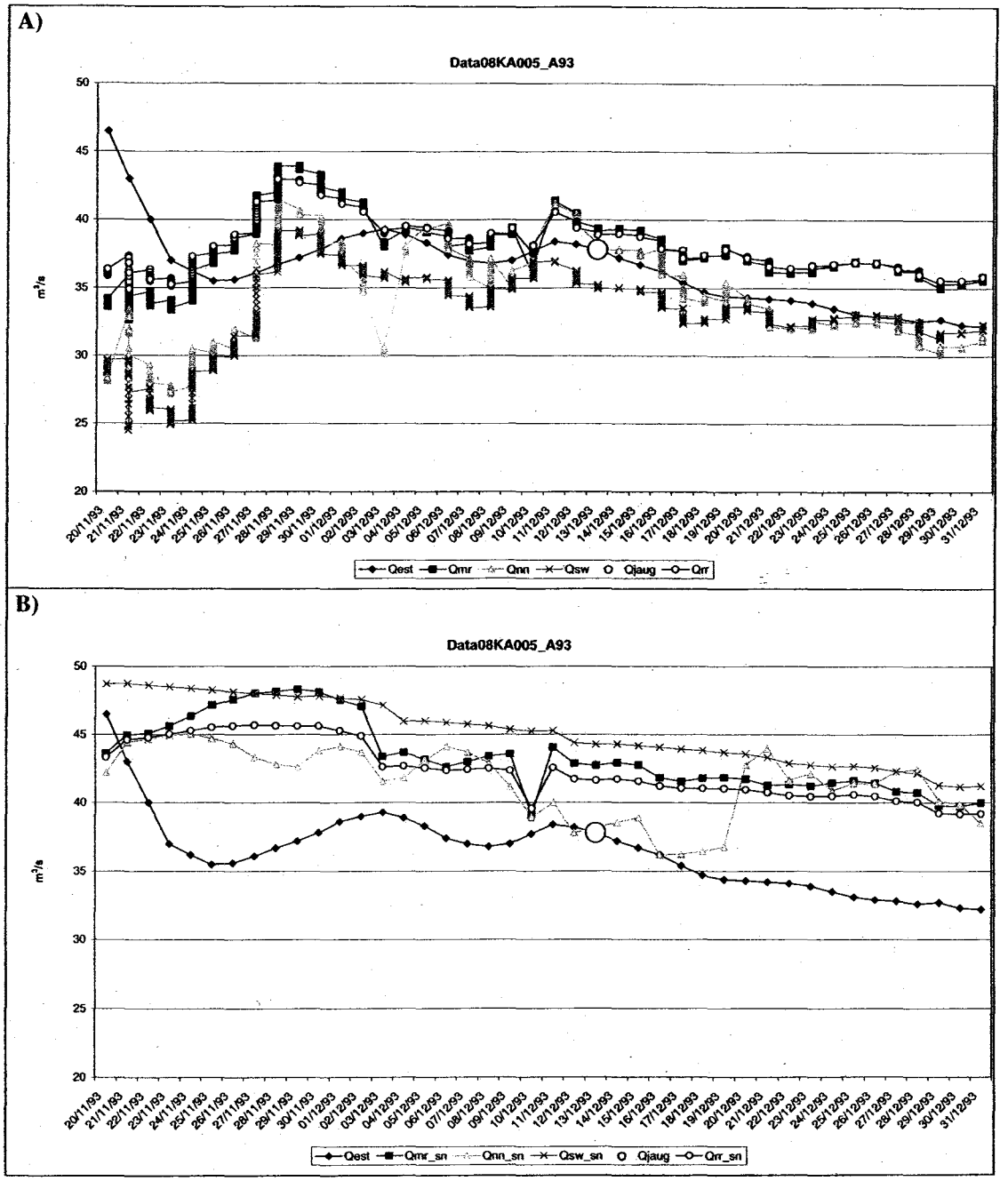


Fig. A.5. 14 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08KA005 pendant l'automne 1993 : A) simulations incluant le niveau; B) simulations sans le niveau

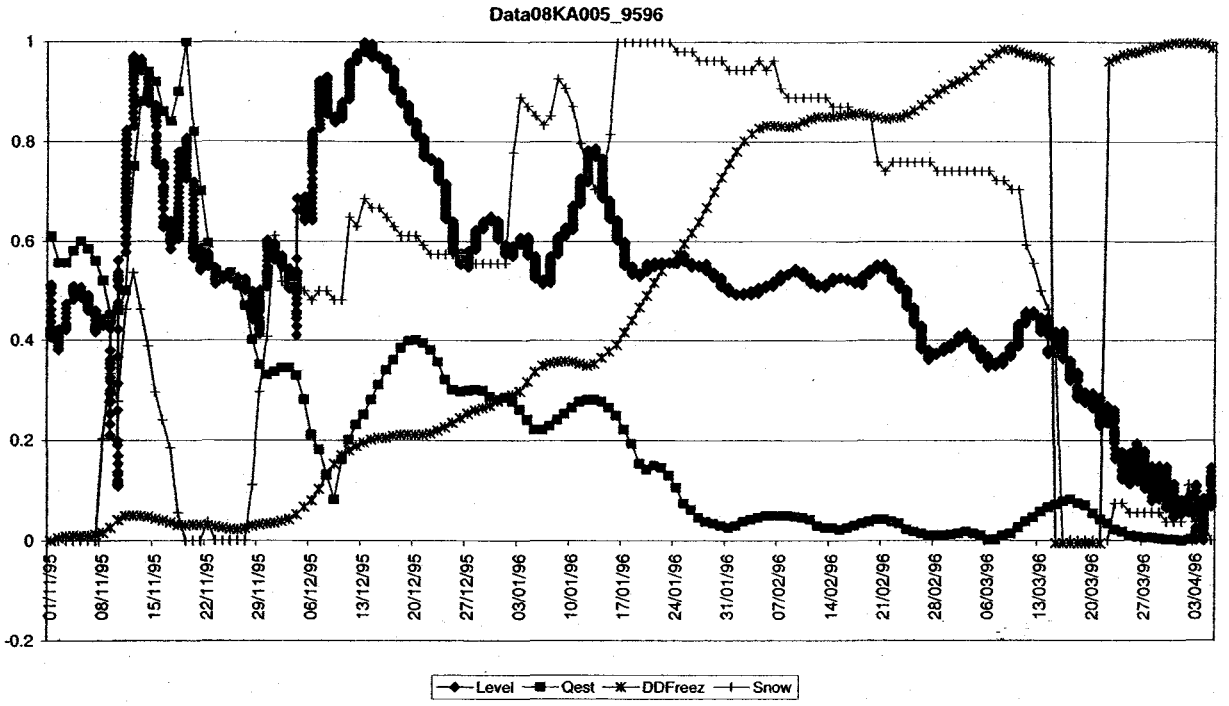


Fig. A.5. 15 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08KA005 pendant l'hiver 1995-1996.

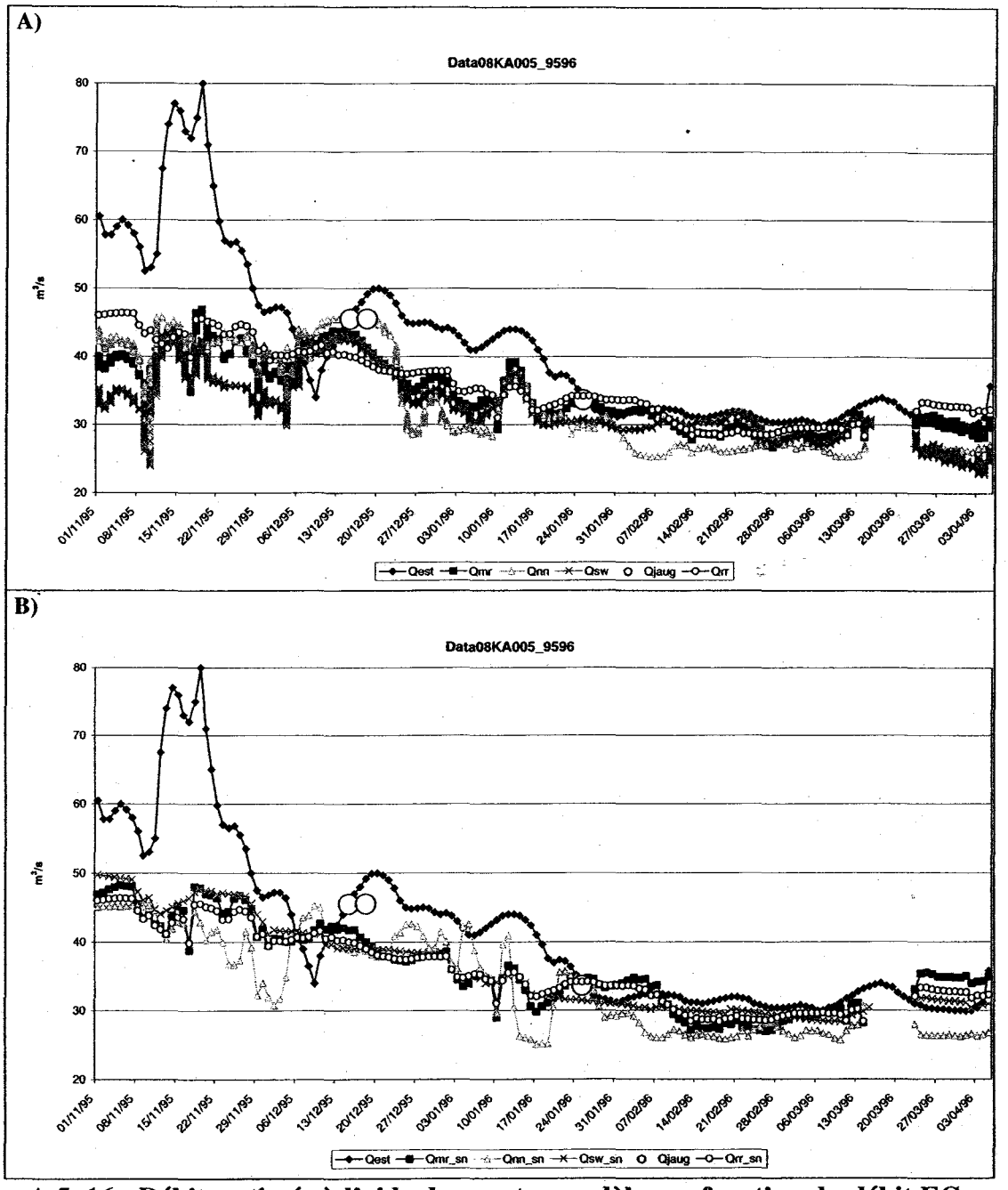


Fig. A.5. 16 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08KA005 pendant l'hiver 1995-1996 : A) simulations incluant le niveau; B) simulations sans le niveau

