Titre: Title:	Prévision des augmentations de turbidité à l'eau brute de la ville de Montréal par des réseaux de neurones artificiels
Auteur: Author:	Geneviève Tremblay
Date:	2003
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Tremblay, G. (2003). Prévision des augmentations de turbidité à l'eau brute de la ville de Montréal par des réseaux de neurones artificiels [Master's thesis, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/7303/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/7303/
Directeurs de recherche: Advisors:	Raymond Desjardins
Programme: Program:	Unspecified

UNIVERSITÉ DE MONTRÉAL

PRÉVISION DES AUGMENTATIONS DE TURBIDITÉ À L'EAU BRUTE DE LA VILLE DE MONTRÉAL PAR DES RÉSEAUX DE NEURONES ARTIFICIELS

GENEVIÈVE TREMBLAY
DÉPARTEMENT DES GÉNIES CIVIL, GÉOLOGIQUE ET DES MINES
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE CIVIL)
DÉCEMBRE 2003

Your file Votre référence
ISBN: 0-612-90861-5
Our file Notre référence
ISBN: 0-612-90861-5

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

PRÉVISION DES AUGMENTATIONS DE TURBIDITÉ À L'EAU BRUTE DE LA VILLE DE MONTRÉAL PAR DES RÉSEAUX DE NEURONES ARTIFICIELS

présenté par : TREMBLAY Geneviève
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

Mme MILLETTE Louise, Ph.D., présidente
M. DESJARDINS Raymond, M. Sc., membre et directeur de recherche
M. LABIB Richard, Ph.D., membre

À la vie, qui est belle avant tout...

REMERCIEMENTS

Je tiens à remercier mon directeur de recherche, M. Raymond Desjardins, et la titulaire de la Chaire CRSNG en eau potable, Mme Michèle Prévost, pour la confiance qu'ils m'ont accordée. Je tiens également à remercier Mme Louise Millette, directrice du département de génie civil, géologique et des mines et M. Richard Labib, professeur au département de mathématiques appliquées et de génie industriel, d'avoir accepté de juger ce mémoire. J'aimerais adresser des remerciements particuliers à M. Vincent Gauthier, chercheur invité à la Chaire en eau potable, de m'avoir guidé dans la réalisation de la première partie de ce projet et à M. Benoît Barbeau, associé de recherche à la Chaire en eau potable, pour ces judicieux conseils et son support lors de la complétion du projet et de la rédaction du mémoire. Je voudrais aussi remercier tout le personnel de la Chaire en eau potable ainsi que les autres étudiants pour leurs encouragements ainsi que les partenaires (CRSNG, Ville de Montréal, Ville de Laval, Triax, BPR, Vivendi Water, Groupe John Meunier) pour leur soutien financier.

Comme ce projet n'aurait pu être réalisé sans la collaboration de la Ville de Montréal, j'aimerais remercier M. Michel Gagné, ingénieur et directeur des usines de production d'eau potable, Mme Anne-Marie Bernier, chef de section du contrôle de la qualité de l'eau, M. Robert Millette, ingénieur à l'usine de production d'eau potable, Alain Champagne, microbiologiste. J'aimerais aussi adresser des remerciements à Jean-François Cantin, hydrologiste régional à Environnement Canada, et aux directeurs et aux opérateurs des stations de traitement de l'eau potable de Hawkesbury, de Coteau-du-Lac, de Salaberry de Valleyfield, de Grande-Île, de Sainte-Anne-de-Bellevue, de l'Île-Perrot, de la Communauté urbaine de l'Outaouais et de Laval qui m'ont fourni si généreusement les données nécessaires à la réalisation du projet.

Finalement, je tiens à remercier mon conjoint et mes amis pour le support moral et les encouragements fournis sans lesquels ce mémoire n'aurait jamais vu le jour.

RÉSUMÉ

Règle générale, la turbidité de l'eau brute à la prise d'eau de la Ville de Montréal est faible car le choix judicieux de l'emplacement de la prise d'eau dans le fleuve St-Laurent permet de disposer d'une eau brute globalement de très bonne qualité. Mais au printemps et à l'automne, des pointes de turbidité sont enregistrées à l'eau brute et les traitements de filtration directe réalisés actuellement aux deux usines de traitement ne permettent pas toujours de respecter le Règlement sur la Qualité de l'Eau Potable du Québec (2001). Pour pallier à ce problème, de nouveaux traitements de coagulation avec ou sans décantation devront être implantés aux deux stations de traitement de la Ville de Montréal. Comme ce type de traitement implique l'ajout d'une concentration variable de coagulant en fonction de la turbidité ou de la couleur de l'eau, il deviendrait très utile d'anticiper les augmentations soudaines de turbidité afin d'ajuster de façon optimale les traitements. Ce projet visait donc à prévoir les augmentations de turbidité de l'eau brute à la station de traitement DesBaillets de la Ville de Montréal. L'outil de modélisation retenu était le réseau de neurones artificiels, une technique empirique de plus en plus utilisée dans le domaine du génie de l'environnement. Les objectifs spécifiques de ce projet consistaient à (1) identifier et caractériser les événements responsables des variations de qualité d'eau brute à la prise d'eau de la Ville de Montréal et (2) développer un modèle de réseaux de neurones artificiels capable de prévoir l'occurrence et la magnitude des augmentations de turbidité.

Pour atteindre les objectifs du projet, il a fallu tout d'abord accroître la compréhension jusque-là limitée des augmentations de turbidité observées à l'eau brute à la Ville de Montréal et des facteurs explicatifs en lien avec ces dernières. Il a fallu aussi proposer une approche méthodologique permettant de développer des modèles de prévision neuronaux performants. L'acquisition de connaissances au sujet des augmentations de turbidité et de leurs causes a été effectuée par une analyse quantitative de données de turbidité représentatives et de différentes variables en lien avec les facteurs explicatifs

potentiels identifiés par une revue de littérature. Il s'est avéré que le phénomène d'inversion thermique, aussi dénommé renversement, ne joue qu'un rôle secondaire tant au printemps qu'à l'automne, contrairement à ce qui était présupposé avant la réalisation de ce projet de recherche. C'est plutôt l'augmentation de débit de la rivière des Outaouais résultant de la fonte des neiges qui joue le plus grand rôle au printemps. À l'automne, ce sont les tempêtes de vent qui expliquent la majorité des hausses très rapides de la turbidité.

L'approche méthodologique théorique a été développée à partir de démarches suggérées par des équipes de chercheurs oeuvrant dans le domaine de l'environnement et des ressources hydriques. Son élaboration s'est avérée nécessaire en raison de l'absence de protocole généralement accepté pour la construction de modèles utilisant des réseaux de neurones artificiels. L'approche méthodologique comprenait six étapes, soit l'identification des besoins à combler et des ressources disponibles, le choix du critère de performance, le développement et l'organisation de la base de données, la construction des modèles candidats et le choix du modèle final. Toutes les étapes de l'approche méthodologique se sont révélées être importantes pour obtenir un modèle de prévision performant et adéquat.

La mise en application de la méthodologie théorique a permis de développer trois modèles de prévision différents dont la performance est satisfaisante. Le premier est un modèle de régression (réseau de neurones de régression généralisée ou GRNN 8:740:2:1) prévoyant la différence entre la turbidité d'aujourd'hui et celle de demain. La justesse de ses prévisions, évaluée par le coefficient de corrélation de Pearson, était de 0,94 pendant les augmentations de turbidité et de 0,86 pour l'ensemble de l'année. Le second est un modèle de classification (perceptron multicouche 6 :5 :1) prévoyant la classe à laquelle appartient la turbidité du lendemain. La justesse de ses prévisions, évaluée par le taux de classification correcte, était de 91% pendant les augmentations de turbidité et de 86% pour l'ensemble de l'année. Le troisième modèle développé, le modèle fonctionnel, n'utilise pas de réseaux de neurones artificiels (RNA) mais intègre

les résultats des deux modèles connexionnistes. La décision de développer deux modèles de prévision connexionnistes prévoyant tous deux la turbidité du lendemain mais de façon différente s'est appuyée sur les besoins des opérateurs d'une station de traitement pour qui la fiabilité et la justesse des résultats est primordiale. En utilisant de façon complémentaire les deux modèles, la robustesse des prévisions a été accrue. Ainsi, en comparant les prévisions correctes et incorrectes pour les trois modèles, on constatait que la performance combinée des deux modèles au sein du modèle fonctionnel était meilleure (94,7%) que celle des deux modèles individuels, particulièrement dans le cas du modèle de régression (80% par rapport à 92,6% pour le modèle de classification). Mais même si sa performance était moindre, le modèle de régression demeurait intéressant car il permettait d'établir une tendance dans le temps comme une hausse graduelle et progressive au printemps. Il permet aussi de prévoir la magnitude des augmentations de turbidité.

Les variables utilisées dans les modèles de prévision retenus étaient peu nombreuses, les modèles de régression et de classification retenus utilisant respectivement huit et six variables d'entrée. Les variables portaient sur la qualité de l'eau brute à la station DesBaillets et sur les deux causes explicatives principales des augmentations de turbidité, soit le débit de la rivière Outaouais et les vents de la région de Montréal. La seule variable d'index retenue, la variable de saison, a probablement permis le développement de modèles de prévision pour l'ensemble de l'année plutôt que par saison.

Ce projet de recherche a prouvé l'utilité des réseaux neuronaux pour développer relativement facilement des outils de prévision performants et utiles. Ces modèles pourraient être éventuellement implantés en ligne à la station de traitement DesBaillets de la Ville de Montréal. Grâce à l'approche méthodologique novatrice proposée, ce projet de recherche a ouvert la voie à plusieurs autres applications possibles dans le domaine de la prévision de paramètres de qualité de l'eau brute ou de l'eau traitée dans les différentes filières de traitement.

ABSTRACT

In general, the turbidity of Montreal raw water is low because of the judicious localization of the water intake in the St.Laurence River, which supplies globally high quality water. But during spring and fall, turbidity peaks are observed in raw water and the direct filtration treatment performed at the two treatment plants of Montreal do not always allow the respect of the new *Règlement sur la Qualité de l'Eau Potable du Québec* (2001). To overcome this problem, new coagulation treatments will have to be implanted at the two water treatment plants. Since the type of treatment would require the addition of a variable concentration of coagulant according to the turbidity or the color of the raw water, it would become very useful to anticipate the sudden turbidity peaks to optimally adjust the treatments. Hence, the main objective of this research was to forecast the turbidity peaks observed at the raw water intake of the DesBaillets treatment plant of the City of Montreal. The modelling tool used was the articificial neural network, an empirical technique starting to be used more frequently in the environmental engineering field. The specific objectives of this project were to (1) identify and characterize the different factors potentially related to the turbidity peaks at the Montreal water intake and (2) to develop an artificial neural networks model capable of forecasting the occurrence and the magnitude of the turbidity peaks at the DesBaillets treatment plant of the City of Montreal.

In order to reach the objectives of the project, it was first necessary to increase the limited knowledge of the turbidity peaks observed at the water intake of the City of Montreal and of the explicative factors related to them. It was also required to elaborate a methodological approach in order to develop performing forecasting neural networks models. The acquisition of knowledge about turbidity peaks and related explanatory events was achieved through a quantitative analysis of representative turbidity and related events data. The events were identified previously by a literature review. It turned out that the temperature inversion, also called turnover, played only a secondary

role in explaining turbidity peaks in spring and fall, contrary to what was supposed prior to this research project. Instead, it was the increase in the Outaouais River flow resulting from the spring snowmelt that proved to be the leading explanatory factor during springtime. During fall, the wind storms explained most of the sudden turbidity peaks.

The theoretical methodology was elaborated from the approaches adopted by research teams working in the environmental and water resources fields. This work has proven to be necessary because of the absence of generally accepted protocol for the construction of models based on artificial neural networks (Baxter *et al.,* 2001). The methodology is made of six main steps: the identification of the needs and the assessment of available resources, the choice of the performance criteria, the development and organization of the database, the construction of neural networks models and the final model choice. All the steps were important to obtain a performing and adequate forecasting model.

The use of the theoretical methodology has allowed the development of three different forecasting models for which the performance was satisfying. The first one was a regression model (generalized regression neural network or GRNN 8:740:2:1) forecasting the difference between today's and tomorrow's turbidity. The accuracy of the predictions, measured by the Pearson r correlation, was 0.94 during the turbidity peaks and 0.86 for the year in general. The second model is a classification one (multilayer perceptrons 6:5:1) that forecasted the class to which belongs tomorrow's turbidity. The accuracy of the prediction, measured with a correct classification rate, was of 91% during the turbidity peaks and 86% for the year in general. The third model, the functional model, did not use artificial neural networks. Instead, it was integrating the results of the two other neural networks. The decision to develop two models forecasting tomorrow's turbidity but in two different ways relied on the needs of the operators at the water treatment plant, for whom the reliability and the accuracy of results is fundamental. By using in a complementary manner the two models in a

functional model, the robustness of the predictions was enhanced. So, by comparing the correct and incorrect predictions of the three models, the combined performance of the two models is better (94.7%) than the individual models, particularly in the case of the regression model (80% against 92.6% for the classification model). But event though the performance of the regression model is less, it remains interresting because it allows to identify trends in time like a gradual increase of turbidity over spring. Moreover, it allows forecasting the turbidity peaks magnitude.

Few inputs were finally used to develop the forecasting models, the regression model using 8 inputs and the classification model, 6 inputs. The inputs were related to the raw water quality at the DesBaillets water treatment plant and to the two main explanatory factors of the turbidity peaks, being the Outaouais river flow and the winds in the Montreal region. The only index parameter udes, the season index, probably allowed the development of a model for the year instead of per season.

This research project has proven the usefulness of artificial neural networks to develop quite easily performing and useful forecasting models. Those models could eventually be implemented in line at the DesBaillets water treatment plant of the City of Montreal. Thanks to the innovative methodology proposed and tested, this project is leading the way to a number of other applications of raw water and treated water quality forecasting. The artificial neural networks represent quite new tools, but the drinking water industry has everything to gain in using them to improve the quality of water dirstributed to consumers.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

AA	algorithme d'apprentissage
ACP	analyse en composantes principales
AG	algorithmes génétiques
ANOVA	analyse de variance
ARMA	*autoregressive moving average*
ARP	algorithme de rétro-propagation
B/H	rapport entre la largeur et la profondeur
°C	degré Celcius
CE	construction par étape de réseaux bi-variables
COT	carbone organique total
GN-RNA	combinaison d'algorithmes génétiques et de réseaux de neurones artificiels
GRNN	réseau de neurones de regression généralisée ou *generalized regression neural network*
ha/an	hectare par année
Hz	hertz
IPS	*intelligent problem solver*
km	kilomètre
km/h	kilomètre par heure
PMC	Perceptron multi-couches
m	mètre
$mg \cdot L^{-1}$	milligramme par litre
m^3/s	mètre cubique par seconde
MES	matière en suspension
$ML \cdot d^{-1}$	méga-litre par jour
mm	millimètre
NARMA	*nonlinear autoregressive moving average*
N_C	nombre de neurones de la couche cachée
N_E	nombre de neurones de la couche d'entrée
N_{EA}	nombre d'échantillons d'apprentissage
PB	design expérimental plackett-Burman

PNN	réseau de neurones probabiliste ou *probabilistic neural network*
r	coefficient de corrélation de Pearson
R^2	coefficient de détermination
RBF	réseau à fonction radiale ou*radial basis function*
RMSE	erreur quadratique moyenne ou *root-mean-square error*
RNA	réseau de neurones artificiels
RNN	réseau de neurones naturels
SOM	*self-organizing map*
STD	écart-type
UC	unité de couleur vraie
$UC \cdot d^{-1}$	taux de chagement de couleur
UTN	unité de turbidité néphélémétrique

LISTE DES ANNEXES

INTRODUCTION

La santé humaine dépend grandement de la qualité de l'eau potable consommée. Selon l'Organisation Mondiale de la Santé, 80% de toutes les maladies affectant la population des pays en voie de développement sont dues à la mauvaise qualité de l'eau potable (UNICEF, WHO and Water Supply and Sanitation Collaborative Council, 2000). Les organismes pathogènes trouvés dans l'eau sont responsables de maladies telles que la diarrhée, qui cause environ 2,2 millions de décès par année, surtout parmi les enfants en bas âge, les vers intestinaux, qui affectent environ 10% de la population des pays en voie de développement, le trachome, responsable de la cécité de 6 millions de personnes, la bilharziose, qui affecte 200 millions de personnes et le choléra, qui sévit souvent suite à des catastrophes naturelles ou à la guerre (UNICEF, WHO and Water Supply and Sanitation Collaborative Council, 2000). Dans les pays industrialisés, des séquelles sérieuses, incluant la mort, sont rarement associées aux maladies causées par les microorganismes présents dans l'eau potable. Néanmoins, les éclosions et les épidémies sont suffisamment nombreuses pour nous rappeler qu'il y a un risque bien réel. Au Canada seulement, entre 1986 et 1993, environ 150 éclosions infectieuses liées à l'eau potable ont été rapportées au Laboratory Center for Disease Control (Aramini *et al.*, 2000). Aux Etats-Unis, en 1995-1996, 22 épidémies associées à l'eau potable ont été documentées, affectant environ 2 567 personnes (Aramini *et al.*, 2000). En plus de la perte de confiance de la population envers l'eau consommée, les coûts directs, tels les soins de santé, et indirects, tels l'absentéisme au travail, sont considérables.

La présence de microorganismes dans l'eau potable est intimement liée à la concentration des particules en suspension dans l'eau brute et dans l'eau traitée. Les microorganismes présents dans l'eau potable peuvent être fixés à la surface de particules en suspension, au sein d'amas de bactéries ou transiter dans l'appareil digestif des micro-invertébrés ou des protozoaires, considérés eux-mêmes comme des particules en suspension en raison de leur taille (Gauthier, 1998). Les problèmes découlant de l'association entre particules et microorganismes sont atténués par le fait que,

normalement, une très grande proportion des particules en suspension dans l'eau est éliminée par les usines de potabilisation par les procédés de coagulation, de décantation, de filtration et de désinfection. Malheureusement, il arrive que ce ne soit pas le cas lors d'un mauvais fonctionnement de l'usine de traitement ou lors de l'accroissement soudain de la turbidité de l'eau brute, par exemple. Lorsque la concentration de particules en suspension dans l'eau traitée est élevée, le risque que des organismes pathogènes pénètrent dans le réseau augmente parfois considérablement, qu'ils soient associés ou non à des particules (Aramini *et al.*, 2000). Il est possible alors que les consommateurs soient exposés à des microorganismes pathogènes, comme ce fut le cas à Milwaukee, en 1993. En raison d'une déficience dans l'étape de coagulation, l'usine de traitement ne parvenait pas à bien éliminer les particules en suspension de l'eau brute, dont des oocystes de *cryptosporidium*. Cette épidémie a résulté en 400 000 cas de cryptosporidiose, 4000 hospitalisations et 104 décès (Morris *et al.*, 1996).

Des normes strictes régissent le traitement de l'eau potable afin de minimiser les risques présentés par les microorganismes. La sévérité des normes dépend de la qualité de l'eau brute. Au Québec, le Règlement sur la Qualité de l'Eau Potable du Québec (2001) stipule que si les eaux de surface à potabiliser sont bien protégées, de très bonne qualité et surtout peu turbides, peu chargées en bactéries coliformes fécales et totales et en carbone organique total (COT), l'élimination physique de la matière particulaire par un traitement minimal de filtration n'est pas requise (Ministère de l'Environnement du Québec, 2001). La désinfection reste cependant nécessaire pour atteindre les objectifs d'élimination des virus, des kystes de *Giardia* et des oocystes de *Cryptosporidium* fixés par le règlement. Pour les stations de traitement ayant obtenu des dispenses de filtration, des variations de la charge particulaire de l'eau sont acceptées tant que la turbidité ne dépasse pas 5 UTN et soit inférieure à 1 UTN, 90% du temps. Pour les stations mettant en œuvre un traitement de filtration, la turbidité de l'eau filtrée ne doit pas excéder la valeur de 0,5 UTN pour 95% des valeurs mesurées aux 4 h durant 30 jours consécutifs. Des augmentations de la turbidité de l'eau filtrée sont tolérées, pour

autant qu'elles n'excèdent pas 36 heures par mois et que la turbidité soit toujours inférieure à 5 UTN.

À Montréal, le fleuve St-Laurent est la source d'eau brute pour la production d'eau potable. Règle générale, la turbidité de l'eau brute est faible car le choix judicieux de l'emplacement des prises d'eau permet de disposer d'une eau brute globalement de très bonne qualité. Néanmoins, à certaines périodes de l'année, au printemps et à l'automne, des pointes de turbidité sont enregistrées à l'eau brute : comprise normalement entre 0,3 et 3 UTN, la turbidité peut atteindre des valeurs aussi « élevées » que 10 à 30 UTN. Les traitements de filtration directe (sans coagulation, ni décantation) réalisés aux deux usines de traitement éliminent une grande partie de cette charge particulaire, mais la turbidité de l'eau traitée peut temporairement être supérieure à 0,5 UTN. De tels événements de turbidité modérés et transitoires ont un impact sur la qualité de l'eau dans le réseau de distribution (selon l'article de Gauthier *et al.* 2003, reproduit à l'Annexe 1). Les analyses microbiologiques réalisées sur l'eau n'ont pas permis de retrouver d'indicateurs de contamination (coliformes, spores de *Clostridium perfringens*) dans l'eau traitée et distribuée au moment des augmentations de turbidité. Néanmoins, il a aussi été trouvé que la concentration de spores de bactéries aérobies est bien corrélée avec la turbidité de l'eau brute. Présentement, lors des augmentations rapides de turbidité, le traitement de filtration en place à Montréal peut difficilement être ajusté pour maintenir la turbidité en-deça de 0,5 UTN. Afin de respecter le Règlement sur la Qualité de l'Eau Potable du Québec (2001), des traitements d'appoint ou permanents tels que la coagulation et la flocculation devront probablement être ajoutés aux deux stations de traitement de la Ville de Montréal. Ce type de traitement implique l'ajout d'une concentration variable de coagulant en fonction de la turbidité ou de la couleur de l'eau. Il deviendrait alors très utile d'anticiper les augmentations soudaines de turbidité afin d'ajuster de façon optimale les traitements ou de mettre en fonction les traitements d'appoint au moment opportun. Ce projet vise donc à prévoir les variations de turbidité de l'eau brute à la station de traitement DesBaillets de la Ville de Montréal. L'outil de modélisation consiste en les réseaux de neurones artificiels, une

technique empirique de plus en plus utilisée dans le domaine du génie de l'environnement. Afin d'améliorer la performance des réseaux de neurones artificiels, les différents événements possiblement liés aux augmentations de turbidité à la prise d'eau de la Ville de Montréal doivent être identifiés et caractérisés au préalable. Les objectifs spécifiques de ce projet consistent donc à (1) identifier et caractériser les événements responsables des variations de qualité d'eau brute à la prise d'eau de la Ville de Montréal et (2) développer un modèle de réseaux de neurones artificiels capable de prévoir l'occurrence et la magnitude des augmentations de turbidité à la station de traitement DesBaillets de la Ville de Montréal.

Le premier chapitre de ce mémoire présente la revue de la littérature et comporte deux sections principales. La première présente les différents événements potentiellement responsables ou liés aux variations de turbidité. Ces événements comprennent les fluctuations des débits de la rivière des Outaouais et du fleuve St-Laurent, le mélange de ces masses d'eau, les inversions thermiques printanière et automnale, les tempêtes de vent et les fortes pluies. La seconde section porte sur les réseaux neuronaux. Les caractéristiques générales et les notions de base du fonctionnement des réseaux neuronaux y sont tout d'abord introduites, avant d'examiner la pertinence de recourir aux réseaux de neurones artificiels pour ce projet. La section se clot sur une synthèse des approches méthodologiques utilisées pour le développement de modèles et des exemples d'applications dans le domaine de l'environnement et des ressources hydriques. Le deuxième chapitre reprend les causes des variations de turbidité identifiées dans la revue de littérature afin d'examiner la concordance entre l'occurrence de différents événements en amont et les augmentations de turbidité observées à la prise d'eau de la Ville de Montréal. Les événements turbides y sont caractérisés et la synergie pouvant exister entre les différents événements en amont est examinée. Le chapitre suivant présente les travaux nécessaires à la construction d'une banque de données, à l'ajustement des paramètres du réseau neuronal ainsi qu'à l'évaluation des modèles développés. Enfin, une discussion et conclusion générale constituent les derniers chapitres de ce mémoire.

CHAPITRE 1 - REVUE DE LITTÉRATURE

Prévoir les augmentations de turbidité dans l'eau brute de la Ville de Montréal constitue une tâche complexe car les augmentations observées résultent de l'occurrence de plusieurs événements sporadiques ou saisonniers agissant en parallèle. Afin d'identifier et de comprendre ces événements et ensuite de prévoir les augmentations de turbidité de l'eau brute, ce chapitre présente une brève revue de littérature des différents événements en lien avec les variations de qualité d'eau du tronçon fluvial en amont de Montréal, suivie d'une revue présentant les réseaux de neurones artificiels et leur utilisation dans le domaine du génie de l'environnement.

1.1 Événements à l'origine des augmentations de turbidité de l'eau brute du fleuve St-Laurent

Les fluctuations saisonnières du débit et le mélange des masses d'eau de qualité différente du fleuve St-Laurent et de la rivière des Outaouais, l'inversion thermique dans les lacs du tronçon fluvial, les tempêtes de vent et les fortes pluies sont autant d'événements qui peuvent contribuer, seuls ou en synergie, à augmenter la turbidité de l'eau en transportant ou en remettant en suspension des particules minérales ou organiques.

1.1.1 Mise en contexte

La prise d'eau de la Ville de Montréal est située à 610 m de la rive nord du fleuve, à LaSalle, en amont des rapides de Lachine et en aval du lac St-Louis (Figure 1-1). Le lac Saint-Louis est formé par un élargissement naturel du fleuve à sa confluence avec la rivière des Outaouais. Il a une forme triangulaire, étant constitué à la base par les principaux affluents, et au sommet, par l'exutoire via les rapides de Lachine. Il couvre une superficie de 148 km^2, soit environ 23 km de longueur par 10 km de largeur dans ses plus grandes dimensions (Centre Saint-Laurent, 1993 et Fortin *et al.* 1994).

Le lac Saint-Louis reçoit les eaux du lac Saint-François, qui s'écoulent en grande partie par le canal de Beauharnois (84% du débit en moyenne) le long de la rive sud et par le lit naturel du fleuve en bordure de la rive nord (Fortin *et al.* 1994). Cette section du fleuve reçoit aussi les eaux de la rivière des Outaouais par l'intermédiaire du lac des Deux Montagnes qui se décharge en partie dans le lac Saint-Louis par le canal de Vaudreuil, à l'ouest de l'île Perrot, et par le canal Sainte-Anne, au nord de l'île Perrot (Fortin *et al.* 1994).

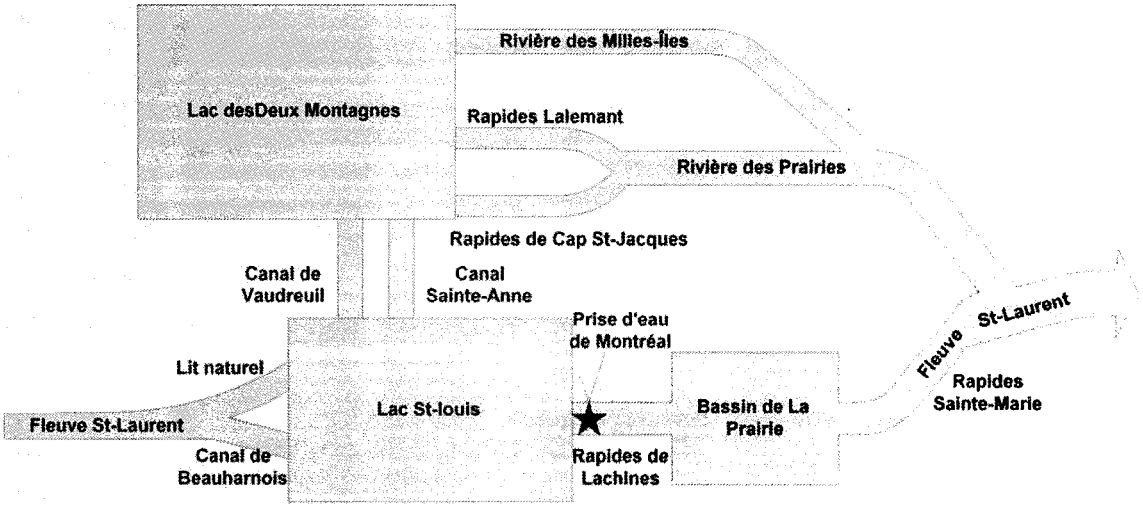


Figure 1-1 : Schéma du réseau hydraulique de la région de Montréal

1.1.2 Fluctuations des débits

Le débit du fleuve Saint-Laurent est régularisé suivant un plan géré par la commission mixte internationale, un organisme bilatéral responsable de l'application du traité relatif aux eaux limitrophes entre le Canada et les États-Unis. Les débits présentent des périodes successives de forte et de faible hydraulicité suivant un cycle dont la durée s'étend sur plusieurs années (Hydro-Québec, 1985a). La fluctuation à court terme du débit du fleuve Saint-Laurent est régularisée naturellement par la rétention lacustre des Grands-Lacs et artificiellement par les barrages Iroquois, Moses-Saunders et Long-Sault en amont de Cornwall. Ainsi, la crue printanière du fleuve Saint-Laurent, contrairement à celle des autres cours d'eau, s'étend sur plusieurs mois et a peu d'effet sur le débit

moyen mensuel. Par exemple, à l'exutoire Beauharnois, le débit moyen mensuel fluctue entre 5475 m^3/s au mois de décembre et 5700 m^3/s au mois de juin. À l'exutoire de Pointe-aux-Cascades, le fleuve qui emprunte son lit naturel est moins régularisé et le débit moyen mensuel varie entre 1029 m^3/s au mois de novembre et 1846 m^3/s au mois d'avril (période d'observation de 1949 à 1983, Hydro-Québec, 1985a).

Par contre, le débit du fleuve à la sortie du lac Saint-Louis augmente sensiblement au printemps. Alors que le débit mensuel moyen du fleuve à l'entrée du lac fluctue de 6,4% de mars à juin, il fluctue de 15% à la sortie. Cette augmentation s'explique par le fait que le débit du fleuve à la sortie du lac St-Louis est fortement influencé par celui de la rivière des Outaouais. Le débit moyen de la rivière Outaouais mesuré au Barrage de Carillon est d'environ 2000 m^3/s (période d'observation de 1962 à 1989) (Fortin *et al.* 1994), mais il peut varier entre des extrêmes saisonniers de 306 m^3/s en période d'étiage (7 sept. 1971) et de 8190 m^3/s en période de crue (4 avril 1971) (Hydro-Québec, 1985a). Le débit journalier de l'Outaouais a même atteint 9230 m^3/s en avril 1951, soit une valeur équivalente au débit annuel moyen du fleuve Saint-Laurent à LaSalle (Frenette *et al.*, 1989).

L'accroissement des débits de la rivière des Outaouais et du fleuve St-Laurent au printemps dégrade la qualité de l'eau puisée à la prise d'eau de la Ville de Montréal, particulièrement au printemps. Frenette et Frenette (1992) ont déterminé les périodes sédimentologiques actives du fleuve et de ses tributaires à l'aide de sédimentogrammes. On remarque ainsi une pointe sédimentologique lors des crues printanières des tributaires (avril-mai) et une seconde pointe, de moindre importance, lors des crues automnales (octobre-novembre). Entre ces maxima, des pointes tertiaires associées aux différentes averses d'été et d'automne (juin-septembre) apparaissent, tandis que les charges solides d'hiver demeurent très faibles (décembre-mars). En tout, la période de crue printanière serait responsable de 60% et 70% de la charge sédimentaire.

C'est la charge sédimentaire de la rivière des Outaouais qui contribue le plus significativement à l'augmentation des apports en matières solides au lac St-Louis,

particulièrement au printemps car les variations saisonnières des concentrations de particules en suspension sont aussi beaucoup plus prononcées dans la rivière des Outaouais que dans le fleuve St-Laurent (SCN-Procéan, 1992). La contribution est exacerbée par le fait que les rapides de Sainte-Anne-de-Bellevue et de Vaudreuil dans le lac Saint-Louis évacuent environ 45% du débit de l'Outaouais en période de crue, ce qui représente environ 30% du débit total du Saint-Laurent (Hydro-Québec, 1985a). En période d'étiage, les rapides évacuent plutôt 25% du débit de la rivière, ce qui représente à peine 2% du débit total du fleuve.

La plus grande force exercée par l'eau lors de l'accroissement des débits du fleuve mais surtout de la rivière des Outaouais remet en suspension des particules s'étant déposées sur le lit des cours d'eau et des lacs fluviaux au cours de la fin de l'automne et de l'hiver. Fortin *et al.* (1994) ont constaté, dans leur étude, que la classe dominante dans le lac St-Louis est celle de sable et gravier, ce qui indique qu'une partie importante du lac présente des conditions d'érosion des particules fines, qui ont été remises en suspension (Figure 1-2). L'érosion des rives et, dans une moindre mesure, celle du lit du fleuve et de la rivière des Outaouais par la force de cisaillement exercée par l'eau en mouvement contribuerait aussi à la charge sédimentaire printanière (Loiselle *et al.*, 1997).

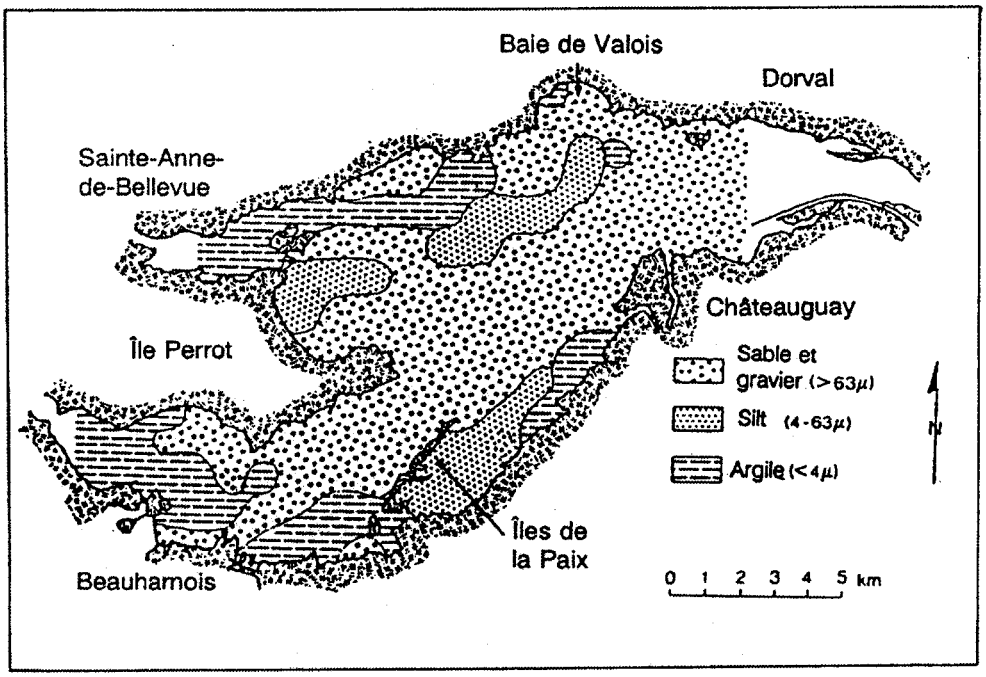
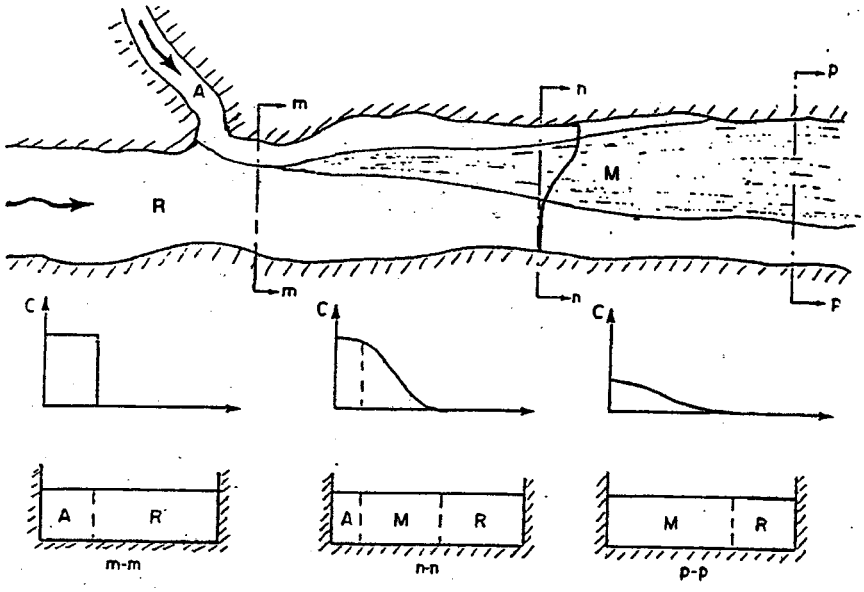


Figure 1-2 Granulométrie des sédiments du Lac St-Louis
(Source : Fortin *et al.*, 1994)

1.1.3 Mélange des masses d'eau

Les eaux du Saint-Laurent ne représentent pas une masse homogène uniforme. Les eaux du fleuve sont formées d'une série de couloirs contiguës aux qualités distinctes désignés par « masses d'eau ». Même si la vitesse des affluents peut être grande, la force d'impact des affluents est très faible par rapport à l'inertie du fleuve (Verrette, 1990). Ainsi, dès leur arrivée dans le fleuve, les affluents comme la rivière Outaouais sont rabattus sur les rives avant de poursuivre leur parcours vers l'aval tout en se mélangeant graduellement avec les eaux adjacentes (Figure 1-3). Malgré les capacités élevées de mélange et de diffusion des eaux du fleuve, le mélange des affluents nécessite de très longues distances étant donné, en particulier, la valeur très élevée du rapport largeur-profondeur (B/H). En effet, la profondeur du fleuve est très faible relativement à la largeur. Par exemple, à la hauteur du lac St-Pierre, le rapport B/H peut atteindre 2500 (Verrette, 1990). La position des masses d'eau et le mélange de celles-ci dépendent aussi de la fluctuation des débits selon les saisons, de l'alternance des lacs,

des rapides, des battures et des îles (Hydrotech, 1988), de même que par le cycle de croissance des plantes aquatiques et le vent (Fortin *et al.*, 1994).



A : eaux de l'affluent; R : eaux de la rivière; M : eaux du mélange

Figure 1-3 : Schématisation du mélange d'un affluent au Saint-Laurent (Source : Verrette, 1990)

Le lac Saint-Louis est baigné par deux grandes masses d'eau: les eaux des Grands Lacs et celles de la rivière des Outaouais. Les eaux du fleuve sont vertes et bien minéralisées et s'écoulent principalement dans la voie maritime et près de la rive sud. Ces eaux sont peu chargées en particules en suspension, contrairement aux eaux de la rivière des Outaouais, brunes et peu minéralisées qui s'écoulent le long de la rive nord (Centre Saint-Laurent, 1993 et Fortin *et al.*., 1994). En tenant compte de la conductivité, de la dureté totale et de la fluorescence naturelle, trois indicateurs utilisés pour identifier les masses d'eau, le nombre de masses d'eau est porté à quatre avec l'ajout d'une zone de transition entre les eaux de l'Outaouais et celles des Grands Lacs le long de la rive nord et de la gyre du lac Saint-Louis observée au sud de l'île Perrot (Figure 1-4).

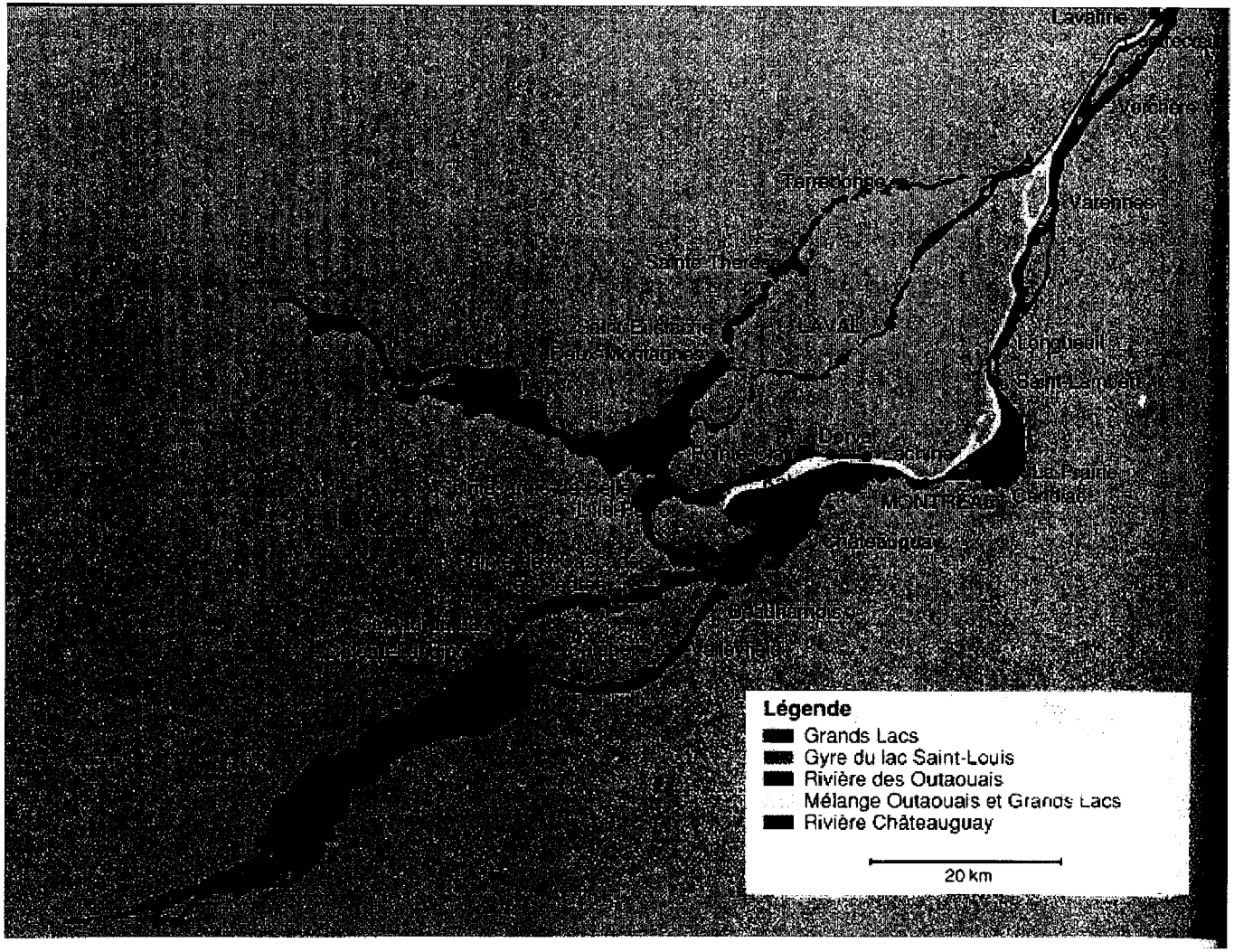


Figure 1-4 : Emplacement des quatre masses d'eau dans l'ensemble du couloir fluvial du lac Saint-François aux îles de Sorel et des prises d'eau dans le fleuve Saint-Laurent, le long du couloir fluvial (Source : Lamarche, 1992)

La formation de la gyre est un phénomène hydrodynamique résultant de la confluence des eaux de canal de Beauharnois avec celles de l'Outaouais (Fortin *et al.*, 1994). Alors que les eaux de la rivière Outaouais pénétrant le lac Saint-Louis par le canal Sainte-Anne-de-Bellevue longent la rive nord en un étroit couloir, les eaux pénétrant par le chenal de Vaudreuil sont confrontées à la grande masse des eaux du Saint-Laurent sortant de la centrale Beauharnois, ce qui a pour effet d'emprisonner les eaux de Vaudreuil dans ce secteur et d'en favoriser le mélange avec celles du Saint-Laurent.

Les masses d'eau de l'Outaouais, de la bande de transition et de la gyre peuvent dégrader la qualité de l'eau brute de la Ville de Montréal si elles atteignent la prise d'eau. Selon des simulations effectuées par Hydro-Québec (1985a), les eaux de l'Outaouais et celles de la bande de transition entre les deux masses d'eau évitent donc dans les deux cas la prise d'eau de la Ville de Montréal, située à 610 m de la rive nord car la largeur des eaux de l'Outaouais provenant du canal Sainte-Anne-de-Bellevue ne dépasse 250 m lors de condition moyenne (débit du fleuve de 8 400 m^3/s à l'entrée du bassin LaPrairie) et 390 m en condition de forte crue (débit du fleuve de 14 000 m^3/s à l'entrée du bassin LaPrairie) à la hauteur du pont Mercier, juste en amont de la prise d'eau de la Ville de Montréal (Figure 1-5 et Figure 1-6).

Les eaux provenant du chenal Vaudreuil ont un comportement tout autre. Mélangées aux eaux du fleuve dans la gyre du lac Saint-Louis en conditions moyennes, elles dégradent plus ou moins la qualité de l'eau puisée à la prise d'eau de Montréal, dépendamment du débit et de la qualité des deux cours d'eau (Hydro-Québec, 1985a). En condition de forte crue, le comportement des eaux de l'Outaouais provenant du canal Vaudreuil change. La masse d'eau ne se mélange plus aux eaux du fleuve mais longe le sud de l'île Perrot, s'ajoutant aux eaux provenant de Sainte-Anne-de Bellevue (Figure 1-6). Le couloir d'écoulement des eaux de l'Outaouais s'élargit, mais, selon les simulations effectuées par Hydro-Québec (1985a), ce comportement des masses d'eau n'affecteraient toujours pas la prise d'eau de la Ville de Montréal. Ce pourrait par contre être le cas pour des débits du fleuve supérieurs à 14 000 m^3/s.

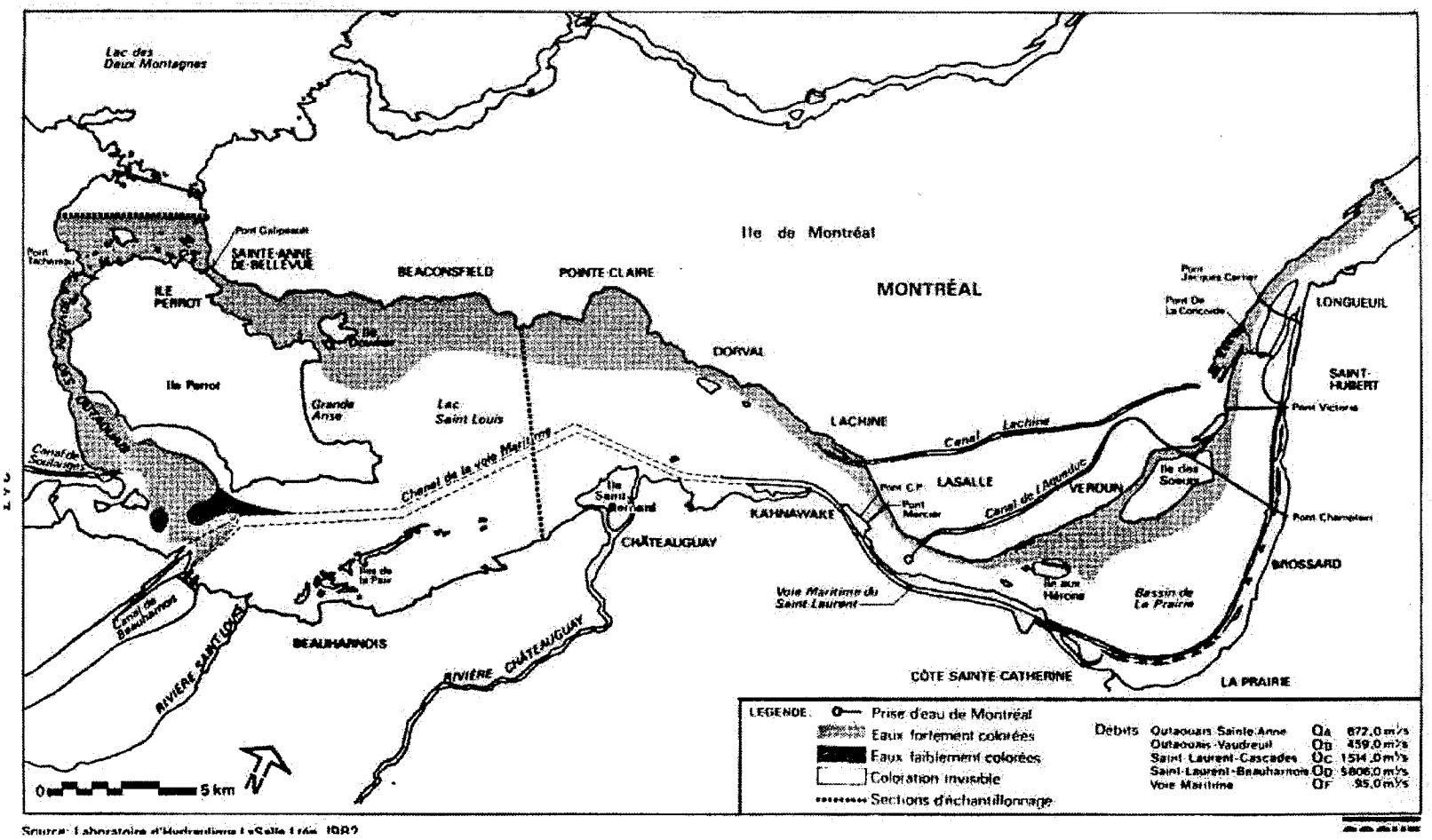


Figure 1-5 : Mélange des eaux au débit moyen de 8 400 m^3/s) (Source : Hydro-Québec, 1985a)

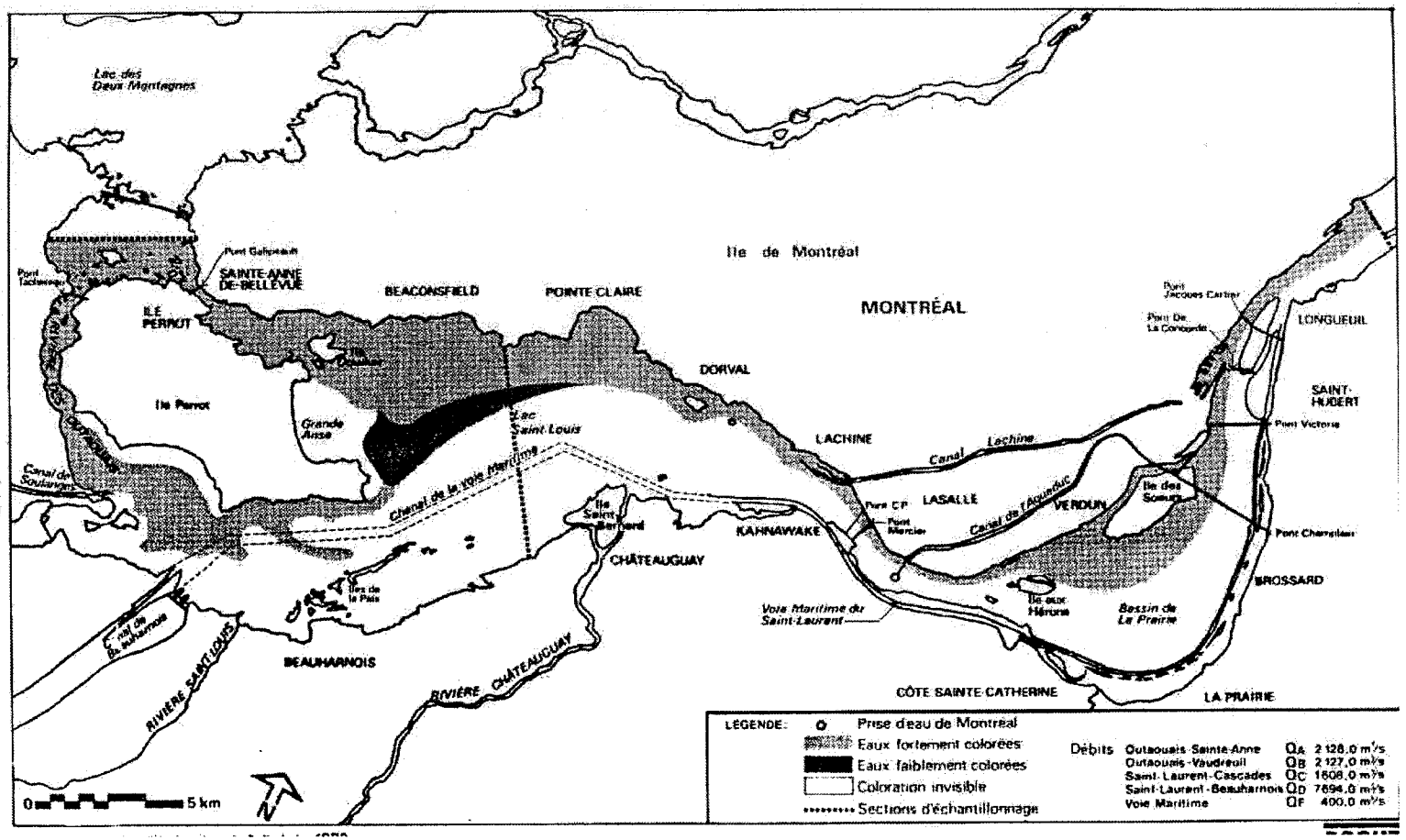


Figure 1-6 : Mélange des eaux au débit d'une forte crue de 14 000 m^3/s (Source : Hydro-Québec, 1985a)

1.1.4 Inversion thermique

Lorsque le régime hydrodynamique d'un cours d'eau se rapproche d'un régime lacustre, il est possible qu'il y ait stratification thermique de la masse d'eau. Lorsque la température de l'eau de la couche supérieure atteint 4°C, cette couche a une densité maximale supérieure à celles au-dessous et il se produit une inversion thermique, communément appelée «renversement ». Ce phénomène, observé au printemps et à l'automne, peut remettre en suspension les sédiments de surface, augmentant momentanément la turbidité de l'eau.

Pour observer une inversion thermique accompagnée d'une remise en suspension de sédiments, la profondeur de l'eau doit être suffisante et les courants doivent être faibles. Ces endroits, désignés comme zones d'accumulation permanente, sont localisées à l'extérieur du chenal principal, là où les courants sont inférieurs à 0,3 m/s et les hauteurs d'eau supérieures à 4,5 m (Carignan *et al.*, 1993). Ces conditions sont réunies dans les lacs St-François, St-Louis et St-Pierre, où près de 12% de la surface totale des lacs sont occupés par des zones d'accumulation permanente (Figure 1-7). La masse de sédiments déposés en permanence dans les lacs fluviaux est cependant faible pour l'ensemble du fleuve (Frenette *et al.,* 1989). Les secteurs latéraux des lacs fluviaux, les endroits propices à la sédimentation des MES en raison de la faible vitesse des courants dans ces secteurs, du temps de résidence des eaux plus élevé (2 à 5 jours) que dans les chenaux (8 à 14 heures) et à la présence d'herbiers aquatiques durant la période estivale, sont peu profonds et peu susceptibles de montrer une stratification thermique (Loiselle *et al.,* 1997). De plus, le régime hydrodynamique des lacs Saint-François et Saint-Louis est plus près d'un régime fluvial que d'un régime lacustre. Dans le cas du lac Saint-Louis, le temps de séjour moyen de l'eau dans le tronçon fluvial du lac Saint-louis n'est que de 12 heures (Frenette *et al.,* 1989). Même si le temps de séjour peut atteindre 2 jours dans les secteurs peu profonds situés de part et d'autre de la Voie maritime du Saint-Laurent (Carignan *et al.* ,1993), le lac Saint-Louis ne montre pas de stratification thermique (Champoux et Sloterdijk, 1988), éliminant l'occurrence d'inversion thermique.

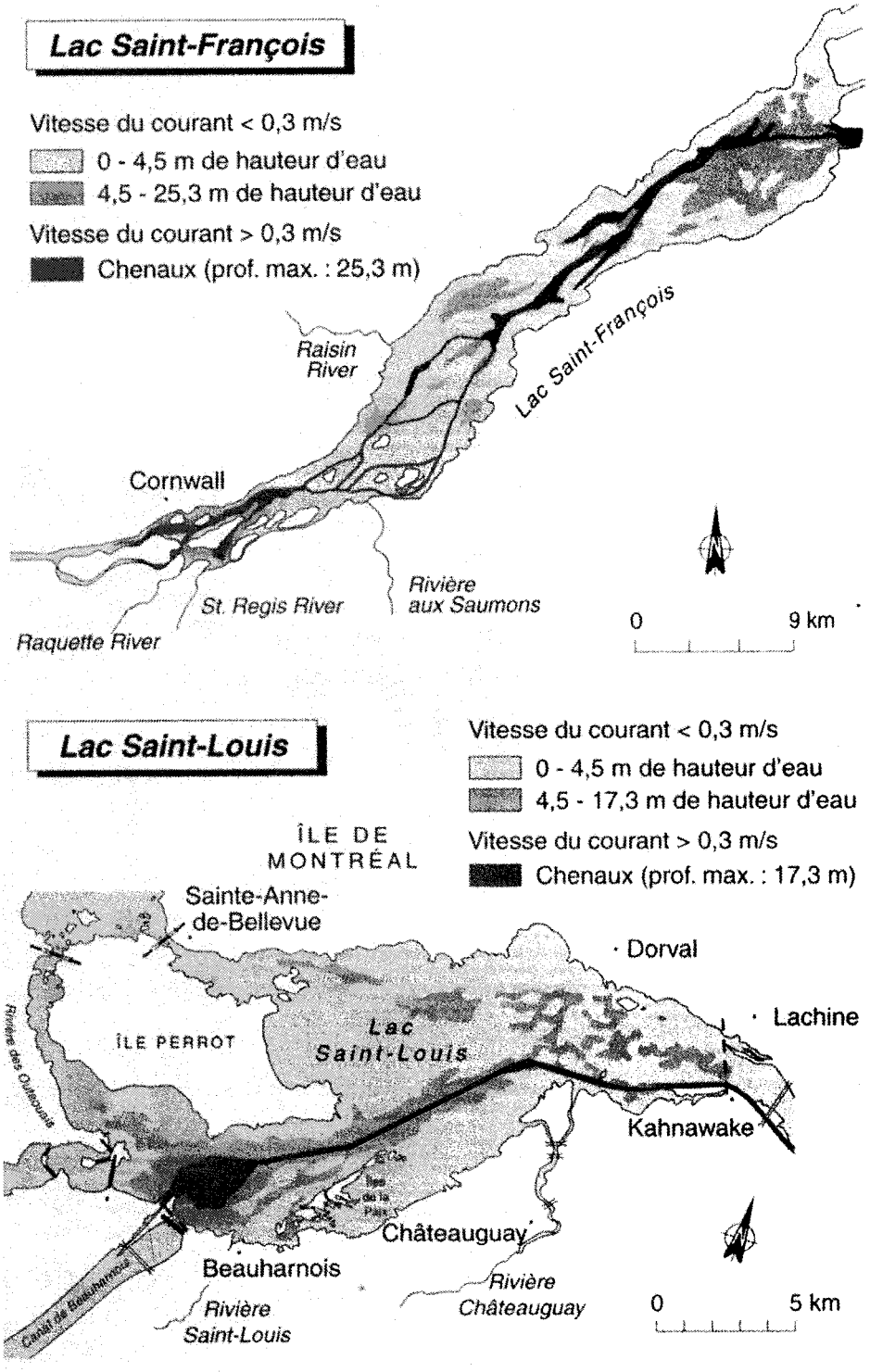


Figure 1-7: Bathymétrie et vitesse du courant dans le lac St-François et le lac St-Louis (Source : Loiselle *et al.*, 1997)

Par contre, le lac des Deux Montagnes, dans lequel se déverse la rivière des Outaouais, est le plan d'eau parmi les lacs du tronçon fluvial, dont le régime hydrodynamique se rapproche le plus d'un régime lacustre (Hydro-Québec, 1985b). Les baies comme celle de Carillon et de Vaudreuil présentent des vitesses de courant très faibles ce qui rend la sédimentation possible. Une vaste zone (de quelque 2 km de largeur) de la partie centrale du lac, de la Pointe aux Bleuets vers la baie de Vaudreuil, au pont de l'Île-aux-Tourtes, constitue une zone de sédimentation permanente depuis 30 à 50 ans avec des taux de sédimentation de l'ordre de 3 mm par année (Hydro-Québec, 1985b). Il est donc possible qu'une inversion thermique se produise dans les baies du lac des Deux Montagnes, remettant en suspension des sédiments et augmentant momentanément la turbidité de l'eau au printemps et à l'automne.

1.1.5 Tempêtes de vent et fortes pluies

Les tempêtes de vent et les fortes pluies seraient deux autres événements qui, seuls ou en conjonction avec d'autres, peuvent contribuer à accroître la quantité de particules en suspension dans les cours d'eau. Les fortes pluies érodent les berges, souvent déboisées, et charrient des particules par ruissellement jusque dans les cours d'eau. En 1994, un inventaire des rives entre Cornwall et l'île d'Orléans a permis de constater que sur les 1532 km de rives de ce tronçon, 28% étaient touchées à des degrés divers par l'érosion. Cette proportion augmente à 47% lorsque l'on considère uniquement les 848 km de rives naturelles présentes dans cette portion du fleuve (Loiselle *et al.,* 1997). Pour le seul tronçon Montréal – Trois-Rivières, 40 ha/an auraient été perdus par érosion des berges au cours des vingt dernières années (Loiselle *et al.,* 1997).

Les tempêtes de vent entraînent la formation de vagues, qui érodent les berges et remettent en suspension la fine couche de sédiments susceptibles de se déposer en périodes tranquilles dans les zones peu profondes (Frenette et Frenette, 1992). La vitesse moyenne lors des tempêtes a été estimée à 40 km/h (Frenette *et al.,* 1989). Des bourrasques de 60 à 80 km/h sont observées régulièrement sur le fleuve. Ces vents importants accroissennt la hauteur des vagues : en temps normal, les vagues atteignent

une hauteur de 0.55 à 1.25 m mais elles peuvent doubler de hauteur sous l'action de vents de 80 km/h (Fortin *et al.,* 1994).

Les zones où la hauteur d'eau est inférieure à 4,5 m sont particulièrement vulnérables à l'automne et au début de l'hiver, après la disparition des plantes aquatiques et des grands herbiers de macrophytes (Loiselle *et al.* 1997). Les effets des tempêtes de vent peuvent donc se faire ressentir dans les lacs fluviaux St-Louis et St-François, où les zones dont la hauteur d'eau est inférieure à 4,5 m sont nombreuses (Figure 1-7). Au sud-est du lac Saint-Louis, les îles de La Paix sont soumises à une érosion importante causée par les vagues d'origine éolienne (Loiselle *et al.,* 1997). Quant au lac des Deux-Montagnes, certaines zones riveraines constituées de sédiments plus grossiers laissent croire à une remise en suspension due aux vents (Hydro-Québec, 1985b).

1.1.6 Synthèse

Au printemps, plusieurs événements peuvent être à l'origine des augmentations de turbidité observées à la prise d'eau de la Ville de Montréal. Tout d'abord, elles pourraient être causées par l'accroissement du débit de la rivière des Outaouais. En plus de contribuer significativement à la charge sédimentaire du lac St-Louis, la rivière des Outaouais accroît le débit total du fleuve à la hauteur de Lasalle, qui autrement serait assez constant en raison de la régulation exercée en amont par des barrages. Ces augmentations de débit peuvent remettre en suspension les particules déposées dans les lacs St-Louis et des Deux Montagnes et peuvent éroder les berges et le lit de la rivière et du fleuve. Lors de fortes crues printanières, les masses d'eau de l'Outaouais, de la bande de transition et de la gyre peuvent aussi dégrader la qualité de l'eau brute de la Ville de Montréal en atteignant la prise d'eau de la Ville de Montréal. Même si ce n'est pas le cas selon les simulations effectuées par Hydro-Québec (1985a), il s'agit d'une possibilité à considérer, surtout lors d'une crue importante de la rivière des Outaouais.

Une inversion thermique pourrait aussi être à l'origine d'une pointe de turbidité au printemps. Le lac des Deux Montagnes est le plan d'eau le plus susceptible parmi les lacs du tronçon fluvial de permettre une inversion thermique accompagnée d'une remise en suspension de sédiments. Les hauteurs d'eau y sont suffisantes, les courants y sont faibles et on y trouve des zones de sédimentation importantes. Finalement, les tempêtes de vent et les fortes pluies seraient deux autres événements qui peuvent contribuer à accroître la turbidité de l'eau au printemps, à la condition que le couvert de glace sur une bonne partie des lacs du tronçon fluvial soit brisé.

À l'automne, les fluctuations de débit ne semblent être en cause pour expliquer les augmentations de turbidité observées à la prise d'eau de la ville de Montréal. Les tempêtes de vent et les fortes pluies seraient plutôt les événements significatifs, auxquels s'ajoute l'inversion thermique.

Les événements turbides printaniers et automnaux se distinguent donc seulement par l'accroissement ou non des débits de l'Outaouais et du fleuve St-Laurent. L'accroissement des débits étant graduelle, on s'attend à ce que les variations de turbidité de l'eau brute au printemps le soient aussi. Par contre, à l'automne, les événements explicatifs comme les tempêtes de vents, les précipitations et l'inversion thermique sont ponctuels et de courte durée, ce qui résulterait en des augmentations de turbidité soudaines et courtes. Plusieurs événements peuvent aussi se produire en même temps, amplifiant les effets sur la turbidité de l'eau brute. Au printemps, cela pourrait se traduire par des pointes de turbidité superposées à l'augmentation de turbidité plus graduelle, tandis qu'à l'automne, l'occurrence simultanée d'événements pourrait engendrer des pointes de turbidité de magnitude et/ou de durée plus importantes.

