Predicting trihalomethane formation in chlorinated waters using multivariate regression and neural networks

Manuel J. Rodriguez, Julie Milot and Jean-B. Sérodes

ABSTRACT

Recently, there has been increased interest in modelling disinfection by-products (DBP) in order to better understand and manage the presence of these compounds in drinking water. In this paper, the use of artificial neural networks (ANN) to predict trihalomethane (THM) formation resulting from chlorination bench-scale experiments is investigated and compared with the use of classical multivariate linear regression (MLR). ANN and MLR were developed from three databases which were generated through bench-scale chlorination essays carried out in the US and Canada. A detailed analysis of modelling results shows that for all three databases, ANNs have in general a greater ability than MLRs to predict THM formation for most water quality and chlorination conditions, with the exception of instantaneous THMs (formation immediately following chlorine addition).

Key words | chlorination, modelling, multivariate regression, neural networks, trihalomethanes

Manuel J. Rodriguez (corresponding author)
Département d'Aménagement,
1624 F. A. Savard,
Université Laval,
Québec, QC,
Canada, G1K 7P4
Tel: (418) 656-2131 ext. 8933
Fax: (418) 656-2018
E-mail: *manuel.rodriguez@ame.ulaval.ca*

Julie Milot
Centre de Recherche en Aménagement et
Développement (CRAD),
1636 F. A. Savard,
Université Laval,
Québec, QC,
Canada, G1K 7P4

Jean-B. Sérodes
Département de Génie Civil,
1916 Pouliot,
Université Laval,
Québec, QC,
Canada, G1K 7P4

INTRODUCTION

Chlorination by-products (CBPs) are generated during water disinfection due to the reaction of chlorine with natural organic matter (NOM) contained in raw waters (Rook 1974; Symons *et al.* 1975). Among these, trihalomethanes (known as THMs: chloroform, bromodichloromethane, dibromochloromethane and bromoform) have been the focus of particular attention because they are considered potentially carcinogenic (Cantor *et al.* 1987). Concerns about health risks associated particularly with chloroform and bromodichloromethane have resulted in the establishment of maximum acceptable levels for THM concentrations in drinking water by the World Health Organization and by several industrialized countries (Rodriguez & Sérodes 2001). The US Environmental Protection Agency (USEPA) proposed a two-stage disinfection (D/DBP) rule in which maximum contaminant levels of total THMs are 80 μg/l based on an annual running average (the second stage standard will be based on locational running annual average) (USEPA 1994; Sharfenaker 2001). Canada recently set out drinking water guidelines stating a total THM maximum acceptable level of 100 μg/l (Health Canada 1996) which is the same standard established by the European Community (Conseil Européen 1998).

In recent years, some research effort has been made to develop predictive models for the formation of THMs in water. Models for THMs are aimed at better understanding the factors affecting their formation and thus are useful as decision-making tools. They may be used by water planners for comparing alternative water sources for city supply (by comparing the potential for THM formation of different sources of waters), or for comparing alternatives in treatments for NOM removal. They can also be applied as decision-making tools by government officials when evaluating the feasibility of updating THM standards in drinking water. Finally, models may be useful in epidemiological studies concerning health effects of DBPs, by providing estimations of a population's exposure to these substances in drinking water.

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

The modelling of THMs consists of establishing relationships between THM levels in water and the parameters of water quality and operational control which can influence their formation. The most important factors for THM formation are the levels of organic matter in water (generally designated by total or dissolved organic carbon, TOC or DOC, and by UV-254 nm absorbance), the applied chlorine dose, water pH, water temperature and the reaction time of residual chlorine in water. Concentrations of bromide are also generally considered because they influence the distribution of the four THM compounds. In different studies, THM formation models have been developed both from data generated in full-scale studies at real water utilities or at laboratory-scale using controlled chlorination conditions. The latter approach is currently being used in testing for THM formation potential in waters, with the aim of establishing the susceptibility of a water to form THMs due to its natural quality (APHA, AWWA & WPCF 1995). Some researchers have developed models to describe the formation of THMs based on kinetics involved during chlorine reactions (Engerholm & Amy 1983; Racaud & Rauzy 1994; Clark & Sivaganesan 1998). However, most of the models presented in literature are empirical and are based on statistical regression equations which predict the levels of THMs from a number of operational and water quality parameters (Amy *et al*. 1987; Montgomery Watson 1993; Rathbun 1996; Rodriguez *et al*. 2000; Milot *et al*. 2000). Although regression-based models have shown acceptable predictive capacity for THM formation, they do not allow for consideration of the complex and non-linear interactions between operational and water quality parameters used as estimators for THMs.