08MC018

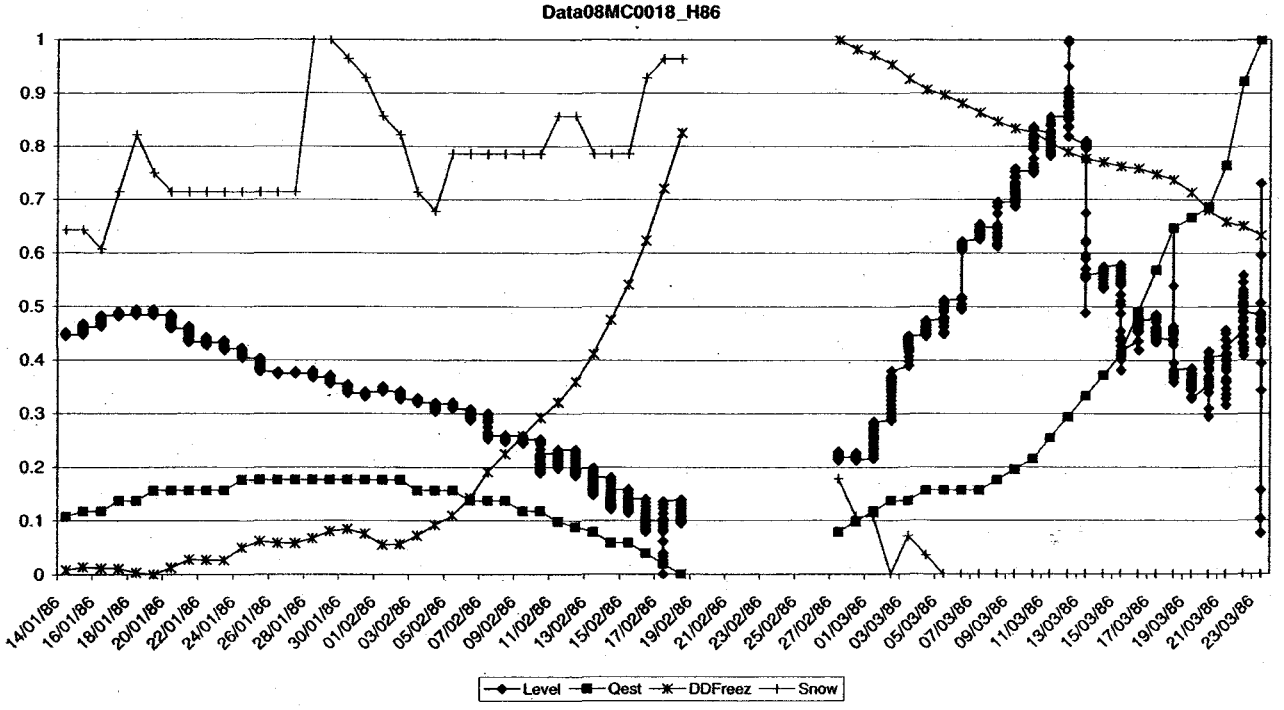


Fig. A.5. 17 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08MC018 pendant l'hiver 1986.

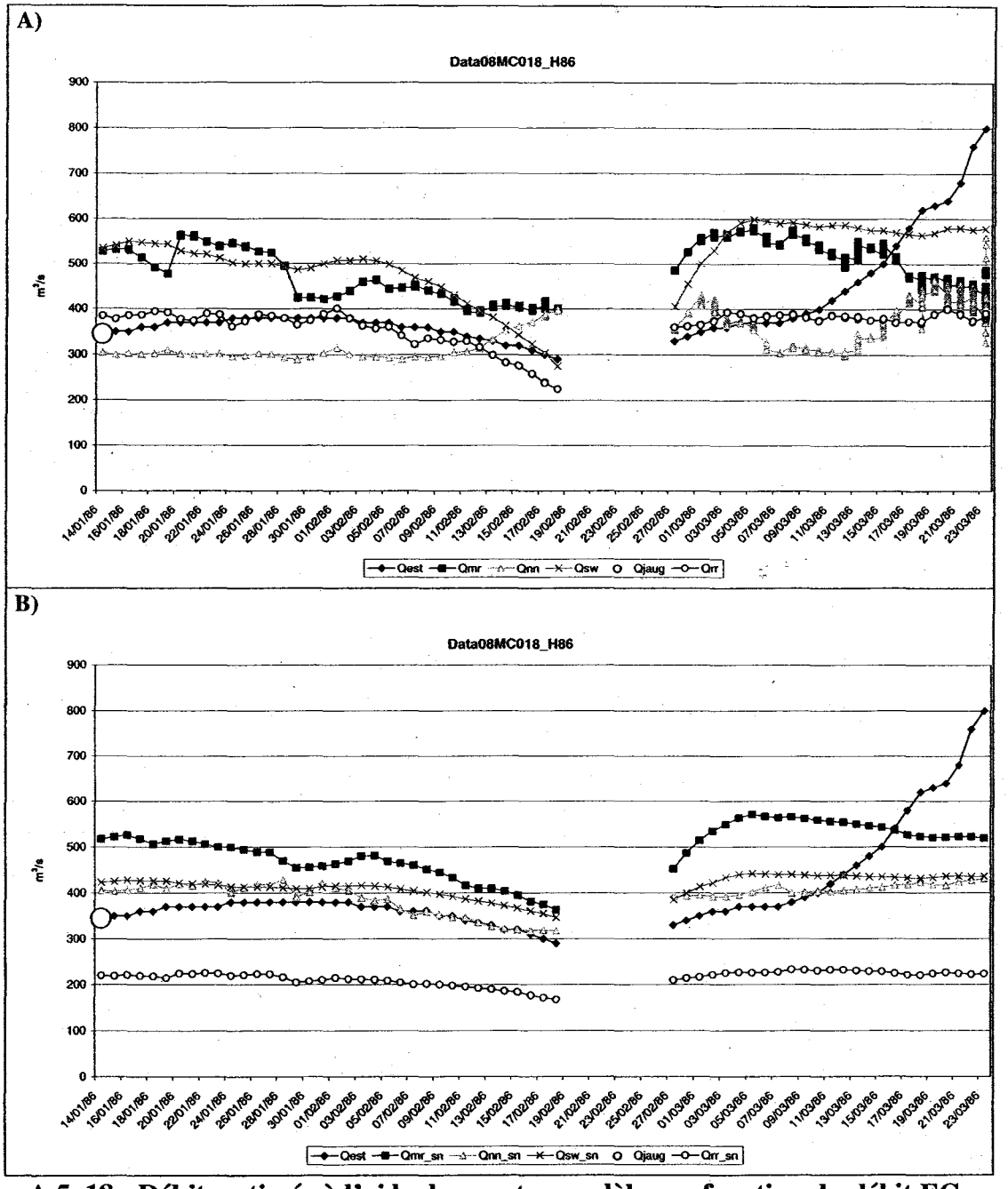


Fig. A.5. 18 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08MC018 pendant l'hiver 1986 : A) simulations incluant le niveau; B) simulations sans le niveau

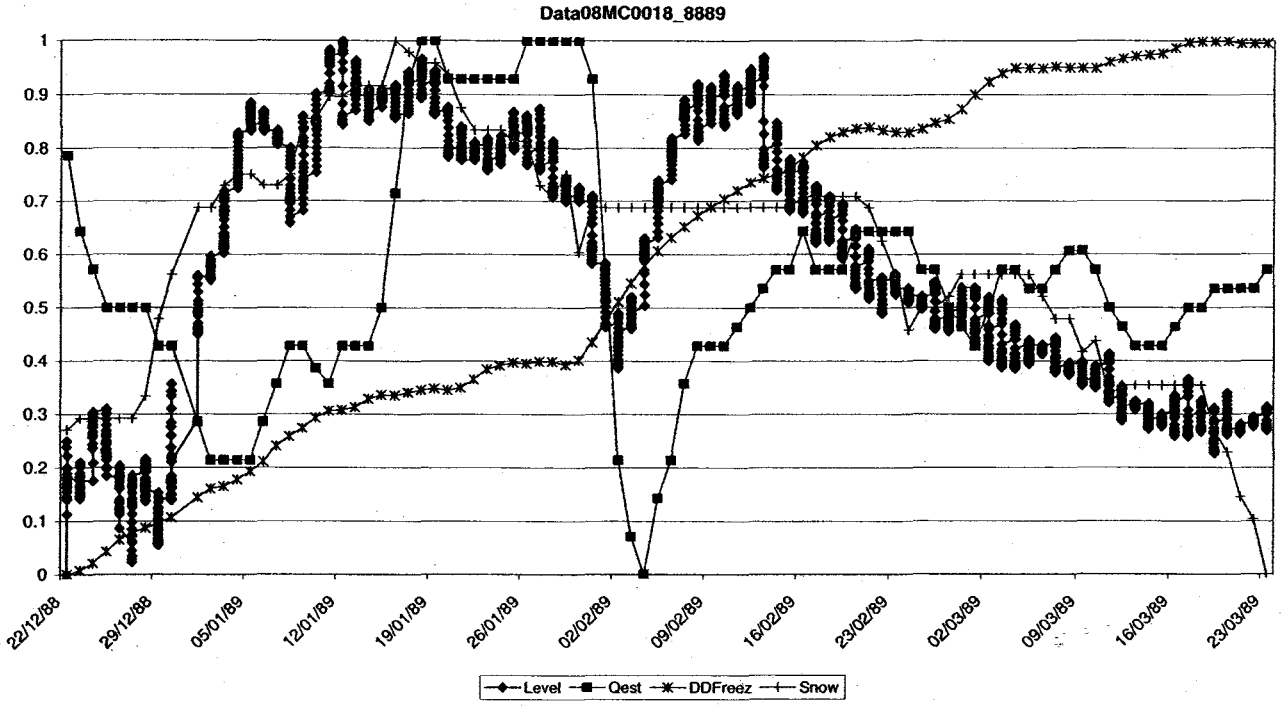


Fig. A.5. 19 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08MC018 pendant l'hiver 1988-1989.