1.2 Réseaux de neurones artificiels

Les réseaux de neurones artificiels (RNA) sont une technique analytique inspirée du fonctionnement du cerveau et du système neuronal. Comme le démontrent les différentes études consultées dans le cadre de ce projet de recherche, les RNA présentent de nombreux avantages qui permettent d'être mis à profit pour prévoir des phénomènes complexes dans le domaine de l'environnement et des ressources hydriques. Toutefois, développer des modèles reposant sur les RNA n'est pas simple car aucune méthologie n'est communément acceptée.

Afin d'éclairer le lecteur sur le fonctionnement des RNA et de justifier leur utilisation dans ce projet, cette section présente tout d'abord les caractéristiques générales et les concepts de base nécessaires à la compréhension des RNA, incluant le fonctionnement du neurone artificiel, l'architecture des RNA et leur processus d'apprentissage. Ensuite, la pertinence du recours au RNA sera examinée en comparant les modèles utilisant les RNA aux modèles stochastiques et mécanistiques selon leurs avantages, leurs inconvénients et leur habilité à résoudre la problématique de la prévision hydrologique. Pour pallier au manque de cohésion sur la méthodologie à employer pour développer des modèles utilisant les RNA, une synthèse des publications d'équipes de chercheurs s'étant penchées sur cette problématique sera présentée, avant de clore la section avec une revue de différentes applications dans le domaine du génie de l'environnement et des ressources hydriques. La synthèse des approches méthodologiques et la revue des applications serviront de base à la méthodologie utilisée pour mener à bien ce projet.

1.2.1 Caractéristiques générales

Les RNA constituent une technique de modélisation très puissante qui simule le processus de résolution de problèmes du cerveau humain (Zhang et Stanley, 1997). Tout comme nous apprenons de nos expériences passées, les RNA extraient la structure d'exemples existants afin de les transférer à de nouveaux exemples auxquels ils sont exposés (Coulibaly *et al.* 1999). Selon Haykin (1994), les RNA ressemblent au cerveau

de deux façons: (i) la connaissance est acquise grâce à un processus d'apprentissage et (ii) la force des connexions interneuronales, ou poids synaptiques, est utilisée pour conserver la connaissance acquise. Alors que certains utilisateurs de RNA allèguent que cette technique constitue une forme d'intelligence artificielle, d'autres, comme Sarle (1994), considèrent que si c'est le cas, alors plusieurs méthodes statistiques doivent aussi être considérées intelligentes. Bien que les réseaux de neurones artificiels (RNA) s'appuient sur la structure des réseaux de neurones naturels (RNN), ils démontrent cependant des capacités moindres (Maier et Dandy, 1996). Alors que le cerveau humain possède environ cent milliards de neurones, les RNA en ont rarement plus de quelques centaines ou quelques milliers. Créer des RNA comparables au cerveau humain en complexité requéreraient des ordinateurs bien plus puissants que ce qui existe de mieux de nos jours (Sarle, 1994). Mais, malgré le débat qui anime le milieu des statisticiens et des utilisateurs de RNA au sujet du statut des RNA, ces derniers ont fait leurs preuves : ils ont été utilisés avec succès dans des domaines variés afin de résoudre des problèmes de classification, d'approximation de fonction, de prédiction et de prévision, d'optimisation, de reconnaissance et de contrôle (Baxter *et al.* 2001, Garceau, 2000). Dans le domaine hydrométéorologique, qui nous intéresse plus particulièrement, Coulibaly *et al.* (1999) rapportent que des modèles ont été réalisés pour la classification des données hydrologiques, la prévision des débits des rivières (crue et étiage), l'évaluation et la prévision de la qualité de l'eau, la prévision de la consommation d'eau, l'estimation des précipitations, la prévision des apports naturels aux réservoirs d'irrigation ou de production hydroélectrique.

1.2.2 Concepts de base

Bien que les réseaux neuronaux soient constitués d'unités effectuant des opérations simples, les possiblités engendrées par leur disposition et leur interconnexion sont nombreuses et complexes. Le processus d'apprentissage est également tributaire de nombreuses composantes. Il est important de bien comprendre le fonctionnement général des réseaux neuronaux afin de développer des modèles aux résultats optimaux. Cette section présente donc les notions fondamentales permettant de bien comprendre le

fonctionnement du neurone artificiel, l'architecture du réseau neuronal et le processus d'apprentissage permettant de développer le modèle.

1.2.2.1 Fonctionnement du neurone artificiel

Les neurones artificiels sont disposés en plusieurs couches : une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie (Figure 1-8). Les neurones peuvent être partiellement ou entièrement interconnectés. Le fonctionnement du neurone artificiel est relativement simple et s'apparente à celui du neurone naturel. Le neurone naturel reçoit des signaux combinés de plusieurs autres neurones par des chemins appelés dendrites. Si le signal résultant est assez puissant, le neurone est activé et produit un signal de sortie, qui est transmis par des structures de sortie aux synapses d'autres neurones, éléments-clé de l'apprentissage (Maier et Dandy, 1996). De façon similaire, le neurone artificiel reçoit des entrées des neurones des couches précédentes sous forme vectorielle, effectue une somme pondérée, et génère à l'aide d'une fonction d'activation linéaire ou non, un signal de sortie qui est envoyé à la couche suivante (Coulibaly *et al.,* 1999). On retrouve typiquement quatre types de fonction d'activation, soit les fonctions linéaire, de seuillage, sigmoïde et radiale. Les réseaux de neurones, souvent utilisés pour modéliser des paramètres non linéaires, utilisent généralement la fonction d'activation sigmoïde, qui est une fonction de régression non linéaire (Coulibaly *et al.,* 1999).

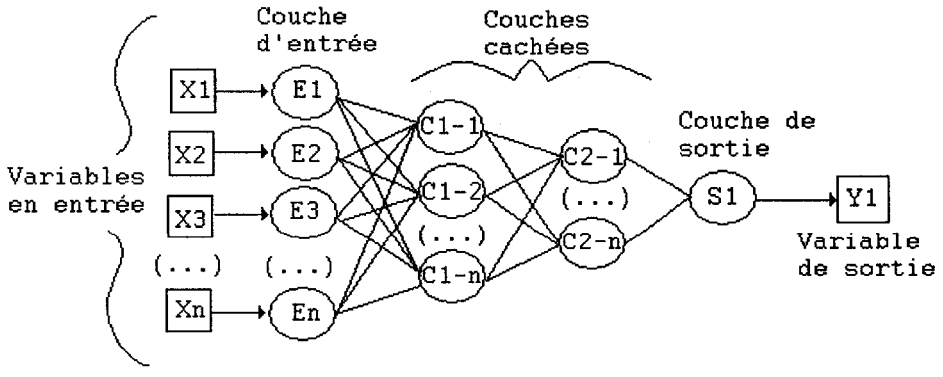


Figure 1-8 : Représentation schématique d'un réseau de neurone artificiel

Le poids de connexion, qui modélise la fonction de la synapse biologique, joue un rôle primordial dans le fonctionnement parallèle et adaptatif des neurones (Coulibaly *et al.*, 1999). Ce poids est ajustable : par exemple, un poids nul représente une absence de connexion alors qu'un poids négatif représente une relation inhibitoire entre deux neurones (Maier et Dandy, 1996). Au départ, les poids de connexion sont souvent attribués de façon aléatoire entre des valeurs limites comme -1 et 1 par exemple (Baxter *et al.*, 2002). L'importance du rôle des poids de connexion dans le fonctionnement des RNA se réflète dans l'appellation « modèles connexionnistes » utilisée par certains chercheurs pour désigner les modèles basés sur les RNA.

1.2.2.2 Architecture

Assembler des nombres variables de neurones arrangés et interconnectés de différentes façons permet de former différents types d'architecture de RNA. Les possibilités sont nombreuses : Coulibaly *et al.* (1999) rapportent que plus de 190 architectures différentes de RNA ont été étudiées ces dernières années. Malgré la diversité des architectures existantes, les RNA peuvent tout de même être regroupés en deux grande catégories, déterminées par la direction employée pour le transfert de l'information au sein du réseau : il s'agit des architectures de topologie « réaction positive » (*feedforward*) et celles de topologie récursive ou récurrente. Le réseau feedforward est un réseau non bouclé qui ne présente aucune rétroaction entre les neurones des différentes couches. Le fonctionnement du neurone artificiel décrit dans la section 1.2.2.1 réfère à la topologie feedforward, car les activations sont projetées vers l'avant. Quant au réseau récursif, il peut être totalement ou partiellement bouclé, les neurones d'une couche étant connectés aux noeuds de la couche suivante, de la couche précédente et de la même couche. Les réseaux feedfoward sont donc des cas spéciaux de réseaux récurrents.

Traditionnellement, les modèles connexionnistes de prévision et de prédiction ont recours aux réseaux « feedfoward ». L'architecture feedforward la plus utilisée de toutes est le perceptron multicouche (PMC) (Maier *et al.*, 2001). C'est pourquoi les

concepts présentés par la suite dans ce chapitre portent principalement sur ce type d'architecture. Le réseau fonction radiale de base (*Radial Basis Function* ou RBF), le réseau de neurones de régression généralisée (*General Regression Neural Network* ou GRNN) et le réseau de neurones probabiliste (*Probabilistic Neural Network* ou PNN) sont trois autres architectures rencontrées couramment. Pour plus de détails sur ces architectures, l'excellent ouvrage de référence de Haykin (1994) peut être consulté.

Récemment, les réseaux récurrents ont été proposés comme alternatives aux réseaux feedforward, car ils seraient particulièrement efficaces pour modéliser les séries temporelles, communes aux applications de prévision et de prédiction (Zhang et Stanley, 1997, Maier et Dandy, 2001, Wilson et Recknagel, 2001). Ils présentent également l'avantage non négligeable de modéliser les propriétés dynamiques implicitement, alors que les réseaux feedfoward ont besoin que les systèmes dynamiques soient traités explicitement par l'ajout de variables de décalage (Maier et Dandy, 2001). Mais ils prennent aussi plus de temps à se développer car leur vitesse d'apprentissage est plus lente.

1.2.2.3 Apprentissage

L'apprentissage, ou phase d'optimisation, est une procédure adaptative par laquelle les poids de connexion des neurones sont ajustés afin d'extraire la structure et les tendances des données historiques qui leur sont soumises. Cette procédure est déterminante pour l'efficacité du réseau neuronal (Brion *et al.*, 2001). Il existe deux types d'apprentissage que l'on emploie avec différents types d'architectures, soit les apprentissages supervisé et non-supervisé. L'apprentissage supervisé se déroule comme suit : des paires de données d'entrée et de sortie, regroupées au sein d'une base de données, sont présentées au réseau d'une architecture donnée; un algorithme d'apprentissage calcule l'écart entre les valeurs de sortie réelles et les valeurs prédites par le réseau; si l'écart est trop important, le poids de chacune des connexions est modifié jusqu'à ce que l'écart atteigne une valeur minimale acceptable. Dans le cas d'apprentissage non-supervisé, le réseau ne requiert pas de données de sortie. Il ajuste les poids de connexion uniquement

à partir des données d'entrée afin que les valeurs de sortie représentent les régularités statistiques des données. Par l'apprentissage non-supervisé, le réseau cherche à détecter les similarités et les différences entre les variables présentées dans la base de données et à refléter ces propriétés dans les valeurs de sortie, d'où son utilisation pour la reconnaissance des patrons, le traitement du signal et l'analyse factorielle (Coulibaly *et al.,* 1999).

Peu importe le type d'apprentissage, les poids de connexion des neurones sont modifiés par un algorithme d'apprentissage (AA) dans le but de trouver le minimum de la fonction d'erreur dans un espace multidimensionnel (Coulibaly *et al.*, 1999). La fonction d'erreur la plus couramment utilisée est la fonction des moindres carrés (Maier et Dandy, 2000b). L'AA explore la surface d'erreur à la recherche de son point le plus faible à l'aide de méthodes locales ou globales. Les méthodes locales comprennent celles de premier-ordre, qui s'appuient sur le modèle linéaire du gradient croissant (*gradient-descent*), et celles du second-ordre, basées sur le modèle quadratique comme la méthode de Newton (Maier et Dandy, 2000b). Les méthodes globales s'appuient sur des approches statistiques ou génétiques (Coulibaly *et al.*, 1999). Plus de détails sur les différentes méthodes peuvent être trouvés dans Maier et Dandy (2000b). Les méthodes locales sont les plus utilisées. Dans tous les cas, ce sont des techniques itératives qui minimisent la fonction d'erreur. L'équation de mise-à-jour des poids de connexion est de la forme générale suivante :

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \gamma_n \mathbf{d}_n \quad \text{(Équation 1)}$$

où w_n est le vecteur des poids de connexion, γ_n est la largeur du pas, d_n est le vecteur définissant la direction de la descente et l'indice n dénote le nombre d'itérations. La différence essentielle entre les différents algorithmes est le choix de d_n, qui détermine le taux de convergence et la complexité informatique.

L'algorithme de rétropropagation (ARP) (*backprograpagtion*) est l'algorithme le plus utilisé en prévision (Coulibaly *et al.,* 1999). L'ARP utilise la méthode du gradient croissant pour minimiser les erreurs dans le cadre d'un apprentissage supervisé. Il sert

surtout à entraîner les réseaux *feedforward* d'architecture perceptrons mutli-couches. L'ARP standard est lent et très sensible aux variations de la vitesse d'apprentissage (*learning rate)*, qui fixe la largeur des sauts sur la surface d'erreur, au *momentum*, qui permet d'augmenter la vitesse et ainsi d'échapper aux minima locaux, aux valeurs initiales des poids et à la taille de l'ensemble d'apprentissage (Garceau, 2000 et Coulibaly *et al.,* 1999). Pour palier à ces problèmes, des variantes de l'ARP ont été développés, notamment les fonctions d'optimisation non-linéaire de *Levenberg-Marquardt* et de gradient croissant conjugué (*Conjugated gradient descent*), mais elles ne sont applicables que dans certaines circonstances, notamment avec des données abondantes ou une seule variable de sortie (Bishop, 1995; Shepherd, 1997, cités par Garceau, 2000). Mais si la vitesse d'exécution n'est pas un critère important pour l'utilisateur de RNA, le recours à l'ARP est acceptable.

L'architecture du RNA joue aussi un rôle important dans le processus d'apprentissage. Si le nombre de neurones dans la couche cachée est insuffisant, le RNA peut manquer d'opportunités pour bien capturer les relations intrinsèques entre les variables de la base de données. Par contre, trop de neurones dans la couche cachée peut résulter en un trop grand nombre d'itérations et nuire à la capacité de prévision ou de classification du RNA : au lieu d'apprendre à généraliser à partir des paires de données historiques soumises et ainsi capturer les caractéristiques générales présentées par l'ensemble de données, le réseau se concentre sur les caractéristiques des points individuels (Brion *et al.,* 2001). On dit alors que le réseau « apprend par cœur » ou mémorise la fonction qu'il tente de modéliser (*over-learning*). L'approche la plus efficace pour déceler le « surapprentissage » par le réseau est la validation croisée (*cross-validation)* : une partie des données, qu'on désigne par ensemble de test, sert à stopper la progression de l'AA durant l'apprentissage si l'erreur d'apprentissage diminue ou stagne alors que l'erreur de vérification augmente, signifiant que le réseau commence à « apprendre par cœur » (Garceau, 2000).

L'erreur de validation, obtenue par la validation croisée, constitue un bon indice de la performance d'un réseau. Mais comme l'ensemble de test utilisé lors de la validation croisée fait partie du processus d'apprentissage, cet indice de performance s'avère biaisé à moins qu'il ne soit prévu de réserver un autre ensemble pour valider de façon indépendante le modèle suite à l'apprentissage en évaluant la justesse des prévisions effectuées par le modèle développé (Baxter *et al.*, 2002). L'erreur de cet ensemble de données, désigné par ensemble de production ou de validation, peut être utilisé comme critère d'arrêt afin de déterminer le meilleur moment pour cesser l'apprentissage, tout en permettant de comparer l'habilité de généralisation de différents modèles (Maier et Dandy, 2000b).

1.2.3 Justification du recours aux RNA

Développer un modèle de prévision robuste qui prend en compte toutes les corrélations disponibles et les interdépendances existant dans les séquences chronologiques des variables est une tâche difficile. Outre les modèles connexionnistes, des modèles stochastiques et mécanistiques pourraient aussi être utilisés pour effectuer les prévisions de qualité d'eau brute. Il est donc primordial de s'assurer que les RNA sont la technique la plus pertinente pour résoudre la problématique de la prévision de qualité de l'eau brute.

La prévision de la qualité de l'eau est un exemple de prévision hydrologique, définie par Coulibaly *et al.* (1999) comme « l'estimation des conditions futures des phénomènes hydrologiques pour une période donnée, à partir des observations passées et actuelles ». Ils distinguent clairement la prévision de la prédiction, la première ayant comme objectif général de fournir les meilleures estimations de ce qui peut arriver en un point donné à une date future précise, contrairement à la prédiction, qui vise l'estimation de conditions futures sans référence à un temps spécifique. Toujours selon Coulibaly *et al.* (1999), la problématique de la prévision hydrologique est constituée des éléments suivants:

- la variable à prévoir et les variables explicatives;
- l'horizon de prévision;
- les méthodes de calcul ou d'estimation (i.e. la fonction d'activation);
- l'objectif de la prévision (alerte de crue, planification de l'opération des réservoirs, projets d'irrigation ou de navigation);
- le type de résultats désirés (valeurs numériques, graphiques, ou distribution de probabilités).

Deux approches peuvent prendre en compte tous les éléments de la problématique de la prévision hydrologique. La première est l'approche mécanistique (ou déterministe) et la seconde est l'approche empirique (ou stochastique). L'approche mécanistique s'appuie sur la simulation physique du système. Elle requiert que tous les phénomènes physiques fondamentaux sous-jacents au système soient bien compris et puissent être décrits mathématiquement (Maier et Dandy, 2000a). Les modèles mécanistiques ont donc l'avantage d'avoir un large domaine d'applicabilité. Malheureusement, les modèles mécanistiques sont limités par le très grand nombre de paramètres à mesurer et par les limites des connaissances actuelles des systèmes naturels complexes (Coulibaly *et al.,* 1999).

L'approche empirique permet de passer outre les limites des connaissances physiques du système en prenant seulement en compte les relations empiriques entre les paramètres de qualité de l'eau et les nombreuses variables environnementales. Contrairement aux modèles mécanistiques, les modèles empiriques fonctionnent comme des « boîtes noires » parce qu'il ne prennent pas en compte la structure interne du système et ne livrent pas de représentation mathématique explicite du système (Coulibaly *et al.*, 1999 et Baxter *et al.*, 2001). La forme du modèle, qui comprend par exemple la régression linéaire, la régression logistique, la méthode autorégressive à moyenne mobile (*ARMA* : *autoregressive moving average*), la méthode autorégressive non-linéaire à moyenne mobile (*NARMA : nonlinear autoregressive moving average*),

est sélectionnée *a priori* et les paramètres inconnus du modèle sont estimés par la suite en minimisant l'erreur entre les prévisions du modèle et les valeurs historiques connues (Maier et Dandy, 2000a). Contrairement aux modèles mécanistiques, l'applicabilité des modèles empiriques est souvent restreinte par le recours à des données historiques qui réflètent une réalité temporelle et spatiale donnée (Baxter *et al*, 2001). Les différences entre les modèles mécanistiques et empiriques sont présentées dans le Tableau 1-1.

Tableau 1-1 Les différences majeures entre les modèles mécanistiques et statistiques (adapté de Maier et Dandy, 2000a)

MODÈLES MÉCANISTIQUES	MODÈLES EMPIRIQUES
• S'appuie sur des processus physiques sous-jacents compris et décrits mathématiquement • Requiert généralement des collectes de données intensives et de courte durée • Requiert beaucoup de données • Flexible en terme de résolution spatiale et temporelle • Grand domaine d'applicabilité, mais requiert une calibration avec des données locales • Ne requiert pas beaucoup de pré-traitement de données • Utilisé généralement si de grands changements physiques doivent être apportés à un système • N'est généralement pas utilisé pour la prévision • N'est généralement pas utilisé pour générer des données synthétiques	• Les processus physiques sous-jacents n'ont pas à être compris explicitement • Utilise généralement des bases de données existantes • Requiert moins de données • Résolution spatiale et temporelle contrainte par les données existantes • Spécifique au site • Requiert généralement beaucoup de pré-traitement de données • N'est généralement pas utilisé pour évaluer l'impact de grands changements à un système • Utilisé fréquemment en prévision • Utilisé fréquemment pour générer des données synthétiques

Les RNA peuvent être considérés comme un outil nouveau dans le domaine de la prédiction et de la prévision (Maier et Dandy, 2000b). La philosophie sous-jacente à la modélisation par des RNA est similaire à celle des approches empiriques traditionnelles, car dans les deux cas, les paramètres inconnus du modèle (soit les poids de connexion dans le cas des RNA) sont ajustés afin d'obtenir la meilleure correspondance possible entre les données historiques et les données prédites (Maier et Dandy, 2000a). En fait, des études examinant les RNA d'un point de vue statistique indiquent que les modèles

connexionnistes avec certaines géométries, connexions et paramètres internes sont soit équivalents ou très semblables à des modèles statistiques existants (Sarle, 1994). C'est le cas des modèles empiriques de type *ARMA* (Maier et Dandy, 2000a), qui ont été utilisés traditionnellement pour modéliser les séries temporelles de ressources hydriques (Maier et Dandy, 1997).

Pourquoi recourir aux RNA plutôt qu'à une méthode empirique traditionnelle pour prévoir les variations de qualité de l'eau? Une grande controverse handicape présentement la comparaison objective des approches connexionniste et empirique (Coulibaly *et al.,* 1999). Maier et Dandy (2000a) avancent néanmoins une piste intéressante : « même si certains modèles connexionnistes ne sont pas très différents de plusieurs *modèles* statistiques standards, ils sont extrêmement utiles car ils offrent une façon flexible de les *mettre en place* ». Il est vrai que, contrairement à plusieurs modèles statistiques, les RNA peuvent intégrer facilement plus de variables d'entrée et s'adapter à la non-linéarité. La complexité du modèle peut aussi être modifiée simplement en altérant la fonction d'activation ou l'architecture du modèle. De plus, Maier et Dandy (2000a) soulèvent un fait pouvant expliquer en partie la popularité grandissante des RNA pour la modélisation : « L'approche adoptée dans la modélisation connexionniste est différente de celle utilisée dans le développement de modèles statistiques traditionnels. La raison est que les RNA ont été développés par les ingénieurs et les informaticiens, plutôt que par des statisticiens. Les utilisateurs de RNA sont concernés par la justesse des prévisions et les méthodes qui fonctionnent, alors que l'objectif principal des statisticiens est de développer une méthodologie universelle et d'atteindre l'optimalité statistique.

En résumé, voici les avantages de la technique des RNA par rapport aux méthodes de modélisation conventionnelles qui font des RNA une alternative de choix dans le domaine de l'eau potable:

1. Les RNA ne nécessitent aucune hypothèse quant au phénomène ou à la relation à modéliser (Haykin, 1994). Les modèles connexionnistes peuvent être

développés sans quantifier les interactions à micro-échelle. Dans le traitement de l'eau potable, ces interactions sont souvent peu comprises, rendant le développement de modèles mécanistiques très difficiles (Baxter *et al.*, 2001).

2. Les RNA profitent de la non-linéarité du système grâce à la structure non-linéaire des données et au procédé informatique (Zhang et Stanley, 1997; Haykin, 1994). Plusieurs variables de qualité de l'eau variant de façon cyclique ou saisonnière, elles sont non-linéaires et difficiles à modéliser par des techniques standard (Maier *et al.,* 1998).

3. L'approche par RNA est rapide et flexible. La construction de modèles physiques et la conduite d'échantillonnage et de tests de laboratoire sont inutiles (Zhang et Stanley, 1997).

4. Les RNA s'adaptent bien aux changements (Haykin, 1994). Le modèle peut intégrer des modifications en étant ré-entraîné avec de nouvelles données (Zhang et Stanley, 1997). Dans le domaine de l'eau potable, les modifications sont fréquentes. Or, avec des modèles conventionnels, des changements aux phénomènes à représenter rendant souvent les modèles invalides.

5. Les RNA tolèrent bien les discontinuités dans les données, les différents niveaux de précision des données, le bruit et la dispersion des données (*data scatter*) (Baxter *et al.*, 2001).

6. Finalement, puisque les modèles de RNA sont développés en utilisant des données opérationnelles « *full-scale* », les problèmes de « scale-up » associés souvent aux modèles empiriques *bench-scale* et *pilot-scale* sont éliminés (Baxter *et al.*, 2001).

1.2.4 Méthodologie pour le développement de modèles

Il n'existe encore aucun protocole généralement accepté pour la construction de modèles connexionnistes (Baxter *et al.*, 2001). En fait, dans la plupart des applications des réseaux de neurones dans le domaine de l'hydrologie et de l'environnement, la méthodologie employée pour construire les modèles est peu décrite ou tout simplement omise. Maier et Dandy (2000b) ont passé en revue 43 articles exposant les résultats de l'utilisation de RNA pour la modélisation hydrologique et peu faisaient mention de la méthodologie employée. La raison est simple : la plupart des scientifiques et ingénieurs considèrent les réseaux de neurones comme des boîtes noires. Ils développent donc les modèles connexionnistes selon des méthodes heuristiques d'essais et erreurs, en ignorant les règles s'appliquant à la modélisation statistique traditionnelle. Or, les réseaux neuronaux et les modèles statistiques sont étroitement liés (Maier et Dandy, 2000a). Des études récentes indiquent d'ailleurs que la considération de principes statistiques dans le processus de construction du modèle peut améliorer la performance du modèle connexionniste (Maier et Dandy, 2000b).

Deux équipes de chercheurs oeuvrant dans le domaine de la modélisation environnementale et de l'eau potable tentent d'élaborer depuis quelques années une méthodologie rigoureuse prenant en compte des principes statistiques. L'équipe de Maier, Dandy et Bowden afine depuis plusieurs années des modèles connexionnistes permettant de prévoir la salinité et la présence de cyanobactéries dans la rivière Murray, au sud de l'Australie. Suite à leur travail avec les réseaux neuronaux, ils ont publié plusieurs articles, revues de littérature et chapitres de manuels de référence portant sur la méthodologie de la modélisation connexionniste dans le domaine environnemental (Maier et Dandy, 1997, 2000a, 2000b, Bowden *et al.*, 2000a, 2000b, 2001, 2002). Leur revue de 43 applications dont il est fait mention au début de la section (Maier et Dandy, 2000b) est particulièrement utile pour ce projet. Une synthèse des détails méthodologiques de ces articles est donc présentée à l'Annexe 2. L'équipe de Baxter, Stanley et Zhang, quant à elle, a développé, appliqué et révisé une méthodologie qui a

été utilisée avec succès pour développer des modèles RNA pour la prévision de la couleur de l'eau brute, de la demande en eau et la modélisation des procédés de coagulation et de filtration de l'eau de la rivière Saskatchewan Nord, à Edmonton, en Alberta (Zhang et Stanley,1997; Zhang et Stanley; 1999; Baxter *et al.,* 2002).

Une synthèse des approches méthodologiques suggérées par ces deux équipes de chercheurs est présentée dans la suite de la section. La synthèse aborde en détail les différentes étapes de modélisation, soit l'évaluation des besoins et des ressources disponibles, le choix du critère de performance, la constitution de la base de données, la construction du modèle et son évaluation. Cette synthèse constitue la base de la méthodologie employée dans ce projet pour développer les modèles connexionnistes servant à prévoir les pointes de turbidité à l'eau brute à la Ville de Montréal. Elle représente aussi un excellent point de départ pour tous les utilisateurs des RNA voulant s'appuyer sur une méthodologie rigoureuse.

1.2.4.1 Évaluation des besoins et des ressources disponibles

Avant de recourir aux RNA pour modéliser des traitements de l'eau potable ou pour prévoir des variations de différents paramètres affectant la qualité de l'eau à traiter, il faut évaluer les besoins de la station de traitement face à un modèle et à ses applications pour s'assurer que les RNA sont la technique la plus appropriée pour le problème à résoudre. Quand les variations de qualité de l'eau ou les procédés de traitement sont bien décrits par des modèles mécanistiques existants ou par des modèles empiriques spécifiques au site, recourir à la technologie des RNA peut être superflue. Par contre, pour modéliser les procédés non-linéaires complexes, où les interactions entre les entrées et les sorties des procédés sont peu comprises, la modélisation connexionniste est appropriée (Baxter *et al.,* 2002).

Il faut aussi s'assurer de disposer de toutes les ressources nécessaires à la modélisation par RNA pour en tirer réellement profit. Des données pertinentes pour décrire le modèle à développer doivent être disponibles dans un format électronique utilisable. Un

logiciel commercial approprié de RNA doit également être disponible, le choix du meilleur logiciel pour chaque station étant dicté par les restrictions spécifiques aux systèmes d'exploitation des ordinateurs et les formats de données de même que par les options désirées. Quant aux besoins en équipement informatique, un processeur de 500 MHz et 128 mégaoctets de mémoire devrait être considéré comme les exigences minimales afin de s'assurer d'une performance optimale (Baxter *et al.,* 2002). Finalement, les exigences en ressources humaines comprennent un développeur de modèles qui ait une connaissance approfondie du procédé ou du phénomène à modéliser ou à prévoir et qui comprenne l'heuristique de base de la modélisation connexionniste (Baxter *et al.,* 2002).

1.2.4.2 Choix du critère de performance

L'évaluation des besoins permet aussi de définir le critère de performance. Selon Maier et Dandy (2001), l'évaluation du modèle inclut l'un ou plusieurs des critères suivants : la justesse de la prévision, la vitesse d'apprentissage et le délai nécessaire pour obtenir une prévision une fois le modèle développé. La justesse de la prévision est de loin le critère de performance le plus utilisé. Des 43 applications examinées par Maier et Dandy (2000b), seulement deux ont choisi la vitesse d'apprentissage plutôt que la justesse de la prévision. Ces applications consistaient d'ailleurs à trouver des alternatives à des modèles chronophages, i.e. qui nécessitent beaucoup de temps à développer.

Plusieurs mesures de la justesse de la prévision sont proposées dans la littérature, la plus commune étant la mesure de l'erreur de l'ensemble de test (voir section 1.2.2.3). Les mesures d'erreur les plus utilisées sont l'erreur quadratique moyenne (*root-mean-square error* ou *RMSE*) et le R^2 (Zhang et Stanley, 1997). L'erreur quadratique moyenne est la moyenne du carré de la différence entre les valeurs de sortie historiques et les valeurs de sortie prédites par le RNA. Par le carré, les erreurs importantes sont pénalisées et l'effet des différences négatives est annulé. Cette mesure est particulièrement utile pour évaluer la justesse des prévisions (Maier *et al.*, 1998). Quant au R^2, cette mesure

statistique est un indicateur de la capacité du modèle à prendre en compte la variabilité des variables spécifiées dans le modèle (Statsoft, 1998). Le R^2 est obtenu selon l'équation suivante :

$$R^2 = 1 - [(x_{obs} - x_{préd})^2 / x_{obs}^2] \quad \text{(Équation 2)}$$

où x_{obs} représente les valeurs observées historiquement et $x_{préd}$ correspond aux valeurs correspondantes prédites par le modèle.

Toutes les mesures de la justesse de la prévision ont en commun d'évaluer l'habilité de généralisation du réseau, définie comme la capacité du modèle de bien fonctionner avec des données qui n'ont pas été utilisées durant l'apprentissage. Selon Maier et Dandy (2000b), la justesse de la prévision est affectée par l'algorithme d'apprentissage utilisé, lequel se distinguent par leur capacité à échapper aux minima locaux. La justesse de la prévision dépend aussi du rapport entre le nombre d'échantillons d'apprentissage et le nombre de connexions : si ce rapport est trop faible, continuer l'apprentissage peut résulter en surapprentissage, affectant alors grandement la capacité de généralisation du réseau. Finalement, la justesse de la prévision est influencée par la représentativité des ensembles d'apprentissage et de validation et par le critère d'arrêt utilisé, qui permet de mettre fin au processus d'apprentissage (voir section 1.2.4.4).

Bien que la vitesse d'apprentissage ne soit pas souvent choisie comme critère de performance, elle n'en demeure pas moins une préocupation importante pour les chercheurs, car plusieurs en font mention dans leurs articles. Maier et Dandy (2000b) suggèrent d'ailleurs, sans nécessairement en évaluer formellement la performance, de tenir compte de la vitesse d'apprentissage en gardant (i) le nombre de poids de connexion le plus petit possible, (ii) les connexions entre elles les plus simples possible et (iii) en évitant la redondance dans les données d'apprentissage.

1.2.4.3 Constitution de la base de données

Collecte des données

Pour développer des modèles qui produisent de bons résultats, une attention particulière doit être portée à la collecte et à l'analyse des données. D'abord et avant tout, les données des variables susceptibles d'être liées aux phénomènes à prévoir ou aux procédés à modéliser doivent être représentatives. Elles doivent donc couvrir l'échelle de valeurs de ces variables, car le modèle ne peut extrapoler au-delà des valeurs auxquelles il a été exposé (Bowden *et al.,* 2001). Évidemment, les données doivent aussi être disponibles, fiables et du format électronique approprié (Baxter *et al.,* 2002). La quantité de données requise pour développer un modèle dépend du site et des fluctuations saisonnières de la qualité de l'eau brute. Pour s'assurer que les données soient représentatives, au moins un cycle complet de données doit être disponible (Baxter *et al.,* 2002). Par exemple, dans les régions tempérées avec quatre saisons, un cycle complet représente une année.

Analyse des données

Une fois l'ensemble de données sélectionné, il doit être caractérisé et assujetti à une analyse statistique afin d'identifier les limites du domaine d'étude de même que les déficiences potentielles de l'ensemble de données (Baxter *et al.,* 2002). La caractérisation des données comprend une évaluation qualitative des tendances saisonnières de chaque variable potentielle du modèle. L'analyse statistique consiste à effectuer les mesures de tendance centrale, des mesures de la variation et une analyse des percentiles (Baxter *et al.,* 2002). Il faut également examiner les variables pour déterminer si elles sont fortement autocorrélées (une forte autocorrélation signifie que les valeurs d'aujourd'hui sont fortement liées à celles des jours précédents). Une forte autocorrélation de variables d'entrée mais surtout de sortie peut faire perdre au modèle sa capacité de bien généraliser (Zhang et Stanley, 1997). Finalement, les observations aberrantes (*outliers*), les entrées erronées et les entrées manquantes doivent être

identifiées pour les corriger ou les éliminer si nécessaire. La détection des observations aberrantes est grandement subjective. Baxter *et al.,* (2002) suggèrent d'utiliser le graphique de nuages de points de chaque variable pour détecter les observations aberrantes ou d'exclure toutes les valeurs dépassant l'écart de ±2 écarts-type (STD) de la moyenne de la variable.

Organisation des données

Maier et Dandy (2000a) préfèrent organiser les données avant de sélectionner les variables d'entrée et de sortie alors que Baxter *et al.* (2002) préfèrent le faire après la sélection. L'important, c'est de le faire correctement. Si la technique de la validation croisée est employée (voir section 1.2.2.3), les données doivent être divisées en trois ensembles : l'ensemble d'apprentissage, l'ensemble de test et l'ensemble de validation. L'ensemble d'apprentissage est utilisé pour trouver l'ensemble optimal de poids de connexion, l'ensemble de test, pour choisir la meilleure configuration de réseau en évitant le surapprentissage et, une fois le réseau optimal trouvé, l'ensemble de validation pour permettre de déterminer le meilleur moment pour cesser l'apprentissage et pour tester l'habilité réelle de généralisation du modèle.

La formation de trois sous-ensembles requiert beaucoup de données. Quand les données disponibles sont limitées, il peut être difficile d'avoir un ensemble de validation représentatif. Maier et Dandy (2000b) suggèrent alors d'employer la méthode *holdout*, dans laquelles les données sont divisées en seulement deux sous-ensembles, soit un ensemble d'apprentissage et un ensemble indépendant de validation, qui joue aussi le rôle d'ensemble de test. Dans cette méthode, un petit sous-ensemble de données est préservé pour la validation et le reste des données est utilisé pour l'apprentissage. Une fois l'habilité de généralisation du réseau établie à l'aide de l'ensemble de validation, un sous-ensemble différent est retenu et le processus est répété, jusqu'à ce que l'habilité de généralisation ait été établie pour toutes les données disponibles. Cette méthode est plus fastidieuse que la méthode de validation croisée. Elle est donc utilisée plus rarement.

Peu importe la méthode employée, la façon dont les données sont subdivisées peut avoir une influence significative sur la performance du réseau (Flood et Kartam, 1994a). Parce que le réseau est typiquement incapable d'extrapoler au-delà de l'échelle de données utilisées pour l'apprentissage, les ensembles d'apprentissage et de validation doivent être représentatifs de la même population (Bowden *et al.*, 2001). Malgré son importance, aucune approche systématique pour la division optimale des données n'a été développée (Bowden *et al.*, 2001). En général, les données sont subdivisées arbitrairement entre les différents sous-ensembles selon un ratio pré-déterminé. Un ratio de 3 :1 :1 s'est révélé efficace pour plusieurs modèles (Baxter *et al.*, 2002). Les données peuvent être divisées par année ou par mois, particulièrement si les données réflètent un phénomène cyclique. C'est ce qu'ont fait Maier et Dandy (1996) en utilisant à tour de rôle une des quatre années de données disponibles comme ensemble de test pour développer et comparer quatre réseaux différents prévoyant la salinité de l'eau de la rivière Murray, en Australie.

Si les données ne sont pas cycliques ou semblables d'une période à l'autre, Baxter *et al.* (2002) suggèrent d'organiser les données en triant d'abord les patrons de données selon la valeur de la variable de sortie et en assignant ensuite les patrons de données aux sous-ensembles selon le ratio pré-déterminé. Par une analyse de variance ou d'autres mesures statistiques, on peut s'assurer que les trois sous-ensembles de données sont similaires et représentatifs de l'ensemble des données. Des techniques analytiques comme les algorithmes génétiques (AG) et « Self-organizing Map » (SOM), utilisées pour déterminer les variables d'entrée des modèles connexionniste multi-variables et décrites à la section 1.2.4.4, peuvent aussi servir à diviser les données dans des sous-ensembles représentatifs (Bowden *et al.*, 2001).

Maier et Dandy (2000b) rapportent que la division des données est incorrecte dans la majorité des applications examinées dans leur revue. En général, la division des données a été effectuée de façon arbitraire et les propriétés statistiques des ensembles de données ont été peu considérées. De plus, même si au moins deux ensembles de

données ont été utilisés dans pratiquement tous les articles, les données de validation ont souvent été utilisées dans le processus d'apprentissage ou pour optimiser les variables d'entrée, la géométrie du réseau ou les paramètres internes. De plus, les données de l'ensemble de test ont parfois été utilisées comme critère d'arrêt. Toutes ces pratiques soulèvent des doutes quant à l'optimalité des résultats obtenus.

Traitement des données

Une fois les données réparties dans leurs sous-ensembles respectifs, on peut procéder à la transformation des données. Généralement, les donnéess ont des échelles différentes. Afin de recevoir une attention égale lors de l'apprentissage du réseau, il est préférable de normaliser les données (Maier et Dandy, 2000a), particulièrement si la méthode d'apprentissage utilise l'algorithme de rétro-propagation (Brion *et al.*, 2001). Par exemple, les données peuvent être normalisées pour avoir une moyenne de zéro et un écart-type de 1 (Wilson et Recknagel, 2001). Il est aussi suggéré de mettre les données à l'échelle afin qu'elles soient dans les limites des fonctions d'activation utilisées dans la couche de sortie (Bowden *et al.,* 2001). La transformation linéaire est de loin la technique de transformation de données la plus utilisées dans les applications de RNA (Bowden *et al.,* 2001). L'ensemble de données est habituellement mis à l'échelle dans une étendue de 0 à 1 ou de -1 à 1 en utilisant l'étendue des données originales comme scalaire. Par exemple, avec la fonction de transfert sigmoïdale où les valeurs de sortie sont entre 0 et 1, Maier et Dandy (2000a) proposent de mettre les valeurs à l'échelle entre 0,1-0,9 ou 0,2-0,8, en évitant les valeurs extrêmes susceptibles de résulter en des optima locaux lors de l'apprentissage.

Dans les faits, la transformation de données n'est pas souvent effectuée car il est couramment perçu dans la littérature que les données utilisées par les modèles RNA n'ont pas besoin d'être transformées (Bowden *et al.*, 2001). Des 43 articles étudiés par Maier et Dandy (2000b) dans leur revue de littérature, seulement 18 font mention d'une mise à l'échelle des données à l'aide d'une transformation linéaire. Bowden *et al.* (2001) ont tout de même mené une étude suggérant que certaines transformations

pouvaient améliorer la performance des modèles RNA. Ils ont développé et comparé des modèles utilisant différentes techniques de transformation de données afin de prévoir la salinité dans la rivière Murray (Australie du sud) 14 jours en avance. Le modèle développé en utilisant la transformation linéaire produisait en l'erreur de prévision la plus faible. Aucune amélioration de l'habilité de prévision du modèle n'a été obtenue en utilisant les transformations logarithmiques, saisonnières et de normalité, trois autres techniques décrites avec détails dans Bowden *et al.* (2001).

1.2.4.4 Construction du modèle

Sélection des variables d'entrée et de sortie

La tâche la plus difficile mais aussi la plus importante dans le développement de modèles de prévision est la sélection des variables d'entrée et de sortie appropriées (Maier et Dandy, 1997, Bowden *et al.,* 2000a). Habituellement, on débute en sélectionnant la ou les variables de sortie du modèle. Comme les RNA donnent de meilleurs résultats quand une seule variable de sortie est modélisée, il est préférable de s'en tenir à une seule variable de sortie et de développer des modèles différents pour chacune des variables de sortie s'il doit y en avoir plusieurs (Baxter *et al.,* 2002). La variable de sortie est sélectionnée sur la base de l'usage éventuel du modèle, la littérature scientifique et la disponibilité des données (Baxter *et al.,* 2002). La variable de sortie utilisée peut ne pas être la valeur exacte de la mesure souhaitée: dépendamment de l'usage du modèle, il peut être préférable que la variable de sortie soit une valeur de classification ou la différence entre la valeur d'aujourd'hui et de demain. Par exemple, si l'analyse des données effectuée préalablement révèle que la variable de sortie est fortement auto-corrélée, comme c'est souvent le cas de paramètres comme la couleur ou la turbidité de l'eau brute, il est préférable de ne pas prévoir directement la valeur de la variable, mais plutôt la différence entre la valeur d'aujourd'hui et de demain (Zhang et Stanley, 1997).

Une fois la variable de sortie sélectionnée, les variables d'entrée du modèle sont choisies parmi les variables disponibles dans la base de données. La sélection s'effectue en s'appuyant sur l'existence de relations connues ou suspectées avec la variable de sortie, la littérature scientifique et la disponibilité des données (Baxter *et al.,* 2002). Dans la plupart des applications des RNA, peu d'attention semble avoir été portée à la sélection des variables d'entrée. La raison principale est que les RNA appartiennent à la classe des approches « data-driven », alors que les méthodes statistiques conventionnelles sont « model-driven ». Contrairement aux approches « model-driven », les approches « data-driven » ont l'habilité de déterminer quels variables d'entrée du modèle sont critiques (Bowden *et al.*, 2000a; Maier et Dandy, 2000b). On peut donc inclure au départ toutes les variables disponibles. Les variables qui s'avéreront redondantes ou peu importantes seront enlevées lors des essais subséquents (Baxter *et al.,* 2002).

Présenter un grand nombre de variables d'entrée aux modèles RNA requiert généralement des réseaux de grande taille, ce qui a pour conséquence de nécessiter plus de données et de réduire la vitesse d'apprentissage (Bowden *et al.,* 2000a). Ce problème est exacerbé dans les applications de séries temporelles, où des variables de décalage temporel sont ajoutées pour chacune des variables d'entrée. Une connaissance experte du système à modéliser peut alors être mise à profit pour réduire le nombre de variables d'entrée et choisir le décalage maximal des variables sélectionnnées. Par exemple, l'expert peut inspecter visuellement la représentation graphique des variables pour détecter les relations potentielles entre les variables d'entrée et de sortie et les décalages à utiliser (Bowden *et al.,* 2000). Cette identification *a priori* des variables significatives est grandement répandue dans la modélisation connexionniste, mais elle est aussi très subjective et dépendante de la connaissance de l'expert. Des 43 articles examinés par Maier et Dandy (2000b), la majorité s'appuie sur la connaissance *a priori* du phénomène à modéliser pour choisir les variables d'entrée. Quant aux variables de décalage, requises pour prendre en compte la dynamique temporelle dans le cas des réseaux *feedforward*, elles n'ont souvent pas été considérées ou ont été choisies selon

les connaissances *a priori* des chercheurs. Des doutes peuvent alors être soulevés quant à l'optimalité des ensembles de variables d'entrée utilisées pour développer les modèles (Bowden *et al.*, 2000a). Conséquemment, il y a des avantages distincts à utiliser des techniques analytiques afin de déterminer les variables d'entrée des modèles connexionnistes multi-variables (Maier et Dandy, 2000a).

Les techniques analytiques peuvent être divisées en deux catégories, soit les techniques non-supervisées et les techniques supervisées. Les premières permettent de réduire le nombre de variables à évaluer en identifiant les similitudes entre les variables. Elles comprennent la technique « Self-Organizing Map » (SOM) et la technique de l'analyse en composantes principales (ACP). La SOM a été développé par Kohonen en 1982 et est un type de réseau de neurones à l'apprentissage non-supervisé. La SOM classe les variables par catégorie en projettant les données de façon non-linéaire dans un espace dimensionnel moindre et en les rassemblant en sous-groupes. Dans des applications de modélisation environnementale, la SOM a permis de classer des données selon leur similarité et ensuite d'éliminer les variables redondantes ou très corrélées (Bowden *et al.*, 2000a).

L'ACP permet aussi de réduire le nombre de variables, mais en transformant l'ensemble original de variables en un nouvel ensemble beaucoup plus petit où les nouvelles variables, ou composantes principales, véhiculent l'essentiel de l'information contenue dans l'ensemble original de données (Bowden *et al.*, 2000a). L'ACP localise de façon linéaire les directions de variance maximale dans les données d'entrée originales et fait une permutation circulaire des données le long des axes. Les composantes principales calculées sont indépendantes et ne contiennent pas de redondance (Bowden *et al.*, 2000a). En examinant la contribution des variables originales aux nouvelles composantes principales, l'ACP permet donc d'identifier facilement les variables qui véhiculent une information similaire. Le principal désavantage de l'ACP est d'être une technique linéaire, qui peut faillir à identifier des tendances non-linaires importantes véhiculées par les données (Statsoft, 1998).

Contrairement aux techniques non-supervisées, qui ne servent qu'à réduire le nombre de variables à examiner, les techniques supervisées permettent de sélectionner les variables significatives selon leur impact sur l'habilité de prévision du RNA. Elles comprennent les algorithmes génétiques (AG), une technique puissante grandement utilisée pour optimiser les variables dans le domaine des ressources en eau et en environnement (Bowden, 2000) et la construction par étape de RNA bi-variables (CE), une technique utilisée avec succès tant par Zhang et Stanley (1997) que par Maier et Dandy (2000b).

Les AG fonctionnent en créant tout d'abord une population initialement aléatoire de chaînes binaires représentant des sous-ensembles de variables. Ils évaluent ensuite leur justesse, i.e. la qualité de la solution qu'elles représentent, et éliminent les chaînes les moins prometteuses. Le processus d'évaluation est répété pour un nombre pré-déterminé de générations modifiées par des mutations et des croisements, à la fin duquel la meilleure chaîne est sélectionnée, d'où l'appellation d'algorithme génétique. L'évaluation des chaînes est faite en construisant des réseaux neuronaux probabilistes (*probabilistic neural networks* ou PNN) ou des réseaux neuronaux de régression généralisée (*generalized regression neural network* ou GRNN), selon que les AG soient utilisés pour les problèmes de classification ou les problèmes de régression (Statsoft, 1998).

À l'instar des AG, la technique de la construction par étape de RNA bi-variables (CE) tente d'identifier le meilleur sous-ensemble de variables d'entrée. Mais contrairement aux AG, qui examinent des sous-ensembles déjà construits, la CE examine les variables individuellement. La méthode employée consiste à développer N modèles bi-variables, chacun utilisant une seule des N variables d'entrée disponibles et la variable de sortie. Le modèle ayant la meilleure performance est retenu et l'effet de l'ajout d'une seconde variable sur la performance du modèle est évalué. Cette procédure est répétée en utilisant des modèles avec un nombre croissant de variables, jusqu'à ce que l'ajout d'une variable n'améliore pas plus la performance du modèle (Zhang et Stanley, 1997). Les désavantages de cette méthode sont qu'elle est exigeante en temps et en ressources

et qu'elle est incapable de saisir les effets synergiques de certaines combinaisons de variables, qui peuvent être insignifiantes individuellement (Maier et Dandy, 2000b).

Les techniques non-supervisées et supervisées sont souvent utilisées conjointement afin de réduire tout d'abord le bassin des variables potentielles pour ensuite faciliter la sélection des variables significatives. Bowden *et al.* (2000a) ont comparé plusieurs combinaisons de techniques non-supervisées et supervisées pour déterminer le sous-ensemble optimal de variables d'entrée pour la prévision de la concentration d'algues bleues-vertes (*Anabaena*) dans la rivière Murray à Morgan, 4 semaines en avance. Les techniques non-supervisées comprenaient le « self-organizing map » (SOM) et l'analyse en composantes principales (ACP), et les techniques supervisées, une combinaison d'algorithmes génétiques (AG) et de RNA (GN-RNA) et une approche de modélisation par étape de RNA. Ces techniques ont aussi été comparées à l'approche consistant à identifier des variables *a priori*, combinée avec l'approche AG-RNA et l'approche de modélisation par étape de RNA. Six modèles connexionnistes ont été construits par Bowden *et al.* (2000a) à partir des variables d'entrée identifiées par les combinaisons de techniques et ont permis d'évaluer la performance des techniques examinées.

Ils ont trouvé que la connaissance *a priori* du système à modéliser était la technique non-supervisée la plus efficace pour réduire le nombre de variables, suivie par l'ACP. Les meilleures prévisions ont été obtenues avec les AG plutôt que par l'approche par étapes. Bowden *et al.* (2000a) expliquent ces résultats par le fait que les AG sont capables de découvrir des combinaisons synergiques entre les variables d'entrée qu'il est difficile de trouver en utilisant des approches par essais-et-erreurs. Conséquemment, le modèle développé en utilisant les variables d'entrée identifiées *a priori* en combinaison avec l'hybride AG-RNA donne de meilleurs résultats que les autres modèles développés avec les variables d'entrée identifiées à partir des autres combinaisons de techniques non-supervisées/supervisées (Bowden *et al.*, 2000a).

Sélection des caractéristiques de l'architecture du réseau

Sélectionner les caractéristiques du réseau consiste à choisir le type de connexion et la géométrie du réseau. Il s'agit d'une tâche importante et difficile en raison des multiples possibilités. Les types de connexion, décrites brièvement à la section 1.2.2.2, comprennent les réseaux *feedforward* et récurrents. Malgré les avantages indéniables des réseaux récurrents pour les séries temporelles, Maier et Dandy (2000b) suggèrent de n'utiliser que les réseaux *feedforward* car :

- ils performent bien comparativement aux réseaux récurrents dans la plupart des applications pratiques;
- ils ont été utilisés presque exclusivement pour la prévision et la prédiction des variables de ressouces hydrauliques. D'ailleurs, la quasi-totalité des 43 applications examinées par Maier et Dandy (2000b) ont recours aux réseaux *feedforward* (Annexe 2);
- leur vitesse de traitement des données est parmi les plus rapides de tous les réseaux présentement en usage;
- les réseaux récurrents ne présentent pas clairement d'avantage pratique par rapport aux réseaux *feedforward* dont la fenêtre de temps est limitée.

Une fois le type de connexion sélectionné, le nombre de couches cachées est généralement fixé avant de déterminer le nombre de neurones dans chaque couche. Il s'agit de l'ordre habituel qu'emploient la majorité des chercheurs. Au départ, Maier et Dandy (2000b) suggèrent de n'utiliser qu'une seule couche cachée car les réseaux de neurones avec une couche cachée peuvent approximer n'importe quelle fonction continue, s'ils disposent de suffisamment de neurones. Mais dans la pratique, plusieurs fonctions sont difficiles à approximer avec une seule couche cachée en raison du nombre prohibitif de neurones requis. Le recours à plus d'une couche cachée est alors

souhaitable. Flood et Kartam (1994a) suggèrent par ailleurs d'utiliser deux couches cachées comme point de départ car leur utilisation offre une plus grande flexibilité et permet d'approximer des fonctions complexes avec moins de neurones. Dans les faits, aucune règle formelle ne peut être dégagée car la géométrie optimale du réseau dépend grandement de la problématique à modéliser. Le mieux reste donc de choisir le nombre de couches qui résulte en l'architecture la plus simple.

Une fois le nombre de couches cachées fixé, le nombre de neurones dans chaque couche est choisi. Le nombre de neurones de la couche d'entrée est égal au nombre de variables en entrée du modèle, alors que le nombre de noeuds dans la couche de sortie correspond au nombre de variables de sortie du modèle (Maier et Dandy, 2001). L'étape critique consiste donc à sélectionner le nombre optimal de neurones dans la (ou les) couche(s) cachée(s), l'optimalité étant définie comme le plus petit réseau pouvant capturer adéquatement la relation dans les données d'apprentissage. Mais les petits réseaux ont leurs limites : si le nombre de neurones dans la couche cachée est insuffisant, le réseau peut être incapable de converger lors de l'apprentissage. Wilson et Recknagel (2001) ont d'ailleurs conclu, après avoir exploré dans leur essais de modélisation l'impact du nombre de neurones dans la couche cachée, que la présence des neurones dans la couche cachée améliore la performance prédictive des modèles. Néanmoins, il est important de ne pas utiliser trop de neurones, car le réseau risque « d'apprendre par cœur » et ainsi de perdre sa capacité de généralisation (Maier et Dandy, 2001).

Comment trouver alors le juste milieu? Selon Maier et Dandy (2001), le nombre optimal de neurones des couches cachées est généralement trouvé par une approche d'essais et d'erreurs. Il existe des lignes directrices pouvant être suivies, mais ces dernières sont souvent contradictoires d'un chercheur à l'autre. Maier et Dandy (2000b, 2001) suggèrent de s'appuyer sur les travaux de Hecht-Nielsen (1987).

Selon ce dernier, la limite supérieure permettant de s'assurer que les réseaux de neurones soient capables d'approximer n'importe quelle fonction est:

$$N_C \leq 2\,N_E + 1 \quad \text{(Équation 3)}$$

où N_C est le nombre de neurones de la couche cachée et N_E est le nombre de neurones de la couche d'entrée (soit le nombre de variables en entrée). Cependant, afin d'éviter que les réseaux n' « apprennent par cœur » en raison d'un trop grand nombre de neurones, il est crucial de considérer le rapport entre le nombre d'échantillons d'apprentissage et le nombre de variables d'entrée. Roger et Dowla (1994) recommandent d'utiliser la limite supérieure suivante :

$$N_C \leq \frac{N_{EA}}{(N_E + 1)} \quad \text{(Équation 4)}$$

où N_{EA} est le nombre d'échantillons d'apprentissage. Le nombre de noeuds de la couche cachée peut être déterminé selon la plus petite valeur de N_C des deux équations ci-dessus. Maier et Dandy (2001) ont utilisé ces équations pour déterminer les géométries des modèles développés pour prévoir l'occurrence d'algues bleues-vertes dans la rivière Murray (Australie). La géométrie du modèle final sélectionné, qui utilisait le débit, la température et la couleur et les variables de décalage correspondantes en entrée, était de 20 neurones en entrée, 17 neurones cachés et 1 neurone en sortie (20-17-1). L'effet d'utiliser plutôt 5, 10, 23, 30 et 35 neurones cachés au lieu de 17 a aussi été examiné. Il a été trouvé que les différentes géométries ont eu un impact négligeable sur la performance du modèle. La même chose a été constatée pour l'étude portant sur la prévision de la salinité de la rivière Murray, pour laquelle trois différentes géométries ont été essayées (25-5-1, 25-15-1, 25-30-1).

Ces travaux s'inscrivent dans la même voie que les 43 applications examinées par Maier et Dandy (2000b). La vaste majorité n'a utilisé qu'une seule couche cachée et le nombre optimal de neurones dans la couche cachée a été obtenu généralement par essais et erreurs (Annexe 2). Dans certains cas, le nombre de neurones dans la couche cachée était supérieur aux limites théoriques suggérées par la littérature et dans d'autres, des

géométries fixes ont été utilisées sans justification des choix, soulevant des doutes quant à l'optimalité des résultats obtenus. Afin de pallier à ces problèmes, beaucoup d'efforts ont été investis dans le développement de procédures ajustant automatiquement la géométrie du réseau durant l'apprentissage. Le but de ces méthodes est de déterminer le plus petit réseau capable de représenter adéquatement la relation à modéliser. Une bonne revue des différentes procédures est présentée par Bebis and Georgiopoulos (1994) et résumée par Maier et Dandy (2000b). Elles comprennent les algorithmes d'élagage (*pruning*) et constructifs (*constructive*). Bien que ces méthodes soient avantageuses, elles ne sont pas sans problèmes. Par exemple, Maier et Dandy (2000b) mentionnent qu'avec ces méthodes, on n'obtient pas une indication juste de l'habilité de généralisation du réseau car l'impact de l'ajout ou du retrait d'un neurone sur la performance du réseau est évalué en utilisant les données d'apprentissage du réseau. De plus, les algorithmes d'élagage sont sensibles aux valeurs attribuées aux paramètres internes et leur usage est restreint aux réseaux d'une couche cachée. Des 43 applications révisées par Maier et Dandy (2000b), six ont eu recours à des algorithmes qui optimisent automatiquement la géométrie du réseau.

Plusieurs logiciels commerciaux sont disponibles afin d'assister les modélisateurs dans la sélection des architectures optimales des réseaux. Règle générale, les logiciels qui ont une philosophie « boîte noire » où peu de choix peuvent être effectués dans la construction du modèle devraient être évités en faveur des logiciels qui donnent plus de flexibilité au modélisateur. Baxter *et al.* (2002) proposent les logiciels NeuroShell 2 de Ward Systems Group Inc. (Frederick, Maryland) et Statistica Neural Networks de Statsoft Inc. (Tulsa, Oklahoma), tous deux utilisés avec succès dans le développement de modèles de RNA dans l'industrie du traitement de l'eau. Ces deux logiciels intègrent des outils qui utilisent des techniques d'optimisation non-linéaire permettant de trouver par recherche heuristique les meilleures architectures. Sinon, il est toujours possible de débuter la construction du modèle avec une architecture perceptron multicouches de trois couches, une architecture standard largement utilisée (Baxter *et al.,* 2002).

Choix de la méthode d'apprentissage

Choisir la méthode d'apprentissage consiste à sélectioner l'algorithme d'apprentissage qui optimisera les poids de connexion entre les neurones. L'algorithme de rétro-propagation est de loin la méthode la plus populaire. Des 43 articles examinés par Maier et Dandy (2000b), 36 y ont eu recours, mais très peu ont justifié leur choix (Annexe 2). Comme la justesse de la prévision a été adoptée comme critère de performance dans la majorité des applications, le choix de l'algorithme de rétro-propagation est tout de même justifié, son désavantage principal étant sa vitesse d'apprentissage peu rapide. Le choix du meilleur algorithme d'apprentissage peut être facilité en utilisant un logiciel commercial, car ce dernier recherche automatiquement l'algorithme le plus approprié, tout en testant différentes valeurs de paramètres internes.

Choix du critère d'arrêt

Le critère d'arrêt sert à décider quand arrêter le processus d'apprentissage. Idéalement, l'apprentissage doit cesser quand le modèle est optimal mais il arrive aussi qu'il soit arrêté quand il est sous-optimal, comme c'est le cas lorsqu'il y a surapprentissage. Les logiciels commerciaux informent alors l'utilisateur de la suboptimalité du modèle. En général, l'apprentissage est arrêté quand l'erreur d'apprentissage a atteint une valeur suffisamment faible et que les variations entre les valeurs d'erreur sont faibles (Maier et Dandy, 2000b). La validation croisée est une excellente méthode pour déterminer le moment où l'apprentissage est complété (voir section 1.2.2.3). Dans cette méthode, le critère d'arrêt consiste à cesser l'apprentissage quand l'erreur de l'ensemble de test augmente de nouveau après avoir maintenu un plateau de faibles valeurs d'erreur. Mais comme il est difficile de savoir si le minimum atteint n'est pas un minimum local et que l'apprentissage n'a pas été arrêté prématurément, Maier et Dandy (2000b) suggèrent de continuer un peu le processus d'apprentissage une fois que l'erreur de l'ensemble de test commence à augmenter.

Malgré son efficacité, la méthode de validation croisée est peu utilisée comme critère d'arrêt. Seulement 8 articles sur 43 en font mention dans la revue de Maier et Dandy (2000b). Dans près du tiers des articles, l'apprentissage a été arrêté suite à un nombre pré-fixé d'itérations tandis que dans 11 autres articles, l'apprentissage a cessé une fois que l'erreur d'apprentissage a atteint une valeur pré-déterminée. Finalement, 11 articles ne font mention d'aucun critère d'arrêt, et ce, malgré son importance pour justifier l'optimalité des modèles développés.

1.2.4.5 Évaluation de la performance des modèles candidats

Une fois l'apprentissage complété, la performance de chaque modèle candidat a besoin d'être validée en utilisant le critère de performance sélectionné au départ (voir section 1.2.4.2.). Si l'erreur obtenue avec l'ensemble de données de validation est significativement différente de celle obtenue avec l'ensemble de données d'apprentissage, il est probable que les deux ensembles de données ne soient pas représentatifs de la même population ou que le modèle ait souffert de surapprentissage. Selon Maier et Dandy (2000b), une mauvaise validation peut être due à l'architecture du modèle ou à un prétraitement des données et à une normalisation des données d'apprentissage ou de validation manquants ou inadéquats.

Il faut aussi s'assurer que la performance de prévision des modèles candidats est indépendante de la manière dont les données ont été séparées en trois ensembles. Baxter *et al.* (2002) suggèrent de redistribuer différemment les données entre les trois ensembles et de ré-entraîner les modèles candidats avec ces nouveaux ensembles de données afin de comparer les erreurs. Une augmentation significative de l'erreur de prévision du nouvel ensemble de validation est une indication de l'instabilité du modèle. Les meilleurs modèles candidats auront des erreurs semblables lorsque les ensembles de données sont redistribués.

Finalement, le choix du meilleur modèle parmi les modèles candidats développés demeure une étape difficile. Selon Coulibaly *et al.* (1999), de nombreuses méthodes

heuristiques ont été proposées pour optimiser l'identification du meilleur modèle neuronal, mais toutes ces approches méthodologiques sont expérimentales et s'appliquent à des cas spécifiques. En fait, il n'existerait pas de méthode universelle pour l'identification a priori du meilleur modèle neuronal, parce que la complexité de chaque modèle est inhérente à celle du système considéré. Il faut donc que l'utilisateur choisisse selon son jugement, selon l'usage qui sera fait du modèle, la facilité à obtenir les variables utilisées par le modèle et la complexité du réseau, le plus simple étant souvent le mieux. Pour aider les utilisateurs à choisir, Zhang et Stanley (1997) et Maier *et al.* (1998) suggèrent d'analyser graphiquement les résultats des modèles candidats afin de détecter visuellement quel modèle prédit le mieux les larges fluctuations et présente le moins de décalages, particulièrement lorsque les performances des modèles candidats sont semblables.

1.2.5 Applications dans le domaines du génie de l'environnement et des ressources hydriques

Selon Maier et Dandy (1997), French *et al.* (1992) ont été les premiers à publier dans un journal scientifique d'envergure internationale un article portant sur l'application des RNA à un problème hydrologique. Il s'agissait de prévoir les patrons temporel et spatial des précipitations de pluie. Depuis, plusieurs chercheurs ont utilisé les RNA pour modéliser, prévoir ou prédire des phénomènes environnementaux complexes. Maier et Dandy (2000a), Coulibaly *et al.* (1999) et Adeli (2001) ont effectué des revues d'articles publiés dans des journaux scientifiques internationaux qui portaient respectivement sur la modélisation connexionniste de ressources hydrauliques, de variables hydrométéorologiques et d'applications en génie de l'environnement et des ressources hydriques. La distribution des 43 articles recencés par Maier et Dandy (2000b) entre 1992 et 1998 illustre bien l'accroissement de la popularité des RNA à travers les années (Tableau 1-2). En raison du nombre imposant d'articles, il est difficile d'en faire une revue exhaustive dans cette section. Conséquemment, seules les applications de prévision hydro-météorologique dont l'approche, la méthodologie, les

résultats ou les conclusions sont particulièrement pertinents à la conduite du présent projet de maîtrise seront examinées. Les autres articles consultés, jugés moins pertinents mais néanmoins intéressants, seront mentionnés afin d'être consultés si désiré.

Tableau 1-2 : Distribution des articles rescencés par Maier et Dandy (2000b)

Année de la publication	Nombre d'articles
1998	10
1997	17
1996	7
1995	4
1994	2
1993	1
1992	2

1.2.5.1 Prévision du débit des rivières

Plus de la moitié des applications répertoriées concernent la prévision des débits, crues et étiages des rivières, de même que les apports naturels dans les réservoirs hydroélectriques (Coulibaly *et al.*, 1999). Les motivations derrière le développement de modèles connexionnistes pour prévoir les débits sont souvent d'ordre économique, humain ou écologique. Par exemple, l'équipe de Elshorbagy *et al.* (2000) a évalué la performance des réseaux neuronaux pour la prévision du ruissellement printanier dans la région du Sud du Manitoba, une région affectée par des débordements importants de la rivière Rouge survenant à la suite d'une fonte des neiges trop rapide combinée à de fortes précipitations printanières. Les modèles connexionnistes développés par Elshorbagy *et al.* (2000) ont été comparés à des techniques traditionnelles de régression linéaire et non-linéaire. Selon leur étude, la performance des modèles connexionnistes s'est avérée comparable ou supérieure dans tous les cas examinés. Liong *et al.* (2000) ont utilisés les RNA afin de prévoir avec exactitude le niveau de l'eau de la rivière Buriganga, près de Dhaka, au Bangladesh. Les auteurs concluent que les réseaux neuronaux constituent un outil de prévision permettant d'émettre efficacement des alertes de crue tout en réduisant les coûts associés aux collectes de données aux stations de jaugeage.