The aim of this paper is to evaluate the ability of a recent non-linear modelling technology, artificial neural networks (ANNs), to predict the concentrations of THMs formed under controlled chlorination conditions. ANN modelling will be assessed by comparing its ability to predict THM levels in comparison with the classical multivariate linear regression (MLR) approach. The development of the model is based on different databases generated through chlorination experiments with natural and treated waters of the US and the province of Quebec (Canada).

METHODOLOGY

Overview of MLR and ANNs

MLR analysis is a well-known modelling methodology used in many research fields to establish the strength of a linear relationship between a dependent variable and a set of independent variables (Menard 1995). The relationship between variables can be described using an equation in the following form:

$$Y = \sum \beta_0 + \sum_{i=1}^{m} \beta_{\mathrm{i}} X_{\mathrm{i}} \tag{1}$$

where Y is the dependent variable, X_{i} represents the independent variables with m denoting the number of independent variables considered, β_0 the intercept and β_{i} the partial slope coefficients providing a partial explanation or prediction for the value of Y. The parameters of the MLR model are generally estimated using the ordinary least squares (OLS) method which results in a line that minimizes the sum of squared vertical distances from the observed data points to the line (Lewis-Beck 1980; Neter *et al*. 1990).

ANNs are a modelling technique inspired by studies of the brain's nervous system. They are capable of learning by example from representative data which describe a physical phenomenon or a decision process (Rumelhart *et al.* 1994). An ANN model provides certain theoretical advantages over conventional approaches such as MLR, including its high capacity for generalization and its increased tolerance to noisy data (Hammerstrom 1993). An ANN consists of several layers of processing elements (Figure 1): one input layer that receives a signal input, one or many hidden layers for processing information, and one output layer containing the response of the network. Elements between layers are highly interconnected by weighted links through which information may pass. The number of elements contained in the input and the output layer depends, respectively, on the number of input variables and output variables used in the specific problem to be solved. *Back-propagation* is the most commonly used algorithm in the ANN learning process (Jones & Hoskins 1987; Cook & Wolfe 1991). The learning process (also

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

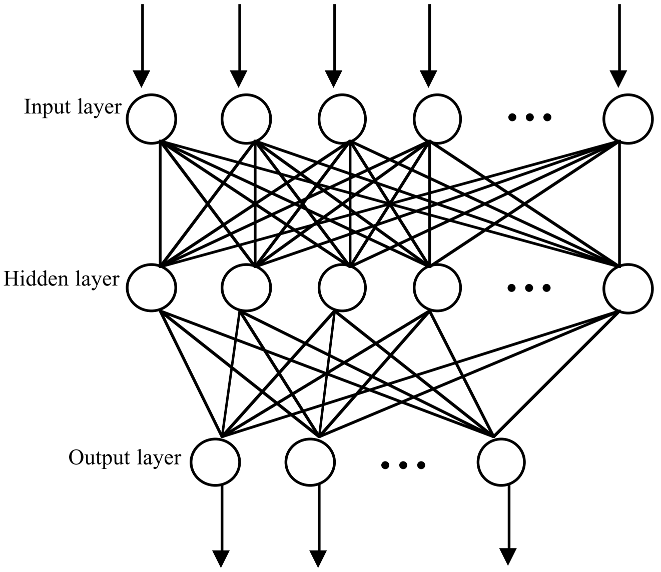

Figure 1 | Three layer ANN.

called training) consists of presenting pairs of input-output examples to the network a sufficient number of times (iterations) until differences between the desired response and the calculated response given by the ANN are minimized (Caudill 1991; Hammerstrom 1993). During the learning process, three parameters have to be estimated by experimentation: the number of hidden elements, the learning rate and the momentum. Mathematical details of learning and parameter estimation have been broadly described in literature (Lippmann 1987; Simpson 1992; Rodriguez & Sérodes 1994). The development of an ANN is generally carried out by dividing the original database into three different data sets: the learning set for parameter adjustment, a cross-validation set for model verification during training (ensuring model generalization) and a verification set for overall test of the trained model.