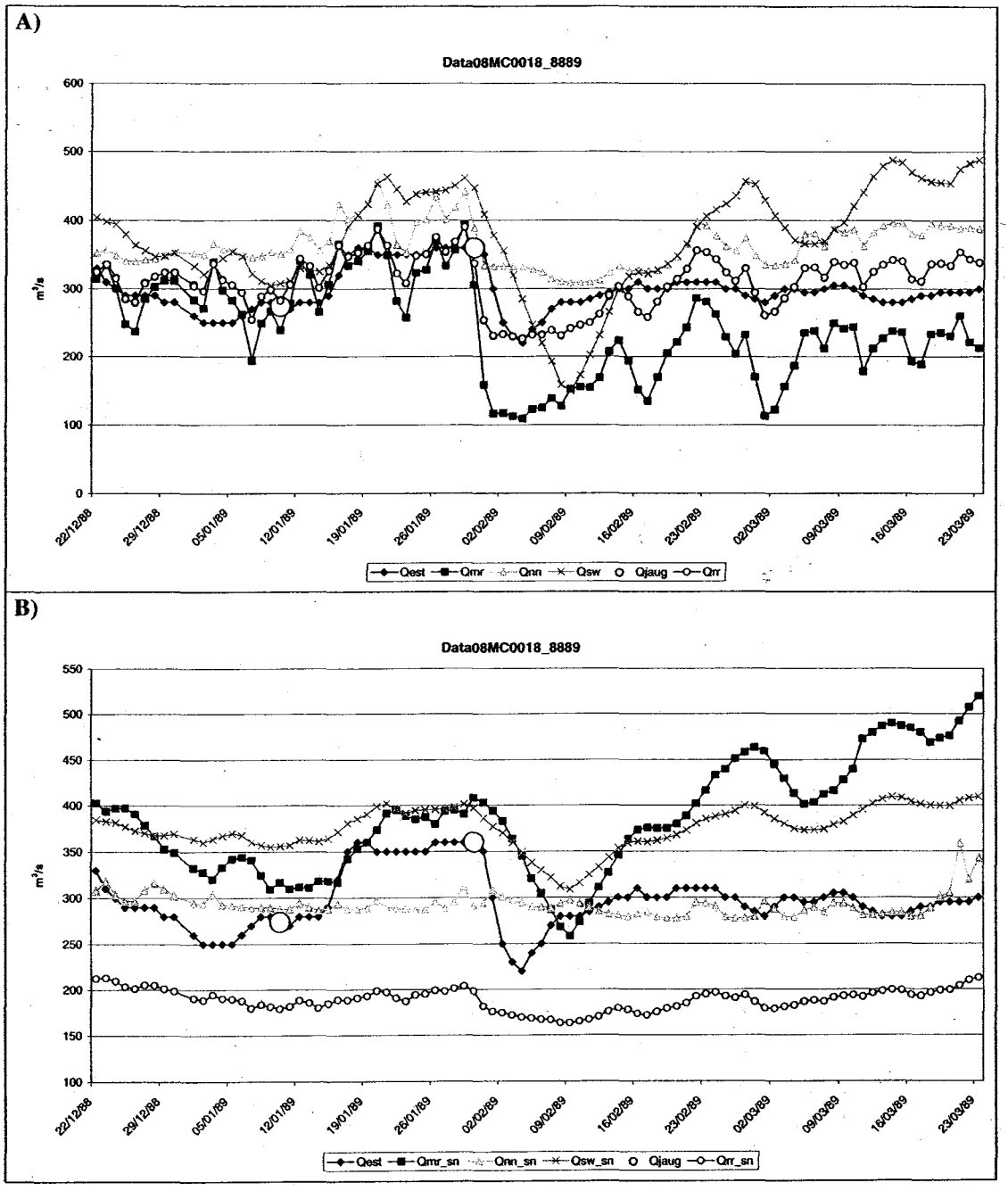


Fig. A.5. 20 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08MC018 pendant l'hiver 1988-1989 : A) simulations incluant le niveau; B) simulations sans le niveau

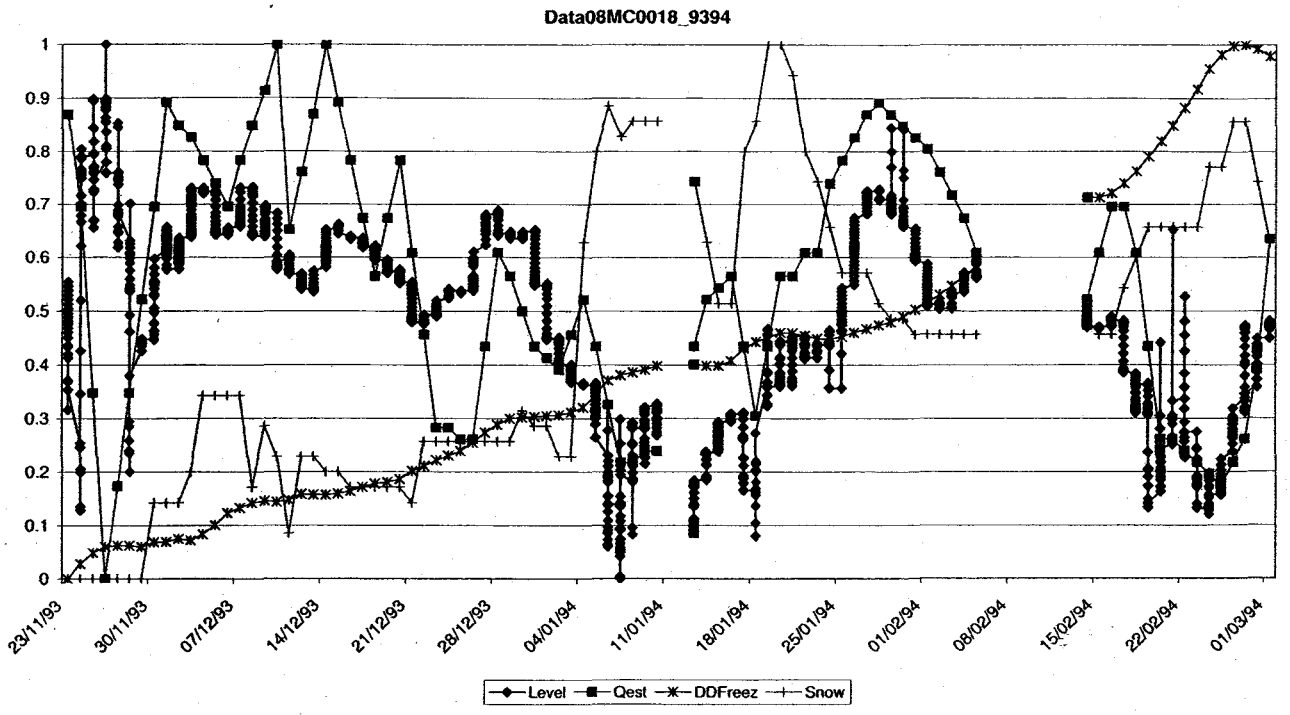


Fig. A.5. 21 : Les variables explicatives les plus significatives pour le débit jaugé en fonction du débit EC pour la station 08MC018 pendant l'hiver 1993-1994.

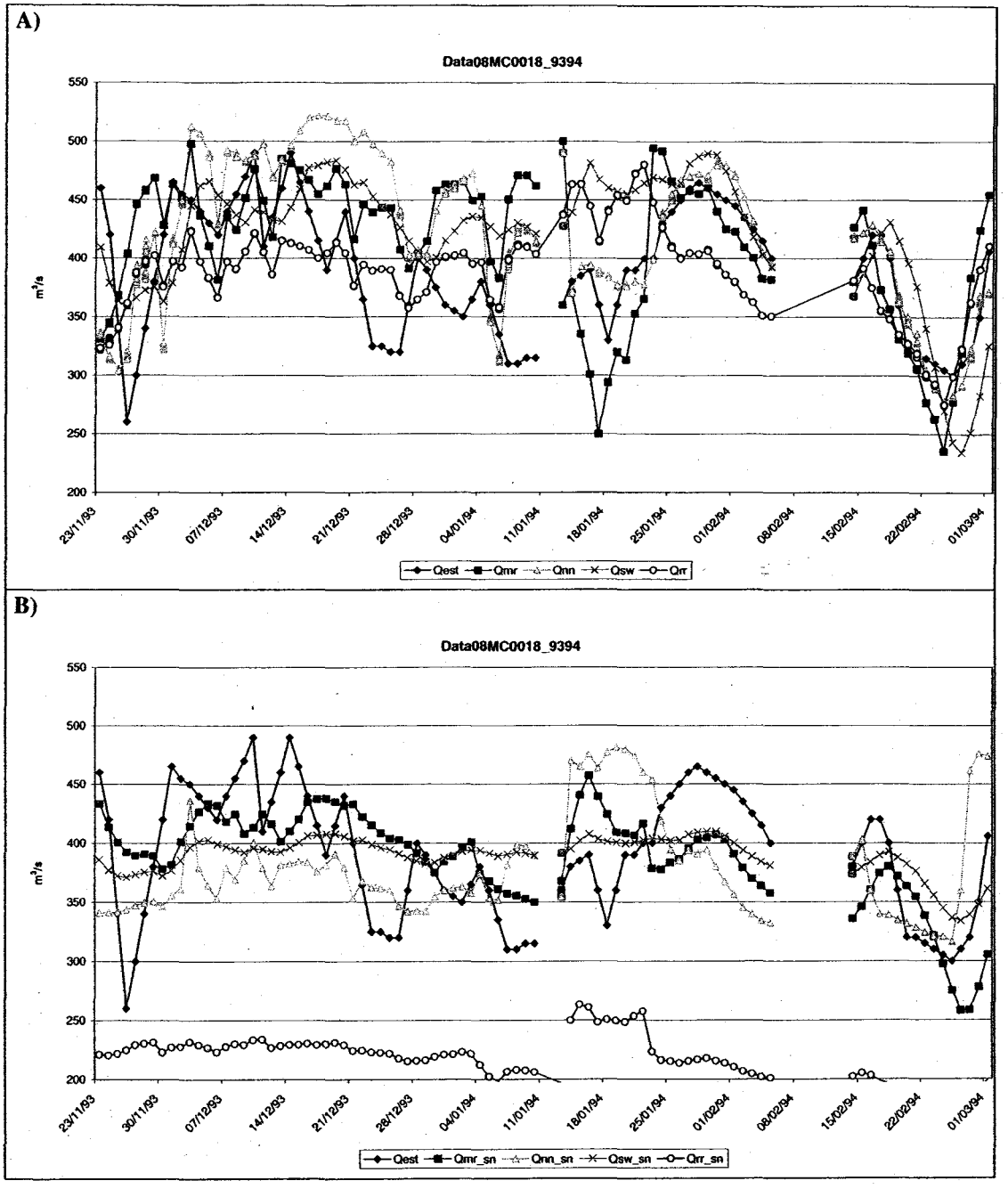


Fig. A.5. 22 : Débits estimés à l'aide des quatre modèles en fonction du débit EC pour la station 08MC018 pendant l'hiver 1993-1994 : A) simulations incluant le niveau; B) simulations sans le niveau