D'un point de vue écologique, on peut évaluer l'impact des fluctuations de débit sur différentes espèces animales, végétales ou aquatiques. Par exemple, Karunanithi *et al.* (1994) ont utilisés les RNA pour prévoir les fluctuations de débit à une station non-jaugée de la rivière Huron (Michigan), qui abrite des achigans à petite bouche. Contrairement à la majorité des chercheurs qui développent les RNA grâce à l'algorithme de rétro-propagation, l'équipe de Karunanithi *et al.* (1994) ont utilisé l'algorithme de « cascade-correlation » pour développer leur modèle. Cet algorithme a l'avantage de modifier la topologie du modèle lors de l'apprentissage. Les résultats obtenus par le modèle connexionniste ont été comparés à ceux d'un modèle de puissance. Le modèle connexionniste s'avère supérieur, particulièrement pour la prévision des débits de pointe.

1.2.5.2 Prévision des précipitations

La prévision des précipitations a été le sujet d'étude de plusieurs autres chercheurs car une bonne estimation de la distribution spatio-temporelle des précipitations représente un atout capital pour la prévision des débits et crues des rivières et l'évaluation du bilan des grands bassins versants (Coulibaly *et al.,* 1999). Depuis plus de quarante ans, les précipitations sont estimées en utilisant des relations empiriques prenant en compte des mesures effectuées au radar ou selon des relations entre la distribution probalistique des précipitations au sol et les mesures au radar. Xiao et Chandrasekar (1997) ont introduit les réseaux neuronaux pour établir de nouvelles relations entre les mesures au radar et les précipitations. Les résultats obtenus à partir des modèles connexionnistes sont meilleurs que ceux obtenus avec les méthodes traditionnelles, offrant ainsi une approche alternative très intéressante à la prévision météorologique.

1.2.5.3 Prévision de paramètres indicateurs de qualité de l'eau brute

Selon Coulibaly *et al.* (1999), la prévision de la qualité de l'eau représente environ 20% des applications des RNA en hydrologie. Les paramètres indicateurs examinés comprennent entre autres les concentrations d'algues (Scardi, 2001; Wilson et

Recknagel, 2001; Maier *et al.,* 1998), de bactéries (Brion et Lingireddy, 1999) et de protozoaires (Brion *et al.*, 2001 et Neelakantan *et al.,* 2001), de niveaux de salinité (Maier et Dandy, 1996) et de la couleur (Zhang et Stanley, 1997). Bien qu'il soit très difficile d'en être certain en raison du nombre impressionnant de journaux scientifiques internationaux, aucune étude portant sur un sujet identique à notre problématique, i.e. la prévision à court-terme de la turbidité de l'eau brute, n'a été relevée.

L'étude de Zhang et Stanley (1997) est celle dont la problématique se rapproche le plus de celle exposée dans ce mémoire. Elle sera donc examinée en détail. Zhang et Stanley (1997) ont développé un modèle connexionniste pouvant prévoir 24h à l'avance la couleur de l'eau brute de la rivière Saskatchewan Nord, une rivière importante qui approvisionne la station de traitement Rossdale, à Edmonton (Alberta). La qualité globale de l'eau brute de la rivière est bonne, mais durant les orages estivaux et le dégel printanier, les paramètres de qualité de l'eau brute telle la couleur peuvent varier de plusieurs ordres de grandeur en une seule journée. Le but du modèle était de fournir aux opérateurs de la station un avertissement précoce des fluctuations de qualité de l'eau brute et, ainsi, d'augmenter l'efficacité du traitement.

Zhang et Stanley (1997) ont choisi d'utiliser la modélisation connexionniste en raison de la non-linéarité inhérente des valeurs de la couleur à travers l'année et de la disponibilité de données de qualité de l'eau. Au départ, les auteurs ont développé un modèle neuronal de type récurrent car la couleur du jour même étant fortement corrélée à la valeur des journées précédentes. Mais ce modèle a été un échec car, plutôt que de prévoir la valeur de la couleur du lendemain, le réseau prévoyait une valeur très près de la valeur de la journée même. La forte auto-corrélation des données de couleur semblait confondre le réseau récurrent. Pour pallier à ce problème, ils ont construit des modèles candidats de type semi-récurrent qui prédisaient la valeur du lendemain en ajoutant à la valeur historique d'aujourd'hui la prévision de la différence entre la valeur d'aujourd'hui et du lendemain.

Quatre ans de données quotidiennes ont été utilisés afin de développer les modèles et un an pour les évaluer. En observant les données, les auteurs ont répertorié quatre « saisons » distinctes dont le début varie d'une année à l'autre mais dont les caractéristiques de la couleur de l'eau brute sont assez consistantes (Figure 1-9). Les données ont été réparties en trois sous-ensembles selon les « saisons » afin que les sous-ensembles d'apprentissage et de test aient tous deux des patrons représentatifs de variations de couleur de l'eau brute.

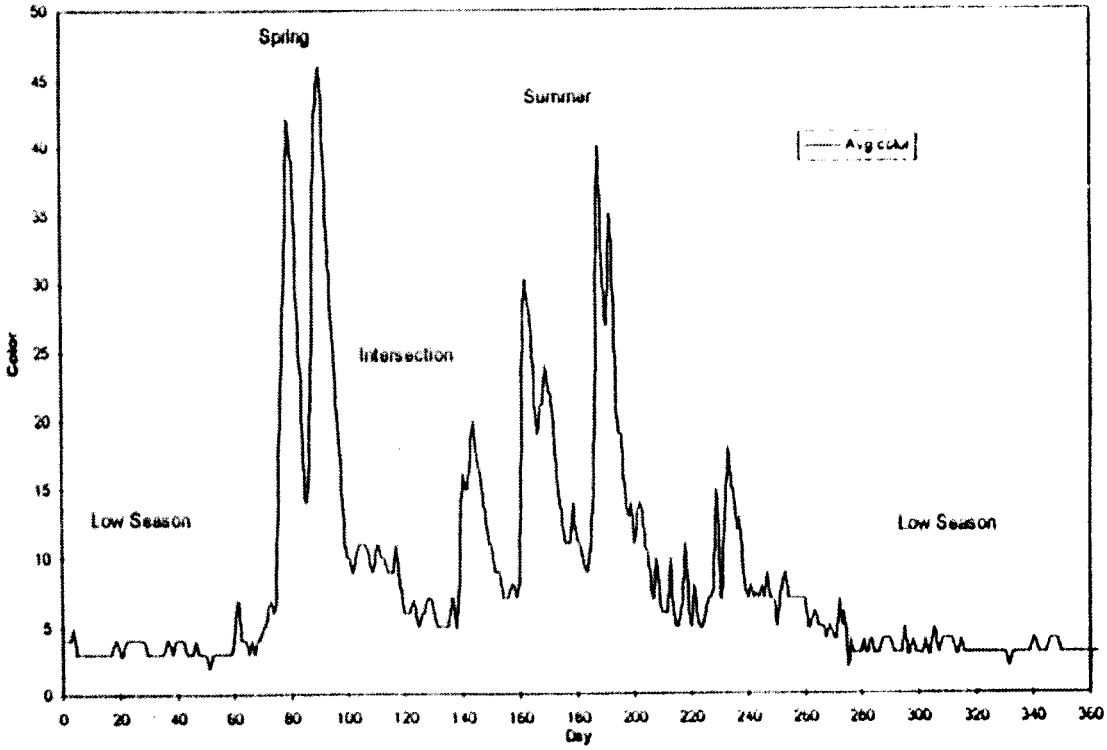


Figure 1-9 Cycle annuel des patrons de couleur de l'eau brute en 1994 (provient de Zhang et Stanley, 1997).

Aucune information quant à l'architecture des modèles n'est fournie dans l'article. Les modèles candidats résultent de divisions différentes des données et des variables d'entrée, et non des architectures différentes à tester. Le meilleur modèle a été développé en utilisant 11 variables d'entrée (Tableau 1-2). Les paramètres de qualité d'eau brute ont été utilisés pour identifier les conditions actuelles de la qualité de l'eau de la rivière. Les paramètres de série chronologique, qui comprennent les paramètres de

décalage et les paramètres de taux de changement, sont utilisés pour prendre en compte la corrélation existant entre les valeurs successives de couleur. Les paramètres environnementaux aident le modèle à intégrer les contributions importantes à la couleur dues à des phénomènes saisonniers comme le dégel printanier et au ruissellement. Les paramètres d'index permettent au modèle d'établir une ligne de base saisonnière. Chaque paramètre d'index a une valeur de 0 ou 1. Au printemps, quand la couleur de la rivière est due principalement au ruissellement, l'index du printemps a une valeur de 1 alors que le paramètre de l'automne a une valeur de 0. Durant l'été, alors que la couleur de l'eau de la rivière est due principalement aux pluies et au ruissellement, l'index d'été est de 1 et celui du printemps, 0.

Tableau 1-2: Paramètres d'entrée du modèle prévoyant la couleur de l'eau brute de la rivière Saskatchewan Nord à Edmonton (Zhang et Stanley, 1997)

Paramètre	Classification
Débit de la rivière (m^3/s)	Paramètre de qualité d'eau brute
Couleur à la prise d'eau Rossdale (UC)	Paramètre de qualité d'eau brute
Débit de la rivière (m^3/s) de la veille	Paramètre de série chronologique
Couleur à la prise d'eau Rossdale (UC) de la veille	Paramètre de série chronologique
Taux de changement de couleur (UC d^{-1})	Paramètre de série chronologique
Turbidité à la prise d'eau Rossdale (UTN) de la veille	Paramètre de série chronologique
Turbidité à la prise d'eau Rossdale (UTN) d'il y a deux jours	Paramètre de série chronologique
Précipitations de pluie (mm)	Paramètre environnemental
Température degré jour (°C)	Paramètre environnemental
Index du printemps	Paramètre d'index
Index de l'été	Paramètre d'index

Le meilleur modèle développé a prédit la couleur de l'eau brute à la prise d'eau de la station Rossdale 24 heures à l'avance avec une erreur quadratique moyenne de 0.962 UC. Lorsqu'il a été exposé aux données de test, le modèle a été capable de prévoir adéquatement les augmentations subites de couleur, sans phénomène de décalage. En tout, le modèle a été capable de bien prévoir 355 points des 365 auxquels il n'a pas été exposé lors de l'apprentissage. Depuis son développement, le modèle a été implanté en

ligne à la station Rossdale, permettant aux opérateurs d'anticiper avec succès les fluctuations importantes de la qualité de l'eau brute et d'ajuster en conséquence les opérations de procédé en accord.

Zhang et Stanley (1997) ont présenté des pistes intéressantes pour le choix des variables d'entrée et une façon d'exprimer la valeur de sortie pour éviter des problèmes d'auto-corrélation. L'équipe de Brion et Lingirreddy (1999), à laquelle s'ajoute ensuite Neelakantan (Neelakantan *et al.,* 2001), propose une autre façon fort intéressante d'exprimer les valeurs de sortie. Brion et Lingirredy (1999) ont utilisé les RNA pour distinguer la contamination fécale d'origine urbaine de celle d'origine agricole présente dans un réservoir d'eau potable à Lexington, Kentucky. Brion et Lingireddy (2001) et Neelakantan *et al.* (2001) ont aussi entraîné des modèles permettant de prévoir les concentrations de pointe d'oocystes de *Cryptosporidium* et de kystes de *Giardia* à la prise d'eau d'une station de traitement de la rivière Delaware, dans le New Jersey. Ce qui est particulier dans les deux cas, c'est que le but du modèle était de prévoir qualitativement plutôt que quantitativement la source de contamination et les concentrations de kystes de protozoaires. Ainsi, pour identifier l'origine de la contamination fécale, les valeurs de la variable de sortie se sont vues attribuées les valeurs de 1,0, 0,5 et 0,0 pour les sites urbains, mixtes et agricoles, respectivement. Lors de l'évaluation du modèle, si la prévision tombait entre 0,0 et 0,33, la nature de la source de contamination était classifiée comme agricole, entre 0,34 et 0,66, comme mixte et entre 0,67 et 1,0, comme urbaine. Dans les études portant sur la prévision des concentrations de *Cryptosporidium* et de *Giardia*, les percentiles calculés à partir des valeurs de concentrations de kystes et d'oocystes de la base de données et la présence de kystes et d'oocystes avec présence d'organes internes ont servi à définir des classes, dans lesquelles ont été réparties les valeurs de concentrations. Les classes consistaient en une classe « valeur de fond » (*background*), en une classe « valeur normale » et une dernière classe « valeur élevée ». L'échelle différait pour les deux types de protozoaires car un modèle distinct était développé pour chacun. Les modèles finaux prédisaient deux conditions de concentrations, soit la condition « valeur de fond » (*background*),

qui comprenait la première classe et la condition « supérieure à la valeur de fond » (*above-background*), dont faisaient partie les deux autres classes. Ces conditions étaient désignées respectivement par 1 et 0 et les prévisions correspondantes étaient comprises dans les échelles 0,7-1,0 et 0,0-0,3.

La façon de classifier les données de certaines variables d'entrée est aussi intéressante. Dans l'étude sur l'identification de l'origine de la contamination fécale, les données de pluies ont été classées selon l'intensité des précipitations: les journées pluvieuses, définies comme celles où plus de 2 cm de pluie étaient tombées dans les derniers 24 heures ou plus de 3 cm dans les derniers 48 heures, se voyaient accordées une valeur de 1 et les autres journées non ou peu pluvieuses, une valeur de 0.

L'architecture des réseaux développés est standard (perceptron multicouches à une couche cachée). Le nombre de neurones de la couche cachée était fixé au double du total du nombre de variables d'entrée et de sortie. Dans les études sur la prévision de la concentration des kystes de protozoaires, les auteurs ont tout de même testé l'effet d'utiliser un nombre de neurones cachés égal au nombre de variables d'entrée, mais les résultats étaient insatisfaisants. Trois sous-ensembles ont été constitués de façon aléatoire à partir de 106 échantillons pour l'identification de l'origine de la contamination et de 68 échantillons pour la prédiction de la concentration des kystes de protozoaires. Selon l'expérience des auteurs, un bon ensemble d'apprentissage devrait plutôt comprendre entre 200 et 400 échantillons pour des réseaux de la taille utilisée. C'est pourquoi, dans ce contexte, il peut être profitable d'assigner des classes de valeurs plutôt que des valeurs singulières aux variables de sortie afin d'améliorer la performance des modèles.

La sélection des variables d'entrée est aussi intéressante. Pour l'identification des sources de contamination, les variables employées sont des paramètres microbiologiques mesurés lors d'échantillonnage de routine (coliformes fécaux, streptocoques fécaux, coliformes totaux) et des paramètres météorologiques (pluie) facilement accessibles au grand public. C'est d'ailleurs cette disponibilité des données

qui a justifié le recours aux RNA plutôt qu'à d'autres indicateurs, plus reconnus mais aussi plus coûteux et complexes. Dans les études portant sur la prévision de la concentration des kystes de protozoaires, les variables d'entrée des modèles ont été sélectionnés selon leur corrélation avec les concentrations d'oocystes de *Cryptospsoridium* et de kystes de *Giardia* et selon les résultats d'une analyse de sensibilité qui détermine l'importance de chacune des variables d'entrée potentielles. La sélection des variables d'entrée selon leur corrélation avec la variable de sortie est une méthode plutôt novatrice car, habituellement, les variables d'entrée sont choisies selon une relation de cause-à-effet présumée avec la variable de sortie (section 1.2.4.4). Scardi (2001) utilise aussi les corrélations entre les variables pour améliorer la performance d'un modèle développé précédemment afin de prévoir la production du phytoplancton. Il désigne par le terme « co-prédicteurs » les variables d'entrée du modèle qui ne sont pas directement liées à la variable de sortie, mais qui fournissent des informations permettant d'améliorer l'exactitude des prévisions. Par exemple, l'information bathymétrique, qui est corrélée de façon significative avec la production primaire de phytoplancton (r = -0.318, P >0.01, n=2522), permet d'améliorer les prévisions quand elle est ajoutée au modèle neuronal développé précédemment, même si la profondeur n'affecte pas directement le processus de la photosynthèse.

La portabilité d'un modèle à d'autres sites de caractéristiques semblables est une préoccupation importante pour de nombreux chercheurs. Ainsi, Brion et Lingirreddy (1999) ont évalué la performance du modèle d'identification des sources de contamination alors qu'il était exposé à des données d'un autre bassin versant que celui utilisé pour développer le modèle, mais dont les caractéristiques étaient similaires (taille, conditions climatiques, location géographique, etc.). Les auteurs ont été impressionnés par la bonne performance du modèle, qui a su bien réagir à la variabilité présentée par les données de ce site.

Wilson et Recknagel (2001) sont allés plus loin dans la tentative de développer un modèle pouvant s'appliquer à différents sites. Ils ont développé un modèle

connexionniste générique de prévision de croissance algale en utilisant des données de six lacs et réservoirs de pays différents. En utilisant des données historiques de variables-clé de la croissance algale comme le phosphore, l'azote, la lumière subaquatique et la température de l'eau de six lacs d'eau douce différents, ils ont développé deux modèles : un pour des prévisions pour la journée même et un autre pour des prévisions dans 30 jours. Le modèle prédictif immédiat a produit des résultats satisfaisants, supérieurs aux résultats obtenus avec le modèle prédictif de 30 jours. Wilson et Recknagel (2001) suggéraient différentes façons d'améliorer le modèle de prévision à moyen-terme, dont réduire le délai de prévision, modéliser la différence entre la production algale de deux journées consécutives plutôt que la production absolue, utiliser plus de variables de décalage pour mieux capturer les changements pendant la période de prévision et utiliser un réseau récurrent, plus approprié théoriquement dans le cas des séries chronologiques.

Le recours aux variables de décalage temporel pour améliorer la performance des réseaux feedforward est mentionné par plusieurs chercheurs, mais surtout pour les prévisions à moyen et long terme. Par exemple, Maier *et al.* (1998) ont examinés l'impact de l'utilisation de variables de décalage sur la performance des RNA dans l'étude où ils prévoient avec succès, quatre semaines à l'avance, l'incidence et l'amplitude des pointes de croissance de cyanobactéries dans la rivière Murray, en Australie. Ils concluent que les variables de décalage améliorent grandement la représentation des patrons véhiculés par les données historiques.

L'équipe de Maier, Dandy et Bowden (Maier et Dandy, 2000a, 2000b, Bowden *et al.*, 2000a, 2000b, 2001, 2002) ont aussi exploré d'autres avenues dans les études qu'ils ont conduits et qui ont mené à la rédaction des ouvrages et articles ayant en partie nourri la revue méthodologique de la section précédente (voir section 1.2.4). Outre la prévision à moyen terme de la concentration de cyanobactéries, Maier et Dandy (1996) se sont aussi intéressés à la salinité, un autre paramètre qui a un impact significatif sur la qualité de l'eau brute de la rivière Murray. Au départ, tant pour la prévision du degré de salinité

que de la croissance des cyanobactéries, ces chercheurs ont préféré employer les approches les plus simples côté méthodologie. Ils ont donc utilisé leurs connaissances des phénomènes en cause pour choisir les variables d'entrée et les variables de décalage. La géométrie des réseaux a été fixée selon les indications données par la littérature (voir section 1.2.4.4). Les valeurs des paramètres internes des réseaux ont été déterminées selon les valeurs suggérées par défaut dans le logiciel commercial utilisé. L'algorithme d'apprentissage était dans les deux cas l'agorithme de rétro-propagation. L'erreur quadratique moyenne et la validation croisée servaient respectivement de critère de performance et de critère d'arrêt. Ce n'est que par la suite que les chercheurs ont tenté d'améliorer les modèles de diverses façons. Pour la prévision de la salinité de la rivière Murray, Maier et Dandy (1997) ont raffiné le choix des variables d'entrée afin de réduire leur nombre et d'améliorer la performance du modèle. Ils ont ensuite étudié l'effet de la géométrie et des paramètres internes sur la performance du modèle (Maier et Dandy, 1998b) et tenté d'optimiser la performance de l'algorithme d'apprentissage de rétro-propagation (Maier et Dandy, 1998a). Ils ont ensuite comparé différentes méthodes d'apprentissage pour le réseau (Maier et Dandy, 1999), avant de se pencher sur la division optimale des données (Bowden *et al.,* 2002) et leur transformation (Bowden *et al.,* 2001). Pour la prévision des cyanbactéries *Anabaena*, plusieurs techniques ont été comparées afin d'optimiser l'ensemble de variables d'entrée du modèle (Bowden *et al.,* 2000a). Ensuite, des modèles développés à partir de réseaux « mémoire associative B-spline » ont été comparés à des modèles développés à partir de réseaux perceptrons multicouches, les premiers ayant comme avantage d'exprimer explicitement les relations existant entre les variables d'entrée et de sortie.

Les faits saillants et les conclusions pertinentes de toutes ces études sont présentés dans la section portant sur la méthodologie (section 1.2.4). Ce qui ressort du survol rapide des articles, c'est la démarche des chercheurs, qui pourrait se résumer ainsi : il vaut mieux commencer simplement et raffiner les modèles par la suite selon les problèmes rencontrés lors du développement du modèle initial. Les façons d'améliorer un modèle

sont nombreuses, mais elles sont souvent spécifiques à la situation examinée, comme le mentionnent souvent d'ailleurs les chercheurs.

1.2.5.4 Prévision et modélisation de procédés de traitement de l'eau

L'utilisation des RNA dans l'industrie du traitement de l'eau est aussi à la hausse. Les applications incluent la prévision de la demande en eau potable (Baxter *et al.,* 2001), la prévision de l'enlèvement de la couleur et de la turbidité lors de la coagulation (Baxter *et al.*, 2001), la prévision des doses l'alun et de polymères à utiliser lors du procédé de coagulation (Joo *et al.*, 2000, Baxter *et al.,* 2001), la prévision de performance de la filtration (Baxter *et al.,* 2001), la prévision de l'adoucissement de l'eau (Baxter *et al.,* 2001) et la prévision du chlore résiduel dans le réseau de distribution (Sérodes et Rodriguez, 1996, Rodriguez et Sérodes, 1999, Skipworth *et al.,* 1999). Outre la prévision et la prévision, les RNA peuvent aussi être utilisés pour modéliser et simuler des phénomènes complexes ou peu compris, pour détecter les défaillances et désigner et implanter des stratégies de contrôle des procédés (Adgar *et al.,* 2000).

Dans le domaine du traitement de l'eau, les motivations derrière le développement de modèles connexionnistes prédictifs sont souvent d'ordre économique. Par exemple, dans l'industrie du traitement de l'eau, le coût de l'électricité dans les systèmes de distribution représente une fraction importante des budgets d'opération des stations de traitement. En prévoyant la demande, il devient possible de prendre avantage des différentes tarifications pour réduire les coûts en énergie. Baxter *et al.* (2001) ont développé deux modèles connexionnistes pour tenter de prévoir la demande en eau de la Ville d'Edmonton (Alberta) 24 h et 12 jours à l'avance. À l'instar de la prévision de la qualité d'eau brute (voir section 1.2.5.3), les variables d'entrée sont des paramètres météorologiques et des paramètres d'index sélectionnés selon leur disponibilité et la connaissance des auteurs du phénomène à prévoir (Tableau 1-3). Les modèles de contruction standard (une couche cachée, algorithme de rétro-propagation) ont donné d'excellents résultats dans le cas de la prévision 24 h (r^2 de 0.90) mais de moins bons résultats dans le cas du modèle de prévision de 12 jours (r^2 de 0.49).

Tableau 1-3 Paramètres d'entrée du modèle retenu pour la prévision de la demande en eau à Edmonton (Baxter *et al.*, 2001)

Demande en eau quotidienne	**Demande en eau dans 12 jours**
Température quotidienne minimale (°C)	Température quotidienne minimale (°C)
Température quotidienne maximale (°C)	Température quotidienne maximale (°C)
Précipitations du jour précédent (mm)	Précipitations du jour précédent (mm)
Précipitations des cinq jours précédents (mm)	Précipitations des cinq jours précédents (mm)
Précipitations des trente jours précédents (mm)	Précipitations des trente jours précédents (mm)
Index de l'été	Index de l'été
Index de semaine / fin de semaine	Index de semaine / fin de semaine
Demande en eau de 9h00 à 10h00	

Réduire la quantité de produits chimiques utilisés dans les différentes filières de traitement est également une préoccupation constante tant du point de vue économique que sanitaire. À l'étape de la coagulation, un coagulant chimique comme l'alun est ajoutée à l'eau brute pour déstabiliser la matière particulaire en suspension afin qu'elle puisse éventuellement former des flocs assez gros pour sédimenter dans le bassin de décantation. Les doses de coagulant sont souvent déterminées par des essais de flocculation en laboratoire et selon l'expérience de l'opérateur. Lorsque la qualité de l'eau brute fluctue, il est difficile d'ajuster rapidement le dosage de coagulant. Récemment, des modèles connexionnistes ont été développés afin de prévoir rapidement et précisément la dose de coagulant à utiliser. Joo *et al.* (2000) ont utilisé 142 échantillons choisis à partir de deux ans de données pour entraîner un modèle neuronal à l'architecture simple (5-10-1). Les variables d'entrée comprenaient un paramètre de procédé, quatre paramètres de qualité d'eau brute et aucun paramètre chronologique (Tableau 1-4). Les valeurs de turbidité dans l'application examinée varient grandement, surtout pendant la saison des pluies, ce qui rend l'ajustement de la dose de coagulant particulièrement difficile.

Afin de comparer la performance du modèle neuronal, un modèle de régression multi-variable a aussi été développé en utilisant les mêmes données historiques. L'erreur de prévision du modèle neuronal était de 52% inférieure à celle obtenue par le modèle de

régression multi-variable. Le modèle neuronal a donné aussi de meilleurs résultats lorsqu'il y a des changements drastiques dans le dosage de coagulant.

Tableau 1-4 Paramètres d'entrée des modèles retenus pour la prévision des doses de coagulant (à partir de Joo *et al.*, 2000)

Variables d'entrée du modèle de prévision de la turbidité	**Classification**
pH de l'affluent	Paramètre de qualité d'eau brute
Turbidité de l'affluent (UTN)	Paramètre de qualité d'eau brute
Température de l'affluent (°C)	Paramètre de qualité d'eau brute
Alcalinité de l'affluent ($mg \cdot L^{-1}$)	Paramètre de qualité d'eau brute
Débit d'injection du coagulant	Paramètre de procédé

Baxter *et al.* (2001) ont aussi développé un modèle neuronal prédisant la dose de coagulant à utiliser, mais en utilisant une approche différente de Joo *et al.* (2000). Plutôt que de développer un seul modèle prédisant directement la dose de coagulant, ils ont développé deux modèles, soit un modèle de prévision de la turbidité de l'eau à la sortie du bassin de décantation et un modèle de prévision de la dose d'alun, le coagulant employé dans cette étude. Les deux modèles ont été développés en utilisant trois ans de données quotidiennes. Les variables d'entrée utilisées pour chaque modèle sont présentées dans le Tableau 1-5. Les variables d'entrée du modèle de prévision de la turbidité de l'effluent incluent des paramètres de qualité d'eau brute, des paramètres opérationnels et des paramètres de série chronologique. Pour le modèle de prévision du dosage d'alun, la turbidité de l'effluent du clarificateur devient une variable d'entrée et le dosage d'alun devient la variable de sortie.

Tableau 1-5 Paramètres d'entrée des modèles retenus pour la prévision de la turbidité de l'effluent du bassin de décantation et des dosages de coagulant (alun) (Baxter *et al.*, 2001)

Variables d'entrée du modèle de prévision de la turbidité	Variables d'entrée du modèle de prévision de la dose d'alum
pH de l'affluent	pH de l'affluent
Turbidité de l'affluent (UTN)	Turbidité de l'affluent (UTN)
Température de l'affluent (°C)	Température de l'affluent (°C)
Couleur de l'affluent (UTC)	Couleur de l'affluent (UTC)
Alcalinité de l'affluent ($mg \cdot L^{-1}$)	Alcalinité de l'affluent ($mg \cdot L^{-1}$)
Dose d'alun ($mg \cdot L^{-1}$)	Turbidité de l'effluent du clarificateur (UTN)
Dose de chabon actif en poudre ($mg \cdot L^{-1}$)	Dose de chabon actif en poudre ($mg \cdot L^{-1}$)
Vitesse de déversement ($m^{-3} \cdot d^{-1}$)	Vitesse de déversement ($m^{-3} \cdot d^{-1}$)
Décalage-1 de la turbidité de l'affluent (UTN)	Décalage-1 de la turbidité de l'affluent (UTN)
Décalage-1 de la couleur de l'affluent (UTC)	Décalage-1 de la couleur de l'affluent (UTC)

Les modèles ont produits de bons résultats lorsqu'ils ont été exposés à de nouvelles données, l'erreur absolue moyenne pour la turbidité de l'effluent étant de moins de 0,77 UTN et l'erreur absolue moyenne pour la dose d'alun étant de moins de 1,8 mg L^{-1}. Les modèles développés ont été implantés en ligne pour complémenter les résultats des tests de laboratoire et prévoir plus rapidement le dosage approprié d'alun requis pour produire un effluent d'une qualité donnée.

Baxter *et al.* (2001) ont aussi développé avec succès un modèle prévoyant la dose de chaux à utiliser pour adoucir l'eau traitée. La procédure employée est similaire à celle utilisée pour la prévision de la dose d'alun. L'objectif consitait à prévoir la dureté totale de l'effluent du bassin d'adoucissement et le dosage de chaux requis. Les variables d'entrée des deux modèles sont les mêmes (Tableau 1-6), sauf que la dureté de l'effluent devient une variable d'entrée et le dosage d'alun, la variable de sortie dans le cas du modèle de prévision de la dose de chaux requise. Développés en utilisant 8 mois de données quotidiennes, les modèles ont prédit la dureté totale du décanteur avec un r^2 de 0,84 et la dose de chaux à utiliser avec un r^2 de 0,95.

Tableau 1-6 Paramètres d'entrée des modèles retenus pour la prévision de la turbidité de la dureté de l'effluent et des doses de chaux (Baxter *et al.*, 2001)

Variables d'entrée du modèle de prévision de la turbidité	**Variables d'entrée du modèle de prévision de la dose d'alum**
Température de l'eau brute (°C)	Température de l'eau brute (°C)
pH de l'eau brute	pH de l'eau brute
Dureté de l'eau brute (mg $CaCO_3 \cdot L^{-1}$)	Dureté de l'eau brute (mg $CaCO_3 \cdot L^{-1}$)
Alcalinité de l'eau brute (mg $\cdot L^{-1}$)	Alcalinité de l'eau brute (mg $\cdot L^{-1}$)
Débit de la station de traitement (ML $\cdot d^{-1}$)	Débit de la station de traitement (ML $\cdot d^{-1}$)
Dose d'alun (mg $\cdot L^{-1}$)	Dose d'alun (mg $\cdot L^{-1}$)
Dose de chaux (mg $\cdot L^{-1}$)	Dureté totale de l'effluent (mg $CaCO_3 \cdot L^{-1}$)
pH de l'effluent du clarificateur d'adoucissement	pH de l'effluent du clarificateur d'adoucissement

Du point de vue sanitaire, les stations de traitement de l'eau doivent se plier à des réglements de plus en plus sévères sur la qualité de l'eau traitée. C'est le cas avec les sous-produits de désinfection, dont la matière organique est le précurseur. En prévoyant la couleur de l'eau traitée, la couleur étant un paramètre indicateur de la matière organique présente dans l'eau, les opérateurs peuvent ajuster les procédés de traitement afin d'enlever le plus de matière organique possible de l'eau avant de la désinfecter et ainsi réduire la production de sous-produits de désinfection. Baxter *et al.* (2001) ont mis au point des modèles de prévision de la couleur pour les deux stations de traitement de la ville d'Edmonton. Les variables d'entrée du modèle sont similaires à celles utilisées pour le modèle de prévision de la turbidité de l'eau traitée présenté précédemment (Tableau 1-7). Encore une fois, les variables d'entrée ont été sélectionnées selon leur disponibilité et les relations suspectées de cause à effet avec la couleur de l'effluent du bassin de décantation, qui est la variable de sortie.

Tableau 1-7 Paramètres d'entrée du modèle retenu pour l'enlèvement de la couleur par la coagulation (Baxter *et al.*, 2001)

Paramètre	Classification
pH de l'affluent	Paramètre de qualité d'eau brute
Turbidité de l'affluent (UTN)	Paramètre de qualité d'eau brute
Température de l'affluent (°C)	Paramètre de qualité d'eau brute
Couleur de l'affluent (UC)	Paramètre de qualité d'eau brute
Dureté de l'affluent (mg $CaCO_3 \cdot L^{-1}$)	Paramètre de qualité d'eau brute
Alcalinité de l'affluent ($mg \cdot L^{-1}$)	Paramètre de qualité d'eau brute
Dose d'alun ($mg \cdot L^{-1}$)	Paramètre de procédé
Dose de chabon actif en poudre ($mg \cdot L^{-1}$)	Paramètre de procédé
Dose de polymère ($mg \cdot L^{-1}$)	Paramètre de procédé
Décalage-1 de la turbidité de l'affluent (UTN)	Paramètre de série chronologique
Décalage-1 de la couleur de l'affluent (UC)	Paramètre de série chronologique

Les résultats des modèles sont excellents : l'erreur absolue moyenne avec l'ensemble de test est de moins 0.32 UC, ce qui est inférieur à l'erreur associée à l'instrument utilisé pour mesurer la couleur dans le clarificateur. Les modèles ont aussi été testés en ligne pendant la fonte du printemps de 1998 et l'erreur était de moins de 0.35 UC. Les données générées par ces modèles peuvent être intégrées au logiciel de contrôle des stations de traitement afin de prévoir à l'avance les doses d'alun, de charbon actif en poudre et de polymères appropriées pour obtenir un effluent de qualité pour tout le spectre de qualité d'eau brute sans avoir à conduire des tests de laboratoire.

Les RNA peuvent aussi servir à modéliser la performance de la filière de filtration, une autre étape importante pour enlever de l'eau les fines particules en suspension. Baxter *et al.* (2001) ont développé un modèle prédisant le compte de particules à la sortie du filtre. Les variables d'entrée sont présentées dans le Tableau 1-8 et ont été sélectionnées selon la disponibilité des données et l'importance de leur effet sur la performance de la filtration, tel que suggéré par la littérature scientifique. Le modèle a démontré d'excellentes capacités prédictives, le r^2 étant de 0,79 et l'erreur absolue moyenne de seulement 2,3 particules·mL^{-1}.

Tableau 1-8 Paramètres d'entrée du modèle retenu pour la prévision du compte de particules du filtre (Baxter *et al.*, 2001)

Variables d'entrée du modèle	Classification
Température de l'eau brute (°C)	Paramètre de qualité d'eau brute
Turbidité de l'affluent (UTN)	Paramètre de qualité d'eau brute
Température de l'affluent (°C)	Paramètre de qualité d'eau brute
Alcalinité de l'affluent (mg · L^{-1})	Paramètre de qualité d'eau brute
Dureté de l'affluent (mg $CaCO_3$ · L^{-1})	Paramètre de qualité d'eau brute
Débit de la station de traitement (ML · d^{-1})	Paramètre de procédé
Dose d'alun (mg · L^{-1})	Paramètre de procédé
Dose de charbon actif en poudre (mg · L^{-1})	Paramètre de procédé
Dose de chaux (mg · L^{-1})	Paramètre de procédé

Finalement, une fois l'eau traitée, une concentration adéquate de désinfectant doit y être maintenue lors de sa distribution. Sérodes et Rodriguez (1996) se sont intéressés à la question. Ils ont développés un modèle à base de RNA pour la prévision et l'évolution du chlore résiduel dans un réservoir d'un système de distribution. L'élaboration d'un tel modèle empirique est fondée sur les données historiques de paramètres opérationnels de qualité, sélectionnés préalablement par une analyse de régression (Tableau 1-9). Les auteurs ont trouvé par essais et erreurs le nombre de variables de décalage adéquat pour bien saisir l'évolution de chlore résiduel dans le temps.

Tableau 1-9 Paramètres d'entrée du modèle retenu pour la prévision de la concentration de chlore à la sortie d'un réservoir (à partir de Rodriguez et Sérodes, 1996)

Variables d'entrée du modèle	Classification
Température de l'eau dans le réservoir (°C)	Paramètre de qualité d'eau
Concentration du chlore résiduel à l'entrée du réservoir (mg · L^{-1})	Paramètre de qualité d'eau
Concentration du chlore résiduel à la sortie du réservoir (mg · L^{-1})	Paramètre de qualité d'eau
Débit de l'eau admise dans le réservoir (ML · d^{-1})	Paramètre de procédé
Dose de re-chloration (mg · L^{-1})	Paramètre de procédé

Deux modèles distincts ont été développés pour prendre en compte la variabilité saisonnière entre l'hiver et l'été. Trois années de données quotidiennes ont été utilisées, mais seuls les échantillons de données où la dose de chlore était adéquate ont été retenus afin de développer un modèle aux prévisions justes. Environ 10% des données ont été réservées pour tester les modèles. Les modèles finaux ont une architecture 15-9-1 et l'apprentissage a été effectué avec l'algorithme de rétro-propagation. Les résultats démontrent la capacité des RNA à reconnaître la dynamique de l'évolution du chlore résiduel dans les réservoirs et donc le potentiel de la technique dans la modélisation de la qualité de l'eau dans les réseaux de distribution.

1.2.6 Synthèse

Cette seconde partie de la revue de littérature a permis de cerner la majorité des aspects relatifs à l'usage des réseaux neuronaux dans le domaine du génie hydrique. Pour bien comprendre la technique, les concepts généraux tels que le fonctionnement des réseaux de neurones artificiels, leur architecture et leur mode d'apprentissage ont été présentés. Ensuite, la pertinence du recours à la modélisation connexionniste plutôt qu'aux autres méthodes possibles pour la prévision hydrologique a été justifiée. L'approche méthodologique étant souvant négligée par les modélisateurs connexionnistes, une revue approfondie de la littérature a été réalisée afin d'identifier les différentes étapes de la modélisation avec des réseaux de neurones artificiels (Figure 1-10). Finalement, des exemples d'applications relevées dans la littérature ont illustré les différents usages des réseaux neuronaux dans le domaine du génie de l'environnement et des ressources hydriques.

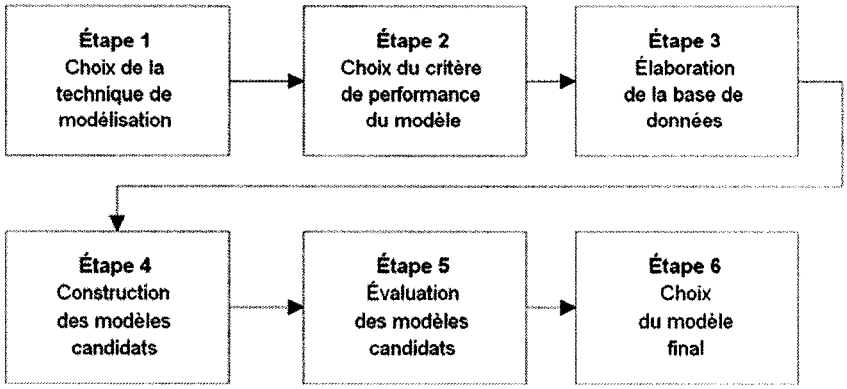


Figure 1-10 : Schéma global présentant les étapes principales de la modélisation connexionniste

Afin de fournir un outil simple aux utilisateurs potentiels de réseaux neuronaux, les différentes décisions à prendre, les choix à effectuer et les questions à se poser à toutes les étapes de la modélisation connexionniste sont schématisés aux Figures 1-11 à 1-14. Cette approche méthodologique, élaborée à partir de la revue de littérature, sera mise en pratique dans les chapitres suivants afin de développer un modèle connexionnsite capable de prévoir efficacement les augmentations de turbidité à l'eau brute à la Ville de Montréal.

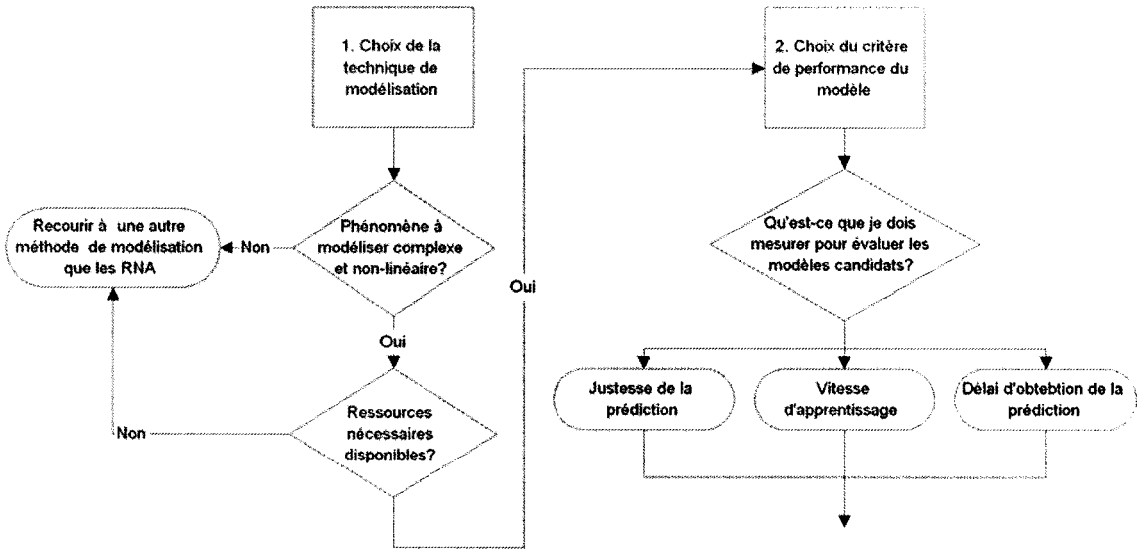


Figure 1-11 : Synthèse schématisée pour le choix de la technique de modélisation et le choix du critère de performance

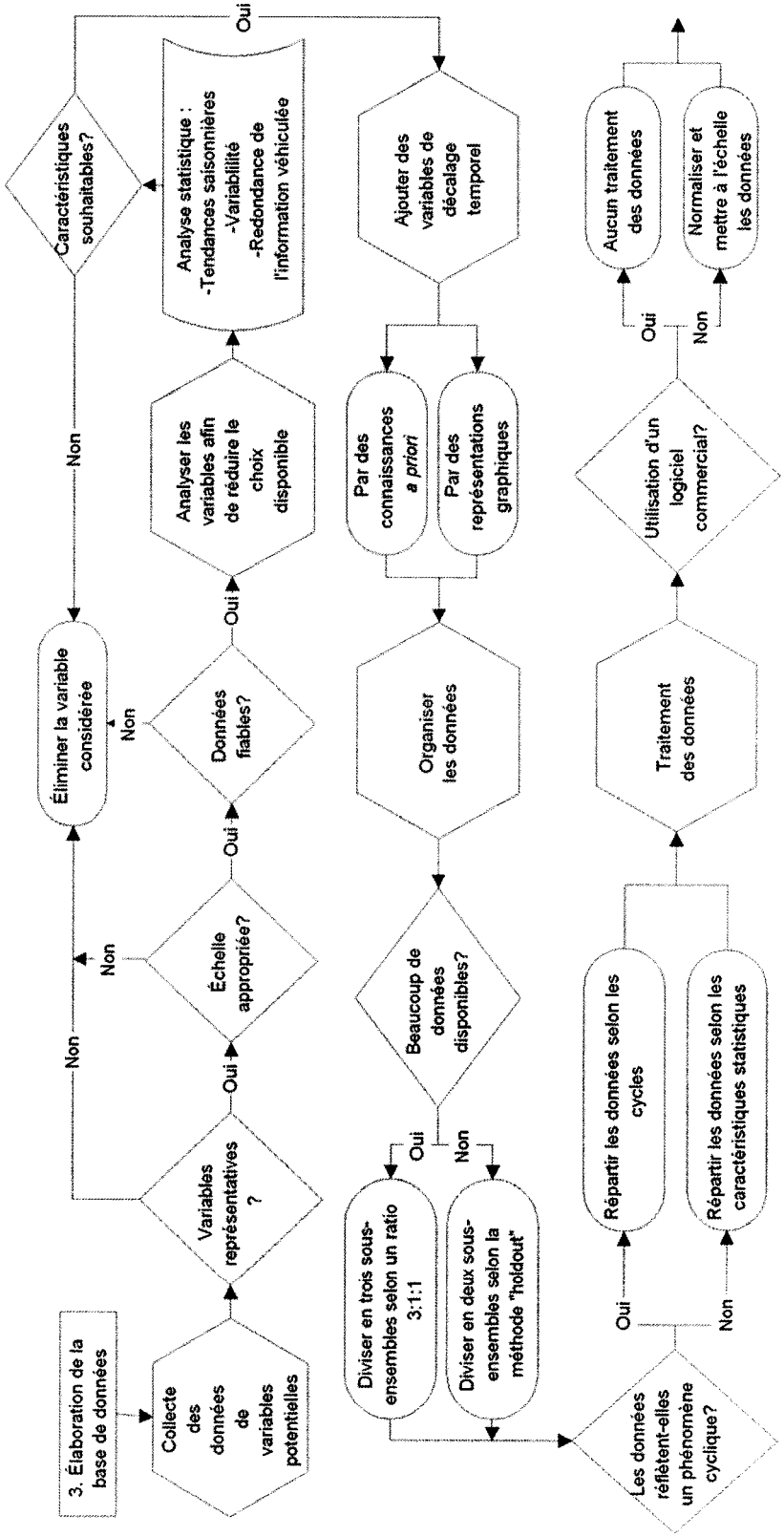


Figure 1-12 : Synthèse schématisée pour l'élaboration de la base de données

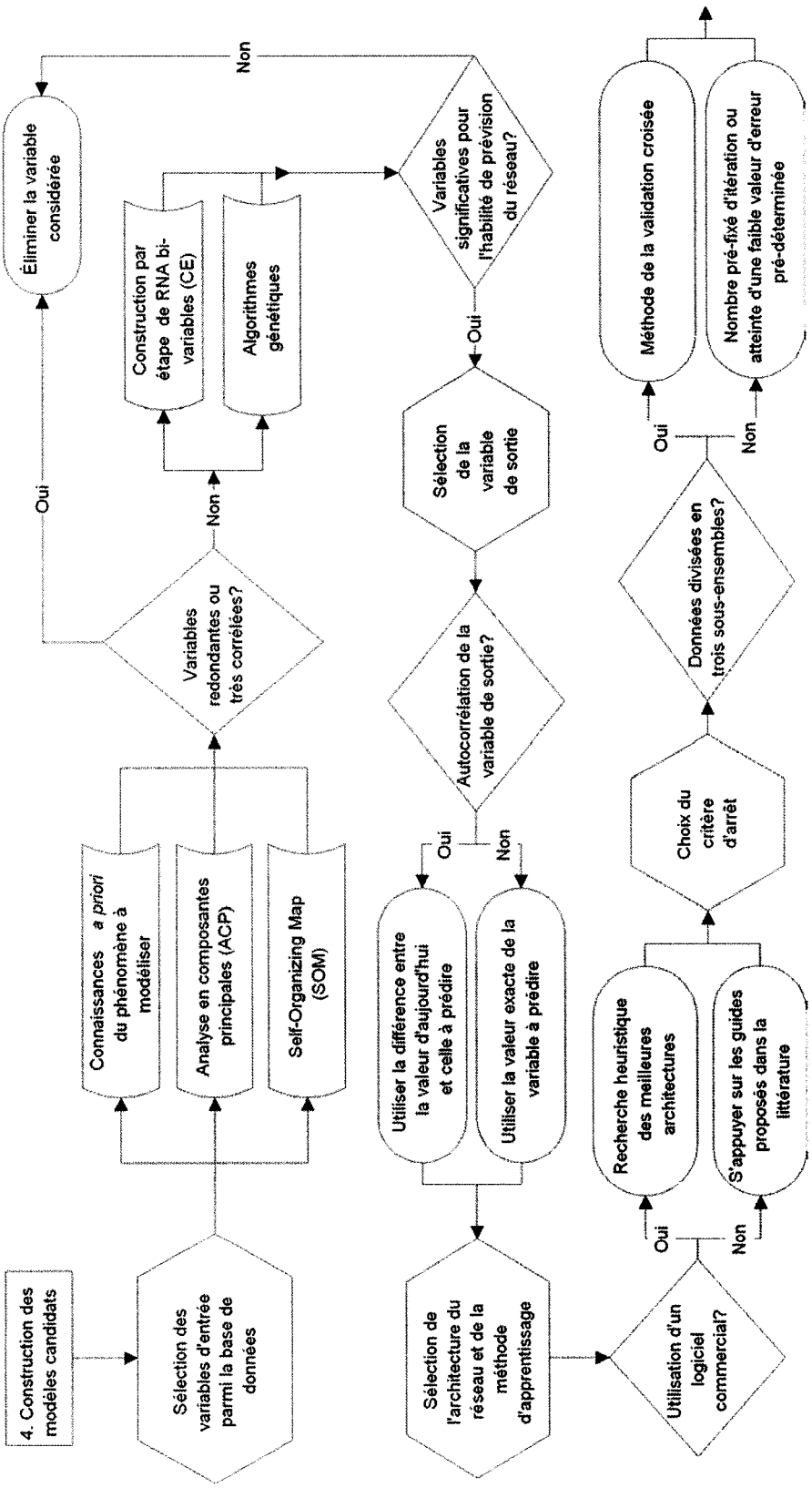


Figure 1-13 : Synthèse schématisée pour la construction des modèles candidats

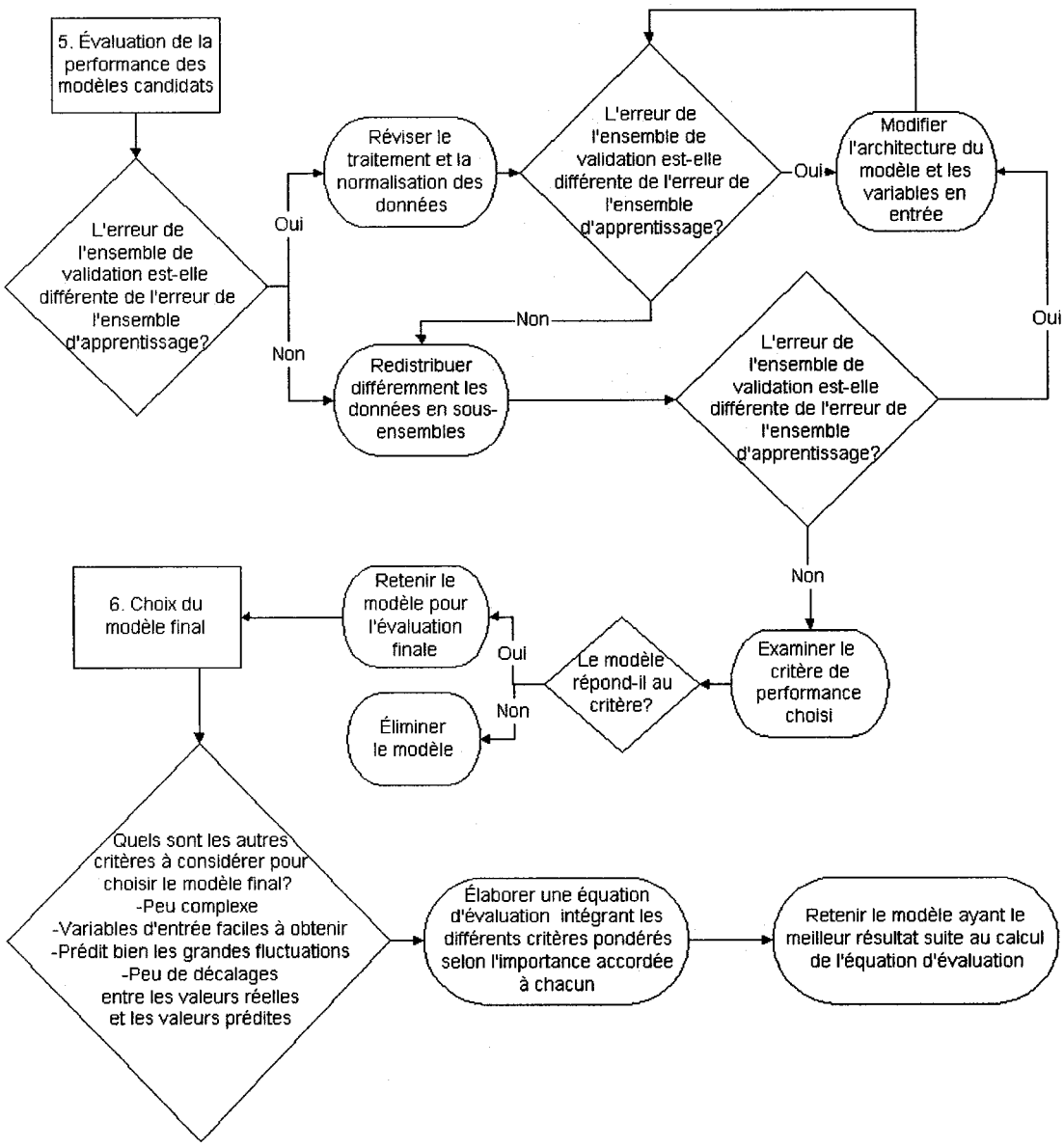


Figure 1-14 : Synthèse schématisée pour l'évaluation des modèles candidats et le choix du modèle final

CHAPITRE 2 - IDENTIFICATION DES ÉVÉNEMENTS CAUSANT LES AUGMENTATIONS DE TURBIDITÉ

Les travaux de ce chapitre font partie des premières étapes de la modélisation connexionniste, soit de bien connaître le phénomène à modéliser, de constituer une base de données représentative et d'identifier des variables d'entrée potentielles du modèle. Dans le chapitre précédent, la revue de littérature a permis de cerner certains facteurs explicatifs susceptibles d'influencer ou de causer les augmentations de turbidité, soit l'accroissement du débit de la rivière des Outaouais, une inversion thermique dans le lac des Deux Montagnes, les tempêtes de vent et les fortes pluies. Ces facteurs explicatifs étant hypothétiques, ce chapitre permet de justifier leur usage en modélisation connexionniste démontrant qu'ils expliquent la très vaste majorité des augmentations de turbidité durant une période de temps donnée. Plus précisément, les travaux de ce chapitre consistent à:

- Décrire les caractéristiques de la turbidité de l'eau brute à la prise d'eau de la Ville de Montréal, dégagées à partir de données quotidiennes de la turbidité de janvier 1998 à avril 2001.

- Identifier et caractériser les événements turbides, définis comme des journées consécutives au cours desquelles la turbidité excède une intensité donnée. Ce sont ces événements turbides qui seront expliqués à l'aide du modèle connexionniste.

- Sélectionner les variables d'entrée potentielles dans la base de données préliminaire, qui fournira les paramètres indicateurs servant à illustrer les différents facteurs explicatifs identifiés par la revue de littérature et les variables d'entrée nécessaires au développement de modèles connexionnistes.

- Explorer les liens entre les événements turbides et les facteurs explicatifs pour identifier les causes des augmentations de turbidité et aider à choisir les variables d'entrée des modèles connexionnistes.

2.1 Analyse descriptive de la turbidité

Avant de pouvoir comprendre ce qui cause les variations de la turbidité de l'eau brute, il faut tout d'abord bien connaître les caractéristiques de la turbidité. Pour ce faire, des données quotidiennes de turbidité de janvier 1998 à avril 2001 ont été rassemblées. Ces données sont facilement accessibles et d'un format électronique approprié. Selon des discussions informelles avec les responsables des stations de traitement de la Ville de Montréal, elles sont aussi représentatives de l'échelle de valeurs pouvant être observées pendant une année et d'une année à l'autre.

Dans cette section, ces données de turbidité de l'eau brute sont examinées sous différents angles afin d'en dresser un portrait global utile pour la suite du projet. Tout d'abord, les périodes critiques sont identifiées. Ensuite, les données individuelles sont soumises à une analyse statistique pour évaluer la variabilité de la turbidité d'une période à l'autre et définir des classes de turbidité. Ces classes de turbidité permettront de comparer la turbidité d'une année à l'autre et d'une période turbide à l'autre.

2.1.1 Identification des périodes critiques

L'inspection visuelle rapide de la représentation graphique des trois ans et demi de données de turbidité de l'eau brute permet de délimiter de manière heuristique quatre périodes distinctes (Figure 2-1), dont les caractéristiques et les dates de début et de fin sont présentées au tableau 2.1. On y constate facilement que les périodes d'intérêt sont celles du printemps et de l'automne.

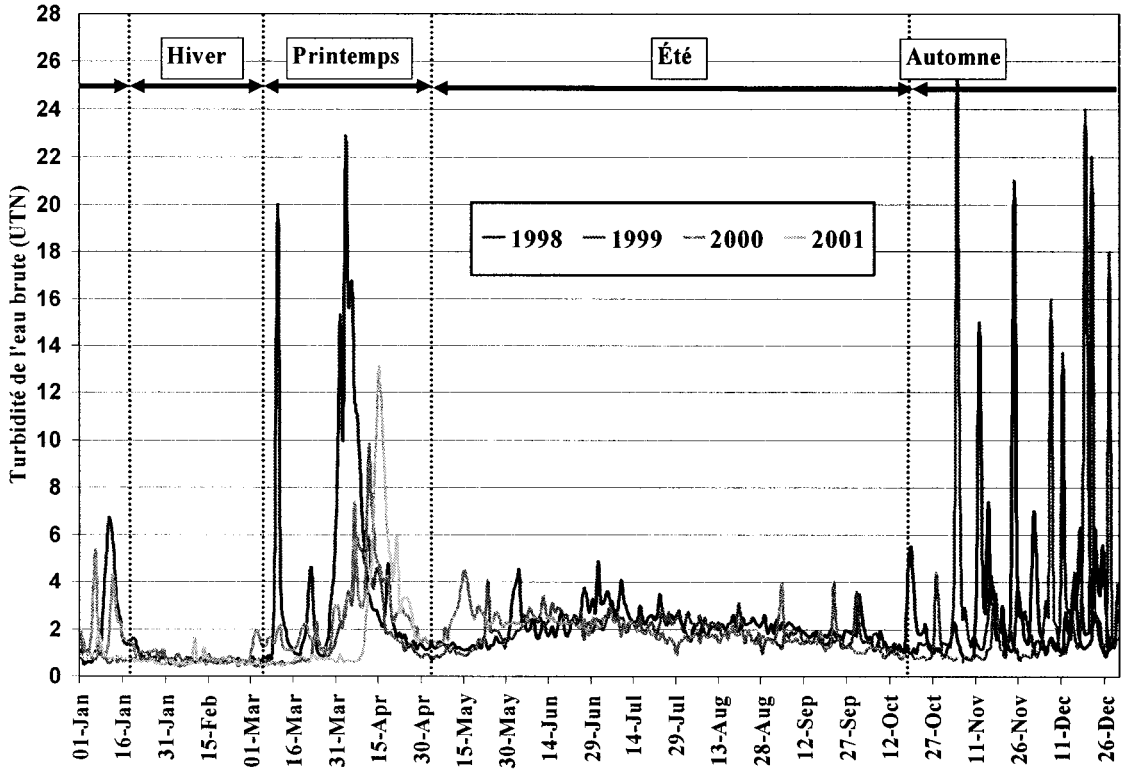


Figure 2-1 : Variations journalières de la turbidité de l'eau brute de la station de traitement Desbaillets, de janvier 1998 à avril 2001

Tableau 2-1 : Description des périodes de l'année, définies selon les caractéristiques de la turbidité de l'eau brute

Nom de la période	Caractéristiques générales	Dates de début et de fin
Hiver	▪ Turbidité très faible ▪ Très peu de fluctuations	15 janvier - 28 février
Printemps	▪ Pointes de turbidité de grande magnitude (relativement à la moyenne) ▪ Augmentations soudaines et de durée prolongée	1er mars – 30 avril
Été	▪ Turbidité faible, mais plus élevée qu'à l'hiver ▪ Nombreuses fluctuations, résultant parfois en des pointes de turbidité de faible magnitude	1er mai – 14 octobre
Automne	▪ Nombreuses pointes de turbidité de grande magnitude (relativement à la moyenne) ▪ Augmentations soudaines et de courte durée	15 octobre - 14 janvier

2.1.2 Analyse statistique des données et définition des classes de turbidité

Afin de caractériser la turbidité de l'eau brute de la Ville de Montréal, les données quotidiennes de turbidité sont comparées d'une période à l'autre et d'une année à l'autre. Pour ce faire, les données sont soumises à une analyse statistique, qui comprend la mesure de la tendance centrale et de la variation et une analyse des percentiles.

2.1.2.1 Variabilité de la turbidité

La mesure de la tendance centrale et de la variation permet de déterminer à quel point la turbidité fluctue au cours de l'année et d'une année à l'autre. La turbidité de l'eau brute de la ville de Montréal est relativement faible, la moyenne annuelle n'étant que de 1,97 UTN pour l'ensemble des données de janvier 1998 à avril 2001. Néanmoins, les résultats de l'analyse statistique par période révèle qu'il existe des tendances saisonnières qui amènent la turbidité à fluctuer de façon appréciable au cours du printemps et de l'automne. Les résultats sont présentés au Tableau 2-2 et illustrés par une représentation graphique Box-Whisker à la Figure 2-2. On y constate que la turbidité moyenne de l'eau brute est très faible durant l'hiver et l'été et à peine plus élevée durant le printemps et l'automne. Par contre, les écart-types du printemps et de l'automne sont très élevés, illustrant bien la grande variabilité des valeurs de turbidité lors de ces périodes comparativement à celles de l'hiver et de l'été.

Tableau 2-2 : Moyenne et écart-type des données de turbidité par période

Paramètres	Printemps (1er mars au 30 avril)	Été (1er mai au 14 octobre)	Automne (15 octobre au 14 janvier)	Hiver (15 janvier au 28 février)
Moyenne	2,48 UTN	1,96 UTN	2,34 UTN	0,72 UTN
Écart-type	3,09 UTN	0,72 UTN	3,27 UTN	0,30 UTN

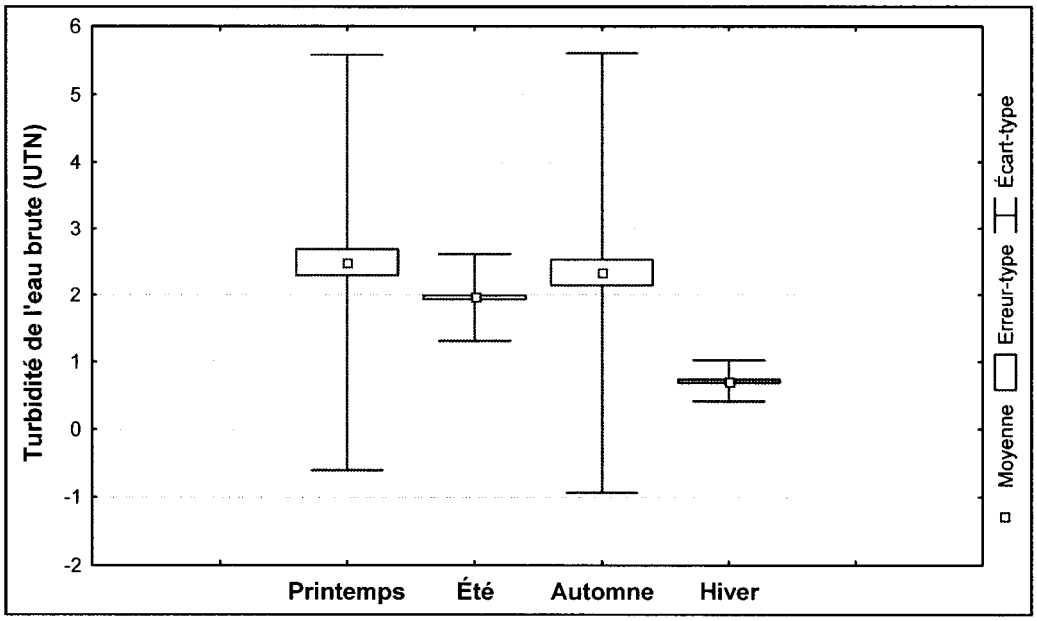


Figure 2-2 : Représentation graphique Box-Whisker, illustrant la moyenne, l'erreur-type et l'écart-type des valeurs de turbidité des quatre périodes

2.1.2.2 Classes de turbidité

L'analyse des percentiles, quant à elle, permet de fixer des seuils à partir desquels les valeurs de turbidité sont réparties en cinq classes d'intensité (Tableau 2-3). Ce sont ces classes qui seront utilisées pour la suite du projet, car les classes permettent d'analyser et de comparer les résultats plus facilement et clairement que les valeurs exactes de turbidité.

Tableau 2-3 : Définition des classes d'intensité de la turbidité de l'eau brute à la prise d'eau de la Ville de Montréal

Identification	Percentiles	Valeurs de turbidité correspondant aux percentiles
Classe 1	0 à 0,75	Turbidité < 2,3 UTN
Classe 2	0,75 à 0,90	2,3 UTN < Turbidité < 3,2 UTN
Classe 3	0,90 à 0,95	3,2 UTN < Turbidité < 4,4 UTN
Classe 4	0,95 à 0,99	4,4 UTN < Turbidité < 13,7 UTN
Classe 5	Entre 0,99 et 1,00	13,7 UTN < Turbidité < 26,0 UTN

2.1.3 Comparaison de la turbidité entre les périodes et entre les années

Les périodes du printemps et de l'automne présentent le plus de valeurs quotidiennes de turbidité appartenant aux classes 3 à 5 (plus de 3,2 UTN) que la moyenne des périodes, soit 18% des valeurs pour le printemps et 16% pour l'automne. En été, seulement 5% des valeurs de turbidité dépassent 3,2 UTN alors qu'en hiver, il n'y en a pas.