The ANN modelling approach has been applied in several fields for solving classification problems, for forecasting phenomena and for pattern recognition (Rumelhart *et al*. 1994; Wen & Lee 1998). In the field of drinking water management they have been used for assessing the quality of raw water (Zhang & Stanley 1997; Maier & Dandy 2000), for predicting urban water demand (Crommelynck *et al*. 1992; Heller & Singh Thind 1994), for establishing coagulation dosage (Collins *et al*. 1991; Baxter *et al*. 1999; Joo *et al*. 2000), for predicting peaks of parasite occurrence (Brion *et al*. 2001) and for predicting residual chlorine decay in distribution systems (Sérodes & Rodriguez 1996; Sérodes *et al*. 2001). Considering how much interest in modelling THMs in drinking water has grown in recent years and the ability of the ANN to model complex and non-linear phenomena, it becomes important to assess the abilities of this approach in predicting THM formation in chlorinated waters.

Description of databases for modelling THMs

Three databases describing THM formation in experimental bench-scale conditions are used to develop MLR and ANN models. Two databases were developed by researchers in the US (Amy *et al*. 1987; Rathbun 1996) whereas the third has recently been developed by the authors in Quebec (Canada). Table 1 presents the information about the procedure for the development of the three THM formation databases.

Although all three databases were developed following controlled chlorination experiments at bench-scale conditions, they are not entirely comparable. Databases 1 and 3 were developed following experiments with waters directly collected from drinking water utilities (raw and/or treated waters) whereas experiments for Database 2 were undertaken using river waters with variable levels of pollution, thus not necessarily usable as sources for drinking water supply. In addition, chlorination conditions associated with Database 1 and Database 2 are similar to those of tests for THM formation potential: relatively high chlorine doses (moderately high for Database 1 and very high for Database 2) and long reaction times. Chlorination conditions for bench-scale experiments which led to the generation of Database 3 are associated with experimentation conditions closer to those encountered in full-scale water utilities. For Database 1, pre-established chlorine doses were used in experiments, whereas for Databases 2 and 3, applied chlorine doses were determined according to the organic matter content of water (TOC as an indicator). The latter strategy

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

Table 1 | Characteristics of databases resulting from bench-scale chlorination experiments

	Database 1	Database 2	Database 3
Origin of waters	Surface waters collected at the uptake of nine utilities across the U.S (raw waters)	Surface waters collected at 14 locations of the Mississipi river and its affluents	Raw waters and treated waters (following physico-chemical treatment) of six utilities of the Quebec city region
Period of collection	From August 1982 to February 1984	From June 1991 to April 1992	From May to September 1999
Conditions for chlorination tests (experimental matrix)	Variable chlorine dose (proportional to organic carbon content), three different water temperatures, natural pH. Samples taken at 10 different contact times over a 7-day incubation period (from 0.1 to 168 h)	Three different chlorine doses (15, 30 and 50 mg/l), water temperature adjusted at 25°C, three different adjusted pHs. Samples taken at nine different contact times over a 7-day incubation period (from 0 to 168 h)	Three different chlorine doses (proportional to TOC), ambient water temperature (20°C), natural pH. Samples taken at eight different contact times over a 2-day incubation period (from 0.16 to 48 h)
Method for analysis of water quality parameters	*TOC*: TOC analyser (Xertex Corp., model DC-80) *UV-absorbance* (254 *nm*): UV/visible spectrophotometry (Perkin Elmer, model 200) *Bromide*: Ion chromatography (Dionex model 10) *THMs*: Gas chromatography with liquid-liquid extraction (Hewlett Packard, model 5794)	*DOC*: Wet oxidation method *UV-absorbance* (254 *nm*): UV/visible spectrophotometry (Spectronics, model 2000) *Bromide*: Segmented flow automated colorimetry *THMs*: Gas chromatography with liquid-liquid extraction (Hewlett Packard, model 5780)	*TOC*: TOC analyzer (Shimadzu, model 500) *UV-absorbance* (254 *nm*): UV/visible spectrophotometry (Pharmacia, model 80-2097-62) *Bromide*: Ion chromatography (Dionex model) *THMs*: Gas chromatography with liquid-liquid extraction (Perkin Elmer, Autosystem XL)
Number of THM measurements in the database	1,025*	581*	270
Models for THMs published in literature**	$THM = 0.0031\ (pH-2.6)^{0.715}\ (UV\times TOC)^{0.440}\ (D)^{0.409}\ (t)^{0.265}\ (T)^{1.06}\ (Br+1)^{0.036}$ $N = 995$ $R^2 = 0.903$	$THM^{***} = 14.6\ (pH-3.8)^{1.01}\ (UV)^{0.849}\ (D)^{0.206}\ (t)^{0.306}$ $N = 685$ $r^2 = 0.985$	n.p.****