ANNEXE 6: *UNICCO* USER'S GUIDE

1 SOFTWARE OVERVIEW

UNICCO is a convivial data-processing tool and flexible device for the visualization and the manipulation of the hydrometric and meteorological available data, the calibration of the various models implemented within the tool as well as the estimation on a real time basis of the streamflow corrected for ice effect. It was developed within the Matlab (version 6.1) environment. The software is setup by copying the necessary files on the desired directory (some with the extension ".m" are for the programs and others with the ".fig" extension are for the interfaces). Prior to using UNICCO , the user must install first the Matlab 6.1 (or any latter version) on his computer. To execute the software, the user need to run the Matlab software and then type "unicco_principal" into the Matlab command window. The UNICCO main menu window appears as shown in Fig.A.6. 1

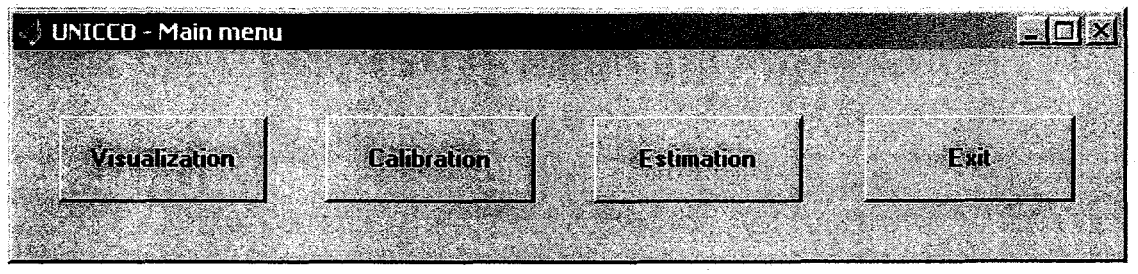
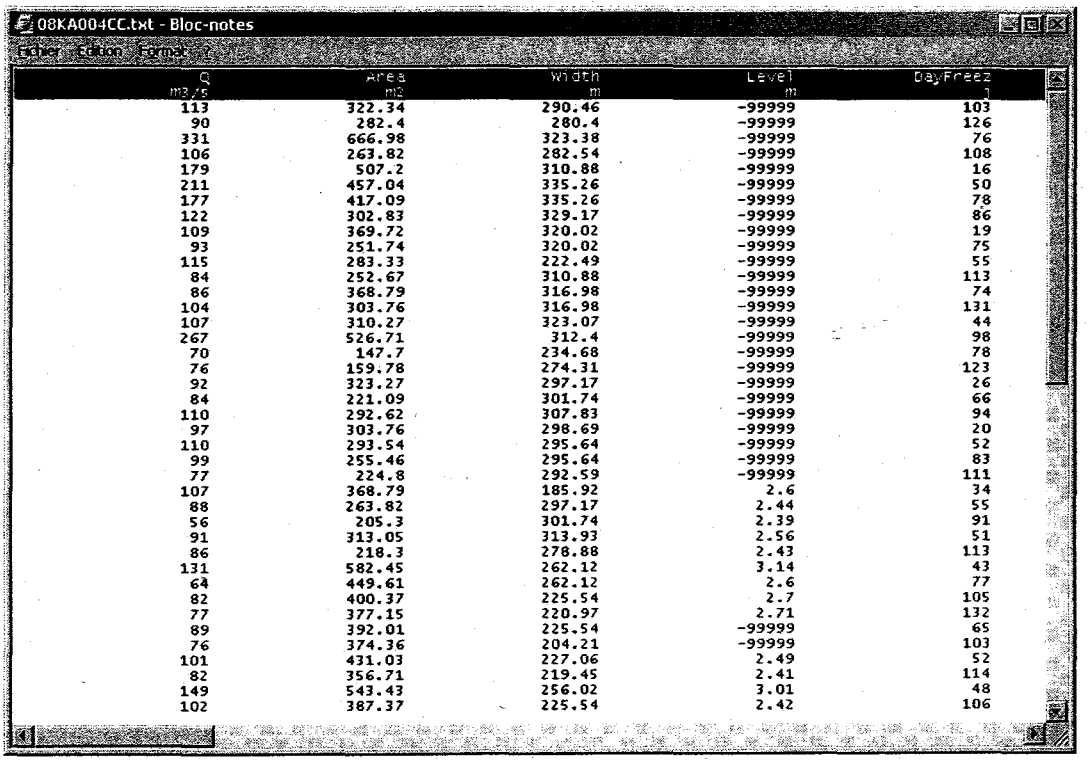


Fig.A.6. 1 : The *UNICCO* main menu window.

The main menu window consists of four buttons: an "Exit" button to end the program and three buttons to launch the visualization, calibration and the estimation modules. The sections 2.1, 2.2 and 2.3 of the present annex explain respectively the operation mode of each module.

UNICCO accepts inputs only in the ASCII text file format: tab separated data ("txt" extension), space separated data ("prn" extension) and semi-colon separated data ("csv" extension). Each column represents a variable.

As shown in the Fig.A.6. 2, the first and the second line of the input file is dedicated respectively to variables names and units. They consist of a sequence of alphanumerical characters. Any of the characters may not be a space character. The rest of the file consists of the variables values. Missing data should be represented by the code "-99999". This flexible format allows to use any number of dependent or independent variables.



08KA004CC.txt - Bloc-notes

Fichier Edition Format ?

Q	Area	Width	Level	DayFreez
m3/s	m2	m	m	j
113	322.34	290.46	-99999	103
90	282.4	280.4	-99999	126
331	666.98	323.38	-99999	76
106	263.82	282.54	-99999	108
179	507.2	310.88	-99999	16
211	457.04	335.26	-99999	50
177	417.09	335.26	-99999	78
122	302.83	329.17	-99999	86
109	369.72	320.02	-99999	19
93	251.74	320.02	-99999	75
115	283.33	222.49	-99999	55
84	252.67	310.88	-99999	113
86	368.79	316.98	-99999	74
104	303.76	316.98	-99999	131
107	310.27	323.07	-99999	44
267	526.71	312.4	-99999	98
70	147.7	234.68	-99999	78
76	159.78	274.31	-99999	123
92	323.27	297.17	-99999	26
84	221.09	301.74	-99999	66
110	292.62	307.83	-99999	94
97	303.76	298.69	-99999	20
110	293.54	295.64	-99999	52
99	255.46	295.64	-99999	83
77	224.8	292.59	-99999	111
107	368.79	185.92	2.6	34
88	263.82	297.17	2.44	55
56	205.3	301.74	2.39	91
91	313.05	313.93	2.56	51
86	218.3	278.88	2.43	113
131	582.45	262.12	3.14	43
64	449.61	262.12	2.6	77
82	400.37	225.54	2.7	105
77	377.15	220.97	2.71	132
89	392.01	225.54	-99999	65
76	374.36	204.21	-99999	103
101	431.03	227.06	2.49	52
82	356.71	219.45	2.41	114
149	543.43	256.02	3.01	48
102	387.37	225.54	2.42	106

Fig.A.6. 2 : An example of the *UNICCO* input file

2 SOFTWARE OPERATION

2.1 VISUALISATION MODULE

The visualisation module is devoted to visualization, exploration, manipulation and transformation of variables within the input file. The visualization module is executed by pushing the main menu visualization button, which bring up the window entitled "Unicco-Variable visualization", as shown in Fig.A.6. 3. At the top of the window, is located the input file path and filename capture area. The file path and name can be entered manually or selected using the Windows file explorer through the "Browse" button. The user can then list the variables names contained in the selected input file by pressing the "Update list" button. When one or more variables are selected, the sample size is shown, from which missing data is excluded.

It is also possible to delete one or more variables. This action will remove definitively the selected variables from the file. So, when the delete button is pressed, a dialogue window appears to confirm the action in order to prevent an involuntarily variable removal. After deleting the variables, the user should press the "Update list" button in order to refresh the variables list.

For any selection of variables, the "Statistics" option allows to calculate the principal descriptive statistics (mean, median, maximum, minimum, standard-deviation, skewness and kurtosis) as well as the correlation matrix (Fig.A.6. 4). It is possible then to export these statistics and save them into a text file.