Si on examine plus particulièrement les données du printemps et de l'automne par année, on constate que les années sélectionnées sont différentes les unes des autres (Figure 2-3). Pour la période du printemps, l'année 1998 a été inhabituelle car la turbidité de l'eau brute a atteint 22,8 UTN, une valeur très élevée pour la prise d'eau de la Ville de Montréal. Il s'agit d'ailleurs de la seule année où des valeurs de turbidité appartiennent à la classe d'intensité 5. Pour la période de l'automne, les données de trois années sont assez semblables, bien que l'année 2000 présente moins de valeurs de turbidité appartenant aux classes 3 à 5 que les années 1998 et 1999.

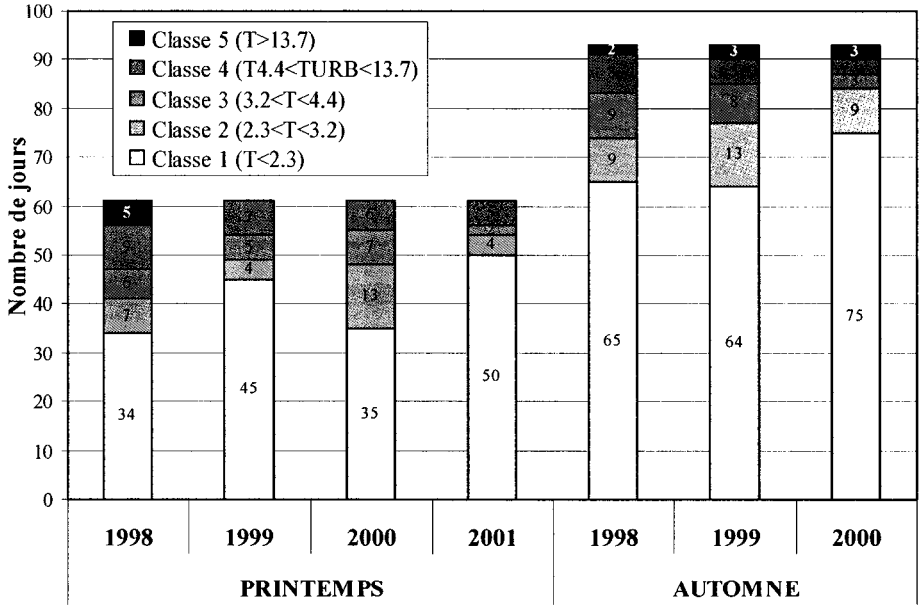


Figure 2-3 : Distribution des valeurs de turbidité appartenant aux différentes classes d'intensité par année et par période

2.2 Analyse qualitative des événements turbides du printemps et de l'automne

Alors que la première section examinait globalement les données journalières de turbidité de l'eau brute, cette section s'intéresse plutôt aux augmentations significatives de turbidité, qui seront l'objet d'étude pour la suite du projet.

2.2.1 Définition et identification des événements turbides

Désignées par l'appelation « événements turbides », les augmentations de turbidité sont définies comme un ou plusieurs jours consécutifs au cours desquels la turbidité s'accroît et excède 3,2 UTN (classes 3 à 5).

Les événements turbides du printemps et de l'automne de chaque année sont identifiés par inspection visuelle des représentations graphiques des classes d'intensité quotidiennes de la turbidité. La durée de l'événement et l'intensité maximale atteinte sont examinées et notées. La durée détermine le type d'événement turbide. Lorsque la durée de l'événement excède cinq jours, l'événement est désigné comme « événement de fond ». Sinon, il s'agit d'un « événement de courte durée ». Les événements de courte durée peuvent être isolés ou se superposer à un événement de fond. Les Figures 2-4 et 2-5 présentent en exemple les représentations graphiques des classes de turbidité et des valeurs numériques correspondantes pour le printemps et l'automne 1998, sur lesquelles ont été identifiées les événements turbides. Les représentations graphiques de toutes les autres périodes étudiées sont présentées à l'Annexe 3.

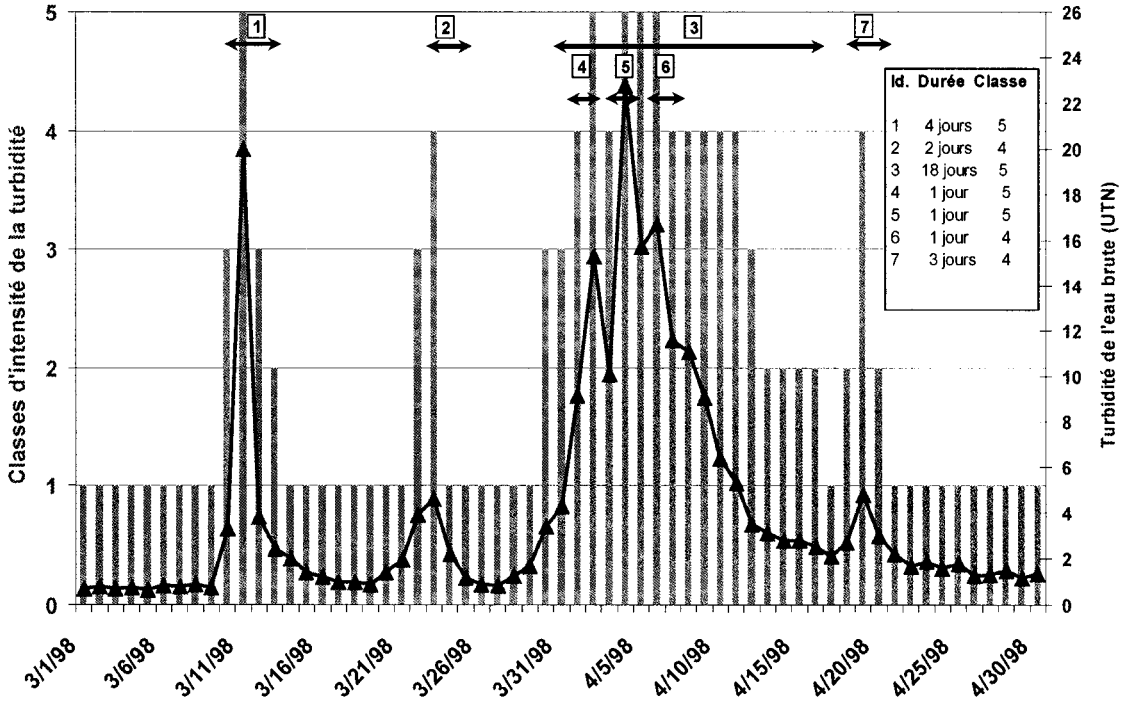


Figure 2-4 : Identification des événements turbides du printemps 1998

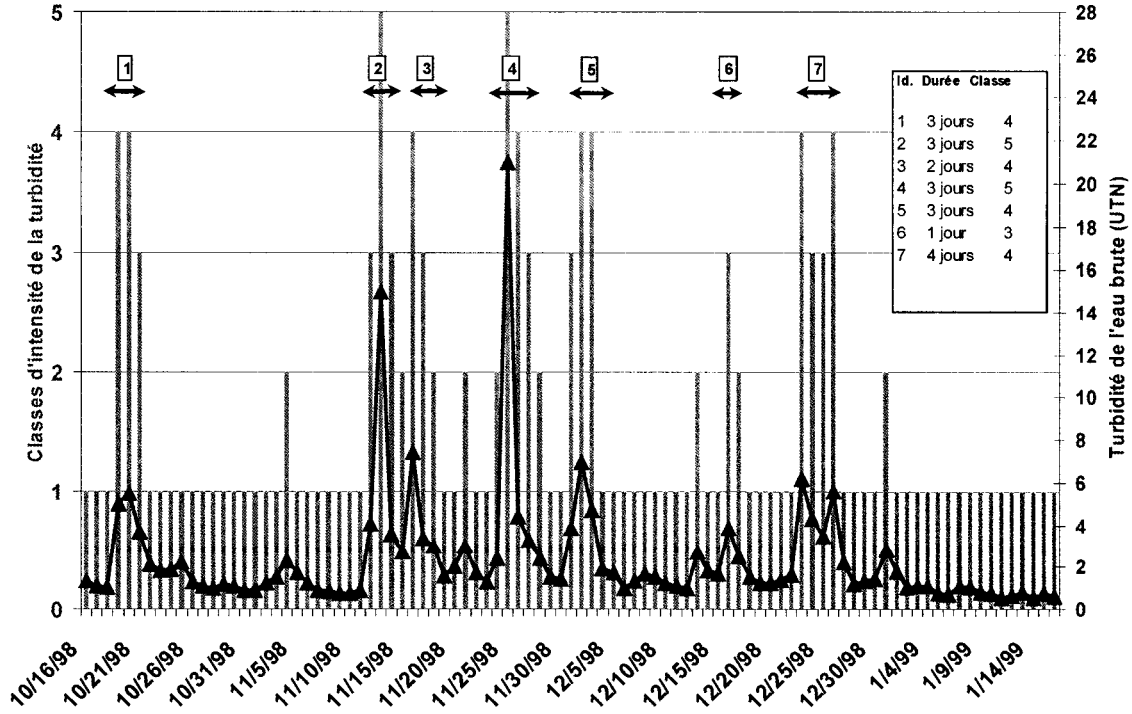


Figure 2-5 : Identification des événements turbides de l'automne 1998

2.2.2 Comparaison des événements turbides répertoriés

Le Tableau 2-4 présente le nombre d'événements turbides de chaque type répertoriés au cours des quatre périodes printanières et des trois périodes automnales examinées.

Tableau 2-4 : Événements turbides répertoriés par type et par période

TYPE D'ÉVÉNEMENT TURBIDE	PRINTEMPS				AUTOMNE		
	98	*99*	*00*	*01*	*98*	*99*	*00*
Événement turbide de fond	*1*	*1*	*1*	*1*	*0*	*0*	*0*
	4				**0**		
Événement turbide court superposé	*3*	*4*	*2*	*1*	*0*	*0*	*0*
	10				**0**		
Événement turbide court isolé	*3*	*0*	*0*	*0*	*7*	*10*	*6*
	3				**23**		

Les périodes du printemps et de l'automne se dinstinguent l'une de l'autre par la durée et l'intensité des événements turbides. Au cours des printemps 1998 à 2000, des événements de courte durée (moins de 5 jours) se superposent à un seul événement de fond (plus de cinq jours). À ceci s'ajoutent parfois quelques événements de courte durée isolés de l'événement de fond. Par contre, au cours des automnes 1998 à 2000, il n'y a aucun événement turbide de fond. Il n'y a que des événements courts, isolés les uns des autres.

En plus des types d'événements turbides, les différences entre le printemps et l'automne s'inscrivent aussi au niveau de l'intensité des événements turbides, comme le démontre la Figure 2-6. On peut y remarquer qu'au printemps, la turbidité des événements est majoritairement de classe 4 (turbidité entre 4,4 et 13,7 UTN), tant pour les événements turbides de fond que les événements courts. On retrouve très peu d'événements de classe 3 (turbidité entre 3,2 et 4,3 UTN). Par contre, à l'automne, les événements se répartissent pratiquement également entre les trois classes d'intensité de turbidité, avec une légère prédominance pour la classe 5 (turbidité excédant 13,7 UTN).

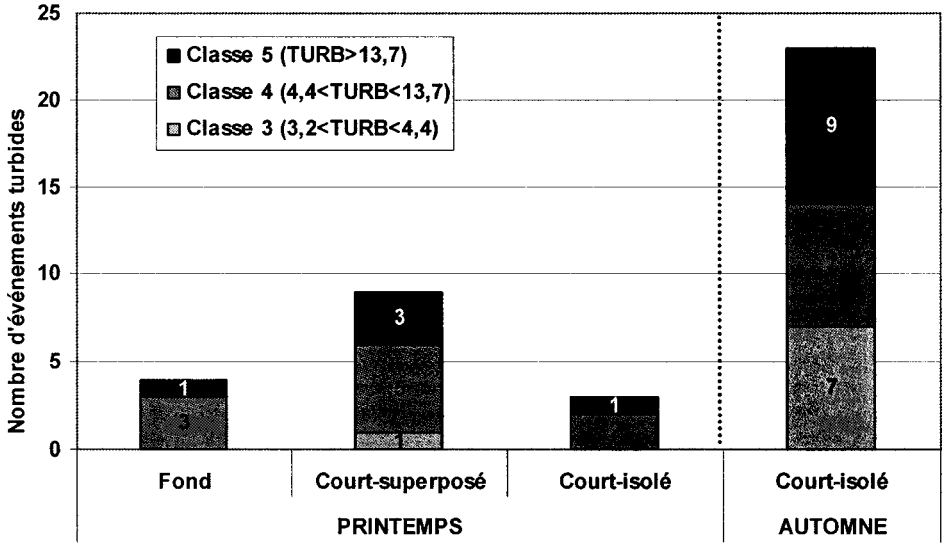


Figure 2-6 : Intensité des différents types d'événements turbides répertoriés

2.3 Sélection de variables d'entrée potentielles

Avant de pouvoir expliquer les événements turbides identifiés précédemment, une base de données préliminaire de variables d'entrée potentielles doit être constituée afin que puissent y être choisis, dans la prochaine section, les paramètres indicateurs servant à illustrer les différents facteurs explicatifs identifiés par la revue de littérature et les variables d'entrée nécessaires au développement de modèles connexionnistes. Cette section présente donc tout d'abord l'ensemble des variables disponibles et explique ensuite le processus de sélection des variables qui permet de constituer la base de données préliminaire.

2.3.1 Présentation des variables disponibles

La base de données doit rassembler des données quotidiennes de variables pouvant être liées aux différents événements explicatifs identifiés dans la revue de littérature, soit les variations de débit du fleuve St-Laurent et de la rivière des Outaouais, l'inversion thermique dans les lacs fluviaux, les tempêtes de vent et les précipitations abondantes. Toutes les variables disponibles pour la sélection sont présentées par catégorie aux

Tableaux 2-3 à 2-5. Ces variables ont été rassemblées selon la connaissance de l'auteur des événements explicatifs et selon leur fiabilité et leur accessibilité.

Tableau 2-5 : Variables de qualité de l'eau brute et traitée disponibles

Variables	Ville de Montréal (station DesBaillets)	Ville de Coteau-du-lac	Communauté urbaine de l'Outaouais	Ville de Hawkesbury
Turbidité de l'eau brute	X	X	X	X
Couleur de l'eau brute	X			X
Conductivité de l'eau traitée	X			
pH de l'eau brute	X		X	
Température de l'eau brute	X	X		X

Tableau 2-6: Variables météorologiques disponibles

Variables	Dorval	Ottawa	Sainte-Anne-de-Bellevue	Lac St-François
Température moyenne de l'air	X	X	X	X
Hauteur de pluie	X	X		
Précipitation totale	X	X	X	X
Ensoleillement	X			
Vitesse horaire maximale du vent	X		X	X
Vitesse horaire moyenne du vent	X		X	X
Direction horaire maximale du vent	X		X	X
Direction horaire moyenne du vent	X		X	X

Tableau 2-7 : Variables de débit des cours d'eau disponibles

Stations d'échantillonage du débit	Rivière Outaouais	Fleuve St-Laurent	Rivière Raisin	Rivière Beaudette
Barrage de Carillon	X			
Couloir Ste-Anne-de-Bellevue	X			
Couloir Vaudreuil	X			
Lit naturel à Les Cèdres		X		
Canal de Beauharnois, à Beauharnois		X		
Lasalle		X		
Williamstown			X	
Glen Navis				X

2.3.2 Constitution de la base de données préliminaire

Afin de réduire le nombre de variables d'entrée potentielles et de faciliter ainsi le travail subséquent de sélection des variables d'entrée des modèles connexionnistes, une sélection est effectuée parmi toutes les variables disponibles.

Une première sélection de variables est effectuée suite à l'examen des mesures de la tendance centrale et de la variation de toutes les variables disponibles. Les résultats complets de l'analyse statistique sont présentés à l'Annexe 4. Seules les variables présentant des tendances saisonnières, i.e. dont la moyenne varie d'une saison à l'autre et dont la variation est plus importante au printemps et à l'automne, ont été retenues. Ainsi, les variables dont la moyenne d'une période à l'autre ne diffère pas beaucoup, telles que le pH à la station DesBaillets et la direction des vents à toutes les locations, ont été éliminées.

D'autres considérations permettent d'éliminer des variables moins pertinentes. C'est le cas des variables météorologiques de régions plus éloignées comme Ottawa, car ces variables sont moins susceptibles d'influencer directement les fluctuations de turbidité à la prise d'eau Montréal. D'autres variables véhiculant des informations redondantes ont aussi été éliminées, comme les données météorologiques de hauteur d'eau, éliminées au profit des précipitations totales, dont on dispose des données pour plus de locations géographiques. Dans le même ordre d'idée, l'ensoleillement, dont les informations véhiculées sont complémentaires à celles des variables de précipitation, n'a pas été retenu.

Des variables nouvelles, élaborées à partir de variables disponibles, sont venues complétées la base de données préliminaire. Deux variables de « température-degré » ont été calculées à partir de la température de l'air à Dorval pour représenter, dans un cas, la croissance de la végétation (température-degré-croissance = nombre de degrés excédant 5°C en moyenne pour la journée) et dans l'autre, le dégel (température-degré-

dégel = nombre de degrés excédent 0°C en moyenne pour la journée). Le Tableau 2-8 présente toutes les variables retenues pour constituer la base de données préliminaire.

Tableau 2-8 : Variables retenues pour constituer la base de données préliminaire

Variables de QUALITÉ DE L'EAU	Variables MÉTÉOROLOGIQUES	Variables de DÉBIT
TURBIDITÉ • Turbidité de l'eau brute à Coteau-du-lac • Turbidité de l'eau brute à Hawkesbury **TEMPÉRATURE DE L'EAU** • Température de l'eau traitée à Hawkesbury **COULEUR** • Couleur de l'eau brute à DesBaillets • Couleur de l'eau brute à Hawkesbury **CONDUCTIVITÉ** • Conductivité de l'eau traitée à DesBaillets	**TEMPÉRATURE DE L'AIR** • Température moyenne de l'air à Dorval • Température-degré de croissance à Dorval • Température-degré de dégel à Dorval **PRÉCIPITATIONS** • Précipitations totales à Dorval • Précipitations totales à Sainte-Anne-de-Bellevue • Précipitations totales au lac St-François **VENTS** • Vitesse maximale du vent à Dorval • Vitesse maximale du vent à Sainte-Anne-de-Bellevue • Vitesse maximale du vent au lac St-François • Vitesse moyenne du vent à Dorval • Vitesse moyenne du vent à Sainte-Anne-de-Bellevue • Vitesse moyenne du vent au lac St-François	**OUTAOUAIS** • Débit moyen quotidien de la rivière des Outaouais au barrage Carillon • Débit moyen quotidien de la rivière Outaouais s'écoulant par le canal Sainte-Anne • Débit moyen quotidien de la rivière Outaouais s'écoulant par le chenal de l'Île Perrot • Moyenne des débits moyens quotidiens de la rivière Outaouais s'écoulant par les canaux Sainte-Anne et Île Perrot • Proportion des eaux de l'Outaouais dans les eaux totales du fleuve à Lasalle **FLEUVE ST-LAURENT** • Débit moyen quotidien du fleuve St-Laurent s'écoulant par le lit naturel à Des Cèdres • Débit moyen quotidien du fleuve St-Laurent s'écoulant par le canal à Beauharnois • Débit moyen quotidien du fleuve St-Laurent à Lasalle **TRIBUTAIRES SECONDAIRES DU LAC ST-FRANÇOIS** • Débit moyen quotidien de la rivière Raisin à Williamstown • Débit moyen quotidien de la rivière Beaudette à Glen Navis

Finalement, tel que suggéré dans la revue de littérature (section 1.2.4.3), la base de données préliminaire est complétée en identifiant et en corrigeant au besoin les observations aberrantes (*outliers),* les entrées erronées et les entrées manquantes de chaque variable sélectionnée.

2.4 Exploration des liens entre les facteurs explicatifs et les événements turbides du printemps et de l'automne

Explorer les liens existant entre les événements turbides identifiés et les facteurs explicatifs est une étape cruciale tant pour l'identification des causes des augmentations de turbidité que pour le choix des variables d'entrée des modèles connexionnistes.

2.4.1 Méthodologie

La méthodologie employée consiste à comparer l'occurrence des facteurs explicatifs à celle des événements turbides pour déterminer s'il y a concordance entre les deux. Le décalage entre l'occurrence du facteur explicatif et de l'événement turbide de même que l'occurrence simultanée ou rapprochée de plusieurs facteurs explicatifs sont aussi examinés, afin de guider le choix des variables de décalage, nécessaires au développement prochain de réseaux de neurones feedforward.

Les facteurs explicatifs sont représentés graphiquement par des paramètres indicateurs, sélectionnés à partir des variables de la base de données préliminaire (Tableau 2-9). Aux quatres facteurs explicatifs mentionnés dans la revue de littérature, l'auteur a choisi d'ajouter la fonte des neiges et la prise de la glace sur les rivières, deux événements qui influencent grandement le moment de l'occurrence ou l'impact des autres facteurs explicatifs sur la turbidité de l'eau brute. Les raisons justifiant la sélection des paramètres indicateurs et des seuils et critères d'évaluation utilisés sont présentés à l'Annexe 5.

Tableau 2-9 : Paramètres indicateurs sélectionnés pour représenter les différents événements explicatifs

Événement explicatif	Paramètres indicateurs
Débit accru de l'Outaouais et augmentation de la contribution de l'Outaouais dans le fleuve St-Laurent	➢ Débit de l'Outaouais au barrage Carillon ➢ Contribution de l'Outaouais au débit total du fleuve à LaSalle, soit le rapport entre le débit total de l'Outaouais par les chenaux Sainte-Anne et Île-Perrot et le débit du fleuve à Lasalle
Fonte des neiges et présence du couvert de glace	➢ Index de température-degré de dégel (Dorval), soit le nombre de degrés Celcius au-dessus du point de congélation à Dorval ➢ Débit de la rivière Raisin ➢ Débit de la rivière Beaudette ➢ Précipitations locales au lac St-François
Inversion thermique dans le lac des Deux Montagnes	➢ Température de l'eau brute à Hawkesbury
Fortes pluies	➢ Précipitations totales à Dorval, à Sainte-Anne-de-Bellevue et au lac St-François
Tempêtes de vent	➢ Intensité moyenne et maximale du vent à Dorval ➢ Intensité moyenne et maximale du vent au lac St-François

Tous les choix de paramètres, de seuils et de critères, effectués en s'appuyant sur la connaissance experte de l'auteur, sont révisés et ajustés lors d'un processus itératif. Si les seuils et critères d'évaluation retenus ne sont pas assez sévères, l'occurrence de certains facteurs explicatifs ne résultera en aucune augmentation de turbidité. Par contre, si les seuils ou critères sont trop restrictifs, certains facteurs turbides ne pourront être expliqués. Le processus itératif d'ajustement des seuils et critères cesse idéalement lorsque tous les événements turbides sont expliqués et que tous les facteurs explicatifs résultent en une augmentation de turbidité.

Mais si tous les seuils et critères d'évaluation semblent adéquats et que des événements turbides restent inexpliqués, alors le processus itératif nous indique qu'il y aurait un

autre facteur explicatif non identifié préalablement par la revue de littérature. L'hypothèse selon laquelle tous les événements turbides observés à la prise d'eau de la Ville de Montréal sont causés par un ou plusieurs facteurs identifiés dans la revue de littérature se révélerait alors inexacte.

Finalement, en complément à l'examen visuel des représentations graphiques des paramètres indicateurs, une matrice de corrélation entre les données des paramètres indicateurs et de la turbidité à différents décalages est développée et examinée. Les corrélations permettent de confirmer les valeurs de décalage entre les paramètres indicateurs et la turbidité, dont l'examen visuel des représentations graphiques donne déjà un aperçu. La matrice de corrélation permet aussi de déterminer quelles variables non utilisées comme paramètres indicateurs sont les plus corrélées à la turbidité. Bien que ces variables ne soient pas en lien direct avec les facteurs explicatifs, elles peuvent transmettre aux réseaux de neurones artificiels des informations très utiles en raison de leur lien avec la variable à prévoir.

2.4.2 Résultats

Les liens entre les facteurs explicatifs et les événements turbides sont de deux ordres. Tout d'abord, les liens établissent une relation de cause-à-effet entre l'occurrence d'un ou de plusieurs facteurs explicatifs et une augmentation de turbidité. Deuxièmement, les liens révèlent le décalage temporel avant qu'un facteur explicatif affecte la turbidité de l'eau brute. Ces deux aspects différents sont présentés en première partie de cette section. Ensuite, les variables d'entrée potentielles non utilisées comme paramètres indicateurs les plus corrélées à la turbidité sont identifiées afin d'être ajoutées à l'ensemble de variables d'entrée proposées pour le développement de modèles connexionnistes. L'ensemble complété est d'ailleurs présenté à la fin de la section, parmi les perspectives soulevées par les travaux de ce chapitre.

2.4.2.1 Identification des causes des événements turbides printaniers et automnaux

L'analyse des représentations graphiques des paramètres indicateurs a révélé que les liens entre les facteurs explicatifs et les événements turbides sont plus complexes au printemps qu'à l'automne. Alors qu'à l'automne, on ne retrouve que des événements courts isolés pour lesquels les causes sont facilement identifiables, les trois types d'événements turbides coexistent au printemps, avec des causes parfois différentes, parfois similaires. De plus, plusieurs facteurs explicatifs agissent en synergie au printemps pour chacun des types d'événements turbides. La Figure 2-7 compare les causes des événements turbides de fond, de courte durée isolés et de courte durée superposés du printemps et de l'automne. Un exemple de l'analyse des représentations graphiques des paramètres indicateurs ayant mené à l'identification des causes est présenté à l'Annexe 6.

Au printemps, on constate que tous les événements turbides de fond sont causés par la hausse graduelle du débit de la rivière des Outaouais suite à la fonte des neiges, reflétée par la hausse soudaine des tributaires secondaires du fleuve St-Laurent. Les événements courts se superposant à ces événements de fond ont des causes plus diverses. Néanmoins, il ressort nettement que plusieurs sont aussi liés aux augmentations de débit : alors que la hausse des débits se réflète par une hausse graduelle de la turbidité de l'eau brute, les pointes de débit de la rivière des Outaouais ou des tributaires secondaires résultent en des pointes soudaines de turbidité. Ces pointes de débit peuvent agir seules ou en synergie avec d'autres événements turbides comme les tempêtes de vents avec ou sans précipations ou le renversement.

Le renversement, que l'on pensait être à la source des augmentations de turbidité printanières avant que ce projet de mémoire soit entrepris, se révèle n'être qu'un facteur explicatif secondaire expliquant quelques événements turbides courts, isolés, à l'automne ou superposés à l'événement turbide de fond, au printemps. De plus, le renversement n'est jamais la cause unique d'un événement turbide : il est toujours en

lien avec d'autres facteurs explicatifs comme des pointes de débit ou des tempêtes de vents avec ou sans précipitations.

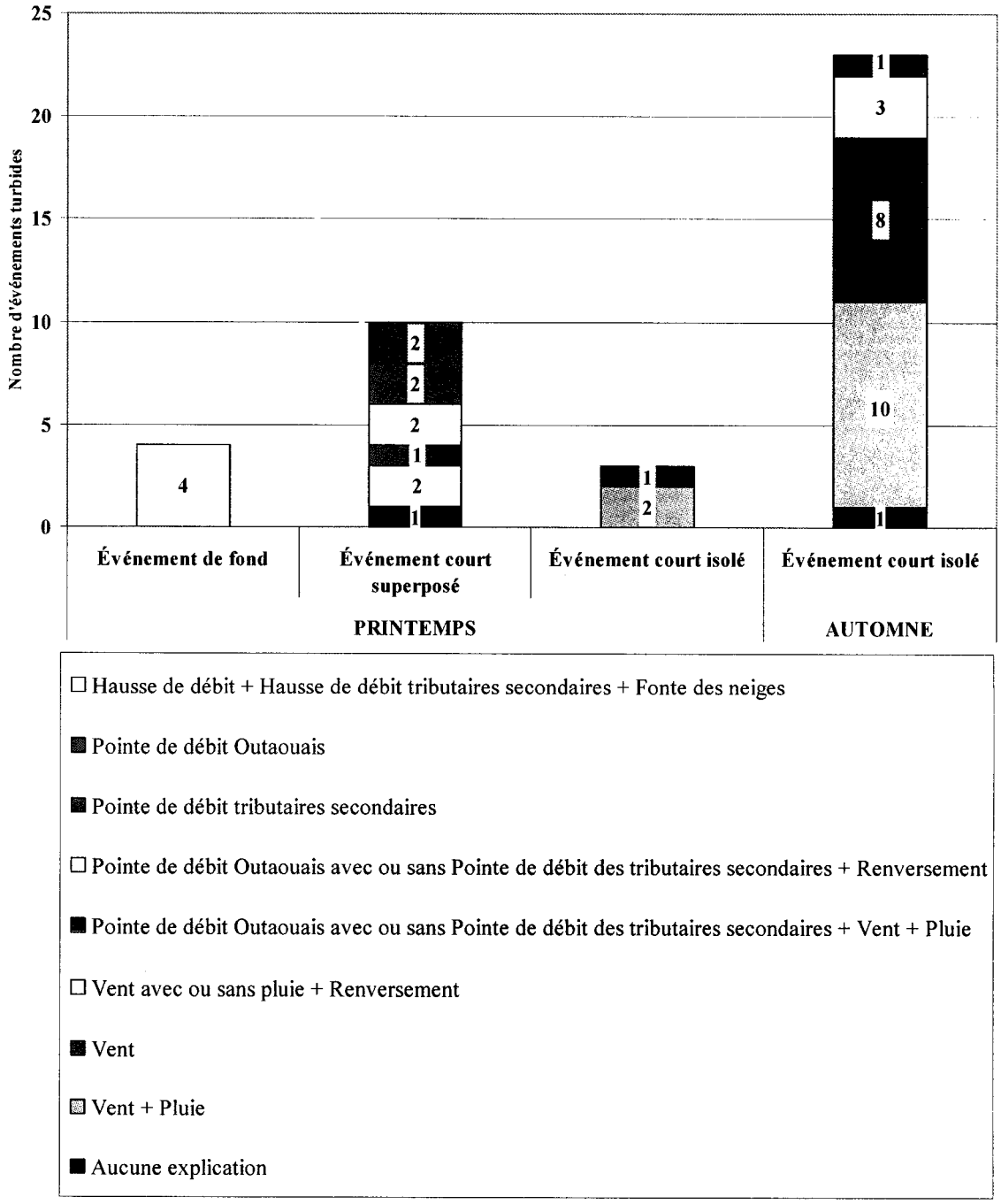


Figure 2-7 : Facteurs explicatifs en lien avec les événements turbides

Les quelques événements turbides courts isolés du printemps sont causés principalement par les tempêtes de vent avec précipitations qui agissent seules ou, dans un cas, simultanément avec une pointe du débit de la rivière des Outaouais. Ces résultats s'apparentent à ceux de l'automne, tel que le révèle la Figure 2-7. On y constate que les facteurs turbides de l'automne, tous de courte durée et isolés, résultent en grande majorité de tempêtes de vent avec ou sans précipitations agissant seuls. Dans quelques cas, le renversement automnal ou une pointe de débit de l'Outaouais s'y joignent.

Pour résumer, le Tableau 2-10 présente à quel point les facteurs explicatifs sont liés à chacun des types d'événement turbide du printemps et de l'automne. Par exemple, les tempêtes de vent font sentir leur influence dans respectivement 95% et 100% des événements turbides de courte durée isolés du printemps et de l'automne, mais dans seulement 30% des événements courts superposés du printemps. Ils n'ont aucune influence dans les événements turbides de fond.

Tableau 2-10 : Importance des facteurs explicatifs pour expliquer les hausses de turbidité du printemps et de l'automne

Facteurs explicatifs	PRINTEMPS Événement de fond	PRINTEMPS Événement court superposé	PRINTEMPS Événement court isolé	AUTOMNE Événement court isolé
Tempête de vent	0%	30%	**100%**	**95%**
Précipitations	0%	20%	**100%**	52%
Renversement	0%	40%	0%	12%
Débit	**100%**	**70%**	33%	5%
Fonte des neiges	**100%**	0%	0%	0%
Aucune explication	0%	10%	0%	4%

Finalement, il est important de noter que deux événements turbides courts demeurent inexpliqués, un au printemps et un autre, à l'automne. Ils peuvent résulter d'un autre facteur explicatif qui n'aurait pas été identifiée par la revue de littérature. Les facteurs

explicatifs examinés dans ce projet sont les plus communs. Mais il est possible, sinon probable, que d'autres facteurs d'occurrence plus rare, comme le dragage du lit fluvial, n'aient pas été pris en compte. Il faut aussi envisager la possibilité que quelques données de la base de données soient fausses malgré le soin minutieux employé pour son élaboration. Par exemple, si un épisode de vent important n'a pas été enregistré adéquatement à l'automne, cette donnée manquante expliquerait peut-être l'événement turbide automnal.

2.4.2.2 Identification des décalages

L'analyse des représentations graphiques des paramètres indicateurs a aussi permis d'identifier les décalages entre les variables d'entrée et de sortie et de ne retenir les variables de décalage les plus pertinentes. Ces résultats sont comparés à ceux obtenus par une matrice de corrélation entre les différentes variables potentielles et la turbidité à différents décalages (Tableau 2-11).

Les résultats obtenus avec les deux méthodes concordent bien, tant au printemps qu'à l'automne. Néanmoins, les décalages sont identifiés avec plus de certitude à l'automne avec l'analyse des représentations graphiques. Les événements explicatifs en cause étant moins nombreux qu'au printemps, il y a plus de cas pour déduire le décalage entre les paramètres indicateurs et les événements turbides. On constate aussi que les décalages pour le printemps et l'automne sont relativement similaires et que la majorité des décalages sont faibles, se situant entre 0 et 3 jours. La variable de température-degré du dégel et les débits des rivières Raisin et Beaudette représentent l'exception, variant de 4 à 17 jours selon les années pour la première et de 2 à 8 jours pour les seconds.

Tableau 2-11 : Décalage entre les paramètres indicateurs et les événements turbides

Paramètres indicateurs	Printemps		Automne	
	Décalage selon l'examen des graphiques	**Décalage selon la table de corrélation**	**Décalage selon l'examen des graphiques**	**Décalage selon la table de corrélation**
Débit de l'Outaouais au barrage Carillon	1 jours	1 jour	---	---
Contribution de la rivière Outaouais au débit total du fleuve à LaSalle	* aucun 1 jour 3 jours	1 jour	---	---
Index de température-degré de dégel (Dorval)	4 à 17 jours *10 jours	5 jours	* 3 jours	---
Débit de la rivière Raisin	2 à 8 jours	* 5 jours * 6 jours	* 1 jour	1 jour
Débit de la rivière Beaudette	2 à 7 jours	* 5 jours * 6 jours	* 1 jour	1 jour
Température de l'eau brute à Hawkesbury	1 à 4 jours	Non concluant	* 4 jours	Non concluant
Précipitations totales à Dorval	0 à 3 jours	2 jours	* 1 jour * 2 jours	2 jours
Précipitations totales à Sainte-Anne-de-Bellevue	2 à 3 jours	3 jours	* 1 jour * 2 jours	2 jours
Précipitations totales au lac St-François	2 à 3 jours	3 jours	* 1 jour * 2 jours	2 jours
Intensité moyenne du vent à Dorval	0 à 2 jours	Non concluant	1 à 3 jours * 1 jour * 2 jours	1 jour
Intensité maximale du vent à Dorval	0 à 2 jours	Non concluant	0 à 3 jours * 1 jour * 2 jours	1 jour
Intensité moyenne du vent au lac St-François	0 à 2 jours	Non concluant	0 à 2 jours * 1 jour	1 jour
Intensité maximale du vent au lac St-François	0 à 2 jours	Non concluant	1 à 2 jours * 1 jour	1 jour

* Cette valeur de décalage domine ou co-domine avec une autre valeur de décalage

2.4.3 Corrélation entre les variables d'entrée potentielles et la turbidité

La matrice de corrélation examinée pour déceler les décalages entre les paramètres indicateur et la turbidité récèle aussi des informations très intéressantes sur les liens existant entre la turbidité et les variables d'entrée potentielles non utilisées comme paramètres indicateurs. Le Tableau 2-12 présentent les variables les plus corrélées à la turbidité à un décalage donné pour le printemps et l'automne.

Tableau 2-12 : Corrélation et décalage entre la turbidité et différentes variables d'entrée potentielles

Paramètres indicateurs	Printemps		Automne	
	Décalage	Corrélation	Décalage	Corrélation
Couleur de l'eau brute à Desbaillet	aucun	**0,88**	aucun	**0,48**
Couleur de l'eau brute à Hawkesbury	1 jour	**0,45**	aucun	0,15
Turbidité de l'eau brute à Hawkesbury	1 jour	**0,63**	1 jour	0,32
Conductivité de l'eau à DesBaillets	Aucun	**-0,71**	Aucun	-0,18
Variables de débit de la rivière Outaouais (à Carrillon et aux canaux Sainte-Anne et Île-Perrot)	1 jour	**0,75**	Aucun	0,13
Contribution de la rivière des Outaouais au fleuve St-Laurent	1 jour	**0,71**	1 jour	0,16
Débit de la rivière Raisin	5 à 6 jours	**0,73**	1 jour	0,13
Débit de la rivière beaudette	4 à 5 jours	**0,77**	1 jour	0,14
Vitesse maximale du vent				
o à Dorval	1 jour	0,06	1 jour	**0,50**
o à Sainte-Anne-de-Bellevue	2 jours	0,09	1 jour	**0,45**
o au lac St-François	3 jours	0,09	1 jour	**0,40**
Vitesse moyenne du vent				
o à Dorval	1 jour	0,09	1 jour	**0,45**
o à Sainte-Anne-de-Bellevue	1 jour	0,09	1 jour	**0,43**
o au lac St-François	1 jour	0,06	1 jour	**0,45**
Précipitations locales				
o à Dorval	2 jour	0,20	2 jours	0,15
o à Sainte-Anne-de-Bellevue	3 jours	0,17	3 jours	0,22
o au lac St-François	3 jours	0,16	2 jours	0,15

Les corrélations significatives, identifiées en caractère gras dans le Tableau 2-12, indiquent clairement que les variables de qualité comme la couleur de l'eau brute à DesBaillets et à Hawkesbury, la turbidité de l'eau à Hawkesbury et la conductivité de l'eau à Desbaillets sont fortement corrélées à la turbidité de l'eau brute à DesBaillets principalement au printemps. C'est le cas également des variables de débit, venant ainsi appuyer les résultats précédents voulant que le débit joue un rôle important pour expliquer les événements turbides au printemps, et non à l'automne. De même, les données concernant le vent sont corrélées aux valeurs de turbidité à l'automne et non au printemps, ce qui concorde avec les résultats précédents identifiant les tempêtes de vent comme facteur explicatif principal des événements turbides automnaux. On peut noter que les précipitations locales sont peu corrélées à la turbidité, tant au printemps qu'à l'automne. Comme les précipitations importantes ne sont pas illustrées comme facteur explicatif significatif dans les travaux précédents, on peut donc déjà éliminer ces variables de l'ensemble de variables d'entrée proposées.

2.4.4 Perspectives pour le développement de réseaux de neurones artificiels

Les travaux de ce chapitre offrent des perspectives intéressantes pour le développement de réseaux neuronaux, particulièrement pour la sélection des variables d'entrée des modèles. Tout d'abord, la sélection de variables d'entrée potentielles, l'identification de paramètres indicateurs représentatifs des différents événements explicatifs et l'examen des corrélations entre les autres variables d'entrée potentielles et la turbidité orientent tous la sélection finale des variables d'entrée des modèles prédictifs. L'identification des décalages entre les données de turbidité et différentes variables d'entrée potentielles permet aussi de réduire le nombre de variables d'entrée potentielles en ne retenant que les variables au décalage opportun.

Une autre perspective à considérer est l'ajout de paramètres d'index et environnementaux. Afin d'éviter de développer deux modèles prédictifs distincts pour prendre en considération les différences entre les événements turbides et les facteurs explicatifs prédominants du printemps et de l'automne, des paramètres

environnementaux et des paramètres d'index réflétant ces différences peuvent être utilisés, tel que démontré par Zhang et Stanley (1997) dans la prévision de la couleur de l'eau brute (section 1.2.5.3). Les paramètres environnementaux aident le modèle à intégrer les contributions importantes à la couleur dues à des phénomènes saisonniers comme le dégel printanier tandis que les paramètres d'index permettent au modèle d'établir une ligne de base saisonnière. Ainsi, pour développer un seul modèle prédictif, il faudra ajouter aux variables d'entrée les paramètres environnementaux et d'index suivants :

- *Index de saison* : Paramètre reflétant les saisons de turbidité.
    - Index-saison = 0 au printemps, du 1er mars au 30 avril
    - Index-saison = 1 à l'automne, du 15 octobre au 14 janvier
    - Index-saison = 0,5 à l'été et à l'hiver, du 1er mai au 14 octobre et du 15 janvier au 28 février.

- *Index de gel / dégel :* Paramètre indiquant au modèle la période de gel et la présence d'une couverture de glace sur les rivières et le fleuve, période pendant laquelle l'occurrence d'événements explicatifs ne cause plus d'augmentations de turbidité.
    - Index-Gel = 1 si la température-degré-dégel d'il y a 5 jours = 0
    - Sinon, Index-Gel = 0

- *Index de fonte des neiges :* Paramètre indiquant au modèle la période de fonte des neiges.
    - Index-Fonte = 1 si la température-degré-dégel d'il y a 10 jours est entre 0 et 10 et que le débit de la rivière Raisin excède 7 m^3/s.
    - Sinon, Index-Fonte = 0

- *Index du renversement* : Paramètre reflétant la température de l'eau et indiquant au modèle l'occurrence possible du renversement avec plus de pertinence que les données brutes de température de l'eau.

- Index-Renversement = 1 si $Temp_{hier}$ et $Temp_{auj}$ était entre 3,5 et 4,5 deg.C il y a 4 jours, qui est la période de décalage estimée.
- Sinon, Index-Renversement = 0.

Les paramètres environnementaux et d'index présentent aussi l'avantage de représenter plus avantagement certaines variables d'entrée potentielles qui peuvent alors être éliminées, tout en prenant en compte les décalages implicitement. C'est le cas notamment de la température de l'eau brute pour le renversement.

En résumé, le Tableau 2-13 présentent les variables d'entrée retenues suite à tous les travaux du présent chapitre. Cet ensemble de variables n'est pas encore définitif. Une dernière évaluation de la pertinence des variables proposées devra être effectuée au chapitre suivant par des méthodes proposées par la revue de littérature, afin de tenter de réduire au maximum le nombre de variables d'entrée des modèles connexionnistes.

Tableau 2-13 : Variables d'entrée proposées pour le développement de modèles connexionnistes

	Variables	Code d'identification	Justification
Variables de qualité de l'eau	Turbidité de l'eau brute à DesBaillets de la veille	TURB_DB1	Pour prendre en compte les caractéristiques dynamiques de la turbidité
	Couleur de l'eau brute à Desbaillets	COUL_DB	Corrélation significative au printemps
	Couleur de l'eau brute à Hawkesbury	COUL_HAW	Corrélation significative au printemps
	Turbidité de l'eau brute à Hawkesbury	TURB_HAW	Corrélation significative au printemps
	Conductivité de l'eau à DesBaillets	COND_DB	Corrélation significative au printemps
Variables de débit	Variable de débit de la rivière Outaouais à Carillon de la veille	OUT_LAG1	Paramètre indicateur significatif Corrélation significative au printemps
	Contribution de la rivière des Outaouais au fleuve St-Laurent de la veille	CONT_LAG1	Paramètre indicateur significatif Corrélation significative au printemps
	Débit de la rivière Raisin d'il y a cinq jours	RAIS_LAG5	Paramètre indicateur significatif Corrélation significative au printemps
	Débit de la rivière Beaudette d'il y a quatre jours	BDT_LAG4	Paramètre indicateur significatif Corrélation significative au printemps
Variables météorologiques	Vitesse maximale du vent à Dorval de la veille	DOR_VITX	Paramètre indicateur significatif Corrélation significative à l'automne
	Vitesse maximale du vent au lac St-François de la veille	LSF_VITX	Paramètre indicateur significatif Corrélation significative à l'automne
	Vitesse moyenne du vent à Dorval de la veille	DOR_VITM	Paramètre indicateur significatif Corrélation significative à l'automne
	Vitesse moyenne du vent au lac St-François de la veille	LSF_VITM	Paramètre indicateur significatif Corrélation significative à l'automne
Variables d'index	Index de saison	IDX_SAIS	Différencie les périodes entre elles
	Index de gel / dégel	IDX_GEL	Indique la période de gel et de couvert de glace
	Index de fonte des neiges	IDX_FONT	Indique la période durant laquelle a lieu la fonte des neiges
	Index du renversement	IDX_RENV	Indique la période durant laquelle le reversement peut se produire au printemps et à l'automne

CHAPITRE 3 - PRÉVISION DES AUGMENTATIONS DE TURBIDITÉ PAR DES RÉSEAUX DE NEURONES ARTIFICIELS

Les travaux de ce chapitre visent à prévoir les augmentations de turbidité à l'eau brute de la Ville de Montréal. Pour ce faire, deux types de modèles sont développés, l'un prévoyant la valeur exacte de la turbidité du lendemain et l'autre, l'occurrence d'une hausse de turbidité, exprimée sous forme de classes de turbidité. Les travaux effectués dans ce chapitre s'appuient directement sur la méthodologie proposée dans la revue de littérature, permettant ainsi d'illustrer sa mise en application et d'en éprouver le potentiel et les limites pour un cas réel. Plus précisément, les travaux de ce chapitre consistent à :

1. *Identifier les besoins et les ressources.* Il s'agit d'une étape préliminaire essentielle car on s'y assure que les réseaux de neurones artificiels représentent une méthode de modélisation appropriée à utiliser pour le cas présent et que le développement du modèle soit orienté de façon à bien répondre aux besoins réels des utilisateurs futurs.

2. *Choisir le critère de performance.* Le choix peut s'effectuer parmi la liste des critères proposés couramment dans la littérature ou, comme c'est le cas pour ce projet, parmi d'autres critères qui prennent mieux en compte les particularités des besoins identifiés.

3. *Organiser la base de données.* Différentes pratiques proposées dans la revue de littérature sont mises à l'épreuve afin d'en évaluer la pertinence et les limites.

4. *Construire les modèles candidats.* Cette étape constitue le cœur du chapitre. Elle comprend tout d'abord la sélection des variables de sortie et des variables d'entrée des modèles, par laquelle les différentes techniques de sélection proposées dans la revue de littérature sont testées comparées, et la construction des modèles candidats en tant que tel grâce à un logiciel commercial recommandé, NeuralNetworks (Statsoft).

5. *Choisir et intégrer les modèles de prévision.* Le choix s'effectue parmi les modèles candidats en s'appuyant sur le critère de performance retenu et les deux modèles sont intégrés à un modèle de prévision opérationnel.

3.1 Évaluation des besoins et ressources disponibles

Évaluer les besoins et les ressources disponibles peut sembler être une étape superflue que plusieurs pourraient être tentés de négliger. Cette étape doit en fait être considérée comme une opportunité d'éviter facilement des écueils importants comme développer des modèles ne répondant pas aux besoins réels des utilisateurs ou utiliser la modélisation connexionniste alors que d'autres méthodes plus simples auraient pu être utilisées plus avantageusement.

3.1.1 Choix de la méthode de modélisation

Il existe un réel intérêt de connaître 24h à l'avance les fluctuations de turbidité de l'eau brute qui pourraient affecter les opérations de traitement de la Ville de Montréal. Néanmoins, des fonds importants ne peuvent être investis pour développer des modèles mécanistiques pouvant prévoir à l'avance ces fluctuations de turbidité. Les réseaux de neurones artificiels (RNA) se révèlent donc être une alternative intéressante du point de vue économique, particulièrement si on considère le fait qu'ils peuvent être intégrés relativement facilement et à faible coût à la plateforme de contrôle de la station de traitement afin d'y fournir en ligne les prévisions de la turbidité de l'eau brute.

Finalement, le travail effectué dans le chapitre précédent a révélé que les augmentations de turbidité sont causées par différents facteurs explicatifs, dont l'impact varie d'une période à l'autre de l'année et d'une année à l'autre. Les augmentations de turbidité sont donc le reflet d'un phénomène non-linéaire, difficile à prévoir. Les réseaux neuronaux articifiels étant particulièrement appropriés pour prévoir des phénomènes non-linéaires, il s'agit d'un autre incitatif important pour utiliser cette méthode plutôt qu'une autre.

3.1.2 Besoins à considérer

Plusieurs types de modèles de prévision par RNA peuvent être développés selon les besoins. Ce projet de maîtrise vise à évaluer le potentiel des RNA pour la prévision de la turbidité : dans cet optique, il est intéressant de développer un modèle prévoyant les valeurs exactes de la turbidité du lendemain. Mais du point d'un vue d'un opérateur, il peut être tout aussi intéressant de prévoir la valeur de la turbidité selon une classe ou de prévoir les augmentations significatives de turbidité en tant que telles, particulièrement si ces modèles sont plus robustes et justes. Ainsi, un opérateur préférera de loin un modèle prévoyant bien la turbidité du lendemain selon une classe de danger relatif (par exemple, vert signifie « eau claire » et rouge, « eau turbide ») dans 99% des cas qu'un modèle prévoyant bien la valeur exacte de la turbidité dans seulement 65% des cas. Deux modèles seront donc développés, l'un prévoyant les valeurs exactes de la turbidité du lendemain et le second prévoyant une classe de danger pour le lendemain et le surlendemain. Ces modèles pourraient d'ailleurs être utilisés de façon complémentaire dans une station de traitement.

3.1.3 Ressources disponibles

Toutes les ressources requises pour développer des modèles connexionnistes sont disponibles : l'étudiante responsable du présent projet dispose d'une connaissance suffisante du phénomène à modéliser et du fonctionnement des réseaux de neurones artificiels, d'un logiciel commercial recommandé, Statistica Neural Networks version 5 de Statsoft Inc. (Tulsa, Oklahoma), et du matériel informatique adéquat, constitué d'un ordinateur personnel avec un microprocesseur Pentium 4 de 2,4GHz et une mémoire vive de 333 MHz.

3.2 Choix du critère de performance

Dans la littérature, le principal critère retenu pour évaluer les modèles est la justesse de la prévision, mesurée le plus souvent par l'erreur quadratique moyenne de l'ensemble de test (RMSE). Dans le cas présent, le critère de performance retenu est aussi la

justesse de la prévision, mais la mesure choisie n'est pas l'erreur quadratique moyenne. Les essais de modélisation préliminaires pour la prévision des valeurs numériques de turbidité ont révélé (1) qu'il existait parfois un décalage d'une journée entre les valeurs prévues et les valeurs réelles et (2) que les modèles prévoyaient bien l'occurrence des fluctuations de turbidité mais non leur amplitude, les valeurs prévues étant souvent moindres que les valeurs réelles. Comme les besoins des utilisateurs portent davantage sur la prévision exacte des augmentations soudaines et importantes, i.e. quand la turbidité excède 3,2 UTN, que sur la prévision des valeurs exactes, il est préférable d'utiliser une mesure qui pénalise les décalages et les manquements à prévoir les augmentations significatives plutôt que les différences importantes entre les valeurs historiques et les valeurs prévues, comme le fait l'erreur quadratique moyenne. La mesure retenue est donc la corrélation entre la valeur réelle et la valeur prédite, exprimée sous la forme du coefficient de Pearson (r), car elle permet d'évaluer à quel point les valeurs sont proportionnelles entre elles plutôt que l'erreur qui existe entre elles.

Dans le cas du modèle de prévision des hausses de turbidité, la mesure de la justesse de la prévision est le taux de classification correcte des événements (en pourcentage). Exprimées sous forme de classes, les prévisions ne peuvent qu'être identiques ou différentes des classes réelles, simplifiant ainsi grandement l'évaluation de la performance. Une attention particulière est portée aux prévisions des événements turbides et plus particulièrement de la première journée de l'événement turbide.

3.3 Organisation de la base de données

La base de données contient les variables sélectionnées suite aux travaux du chapitre précédent (Tableau 2-13). Comme les données sont relativement abondantes (3 ans et 4 mois de données journalières, soit 1216 entrées), les données sont réparties en trois sous-ensembles selon le ratio 3:1:1. Cette façon d'organiser les données, suggéré par Baxter *et al.* (2002), permet d'utiliser la méthode de validation croisée comme critère d'arrêt. Comme les augmentations de turbidité à l'eau brute sont cycliques d'une année

à l'autre et seulement abondantes à certaines périodes de l'année, la répartition des données dans les trois sous-ensembles est effectuée de façon semi-aléatoire : les données correspondant aux événements turbides examinés dans le chapitre précédent sont réparties de façon structurée entre les trois sous-ensembles, alors que les autres données, dont la turbidité excède rarement 3,2 UTN, sont réparties de façon aléatoire.

La structure adoptée pour la répartition des événements turbides diffère pour le printemps et l'automne. Les données des périodes printanières ont été réparties en bloc d'année entre les trois sous-ensembles, un bloc étant constitué des données entre le 1er mars et le 30 avril de chaque année. Les données de deux périodes printanières présentant les augmentations les plus extrêmes (1998 et 1999) ont donc été attribuées à l'ensemble d'apprentissage, car un modèle connexionniste ne peut prévoir au-delà de l'échelle des données dont il dispose pour son développement. Les données des printemps 2000 et 2001 ont été attribuées aux ensembles de test et de validation. Pour respecter le ratio de 3 :1:1, l'idéal aurait été de disposer de données pour cinq périodes printanières plutôt que quatre, comme c'est le cas actuellement. Pour pallier à ce problème, des données correspondant à des périodes non-turbides de l'année ont comblé le manque à gagner pour l'ensemble d'apprentissage.

Les données des périodes automnales ont été attribuées différemment entre les trois sous-ensembles. Contrairement aux événements turbides printaniers, les événements turbides automnaux sont courts et isolés les uns des autres. Ils ont donc été répartis entre les trois sous-ensembles selon le ratio pré-déterminé de façon à ce que chacun ait des événements de toutes les intensités et de toutes les années et aussi de façon à ce que la somme des jours « turbides » attribués aux trois sous-ensembles respecte le ratio retenu de 3:1:1.

Le processus de répartition des données entre les trois sous-ensembles a été répété deux fois afin de subdiviser les données de deux façons différentes. Les modèles candidats sont développés à partir de la première répartition des données. Si le modèle démontre une bonne performance, il est entraîné de nouveau en utilisant la seconde répartition des

données. Cette façon de faire permet de vérifier si la performance des modèles candidats développés est indépendante de la manière dont les données sont séparées en trois sous-ensembles. Les deux répartitions des événements turbides entre les trois sous-ensembles est détaillée à l'Annexe 7.

La distribution des données résulte en quelques sous-ensembles dont les propriétés statistiques diffèrent des autres ($p<0,05$). C'est le cas d'un des sous-ensembles de test (répartition 1), qui diffère de tous les autres sous-ensembles, et des deux ensembles de validation, qui diffèrent significativement l'un de l'autre (Tableau 3-1). Selon Baxter *et al.* (2002), il ne s'agit pas d'une situation souhaitable car la performance de prévision des modèles candidats devrait être indépendante de la manière dont les données ont été séparées en trois sous-ensembles. Dans le cas présent, les différences s'expliquent par la répartition des données des périodes printanières, où les valeurs extrêmes ont été attribuées dans les deux répartitions à l'ensemble d'apprentissage alors que l'idéal aurait été de disposer d'autres données de périodes printanières exceptionnelles comme celle de 1998 afin de les joindre aux sous-ensembles de test et de vérification. Mais comme ce n'est pas le cas, il est préférable que le modèle apprenne bien plutôt que les sous-ensembles de données soient statistiquement semblables. Éventuellement, le modèle pourra être amélioré par l'inclusion d'autres données d'événements turbides printaniers.

Tableau 3-1 : Propriétés statistiques des sous-ensembles d'apprentissage, de test et de validation selon les deux répartitions des données

Mesures statistiques pour TURB_1	Données d'apprentissage		Données de test		Données de validation	
	Répartition 1	Répartition 2	Répartition 1	Répartition 2	Répartition 1	Répartition 2
Moyenne	2,05	2,04	1,62	1,94	2,09	1,80
Écart-type	2,37	2,32	1,57	1,99	2,51	2,18
Maximum	26,0	22,8	16,0	26,0	22,8	24,0
Minimum	0,38	0,38	0,40	0,48	0,41	0,41

La dernière étape avant la construction des modèles candidats réside dans le traitement des données. Ce dernier est effectué automatiquement par le logiciel Neural Networks, qui met à l'échelle et normalise les données des variables d'entrée par la fonction

Minimax, qui utilise un algorithme assignant des coefficients de mise à l'échelle linéaire aux données.

3.4 Construction des modèles candidats

Dans cette section, la méthodologie employée et les résultats obtenus pour développer les modèles connexionnistes sont présentés et discutés. La construction des modèles s'effectue en trois étapes principales, soit la sélection des variables de sortie, la sélection des variables d'entrée des modèles et finalement le développement des modèles candidats en tant que tel avec l'aide du logiciel commercial Neural Networks.

3.4.1 Sélection des variables de sortie

Deux variables de sortie ont été sélectionnées. Des modèles différents seront développés pour chacune, puisque c'est une approche à privilégier selon Baxter *et al.* (2002). La première variable de sortie sélectionnée est la différence entre la turbidité de l'eau brute d'aujourd'hui et de demain (DIFF_1), qui permettra d'estimer la valeur numérique de la turbidité prévue pour le lendemain en ajoutant la différence prévue par le modèle à la valeur réelle de la turbidité d'aujourd'hui. La différence entre la turbidité d'aujourd'hui et du lendemain est retenue comme variable de sortie plutôt que la valeur de la turbidité du lendemain (TURB_1) en raison de l'auto-corrélation significative des valeurs de turbidité. L'auto-corrélation est jugée significative car le coefficient de corrélation de Pearson (r) entre la valeur de la turbidité d'aujourd'hui et celle d'hier excède 0,50 dans le cas de la base de données complète (r = 0,52) et des données des périodes printanières (r = 0,77). Par contre, les données de turbidité à l'automne sont peu auto-corrélées (r de 0,25), indiquant que dans l'éventualité où deux modèles distincts devaient être développés pour prévoir la turbidité au printemps et à l'automne, la valeur de la turbidité de lendemain pourrait être utilisée directement comme variable de sortie dans le cas du modèle automnal.

La seconde variable de sortie a été sélectionnée sur la base de l'usage éventuel du modèle connexionniste. La prévision souhaitée est la valeur numérique de la turbidité

du lendemain, qui nécessite le développement de modèles de régression. Mais comme il s'avère que les besoins des opérateurs ne consistent pas tant à connaître les valeurs exactes de turbidité du lendemain que de prévoir les hausses de turbidité en tant que telles, des modèles de classification peuvent donc être élaborés afin de savoir si la turbidité du lendemain dépassera un seuil donné de turbidité. Dans ce cas-ci, la variable de sortie décrivant la qualité de l'eau du lendemain (EAU_1) consiste en deux classes, attribuées selon la valeur numérique de la turbidité du lendemain : si la turbidité excède 3,2 UTN, la donnée est classée sous « turbide » alors que dans les autres cas, la donnée est classée sous «claire ». Plus de classes auraient pu être utilisées, mais cette possibilité a été rejetée car les essais préliminaires indiquent que la performance des réseaux est diminuée par l'usage de plus de deux classes.

Développer deux types de modèles différents peut sembler un exercice dont l'utilité est discutable, surtout si le premier réseau développé a une bonne performance. Mais en examinant bien les besoins des opérateurs de la station de traitement, on constate que la robustesse du modèle est très importante : manquer la prévision des hausses importantes de turbidité peut résulter en une perte de confiance importante envers le système de prévision, qui pourrait ne plus être utilisé par les opérateurs malgré son utilité pour accroître leur vigilance et améliorer le traitement. En implantant un système de prévision en station qui s'appuie sur deux modèles prévoyant les augmentations de turbidité de façon différente, la justesse des prévisions respectives peut être contre-vérifiée et les erreurs, comme celle de prévoir en retard le début d'un événement turbide important, peuvent être possiblement corrigées. Il est intéressant de noter aussi que développer deux modèles distincts requiert plus de travail au niveau de la modélisation, mais pas au niveau de l'utilisation en ligne à la station de traitement.

3.4.2 Sélection des variables d'entrée

La méthodologie suggérée par la revue de littérature consiste (1) à réduire le nombre de variables d'entrée à examiner, incluant les variables de décalage, à l'aide de techniques non-supervisées et (2) à identifier ensuite les variables d'entrée les plus significatives

selon leur impact sur l'habilité de prévision du RNA à l'aide de techniques supervisées. Dans cette section, différentes techniques non-supervisées et supervisées sont utilisées et comparées entre elles. Les techniques non-supervisées comprennent l'identification *a priori* des variables significatives et l'analyse par composantes principales (ACP), tandis que les techniques supervisées examinées sont les algorithmes génétiques (AG), la construction par étape de RNA bi-variables (CE) et le design expérimental Plackett-Burman.

3.4.2.1 Techniques non-supervisées

Comme le nombre de variables d'entrée avait été considérablement réduit grâce à l'identification *a priori* des variables significatives lors travaux du chapitre 2, la technique analytique non-supervisée retenue pour réduire le nombre de variables à examiner, l'analyse par composantes principales (ACP), n'a pas permis d'identifier clairement des variables redondantes à éliminer. Pour pouvoir éliminer des variables par l'ACP, le modélisateur considère que les trois premières composantes principales doivent représenter ensemble au moins 85% de la variabilité des différentes variables examinées. Les valeurs propres (*Eigenvalues*) cumulatives des trois premières composantes principales doivent donc excéder 85%. Comme ce n'est pas le cas, on peut supposer que les variables ne véhiculent pas la même information et qu'elles devraient toutes être retenues pour l'étape suivante. Ainsi, dans ce cas particulier, le recours à la seule technique d'identification *a priori* semble suffisant pour réduire le nombre de variables en entrée, la technique des ACP n'apportant rien de plus.

3.4.2.2 Techniques supervisées

Pour identifier les variables d'entrée significatives, trois techniques supervisées ont été comparées. Les deux premières techniques examinées, les algorithmes génétiques (AG) et la construction par étape de RNA bi-variables (CE), ont été présentées dans la revue de littérature (voir section 1.2.4.4 pour tous les détails). Le logiciel NeuralNetworks propose un module permettant d'utiliser facilement les AG et la CE (la CE y est

identifiée comme la technique « *forward stepwise* »). Trois essais ont été effectués avec chacune des deux techniques pour chaque variable de sortie, soit DIFF_1 et EAU_1, en utilisant les valeurs des paramètres internes suggérées par défaut par le logiciel. Pour chaque essai, toutes les variables d'entrée potentielles sélectionnées au chapitre précédent (Tableau 2-13) ont été incluses. Les trois essais permettent de s'assurer de la réplicabilité des résultats, i.e. que les variables identifiées par chacune des techniques pour les deux variables de sortie sont les mêmes d'un essai à l'autre. Si les différences sont nombreuses entre les trois essais, ceci peut soit signifier que les variables ne sont pas significatives ou que les paramètres internes utilisés par défaut pour construire les RNA nécessaires sont inadéquats. Les variables considérées comme significatives sont celles qui ont été identifiées comme telles dans au moins deux essais sur trois.

La troisième technique est celle du design expérimental Plackett-Burman (PB). Avec cette technique, 20 RNA sont développés à partir de 20 sous-ensembles différents de variables d'entrée pour chacune des deux variables de sortie. Les RNA sont construits en utilisant l'outil automatisé de Neural Networks, appelé « *Intelligent Problem Solver* » (IPS). Cet outil utilise une technique d'optimisation non-linéaire permettant de trouver par recherche heuristique les meilleures architectures de RNA. Les 20 sous-ensembles de variables d'entrée sont définis par la matrice de design expérimental Plackett-Burman pour 20 essais / 19 variables disponibles. La performance des RNA est notée pour l'ensemble des données et pour les événements turbides seulement, les critères de performance utilisés étant ceux définis à la section 3.2. Ces valeurs de performance sont intégrées à la matrice de design expérimental afin d'effectuer une analyse de variance (ANOVA). Une variable est considérée comme significative lorsque la variabilité qu'elle représente excède 5%. Il est important de noter que les essais ne sont pas répétés trois fois comme pour les AG et la CE car la performance des réseaux développés pour un sous-ensemble de variables donné varie très peu d'un essai à l'autre.

La technique du design expérimental Plackett-Burman (PB) n'a pas été mentionnée lors de la revue de littérature. Elle est néanmoins intégrée à ce chapitre car elle permet d'identifier les variables significatives lors des événements turbides spécifiquement, contrairement aux deux autres techniques. Pour en être capables, les techniques des AG et de la CE devraient utiliser les données des événements turbides uniquement, mais c'est impossible dans le cas présent car le nombre de données liées à des événements turbides est limité alors que le logiciel requiert, pour une opération adéquate, un grand nombre de données disponibles par rapport au nombre de variables d'entrée à évaluer. De plus, la technique PB permet d'identifer l'importance relative de chaque variable d'entrée, ce que les autres techniques ne permettent pas.