*Data obtained from authors.

**Nomenclature: *THM*: total trihalomethanes (μmol/l); *UV*: absorbance UV at 254 nm (cm^{-1}); *TOC*: total organic carbon (mg/l); *D*: chlorine dose (mg/l); *t*: contact time (h); *T*: water temperature (°C); *Br*: bromide (μg/l); *N*: number of observations for model calibration; R^2: multiple coefficient of determination; r^2: adjusted coefficient of determination.

***In μg/l.

****Model not published in literature.

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf by guest

permits the reproduction of practices used within utilities where the applied chlorine dose depends on chlorine demand and thus varies seasonally. Among the three databases, only Database 1 considers the effect of temperature on THM formation (varying from 10°C to 30°C) whereas the others consider controlled water temperature (25°C for Database 2 and 20°C for Database 3). Finally, methods for the measurement of water quality parameters were quite different but comparable. The diversity in the type of waters and the experimental conditions explains why ranges for resulting THM levels are not comparable between the three databases. As presented in Table 2, average and median THM levels are the lowest for Database 3 and the highest for Database 2.

Data for model development

MLR and ANN models were developed based on the three THM formation databases presented earlier. As shown in Table 1, regression-based models have already been developed and published in literature using Databases 1 and 2 by Amy *et al*. (1987) and Rathbun (1996), respectively. However, such models cannot be used as published in comparisons with ANN-based models, for several reasons: (1) the method for regression parameter estimation is different in the model calibration procedures [conventional regression procedure for Amy *et al.* (1987) and regression stepwise procedure for Rathbun (1996)]; (2) the dependent variable (THMs) is expressed in μmol/l by Amy *et al.* (1987) whereas it is expressed in μg/l in the model by Rathbun (1996); (3) both authors excluded some data from the raw database for MLR development (because of lack of validity, because of variable transformation requirements, or to consider only cases with remaining residual chlorine), but such data were not identified in the raw database obtained from the authors; (4) in both cases, the authors calibrated and tested MLR models using the same observations, which is an acceptable strategy for this statistical modelling approach. However, ANN modelling requires the segmentation of the database into three data sets (training, cross-validation and verification sets). Thus, to be able to appropriately compare the ability of both approaches for predicting THM formation, MLR and ANN models must be developed using identical data sets. For these reasons, in addition to developing an MLR using the Quebec data (Database 3), two new MLR models were also developed, using Databases 1 and 2, respectively.

Accordingly, Databases 1, 2 and 3 were subdivided randomly into three subsets for model development in order to comply with methodological requirements for ANN modelling. Each database was separated randomly into a training set (60% of data), a cross-validation set (20% of data) and a verification set (20% of data) (Table 3). Because cross-validation is a requirement for the development of the ANN models only, calibration of the MLR models (that is, estimation of the statistical parameters β_i of Equation 1) was done using the combined training and cross-validation sets.