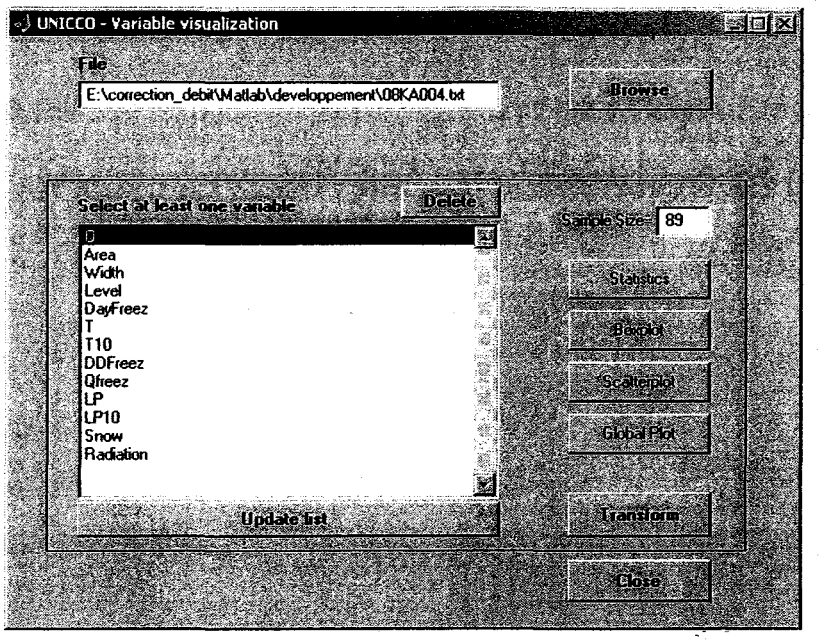
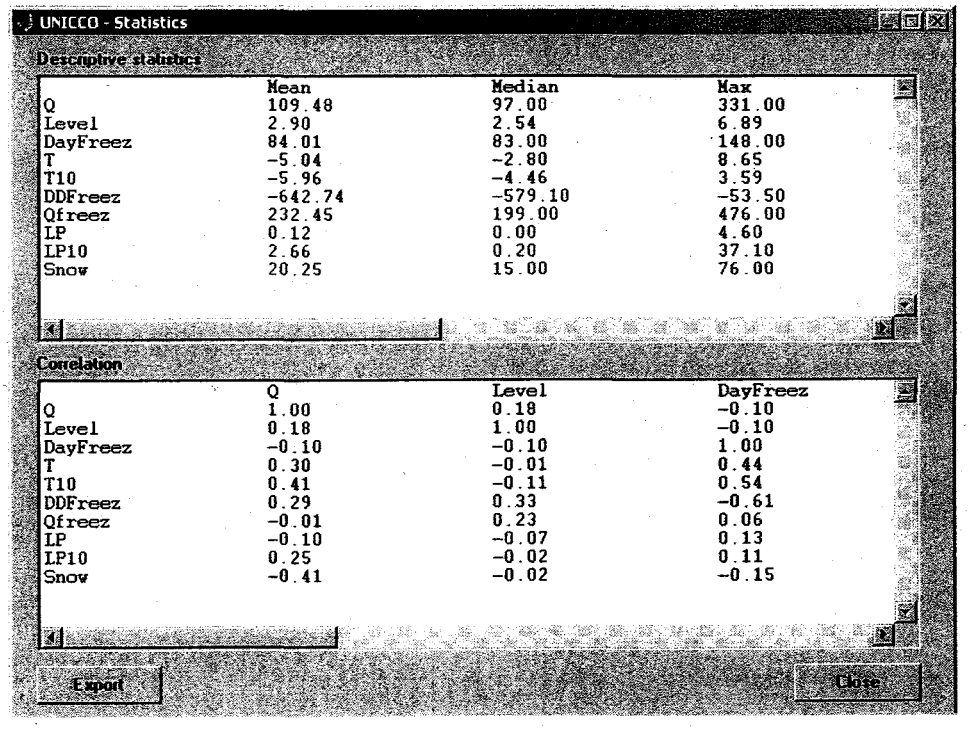


Fig.A.6. 3 : The visualization module window



	Mean	Median	Max
Q	109.48	97.00	331.00
Level	2.90	2.54	6.89
DayFreez	84.01	83.00	148.00
T	-5.04	-2.80	8.65
T10	-5.96	-4.46	3.59
DDFreez	-642.74	-579.10	-53.50
Qfreez	232.45	199.00	476.00
LP	0.12	0.00	4.60
LP10	2.66	0.20	37.10
Snow	20.25	15.00	76.00

	Q	Level	DayFreez
Q	1.00	0.18	-0.10
Level	0.18	1.00	-0.10
DayFreez	-0.10	-0.10	1.00
T	0.30	-0.01	0.44
T10	0.41	-0.11	0.54
DDFreez	0.29	0.33	-0.61
Qfreez	-0.01	0.23	0.06
LP	-0.10	-0.07	0.13
LP10	0.25	-0.02	0.11
Snow	-0.41	-0.02	-0.15

Fig.A.6. 4 : The variables statistics window

As well, it is possible to produce box plots for the selected variables (Fig.A.6. 5). A box plot is graphical representation of the variable distribution. Hence, the mean value (red dash), the confidence interval on the mean value (cuts), the quartiles (the box for 2nd and 3rd quartiles and blue dashes for min and max) and extreme values (blue asterisks) are also shown.

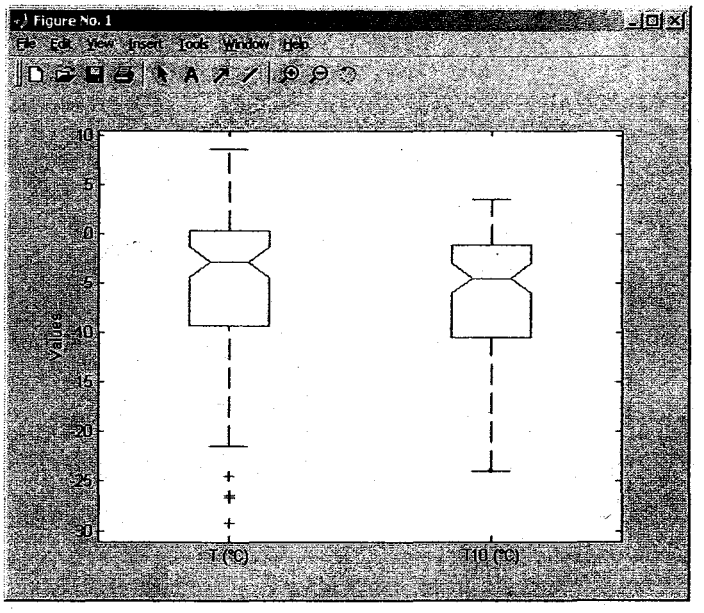


Fig.A.6. 5 : Example of a two variables boxplot.

The "Scatterplot" option permits to plot one selected variable according to another (Fig.A.6. 6). This option works only with a selection of two variables. If the user wants to draw more than two variables, he must choose then the "Global Plot" option. This feature produces a matrix like graph where the diagonal is made up of each variable histogram and the rest of the cells is occupied with two by two variables scatter plots (Fig.A.6. 7). If only one variable is selected, the Global plot consists only of the variable histogram.

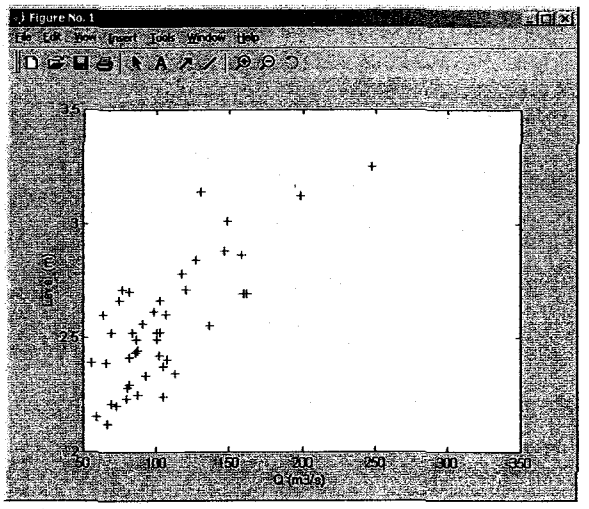


Fig.A.6. 6 : The scatter plot window

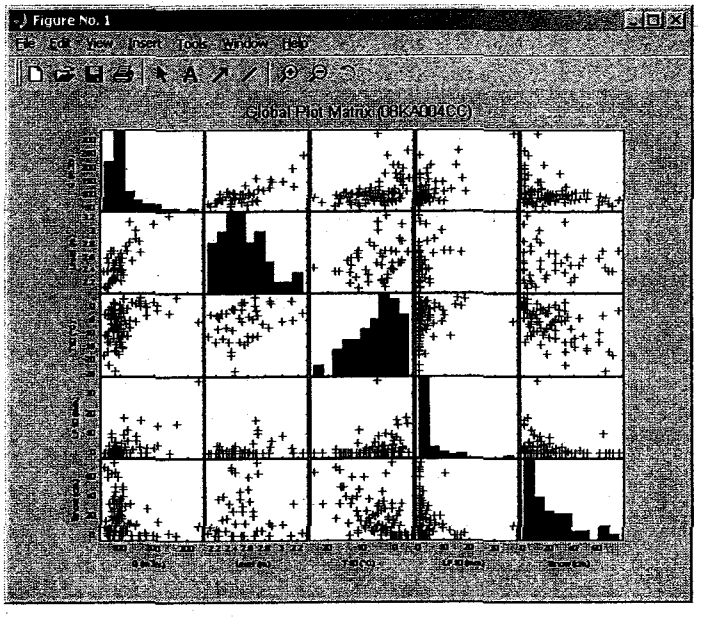


Fig.A.6. 7 : The global plot window

The visualization module allows, if needed, to transform the selected variable. Thus, the user can add a constant "A" to the variable, raise it to the power "Y" or calculate its inverse, absolute value and its logarithm (Fig.A.6. 8). The user has to specify the new variable name and unit. If no unit is need, the user has to enter the code "NA". The "Save" button adds the transformed variable to the input file, closes the transformation window and returns back to the visualization window.