Réplicabilité des résultats avec les techniques des algorithmes génétiques (AG) et de la construction par étape (CE)

La répétition des essais avec les techniques des AG et de la CE a permis de vérifier la réplicabilité des résultats. Pour qu'on puisse considérer que la technique est utile, les résultats doivent être sensiblement les mêmes d'un essai à l'autre pour une variable de sortie donnée. Le Tableau 3-2 révèle que c'est le cas. Par exemple, pour la technique des AG avec la valeur de sortie EAU_1, les trois essais ont donné les mêmes résultats pour 89% des variables d'entrée soumises à l'évaluation. Il est intéressant de noter que l'usage de la variable de sortie EAU_1 résulte en une meilleure réplicabilité des résultats que la variable DIFF_1. Cette différence entre les deux variables de sortie peut s'expliquer par le fait que les techniques requièrent la construction de RNA de types différents pour les deux variables de sortie, soit des réseaux neuronaux probabilistiques (probabilistic *neural networks*) (PNN) pour la variable de classe EAU_1 et des réseaux neuronaux de régression généralisée (*generalized regression neural network*) (GRNN) pour la variable numérique DIFF_1. Dans ce cas-ci, les PNN semblent mieux performer.

Tableau 3-2 : Réplicabilité des résultats des essais avec les algorithmes génétiques et la construction par étape

	Technique des algorithmes génétiques		Technique de la construction par étape	
	EAU_1	**DIFF_1**	**EAU_1**	**DIFF_1**
Résultats identiques pour les trois essais	89%	72%	89%	79%
Résultats différents entre les trois essais	11%	28%	11%	21%

Variables d'entrée identifiées par les trois techniques

Le Tableau 3-3 présentent les variables d'entrée significatives identifiées par les techniques des algorithmes génétiques (AG), de la construction par étape (CE) et du design expérimental de Plackett-Burman (PB) pour les deux ensembles de données analysées (année complète et événements turbides exclusivement). Les variables d'entrée identifiées par les trois techniques diffèrent d'une technique à l'autre et d'une variable de sortie à l'autre. Dans l'optique où on souhaite sélectionner un seul ensemble de variables d'entrée pour construire les deux modèles de prévision, ces résultats souvent contradictoires sont confondants. Néanmoins, il est possible d'en tirer des observations facilitant le choix final des variables d'entrée et nous éclairant aussi sur les particularités de chaque technique supervisée.

Tableau 3-3 : Comparaison des variables significatives identifiées par les trois techniques supervisées examinées

	Technique des algorithmes génétiques	Technique de la construction par étape	Design expérimental Plackett-Burman (pour toute l'année)	Design expérimental Plackett-Burman (pour les événements turbides exclusivement)
EAU_1	COUL_DB TURB_DB COND_DB TURB_DB1 TURB_HAW COUL_HAW OUT_LAG1 CONT_LAG BDT_LAG4 RAI_LAG5 LSF1_VITM LSF1_VITX DOR1_VITX DOR1_VITM IDX_SAIS	COUL_DB TURB_DB COND_DB TURB_DB1 TURB_HAW COUL_HAW OUT_LAG1 CONT_LAG BDT_LAG4 RAI_LAG5 LSF1_VITM LSF1_VITX DOR1_VITX DOR1_VITM IDX_SAIS IDX_FONT	TURB_DB (52,5%) COUL_DB (21,6%) COND_DB (13,5%)	TURB_DB (40%) COUL_DB (22,4%) COND_DB (8,3%) RAI_LAG5 (5,3%) IDX_FONT (4,9%)
DIFF_1	TURB_DB TURB_DB1 COUL_DB LSF1_VITM LSF1_VITX DOR1_VITX	TURB_DB TURB_DB1 COUL_DB LSF1_VITM	TURB_DB (6,4%) TURB_HAW (5,2%) TURB_DB1 (5,1%) COUL_DB (5,1%) COUL_HAW (5,0%) COND_DB (4,9%) OUT_LAG1 (5,4%) CONT_AUJ (5,3%) CONT_LAG (5,1%) RAI_LAG5 (5,1%) BDT_LAG4 (4,9%) DOR1_VITX (5,6%) DOR1_VITM (5,4%) LSF1_VITX (5,3%) LSF1_VITM (5,1%) IDX_RENV (5,4%) IDX_FONT (5,3%) IDX_SAIS (5,2%) IDX_GEL (5,1%)	TURB_DB (22,6%) TURB_HAW (4,9%) OUT_LAG1 (6,1%) CONT_AUJ (5,2%) DOR1_VITX (9,5%) DOR1_VITM (5,8%) LSF1_VITX (5,5%) IDX_SAIS (5,5%)

Note : La description des variables d'entrée est présentée à la page 99.

Tout d'abord, on peut constater que pour une variable de sortie donnée, les variables d'entrée significatives sont sensiblement les mêmes pour les techniques des AG et de la

CE. Mais, elles diffèrent selon la variable de sortie utilisée, les variables significatives étant beaucoup nombreuses avec EAU_1 qu'avec DIFF_1. Cette différence peut s'expliquer à nouveau par le recours à des types de RNA différents selon les deux variables de sortie, soit des RNA de régression et des RNA de classification.

Les résultats obtenus avec la technique de PB varient aussi selon la variable de sortie examinée et aussi selon les données utilisées pour l'analyse, soit celles des événements turbides exclusivement ou celles de toute l'année. Dans le cas de EAU_1, peu de variables sont identifiées, contrairement aux résultats obtenus avec les techniques des AG et la CE. Les variables significatives sont toutes des variables de qualité de l'eau à DesBaillets, la turbidité à l'eau brute (TURB_DB) devançant toutes les autres avec une importance relative de 52,5% (pour toute l'année) et de 40% (pour les événements turbides). On ne retrouve aucune variable de vent et pratiquement aucune variable d'index ou de débit pour les événements turbides, si ce n'est du débit d'il y a cinq jours de la rivière Raisin (RAI_LAG5) et de l'index de la fonte des neiges (IDX_FONT).

Par contre, les résultats sont bien différents avec la variable de sortie DIFF_1. Alors que seules quelques variables de vent et de qualité de l'eau à DesBaillets ont été identifiées avec les techniques des AG et de la CE, la technique de PB analysant les données de toute l'année identifie pratiquement toutes les variables d'entrée comme des variables significatives. L'importance relative est sensiblement la même d'une variable à l'autre, soit entre 4,9% et 6,4%. Dans le cas de l'analyse des événements turbides exclusivement, le sélection effectuée par la technique PB est plus restreinte. La variable TURB_DB est de nouveau la variable la plus significative avec une importance relative de 22,6%, à laquelle s'ajoutent certaines variables de vent, de débit et d'index, dont l'importance relative est plus faible.

Il est intéressant de noter que la sélection effectuée par la technique de PB pour DIFF_1 avec les événements turbides exclusivement est en fait un sous-ensemble des variables identifiées par les techniques des AG et de la CE pour EAU_1. Comme les techniques des AG et de la CE utilisent toutes deux les données de toutes l'année, on aurait pu

penser que ce sont les résultats de la technique de PB pour l'ensemble de l'année et non pour les événements turbides qui auraient été semblables. Ceci peut laisser présager que les tehcniques des AG et de la CE sont capables de déceler les tendances importantes parmi les données de toute l'année.

Sélection finale des variables en entrée

L'ensemble final de variables d'entrée servant de point de départ pour le développement de modèles candidats de prévision de DIFF_1 et de EAU_1 est présenté au Tableau 3-4 selon le type de variables. L'ensemble rassemble toutes les variables identifiées par les techniques des AG et de la CE pour EAU_1. Il s'agit aussi d'un sous-ensemble des variables identifiées par la technique PB (année complète) pour DIFF_1, et d'un ensemble élargi des variables identifiées par PB (événements turbides). Les variables du Tableau 3-4 sont réparties en sous-ensembles pour faciliter les essais de modélisation. Les sous-ensembles sont présentés à l'Annexe 8. Les résultats de l'analyse de sensibilité conduite par l'IPS pour l'ensemble des variables d'entrée sont aussi présentés à l'Annexe 8 car ces résultats sont à la source de certains sous-ensembles de variables d'entrée.

Tableau 3-4 : Sélection finale des variables en entrée

Qualité	Débit	Vent	Index
COUL_DB	OUT_LAG1	DOR1_VITX	IDX_SAIS
TURB_DB	CONT_LAG	DOR1_VITM	IDX_FONT
COND_DB	CONT_AUJ	LSF1_VITX	
TURB_DB1	BDT_LAG4	LSF1_VITM	
TURB_HAW	RAI_LAG5		
COUL_HAW			

Note : La description des variables d'entrée est présentée à la page 99.

3.4.3 Développement des modèles candidats

Trois types de modèles sont développés dans cette section. Les deux premiers prévoient les variables de sortie EAU_1 et DIFF_1, pour lesquels des modèles candidats sont développés avec les variables d'entrée proposés au Tableau 3-4. Le troisième type de modèle développé consiste en un modèle opérationnel, qui combine les résultats des

deux premiers types de modèle afin d'accroître la robustesse de la prévision fournie aux opérateurs de la station de traitement.

La construction et l'évaluation des modèles candidats sont effectuées en suivant la méthodologie proposée dans la revue de littérature et synthétisée aux Figures 1-13 et 1-14. Voici, en résumé, les étapes concrètes effectuées pour construire et évaluer les modèles de régression pour la prévision de DIFF_1 et les modèles de classification pour la prévision de EAU_1 :

1. L'outil automatisé de construction de RNA de Neural Networks, désigné par « *Intelligent Problem Solver* » (IPS), construit dix RNA aux architectures variées pour chaque sous-ensemble de variables d'entrée (voir Annexe 8). La répartition 1 des données en trois sous-ensembles est utilisée (Annexe 7) à cette étape. Le critère d'arrêt est la méthode de la validation croisée. Tous les autres paramètres sont déterminés par recherche heuristique par l'IPS.

2. Pour les meilleurs modèles de deux architectures différentes, l'erreur de l'ensemble de validation est comparée à l'erreur de l'ensemble d'apprentissage pour s'assurer que la différence est de moins de 35%.

3. Si c'est le cas, on examine le critère de performance. On examine aussi la représentation graphique pour identifier la présence de décalages temporels entre les valeurs réelles et les valeurs prédites et la justesse de la prévision des événements turbides faisant partie de l'ensemble de validation. Cet ensemble n'ayant pas servi à construire le modèle, on peut considérer les données qu'il contient comme des données nouvelles aptes à tester le potentiel de prévision du modèle.

4. Si les résultats sont satisfaisants, on ré-entraîne le même modèle avec les données distribuées selon la répartition 2 (Annexe 7). On vérifie que les erreurs sont semblables et que les résultats sont encore satisfaisants. Si oui, on retient le modèle développé comme modèle candidat.

La construction des modèles candidats de régression et de classification s'effectue en parallèle, tel qu'illustrée à la Figure 3-1. Les mêmes variables d'entrée sont utilisées pour développer les deux types de modèles. Cependant, l'amélioration des modèles candidats peut mener à éliminer certaines variables de l'ensemble initial de variables d'entrée, résultant par exemple en un modèle de classification ayant comme variables d'entrée un sous-ensemble des variables d'entrée du modèle de régression retenu.

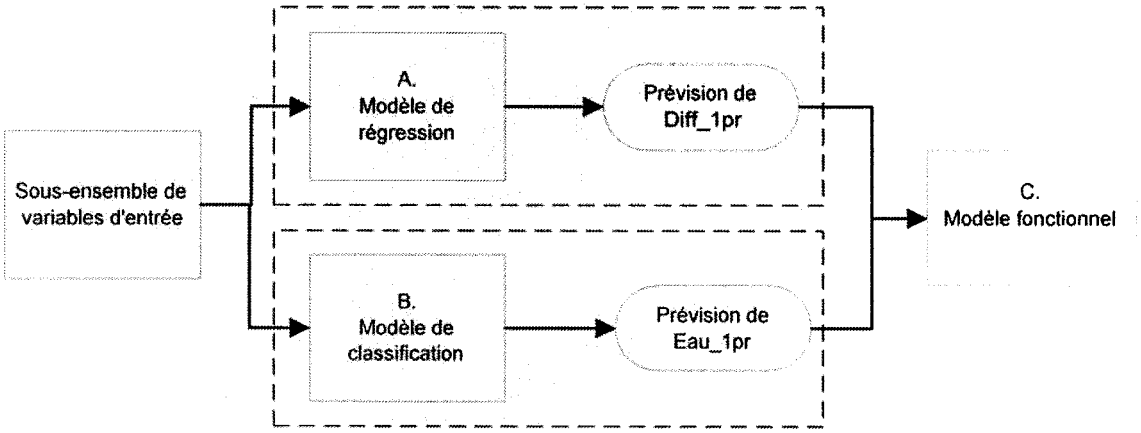


Figure 3-1 : Schéma résumant les différentes étapes et les différents modèles à contruire

3.4.4 Évaluation des modèles candidats

Les résultats de tous les modèles de régression et de classification développés sont présentés respectivement à l'Annexe 9 et à l'Annexe 10. Les résultats comprennent la différence entre l'erreur d'apprentissage et l'erreur de validation, les coefficients de corrélation dans le cas des modèles de régression et les pourcentages de classification correcte dans le cas des modèles de classification, l'évaluation qualitative des décalages temporels et la justesse de prévision des données de validation.

3.4.4.1 Tri préliminaire

Les observations découlant de l'examen attentif des résultats présentés aux Annexes 9 et 10 permettent d'effectuer un tri parmi les modèles développés pour ne retenir que les plus intéressants. On constate tout d'abord la présence de RNA de différentes

architectures au sein de la banque de modèles candidats potentiels, soit les perceptrons multicouches (PMC), les réseaux fonction à radiale de base (*Radial basis function* ou RBF) et les réseaux de neurones de régression généralisée (*General Regression Neural Networks* ou GRNN) pour les modèles de régression. La performance des modèles varie d'un type de modèle à l'autre et, pour un type donné, selon le nombre et la nature des variables d'entrée. Les modèles de chaque type d'architecture sont comparés entre eux afin de ne retenir que les meilleurs, i.e. ceux dont les coefficients de corrélation ou les taux de classification exacte sont les plus élevés et ceux qui présentent le moins de décalage et prévoient le mieux les données de validation.

3.4.4.2 Sélection finale

Les modèles retenus comme modèles candidats pour les variables de sortie DIFF_1 et EAU_1 sont présentés aux Tableaux 3-5 et 3-6. Les critères retenus pour comparer entre eux ces modèles et identifier le meilleur comprennent le nombre et l'accessibilité des variables d'entrée et la justesse des prévisions effectuées. Ces critères et leur pondération respective, choisis de façon heuristique selon le jugement de l'auteur, ont été intégrés dans une équation permettant de déterminer la performance globale des modèles candidats (donnée également aux Tableaux 3-5 et 3-6) :

$$\mathbf{P = 80\% \times C + 10\% \times N + 10\% \times [VF / (VF + VD)]} \qquad \textbf{(Équation 5)}$$

La variable **P** correspond à la performance globale du modèle évalué. La variable **C** correspond au coefficient de corrélation dans le cas des modèles de régression ou au taux de classification exacte pour le début des événements turbides pour les modèles de classification; La variable **N** est un facteur compris entre 0 et 1 dont la valeur est déterminée selon le nombre de variables en entrée du modèle : s'il y a moins de 10 variables, N = 1; s'il y en a entre 10 et 15, N = 0.5; s'il y a plus de 25 variables, N = 0. Finalement, les variables **VF** et **VD** correspondent respectivement au nombre de variables en entrée faciles et difficiles à obtenir. Les variables considérées comme « faciles » sont toutes les variables de qualité à la station DesBaillets et la variable

d'index de saison (IDX_SAIS) alors que toutes les autres sont considérées comme difficiles à obtenir.

Tableau 3-5 : Performance et caractéristiques des modèles candidats retenus pour la variable de sortie DIFF_1

Architect. du RNA	Variables d'entrée	Coefficient de corrélation		Inspection poussée des représentations graphiques		Performance globale du modèle
		Année complète	Évén. turbides	Présence de décalages temporels	Justesse de prév. des données de validation	
PMC 6:19:1	**(s.-ens. 10)** TURB_DB TURB_HAW LSF1_VITM LSF1_VITX DOR1_VITM IDX_SAIS	0,68 (0,67)	0,82 (0,78)	Surtout lors événements turbides soudains	Aut. : 1 / 4 Print. : 1 / 1 mais présence d'oscillations confondantes	**0,79**
GRNN 8:740:2:1	**(s.-ens. 6)** COUL_DB TURB_DB COND_DB OUT_LAG1 LSF1_VITX DOR1_VITX DOR1_VITM IDX_SAIS	0,86 (0,83)	0,94 (0,89)	Surtout lors événements turbides soudains	Aut. : 2 / 4 Print. : 1 / 1 mais avance dans la prévision	**0,90**
RBF 6:60:1	**(s.-ens. 11)** TURB_DB OUT_LAG1 LSF1_VITM LSF1_VITX DOR1_VITM IDX_SAIS	0,67 (0,60)	0,80 (0,73)	Oscillations au printemps qui sont confondantes	Aut. : 1 / 4 Print. : 1 / 1 mais présence d'oscillations confondantes	**0,77**

Note : Les résultats entre parenthèses sont obtenus avec la répartition 2 des données

Tableau 3-6 : Performance et caractéristiques des modèles candidats retenus pour la variable de sortie EAU_1

Architec. du RNA	Variables d'entrée	Classification correcte entre « eau claire » et « eau turbide »				Performance globale du modèle
		Toute l'année	Évén. turbides	Début évén. turbides	Début évén. de valid.	
PMC 17 :3 :1	**(s.-ens. 0)** **COUL_DB** TURB_DB COND_DB TURB_DB1 TURB_HAW COUL_HAW OUT_LAG1 CONT_LAG CONT_AUJ BDT_LAG4 RAI_LAG5 LSF1_VITM LSF1_VITX DOR1_VITX DOR1_VITM IDX_SAIS IDX_FONT	82% (82%)	90% (89%)	87% (77%)	Aut. : 3 / 4 Print.: 1 / 1 (Aut. : 3 / 4) (Print.: 1 / 2)	**0,72**
PMC 12 :4 :1	**(s.-ens. 5)** TURB_DB COND_DB TURB_DB1 OUT_LAG1 CONT_LAG BDT_LAG4 RAI_LAG5 LSF1_VITM LSF1_VITX DOR1_VITM IDX_SAIS IDX_FONT	81% (84%)	90% (89%)	81% (77%)	Aut. : 3 / 4 Print.: 1 / 1 (Aut. : 4 / 4) (Print.: 0 / 2)	**0,73**
PMC 6 :9 :1	**(s.-ens. 10)** TURB_DB TURB_HAW LSF1_VITM LSF1_VITX DOR1_VITM IDX_SAIS	82% (86%)	91% (91%)	77% (81%)	Aut. : 3 / 4 Print.: 1 / 1 (Aut. : 3 / 4) (Print.: 1 / 2)	**0,78 (répart. 2)**
PMC 6 :5 :1	**(s.-ens. 11)** TURB_DB OUT_LAG1 LSF1_VITM LSF1_VITX DOR1_VITM IDX_SAIS	83% (84%)	92% (92%)	81% (81%)	Aut. : 3 / 4 Print.: 1 / 1 (Aut. : 3 / 4) (Print.: 2 / 2)	**0,78**

Note : Les résultats entre parenthèses sont obtenus avec la répartition 2 des données

Le modèle retenu pour prévoir la différence entre la turbidité d'aujourd'hui et demain à DesBaillets (DIFF_1) est le réseau de neurones de régression généralisée (GRNN 8:740:2:1), qui affiche une performance globale de 0,90. Les huit variables d'entrée sont faciles à obtenir : la variable d'index (IDX_SAIS) est calculée à partir de la date, les variables de qualité (COUL_DB, TURB_DB et COND_DB) sont fournies directement à la station de traitement DesBaillets et les variables de débit et de vent (OUT_LAG1, LSF1_VITX, DOR1_VITX, DOR1_VITM) sont disponibles à Environnement Canada. La Figure 3-2 donne un aperçu des prévisions effectuées par le modèle de régression retenu. Toutes les représentations graphiques peuvent être trouvées à l'Annexe 11.

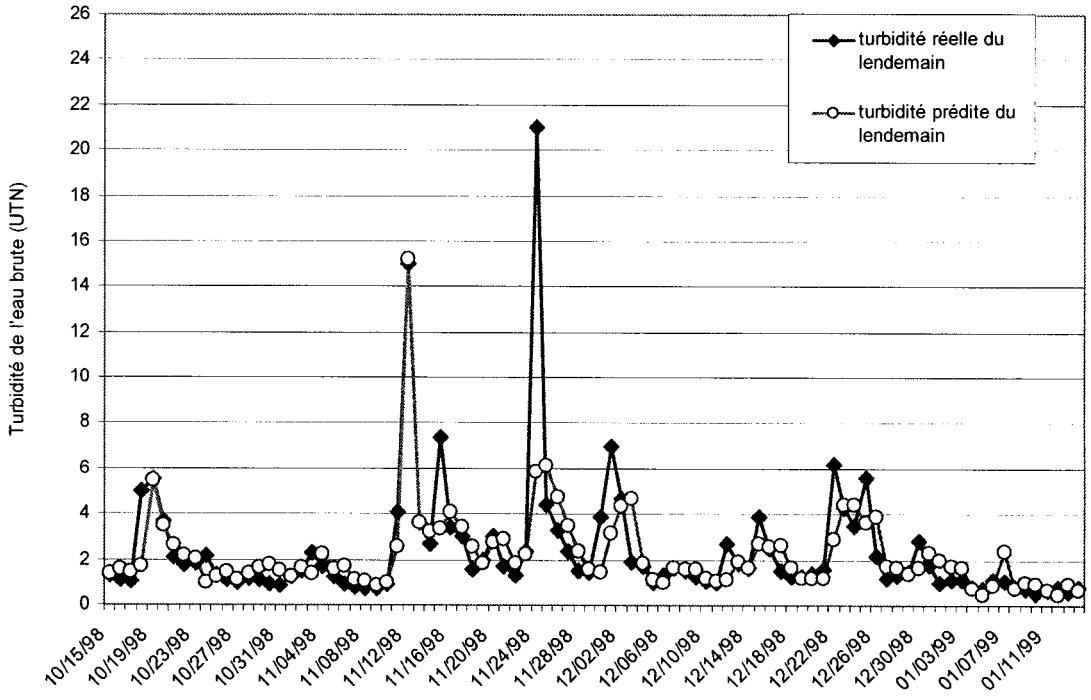


Figure 3-2 : Prévisions du modèle de régression GRNN 8:740:12:1 pour l'automne 1998

Le choix du modèle prévoyant la classe de qualité de l'eau du lendemain (EAU_1) est plus difficile étant donné que deux modèles candidats ont obtenu la même note de performance globale. Il s'agit des deux modèles de type PMC avec six variables en entrée chacun se distinguant l'un de l'autre seulement par le sous-ensemble de variables d'entrée utilisée. Le modèle PMC 6 :5 :1 est finalement retenu car il exclut la turbidité de l'eau brute à Hawkesbury, une variable d'entrée plus difficile à obtenir que les

variables de qualité de l'eau à la station DesBaillets, au profit du débit de la rivière Outaouais (OUT_LAG1), une variable plus facile à obtenir. De plus, le modèle retenu utilise en variables d'entrée un sous-ensemble des variables d'entrée du modèle de prévision de DIFF_1, ce qui est souhaitable et favorisé. La Figure 3-3 donne un aperçu de la performance du modèle de classification. Toutes les représentations graphiques des classes de turbidité réelles et prévues par le modèle de classification sont présentées à l'Annexe 12.

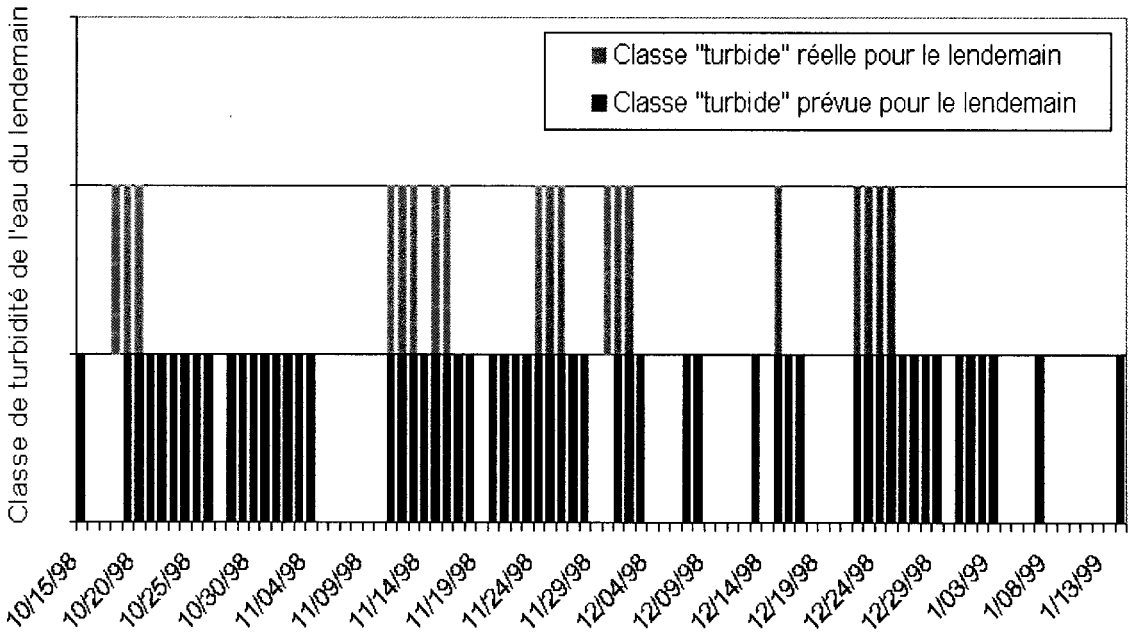


Figure 3-3 : Prévisions du modèle de classification PMC 6 :5 :1 pour l'automne 1998

3.5 Intégration des modèles de prévision à un modèle opérationnel

Le modèle opérationnel combine les résultats des deux premiers types de modèle afin d'accroître la robustesse de la prévision. Cette prévision très simple s'appuie sur un code de couleurs, où « VERT » signifie qu'aucune augmentation significative de la turbidité n'est anticipée pour le lendemain, « JAUNE » signifie qu'une augmentation est probable alors que « ROUGE » signifie qu'une augmentation significative est prévue

pour le lendemain. Le Tableau 3-7 décrit comment le code de couleur est attribué selon les résultats des deux modèles de réseaux de neurones retenus.

Tableau 3-7: Fonctionnement des prévisions du modèle fonctionnel

Valeur de la turbidité du lendemain	Classe prévue de turbidité du lendemain	Prévision intégrée du modèle opérationnel	Signification de la prévision
< 3,2 UTN	Eau claire	VERT	Pas d'augmentation significative
< 3,2 UTN	Eau turbide	JAUNE	Augmentation significative probable
> 3,2 UTN	Eau claire	JAUNE	Augmentation significative probable
> 3,2 UTN	Eau turbide	ROUGE	Augmentation significative prévue

La performance du modèle fonctionnel est très bonne (Tableau 3-8). On dénote une forte majorité de prévisions correctes et très peu de prévisions faussement négatives, i.e. où la prévision indique « VERT » alors qu'elle devrait être « ROUGE ». Si on considère qu'une prévision « JAUNE » au lieu de « VERT » est aussi considérée comme correcte, la proportion de prévisions correctes passe de 83,2% à 95,9%. Le modèle fonctionnel peut donc être considéré comme un outil utile et surtout, fiable.

Tableau 3-8 : Performance du modèle fonctionnel

Statut de la prévision intégrée	Prévision intégrée	Pour l'année complète		Pour les événements turbides seulement	
		% par catégorie	% total	% par catégorie	% total
Correcte	Réel =Vert Prévu =Vert	74,4%	83,2%	0%	94,7%
	Réel = Rouge Prévu =Rouge	7,1%		77,9%	
	Réel = Rouge Prévu= Jaune	1,7%		16,8%	
Faux positif	Réel = Vert Prévu = Jaune	12,7%	15,3%	0%	0%
	Réel = Vert Prévu= Rouge	2,6%		0%	
Faux négatif	Réel=Rouge Prévu =Vert	1,5%	1,5%	5,3%	5,3%

Le Tableau 3-9 compare les taux de prévisions correctes et incorrectes durant les événements turbides pour les modèles fonctionnel, de régression et de classification. Pour le modèle de régression, une prévision est considérée comme correcte si les valeurs turbidité réelle et prévue excèdent toutes deux 3,2 UTN. Lorsqu'on compare les prévisions correctes et incorrectes pour les trois modèles, on constate que la performance combinée des deux modèles est meilleure que celle des deux modèles individuels, particulièrement dans le cas du modèle de régression. Ainsi, le modèle fonctionnel améliore la fiabilité des prévisions effectuées. Son usage est donc souhaitable en station de traitement. La prévision par code de couleurs du modèle fonctionnel peut être couplée de la prévision numérique de turbidité du modèle de régression, donnée à titre indicatif.

Tableau 3-9: Comparaison des prévisions pour les événements turbides des modèles de régression, de classification et fonctionnel

Statut de la prévision	Modèle de régression	Modèle de classification	Modèle fonctionnel
Correcte	80%	92,6%	94,7%
Incorrecte	20%	7,4%	5,3%

CHAPITRE 4 - DISCUSSION

Ce projet de maîtrise a été mis en place afin de comprendre et de prévoir les augmentations de turbidité à la prise d'eau brute de la station de traitement DesBaillets de la Ville de Montréal. Pour ce faire, une démarche méthodologique théorique axée sur la fonctionnalité et la simplicité d'utilisation a été élaborée à partir d'une revue de littérature approfondie. Cette première section évalue donc la pertinence de la démarche méthodologique théorique proposée en fonction de son application à un cas réel. Les résultats obtenus par la mise en application de la méthodologie sont discutés dans la seconde partie du chapitre et portent sur les connaissances acquises au sujet des augmentations de turbidité et leur prévision à l'aide des réseaux de neurones artificiels. Les perspectives découlant du projet réalisé sont exposées dans la troisième et dernière section du chapitre.

4.1 Évaluation de la méthodologie théorique appliquée à un cas réel

Baxter *et al.* (2001) ont souligné qu'il n'existait encore aucun protocole généralement accepté pour la construction de modèles connexionnistes. En fait, la plupart des scientifiques et ingénieurs considèrent les réseaux de neurones artificiels (RNA) comme des boîtes noires. Cette vision utilitaire des RNA permet de développer des modèles connexionnistes qui peuvent bien s'acquitter de leur rôle. Mais du point de vue de la recherche, ce manque de rigueur au niveau de la méthodologie soulève des doutes quant à l'optimalité et même à la fiabilité des résultats obtenus (Maier et Dandy, 2000b).

La démarche méthodologique proposée dans ce projet, élaborée à partir des approches suggérées par des équipes de chercheurs oeuvrant dans le domaine de l'environnement et des ressources hydriques et mise en application concrètement afin de prévoir la turbidité de l'eau brute de la Ville de Montréal, se veut être un point de départ pour tous les utilisateurs des RNA voulant obtenir des résultats dont la fiabilité et l'optimalité sont moins discutables. La démarche méthodologique se veut aussi un outil novateur pour permettre aux chercheurs non familiers avec le fonctionnement des RNA de bien

modéliser un phénomène complexe et non-linéaire à l'aide de cette technologie. Dans cette optique, cette section examine et discute des différentes composantes de la méthodologie en fonction de leur application dans le présent projet et les applications exposées dans la revue de littérature.

4.1.1 Importance de bien connaître le phénomène à modéliser

L'importance de posséder une bonne connaissance *a priori* du phénomène à modéliser avait été soulignée par Maier et Dandy (2000b), qui rapportaient son usage fréquent par les modélisateurs pour choisir les variables d'entrée. C'est pourquoi la première partie de ce projet de maîtrise était consacrée à accroître la compréhension jusque-là limitée des augmentations de turbidité observées à l'eau brute à la Ville de Montréal et des facteurs explicatifs en lien avec ces dernières. La pertinence de ces travaux a été démontrée, car sans une bonne connaissance du phénomène à modéliser, les étapes de modélisation suivantes n'auraient pu être réalisées avec succès:

- Constituer la base de données rassemblant les variables présentant des liens avec la turbidité et dont les données étaient bien représentatives des situations rencontrées. Sans connaissance *a priori*, il aurait été impossible de déterminer quelles variables pouvaient être corrélées ou avoir un lien de cause à effet avec la turbidité à l'eau brute de la Ville de Montréal.

- Organiser la base de données en sous-ensembles d'apprentissage, de test et de validation. Si les caractéristiques distinctes des événements turbides du printemps et de l'automne avaient été ignorées, les données auraient été distribuées de façon aléatoire entre les trois sous-ensembles selon le ratio pré-déterminé. Selon les essais de modélisation préliminaires, cette façon de procéder n'aurait pas résulté en des modèles performants car les données disponibles pour l'apprentissage ne véhiculent pas une information juste du phénomène à modéliser.

- Sélectionner les variables d'entrée des modèles connexionnistes candidats. Comme le proposent Bowden *et al.* (2000a), la connaissance experte du système à modéliser

a été mise à profit pour réduire le nombre de variables d'entrée et choisir le décalage maximal des variables sélectionnées. Ainsi, les liens établis entre les facteurs explicatifs et les événements turbides ont permis d'abord de cerner les variables les plus pertinentes et ensuite d'identifier le décalage temporel existant entre ces dernières et la turbidité. Des variables d'entrée utilisées pour développer les modèles de prévision retenus, au moins la moitié sont des variables de décalage. C'est le cas de la variable de débit de l'Outaouais (OUT_LAG1) et les variables de vent (LSF1_VITM, LSF1_VITX, DOR1_VITM et DOR1_VITX), dont les données présentent un décalage temporel d'une journée par rapport à la variable originale.

4.1.2 Identification des besoins à combler

Le projet a révélé qu'une bonne évaluation des besoins à combler et de la pertinence de recourir à la technologie des RNA permet de développer les modèles les plus appropriés. Si les besoins spécifiques des opérateurs n'avaient pas été considérés dans ce projet, seul un modèle de prévision des valeurs numériques de turbidité du lendemain aurait été développé. Le modèle de prévision de la turbidité selon la classe de qualité de l'eau et le modèle fonctionnel intégrant les prévisions numériques et les prévisions de classe n'auraient pas été développés, privant ainsi les opérateurs d'un outil robuste et fiable pour améliorer le traitement de l'eau potable, l'objectif sous-jacent de ce projet de maîtrise.

4.1.3 Choix du critère de performance

Cette étape s'appuie sur le jugement du modélisateur, car bien que certains critères soient utilisés couramment, rien n'oblige le modélisateur à les adopter. C'est ce qui s'est produit pour le présent projet, où le coefficient de corrélation a été retenu plutôt que l'erreur quadratique moyenne de l'ensemble de test (RMSE), la mesure la plus commune pour les modèles de régression. Des essais de modélisation préliminaires peuvent venir appuyer la décision du modélisateur, mais c'est principalement

l'interprétation subjective des besoins à combler par le modèle qui guide le choix du critère de performance.

Le critère de performance s'est révélé être un outil important, mais dont l'utilité est limitée. En effet, le critère de performance convient bien pour effectuer un tri préliminaire parmi les différents modèles développés. Mais, une fois les modèles candidats potentiels identifiés, la comparaison des valeurs prévues et historiques représentées graphiquement s'est avérée essentielle pour repérer les décalages entre les deux et pour évaluer la justesse de la prévision lors des premiers jours des événements turbides. L'élaboration d'une équation de performance globale qui intègre des critères comme le nombre et l'accessibilité des variables d'entrée et la justesse des prévisions effectuées est aussi un autre outil important sans lequel il aurait été difficile d'identifier les meilleurs modèles parmi l'ensemble des modèles candidats.

4.1.4 Sélection des variables d'entrée

Les travaux effectués dans le cadre du projet appuient Maier et Dandy (1997) et Bowden *et al.* (2000a) qui énoncent que cette étape représente la tâche la plus difficile et la plus importante dans le développement de modèles de prévision. En effet, sélectionner les variables est difficile en raison des nombreuses variables disponibles, y compris les variables de décalage temporel. Mais lorsque bien effectuée, une sélection pertinente de variables d'entrée permet de développer facilement des modèles de prévision performants.

Comme le mentionne Baxter *et al.* (2002), on pourrait utiliser au départ toutes les variables disponibles car les variables qui s'avèrent redondantes ou peu importantes sont retirées au cours des essais de modélisation subséquents. Mais comme le souligne Bowden *et al.* (2000a), présenter un grand nombre de variables d'entrée aux modèles RNA requiert généralement des réseaux de grande taille, ce qui a pour conséquence de nécessiter plus de données et de réduire la vitesse d'apprentissage. De plus, les essais de modélisation préliminaires réalisés dans le cadre de ce projet ont indiqué que les

modèles souffraient de « sur-apprentissage » lorsque le nombre de variables en entrée était trop important comparativement au nombre de données disponibles, soit à partir d'un ratio de 25 variables pour trois ans et quatre mois de données journalières dans le cas présent. Réduire le nombre de variables en entrée se révèle donc être essentiel à la bonne poursuite du projet, mais il est encore plus essentiel de conserver les bonnes variables d'entrée, capables de permettre aux modèles de prévision d'être performants.

4.1.4.1 Réduction du nombre de variables par la connaissance *a priori*

La meilleure compréhension des augmentations de turbidité et des causes explicatives en lien avec ces dernières a joué un rôle prépondérant dans la réduction du nombre de variables d'entrée potentielles. En cernant les variables les plus pertinentes et la valeur du décalage temporel entre ces dernières et la turbidité, le nombre de variables d'entrée potentielles est passé de 43 à 17, ou de 125 à 17 variables, si on inclut les variables de décalage potentielles. L'identification des facteurs explicatifs en lien avec les augmentations de turbidité a aussi mené au développement de paramètres d'index. Inspirés de l'étude de Zhang et Stanley (1997), qui ont eu recours à un paramètre de saison dans leur modèle de prévision de la couleur de l'eau brute, les paramètres d'index du projet ont permis de bien représenter certains facteurs explicatifs comme le renversement, la période de gel et de la fonte des neiges par une seule variable.

4.1.4.2 Identification des variables significatives par des techniques analytiques

En réponse à Bowden *et al.* (2000a) qui soulèvent des doutes quant à l'optimalité des ensembles de variables d'entrées développés à partir de la seule connaissance *a priori* des chercheurs, Maier et Dandy (2000a) proposaient d'utiliser des techniques analytiques pour sélectionner les variables d'entrée. Pour évaluer la pertinence de recourir à ces techniques, qui sont quand même plus complexes que la sélection selon la connaissance des chercheurs, ce projet a mis à l'épreuve une technique non-supervisée permettant de réduire le nombre de variables (analyse en composantes principales ou ACP) et trois techniques supervisées (algorithmes génétiques, construction par étapes et

design expérimental de Plackett-Burman) permettant d'identifier les variables les plus significatives selon leur impact sur l'habileté de généralisation d'un RNA. Mais plutôt que de confronter ces techniques analytiques à la sélection par connaissance *a priori* comme le suggèrent Maier et Dandy (2000a), elles ont été utilisées de façon complémentaire, en utilisant l'ensemble déjà réduit de 17 variables d'entrée plutôt que l'ensemble initial de 125 variables. La raison de cette décision est d'ordre pratique et peut s'appliquer à plusieurs utilisateurs potentiels disposant d'une plage de données limitée: dans le cas présent, le nombre de données disponibles était insuffisant par rapport au nombre de variables à évaluer si elles étaient toutes considérées.

À la lumière des résultats obtenus, il n'est pas pertinent de recourir à la technique des ACP quand le nombre de variables à évaluer est relativement peu élevé. Par contre, les techniques analytiques supervisées se sont révélées plus utiles. Mais le fait que chacune identifie des variables différentes peut être confondant pour le modélisateur. Ainsi, plutôt que d'éliminer les variables les moins significatives, les techniques ont servi surtout à élaborer les sous-ensembles ayant servi ensuite à élaborer les modèles candidats. Cette façon de faire a été fructueuse. Elle a aussi révélé que l'ensemble complet des variables sélectionnées par la connaissance *a priori* n'est pas celui ayant mené au développement des meilleurs modèles candidats, mais bien un sous-ensemble identifié à partir des résultats combinés des trois méthodes analytiques supervisées.

À prime abord, il est difficile de déterminer quelle méthode est la plus efficace et de n'en recommander qu'une seule. En fait, la comparaison des résultats obtenus par différentes techniques est plus instructive que les résultats seuls. Néanmoins, si un choix devait être effectué à partir des observations effectuées dans le présent projet, la technique des algorithmes génétiques et de la construction par étape pourraient être utilisées alternativement car les résultats obtenus sont semblables. Quant à la technique du design expérimental Plackett-Burman, elle a révélé que les résultats obtenus variaient selon la variable de sortie examinée et aussi selon les données utilisées pour l'analyse, soit celles des événements turbides exclusivement ou celles de toute l'année.

Dans le cas présent, il valait mieux utiliser les données se référant exclusivement au phénomène d'intérêt que les données pour l'ensemble de l'année car la sélection de variables significatives était plus consistante. Dans le cas d'autres applications, il serait difficile de recommander l'usage unique de cette technique. En effet, il vaut mieux l'utiliser comme un complément utile pour identifier les variables significatives et leur importance relative.

4.1.5 Sélection des variables de sortie

L'usage éventuel du modèle de prévision, déterminé à partir des besoins à combler, a été l'élément-clé de la sélection des deux variables de sortie utilisées dans ce projet, appuyant les propos de Baxter *et al.* (2002) voulant que la variable de sortie soit sélectionnée sur la base de l'usage éventuel du modèle, la littérature scientifique et la disponibilité des données. Ainsi, pour les besoins de la recherche voulant tester le potentiel des RNA pour la prévision, la première variable de sortie retenue est une valeur numérique (DIFF_1) alors que pour combler les besoins pratiques des opérateurs de la station de traitement, la seconde variable de sortie est une variable de classe générale du niveau de la turbidité de l'eau (EAU_1).

Il est intéressant de noter que les deux variables de sortie portent sur la turbidité de l'eau brute, mais qu'aucune n'utilise directement les valeurs réelles de turbidité. La variable de sortie numérique est la différence entre la valeur de turbidité d'aujourd'hui et de demain. Cette façon de procéder a été adoptée suite aux suggestions de Zhang et Stanley (1997), qui la proposaient dans leur étude lorsque l'analyse des données originales révèle que la variable de sortie est auto-corrélée. Dans ce projet, les résultats des essais de modélisation préliminaires ont aussi démontré sans équivoque la supériorité de la variable différentielle sur la variable de turbidité originale. Il est donc recommandable d'utiliser la différence entre la valeur d'aujourd'hui et de demain comme variable de sortie dans toutes les applications visant à prévoir des paramètres comme la couleur, la turbidité de l'eau brute ou des paramètres corrélés avec ces dernières.

4.1.6 Développement des modèles candidats

Ce projet a révélé que la facilité avec laquelle les modèles candidats sont développés dépend grandement du soin accordé aux étapes préalables de sélection de variables d'entrée et de sortie. Dans le cas présent, la constitution de sous-ensembles de variables d'entrée pertinents à partir de l'ensemble restreint de départ a permis de développer de façon systématique différents modèles candidats. Cette façon de procéder permet à l'utilisateur de RNA d'éviter d'être confondu par les multiples possibilités découlant des combinaisons possibles de variables d'entrée et d'architectures neuronales qui s'offrent à lui. De même, l'usage de variables de sortie choisies avec discernement permet d'éviter plusieurs écueils, comme celui d'obtenir un décalage entre les valeurs réelles et les valeurs prévues si on omet de considérer l'autocorrélation de la variable de sortie.

La méthodologie théorique propose de s'appuyer sur un logiciel commercial pour la recherche heuristique des meilleures architectures. Cette approche méthodologique s'est révélée adéquate et efficace dans le cadre de ce projet. Après avoir choisi de ne développer que des RNA de type « *feedforward* » suite à la forte recommandation de Maier et Dandy (2000b) et d'utiliser comme critère d'arrêt la validation croisée en raison de la subdivision des données en trois sous-ensembles, l'outil automatisé de construction de RNA de Neural Networks, désigné « *Intelligent Problem Solver* » (IPS), a permis de développer facilement de nombreux modèles dont plusieurs se sont révélés être très intéressants.

4.1.7 Évaluation des modèles candidats

Cette dernière étape a été menée avec succès en raison de l'application systématique et rigoureuse de la méthodologie théorique proposée et de l'usage d'un critère de performance permettant d'évaluer adéquatement les modèles candidats. Une des composantes importantes, mais souvent négligée, de la méthodologie de l'évaluation des modèles est la redistribution différente des données entre les trois sous-ensembles

d'apprentissage, de test et de validation. Or, dans ce projet, l'utilisation de deux ensembles de données réparties différemment a permis de s'assurer de la stabilité des modèles candidats (Baxter *et al.*, 2002) et conséquemment, d'accroître la confiance dans les prévisions des modèles retenus dont la performance de prévision s'est révélée indépendante de la distribution des données en trois sous-ensembles.

Une particularité intéressante du processus d'évaluation des modèles candidats a été de considérer la facilité d'obtention des variables d'entrée utilisées pour développer les modèles. Cette considération est d'ordre purement pratique et démontre que le modélisateur doit toujours garder à l'esprit l'usage auquel est destiné le modèle. Dans ce cas-ci, l'utilisation en parallèle de deux modèles distincts implantés en ligne à la station de traitement DesBaillets amène le modélisateur à préférer les modèles utilisant les variables générées à la station même aux données d'autres stations en amont, comme les variables de qualité à l'usine de filtration de Hawkesbury, par exemple.

4.1.8 Commentaires sur les connaissances requises du fonctionnement des RNA et l'utilisation de logiciels commerciaux

Les informations présentées dans la revue de littérature au sujet des caractéristiques générales et des concepts de base concernant le fonctionnement du neurone artificiel, l'architecture des RNA et leur processus d'apprentissage se sont révélées satisfaisantes pour développer des modèles à l'aide d'un logiciel commercial. Évidemment, sans logiciel, une connaissance approfondie du fonctionnement des RNA aurait été nécessaire pour construire et sélectionner les meilleures architectures, mais cette approche n'aurait pas nécessairement résulté en de meilleurs modèles connexionnistes que ceux construits à l'aide d'un logiciel commercial. Ceci ne veut pas dire néanmoins qu'on doit s'appuyer aveuglément sur un logiciel commercial pour développer des modèles performants. À la limite, la recherche heuristique des meilleures architectures peut être traitée comme une « boîte noire », mais pas la sélection des variables d'entrée et de sortie et du critère de performance, pour lesquels la connaissance *a priori* du phénomène à modéliser et le jugement du modélisateur sont essentiels.

L'approche méthodologique employée dans ce projet s'inspire de la philosophie de l'équipe de Maier, Dandy et Bowden (Maier et Dandy, 1997, 2000a, 2000b, Bowden *et al.*, 2000a, 2000b, 2001, 2002) face au développement de réseaux de neurones artificiels pour prévoir des phénomènes environnementaux complexes : il vaut mieux commencer simplement et raffiner les modèles par la suite selon les problèmes rencontrés lors du développement du modèle initial car les façons d'améliorer un modèle sont nombreuses mais elles sont souvent spécifiques à la situation examinée. C'est alors qu'il peut devenir souhaitable de délaisser l'assistance offerte par les logiciels commerciaux pour explorer les concepts avancés du fonctionnement des réseaux de neurones artificiels.

4.2 Discussion des résultats du projet de recherche

Cette section du chapitre discute des résultats obtenus en appliquant la méthodologie discutée précédemment. Les résultats discutés portent sur les grands volets de ce projet de maîtrise, soit l'acquisition d'une meilleure compréhension des hausses de turbidité et des causes à leur origine et le développement de modèles de prévision capable d'anticiper les hausses de turbidité un jour à l'avance.

4.2.1 Meilleure compréhension des hausses significatives de turbidité et des causes à leur origine

Les causes exactes à l'origine des augmentations de turbidité observées à la prise d'eau de la Ville de Montréal étaient peu connues avant la réalisation de ce projet de maîtrise. Traditionnellement, les explications liées aux augmentations de turbidité se résumaient à associer l'augmentation printanière au phénomène de l'inversion thermique, aussi dénommé renversement. Quant aux augmentations de l'automne, on présumait aussi que le renversement était à l'origine de l'une d'entre elles, mais aucune hypothèse n'était vraiment élaborée au sujet des causes de l'ensemble de ces hausses, dont le nombre varie entre six et dix par années pour les périodes automnales de 1998 à 2000. En raison du peu d'informations disponibles mais de la nécessité d'en acquérir afin de développer des modèles de prévision, une revue de littérature s'imposait pour identifier

les facteurs explicatifs potentiellement liés aux augmentations de turbidité printanières et automnales. Ensuite, une analyse quantitative de données de turbidité représentatives et de différentes variables en lien avec la turbidité a permis de cerner parmi les facteurs explicatifs potentiels lesquels jouaient un rôle prépondérant selon la saison.

Un premier constat d'intérêt dégagé du portrait dressé de la turbidité de l'eau brute à la Ville de Montréal est la différence de durée des événements turbides du printemps et de l'automne, les deux périodes d'intérêt à l'étude dans ce projet. Au printemps, on retrouve toujours un événement de fond (plus de cinq jours) auquel se superposent des événements de courte durée (moins de cinq jours). À ceci s'ajoutent parfois quelques événements de courte durée isolés de l'événement de fond, mais cela semble plutôt rare. Par contre, au cours de l'automne, il n'y a aucun événement turbide de fond. Il n'y a que des événements courts, isolés les uns des autres et dont l'intensité semble aléatoire.

Ces différences entre les événements turbides printaniers et automnaux laissaient déjà présager que des causes explicatives différentes sont à l'origine des événements turbides de fond et des événements courts isolés et soulevaient aussi la question à savoir si des facteurs explicatifs communs étaient à l'origine des événements courts isolés de l'automne, des événements courts isolés du printemps et des événements courts superposés à un événement de fond. Il s'avère que l'augmentation de débit de la rivière des Outaouais joue le plus grand rôle au printemps et les tempêtes de vent, à l'automne. Au printemps, tous les événements turbides de fond sont causés par la hausse graduelle du débit de la rivière des Outaouais suite à la fonte des neiges. Les événements courts se superposant aux événements de fond du printemps ont des causes plus diverses, mais il ressort nettement que la plupart sont aussi liés aux fluctuations de débit : les pointes soudaines de turbidité coïncident souvent avec les pointes de débit de la rivière des Outaouais ou des tributaires secondaires, soit seules ou en synergie avec d'autres événements turbides comme les tempêtes de vents avec ou sans précipitations ou le renversement. À l'automne, les tempêtes de vent expliquent la majorité des hausses très rapides de la turbidité. Mais il est important de noter que dès l'apparition du couvert de

glace, elles n'ont plus d'effet. L'analyse des représentations graphiques a aussi permis de déduire que les autres causes explicatives potentielles suggérées par la littérature, soit les fortes précipitations et le renversement, jouent un rôle secondaire. L'occurrence des fortes précipitations non combinée à d'autres causes explicatives ne se traduit jamais en hausse significative de turbidité. Quant au renversement, son occurrence résulte parfois en un événement turbide court à l'automne ou superposé à l'événement turbide de fond, au printemps, mais jamais à l'événement de fond printanier, comme on le présupposait avant la réalisation de ce projet de recherche. Finalement, le fait que deux événements turbides de courte durée demeurent inexpliqués sur les 36 répertoriés durant la période d'étude soulève l'hypothèse que d'autres facteurs explicatifs d'occurrence plus rare n'aient pas été pris en compte.

Il est important de noter que le laps de temps étudié pour acquérir une bonne compréhension des hausses de turbidité et de leurs causes est relativement court, soit trois ans et quatre mois. Cette limitation venait de l'ampleur du travail nécessaire pour construire une base de données rassemblant de nombreuses variables à l'accès et à la disponibilité parfois limités. Néanmoins, les données disponibles se sont révélées suffisantes pour dégager les observations nécessaires à la compréhension des augmentations de turbidité, illustrant le fait qu'il est plus important d'avoir des données représentatives de l'échelle de valeurs pouvant être observées pendant une année et d'une année à l'autre que d'avoir des données en grand nombre.

Les discussions informelles avec les responsables de la station DesBaillets et les opérateurs de station de traitement en amont de Montréal se sont révélées être un outil très utile pour confirmer ou infirmer l'importance des différentes causes explicatives et pour choisir les variables les plus représentatives possibles des différents facteurs explicatifs. C'est notamment le cas des forts vents, qui entraînent la formation de vagues érodant les berges et remettant en suspension la fine couche de sédiments susceptibles de se déposer en périodes tranquilles dans les zones peu profondes (Frenette et Frenette, 1992). L'impact néfaste du vent sur la qualité de l'eau a été

souligné par les opérateurs de la station de Coteau-du-Lac, qui traite l'eau du lac St-François. Une certaine dose d'imagination a aussi été requise dans le cas de certains facteurs explicatifs. C'est le cas de la prise du couvert de glace sur les lacs fluviaux, qui a été représentée conjointement par le débit de petites rivières tributaires du lac St-François, la température de l'air et les précipitations locales.

4.2.2 Développement de modèles de prévision

4.2.2.1 Comparaison de la performance des différents modèles de prévision retenus

Ce projet a permis de développer deux types de modèles connexionnistes différents, soit les modèles de régression avec la variable de sortie DIFF_1 et les modèles de classification avec la variable EAU_1. Le troisième modèle développé, le modèle fonctionnel, n'utilise pas de réseaux de neurones artificiels (RNA) mais intègre les résultats des deux modèles connexionnistes. La comparaison d'un modèle à l'autre est difficile car les critères de performance retenus diffèrent. Pour les modèles de régression, ce sont les coefficients de corrélation qui sont scrutés alors que pour les modèles de classification et les modèles fonctionnels, c'est le taux de classification correcte qui est examiné. De plus, il est difficile de comparer un modèle prévoyant des valeurs numériques à un modèle prévoyant des classes. Malgré ces difficultés, les trois modèles ont été comparés selon leur proportion de prévisions correctes et incorrectes pendant les événements turbides, i.e. lorsque la turbidité excède 3,2 UTN pendant les périodes printanières ou automnales. Le modèle fonctionnel s'est révélé supérieur avec une proportion de prévision correcte de 94,7%. Il était suivi de près par le modèle de classification, qui affichait un taux de succès de 92,6%. La performance du modèle de régression était moindre avec un taux de prévision correcte de 80%. La moins bonne performance du modèle de régression peut s'expliquer par le fait qu'il est plus complexe de prévoir des valeurs numériques pour lesquelles les possibilités de sortie sont pratiquement infinies que de prévoir une classe pour laquelle seulement deux valeurs de sortie sont possibles.

Lorsqu'on compare les modèles entre eux, il faut aussi prendre en compte leurs avantages et inconvénients propres. L'utilité du modèle de régression va notamment au-delà de la simple prévision : en fournissant une valeur numérique plutôt qu'une classe de danger potentiel, il permet d'établir une tendance dans le temps comme une hausse graduelle et progressive au printemps, ce que ne permet pas de faire le modèle de classification. Le modèle de classification, quant à lui, nécessite des réseaux de neurones à l'architecture plus simple que les modèles de régression, probablement parce que la variable de sortie de classe est plus facile à modéliser. Par contre, plusieurs prévisions faussement positives sont observées entre les événements turbides. Ceci peut conduire les opérateurs, mis trop souvent sur un pied d'alerte alors que la qualité de l'eau reste bonne, à perdre confiance dans la justesse des prévisions effectuées par le modèle, même si ce dernier prévoit correctement les hausses de turbidité en tant que telles.

La grande majorité des modèles de prévision examinés dans la revue de littérature sont des modèles de régression. Seule l'équipe de Brion et Lingirreddy (1999), à laquelle s'ajoute ensuite celle de Neelakantan (Neelakantan *et al.,* 2001), utilisent des classes comme valeurs de sortie pour exprimer l'origine de la contamination fécale dans un cas et les concentrations de pointe d'oocystes de *Cryptosporidium* et de kystes de *Giardia* dans l'autre cas. Aucune application signalant l'usage simultané de deux modèles distincts de prévision, notamment par un modèle fonctionnel comme c'est le cas dans ce mémoire, n'a été relevée. Néanmoins, les résultats très encourageants obtenus dans ce projet amènent à penser que les modèles de classification et les modèles fonctionnels devraient être plus utilisés, particulièrement dans les cas où les besoins à combler touchent plus la prévision d'un risque que la prévision d'une valeur numérique exacte. De plus, l'usage d'un modèle fonctionnel, qui améliore la fiabilité et la justesse des prévisions, s'est révélé être une façon relativement simple de gagner la confiance des opérateurs de station de traitement.

4.2.2.2 Comparaison des caractéristiques des différents modèles candidats

Pour chaque sous-ensemble de variables d'entrée, les deux ou trois meilleurs modèles développés par l'outil automatisé du logiciel étaient examinés [1]. Comme leur architecture différait souvent l'un de l'autre et d'un sous-ensemble de variables d'entrée à un autre, des observations intéressantes en ont été retirées. Pour les modèles de régression, on retrouve des modèles de trois architectures différentes, soit les perceptrons multicouches (PMC), les réseaux fonction radiale de base (*Radial basis function* ou RBF) et les réseaux de neurones de régression généralisée (*General Regression Neural Networks* ou GRNN) alors que pour les modèles de classification, on ne retrouve que des réseaux des deux premières architectures.

Les réseaux de type GRNN présentent des résultats intéressants. Tout d'abord, le simple fait d'utiliser ces réseaux suscite de l'intérêt car ils ne sont pas fréquemment utilisés dans le domaine de l'environnement et des ressources hydriques, où les modèles de type PMC dominent largement (Maier *et al.*, 2001). Ensuite, la très bonne performance de quelques modèles candidats, qui en a mené un à être choisi comme modèle final, suggère une nouvelle avenue dans la prévision de valeurs numériques. Cette amélioration de la performance s'explique peut-être par l'utilisation de deux couches cachées et d'un grand nombre de neurones dans la première couche, ce qui permet de modéliser plus facilement un phénomène complexe. Par contre, le recours à grand nombre de neurones peut aussi susciter des problèmes de surapprentissage, identifiés par une différence significative entre l'erreur d'apprentissage et l'erreur de validation. Plusieurs modèles de type GRNN ont d'ailleurs présenté ces problèmes. Il est intéressant de noter que le nombre de neurones dans la couche cachée excédait toujours le nombre maximal recommandé par Roger et Dowla (1994) pour éviter le

[1] Tous les modèles candidats potentiels sont présentés aux Annexes 9 et 10.

surapprentissage[2], selon lesquels le nombre de neurones dans la couche cachée aurait dû être compris entre 41 dans le cas où les 17 variables d'entrée étaient utilisées et 183 avec seulement 3 variables d'entrée plutôt que les 740 neurones actuellement utilisés avec les modèles de type GRNN.

Quant à elle, la performance des modèles à l'architecture de type PMC varie beaucoup, tant dans le cas des modèles de régression que des modèles de classification. Il est difficile de tirer des conclusions valables de la comparaison des différents modèles PMC car le nombre de modèles candidats est limité et les variables d'entrée varient simultanément en nombre et en type. Néanmoins, il est possible de constater qu'en général, la performance des modèles est bonne lorsque le nombre de neurones dans la couche cachée est entre le double ou le triple du nombre de variables en entrée et elle diminue lorsque le nombre de neurones dans la couche cachée est plus petit que le nombre de variables en entrée. Ces observations concordent avec celles de Wilson et Recknagel (2001) qui concluaient, après avoir exploré dans leur essais de modélisation l'impact du nombre de neurones dans la couche cachée, que la présence des neurones dans la couche cachée améliore la performance prédictive des modèles. Elles corroborent aussi les recommandations de Maier et Dandy (2000b, 2001) selon lesquels la limite supérieure permettant de s'assurer que les réseaux de neurones soient capables d'approximer n'importe quelle fonction est égale au double du nombre de variables en entrée plus un neurone[3]. L'architecture RBF est la moins performante de toutes. Dans le cas des modèles de régression, les représentations graphiques de ces modèles

---

[2] Ces recommandations correspondent à l'équation 4 et sont expliquées à la section I.2.4.4 sous la rubrique «sélection des caractéristiques de l'architecture du réseau »

présentent, lors des périodes critiques, des oscillations importantes dans les valeurs prévues. Dans le cas des modèles de classification, les réseaux RBF avaient un excellent taux de classification correcte pour l'ensemble des données et pour les événements turbides. Mais ils performaient nettement moins bien quand venait le temps de prévoir correctement le début des événements turbides, ce qui est essentiel.

4.3 Perspectives

Les modèles de prévision ont démontré le potentiel des réseaux de neurones artificiels pour la prévision de la turbidité de l'eau brute à la Ville de Montréal et leur développement a permis de mettre en application la méthodologie théorique élaborée dans ce projet. Mais il reste encore beaucoup à faire, car les modèles actuels constituent seulement un point de départ démontrant qu'il est possible de prévoir la qualité de l'eau à l'aide de réseaux de neurones artificiels. Cette section présente donc quelques perspectives découlant du projet, soit l'amélioration des modèles actuels, leur implantation en ligne et le développement d'autres modèles prévoyant la qualité de l'eau au sein des différentes filières de traitement de la station DesBaillets.

4.3.1 Amélioration des modèles connexionnistes actuels

Les modèles connexionnistes actuels peuvent être améliorés et même transformés de plusieurs façons afin de prévoir correctement *toutes* les augmentations de turbidité, particulièrement les hausses très soudaines observées à l'automne. Tout d'abord, un réglage de précision (*fine-tuning*) peut être effectué, permettant au modélisateur d'explorer davantage le fonctionnement des réseaux de neurones artificiels et d'en découvrir toute la richesse et le potentiel. Ensuite, des données supplémentaires des variables d'entrée devraient être rassemblées afin de ré-entraîner les modèles de

---

[3] Ces recommandations s'appuient sur les travaux de Hecht-Nielsen (1987) et correspondent à l'équation 3 (voir section I.2.4.4 sous la rubrique « sélection des caractéristiques de l'architecture du réseau »).

régression et de classification actuels. La performance en serait probablement accrue, particulièrement si les données ajoutées comprennent des événements turbides d'intensité exceptionnelle. Il serait alors avantageux de les incorporer aux sous-ensembles de test et de validation. Avec plus de données en main, il serait aussi possible d'explorer la possibilité de développer des modèles sur une base saisonnière. Le modèle automnal pourrait par exemple utiliser seulement les variables en lien avec les facteurs explicatifs à l'origine des événements turbides automnaux comme le vent et le renversement, alors que le modèle printanier utiliserait principalement des variables se rapportant au débit de la rivière des Outaouais et à la fonte des neiges. Cette exploration pourrait permettre d'évaluer si des modèles saisonniers sont plus performants qu'un modèle annuel.

4.3.2 Implantation en ligne des modèles de prévision améliorés

L'implantation en ligne des modèles a été évoquée à quelques reprises dans ce mémoire. Il ne s'agissait pas d'un simple souhait, mais bien d'une éventualité à considérer sérieusement. Pratiquement tous les logiciels commerciaux présentent un module d'interface permettant d'intégrer le modèle à la plateforme de contrôle de la station de traitement. Zhang et Stanley (1997) mentionnent d'ailleurs avoir implanté en ligne à la station de traitement leur modèle de prévision de couleur de l'eau brute. Avant de procéder à l'implantation, il serait tout de même souhaitable d'améliorer les modèles actuels en s'appuyant sur les recommandations suggérées auparavant.

4.3.3 Développement de modèles de prévision de la qualité de l'eau pour les différentes filières de traitement

À l'instar de l'équipe de Zhang et Stanley (1997) qui ont développé plusieurs modèles de prévision de la qualité de l'eau dans différentes filières de traitement, les réseaux de neurones artificiels pourraient être mis avantageusement à profit dans les nouvelles installations de traitement prévues aux stations de traitement de la Ville de Montréal. Les applications possibles incluent la prévision de l'enlèvement de la turbidité lors de la

coagulation (Baxter *et al.*, 2001), la prévision des doses l'alun et de polymères à utiliser lors du procédé de coagulation (Joo *et al.*, 2000, Baxter *et al.,* 2001) et la prévision de performance de la filtration (Baxter *et al.,* 2001).

CHAPITRE 5 - CONCLUSION

Le but de ce projet de maîtrise était de prévoir les augmentations de turbidité de l'eau brute de la Ville de Montréal à l'aide de réseaux de neurones artificiels. Afin d'atteindre ce but, deux réalisations significatives ont été accomplies, soit (1) d'accroître la compréhension jusque-là limitée des augmentations de turbidité observées à l'eau brute à la Ville de Montréal et des facteurs explicatifs en lien avec ces dernières et (2) de proposer une approche méthodologique permettant aux chercheurs non familiers avec le fonctionnement des RNA de bien modéliser un phénomène environnemental complexe et non-linéaire à l'aide de cette technologie.

Les travaux réalisés ont permis de dégager les conclusions suivantes relativement aux connaissances liées aux augmentations de turbidité et à leurs causes :

- Les caractéristiques et les causes explicatives des augmentations de turbidité diffèrent entre les deux périodes critiques du printemps et de l'automne. L'augmentation de débit de la rivière des Outaouais résultant de la fonte des neiges joue le plus grand rôle au printemps et les tempêtes de vent expliquent la majorité des hausses très rapides de la turbidité à l'automne.

- Le phénomène d'inversion thermique, aussi dénommé renversement, ne joue qu'un rôle secondaire tant au printemps qu'à l'automne, contrairement à ce qui était présupposé avant la réalisation de ce projet de recherche.

Les conclusions suivantes ont été tirées suite à la mise en pratique de l'approche méthodologique théorique pour développer les modèles de prévision de la turbidité:

- Toutes les étapes de l'approche méthodologique, soit l'identification des besoins à combler et des ressources disponibles, le choix du critère de performance, le développement et l'organisation de la base de données, la construction des modèles

candidats et le choix du modèle final, sont importantes et doivent être accomplies avec rigueur pour obtenir un modèle performant et adéquat.

- Une connaissance approfondie du phénomène à modéliser est essentielle pour constituer une base de données rassemblant des variables pertinentes et l'organiser adéquatement en sous-ensembles d'apprentissage, de test et de validation et pour réduire le nombre de variables d'entrée des modèles candidats et choisir leur décalage maximal.

- Le modélisateur doit toujours garder à l'esprit l'usage auquel est destiné le modèle afin de développer le modèle de prévision le plus approprié. Les besoins à combler doivent être considérés dans le choix du critère de performance et de la variable de sortie du modèle de prévision.

- L'utilisation d'un logiciel commercial s'avère un choix judicieux pour construire et sélectionner les meilleures architectures de réseaux de neurones artificiels. Il ne faut pas néanmoins s'appuyer aveuglément sur un logiciel commercial pour développer des modèles performants, car la connaissance *a priori* du phénomène à modéliser et le jugement du modélisateur sont essentiels pour sélectionner les variables d'entrée et de sortie et le critère de performance.

- Les façons d'améliorer un modèle sont nombreuses mais elles sont souvent spécifiques à la situation examinée. Il est donc préférable de développer au départ des modèles de prévision simples et de les raffiner par la suite selon les problèmes rencontrés lors du développement du modèle initial.

La Figure 5-1 résume les composantes essentielles au développement d'un modèle de prévision performant et adéquat : une bonne connaissance du phénomène à modéliser, la considération des besoins à combler et l'approche méthodologique rigoureuse

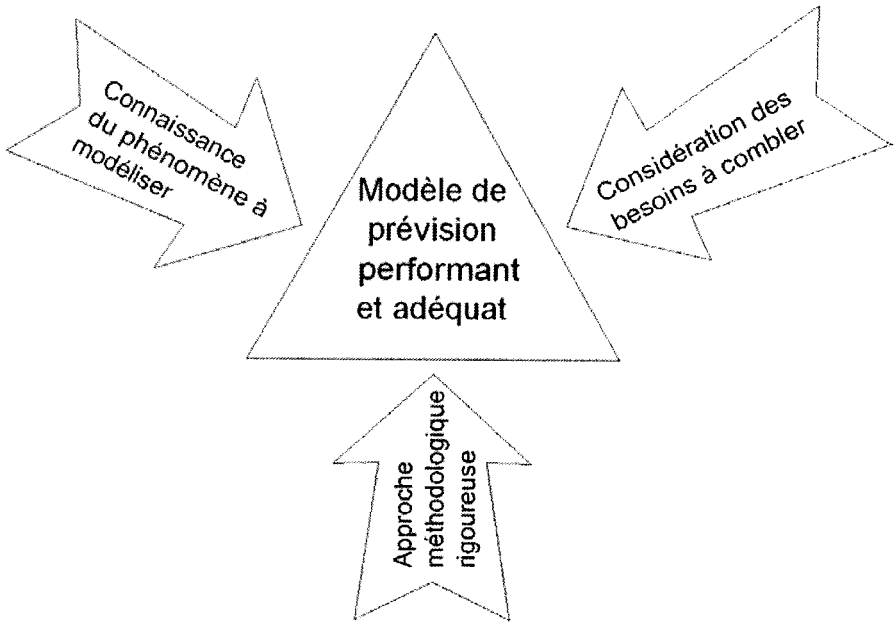


Figure 5-1 : Les ingrédients essentiels au développement de modèles de prévision performants et adéquats

La mise en application de la méthodologie théorique a permis de développer trois modèles de prévision différents dont la performance est satisfaisante. Le premier est un modèle de régression de type GRNN (8:740:2:1) prévoyant la différence entre la turbidité d'aujourd'hui et celle de demain. Le second est un modèle de classification de type PMC (6 :5 :1) prévoyant la classe à laquelle appartient la turbidité du lendemain. Le troisième modèle développé, le modèle fonctionnel, n'utilise pas de réseaux de neurones artificiels (RNA) mais intègre les résultats des deux modèles connexionnistes. Les conclusions suivantes ont été tirées au sujet des modèles de prévision développés :

- La décision de développer plus d'un modèle de prévision connexionniste repose sur les besoins des utilisateurs. Lorsque la fiabilité et la justesse des résultats importent beaucoup comme c'est le cas pour les opérateurs d'une station de traitement, il est judicieux de développer et d'utiliser de façon complémentaire deux modèles prévoyant la turbidité du lendemain différemment.

- Le modèle fonctionnel a deux usages principaux : en intégrant les résultats des deux modèles connexionnistes, il facilite l'interprétation par des utilisateurs et il améliore la robustesses des prévisions par rapport à celles des modèles connexionnistes individuels.

- Il est difficile de comparer la performance de modèles dont les critères de performance diffèrent. Il est néanmoins possible de trouver un critère comparatif adéquat comme la proportion de prévisions correctes et incorrectes par exemple. Selon ce critère comparatif, le modèle fonctionnel s'est révélé supérieur avec une proportion de prévision correcte de 94,7%. Il était suivi de près par le modèle de classification, qui affichait un taux de succès de 92,6%. La performance du modèle de régression était moindre avec un taux de prévision correcte de 80%.

- La comparaison de la performance des modèles doit aussi prendre en compte les avantages et les inconvénients de chacun. Même si la performance du modèle de régression est moindre, il est intéressant car il permet d'établir une tendance dans le temps comme une hausse graduelle et progressive au printemps, il prévoit la magnitude des augmentations de turbidité et il présente peu de prévisions faussement positives comparativement au modèle de classification.

- Les modèles de prévision performants ne requièrent pas de nombreuses variables en entrée mais plutôt des variables d'entrée bien choisies et représentatives du phénomène à modéliser. Les modèles de régression et de classification retenus utilisent respectivement huit et six variables d'entrée représentant la qualité de l'eau brute à la station DesBaillets, les deux causes explicatives principales des augmentations de turbidité, soit le débit de la rivière Outaouais et les vents de la région de Montréal et la saisonnalité à l'aide d'une variable d'index.

Les réseaux neuronaux ont permis de développer des outils de prévision performants et utiles, qui pourraient être éventuellement implantés en ligne à station de traitement DesBaillets de la Ville de Montréal. Grâce à l'approche méthodologique novatrice

proposée, ce projet de recherche ouvre la voie à plusieurs autres applications possibles dans le domaine de la prévision de paramètres de qualité de l'eau brute ou de l'eau traitée dans les différentes filières de traitement. Les modèles de prévision connexionnistes sont des outils relativement nouveaux mais l'industrie de l'eau potable a tout à gagner à les utiliser pour maintenir et même accroître la qualité de l'eau distribuée aux consommateurs.

RÉFÉRENCES

ADGAR A., COX C.S., BÖHME T.J. (2000). Performance improvements at surface water treatment works using ANN-based automation schemes.*Trans IChemE* 78: part A. 1026-1039.

ADELI H. (2001). Neural Networks in Civil Engineering : 1989-2000. *Computer-Aided Civil and Infrastructure Engineering* 16. 126-142

ARAMINI J., MCLEAN M., WILSON J., HOLT J., COPES R., ALLEN B, SEARS W. (2000). Drinking Water Quality and Health Care Utilization for Gastrointestinal Illness in Greater Vancouver. Santé Canada.

BAXTER C.W., STANLEY S.J., ZHANG Q. (1999) Development of a full-scale artificial neural network model for the removal of natural organic matter by enhanced coagulation. *Aqua* 48:4. 129-136.

BAXTER C.W., ZHANG Q., STANLEY S.J., SHARIFF, R., TUPAS, R-R.T., STARK, H.L. (2001). Drinking water and treatment: the use of artificial neural networks. *Canandian Journal of Civil Engineering* 28 (suppl.1). 26-35

BAXTER C.W., STANLEY S.J., ZHANG Q., SMITH D.W. (2002). Developing artificial neural network models of water treatment processes: a guide for utilities. *Journal of Environmental Engineering and.Sciences* 1. 201-211.

BEBIS G., GEORGIOPOULOS M. (1994). Feed-forward neural networks: Why networks size is so important. IEEE Potentials. octobre/novembre. 27-31.

BISHOP C.M. (1995). Neural Networks for Pattern Recognition. Oxford University Press, New York, USA.

BOWDEN G.J., DANDY G.C., MAIER H.R. (2000a). Identification of inputs for artificial neural networks (ANN) based ecological models. 2nd international conference on application of machine learning to ecological modelling. Paper to be published in a book by Springer-Verlag, April 2002.

BOWDEN G.J., DANDY G.C., MAIER H.R. (2000b). Optimal division of data for neural network models in water resources applications, accepted for publication in *Water Resources Research.*

BOWDEN G.J., MAIER H.R. and DANDY G.C. (2002) Optimal division of data for neural network models in water resources applications. *Water Resources Research* 38:2. 1-11.

BOWDEN G.J., DANDY G.C., MAIER H.R. (2001). Data transformation for neural network models in water resources applications, accepted for publication in Journal of *Hydroinformatics.*

BRION G.M., LINGIREDDY S. (1999). A neural network approach to identifying non-point sources of microbial contamination. *Water Research* 33:14. 3099-3106.

BRION G.M., NEELAKANTAN T.E., LINGIREDDY S. (2001). Using Neural Networks to predict peak Cryptosporidium concentrations. *Journal AWWA* Jan. 99-105

CARIGNAN, R., LORAIN, S., LUM, K. (1993).Sediment Dynamics in the Fluvial Lakes of the St.Lawrence River : Accumulation Rates, and Residence Time of Mobile Sediments. Texte soumis à *Geochimica Cosmochimica Acta.*

CENTRE SAINT-LAURENT (1993). Qualité des sédiments et bilan des dragages sur le Saint-Laurent. Document rédigé par Lucie Olivier et Jacques Bérubé. Direction du développement technologique. No. de catalogue En 153-12/1993F

CHAMPOUX, L., SLOTERDIJK (1988). Étude de la qualité des sédiments du lac Saint-Louis 1984-1985. Rapport technique no. 1 : Géochimie et contamination. Environnement Canada, Conservation et Protection, région du Québec.

COULIBALY P., ANCTIL F., BOBÉE B. (1999). Prévision hydrologique par réseaux de neurones artificiels : état de l'art, *Canadian Journal of Civil Engineering* 26. 293-304.

ELSHORBAGY, A., SIMONOVIC, S.P., PANU, U.S., (2000). Performance evaluation of artificial neural networks for runoff prediction. *Journal of Hydrologic Engineering.* October. 424-427.

MINISTÈRE DE L'ENVIRONNEMENT DU QUÉBEC (2001). Règlement sur la qualité de l'eau potable. Gouvernement du Québec. 27 p.

FLOOD, I., KARTAM, N. (1994a). Neural networks in civil engineering. I: Principles and understanding. *Journal of Computing in Civil Engineering* 8:2. 131-148.

FLOOD, I., KARTAM, N. (1994b). Neural networks in civil engineering. I: Systems and applications. *Journal of Computing in Civil Engineering* 8:2. 149-162.

FORTIN G. R., AUCLAIR M.-J., LÉTIENNE-PRÉVOST M., POTVIN P., SÉGUIN D. (1994). Synthèse des connaissances sur les aspects physiques et chimiques de l'eau et des sédiments du lac Saint-Louis – Rapport technique, Zone d'intervention prioritaire. Environnement Canada – Région du Québec, Conservation de l'environnement, Centre Saint-Laurent. Rapport technique. 177 pages.

FRENCH M.N., KRAJEWSKI W.F., CUYKENDALL R.R. (1992). Rainforest forecasting in space and time using a neural network. *Journal of Hydrology* 137. 1-31.

FRENETTE, M., BARBEAU C., VERRETTE, J.-L. (1989). Aspects quantitatifs, dynamiques et qualitatifs des sédiments du Saint-Laurent. Hydrotech inc., pour Environnement Canada et le Gouvernement du Québec.

FRENETTE, R., FRENETTE, M. (1992). Modélisation des biLans sédimantaires du Saint-Laurent tronçon aval : Montréal-Montmagny (modèle Bi-Lavsed). Congrès annuel de la société canadienne de génie civil, Mai 27-29, 1992, Québec.

GARCEAU, P. (2000). Modélisation du contrôle neural des muscles du dos. Mémoire. *École Polytechnique de Montréal*. 157 pages.

GAUTHIER, V. (1998). Les particules dans les réseaux d'eau potable : caractérisation et impact sur la qualité de l'eau distribuée. Thèse. *Université Henri Poincare- Nancy I.* 189 pages.

HAYKIN S. (1994). Neural Networks: A Comprehensive Foundation. Macmillan Publishing, New York, USA.

HECHT-NIELSEN R. (1987). Kolmogorov's mapping neural network existence theorem, First IEEE International Joint Conference on Neural Networks. 11-14.

HYDRO-QUÉBEC (1985a). Avant-projet Archipel. Rapport des études environnementales. Vol. 2, Un milieu à connaître. Tome 1, Milieu physique et biologique.

HYDRO-QUÉBEC (1985b). Dossier synthèse sur le régime hydrosédimentologique des plans d'eau de l'archipel de Montréal. Direction de l'environnement.

HYDROTECH INC. (1988). Fleuve Saint-Laurent : État de connaissance et perspectives sédimentologiques – Synthèse et recoomendations. Document préliminaire produit pour Environnement Canada, région du Québec et le Ministère de l'Environnement du Québec.

JOO, D.-S., CHOI, D.-J., PARK, H. (2000). Determination of optimal coagulant dosing rate using an artificial neural network. *Journal of Water Supply: Research and Technology- AUQA* 49:1. 49-55.

KARUNANITHI, N., GRENNEY, W.J., WHITLEY, D. BOVEE, K. (1994). Neural networks for river flow prediction. *Journal of Computing in Civil Engineering* 8:2. 201-220.

LIONG, S.Y., LIM, W.H., PAUDYAL, G.N., (2000). River-stage forecasting in Bangladesh: Neural network approach. *Journal of Computing in Civil Engineering* 14:1. 1-8.

LOISELLE, C., FORTIN, G. R., LORRAIN, S., PELLETIER, M. (1997). Le Saint-Laurent : dynamique et contamination des sédiments. Environnement Canada, Conservation et Protection – Région du Québec, Centre Saint-Laurent.

MAIER H.R., DANDY G.C. (1996). The use of artificial neural networks for the prediction of water quality parameters. *Water resources research* 32:4. 1013-1022.

MAIER H.R., DANDY G.C. (1997). Determining Inputs for Neural Network Models of Multivariate Time Series. *Microcomputers in Civil Engineering* 12. 353-368.

MAIER H.R., DANDY G.C., BURCH, M.D. (1998). Use of artificial neural networks for modelling cyanobacteria Anabaena ssp. In the River Murray, South Australia. *Ecological Modelling* 105. 257-272.

MAIER H.R., DANDY G.C. (1998a). Understanding the behaviour and optimizing the performance of back-propagation neural networks: an empirical study. *Environmental modelling & Software* 13. 179-191

MAIER H.R., DANDY G.C. (1998b). The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study. *Environmental modelling & Software* 13. 193-209

MAIER H.R., DANDY G.C. (1999). Empirical comparison of various methods for training feed-forward neural networks for salinity forecasting. *Water resources research* 35:8. 2591-2596.

MAIER H.R., DANDY G.C. (2000a). Application of artificial neural networks to forecasting of surface water quality variables : issues, applications and challenges. Artificial Neural Networks in Hydrology. Eds. Govindaraju et Ramachandra. The Netherlands. 287-309

MAIER H.R., DANDY G.C. (2000b). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software* 15. 101-124.

MAIER, H.R., SAYED, T., LENCE, B.J. (2000). Forecasting cyanobacterial concentrations using B-spline networks. *Journal of Computing in Civil Engineering* July. 183-189.

MAIER H.R., DANDY G.C. (2001). Neural network based modelling of environmental variables : A systematic approach. *Mathematical and Computer Modelling* 33:6-7. 669-682.

MAIER H.R., SAYED T., LENCE B.J. (2001). Forecasting cyanobacterium Anabaena ssp. in the River Murray, South Australia, using B-spline neurofuzzy networks. *Ecological Modelling* 146. 85-96.

MORRIS R.D., NAUMOVA E.N., LEVIN R., MUNASINGHE R.L. (1996). Temporal variation in drinking water turbidity and diagnosed gastroenteritis in Milwaukee. *Am. Jour. Public Health* 86:2. 237-239

NEELAKANTAN, T.R., BRION, G.M., LINGIREDDY, S. (2001). Neural network modelling of Cryptosporidium and Giardia concentrations in the Delaware River, USA., *Water Science and Technology* 43:12. 125-132.

ROGER, L.L., DOWLA, F.U. (1994). Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling. *Water Resources Research* 30:2. 457-481.

SARLE W.S. (1994). Neural networks and statistical models. Proceedings of the Nineteenth Annual SAS Users Group International Conference. April 1994.

SCARDI, M. (2001). Advances in neural network modeling of phytoplankton primary production. *Ecological Modelling* 146. 33-45

SHEPHERD A.J. (1997). Second-Order Methods for Neural Networks. Springer, New York, USA.

SNC-PROCÉAN (1992). Caractérisation physico-chimique des sediments du lac Saint-Louis. Environnement Canada, Conservation et Protection, région du Québec, Centre Saint-Laurent, rapport d'étude-pilote.

STATSOFT (1998). Guide d'utilisation du logiciel Statistica Neural Networks. Trajan Software Ltd., Tulsa. 318 pages.

UNICEF, WHO, Water Supply and Sanitation Collaborative Council (2000). Global Water Supply and Sanitation Assessment. Rapport en format électronique disponible à l'adresse *www.who.int/docstore/water_sanitation_health/Globassessment/Global1.htm*

VERRETTE, J.-L. (1990). Délimitation des principales masses d'eau du Saint-Laurent (Beauharnois à Québec). Environnement Canada, Conservation et Protection, région du Québec, Centre Saint-Laurent.

WILSON, H., RECKNAGEL, F. (2001). Towards a generic artificial neural network for dynamic predictions of algal abundance in freshwater lakes. *Ecological Modelling* 146. 69-84.

XIAO, R., CHANDRASEKAR, V. (1997) Development of a neural network based algorithm for rainfall estimation from radar observations. IEEE Transactions on geoscience and remote sensing 35:1. 160-171.

ZHANG Q., STANLEY S.J. (1997). Forecasting raw-water quality parameters for the North Saskatchewan River by neural networking modeling. *Water Research* 31:9. 2340-2350

ZHANG Q., STANLEY S.J. (1999). Real-time water treatment process contro; with artificial neural networks. *Journal of Environmental Engineering.* Février. 153-160.

ANNEXE 1

Impact of raw water turbidity fluctuations on drinking water quality in a distribution system

ANNEXE 1 - IMPACT OF RAW WATER TURBIDITY FLUCTUATIONS ON DRINKING WATER QUALITY IN A DISTRIBUTION SYSTEM

Article accepté pour publication dans le « Journal of Environmental Engineering and Science ».

Vincent GAUTHIER[1,#], Benoit BARBEAU[1*], Geneviève TREMBLAY[1], Robert MILLETTE[2], Anne-Marie BERNIER[3]

[1] Ecole Polytechnique de Montréal, NSERC Industrial Chair on Drinking Water, Civil Geological and Mining Engineering, P.O. Box 6079, Centre-Ville, Montreal (Quebec), Canada, H3C 3A7

[2] City of Montreal, Atwater Water Treatment Plant, 3161 Joseph Road, Verdun (Quebec), Canada, H4G 1H8

[3] City of Montreal, Laboratory, DesBaillets Water Treatment Plant, 8585 De la Verendrye Boulevard, LaSalle (Quebec), Canada, H8N 2K2

[*] Corresponding author: tel: 1(514)340 4711 ext. 2988, fax: 1(514)340 5918, e-mail: bbarbeau@polymtl.ca

[1,#] Current affiliation: Vivendi Water, 18 Malesherbes Boulevard, 75008 Paris, France

Abstract: Turbidity is a widely used parameter around the world for describing drinking water quality. Sometimes, turbidity at water treatment plant outlets may reach high values during short periods of time, and this is acceptable according to some current drinking water regulations. In this study, the quantity and nature (chemical and microbiological) of suspended matter, which may travel throughout a distribution system (DS) during turbid events - affecting both raw water and water treatment- were evaluated. Treated water included filtration with no coagulant addition. During turbid events, the concentration of suspended particles increased in treated water, and a similar increase (quantity and nature) was observed throughout the DS. Bacterial indicators of contamination (total and fecal coliforms, *Enteroccocci*, spores of *Clostridium perfringens*) were not found in either treated water nor in the DS during turbid events. Nevertheless, a higher bacterial aerobic spore concentration was associated with turbid events for raw, treated, and distributed water, therefore suggesting the potential passage of pathogens, if present in raw waters. Cultivable bacteria concentrations remained low in treated and distributed water regardless of the turbidity. These results emphasize the need to carefully monitor raw and treated water quality for utilities using "high quality" water resources with limited treatment barriers, especially when such water resources are affected by even slight turbidity variations.

Keywords: aerobic spore-forming bacteria, distribution system, drinking water, filtration, turbidity, suspended particles, water quality

Introduction

The suspended particulate matter present in surface water is usually largely eliminated in water treatment plants by processes such as coagulation, settling, and filtration. In some cases, for high quality surface water, especially if it is of low turbidity, filtration waivers may be granted to utilities. For example, the cities of New-York, Boston, and Seattle in the U.S.A., and Vancouver in Canada presently have part or all of their supply unfiltered, while some of them are in the process of implementing filtration (Ferguson and Neden 2001). For such unfiltered supplies, treatment barriers for particle removal (coagulation, settling, membrane, or granular media filtration) are avoided and the focus is on source water protection and the use of disinfection treatments. For certain utilities that do not use filtration, significant turbidity events may penetrate the distribution system (DS) from time to time, and this is in agreement with current regulations as long as these turbidity events are "moderate and rare" and as long as an efficient disinfection is applied. For example, US regulations require that for utilities delivering unfiltered surface water, turbidity be kept lower than 5 NTU except during two or less unpredictable events in any twelve consecutive months (USEPA 1998). Filtered water regulations may also tolerate "moderate and rare" treated water turbidity peaks. In Quebec, new regulations require that 95% of filtered water turbidities, measured every 4h, be lower than 0.5 NTU on a monthly basis. Another requirement is that the maximum turbidity always remain lower than 5 NTU (Environnement Quebec 2001).

In Montreal (QC, Canada), the St. Lawrence River is used as a source for the production of drinking water. Its average flow rate is about 9000 m^3 s^{-1}. Due to good raw water quality at the intake location (Payment et al. 2000), and prevailing regulations up to very recently (Environnement Quebec 1984; Environnement Quebec 2001), a relatively simple treatment (filtration without coagulant + disinfection) is performed to produce drinking water. At the two treatment plants, treated water turbidity is usually low, 0.23 NTU and 0.19 NTU respectively (average values from 1,065 daily values for the 1998-2000 period). Nevertheless, suspended particulate matter is introduced into the DS during a small number of turbidity events, especially in the spring and fall. During these events, treated water turbidity may temporarily be higher than 0.5 NTU (Figure 1), despite a rapid filtration treatment (no coagulant or settling). These peaks result from the raw water turbidity, usually ranging from 0.3 to 3 NTU, temporarily reaching values higher than 10 NTU. These turbidity events are attributed to a combination of a number of factors, such as (i) the increased flow due to the spring snow melt or heavy rainfalls, (ii) the spring turnover of water masses in upstream lakes when temperatures reach 4°C, which corresponds to the maximum water density; (iii) bank erosion and sediment resuspension in shallow areas following windy periods, and (iv) the temporary increase of water from the more turbid Outaouais River, which is a tributary of the St. Lawrence River Seaway, just upstream from Montreal. The respective effects of these phenomena are currently being investigated and are not yet fully understood.

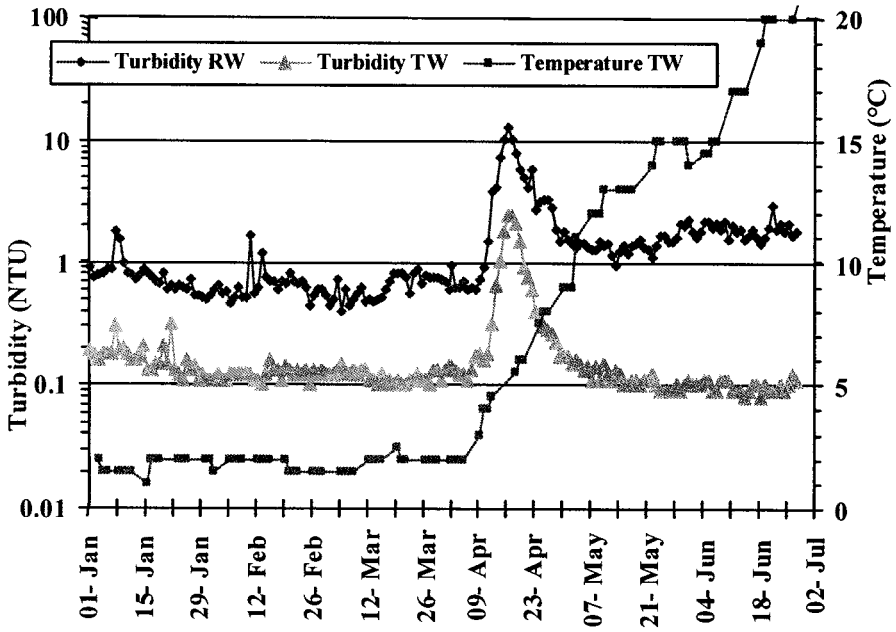


Figure 1: Turbidity and temperature for raw water (RW) and treated water (TW) as measured at one of the two Montreal water treatment plants, from January to July 2001.

High turbidity values reported in the Montreal DS during raw water turbidity events raise questions about the extent and consequences of such an introduction of particulate matter into the DS. In such cases, the issue of disinfection efficiency may become critical due to the potential association of microorganisms with suspended particles (Berman et al. 1988; Gauthier et al. 1999a; Morin et al. 1999). The question of the possible health effects of transitory increased turbidity in drinking water have also been raised recently (Morris et al. 1996; Beaudeau et al. 1999; Aramini et al. 2000).

The objectives of the project presented herein were defined as follows:

- to measure the quantity and nature of particulate matter introduced into the system during transitory turbidity events;
- to assess the microbial content associated with turbidity peaks;
- to evaluate which fraction of these particles travels throughout the DS to the consumer's tap, or is deposited as sediment in pipes and storage facilities.

To achieve these goals, suspended particles in raw, treated, and distributed water were sampled and characterized from 1999 to 2001 for spring turbid events (referred to as turnover, and defined by a treated water turbidity > 0.5 NTU) and for stable reference conditions (treated water turbidity <0.2 NTU). The main parameters that were measured were (i) the concentration of suspended solids as mg L^{-1}, and (ii) their organic and mineral content, derived from the filtration of 2 to 150 L on fiberglass or cellulose

acetate membrane filters. The microbial quality of water was analyzed at the same time, for typical microbial quality indicators (heterotrophic plate counts, fecal and total coliforms, *enterococci*) as well as for more specific parameters such as spores of aerobic bacteria and *Clostridium perfringens* spores.

Material and Methods

Montreal Water Treatment System

The City of Montreal provides water to about 1.5 million people using two treatment plants. The first one (Plant #1, 1,590,000 m^3 d^{-1} capacity) uses filtration and chlorination (no coagulation or settling), and the other one (Plant #2, 1,136,000 m^3 d^{-1} capacity) uses filtration, ozonation, and chlorination (no coagulation or settling). The distributed water quality is very good and complies with local water quality standards (Desjardins et al. 1997), which is related to the very good quality of the St. Lawrence River water at the intake location most of the time (Payment et al. 2000). The distribution system includes six covered storage tanks to regulate pressure at various elevations in the city and to provide additional flow for periods of peak demand. The distributed water is well mineralized (alkalinity = 90 mg $CaCO_3$ L^{-1}, pH 7.8, total hardness 126 mg $CaCO_3$ L^{-1}) and not very aggressive (aggressivity index = 11.9). Consequently, consumer complaints due to red water events are rare.

Sampling points, for the characterization of suspended particulate matter, include raw water (RW1 and RW2) and treated water (TW1 and TW2) for the two plants, 1 and 2. In the DS, the sampling point DS_1 is upstream in the DS, just at the outlet of a major storage tank located about 5 km from the mixing point between TW1 and TW2. DS_2 is in the middle of the distribution system, about 12 to 15 hours downstream from point DS_1 based on hydraulic modeling. DS_3 is a dead end location on an unlined 203 mm (8-in.) cast-iron main. For distribution system sampling, special care was taken to avoid the local impact of household pipes by sampling directly at water mains.

Since the St. Lawrence River has one major turbid period in the spring (referred to as "turnover"), the impact of this turbid period on water quality was evaluated by comparing two sampling periods (Table 1): during the first period ("normal" or "reference" conditions), the typical turbidity of the treated water is less than 0.2 NTU (May 1999; July 2000), while during the "turnover" period, the typical turbidity of the treated water is higher than 0.5 NTU (April 1999, 2000 and 2001).

Table 1: Sampling campaigns to characterize of suspended particulate matter at different times and locations

	Turnover			Reference	
	April 1999	April 2000	April 2001	May 1999	July 2000
Raw water plant 1		✓	✓		✓
Treated 1	✓	✓	✓	✓	✓
Raw water plant 2			✓		
Treated water plant 2			✓		
Distribution system	✓	✓		✓	✓

Characterization of Suspended Particles

Sampling. For treated and distributed water, suspended particulate matter is sampled and concentrated on 47mm filters using a 2-line filtration system (Gauthier et al. 2001) prior to gravimetric, mineral, and visual analyses. Each line consisted of a two-filters assembly series, the first filter retaining the "particulate matter" from the water, and the second acting as a reference filter, both of them being exposed to exactly the same conditions (preparation, type, and quantity of water, analyses). This approach prevents problems such as confusion between particulate and adsorbed material when the filter is analyzed, since the value of the reference filter is systematically subtracted from the value of the exposed filter.

Before each filtration experiment, the system (without filters) is thoroughly washed and rinsed for a 15 minute period with the tap water to be studied. The filters are then placed in the Swinnex filter holders (Millipore, ref. SX00). During the filtration process, the upstream pressure on the filters and the flow are continuously monitored using pressure gauges and rotameters. These parameters vary depending on the degree of clogging in the filters. The flow is adjusted with a throttle valve to limit the pressure on the filter and the maximum flow rates are also limited to 1.3 L min^{-1}. The total volume filtered for each line (23 to 170 L as a function of the suspended solids concentration) is then calculated according to the filtered flow rate and the elapsed time. Following filtration, the water in the filter holders is discarded, and the filters are transferred to a crucible and dried according to the type of analyses to be performed.

For raw water, the concentration of suspended solids is much higher and a 1 to 5 L volume is filtered in the lab under vacuum with a filtration ramp (Standard Methods, 1995) using 10 L grab samples collected on site and thoroughly shaking them by hand so as to mix them before filtration.

The type of filter is selected according to the filtration capacity and type of analyses to be subsequently performed: (a) gravimetric analysis of total and volatile suspended solids (TSS and VSS) requires carbon-free filters in order to avoid any interference with the small quantities to be measured; fiberglass filters (Millipore, ref. AP40) precombusted at 500°C are used for this purpose; (b) mineral analysis requires acid-soluble filters permitting easy recovery of the sample following mineralization; 5-µm cellulose-acetate filters (Millipore MF5, ref. SMWP) are selected for this purpose and are weighed following drying at 37°C (Gauthier et al. 2001).

Total and Volatile Suspended Solids Analysis. The fiberglass filters are weighed following drying at 105°C for 48 hours. They are then heated in a muffle furnace at 500°C for 30 minutes and weighed again. The mass loss due to combustion is attributed to the volatile fraction of particles, i.e. organic matter. Reference filter mass values are subtracted from those of the exposed filter and the total and volatile suspended solid (TSS and VSS) concentrations are expressed as mg L^{-1} or as percentages of the TSS. As the accuracy of the gravimetric procedure is estimated to be better than 0.5 mg and since volumes as high as 170 L may be filtered, the theoretical detection limit for the method is estimated to be < 0.003 mg/L. The accuracy of this method in general is estimated to be around 10% for suspended solids at 37°C and 105°C, and 20% for VSS due to the smaller mass collected on the filters (1 to 27 mg).

Mineral Elementary Analysis. Filters are weighed following drying at 37°C. They are then mineralized in a teflon pot containing 1 mL HCl, 0.5 mL HNO_3 and 0.5 mL HF. The concentrate is then diluted and analyzed for Si, Al, Fe, Ca, and other elements by Flame Atomic Absorption Spectrometry. Results are expressed as mg L^{-1} or as percentages of the TSS. For mass balance purposes, elemental fractions are expressed as oxides and hydroxides (SiO_2, FeOOH, $Al(OH)_3$) or as carbonates ($CaCO_3$).

Microscopic Examination. 1 to 30 L of water are also filtered through a 5-µm cellulose-acetate filter (Millipore MF5, ref. SMWP) to allow microscopic examination of suspended particles retained on the filter. A BX-60 Olympus microscope with 100x to 400x magnification is used for this purpose.

Water Quality Analysis

Traditional water analysis techniques are used. Turbidity is measured using a Hach Ratio laboratory turbidimeter. For measuring the concentration of cultivable bacteria or aerobic spores in water, 1 to 100 mL of sample (or appropriate dilutions) is filtered on a 0.45 µm filter. For cultivable bacteria, the filter is then deposited on R2A agar for 7 days at 20°C (Standard Methods, 1995). For aerobic spore-forming bacteria, appropriate dilutions of the 500 mL sample are filtered on a 0.45 µm membrane which is first pasteurized for 15 min at 75°C and then incubated for 24 hr at 35°C on a pad to which 1.4 mL of Trypticase Soy Broth has been added (Barbeau et al. 1997). Spores of *C. perfringens* were enumerated on m-CP medium as described by Bisson and Cabelli (1979). The experimental medium (batch form, Difco) was supplemented with 60 mg of

beta-D-indoxil glucoside per liter. The plates were incubated in anaerobic conditions in a BBL GasPak pouch at 44.5°C for 24h. Colonies were then exposed to ammonium hydroxide vapors for 20 to 30 s and all yellow colonies turning pink to red were counted. Total coliforms were enumerated on m-Endo agar after 24h at 35°C. Fecal coliforms were enumerated on MFC agar after 24h at 44.5°C. *Enterococci* were enumerated on m-enterococcus agar after 48h at 35°C. Only catalase-negative red colonies were counted. All three methods have a 1 CFU/100 mL detection limit. *Cryptosporidium* and *Giardia* were analyzed using the USEPA-Information Collection Rule (ICR) method. Volumes of 100 to 200 liters of raw water were filtered through 1.0 µm nominal polypropylene filters. The filters were then desorbed and the filtrate concentrated by centrifugation and marked with fluorochrome antibodies. Internal structures were examined under microscope using Nomarsky contrast for confirmation.

Results

Characterization of Untreated and Treated Water

Turbidity and the concentration of suspended solids were measured in untreated and treated water at different dates, covering both turnover and reference conditions. Treated water values for both parameters were 2 to 10 times higher during turnover compared to reference conditions, despite the rapid filtration treatment (Figure 2). During turnover, treated water turbidity was in the 0.5-2.7 NTU range. For non-turnover conditions, treated water turbidity was in the 0.05-0.2 NTU range. Measurements of suspended solids by gravimetry provided very similar results regardless of the type of filter used to collect particles (AP40 fiberglass filter or cellulosic 5 µm membrane). The concentration of suspended solids in raw water was about 1.46 mg L^{-1} for non-turnover conditions, while during turnover, it ranged from 1.81 to 14.7 mg L^{-1}. In treated water, the concentration of suspended solids reached values as high as 0.68 mg L^{-1} during turnover, while it did not exceeded 0.04 mg L^{-1} when sampling in reference conditions. A relationship between turbidity and suspended solids profiles is evident in Figure 2 and this is confirmed by the significant correlation found between these two parameters in raw and treated water, regardless of the type of filter used:

[1] $[TSS(\mu g\ L^{-1})] = 392 \times [TURBIDITY(NTU)]^{1.14}$, with $r^2=0.91$, $n=18$, $p<0.001$

The composition of suspended solids differed between turnover and reference conditions (Figure 3). During turnover, particles were very similar in both raw and treated water, and mostly composed of Si (48-49%), Al (15%) and organic matter (13-19%, measured as VSS). In reference conditions, greater differences were observed between raw and treated water: raw water particles were mostly composed of Si (43%) and organic matter (27%), and treated water particles were mostly organic (68%) with a much smaller fraction of Si (17%).

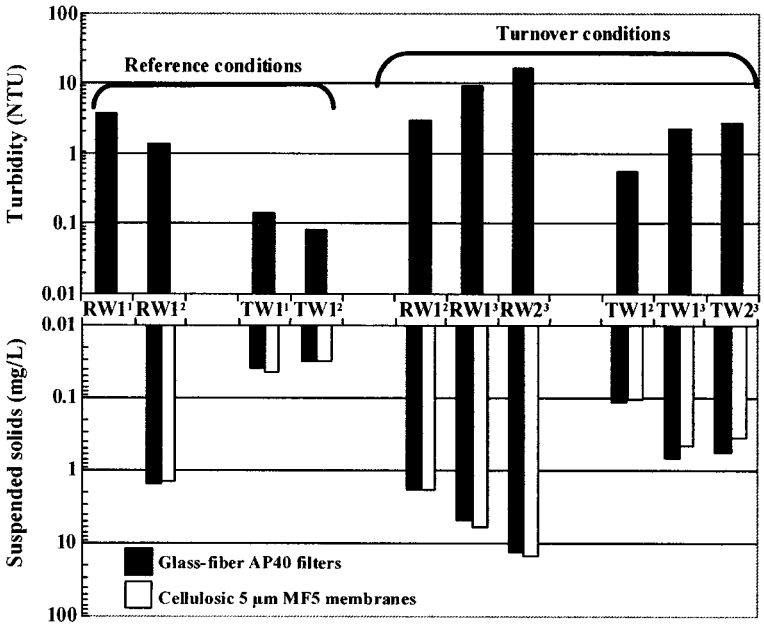


Figure 2: Turbidity and concentration of suspended solids for raw (RW) and treated water (TW) during reference and turnover conditions ([1]: 1999; [2]: 2000; [3]: 2001).

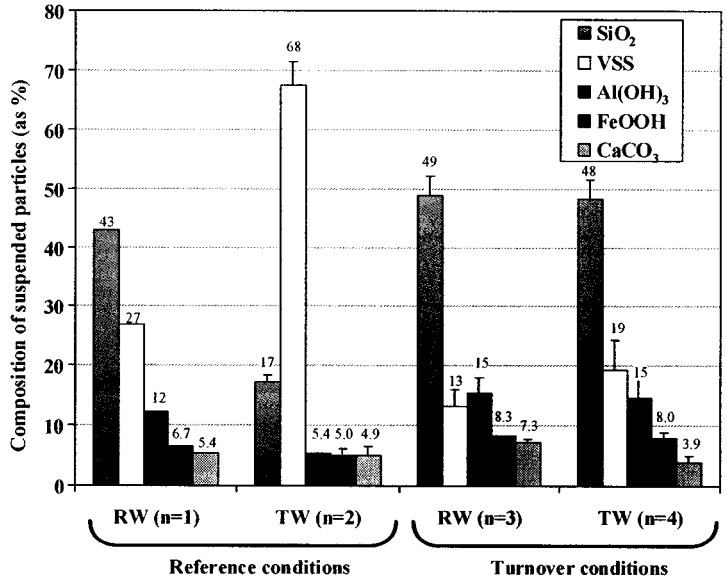


Figure 3: Average composition of suspended particles in raw (RW) and treated (TW) waters during reference and turnover conditions. For mass balance purposes, measured elemental compositions were expressed as oxides and carbonates, the remaining unknown amount being <10% of the total mass of suspended solids (error bars represent standard deviation for n≥2).

Distribution of Particles throughout the DS

The quantity and nature of suspended solids throughout DS were monitored during two successive spring turnovers, and compared to reference conditions found in May 1999 and July 2000. Turbidity measured along the DS shows similar and "flat" profiles for both types of conditions, with values obtained at the entry of the DS being similar to those at the other end (Figure 4). The higher treated water turbidity during turnover seems minimally affected by DS interference (such as settling, resuspension, etc). This is also confirmed by turbidity values measured in reference conditions, for which no degradation is observed from the treatment plant outlet to the DS end (all values are lower than 0.2 NTU).

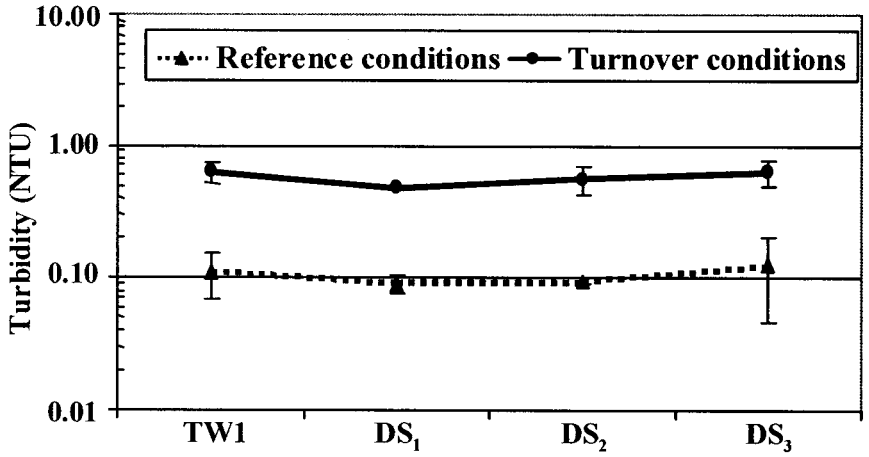


Figure 4: Turbidity profiles in the distribution system for reference and turnover conditions (increasing residence time from TW1 to DS_3; average of 1999 and 2000 data; error bars represent the standard deviation).

Particulate matter composition is also similar along the DS during turnover (Figure 5). Si, Al and organic compounds represent about 53%, 21%, and 13% respectively of suspended solids, regardless of sampling location, with minor amounts (8% and 4%) of Fe and Ca compounds. The concentration of suspended solids also remained stable along the DS, except in the upstream part (between TW1 and DS_1) which included the transit through a large storage tank and where the average concentration decayed by 43% (from 0.14 to 0.08 mg L^{-1}, $p=0.03$). All elemental amounts were similarly affected by such decay.

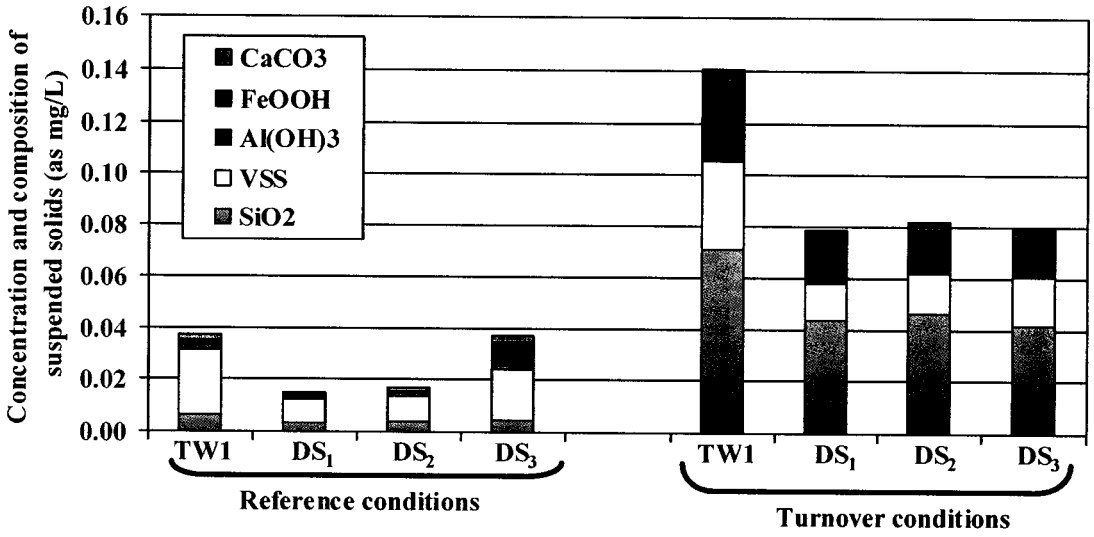


Figure 5: Concentration and composition of suspended solids in the distribution system during reference and turnover conditions (increasing residence time from TW1 to DS_3; average of 1999 and 2000 data, corresponding to the turbidity profiles shown in Figure 4).

In reference conditions (Figure 5), the concentration of suspended solids remained lower than 0.04 mg L^{-1} and in some cases was very close to the quantification limit of the gravimetric method (0.01 mg L^{-1}). The composition of the particles was predominantly organic matter (53 to 67%), while small amounts of Si and Al compounds were in the range of 12-17% and 4-8% respectively. With the exception of calcium, the proportions of SiO_2, VSS, $Al(OH)_3$ and FeOOH were significantly different in treated water during turnover conditions ($p < 0.03$). Similar to turnover conditions, some particles were removed at the upstream portion of the DS, between the outlet of the treatment plant and the exit of the main storage tank (DS_1). At the end of the DS, the concentration of particulate organic matter slightly increased, along with the Fe concentration, probably due to corrosion-enhancing dead end conditions at the DS_3 location.

To evaluate the ability of particles entering the DS during turnover to accumulate as sediment in pipes and storage tanks, the composition of the suspended solids was compared to deposits collected from the finished water storage tank at Treatment Plant 1, and also from DS pipes (Figure 6). Pipe deposits were collected during unidirectional flushing of 150 to 300 mm (6 to 12 in.) pipes mostly made of unlined cast iron (Pinot 2000). Mineral matter proportions were very similar to sediments from the storage tank and to suspended particulate matter found in treated water from the turnover period, particularly the Si compounds (46 versus 48%). In contrast, deposits from water mains were mostly composed of iron oxides and hydroxides with very little Si compounds (9%) and negligible amounts of Al and Ca, indicating that such deposits were mainly from the result of internal pipe corrosion. It is worth noting that the quantity of pipe deposits was relatively low (< 1 g m^{-1} from the flushed pipe), and thus, the quantity of

deposits accumulated in pipes from sedimentation of Si-rich turnover suspended particulate matter should be considered a minor process for the accumulation of deposits in the Montreal DS pipes.

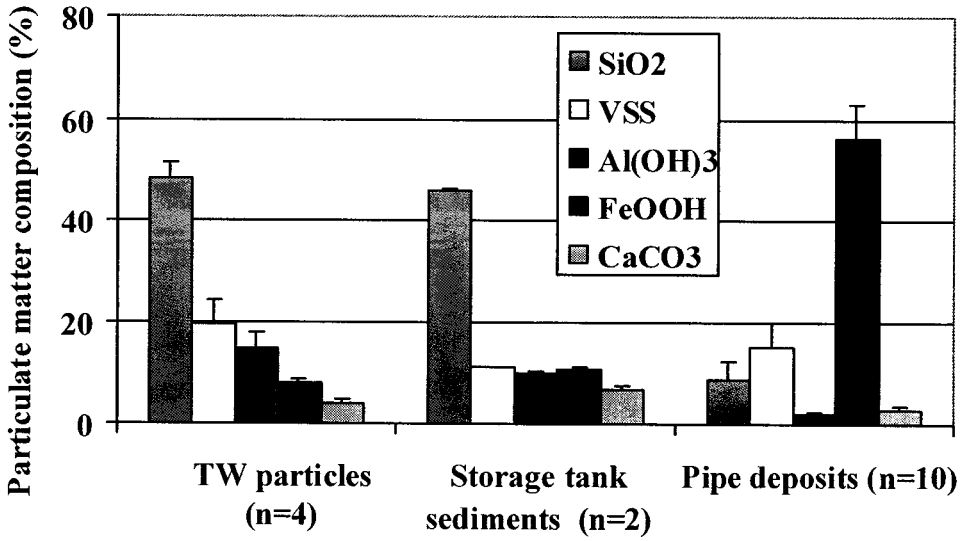


Figure 6: Composition of treated water (TW) suspended particles during turnover events, finished water storage tank sediments, and pipe deposits in the Montreal distribution system.

Microbial Impact of Turbidity Events

The microbial water quality was characterized in order to evaluate (a) the increase in the microbiological load during turnover events, and (b) the potential introduction and transit of microorganisms into the distribution system during such turbid periods. The total coliform concentration in untreated water increased slightly during the turbidity peak (Figure 7), but never exceeded 251 CFU/100 mL during the time of the study, which may be considered a low value (Environnement Quebec, 2001). Filtration decreased this concentration from 0.2 to 1.3 Log, and following disinfection, no coliform could be detected in treated water from either plant (Figure 7). Other bacterial indicators (fecal coliforms, *Enterococci*) were also quite low in untreated water (always <57 and <89 CFU 100 mL^{-1} respectively) during the 1999-2001 spring turnovers. These organisms also remained undetected in treated water and in distribution system sampling points during turnover sampling periods (<1 CFU 100 mL^{-1}). A concentration of 50 CFU mL^{-1} of *Clostridium perfringens* spores was enumerated in raw water in April 2000, while none was detected downstream from the treatment train. Nevertheless, specific sampling during 3 successive days during the 2001 turbidity peak revealed the presence of presumed *Cryptosporidium* and *Giardia* (oo)cysts in 100 L of raw water samples, with 2 and 5 positive samples respectively (total of 6 analyzed samples). This result suggests a higher microbiological risk during high turbidity events. However, it was not possible to confirm the presence of these organisms by examining their internal

structure under a microscope, because of the high concentration of particles in the samples. It is worth noting that such (oo)cysts are usually not detected during routine monitoring or during specific sampling performed in non-turnover conditions (Payment et al. 2000). There was no attempt to search for (oo)cysts in treated nor in distributed water.

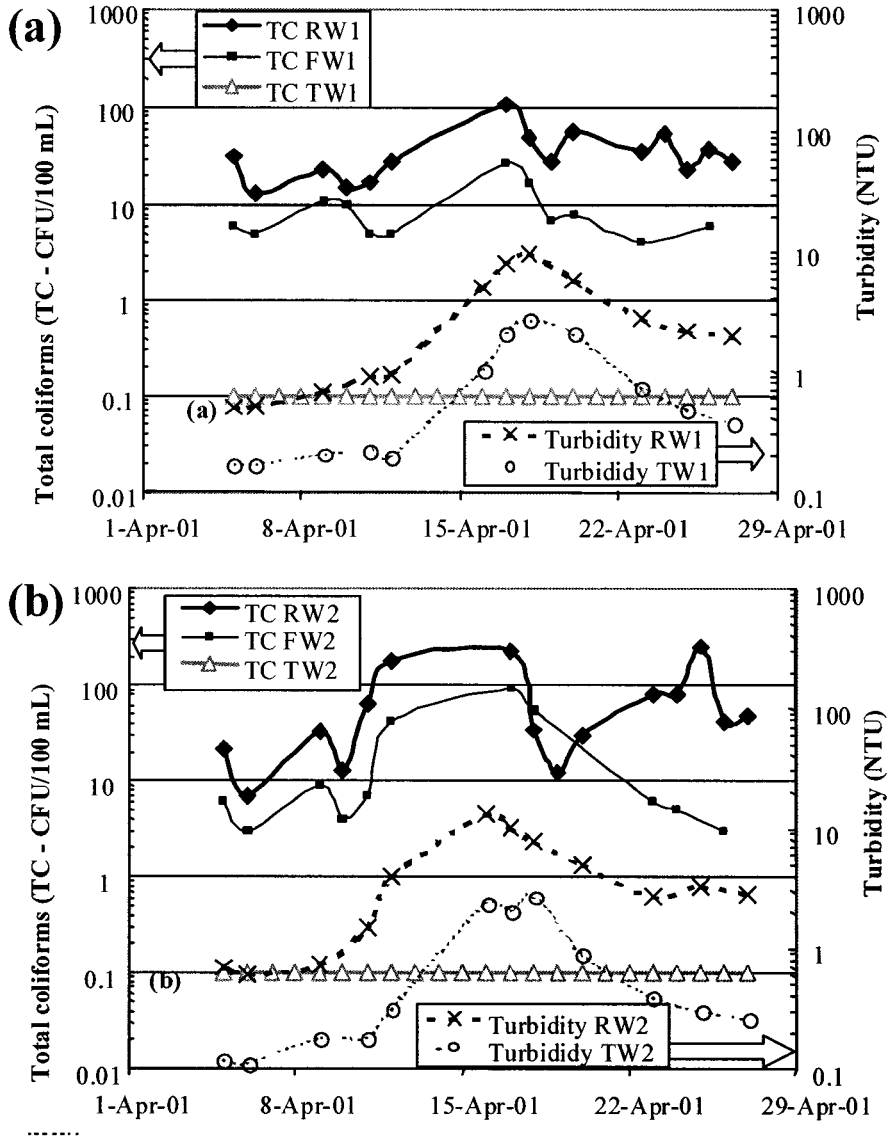


Figure 7: Total coliform bacteria and turbidity variations at (a) treatment plant #1 and (b) treatment plant #2, in April 2001 (RW: raw water; FW: filtered water; TW: treated water)(Total coliform detection limit: 1 CFU/100 mL).

Measurements revealed a higher concentration of aerobic spore-forming bacteria during spring turnover (Figure 8), which strongly correlates to a power law with turbidity in untreated water in the two treatment plants (r^2=0.93, n=24, p<0.0001) (Figure 9). Aerobic spore-forming bacteria concentrations in untreated water supplying the two treatment plants were10 to 20 times higher during turnover, the highest values occurring during the highest turbidities recorded (Figure 8 and 9). Concentrations of aerobic spores in treated water also increased according to untreated water turbidity (Figure 8), but values were about 2 to 5 times lower than in untreated water as a result of treatment. During the turbid event, most of the aerobic spores were removed by filtration, while the disinfection processes provided little additional inactivation (Figure 8). During the turbidity peak, from April 11 to 20, the average removal due to filtration in plant #1 (n=5) and #2 (n=6) was 0.5±0.2 and 0.5±0.1 Log respectively, while at the same time, the inactivation due to disinfection processes was only 0.0±0.1 and 0.1±0.1 Log, respectively. This is probably due to the poor efficacy of chemical disinfection for such microorganisms and such low temperatures (3-7°C) (Barbeau et al. 1999; Finch and Choe 1999). The concentration of aerobic spore-forming bacteria in treated water from the two treatment plants also correlated to a power law with the raw water turbidity (r^2=0.68, n=24, p<0.001). The parallel between the two correlations (Figure 9) plotted along logarithmic scales indicates a nearly constant level of removal efficiency (0.62 to 0.70 Log corresponding to 75-80% removal) by the treatment trains regardless of untreated water turbidity during the period in question.

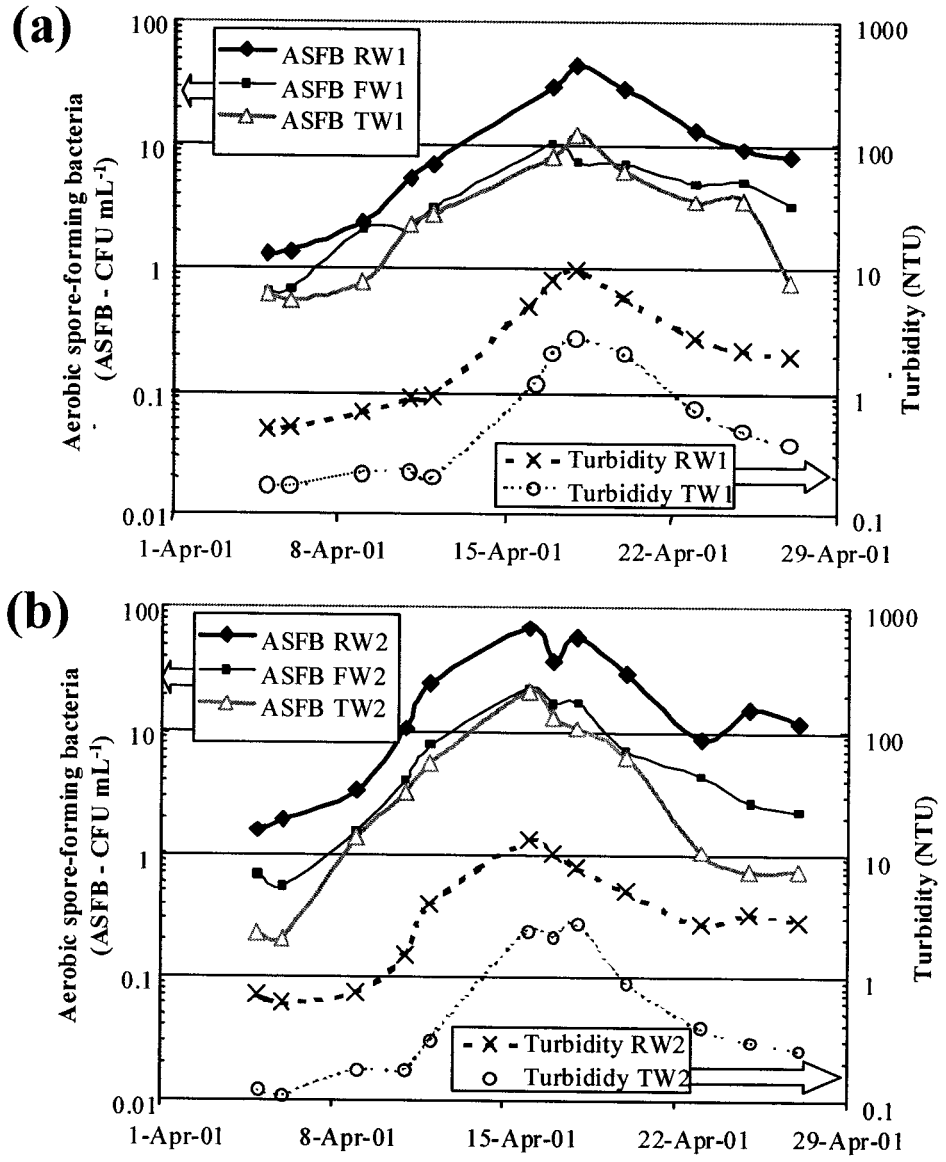


Figure 8: Aerobic spore-forming bacteria and turbidity variations at (a) treatment plant #1 and (b) treatment plant #2, in April 2001 (RW: raw water; FW: filtered water; TW: treated water).

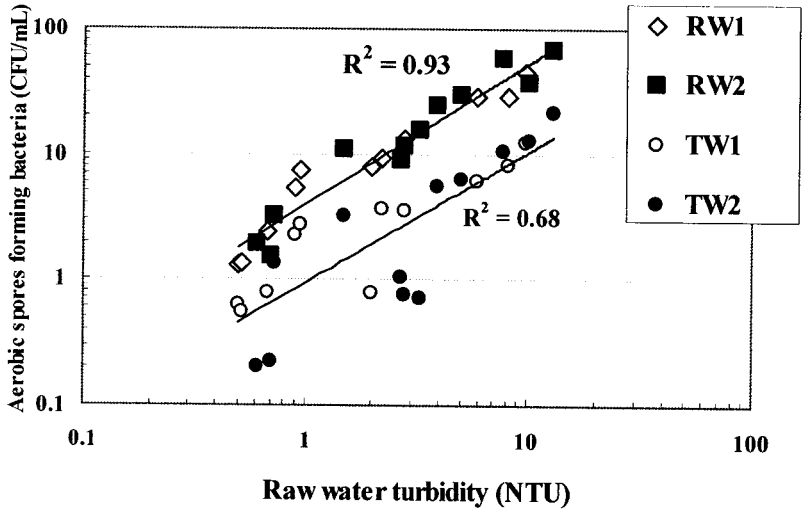


Figure 9: Aerobic spore forming bacteria concentration in the raw and treated water of the two treatment plants during the April 2001 period (including spring turnover) (least square power regressions were used).

In the DS, aerobic spore-forming bacteria are decaying according to residence time (Figure 10a), due to their inactivation by the residual free chlorine and possibly to other factors such as adhesion to pipe walls, biofilms, and deposits, or biodegradation. This decay was significantly ($p<0.01$) faster for reference than for turnover conditions (1.04 Log between TW1 and DS_1 rather than 0.62 Log), probably because of the improved efficiency of disinfection at higher temperatures (Barbeau et al. 1999) (in DS_1, the temperature is 21.5°C in reference conditions versus 7.0°C for turnover, free chlorine residuals are similar, respectively 0.60 versus 0.58 mg Cl_2 L^{-1}). During turnover conditions, since (i) the concentration of aerobic spore-forming bacteria in treated water is higher, and (ii) the decay in the DS is slower, the aerobic spore-forming bacteria concentration is about 100-fold higher along the DS than during the reference conditions.

Cultivable bacteria, measured as R2A-7 days HPC, have quite a different DS profile (Figure 10b). Such microorganisms are more readily inactivated by free chlorine even in cold water, and therefore treatment plants show very high and similar removal for turnover and reference conditions (3.9 and 3.8 Log respectively, corresponding to 99.988% and 99.983%). Therefore, the slightly higher raw water HPC concentration during turnover is also observed in treated water, compared to reference conditions. In the DS (turnover conditions), cultivable bacteria are kept to a lower level than in treated water, this absence of regrowth being associated with the low temperature and moderate chlorine residual (0.34 mg Cl_2 L^{-1} at the most remote sampling point DS_3). For reference (summer) conditions, the temperature in the DS is much higher and the chlorine decay between DS_1 and DS_3 is faster, resulting in a DS_3 free chlorine concentration of 0.11 mg Cl_2 L^{-1}. This may explain why some regrowth is observed, resulting in a 10-fold HPC increase between DS_1 and DS_2.

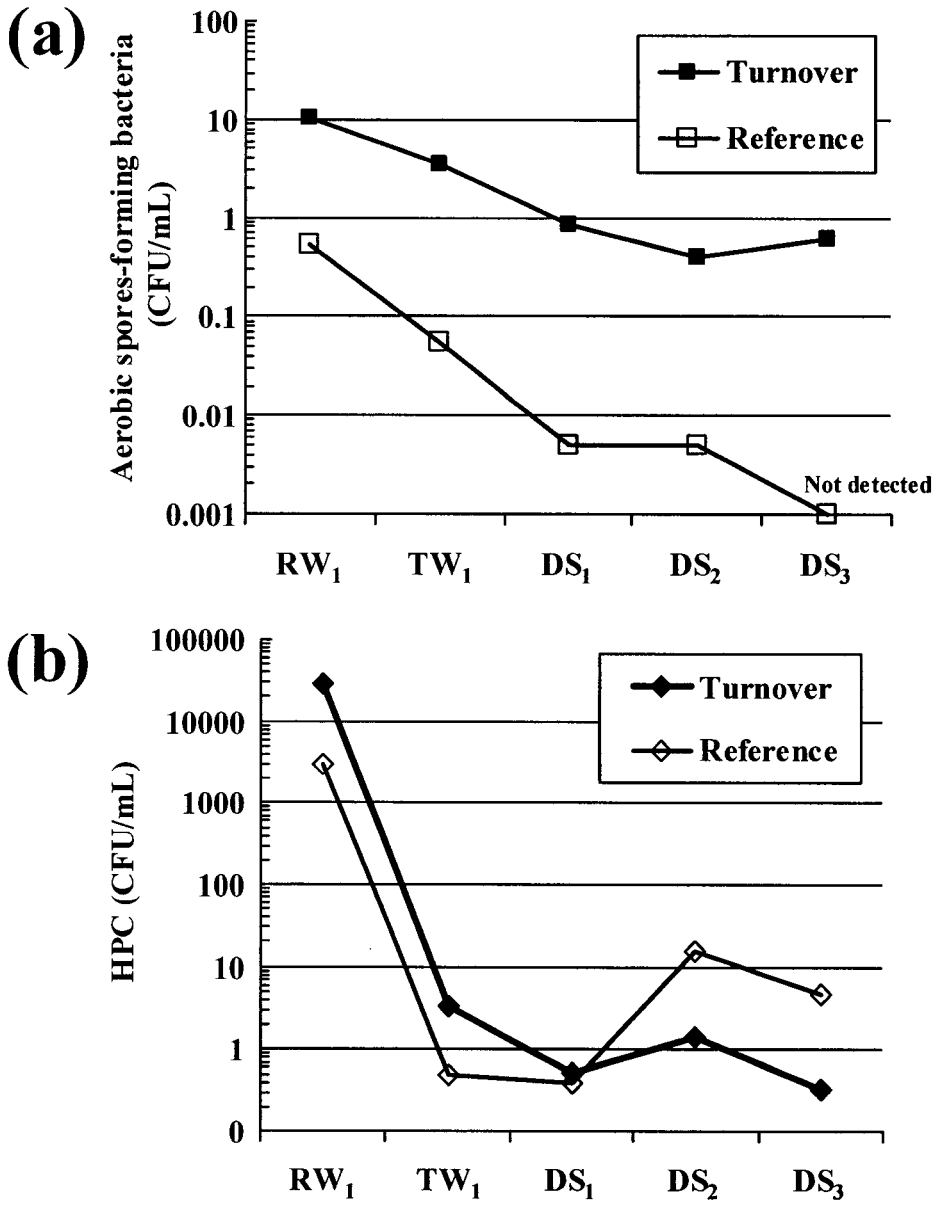


Figure 10: (a) Aerobic spore-forming bacteria and (b) HPC concentration profiles along the distribution system for the year 2000 turnover and reference conditions (increasing residence time in the DS from TW_1 to DS_3).

Discussion

The water quality in a large distribution system was compared under two conditions, namely reference conditions (treated water turbidity <0.2 NTU) and turnover conditions (treated water turbidity >0.5 NTU). Such turnover conditions occur during short periods of time only –typically less than 5% of the time, based on daily turbidity values (data not shown). The quantity and nature of particulate matter penetrating the system during such turbid events were investigated, along with the microbial impact of turnover events on drinking water quality. The particulate nature of such material in treated water may be of concern due to a number of potential consequences:

(a) the accumulation of deposits in the DS may form favorable ecosystems for microorganisms (Gauthier et al. 1999b) or lead to consumer complaints or violation of microbial water quality standards when resuspended (McCoy and Olson 1986; Jones and Tuckwell 1993);
(b) the increased turbidity may be accompanied by a higher microbial content, including not only "free" microorganisms, but also particle-related microorganisms (bacterial aggregates, colonized particles, bacteria inside larger microorganisms) (Lupi et al. 1995; Morin et al. 1999; Schoenen and Hoyer 2000)
(c) the efficiency of disinfection processes may decrease when turbidity increases (LeChevallier et al. 1981; Boyce 1981).