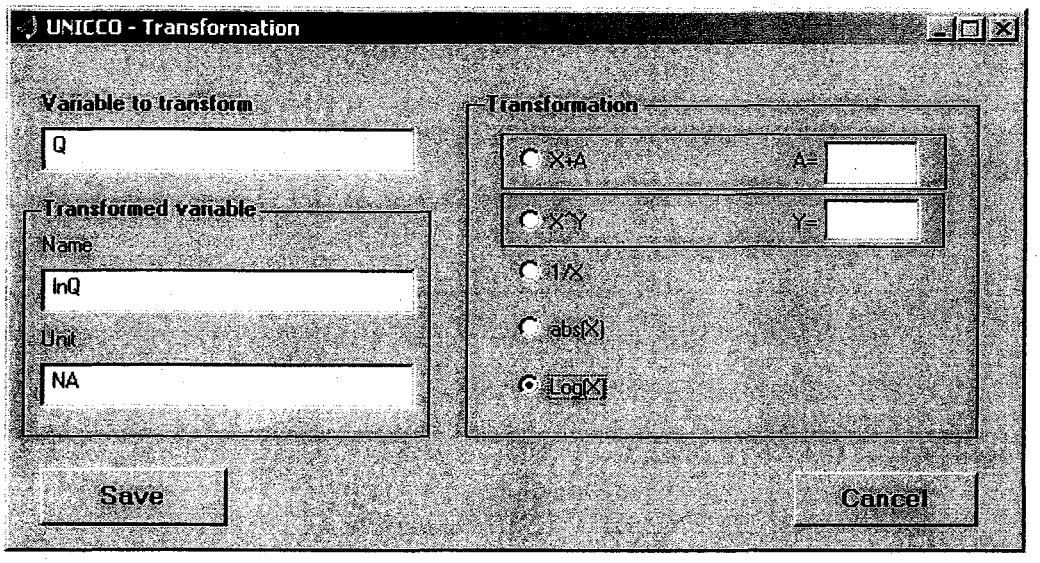


Fig.A.6. 8: The variables transformation window

2.2 CALIBRATION MODULE

The calibration module is the main element of *UNICCO* (Fig.A.6. 9). This module makes it possible to adjust regressive and neural models to the observed data included in the input file. The calibrated model can be saved to a file which can be loaded for later use in order to produce estimations of the dependent variable using a different set of values of the explanatory variables.

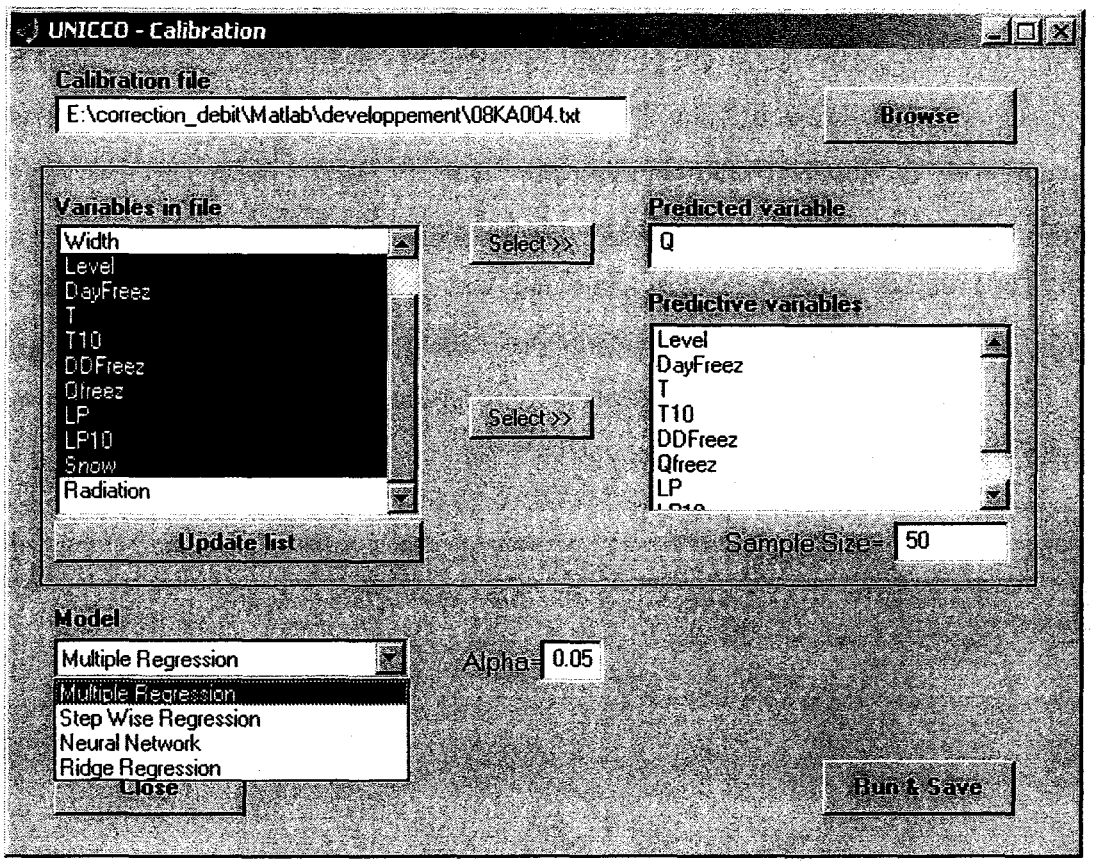


Fig.A.6. 9 : The calibration module window

As for the visualization module, the input file can be specified either by entering manually the path and the file name or by browsing through the computer directories. Once the calibration data file (input file) is loaded, the "Update list" button allows to list the variables within the file. The user has thus to specify the dependent variable (predicted variable), in our case the streamflow (Q) and select the explanatory variables (Predicteve variables) to be included in the model. This choice is made by using the "Select" button next to the respective edition zone. Afterward, the user has to choose the type of model by using the popup menu in the lower left corner. In addition to the neural model, three regressive models are implemented within the *UNICCO* calibration module: multiple regression, stepwise regression and ridge regression.

If one of the regressive models is chosen, it is necessary to specify the statistical significance level (Alpha) which will be used to estimate the confidence intervals of the model parameters. The default value is 0.05. When the calibration procedure of the regressive models is launched after pressing the "Run & Save" button, the user has to specify the name of the ".mat" format file in which the model parameters will be saved for later use. The default suggested file name has the following structure: "[*the model type*]_[*the input file name*].mat". The file name part corresponding to the model type could be "MultipleRegression", "StepWiseRegression", "RidgeRegression" or "NeuralNetwork", according to the selected model. However, the user can choose another file name.

Therefore, a figure showing the estimated values of the dependent variable using the regressive model are posted according to the observed value (Fig.A.6. 10).

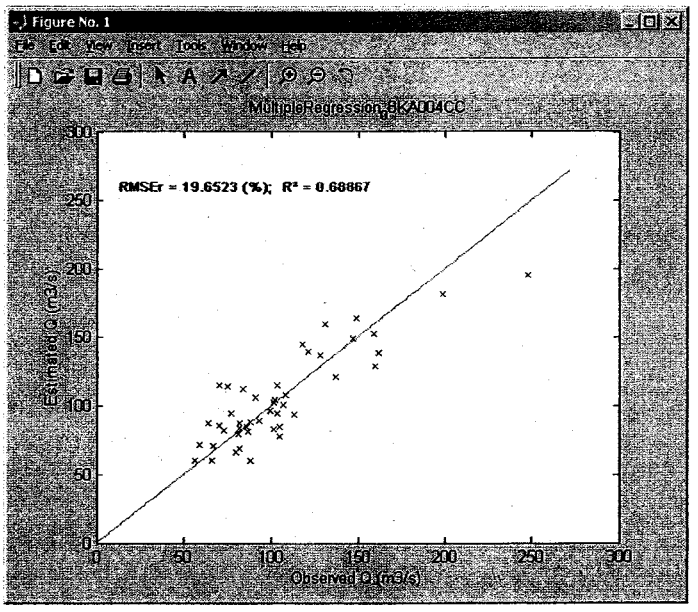


Fig.A.6. 10 : Observed vs estimated values of the dependent variable scatter plot

Also, a window presenting the performances of the regressive model is produced (Fig.A.6. 11). In this window, the model parameters and their confidence intervals, the sample size and the *p* value of the F test are presented as well as the model performance evaluation criteria (adjusted R-square, Nash criterion, RMSE and Bias). In the case of the ridge regression, the optimised *k* parameter is also shown. For the stepwise regression, only the

model parameters corresponding to the explanatory variables selected by the stepwise algorithm are presented. The model performances can be exported and saved to a text file whose name is preceded by the prefix "*perform_*" and the remainder is similar to that of the model file.

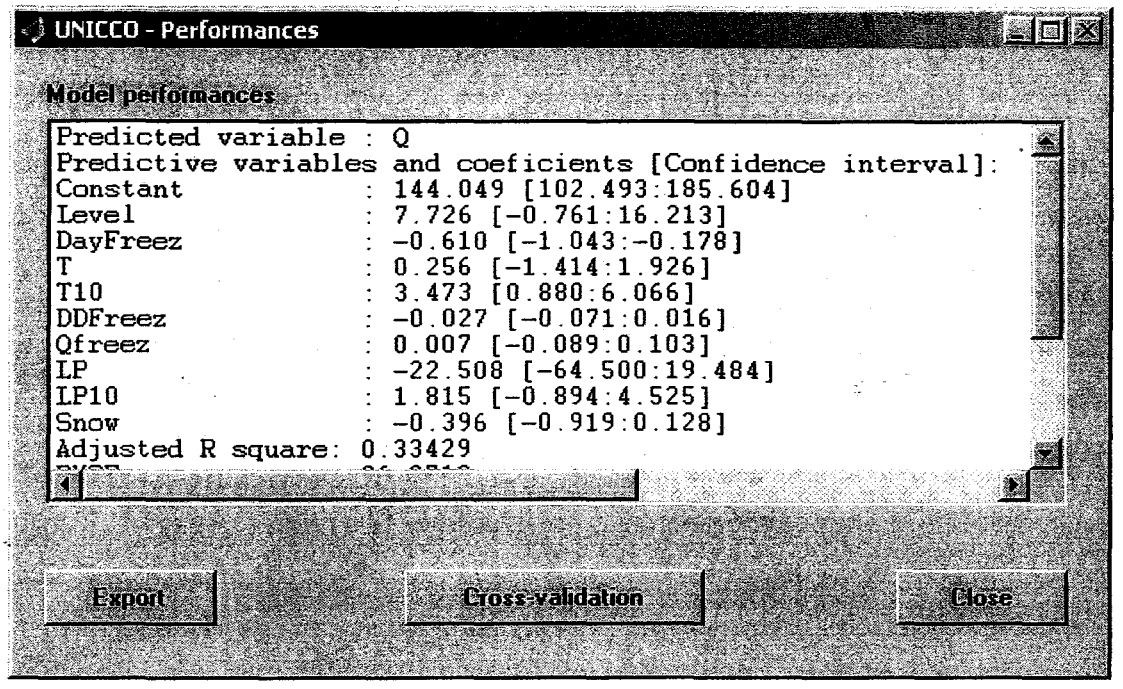


Fig.A.6. 11 : The model performances window

Through the model performances window it is also possible to perform a cross validation of the calibrated model. The results are plotted therefore and the calibration performance criteria are also shown as Fig.A.6. 12.