In this study, the concentration of suspended solids in treated water during turnover (maximum: 0.64 mg L^{-1}) was much lower than dissolved material concentrations (e.g.: total hardness 126 mg $CaCO_3$ L^{-1}), but also much higher than the concentration of suspended solids during reference conditions (0.04 mg L^{-1}). The concentration of suspended solids was also much higher than other values measured in fully treated surface water from systems such as Nancy (France) (Gauthier et al. 1997) and Jonquière (Quebec, Canada) (data not shown), which ranged between 0.01 and 0.09 mg L^{-1} when there is no turbid event.

The relative stability in particulate matter concentration and composition in the DS that was found in this study for reference conditions is consistent with previous studies (Gauthier et al. 1997). However, other studies based on turbidity and particle counting revealed different particle dynamics throughout a Swedish DS, mostly due to the dissolution process of lime particles (Alere and Hanaeus 1997).

During turnover, the nature of suspended particles was similar for all sampling points along the DS, while their concentration revealed only some limited deposition in storage facilities in the upstream part of the system. This observation coincides with the microscopic examination of filters which revealed that most of the suspended particles are very small (in the range of microns) and that the only large particles are algae. The

viability of such algae was not evaluated, but many of them did not exhibit any content and were apparently only algae skeletons. This could explain why such particles, having a density close to that of water, travel well with a minimal sedimentation along the DS.

During turnover events, these particles, introduced into the distribution system, are mostly composed of mineral matter (81%) while organic matter, measured through VSS, accounts for only 19%. Such a high mineral content (combined with the low water temperature (3-7°C), which does not favour microbial proliferation), could possibly minimize the impact of such turbid events on the microbial water quality. This is proven by typical indicators of microbial contamination (total and fecal coliforms, *enterococci,* spores of *Clostridium perfringens*) that remain undetected in treated and distributed water during such a period using cultivation techniques.

Nevertheless, other microbial features support this information, which was obtained in this case only because of the specificity of the implemented measurements (event-based sampling strategy, type of microorganism searched for). The detection of presumptive *Giardia* and *Cryptosporidium* (oo)-cysts in raw water samples was probably strongly related to the fact that samples were taken only during the highest turbidity period, while such organisms usually remain mostly undetected during the monthly routine sampling or in other studies (Payment et al. 2000). Atherholt et al. (1998) already pointed out the potential for higher concentrations of such microorganisms in raw water after rainfalls. More recently, Sobrinho et al. (2001) found that *Giardia* and *Cryptosporidium* (oo)-cysts occurrences were higher in event-based samples in comparison to non-event periods in the Grand River (Ontario, Canada). In this river, *Giardia* concentrations correlated with the raw water turbidity, and the highest concentrations of *Giardia* and *Cryptosporidium* were obtained in cold water (0-6°C). In sand filters (after chemical coagulation), *Giardia* and *Cryptosporidium* removal is correlated with turbidity removal (LeChevallier and Norton,1992), but (oo)-cysts are probably poorly removed by zero-coagulant filtration. This would justify any concerns about high (oo)-cysts concentration during high turbidity events for the filtration plants such as the ones studied here. Unfortunately, the presence of (oo)-cysts was not investigated in treated water during the course of this study.

Aerobic spore-forming bacteria is very useful as an indicator for evaluating treatment performances. Such bacteria may be naturally present or artificially added to water and therefore represent a microbiological tracer for physical removal and chemical disinfection (Toenniessen and Johnson 1972; Rice et al. 1996; Barbeau et al. 2000; Trimboli et al. 2001). In this study, aerobic spore-forming bacteria were naturally present and used in parallel to HPC to monitor the microbial quality in treated and distributed water. They revealed very interesting features due to their much higher resistance to disinfection compared to HPCs: during turnover, their concentration in raw water was high enough to detect them all along the treatment train and in the distribution system (in contrast to *Clostridium perfringens* spores). A very strong correlation was

found between aerobic spore-forming bacteria in raw and treated water and raw water turbidity. In our study, aerobic spore-forming bacteria cannot be completely removed by filtration without chemical assistance (0.62-0.70 Log) and are but little affected by the disinfection process at the treatment plant, while their inactivation is continued inside the distribution system due to additional contact time with chlorine. The total removal between raw water and distribution system sampling points is in the range of 92-96% (1.1-1.4 Log) during turnover periods, while it is higher than 99% (> 2 Log) during reference periods.

Such monitoring of aerobic spore-forming bacteria in distribution systems has not been widely used until now, due to (i) the relatively recent focus on such microorganisms in water treatment and (ii) their concentration, which is often too low in treated water when a conventional treatment is performed (Nieminski and Bellamy 2000). In the case of this study, and probably for most water systems relying on unfiltered surface water, aerobic spore-forming bacteria are a useful indicator of the penetration of microorganisms into the distribution system up to the consumer's tap. Due to the presumed absence of regrowth of such aerobic spore-forming bacteria in treatment facilities and distribution systems, this indicator might indicate a treatment breakthrough or contamination in the distribution systems (pathogen intrusion), and is supports the evidence of other indicators (HPCs, bacterial biomass, biodegradable organic matter) which are actually used to evaluate the water biostability in distribution systems (Mathieu et al. 1995; Servais et al. 1995; Prévost et al. 1998).

The question of potential health risks associated with the turbidity of drinking water has been raised by Morris et al. (1996), who suggested that the increases in turbidity were associated with increases in gastroenteritis illnesses (GI) in Milwaukee (USA), even outside the large waterborne disease outbreak period in late March/early April 1993. Corresponding turbidity values (two week average) outside of the outbreak period were in the range of 0.2-0.5 NTU (Morris et al. 1996). In another study, the fecal contamination of karstic aquifers was suspected since relatively small changes in water turbidity in Le Havre (France) were correlated with subsequent increases in anti-diarrheal drug sales (Beaudeau et al. 1999). Moreover, examining time series data from Vancouver (Canada) - where surface water from a protected watershed is simply chlorinated before distribution - statistically significant turbidity-GI relationships were recently found (Aramani et al. 2000). For most examined cases (i.e. considering several population groups and corresponding water supplies), GI relative rates decreased as turbidity fell below the value of 1 NTU, underscoring that even quite low turbidity values should be studied so that any potential health effects may be known.

These results, and the findings presented in this study, emphasize the need to carefully monitor turbidity variations, especially for surface waters receiving little or no filtration treatment before being distributed (Hunter, 2001). The brevity of certain turbid events and the use of "typical" bacterial indicators for microbial contamination may cause an

inaccurate evaluation of the treatment barrier efficacy and therefore the health risk for certain populations would be underestimated, which is always difficult to evaluate at the endemic level. The nature of turbidity particles is also to be considered important since mineral, organic, and biological amounts may vary greatly over time and according to water treatment, and since pathogen concentrations may not be related to the total particle counts or turbidity. For example, Aramani et al. (2000) hypothesized that a higher proportion of organic matter in turbidity particles could explain the higher GI rate observed for one water supply, by comparison with other supplies with similarly high turbidities.

More studies are also required to better understand the origins of particles found in raw water – and downstream in the water system - during turnover events. In the case of the St. Lawrence River at the Montreal intake location, a combination of factors may be hypothesized, such as (i) the increased flow due to the spring snow melting or heavy rainfalls, (ii) the turnover of water masses in upstream lakes when temperatures reach the 4°C value which correspond to the maximum water density; (iii) sediment resuspension in shallow areas following windy periods and (iv) the temporary increased waterflow from the more turbid Outaouais River which is a major tributary of the St. Lawrence River just upstream from Montreal. Better identification of the main sources for turbidity will provide insights into the origin of the associated microorganisms and thus into the importance of the associated risks. This would help to better define the necessary action to be taken for watershed protection and also for water treatment optimization, to ensure a constant quality of distributed water all year round.

Conclusion

This study allowed for an estimation of the quantity and nature (chemical and microbiological) of suspended matter in treatment plant effluents and in DS's during a turbidity peak. Even if concentrations of suspended matter remained much lower than those of dissolved matter, suspended particles are of particular concern due to (i) the potential protection they can offer microorganisms versus disinfection, and (ii) the higher microbial load, which is usually associated with turbidity events resulting from rainfalls. For the water system that was studied, it was shown herein that:

- during turbidity peaks, the concentration and composition of suspended matter were similar throughout various points in the distribution system, indicating a low or moderate impact of the DS on suspended solids which are released by the treatment plants during turnovers ;
- a higher microbial load may be associated with this cold water turbidity peak (measured here especially for aerobic spore-forming bacteria). In turnover conditions, only microorganisms which are (i) retained by sand filters (operated without any chemicals) or (ii) susceptible enough to cold water disinfection, will be

eliminated by the treatment trains. For other microorganisms, the efficacy of the treatment barrier may be in question;

- typical microbial indicators of contamination were not found in either treated water nor in distribution systems during turnover events. This may be due to the origin of the turbidity peak, which is probably the result of sediment resuspension or a change in the ratio between the more turbid Outaouais and St. Lawrence River waterflow, rather than to runoff following heavy rainfalls.

These conclusions emphasize the need for a better understanding of the role of transient turbidity events on the increase in microbial load of the St. Lawrence River at the Montreal water intake. A careful monitoring of both raw and treated water should be continued, especially during transitory water quality events, since it is probable that microorganisms present in treated water will be transported throughout the DS. It is also very important to determine the risk posed by pathogens which could potentially be associated with these turbidity spikes specific to the Montreal water intake. A better appreciation of the microbial charge at the intake location would be provided by identifying the parameters which govern the occurrence of turbidity peaks, and the origins of suspended particles found in raw water.

Acknowledgements

This research was completed in collaboration with the City of Montreal and supported by the partners of the NSERC Industrial Chair on Drinking Water. R. Plante (CUM) performed the *Clostridium perfringens* spores numeration. The authors would also like to thank J. Baudart, A. Champagne, C. Dallaire, J. Mailly, C. Meunier and P. Simard for their valuable assistance in sampling and analytical work.

References

Alere, I. & Hanæus, J. 1997. Particle dynamics in the drinking water distribution network of Luleå. Vatten **53**(4): 381-390.

Atherholt, T. B., LeChevallier, M. W., Norton, W. D. & Rose, J. S. 1998. Effect of rainfall on *Cryptosporidium* and *Giardia*. J. Am. Water Wks. Assoc. **90**(9) : 66-81.

Aramini, J., McLean, M., Wilson, J., Holt, J., Copes, R., Allen, B. & Sears, W. 2000. Drinking water quality and health care utilization for gastrointestinal illness in greater Vancouver. Health Canada, 78p. http://www.hc-sc.gc.ca/ehp/ehd/catalogue/bch_pubs/vancouver_dwq.htm.

Barbeau, B., Boulos, L., Desjardins, R., Coallier, J., Prévost, M. & Duchesne, D. 1997. A modified method for the enumeration of aerobic spore-forming bacteria. Can. J. Microbiol. **43**: 976-980.

Barbeau, B., Boulos, L., Desjardins, R., Coallier, J. & Prévost, M. 1999. Examining the use of aerobic spore-forming bacteria to assess the efficiency of chlorination. Wat. Res. **33**(13): 2941-2948.

Barbeau, B., Payment, P., Coallier, J., Clément, B. & Prévost, M. 2000. Evaluating the risk of infection from the presence of *Giardia* and *Cryptosporidium* in drinking water. Quantitative Microbiology **2**: 37-54.

Beaudeau, P., Payment, P., Bourderont, D., Mansotte, F., Boudhabay, O., Laubiès, B. & Verdière, J. 1999. A time series study of anti-diarrheal drug sales and tap-water quality. Int. J. Env. Health Res. **9**: 293-311.

Berman, D., Rice, E.W. & Hoff, J.C. 1988. Inactivation of particle-associated coliforms by chlorine and monochloramine. Appl. Environ. Microbiol. **54**(2): 507-512

Bisson, J.W. & Cabelli, V.J. 1979. Membrane filter enumeration method for *Clostridium perfringens*. Appl. Environ. Microbiol. **37**: 55-66.

Boyce, D. S. 1981. The effect of bentonite clay on ozone disinfection of bacteria and viruses. Water Res. **15**: 759-767.

Desjardins, R., Jutras, L., & Prévost, M. 1997. Water quality in Montréal: effect of the distribution system. Rev. Sci. Eau. **10**(2): 167-184.

Environnement Québec 1984. Règlement sur la qualité de l'eau potable. Gazette officielle du Québec, Décret 1158-84: 2123-2129.

Environnement Québec 2001. Règlement sur la qualité de l'eau potable. 27p.

Ferguson, A.M.D. & Neden D.G. 2001. Greater Vancouver's drinking water treatment program. Can J. Civ. Eng. **28**(suppl. 1): 36-48.

Finch, G.R. & Choe K. 1999. Using *Bacillus* spores as surrogates for ozone inactivation of *Cryptosporidium.* Proc. 14th Ozone World Congress, Int. Ozone Assoc., Dearborn, Mi, USA.

Gauthier, V., Portal, J.M., Rosin, C., Block, J.C., Cavard, J., & Gatel, D. 1997. How good are distribution systems for transport of particulate matter? Proc. Water Quality Technol. Conf. of Amer. Water Wks Assoc., Denver, Co, USA, 18 p.

Gauthier, V., Rédercher, S. & Block, J.C. 1999a. Chlorine inactivation of *Sphingomonas* cells attached to goethite particles in drinking water. Appl. Environ. Microbiol. **65**(1): 355-357.

Gauthier, V., Gérard, B., Portal, J.M., Block, J.C., & Gatel, D. 1999b. Organic matter as loose deposits in drinking water distribution systems. Water Res. **33**(4): 1014-1026.

Gauthier, V., Barbeau, B., Millette, R., Block, J.-C. & Prévost, M., 2001. Suspended particles in the drinking water of two distribution systems. Water Supply-Water Science and Technology **1**(4): 237-245.

Hunter, P. 2001. Possible undetected outbreaks of *Cryptosporidiosis* in areas of the North West of England supplied by an unfiltered surface water source. Commun. Dis. Public Health **4**(2): 136-138.

Jones, J.G., & Tuckwell, S.B. 1993. Aesthetic aspects of drinking water quality - what do our customers want? Water Supply **11**(3/4): 37-51.

LeChevallier, M.W., Evans, T.M. & Seidler, R.J. 1981. Effect of turbidity on chlorination efficiency and bacterial persistence in drinking water. Appl. Environ. Microbiol. **42**(1): 159-167.

LeChevallier, M.W., & Norton, W.D. 1992. Examining relationships between particle counts and *Giardia*, *Cryptosporidium* and turbidity. J. Am. Water Wks. Assoc. **84**(12): 54-60.

Lupi, E., Ricci, V. & Burrini, D. 1995. Recovery of bacteria in nematodes isolated from a drinking water supply. J. Water Supply. Res. Technol. - Aqua **44**(5): 212-218.

Mathieu, L., Block, J.C., Prévost, M., Maul, A. & DeBischop, R. 1995. Biological stability of drinking water in the city of Metz distribution system. J. Water Supply. Res. Technol. - Aqua **44**(5): 230-239.

McCoy, W.F. & Olson, B.H. 1986. Relationship among turbidity, particle counts and bacteriological quality within water distribution lines. Water Res. **20**(8): 1023-1029.

Morin, P., Gauthier, V., Saby, S. & Block, J.C. 1999. Bacterial resistance to chlorine through attachement to particles and pipe surfaces in drinking water distribution systems. *In* "Biofilms in the Aquatic Environment", *eds.* C.W. Keevil, A.F. Godfrey, D.M. Holt and C.S. Dow, Royal Society of Chemistry, Cambridge, UK, pp. 171-190.

Morris, R.D., Naumova, E.N., Levin, R. & Munasinghe, R.L. 1996. Temporal variations in drinking water turbidity and diagnosed gastroenteritis in Milwaukee. Am. J. Publ. Health **86**(2): 237-239.

Nieminski, E.C. & Bellamy, W.D. 2000. Application of surrogate measures to improve treatment plant performance. AWWARF report #90811, Am. Water Wks. Assoc. Denver, Co, USA.

Payment, P., Aminata, B., Prévost M., Ménard B. & Barbeau B. 2000. Occurrence of pathogenic microorganisms in the St Lawrence River (Canada) and comparison of health risks for populations using it as their source of drinking water. Can. J. Microbiol. **46**: 565-576.

Pinot, O. 2000. Impact d'une vidange orientée sur la qualité de l'eau dans la zone 4 de la Ville de Montréal, Report, NSERC Ind. Chair on Drinking Water, Ecole Polytechnique de Montreal, 65p.

Prévost, M., Rompré, A., Coallier, J., Servais, P., Laurent, P., Clément, P. & Lafrance, P. 1998. Suspended bacterial biomass and activity in full-scale drinking water distribution systems: impact of water treatment. Wat. Res. **32**(5): 1393-1406.

Rice, E.W., Fox, K.R., Miltner, R.J., Lytle, D.A. & Johnson, C.H. 1996. Evaluating plant performance with endospores. J. Am. Water Wks. Assoc. **88**(9): 122-130.

Schoenen, D. & Hoyer, O. 2000. Contamination of drinking water with coliform organisms by the larves of gnats in the water treatment plant. Acta hydrochim. hydrobiol. **28**: 47-51.

Servais, P., Laurent, P. & Randon, G. 1995. Comparison of the bacterial dynamics in various French distribution systems. J. Water Supply. Res. Technol. - Aqua **44**(1): 10-17.

Sobrinho, J.A.H., Rosen, J.S., LeChevallier, M.W., Frey, M.M. & Clancy, J.M. 2001. Variability of pathogens and indicators in source waters. Proc. Water Quality Technol. Conf. of Amer. Water Wks Assoc., Nashville, Tn, USA, 12p.

Standard methods for the examination of water and wastewater 1995. 19th edition, American Public Health Association/American Water Works Association/Water Environment Federation, Washington, DC, USA.

Toenniessen, G.H. & Johnson, J.D. 1970. Heat shocked *Bacillus subtilis* spores as an indicator of virus disinfection. J. Am.Water Works Assoc. **62**: 589-593.

Trimboli, P., Lozier, J. & Johnson, W. 2001. Demonstrating the integrity of a large scale microfiltration plant using a Bacillus spore challenge test. Water Supply-Water Science and Technology **1**(5-6): 1-12.

USEPA, 1998. Code of Federal Regulations, 40, Part 141—National primary drinking water regulations, http://www.epa.gov/safewater/regs.html.

ANNEXE 2

Synthèse de la revue des applications examinées

par Maier et Dandy (2000b)

Tableau A2.1: Information de base sur les articles examinés par Maier et Dandy (2000b)

Background information				Data					
Ref.	Author(s) and year	Variable	Location(s)	Time step	Forecast length	Data type	Normalisation range	No. train. samples	No. test samples
1	Whitehead et al., 1997	Algal conc.	River Thames (England)	week	?	Real	?	125	31
2	Recknagel et al., 1997	Algal conc.	Lakes in Japan and Finland, River Darling (Australia)	day	+ 1	Real	?	2191–3653	730
3	Yabunaka et al., 1997	Algal conc.	Lake Kasumigaura (Japan)	day	+ 7	Real	?	365	4015
4	Recknagel, 1997	Cyanobact. conc.	Lake Kasumigaura (Japan)	day	+ 1	Real	?	2922	730
5	Maier and Dandy, 1997b	Cyanobact. conc.	River Murray (Australia)	week	+ 2	Real	?	468	52
6	Maier et al., 1998	Cyanobact. conc.	River Murray (Australia)	week	+ 4	Real	?	368	28
7	Crespo and Mora, 1993	Flow	Pisuena River (Spain)	10 days	0	Real	− 1–1	402, 201	0, 201
8	Karunanithi et al., 1994	Flow	Huron River (USA)	day	0	Synth/Real	+ 100	4748	731
9	Hsu et al., 1995	Flow	Leaf River (USA)	day	+ 1	Real	0.1–0.9	365	1826
10	Lorrai and Sechi, 1995	Flow	Araxisi River (Italy)	month	0	Real	?	120	240
11a	Smith and Eli, 1995	Flow (single)	N/A	?	0	Synth	0.1–0.9	250	250
11b	Smith and Eli, 1995	Flow (multiple)	N/A	?	0	Synth	0.1–0.9	750	250
12	Raman and Sunilkumar, 1995	Flow	Reservoirs in India	month	+ 1	Real	0–1	10	2
13	Minns and Hall, 1996	Flow	N/A	hour	0	Synth	0–1	764	794
14	Poff et al., 1996	Flow	Little Patuxent and Independence Rivers (USA)	day	0	Real	?	1096	1096
15	Clair and Ehrman, 1996	Flow, Carb., Nit.	Canadian Rivers	year	0	Real	?	85%	15%
16	Shamseldin, 1997	Flow	Catchments in Nepal, China, Ireland, USA, Australia	day	0	Real	0.1–0.85	1461–2922	365–730
17	Tawfik et al., 1997	Flow	River Nile	day	0	Real	0.05–0.95	70–144	70–144
18	Muttiah et al., 1997	Flow	US River Basins	N/A	N/A	Real	?	75–1000	47–559
19	Sureerattanan and Phien, 1997	Flow	Mae Klong River (Thailand)	day	+ 1	Real	0.05–0.95	1095–2190	365–1460
20a	Dawson and Wilby, 1998	Flow	River Mole (UK)	15 min	+ 24	Real	0–1	9600	9600
20b	Dawson and Wilby, 1998	Flow	River Amber (UK)	15 min	+ 24	Real	0–1	2761	1321
21a	Fernando and Jayawardena, 1998	Flow	Kaminonsha (Japan)	hour	+ 1	Real	?	146	588
21b	Fernando and Jayawardena, 1998	Flow	Kaminonsha (Japan)	hour	+ 1	Real	?	146	588
22a	Jayawardena and Fernando, 1998	Flow	Kaminonsha (Japan)	hour	+ 1	Real	?	146	588
22b	Jayawardena and Fernando, 1998	Flow	Kaminonsha (Japan)	hour	+ 1	Real	?	146	588
23a	Thirumalaiah and Deo, 1998a	Flow	Bhasta River (India)	hour	+ 3	Real	?	560	162
23b	Thirumalaiah and Deo, 1998a	Flow	Bhasta River (India)	hour	+ 3	Real	?	560	162
23c	Thirumalaiah and Deo, 1998a	Flow	Bhasta River (India)	hour	+ 3	Real	?	560	162
24	Golob et al., 1998	Flow	Soca River (Slovenia)	hour	+ 2 – + 6	Real	?	3287	1700

Suite du tableau A2.1: Information de base sur les articles examinés par Maier et Dandy (2000b)

Background information				Data					
Ref.	Author(s) and year	Variable	Location(s)	Time step	Forecast length	Data type	Normalisation range	No. train. samples	No. test samples
25	French et al., 1992	Rainfall	N/A	hour	+ 1	Synth	?	1000	500
26	Allen and le Marshall, 1994	Rainfall	Melbourne (Australia)	day	+ 1/2	Real	0–1	1997	665
27	Goswami and Srividya, 1996	Rainfall	India	year	+ 2– + 15	Real	?	50, 100	15
28	Hsu et al., 1997	Rainfall	Japan, Florida (USA)	hour	0	Real	0–1	?	?
29	Miller, 1997	Rainfall	Western Pacific	?	0	Real	?	300	33000
30	Venkatesan et al., 1997	Rainfall	India	year	0	Real	0–1	49	7
31	Xiao and Chandrasekar, 1997	Rainfall	Central Florida (USA)	min	0	Real	0–1	?	?
32	Tsintikidis et al., 1997	Rainfall	Western Pacific	N/A	0	Synth/Real	0.1–0.9	~ 733	~ 240
33	Chow and Cho, 1997	Rainfall	Hong Kong	hour	+ 1/2	Real	?	96	144
34a	Loke et al., 1997	Rainfall	Denmark	min	N/A	Real	?	1032	4560
34b	Loke et al., 1997	Runoff coeff.	Catchments in Europe and America	N/A	N/A	Real	?	35	7
35	Kuligowski and Barros, 1998a	Rainfall	Mount Carmel (USA)	6 hour	+ 1	Real	0–1	11344	873
36	Kuligowski and Barros, 1998b	Rainfall	Youghiogheny River and Swatare Creek basins (USA)	6 hour	+ 1– + 4	Real	?	649	162
37	Kuligowski and Barros, 1998c	Rainfall	Mid-Atlantic Region (USA)	6 hour	N/A	Real	?	3688	922
38a	DeSilets et al., 1992	Salinity (bottom)	Chesapeake Bay (USA)	N/A	N/A	Real	0–1	395–2712	78–523
38b	DeSilets et al., 1992	Salinity (total)	Chesapeake Bay (USA)	N/A	N/A	Real	0–1	3924–36258	1171–7133
39	Maier and Dandy, 1996b	Salinity	River Murray (Australia)	day	+ 14	Real	?	1461	365
40a	Bastarache et al., 1997	Salinity	Moose Pit Brook (Canada)	day	0	Real	?	285	32
40b	Bastarache et al., 1997	Salinity	Pine Martin Brook (Canada)	day	0	Real	?	356	39
40c	Bastarache et al., 1997	pH	Moose Pit Brook (Canada)	day	0	Real	?	285	32
40d	Bastarache et al., 1997	pH	Pine Martin Brook (Canada)	day	0	Real	?	356	39
41	Yang et al., 1996	Water table level	N/A	day	+ 1	Synth	?	2392	2392
42	Shukla et al., 1996	Water table level	N/A	N/A	N/A	Synth	?	26140	26140
43a	Thirumalaiah and Deo, 1998b	Water level	River Godavari (India)	day	+ 1, + 2	Real	− 0.5–0.5	800	295
43b	Thirumalaiah and Deo, 1998b		River Godavari (India)	day	+ 1, + 2	Real	− 0.5–0.5	800	295
43c	Thirumalaiah and Deo, 1998b		River Godavari (India)	day	+ 1, + 2	Real	− 0.5–0.5	800	295

?, not specified; N/A, not applicable.

Tableau A2.2 : Détails méthodologiques des articles examinés

	Network architecture			Optimisation algorithm							
Ref.[a]	Connect. type	Geometry method	Optimum I-H1-H2-O	Optim. method	Param. method	Transfer function	Learn. rate	Momentum	Epoch size	Initial weights	Stopping criterion
1	FF	?	?	BP	?	?	?	?	?	?	?
2	FF	T&E	?	BP	T&E	HT	0.1–0.9	0.05–0.6	?	?	FI
3	FF	T&E	10-50-0-5	BP	T&E	?	0.15–0.4	?	?	?	CV
4	FF	?	?	BP	?	HT	?	?	?	?	?
5	FF	F	102-120-40-1	BP	?	HT	?	?	?	?	CV
6	FF	T&E	20-17-0-1	BP	T&E	HT	0.004	0.6	16	?	FI
7	FF	F	3-3-2-1	BP	?	HT	?	?	?	?	?
8	FF	CC	15-0.34-0-1	QP	N/A	Log	N/A	N/A	?	− 1, 1	TE
9	FF	LLSSIM	9-3-0-1	LLSSIM	N/A	Log	N/A	N/A	?	?	?
10	FF	F	17-17-17-1	BP	F	Log	0.5	0.9	?	− 0.5, 0.5	FI
11a	FF	T&E	25-15-0-1	BP	F	Log	0.5	0.9	TS	− 0.5, 0.5	TE or FI
11b	FF	T&E	25-50-0-21	BP	F	Log	0.1	0.9	TS	− 0.5, 0.5	TE or FI
12	FF	T&E	4-?-?-2	BP	F	Log	0.5	0.9	?	?	CV
13	FF	F	18-10-0-1	BP	?	Log	?	?	?	?	FI
14	FF	?	?	BP	?	HT	?	?	?	?	
15	FF	?	?	BP	?	?	?	?	?	?	TE
16	FF	F	(1, 3, 5)-2-0-1	CG	N/A	Log	N/A	N/A	?	?	TE
17	FF	T&E	2-2-0-1	BP	F	Lin	0.95	0.1	TS	?	TE
18	FF	CC	?	QP	N/A	Log	N/A	N/A	?	?	TE
19	FF	F	5-2-0-1	BP	F	Log	0.01	0.5	TS	?	TE or FI
20a	FF	T&E	7-20-0-1	BP	F	Log	0.1	?	TS	− 0.29–0.29	FI
20b	FF	T&E	15-10-0-1	BP	F	Log	0.1	?	TS	− 0.13–0.13	FI
21a	FF	OLS	5-14-0-1	OLS	N/A	RBF	N/A	N/A	?	?	TE
21b	FF	T&E	5-6-0-1	BP	F	?	0.05	?	?	?	FI
22a	FF	T&E	6-11-0-1	OLS	N/A	RBF	N/A	N/A	?	?	TE or FI
22b	FF	T&E	6-6-0-1	BP	?	?	?	?	?	?	TE or FI
23a	FF	T&E	5-8-0-1	BP	?	?	?	?	?	?	TE
23b	FF	T&E	5-8-0-1	CG	N/A	?	N/A	N/A	?	− 0.5, 0.5	TE
23c	FF	CC	5-13-0-1	QP	?	?	N/A	N/A	?	?	TE
24	FF	T&E	14-40-0-3	BP	V	Log	V	?	TS	?	CV
25	FF	T&E	625-100-0-625	BP	F	Log	V	?	TS	− 1, 1	FI
26	FF	T&E	6-4-0-1	BP	?	Log	?	?	?	?	FI
27	FF	F	5-5-0-1	BP	?	HT	?	?	?	?	?
28	HYB	F	6-225-225-1	MCP	?	N/A	N/A	N/A	1, TS	?	?
29	FF	?	?	BP	?	?	?	?	?	?	?
30	FF	T&E	2-6-0-1	BP	?	Log	?	?	TS	− 1, 1	FI
31	FF	F	39-47-21-1	RLS	N/A	TL	N/A	N/A	?	?	CV
32	FF	T&E	4-16-0-1	BP	?	Log	V	V	?	− 1, 1	?
33	RC	F	31-21-3	Mod BP		F	HT	0.1	0.01	?	− 0.5, 0.5	FI
34a	FF	T&E	50-10-0-1	Mod BP		T&E	?	?	?	?	?	?
34b	FF	T&E	3-25-0-1	Mod BP		T&E	?	?	?	?	?	?
35	FF	T&E	20-10-0-1	BP	F	HT	0.01	0.001	TS	− 0.1, 0.1	CV
36	FF	T&E	25-11-0-1	BP	F	HT	0.05	0.005	TS	?	CV
37	FF	T&E	5-?-?-1	BP	F	HT	0.01	0.001	TS	?	CV
38a	FF	T&E	6-1-0-1 to	BP	T&E	Log	0.2–0.8	0.1–0.7	1	?	TE or FI
38b	FF	T&E	6-3-0-1	BP	T&E	Log	0.2–0.8	0.1–0.8	1	?	TE or FI
39	FF	T&E	51-45-0-1	BP	T&E	HT	0.02	0.6	?	− 0.1, 0.1	FI
40a	FF	T&E	3-20-0-1	LM	N/A	HT	N/A	N/A	?	?	?
40b	FF	T&E	3-15-0-1	LM	N/A	HT	N/A	N/A	?	?	?
40c	FF	T&E	3-10-0-1	LM	N/A	HT	N/A	N/A	?	?	?
40d	FF	T&E	3-20-0-1	LM	N/A	HT	N/A	N/A	?	?	?
41	FF	T&E	9-6-0-1	BP	?	HT	?	?	16	?	FI
42	FF	T&E	5-6-6-2	BP	F	Log	?	?	?	?	FI
43a	FF	T&E	1-3-0-2	BP	T&E	Log	0.1	0.2	TS	− 0.5, 0.5	TE
43b	FF	?	1-3-0-2	CG	N/A	?	N/A	N/A	TS	− 0.5, 0.5	TE
43c	FF	CC	1-2-0-2	QP	?	?	N/A	N/A	TS	?	TE

[a]Refer to Table 4 for details of reference.

?, not specified; BP, backpropagation; CG, conjugate gradient; CV, cross-validation; CC, cascade correlation; F, fixed; FI, fixed number of iterations; FF, feedforward; HT, hyperbolic tangent; H1, number of nodes in hidden layer 1; H2, number of nodes in hidden layer 2; HYB, hybrid; I, number of nodes in input layer; Lin, linear; LLSSIM, linear least-squares simplex; LM, Levenberg–Marquardt; Log, logistic; MCP, modified counterpropagation; Mod BP, modified backpropagation; Mod Log, modified logistic; N/A, not applicable; O, number of nodes in output layer; OLS, orthogonal least squares; QP, quickprop; RBF, radial basis function; RC, recurrent; RLS, recursive least-squares; T&E, trial and error; TE, training error; TS, training set; TL, threshold logic; V, variable.

ANNEXE 3

Identification graphique des événements turbides au cours des périodes printanières et automnales

ANNEXE 3 - IDENTIFICATION DES ÉVÉNEMENTS TURBIDES

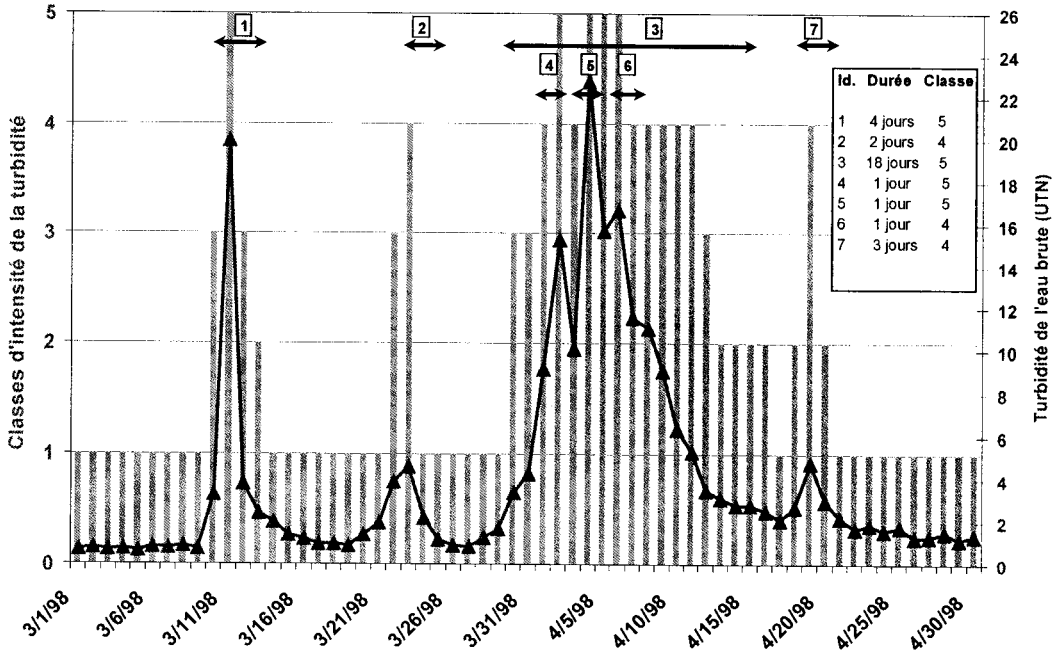


Figure A3-1 : Périodes turbides identifiées pour le printemps 1998

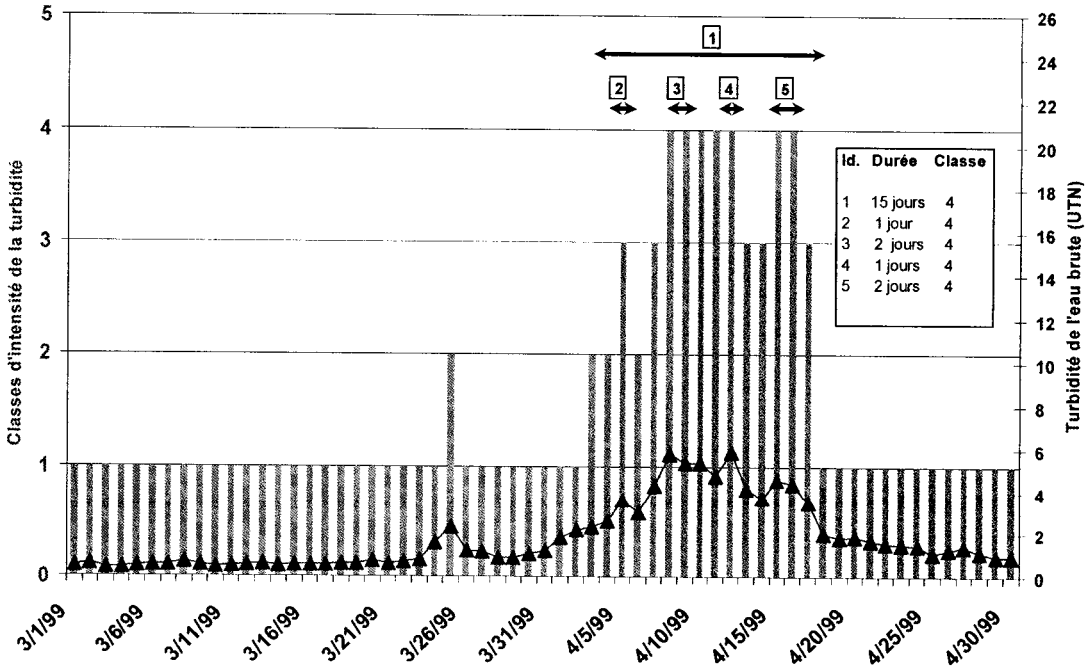


Figure A3-2 : Périodes turbides identifiées pour le printemps 1999

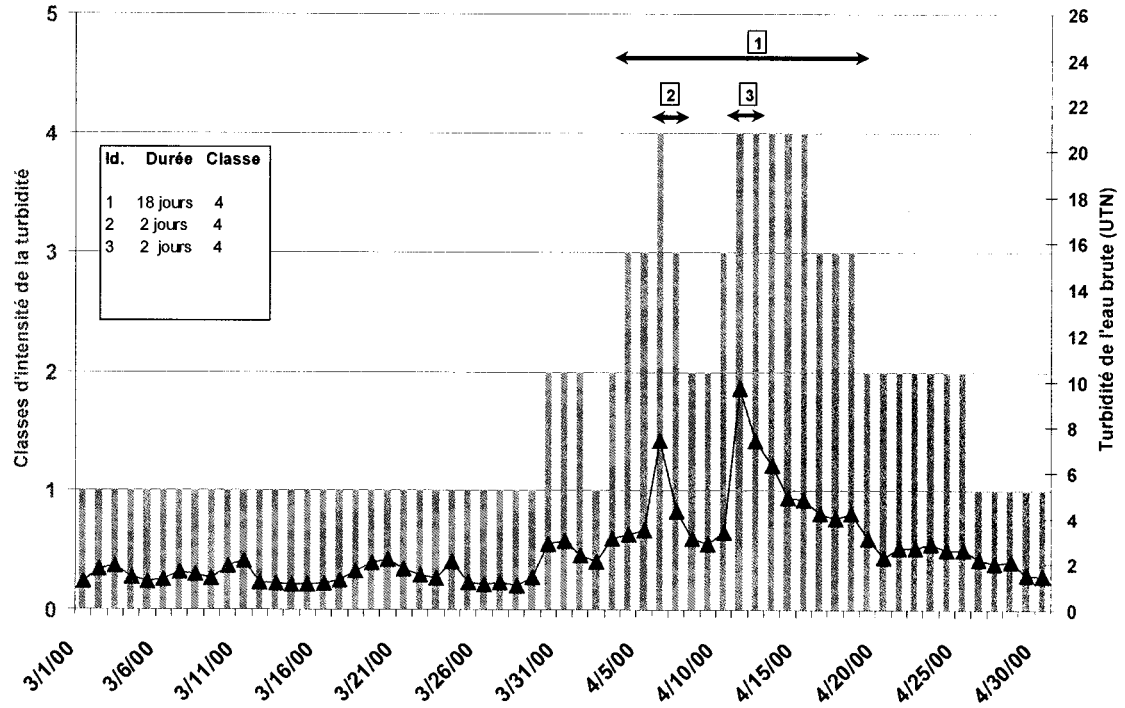


Figure A3-3 : Périodes turbides identifiées pour le printemps 2000

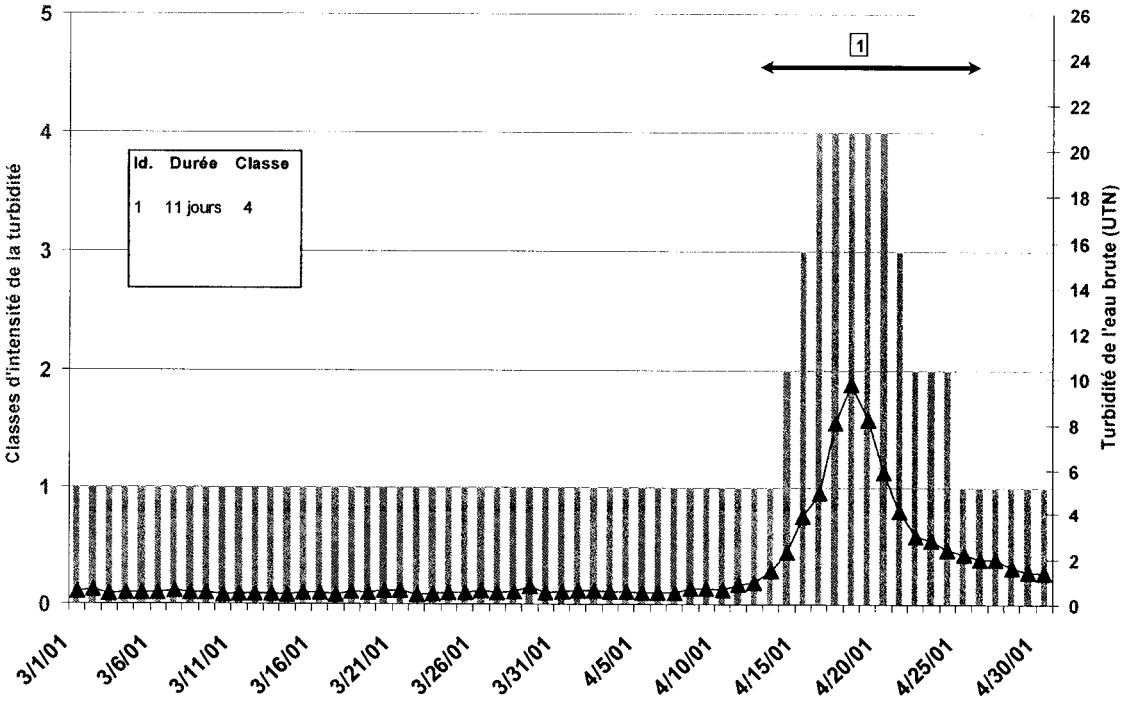


Figure A3-4 : Périodes turbides identifiées pour le printemps 2001

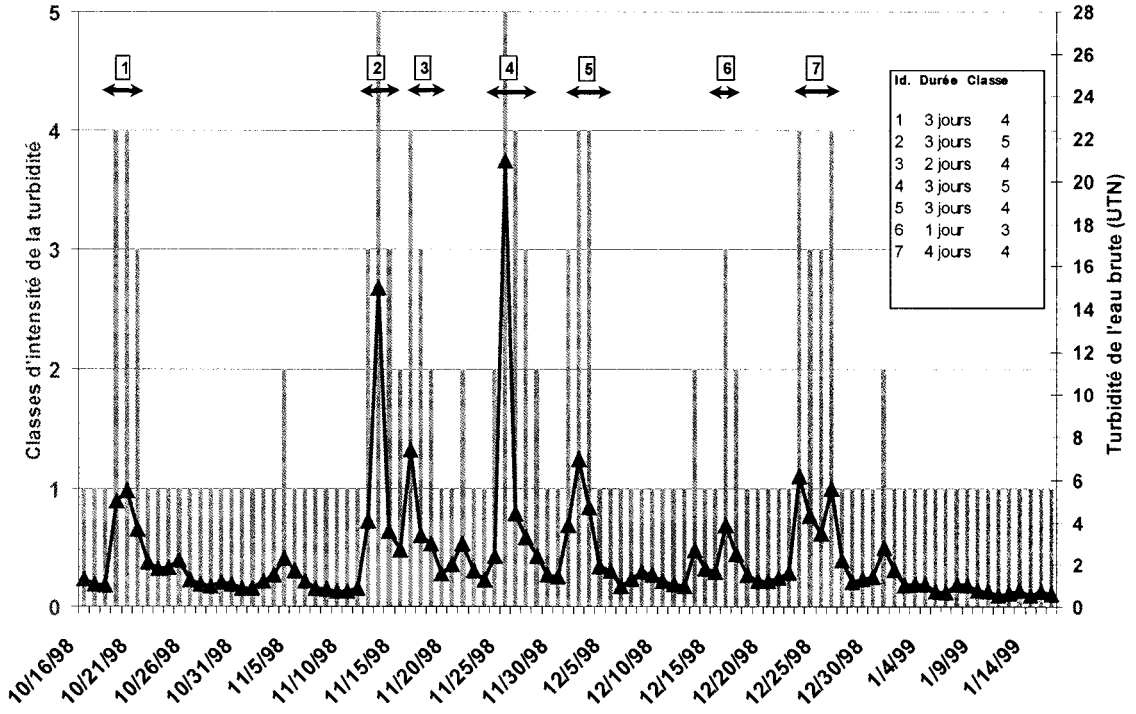


Figure A3-5 : Périodes turbides identifiées pour l'automne 1998

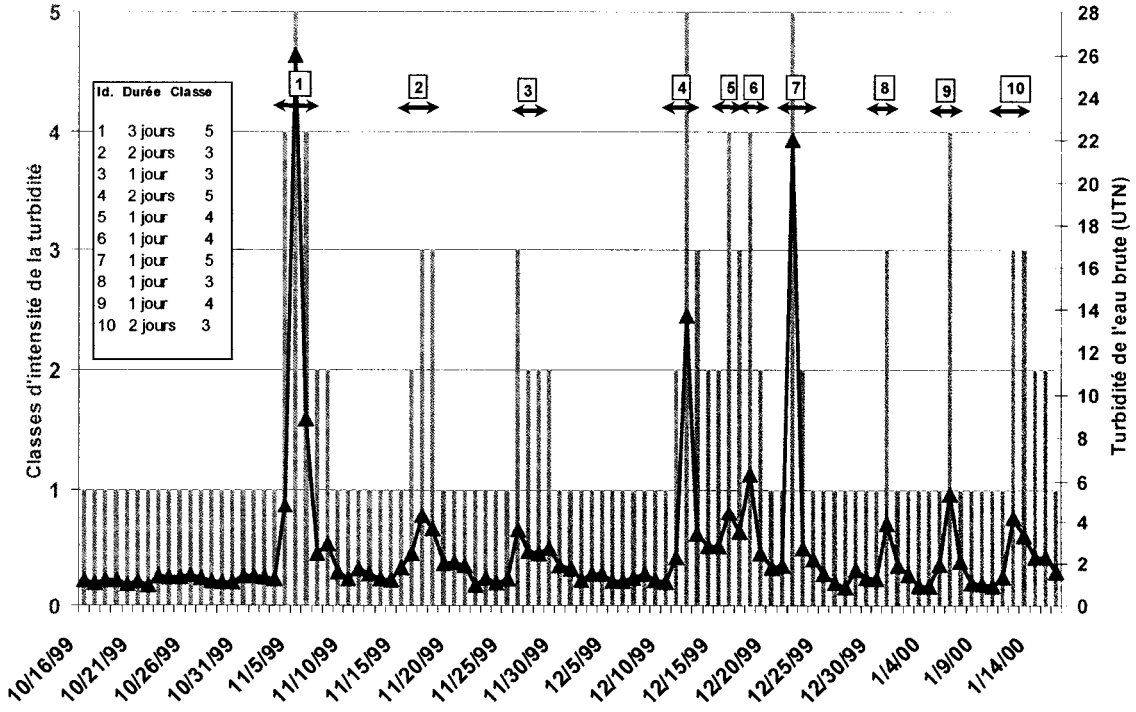


Figure A3-6 : Périodes turbides identifiées pour l'automne 1999

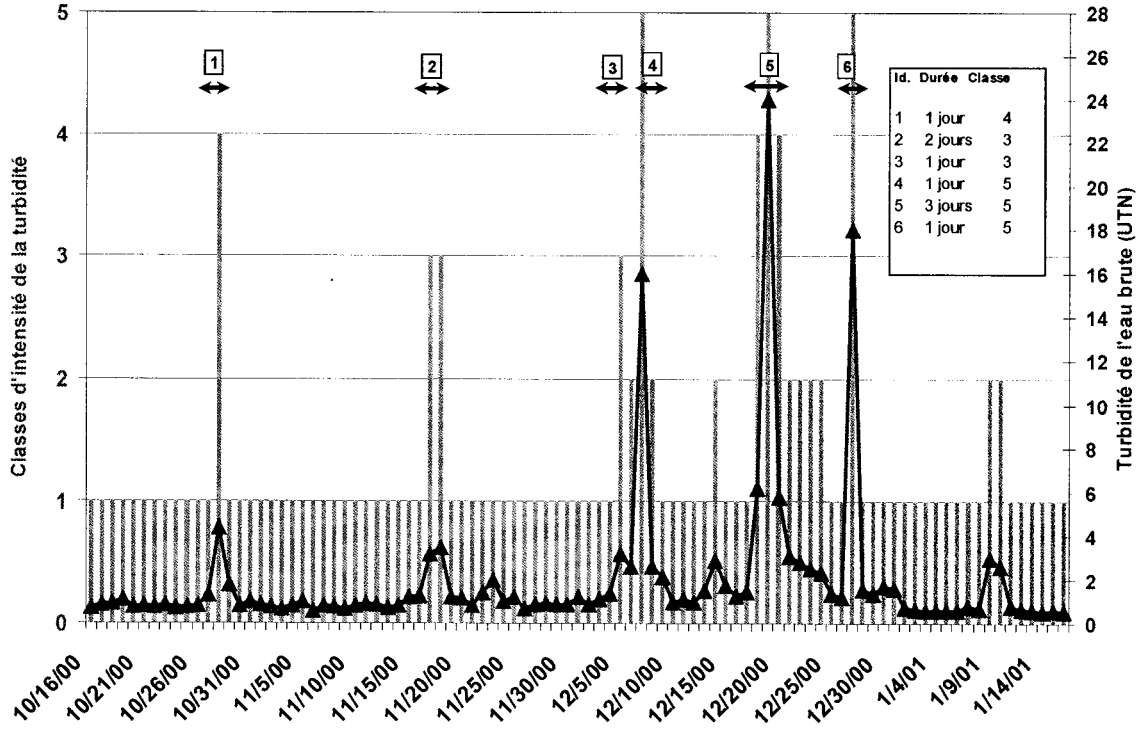


Figure A3-7 : Périodes turbides identifiées pour l'automne 2000

ANNEXE 4

Analyse statistique des variables d'entrée potentielles de la base de données préliminaire

ANNEXE 4 - ANALYSE STATISTIQUE

Tableau A4-1 : Moyenne et écart-type des variables d'entrée disponibles

Nom des variables	Annuel		Automne		Printemps		Été		Hiver	
	Moy.	Écart-type	Moy.	Écart-type	Moy.	Écart-type	Moy.	Écart-type	Moy.	Écart-type
Turbidité de l'eau brute à Coteau-du-lac	1.6	2.5	2.1	3.3	3.0	3.6	0.9	0.5	0.9	1.6
Turbidité de l'eau brute à Hawkesbury	3.5	4.9	2.5	1.2	7.6	9.0	2.7	2.7	1.5	0.4
Couleur de l'eau brute à DesBaillets	7.4	3.6	7.3	3.8	9.9	5.0	6.4	2.3	6.6	2.0
Couleur de l'eau brute à Hawkesbury	33.3	7.6	32.6	5.2	39.3	9.7	31.0	6.7	32.4	5.5
Conductivité de l'eau traitée à DesBaillets	290	12	292	5	282	22	291	5	293	5
pH de l'eau brute à DesBaillets	8.2	0.1	8.2	0.1	8.1	0.1	8.3	0.1	8.1	0.1
Température de l'eau traitée à Coteau-du-Lac	10.0	8.0	6.6	4.2	3.5	2.6	18.3	4.4	1.5	0.9
Température de l'eau brute à Hawkesbury	11.1	8.3	6.8	4.0	4.2	2.9	19.9	3.8	2.6	2.0
Température de l'air à Dorval	6.6	11.6	-0.6	8.1	2.8	6.0	17.8	4.8	-7.4	6.1
Température de l'air à Sainte-Anne-de-Bellevue	6.1	11.4	-0.9	8.1	2.4	5.9	17.0	4.7	-7.9	6.3
Température de l'air au lac St-François	5.9	11.3	-0.7	8.3	2.3	6.2	16.5	4.7	-8.2	6.8
Température moyenne de l'air à Ottawa	6.7	11.6	-0.8	8.4	3.1	6.4	17.7	4.8	-7.2	6.1
Température-degré de croissance à Dorval	5.7	6.8	1.4	6.2	1.3	2.3	12.7	4.9	0.6	7.5
Température-degré de dégel à Dorval	8.9	8.6	3.3	6.9	4.1	4.0	17.7	5.1	0.9	7.5
Hauteur d'eau à Dorval	1.8	5.0	1.4	4.4	1.1	3.6	2.7	6.3	0.6	2.3
Hauteur d'eau à Ottawa	1.9	5.1	1.6	4.6	1.2	3.6	2.8	6.4	0.6	2.5
Précipitations totales à Dorval	2.6	6.0	2.6	6.0	2.3	5.8	2.7	6.3	2.5	5.7
Précipitations totales à Ottawa	2.3	5.4	2.2	5.1	1.9	4.5	2.8	6.4	1.4	3.2
Précipitations totales à Sainte-Anne-de-Bellevue	2.4	5.7	2.5	5.9	2.2	5.2	2.8	6.3	2.3	5.3
Précipitations totales au lac St-François	2.5	5.9	2.2	5.2	2.1	5.1	3.0	7.0	2.1	4.4
Ensoleillement à Dorval	56.7	45.2	31.0	31.1	61.9	46.6	73.7	45.5	43.7	38.8
Vitesse maximale du vent à Dorval	27.2	9.5	28.4	10.7	28.7	9.4	25.9	8.3	26.7	10.0
Vitesse maximale du vent à Sainte-Anne-de-Bellevue	13.9	6.9	16.1	7.8	14.7	6.8	11.6	5.5	15.6	7.1

Suite du Tableau A4-1 : Moyenne et écart-type des variables d'entrée disponibles

Nom des variables	Annuel		Automne		Printemps		Été		Hiver	
	Moy.	Écart-type	Moy.	Écart-type	Moy.	Écart-type	Moy.	Écart-type	Moy.	Écart-type
Vitesse maximale du vent au lac St-François	21.5	8.4	23.0	9.9	23.7	7.9	19.1	6.7	22.8	8.9
Vitesse moyenne du vent à Dorval	13.7	6.6	15.8	7.2	14.7	6.8	11.7	5.6	14.5	6.6
Vitesse moyenne du vent à Sainte-Anne-de-Bellevue	10.7	5.3	12.4	5.9	11.4	5.3	8.9	4.2	11.8	5.4
Vitesse moyenne du vent au lac St-François	11.8	5.7	13.1	6.6	13.6	5.7	9.9	4.2	12.8	6.1
Direction maximale du vent à Dorval	29.3	7.7	29.7	6.8	28.7	9.2	29.3	6.8	29.1	9.0
Direction maximale du vent à Sainte-Anne	24.9	8.3	25.0	7.6	24.8	9.9	24.9	7.4	24.8	9.4
Direction maximale du vent au lac St-François	28.8	7.0	28.1	7.1	28.5	8.1	29.9	5.9	27.6	7.5
Direction moyenne du vent à Dorval	18.2	8.4	19.1	7.8	17.0	9.6	18.8	7.8	16.5	9.0
Direction moyenne du vent à Sainte-Anne-de-Bellevue	19.6	7.9	19.5	7.6	18.9	8.6	20.4	7.2	18.8	8.7
Direction moyenne du vent au lac St-François	18.9	7.4	19.5	7.4	17.6	8.1	19.8	6.6	17.2	7.9
Débit de la rivière des Outaouais au barrage Carillon	1726	909	1609	658	2715	1244	1292	553	1778	270
Débit de la rivière Outaouais par le canal Sainte-Anne	511	314	444	238	857	407	359	190	573	106
Débit de la rivière Outaouais par le chenal de l'Île Perrot	309	242	256	150	571	370	205	125	327	65
Débit moyen de la rivière Outaouais par les canaux Sainte-Anne et Île Perrot	410	277	350	194	714	388	282	157	450	85
Débit du fleuve St-Laurent à Des Cèdres	517	389	392	334	593	477	560	326	498	450
Débit du fleuve St-Laurent s'écoulant par le canal à Beauharnois	6452	891	5897	782	6576	770	6876	808	6000	732
Débit du fleuve St-Laurent à Lasalle	8029	1282	7232	475	8886	1654	8160	1221	7787	873
Proportion des eaux de l'Outaouais dans le fleuve à Lasalle	21.3	9.2	22.2	8.7	30.1	10.4	15.9	6.1	23.1	4.0
Débit de la rivière Raisin à Williamstown	5.7	13.1	2.9	3.6	18.4	23.0	2.2	6.4	2.8	5.3
Débit de la rivière Beaudette à Glen Navis	2.0	4.3	1.1	1.0	6.4	7.8	0.9	1.7	1.0	1.3

ANNEXE 5

Critères de sélection et évaluation des paramètres indicateurs utilisés pour représenter les facteurs explicatifs

ANNEXE 5 - PRÉSENTATION DES PARAMÈTRES INDICATEURS

Tableau A5-1 : Paramètres indicateurs représentant le débit accru de l'Outaouais et l'augmentation de la contribution de l'Outaouais dans le fleuve St-Laurent

Variables sélectionnées	➢ Q_OUT : Débit de l'Outaouais au barrage Carillon ➢ %OUT : Contribution de l'Outaouais au débit total du fleuve à LaSalle, soit le rapport entre le débit total de l'Outaouais par les chenaux Sainte-Anne et Île-Perrot et le débit du fleuve à Lasalle
Critères de sélection des indicateurs	➢ Le débit de l'Outaouais au barrage Carillon est la variable de débit de l'Outaouais la plus près en amont de la prise d'eau de Montréal. Cette variable permet de prendre en compte les fluctuations saisonnières du débit total de l'Outaouais. ➢ La contribution de l'Outaouais au débit total du fleuve permet de prendre en compte les fluctuations du débit du fleuve. Si le débit du fleuve et celui de l'Outaouais augmentent de façon proportionnelle, la contribution de l'Outaouais reste la même. Par contre, si le débit du fleuve est régulé au printemps, la contribution de l'Outaouais devient plus importante.
Seuils et critères d'évaluation	➢ Les seuils à partir dequels les valeurs de débit et de contribution ont été établis à partir de la moyenne, déterminée lors de l'analyse statistique des données : o Q_OUT > 2000 m^3/s o %OUT > 20%

Tableau A5-2 : Paramètres indicateurs représentant la fonte des neiges et la présence du couvert de glace

Variables sélectionnées	➢ TD_DOR : Index de température-degré à Dorval ➢ Q_RAI : Débit de la rivière Raisin ➢ Q_BDT : Débit de la rivière Beaudette ➢ PR_LSF : Précipitations locales au lac St-François
Critères de sélection des indicateurs	➢ La fonte des neiges et le bris du couvert de glace sont deux événements à considérer car leur occurrence influence des événements identifiés dans la revue de littérature. Ainsi, la fonte des neiges est à l'origine de l'augmentation du débit des rivières au printemps. Quant au couvert de glace, sa présence limite grandement l'impact des tempêtes de vents et des précipitations importantes sur la remise en suspension des sédiments et l'érosion des berges. ➢ En combinaison, l'index de température-degré, les débits de tributaires secondaires du Lac St-François et les précipitations locales permettent de déterminer quand le couvert de glace commence à se former et quand débutent la fonte des neiges et le bris du couvert de glace : o Lorsque la température descend sous le point de congélation pendant plusieurs jours consécutifs, quand les débits cessent de fluctuer et que les précipitations locales n'ont plus aucun impact sur les débits qui sont très faibles à ce moment de l'année, alors on peut supposer que le couvert de glace commence à se former. o Au contraire, quand la température reste au-dessus du point de congélation pendant plusieurs jours consécutifs et que les débits s'accroissent soudainement, alors la fonte des neiges a débuté et il est probable que le couvert de glace se brisera prochainement.
Seuils et critères d'évaluation	➢ La prise du couvert de glace à l'automne débute quand: o TD_DOR < 0 pendant au moins dix jours o Q_RAI et Q_BDT très faibles o PR_LSF > 5mm mais Q_RAI et Q_BDT restent constants la journée même et le lendemain. ➢ La fonte des neiges et le bris du couvert de glace au printemps débutent quand: o T_DOR < 0 pendant au moins cinq jours o Q_RAI et Q_BDT s'accroissent soudainement

Tableau A5-3 : Paramètres indicateurs représentant l'inversion thermique dans le lac des Deux Montagnes

Variables sélectionnées	➢ TEMP_HAW : Température de l'eau traitée à Hawkesbury
Critères de sélection des indicateurs	➢ La température quotidienne de l'eau du lac des Deux Montagnes n'a pas pu être obtenue. Bien qu'imparfait, la température de l'eau traitée à Hawkesbury donne un bon indice de la température de l'eau du lac car Hawkesbury est une ville qui puise son eau brute de la rivière des Outaouais, une trentaine de kilomètres en amont du lac des Deux Montagnes. Bien qu'il existe des villes et des villages en amont plus près du lac, les données de température de Hawkesbury semblaient plus précises et fiables.
Seuils et critères d'évaluation	➢ TEMP_HAW = 4°C

Tableau A5-4 : Paramètres indicateurs représentant les tempêtes de vent

Variables sélectionnées	➢ DOR_MOY et DOR_MAX : Intensité moyenne et maximale du vent à Dorval ➢ LSF_MOY et LSF_MAX : Intensité moyenne et maximale du vent au lac St-François
Critères de sélection des indicateurs	➢ Des vents moyens intenses signifient qu'il y a un vent soutenu alors que des vents maximaux intenses signifient qu'il y a des bourrasques importantes. Ces deux conditions étant susceptibles de causer la formation de vagues importantes, elles sont examinées. ➢ Dorval et le Lac St-François ont été choisis car il s'agit de deux endroits en amont relativement distants l'un de l'autre, permettant ainsi d'explorer l'effet de décalage sur l'augmentation de turbidité.
Seuils et critères d'évaluation	➢ Les vents sont considérés significatifs quand : o DOR_MAX ou LSF_MAX > 35 km/h o DOR_MOY ou LSF_MOY > 25 km/h ➢ L'examen des valeurs de vent est effectué en parallèle avec l'examen des paramètres indicateurs de la prise de la glace

Tableau A5-5 : Paramètres indicateurs représentant les fortes pluies

Variables sélectionnées	➢ PR_DOR : Précipitations totales à Dorval ➢ PR_SAB : Précipitations totales à Sainte-Anne-de-Bellevue ➢ PR_LSF : Précipitations totales au lac St-François
Critères de sélection des indicateurs	➢ Les précipitations peuvent être très localisées. C'est pourquoi les précipitations de trois endroits différents sont examinées. Les précipitations à Ottawa, bien que disponibles, n'ont pas été retenues car la région d'Ottawa est trop éloignée pour affectée rapidement et directement la turbidité à la prise d'eau de Montréal. Ces données pourraient se révéler plus utiles ais dans le cas d'un modèle prédictif à long terme.
Seuils et critères d'évaluation	➢ Les précipitations sont considérées significatives quand : o PR_DOR ou PR_SAB ou PR_LSF > 5 mm ➢ L'examen des valeurs des précipitations est effectué en parallèle avec l'examen des paramètres indicateurs de la prise de la glace

ANNEXE 6

Exemple d'analyse des représentations graphiques des paramètres indicateurs

ANNEXE 6 - EXEMPLE D'ANALYSE DES REPRÉSENTATIONS GRAPHIQUES DES PARAMÈTRES INDICATEURS

Cette annexe présente en exemple les représentations graphiques des différents paramètres pour le printemps 1998 (Figures A6-1 à A6-5). Les observations, identifées directement sur les graphiques donnés en exemple, sont notés. Le Tableau A6-1 reproduit comme exemple une partie du tableau d'observation pour le printemps 1998, à paritr duquel les liens entre les facteurs explicatifs et les événements turbides sont établis.

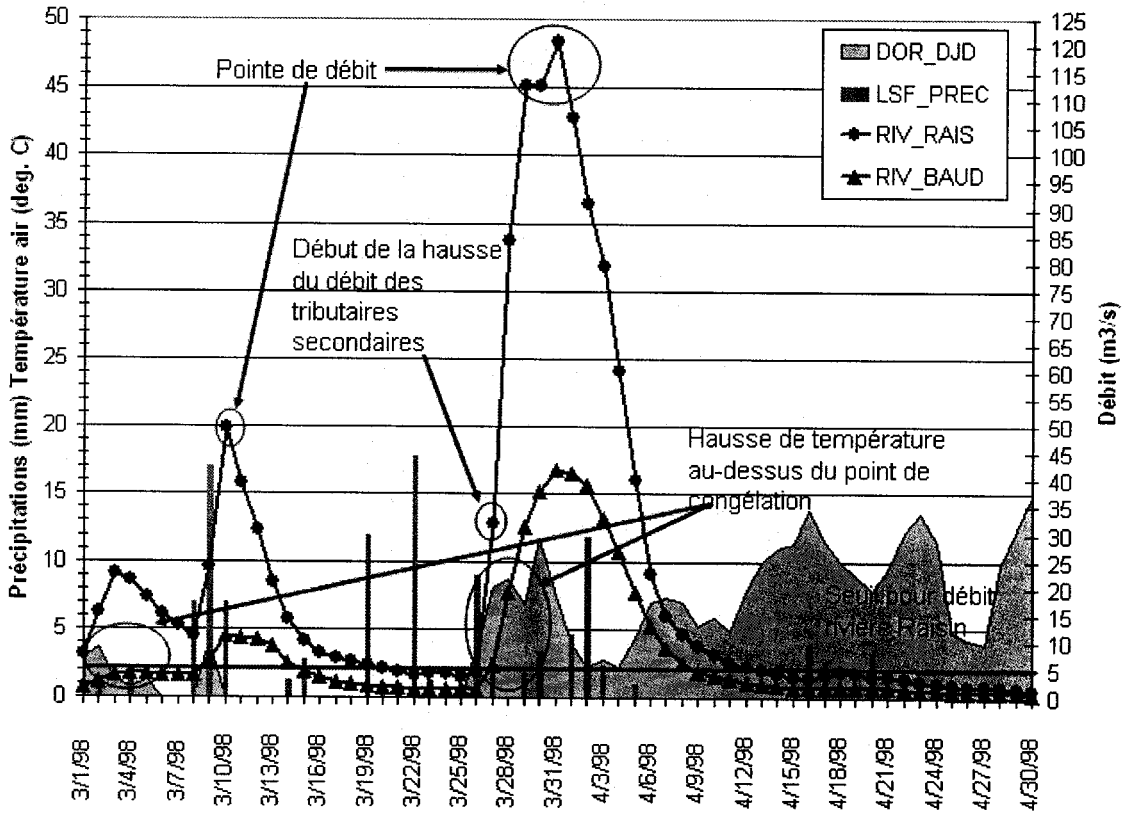


Figure A6-1 : Représentation graphique des paramètres indicateurs de la fonte des neiges

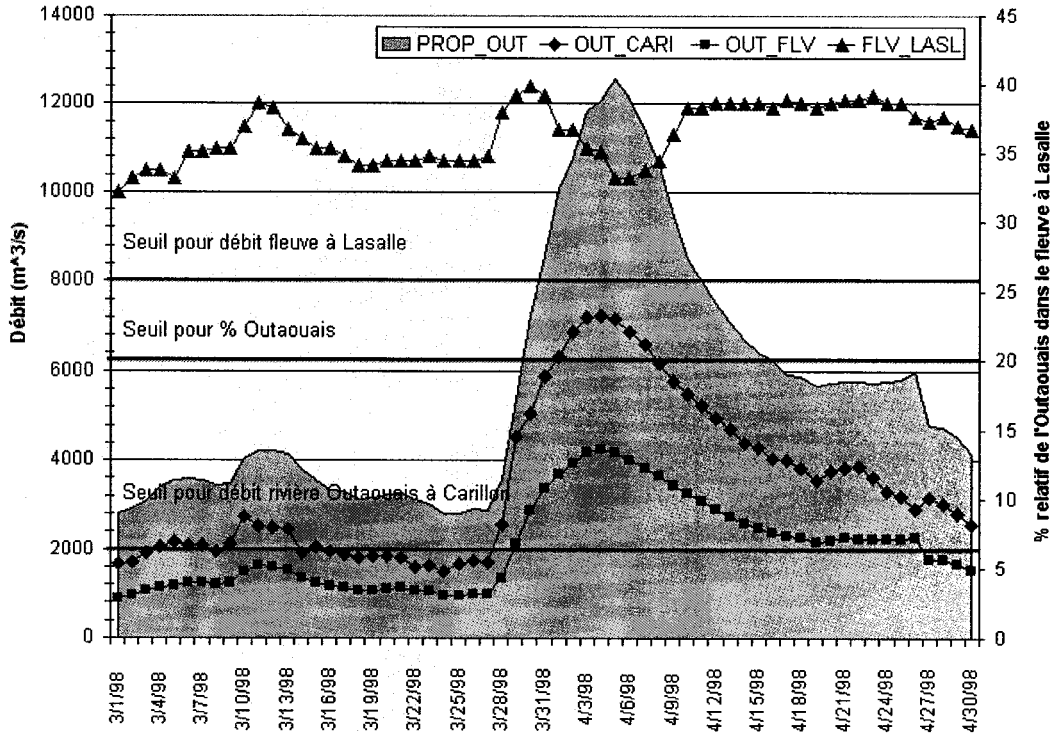


Figure A6-2 : Représentation graphique des paramètres indicateurs de débit

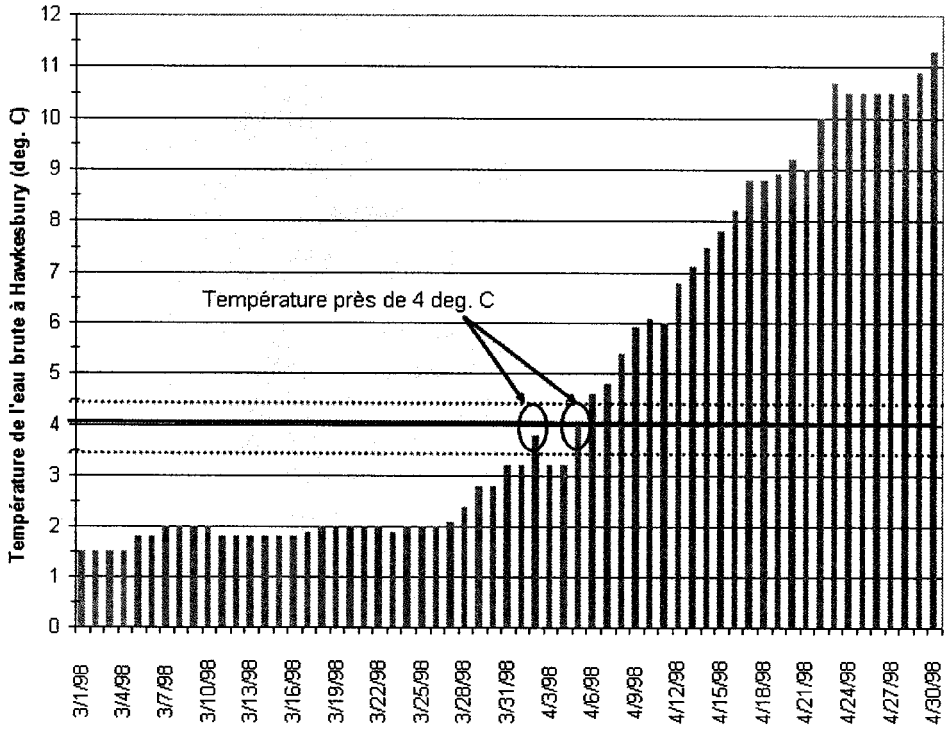


Figure A6-3 : Représentation graphique du paramètre indicateur du renversement

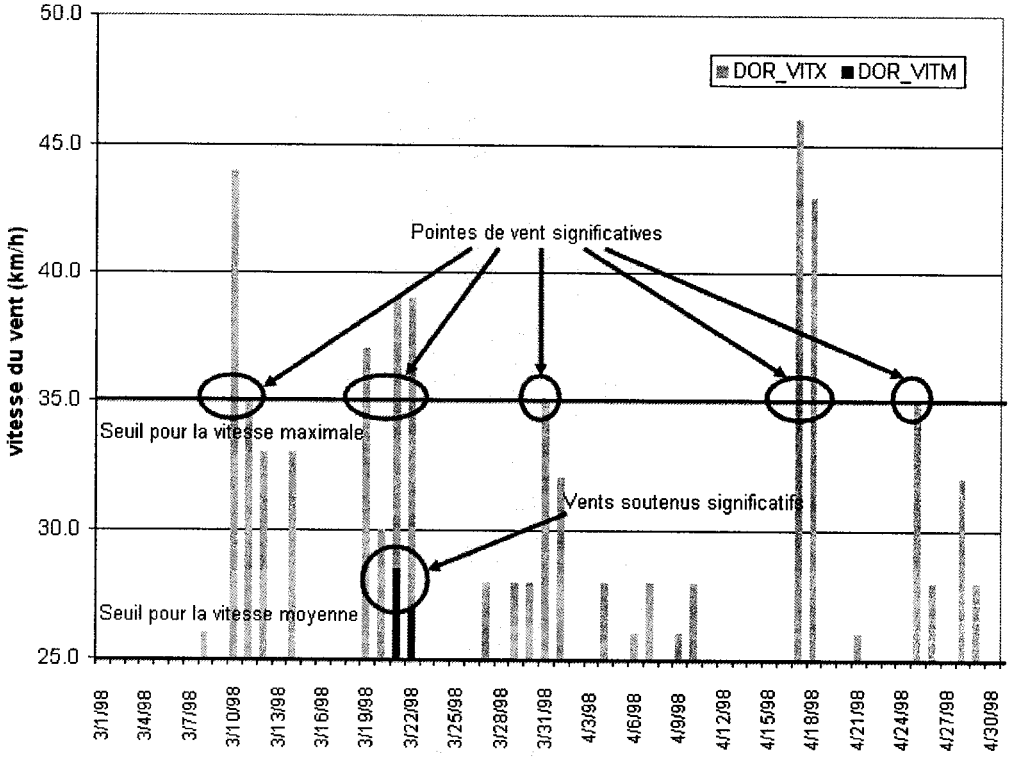


Figure A6-4 : Représentation graphique des paramètres indicateurs de vent à Dorval

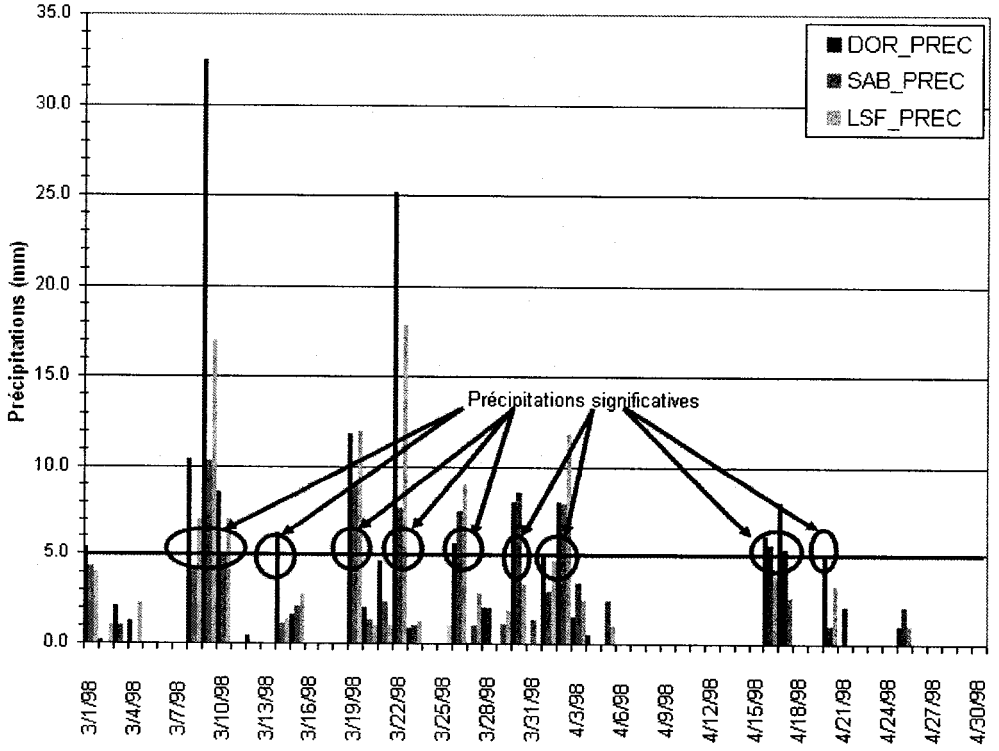


Figure A6-5 : Représentation graphique des paramètres indicateurs des précipitations

Tableau A6-1 : Échantillon des tableaux d'observation utilisés pour l'examen des représentations grapahiques des paramètres indicateurs

	Événement	VENT				PLUIE			RENVERSEMENT			DÉBITS			FONTE DES NEIGES			
	turbide	LSF MOY	LSF MAX	DOR MOY	DOR MAX	LSF	DOR	SAB	T=5	T=4	T=3	%OUT	OUT_CAR	FLV_LASL	RAI>5	BEAU>2	DJD>0	DGC>0
98-03-01														x			x	
98-03-02																	x	
98-03-03																	x	
98-03-04																	x	
98-03-05																	x	
98-03-06																		
98-03-07																		
98-03-08						x	x											
98-03-09						x	xx	x							x	x	x	
98-03-10	/ 3				xx	x	x						x	xx	xx P	xx P		
98-03-11	| 5				x								x	xx	xx P	xx		
98-03-12	\ 3				(x)								x	xx	xx P	x		
98-03-13													x		xx P	x		
98-03-14															x	x		
98-03-15																		
98-03-16																		
98-03-17																		
98-03-18																		
98-03-19																		
98-03-20			(x)		x	x	x	x										
98-03-21																		
98-03-22	/ 3	xx	xx	x	x													
98-03-23	| 4	xx	xx	x	x	xx	xx	x										
98-03-24																		

ANNEXE 7

Répartition des données en sous-ensembles

ANNEXE 7 - RÉPARTITION DES DONNÉES EN SOUS-ENSEMBLES

TableauA7-1 : Répartition des événements turbides du preintemps et de l'automne entre les trois sous-ensembles de données

	Répartition 1	Répartition 2
Ensemble d'apprentissage	**Événements turbides du printemps** Printemps 1998 (03/01/98-04/30/98) Printemps 1999 (03/01/99-04/30/99) **Événements turbides de l'automne** 10/19/98-10/21/98 11/12/98-11/15/98 11/16/98-11/17/98 11/24/98-11/29/98 11/30/98-12/03/98 12/16/98-12/17/98 11/16/99-11/18/99 11/26/99-11/29/99 12/11/99-12/14/99 12/21/99-12/23/99 12/31/99 01/06/00 11/17/00-11/18/00 12/05/00-12/06/00 12/07/00-12/08/00	**Événements turbides du printemps** Printemps 1998 (03/01/98-04/30/98) Printemps 2000 (03/01/00-04/30/99) **Événements turbides de l'automne** 10/19/98-10/21/98 11/12/98-11/15/98 11/16/98-11/17/98 12/16/98-12/17/98 12/23/98-12/26/98 11/04/99-11/08/99 12/11/99-12/14/99 12/15/99-12/16/99 12/17/99-12/20/99 01/06/00 01/12/00-01/14/00 10/28/00 12/07/00-12/08/00 12/18/00-12/24/00 12/27/00
Ensemble de test	**Événements turbides du printemps** Printemps 2000 (03/01/00-04/30/00) **Événements turbides de l'automne** 12/23/98-12/26/98 12/15/99-12/16/99 10/28/00 12/18/00-12/24/00	**Événements turbides du printemps** Printemps 2001 (03/01/01-04/30/01) **Événements turbides de l'automne** 11/24/98-11/29/98 11/16/99-11/18/99 12/21/99-12/23/99 11/17/00-11/18/00
Ensemble de validation	**Événements turbides du printemps** Printemps 2001 (03/01/01-04/30/01) **Événements turbides de l'automne** 11/04/99-11/08/99 12/17/99-12/20/99 01/12/00-01/14/00 12/27/00	**Événements turbides du printemps** Printemps 2001 (03/01/00-04/30/00) **Événements turbides de l'automne** 11/30/98-12/03/98 11/26/99-11/29/99 12/31/99-12/31/99 12/05/00-12/06/00

ANNEXE 8

Sous-ensembles de variables d'entrée des modèles testés

ANNEXE 8 - SOUS-ENSEMBLES DES VARIABLES D'ENTRÉE DES MODÈLES TESTÉS

Les variables d'entrée sélectionnées pour faire partie de l'ensemble final (présentées au Tableau 3-4) ont été réparties en sous-ensembles. Des modèles connexionnistes ont été développés pour chacun de ces sous-ensembles.

Tableau A8-1 : Sous-ensembles de variables d'entrée utilisés pour développer les modèles candidats

	0	1	2	3	4	5	6	7
Qualité	COUL_DB TURB_DB COND_DB TURB_DB1 TURB_HAW COUL_HAW	TURB_DB COND_DB TURB_DB1 TURB_HAW COUL_HAW	COUL_DB TURB_DB COND_DB TURB_HAW	COUL_DB TURB_DB COND_DB TURB_HAW	COUL_DB TURB_DB COND_DB TURB_HAW	TURB_DB COND_DB TURB_DB1	COUL_DB TURB_DB COND_DB	COUL_DB TURB_DB COND_DB
Débit	OUT_LAG1 CONT_LAG CONT_AUJ BDT_LAG4 RAI_LAG5	OUT_LAG1 CONT_LAG BDT_LAG4 RAI_LAG5	OUT_LAG1		OUT_LAG1	OUT_LAG1 CONT_LAG BDT_LAG4 RAI_LAG5	OUT_LAG1	
Vent	LSF1_VITM LSF1_VITX DOR1_VITX DOR1_VITM	LSF1_VITM LSF1_VITX DOR1_VITM	LSF1_VITX DOR1_VITX DOR1_VITM	LSF1_VITX DOR1_VITX DOR1_VITM	LSF1_VITX DOR1_VITX DOR1_VITM	LSF1_VITM LSF1_VITX DOR1_VITM	LSF1_VITM LSF1_VITX DOR1_VITM	LSF1_VITX DOR1_VITX DOR1_VITM
Index	IDX_SAIS IDX_FONT	IDX_SAIS IDX_FONT	IDX_SAIS IDX_FONT	IDX_SAIS		IDX_SAIS	IDX_SAIS	IDX_SAIS

Suite du Tableau A8-1 : Sous-ensembles de variables d'entrée utilisés pour développer les modèles candidats

	8	9	10	11	12	13	14	15
Qualité	TURB_DB TURB_HAW	TURB_DB TURB_HAW	TURB_DB TURB_HAW	TURB_DB	TURB_DB	COUL_DB TURB_DB COND_DB	COUL_DB TURB_DB COND_DB	TURB_DB TURB_HAW
Débit	OUT_LAG1 CONT_LAG BDT_LAG4	BDT_LAG4		OUT_LAG1				OUT_LAG1
Vent	LSF1_VITM LSF1_VITX DOR1_VITM	LSF1_VITM LSF1_VITX DOR1_VITM	LSF1_VITM LSF1_VITX DOR1_VITM	LSF1_VITM LSF1_VITX DOR1_VITM	LSF1_VITM LSF1_VITX DOR1_VITM	LSF1_VITX DOR1_VITX DOR1_VITM		LSF1_VITM LSF1_VITX DOR1_VITM
Index	IDX_SAIS	IDX_SAIS	IDX_SAIS	IDX_SAIS	IDX_SAIS			IDX_SAIS

L'analyse de sensibilité conduite par le logiciel NeuralNetworks classe les variables selon la détérioration dans la performance du modèle si cette variable n'était plus disponible pour son développement. Le logiciel accorde deux valeurs de rang à chaque variable, l'une s'appuyant sur les données de l'ensemble d'apprentissage et l'autre, sur les données de l'ensemble de test. Les rangs accordés selon les deux ensembles de données devraient être sensiblement les mêmes - la constance des résultats est un bon indicateur de la fiabilité des résultats. Les rangs accordés aux différentes variables sont rapportés au Tableau A8-2. Les rangs en caractères gras indiquent les variables les plus importantes pour la bonne performance du modèle.

Tableau A8-2 : Résultats de l'analyse de sensibilité conduite pour l'ensemble des variables d'entrée pour les deux variables de sortie

Variables en entrée	Variable de sortie : DIFF_1		Variable de sortie : EAU_1	
	Répartition 1	Répartition 2	Répartition 1	Répartition 2
COUL_DB	17 / 11	11 / 16	15 / 17	11 / 16
TURB_DB	**1 / 1**	**1 / 1**	**1 / 2**	**1 / 1**
COND_DB	12 / 15	12 / 10	8 / 15	12 / 10
TURB_DB1	10 / 13	15 / 11	**6 / 9**	15 / 11
TURB_HAW	**4 / 2**	6 / 17	13 / 11	6 / 17
COUL_HAW	15 / 14	14 / 13	12 / 13	14 / 13
OUT_LAG1	**7 / 10**	**4 / 9**	**2 / 8**	**4 / 9**
CONT_LAG	**9 / 6**	**2 / 2**	**7 / 3**	**2 / 2**
CONT_AUJ	11 / 16	**3 / 6**	16 / 6	**3 / 6**
BDT_LAG4	**2 / 8**	**8 / 8**	**10 / 7**	**8 / 4**
RAI_LAG5	14 / 17	16 / 14	17 / 12	16 / 14
LSF1_VITM	**3 / 4**	**3 / 6**	**4 / 4**	**5 / 7**
DOR1_VITX	13 / 9	13 / 12	**3 / 5**	**9 / 5**
LSF1_VITX	**5 / 3**	**7 / 3**	**9 / 10**	**7 / 3**
DOR1_VITM	**6 / 7**	**10 / 5**	14 / 16	13 / 12
IDX_SAIS	**8 / 5**	**10 / 4**	(<1)	**10 / 4**
IDX_FONT	16 / 12	17 / 15	**5 / 1**	17 / 15
TURB_1PR	--	--	11 / 14	

ANNEXE 9

Résultats des modèles RNA de régression développés

ANNEXE 9 - RÉSULTATS DES MODÈLES DE RÉGRESSION DÉVELOPPÉS

Tableau A9-1 : Différence entre l'erreur d'apprentissage et l'erreur de validation, coefficients de corrélation et résultats de l'inspection visuelle des graphiques

S.-ens. de var. d'entrée	Architecture du RNA		Différence entre l'erreur d'apprentissage et l'erreur de validation		Coefficient de corrélation				Inspection visuelle des représentations graphiques			
					Événements turbides		Année complète		Présence de décalages temporels		Justesse de prévision des données de validation	
	Rép. 1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2
0	PMC 17 :4 :1	PMC 17 :4 :1	36%	2%	0,76	0,76	0,64	0,68	Moy.	Moy.	Moy.	Moy.
	GRNN 17:740:2:1	GRNN 17:707:2:1	35%	19%	0,81	0,78	0,73	0,72	Beauc.	Beauc.	Moy.	Moy.
1	PMC 14 :2 :1		43%		0,76		0,66		Beauc.		Faible	
	GRNN 14:740:2:1		26%		0,83		0,74		Beauc.		Moy.	
2	PMC 10 :20 :1	PMC 10 :20 :1	38%	6%	0,84	0,78	0,73	0,62	Peu	Moy.	Bon	Moy.
	GRNN 10:740:2:1		37%		0,84		0,75		Moy.		Moy.	
3	RBF 8: 12 :1		43%		0,76		0,66		Moy.		Moy.	
	GRNN 8:740:2:1		38%		0,87		0,79		Moy.		Moy.	
4	PMC 8: 20 :1	PMC 8: 20 :1	23%	1%	0,80	0,75	0,67	0,66	Moy.	Moy.	Moy.	Moy.
	GRNN 8:740:2:1		34%		0,84		0,75		Beauc.		Moy.	
5	RBF 11 :26 :1		46%		0,73		0,64		Beauc.		Faible	
	PMC 11 :1 :1		40%		0,74		0,63		Beauc.		Faible	
	GRNN 11:740:2:1	GRNN 11:707:2:1	22%	71%	0,94	0,89	0,88	0,84	Peu	Peu	Excel.	Bonne
6	RBF 8 :40 :1		42%		0,76		0,66		Moy.		Moy.	
	GRNN 8:740:2:1	GRNN 8:707:2:1	13%	34%	0,94	0,89	0,86	0,83	Peu	Peu	Excel.	Bonne

Suite du Tableau A9-1 : Différence entre l'erreur d'apprentissage et l'erreur de validation, coefficients de corrélation et résultats de l'inspection visuelle des graphiques

S.-ens. de var. d'entrée	Architecture du RNA		Différence entre l'erreur d'apprentissage et l'erreur de validation		Coefficient de corrélation				Inspection visuelle des représentations graphiques			
					Événements turbides		Année complète		Présence de décalages temporels		Justesse de prévision des données de validation	
	Rép. 1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2
7	PMC 7: 20 :1		13%		0,77		0,63		Beauc.		Moy.	
	GRNN 6:740:2:1		39%		0,85		0,77		Beauc.		Moy.	
8	PMC 9: 6 :1	PMC 9: 6 :1	40%	2%	0,80	0,75	0,68	0,64	Moy.	Beauc.	Moy.	Moy.
	GRNN 6:740:2:1	GRNN 6:707:2:1	98%	15%	0,95	0,84	0,88	0,77	Beauc.		Moy.	
9	RBF 7: 36 :1		37%		0,79		0,67		Moy.		Moy.	
	GRNN 6:740:2:1		20%		0,87		0,77		Moy.		Moy.	
10	PMC 6: 19 :1	PMC 6: 19 :1	34%	1%	0,82	0,78	0,68	0,67	Moy.	Moy.	Moy.	Moy.
	GRNN 6:740:2:1		13%		0,87		0,76		Moy.		Moy.	
11	RBF 6: 60 :1	RBF 6: 60 :1	31%	17%	0,80	0,73	0,67	0,60	Moy.		Moy.	
	PMC 6 :29 :1		33%		0,80		0,67		Beauc.		Moy.	
12	PMC 5:7 :1		38%		0,76		0,65		Beauc.		Faible	
	GRNN 6:740:2:1		30%		0,84		0,75		Beauc.		Faible	
13	GRNN 6 :740 :2 :1		35%		0,86		0,77		Beauc.		Moy.	
	PMC 6 :1 :1		38%		0,72		0,62		Beauc.		Moy.	
14	RBF 3 :12 :1		40%		0,73		0,64		Beauc.		Faible	
	PMC 3 :1 :1		42%		0,71		0,59		Beauc.		Faible	
	GRNN 3 :740 :2 :1		30%		0,77		0,68		Beauc.		Faible	

ANNEXE 10

Résultats des modèles RNA de classification développés

ANNEXE 10 - RÉSULTATS DES MODÈLES DE CLASSIFICATION DÉVELOPPÉS

Tableau A10-1 : Différence entre l'erreur d'apprentissage et l'erreur de validation

Sous-ensembles de variables d'entrée	Architecture du RNA		Différence entre l'erreur d'apprentissage et l'erreur de validation	
	Rép. 1	Rép. 2	Rép. 1	Rép. 2
0	PMC 17 :3 :1	PMC 17 :3 :1	20%	4%
	RBF 17 :740 :1		15%	
1	PMC 14 :9 :1	PMC 14 :9 :1	16%	6%
	RBF 14 :29 :2 : 1		9%	
2	PMC 10 20 :1		22%	
	RBF 10 :26 :1		6%	
3	PMC 8 :7 :1		21%	
4	PMC 8 :5 :1		25%	
	RBF 8 :25 :1		7%	
5	PMC 11 :2 :1		30%	
	RBF 11 :13 :1		16%	
5 + IND_FONT	PMC 12 :4 :1	PMC 12 :4 :1	21%	10%
	RBF 12 :18 :1		16%	
6	PMC 8 :6 :1		24%	
7	PMC 7 :19 :1		19%	

Suite du Tableau A10-1 : Différence entre l'erreur d'apprentissage et l'erreur de validation

Sous-ensembles de variables d'entrée	Architecture du RNA		Différence entre l'erreur d'apprentissage et l'erreur de validation	
8	PMC 9 :2 :1	PMC 9 :2 :1	21%	4%
9	PMC 6 :9 :1		21%	
10	PMC 6 :9 :1	PMC 6 :9 :1	21%	7%
10 + IND_FONT	PMC 7 :17 :1		24%	
11	PMC 6 :5 :1	PMC 6 :5 :1	25%	1%
	RBF 6 :26 :1		10%	
11 + IND_FONT	PMC 7 :6 :1		25%	
12	PMC 5 :7 :1		23%	
13	PMC 6 :6 :1		20%	
14	PMC 3 :4 :1		21%	
15	PMC 7 :11 :1			

Tableau A10-2 : Taux de classification correcte pour les événements turbides et pour l'année complète

s.-ens. de var.	Architecture du RNA		Classification correcte entre « eau claire » et « eau turbide »							
			Toute l'année		Événements turbides		Début des événements turbides		Début des événements de validation	
	Rép. 1	Rép. 2	Rép .1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2
0	PMC 17 :3 :1	PMC 17 :3 :1	82%	82%	90%	89%	87%	77%	Aut. : 3 / 4 Print.: 1 / 1	Aut. : 3 / 4 Print.: 1 / 2
	RBF 17 :740 :1		82%		90%		77%		Aut. : 2 / 4 Print.: 1 / 1	
1	PMC 14 :9 :1	PMC 14 :9 :1	82%	83%	91%	89%	84%	77%	Aut. : 3 / 4 Print.: 1 / 1	Aut. : 3 / 4 Print.: 0 / 2
	RBF 14 :29 :2 :1		80%		88%		68%		Aut. : 2 / 4 Print.: 1 / 1	
2	PMC 10 20 :1		85%		91%		74%		Aut. : 2 / 4 Print.: 1 / 1	
	RBF 10 :26 :1		81%		91%		77%		Aut. : 2 / 4 Print.: 1 / 1	
3	PMC 8 :7 :1		85%		90%		71%		Aut. : 2 / 4 Print.: 1 / 1	
4	PMC 8 :5 :1		82%		81%		45%		Aut. : 1 / 4 Print.: 1 / 1	
	RBF 8 :25 :1		83%		86%		61%		Aut. : 2 / 4 Print.: 1 / 1	
5	PMC 11 :2 :1		84%		90%		74%		Aut. : 2 / 4 Print.: 1 / 1	
	RBF 11 :13 :1		78%		90%		77%		Aut. : 2 / 4 Print.: 1 / 1	
5 + IDX_ FONT	PMC 12 :4 :1	PMC 12 :4 :1	81%	84%	90%	89%	81%	77%	Aut. : 3 / 4 Print.: 1 / 1	Aut. : 4 / 4 Print.: 0 / 2
	RBF 12 :18 :1		82%		90%		77%		Aut. : 2 / 4 Print.: 1 / 1	
6	PMC 8 :6 :1		83%		88%		65%		Aut. : 2 / 4 Print.: 1 / 1	

Suite du Tableau A10-2 : Taux de classification correcte pour les événements turbides et pour l'année complète

s.-ens. de var.	Architecture du RNA		Classification correcte entre « eau claire » et « eau turbide »							
			Toute l'année		Événements turbides		Début des événements turbides		Début des événements de validation	
	Rép. 1	Rép. 2	Rép .1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2	Rép. 1	Rép. 2
7	PMC 7 :19 :1		84%		88%		65%		Aut. : 2 / 4 Print.: 1 / 1	
8	PMC 9 2 :1	PMC 9 2 :1	85%	84%	91%	88%	77%	74%	Aut. : 3 / 4 Print.: 1 / 1	Aut. : 3 / 4 Print.: 1 / 2
9	PMC 6 9 :1		82%		91%		74%		Aut. : 2 / 4 Print.: 1 / 1	
10	PMC 6 9 :1	PMC 6 9 :1	82%	86%	91%	91%	77%	81%	Aut. : 3 / 4 Print.: 1 / 1	Aut. : 3 / 4 Print.: 1 / 2
10 + IDX_ FONT	PMC 7 :17 :1		83%		89%		71%		Aut. : 2 / 4 Print.: 1 / 1	
11	PMC 6 5 :1	PMC 6 5 :1	83%	84%	92%	92%	81%	81%	Aut. : 3 / 4 Print.: 1 / 1	Aut. : 3 / 4 Print.: 2 / 2
	RBF 6 26 :1		81%		91%		81%		Aut. : 2 / 4 Print.: 1 / 1	
11 + IDX_ FONT	PMC 7 6 :1		85%		90%		77%		Aut. : 3 / 4 Print.: 0 / 1	
12	PMC 5 7 :1		83%		91%		74%		Aut. : 3 / 4 Print.: 0 / 1	
13	PMC 6 6 :1		82%		83%		52%		Aut. : 1 / 4 Print.: 1 / 1	
14	PMC 3 4 :1		78%		77%		32%		Aut. : 1 / 4 Print.: 0 / 1	
15	PMC 7 :11 :1				90%		74%		Aut. : 2 / 4 Print.: 1 / 1	

ANNEXE 11

Représentations graphiques des modèles candidats choisis

ANNEXE 11 - REPRÉSENTATIONS GRAPHIQUES DES MODÈLES CHOISIS

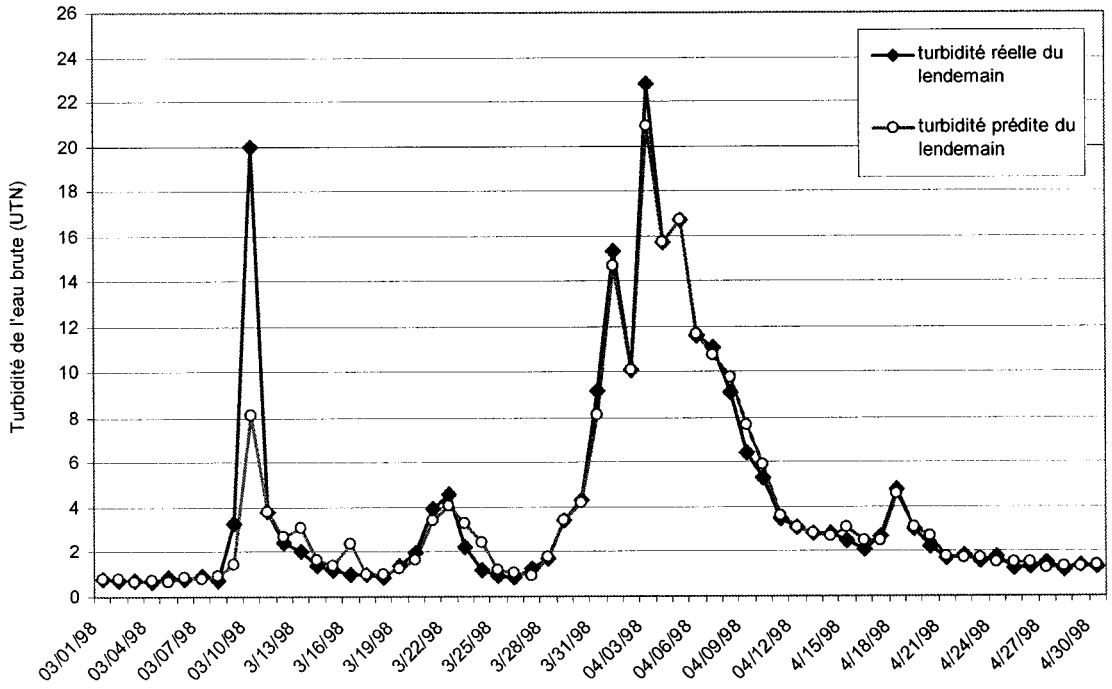


Figure A10-1 : Prévisions du modèle de régression GRNN 8:740:12:1 pour le printemps 1998

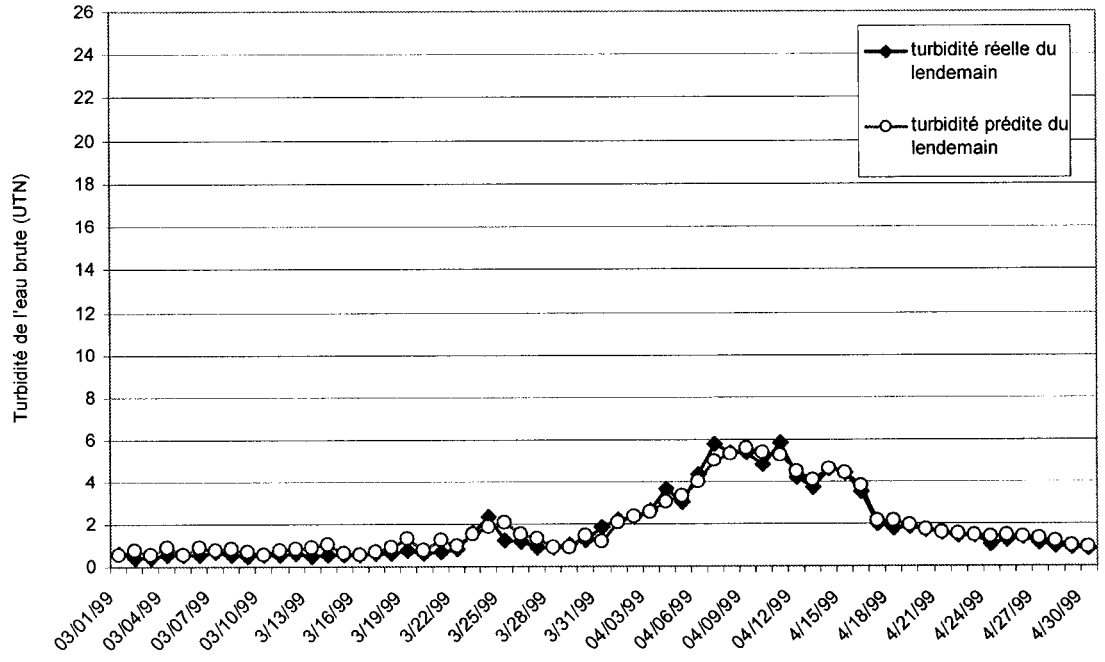


Figure A10-2 : Prévisions du modèle de régression GRNN 8:740:12:1 pour le printemps 1999

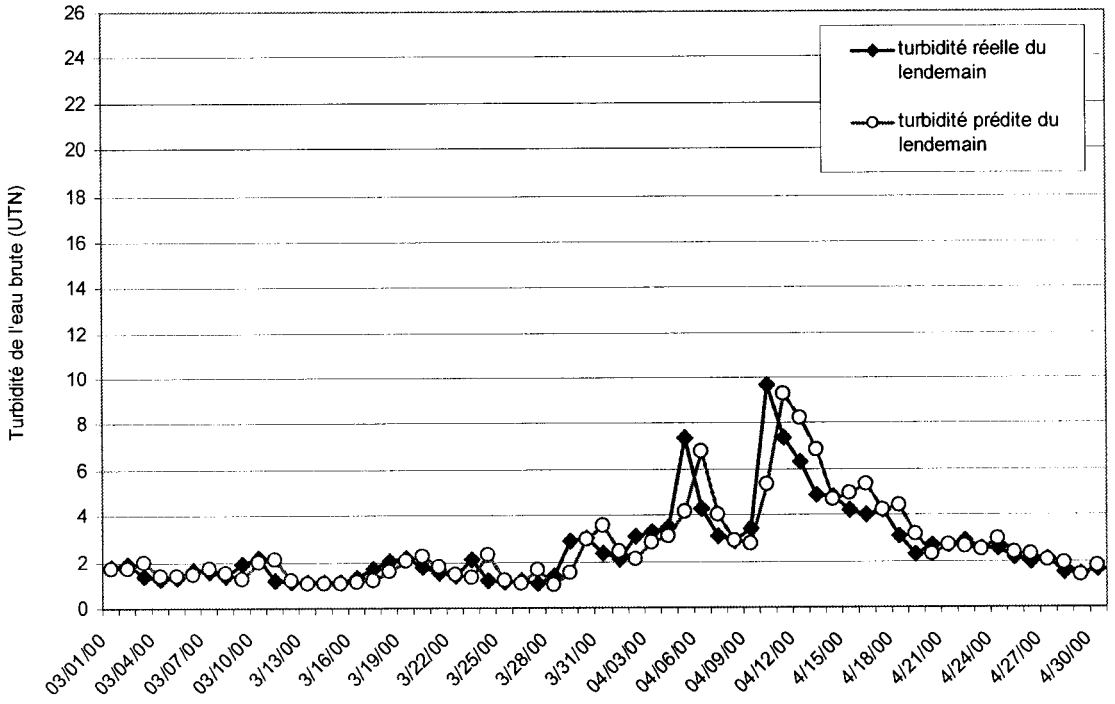


Figure A10-3 : Prévisions du modèle de régression GRNN 8:740:12:1 pour le printemps 2000

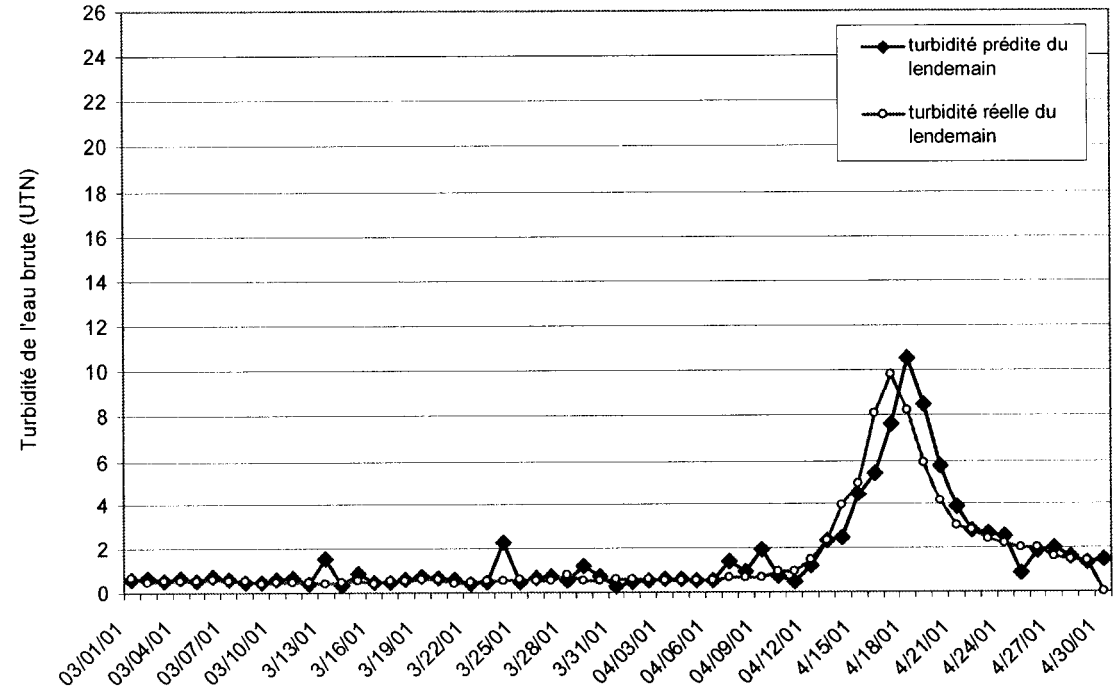


Figure A10-4 : Prévisions du modèle de régression GRNN 8:740:12:1 pour le printemps 2001

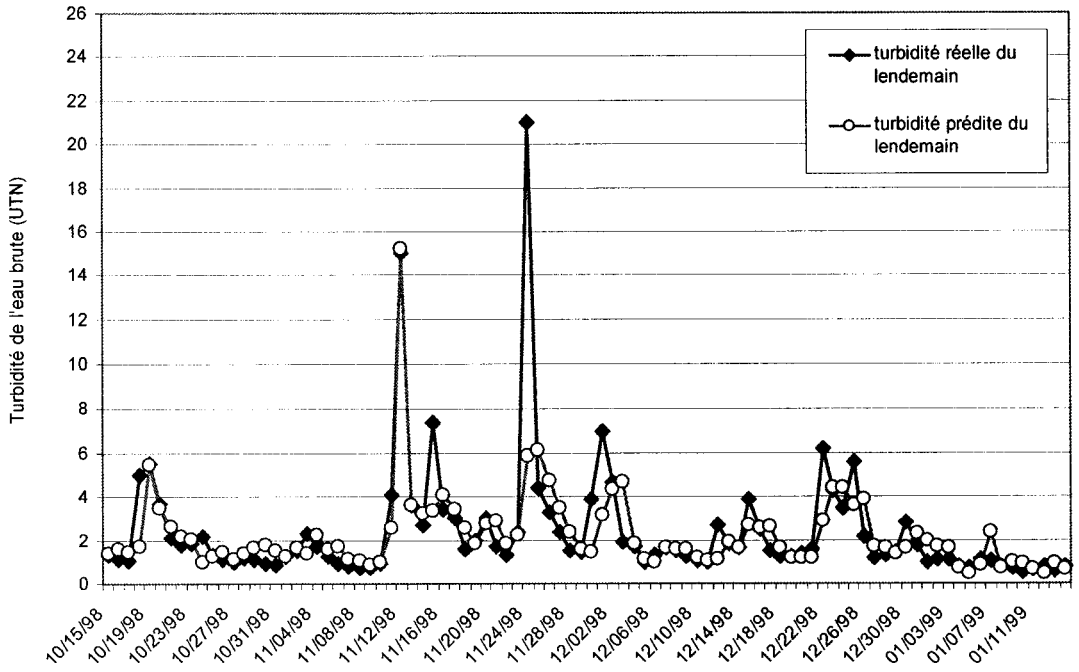


Figure A10-5 : Prévisions du modèle de régression GRNN 8:740:12:1 pour l'automne 1998

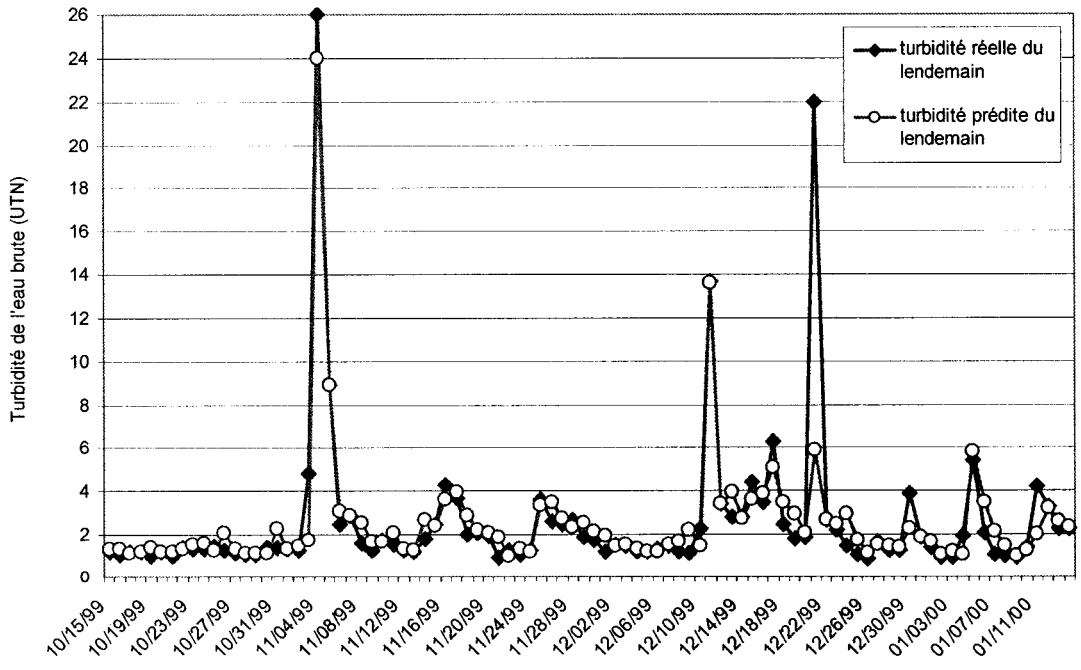


Figure A10-6 : Prévisions du modèle de régression GRNN 8:740:12:1 pour l'automne 1999

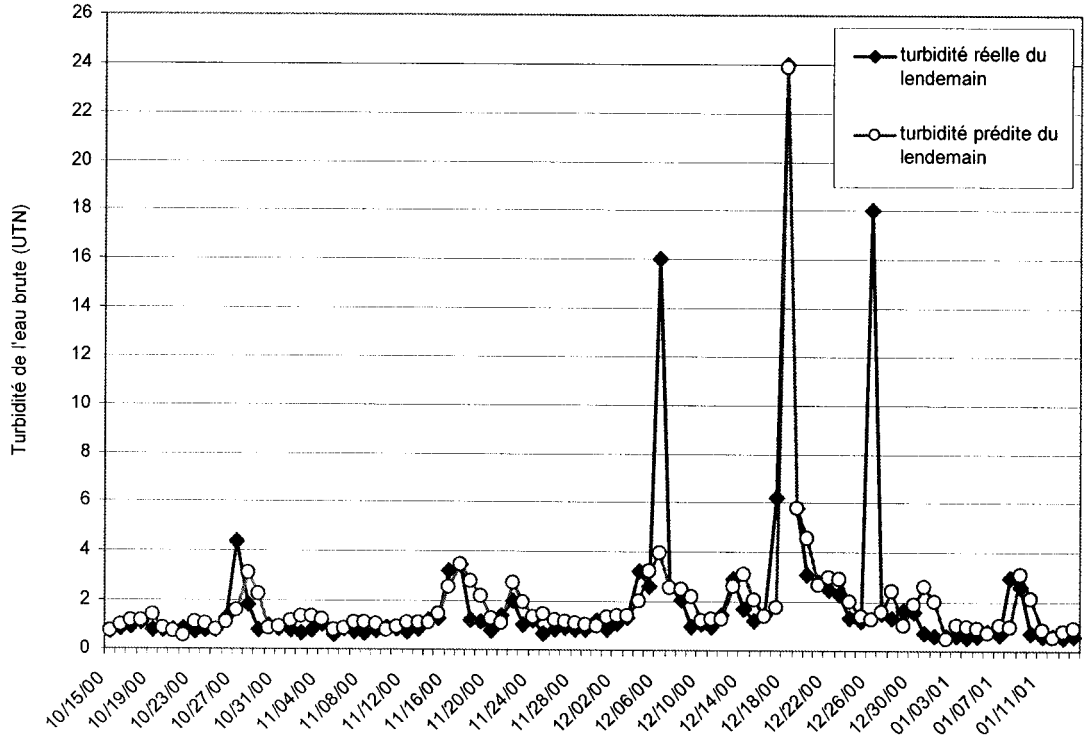


Figure A10-7 : Prévisions du modèle de régression GRNN 8:740:12:1 pour l'automne 2000

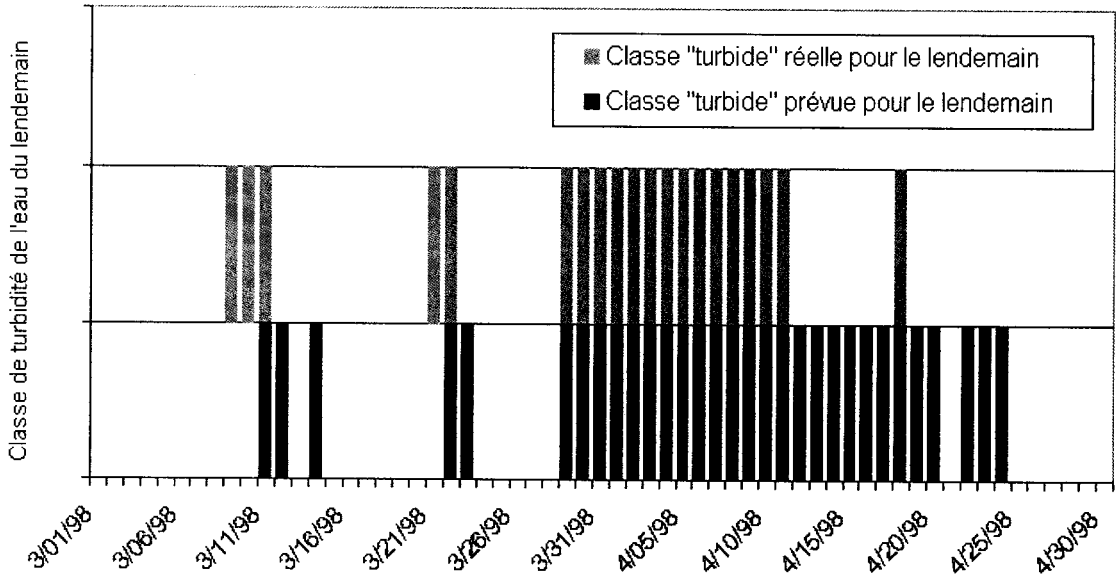


Figure A10-8 : Prévisions du modèle de classification PMC 6 :5 :1 pour le printemps 1998

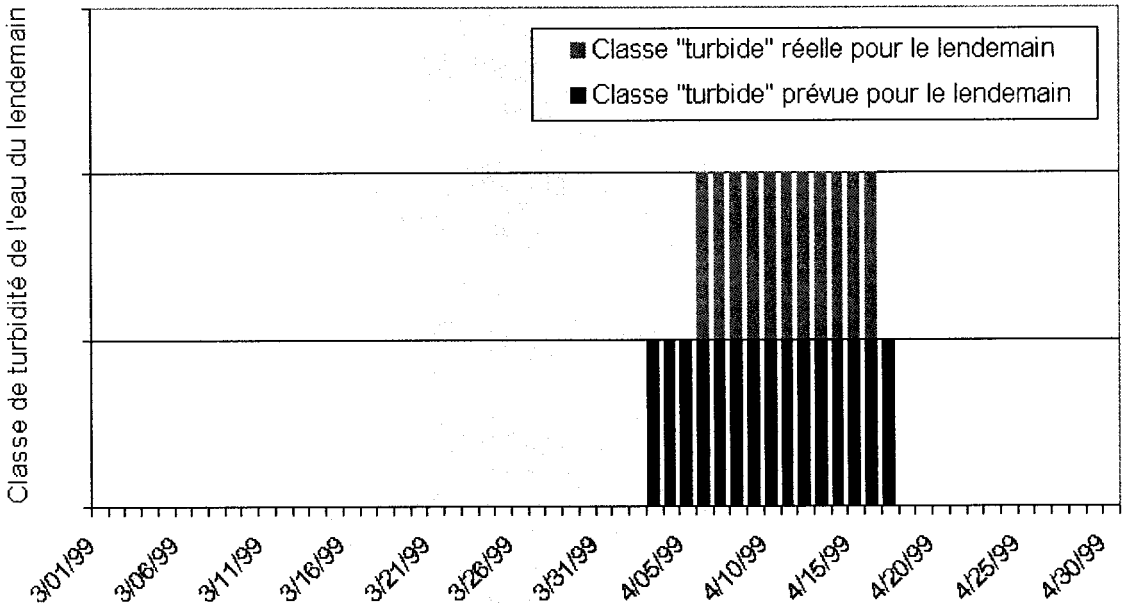


Figure A10-9 : Prévisions du modèle de classification PMC 6 :5 :1 pour le printemps 1999

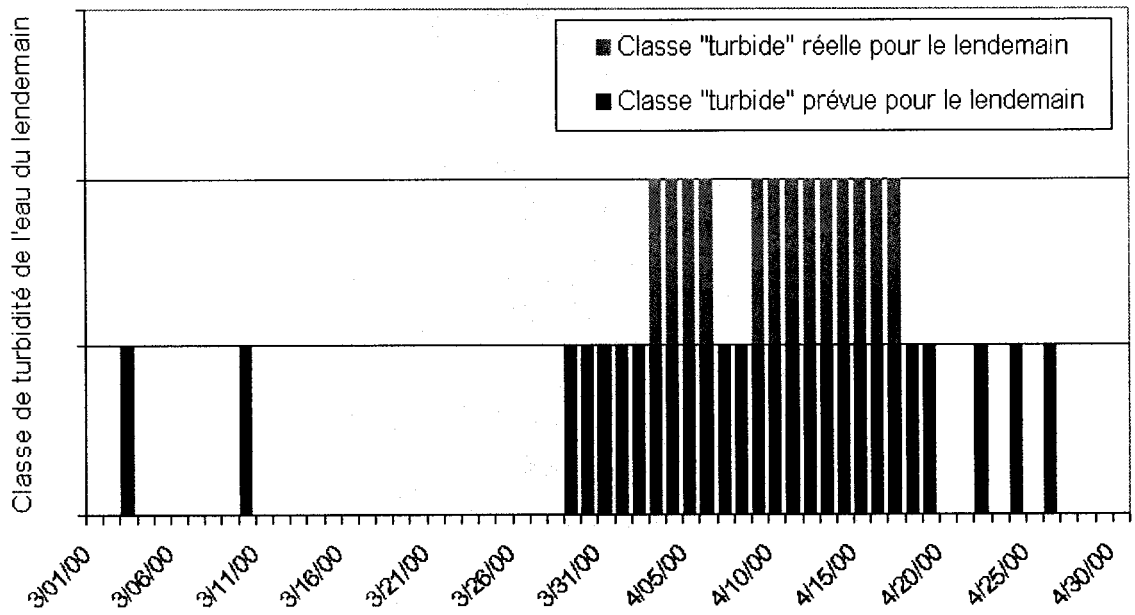


Figure A10-10 : Prévisions du modèle de classification PMC 6 :5 :1 pour le printemps 2000

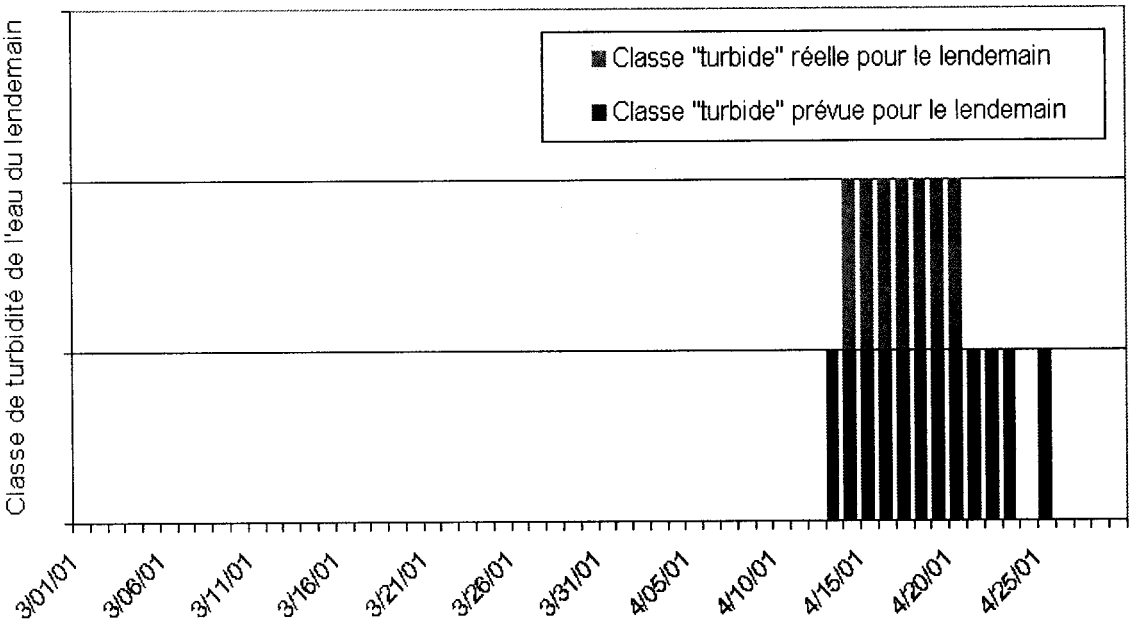


Figure A10-11 : Prévisions du modèle de classification PMC 6 :5 :1 pour le printemps 2001

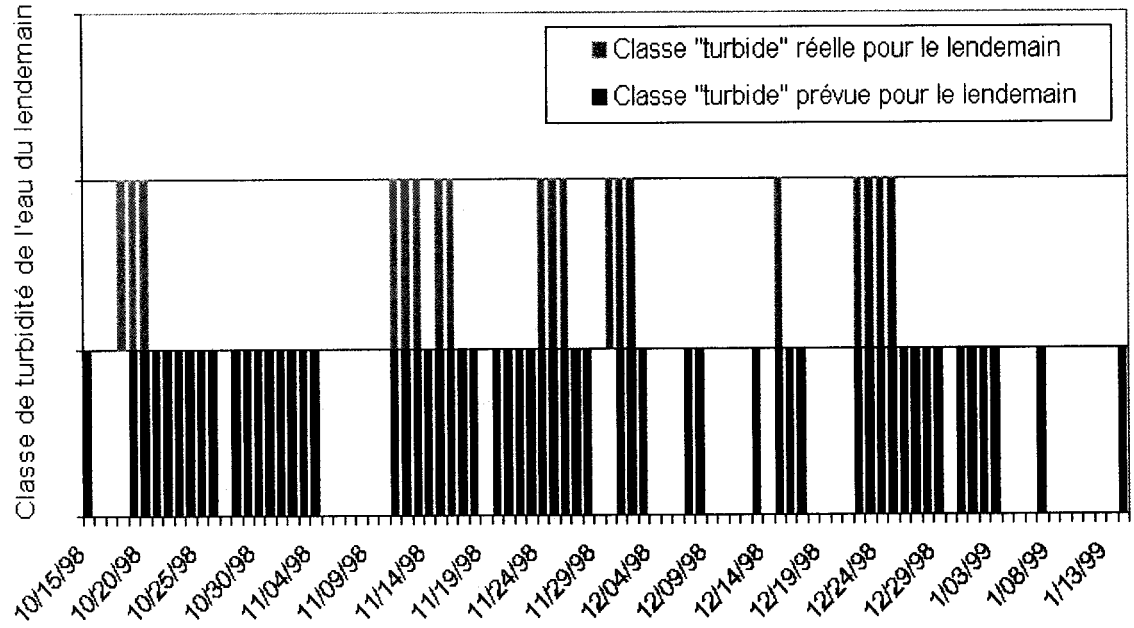


Figure A10-12 : Prévisions du modèle de classification PMC 6 :5 :1 pour l'automne 1998

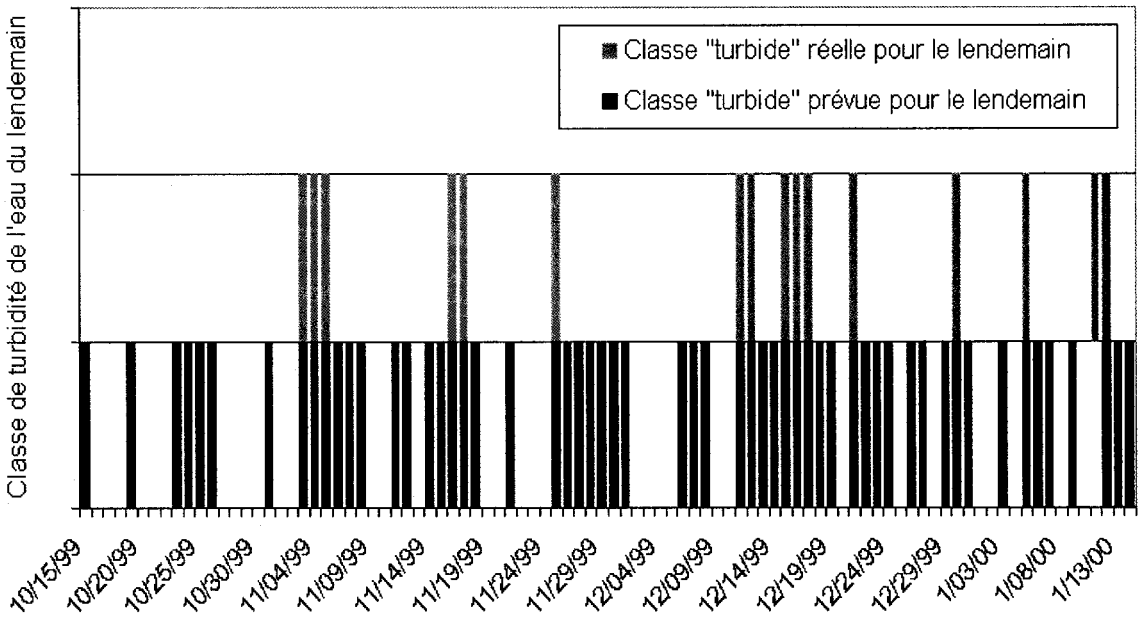


Figure A10-13 : Prévisions du modèle de classification PMC 6 :5 :1 pour l'automne 1999

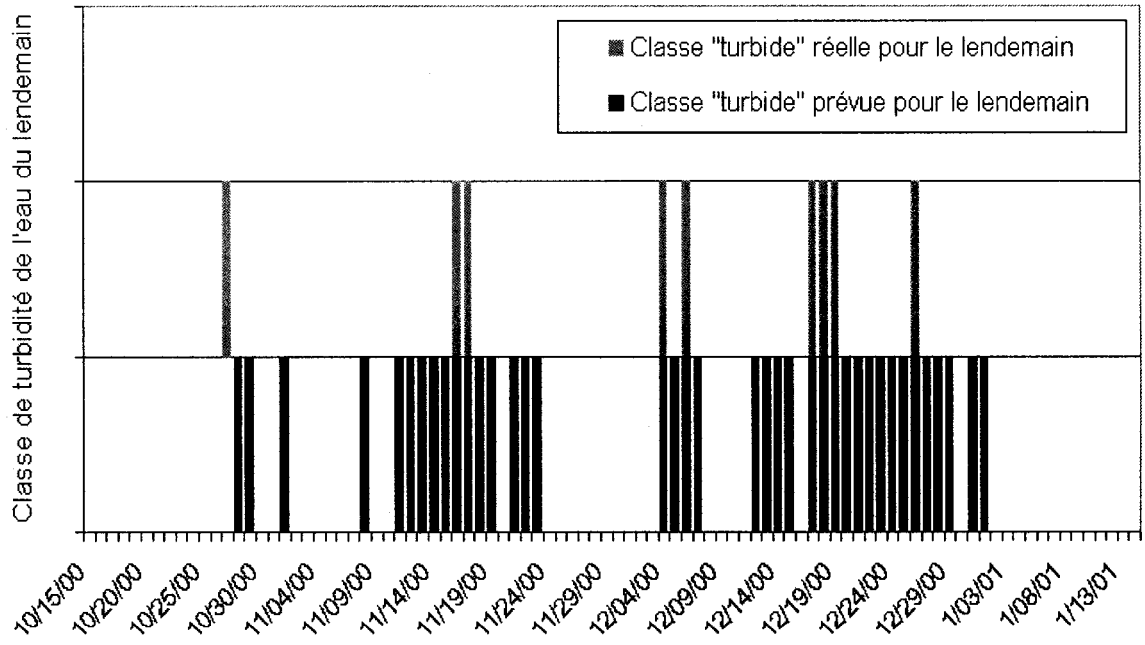


Figure A10-14 : Prévisions du modèle de classification PMC 6 :5 :1 pour l'automne 2000