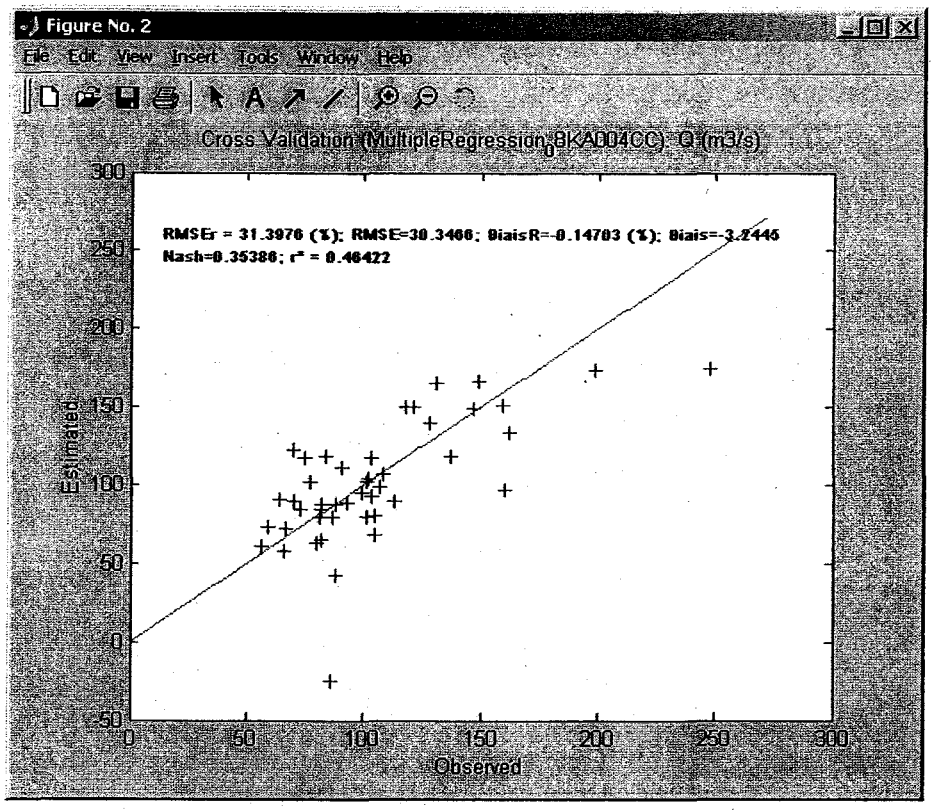


Fig.A.6. 12 : The cross validation scatter plot window

If the neural network model is selected in the popup menu, the calibration window is slightly different from the regressive model case (Fig.A.6. 13). The N parameter represents the number of iteration used by the neural network model in order to optimize the various stages of the neural algorithm. First, the user may start by detecting the possible outliers in the data set. In fact, based on the error estimation of N calibrations of the neural model, the "Detect outliers" option produces a list of the worst five observations (Fig.A.6. 14). The user can therefore select some or all observations to be removed definitively from the input data file. Prior to the neural model training itself, the user should optimize the sample splitting in three distinct groups (training, training and test groups) in order to obtain the best configuration. Indeed, among the N random sample subdivisions, the one leading to the lowest error on the test group is retained as the best group of subsets. The neural subsets are saved into a ".mat" file. The default file name has the "NeuralSubsets" prefix.

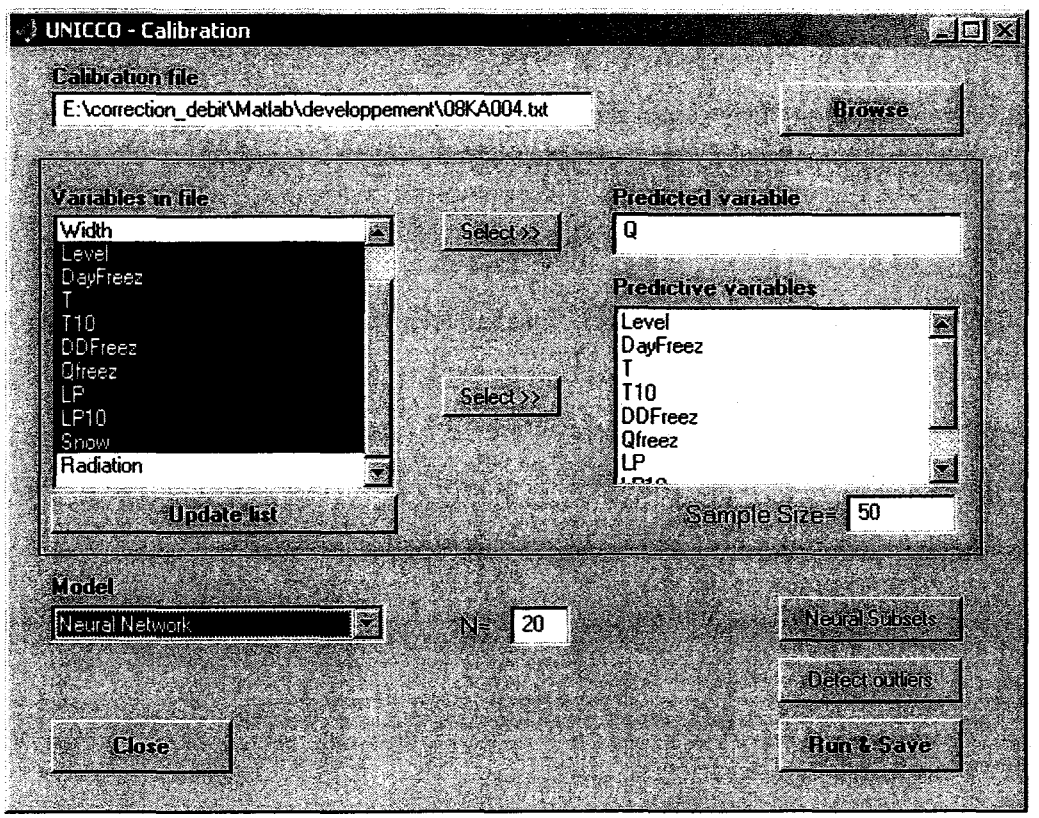


Fig.A.6. 13 : The calibration window in the case of the neural network model

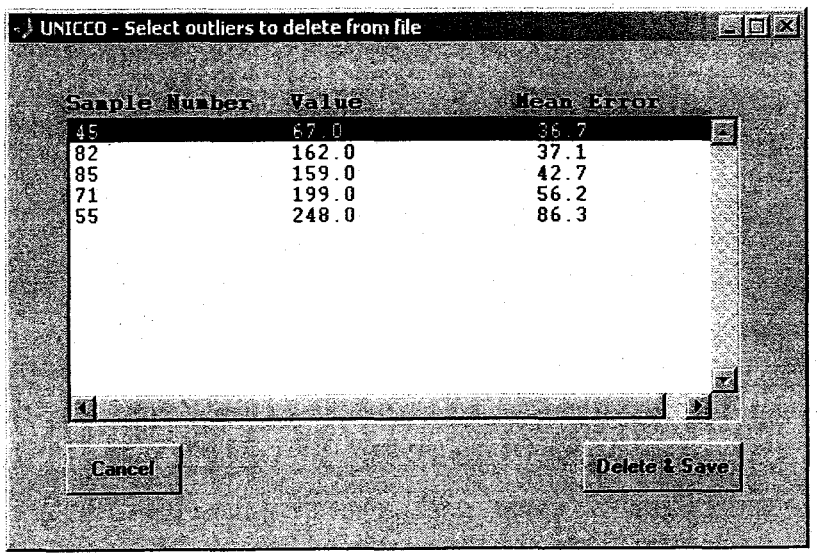


Fig.A.6. 14 : The worst observations window (detection of outliers)

The calibration of the neural model consists of N execution of the training program using the optimal neural subsets, obtained earlier. The training leading to the lowest test group estimation error is retained. The model parameters are thus saved in a ".mat" file to be used for later estimation of the variable of interest.

In the same way as the regressive model, the neural model calibration results are presented in a "Performances" window (Fig.A.6. 15). The model performances can also be exported to a text file. Here, the cross validation option is not available. However, at he same time and for comparison purposes, the regressive models (multiple, stepwise and ridge) are calibrated using the neural model calibration group and applied on the test group. The results are shown in the bottom of the model performances window.

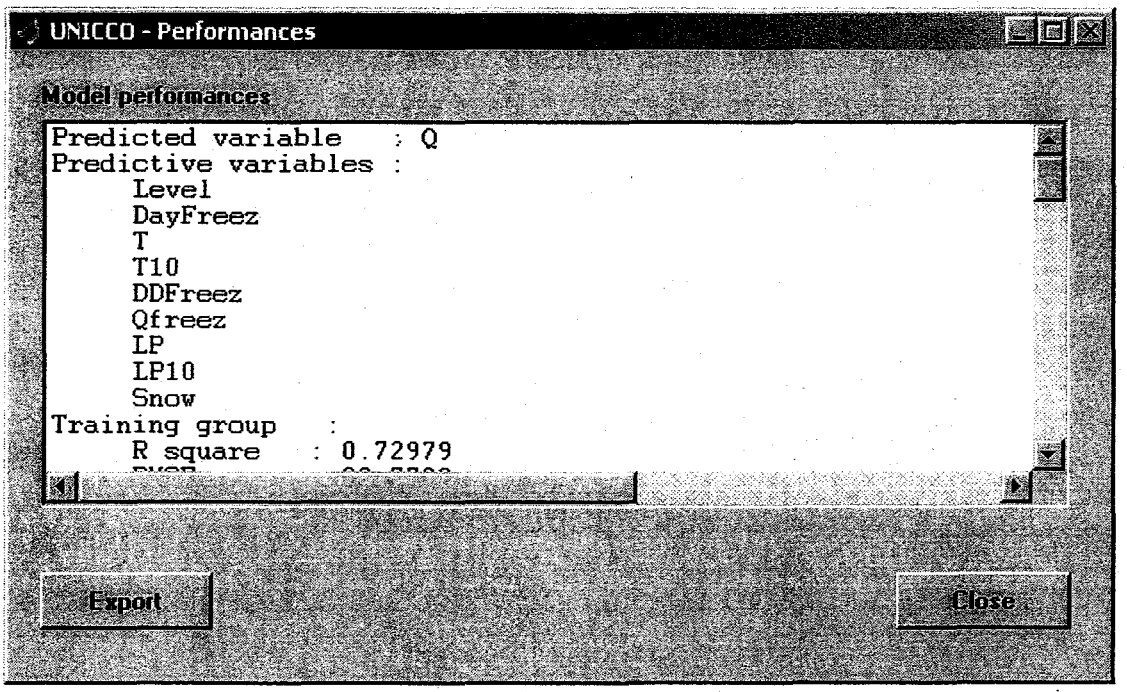


Fig.A.6. 15 : The neural network model performances window.

In addition to the textual form, the model performance is presented under graphical form. Thus, a scatter plot between the observed and estimated values is produced where the various sample subsets are represented using specific symbols (Fig.A.6. 16). Performance criteria for each group is also shown.

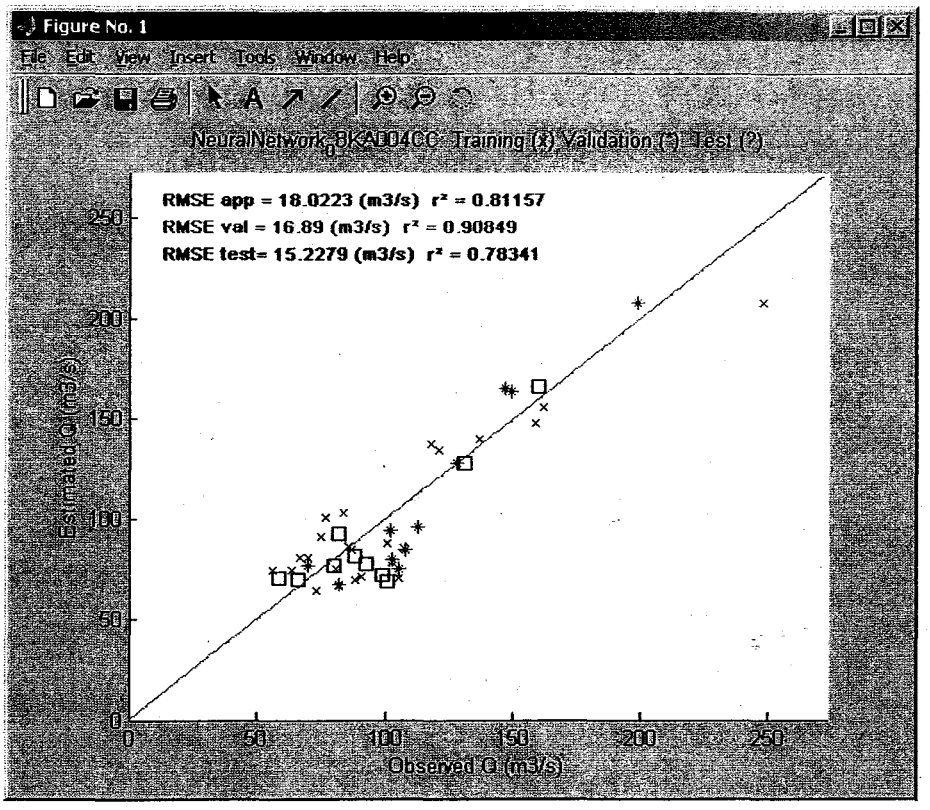
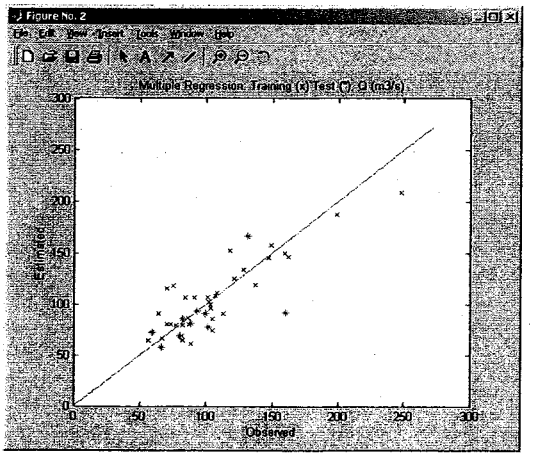
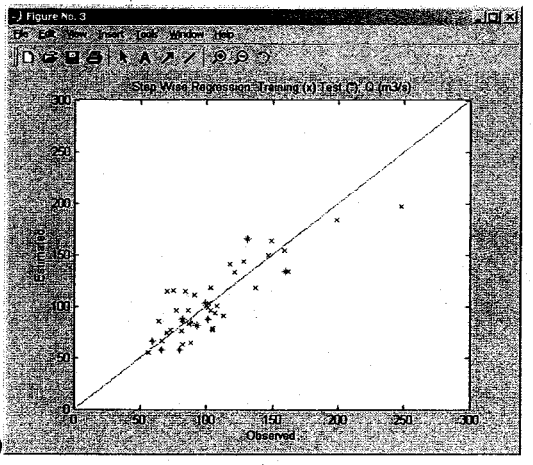


Fig.A.6. 16 : Neural Network model calibration scatter plot window

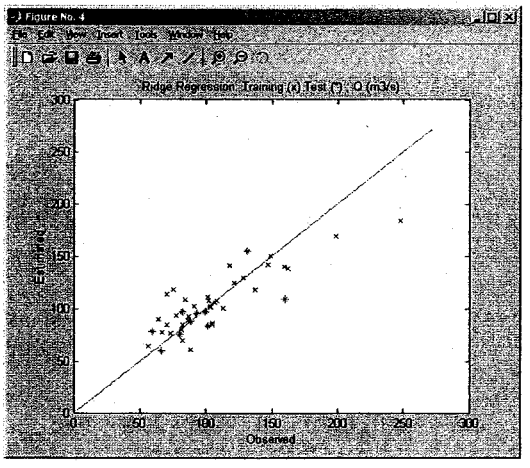
In addition, the regressive model results are plotted and each sample subset is identified using specific symbols (Fig.A.6. 17).



a)



b)



c)

Fig.A.6. 17 : The regressive models calibration scatter plots windows: a) multiple regression, b) stepwise regression and c) ridge regression.

2.3 ESTIMATION MODULE

The estimation module is used to estimate on real time basis the streamflow corrected for ice effect. The estimation is made using any of the models already calibrated and a set of explanatory variables for any continuous time period. The user should firstly load the model file (Fig.A.6. 18). At once, the model type as well as the dependent variables and the explanatory variables names (used in model calibration) are shown. Afterwards, the user should load the data file containing the explanatory variables observations either by entering manually the path and the file name or by browsing through the computer directories. The data file is also an ASCII text file and its content can be listed in the dedicated space using the "Update list" button. The variables can be picked using the "Add" button. The user has to add the explanatory variables in the same order in which they are listed in the model file. Finally, the dependent variable estimates are saved into a text file whose name is specified by the user.

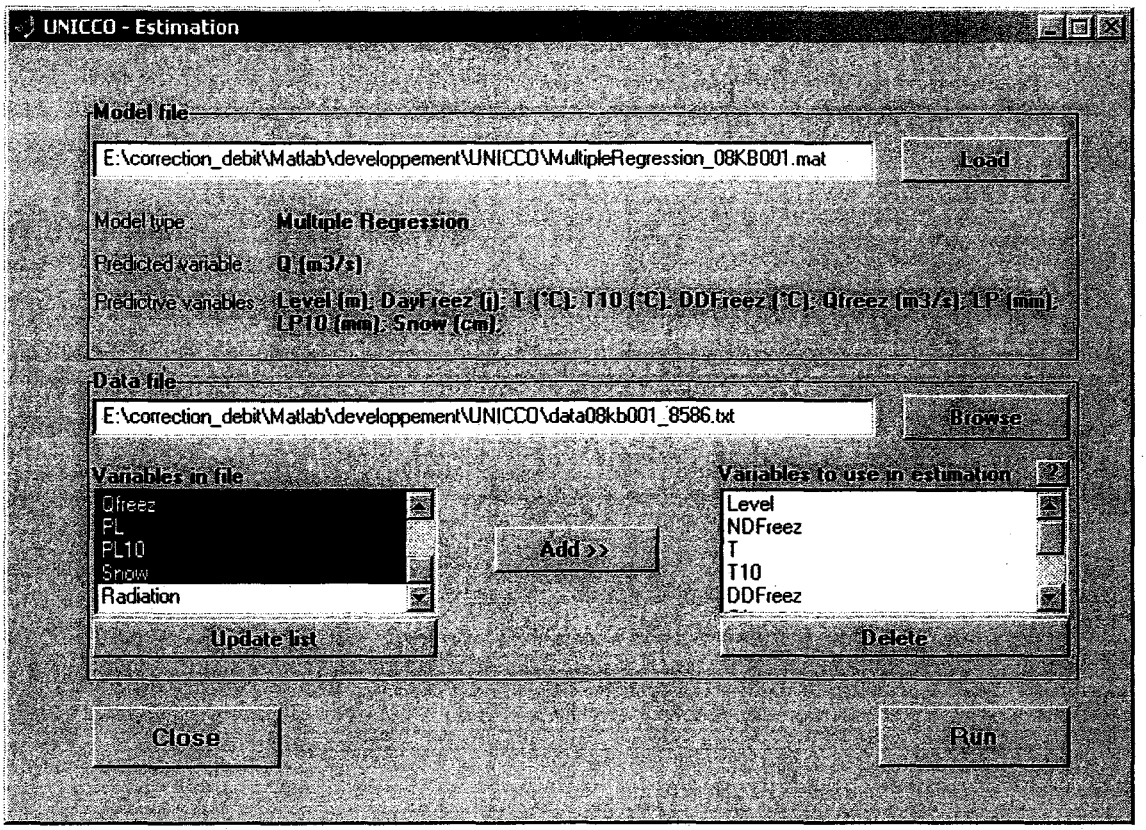


Fig.A.6. 18 : The estimation module window