MODEL DEVELOPMENT AND RESULTS

MLR models

Using the *stepwise* procedure of the statistical software SPSS and based on the OLS method (SPSS France 1997), MLR models for THM formation were developed using the combined training and cross-validation sets of Databases 1, 2 and 3. The method consists of first classifying the predictor variables according to their statistical significance and then including one variable at a time at different steps. Linear models were considered both based on ln-ln variable transformation and without any variable transformation. Working with the model structures obtained by Amy *et al*. (1987) and Rathbun (1996) (presented in Table 1), different forms were used in order to have the models include the pH and the bromide concentrations as well as the combination of organic matter indicators (UV and TOC or DOC). The reason for using a variable transformation for the pH [pH 2.6 for Amy *et al.* (1987) and pH 3.8 for Rathbun (1996)] is to consider values at which THM formation starts, whereas the reason for including a variable increasing the bromide concentration (in mg/l) by the unity (bromide + 1) is to allow for consideration of very low values of bromide within the models as ln-ln transformation is applied. Finally, TOC or DOC and UV were

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

Table 2 | Statistical distribution of water quality parameters for the different databases

Parameter	Database	Minimum	Average	Median	Maximum
pH	Database 1	4.6	7.3	7.3	9.8
	Database 2	5.5	7.5	7.5	10
	Database 3	6.8	7.6	7.6	8.4
UV (cm^{-1})	Database 1	0.063	0.242	0.251	0.490
	Database 2	0.294	0.748	0.598	2.064
	Database 3	0.022	0.121	0.099	0.294
TOC (mg/l)	Database 1	3	7	6	14
	Database 2	n.a.	n.a.	n.a.	n.a.
	Database 3	2	3	3	7
DOC (mg/l)	Database 1	n.a.	n.a.	n.a.	n.a.
	Database 2	2	5	4	12
	Database 3	2	3	2	7
Dose (mg/l)	Database 1	2	22	19	69
	Database 2	8	31	30	90
	Database 3	1	6	4	12
t (h)	Database 1	0.1	32	4	168
	Database 2	0	106	168	168
	Database 3	0.2	8	3	52
T (°C)	Database 1	10	20	20	30
	Database 2	n.a.	n.a.	n.a.	n.a.
	Database 3	n.a.	n.a.	n.a.	n.a.
Br (mg/l)	Database 1	0.010	0.303	0.151	1.245
	Database 2	0.001	0.026	0.026	0.085
	Database 3	n.a.	n.a.	n.a.	n.a.
THM (μg/l)	Database 1	4	280	176	2,843
	Database 2	15	364	310	1,560
	Database 3	16	95	82	349

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

Table 3 | Data (number of observations) for model development

	Database 1	**Database 2**	**Database 3**
MLR			
Calibration	820	465	216
Verification	205	116	54
ANN			
Training	615	349	162
Cross validation	205	116	54
Verification	205	116	54

integrated within one single variable (UV · TOC or UV · DOC) which has been found to be a good predictor of THM formation in drinking waters.

During the MLR development, it quickly became obvious that the model structure which gave the better performances, based on two criteria (coefficient of determination, R^2, and Mean Square Error, MSE), was the one with natural logarithm transformed variables (model ln-ln). This kind of model is linear in the model parameters and can be also represented by the following structure,

$$Y = \beta_0 \prod_{i=1}^{m} X_i^{\beta_i} \tag{2}$$

MLR models were developed for each database, on one hand by manipulating the independent variables pH, bromide, TOC (or DOC) and UV in forms similar to those proposed by Amy *et al.* (1987) and Rathbun (1996), and, on the other hand, by using those variables in their raw form (non-transformed variables that are taken as measured in laboratory). Table 4 presents the regression coefficients obtained for the MLR models and the values for the performance criteria, R^2 and MSE. All models shown in this table were found to be statistically significant. The results shown in Table 4 show the capacity of the MLR model for THM prediction to be quite acceptable (with an R^2 higher than 0.83 in all cases). The best MLR models for Databases 1 and 2 are those which use independent variables in their raw form. In contrast, the best models for Database 3 are those which use the independent variables in the form proposed by Amy *et al.* (1987) and the form proposed by Rathbun (1996). The models for Database 2 have higher values for R^2, in part because, as shown in Table 2, the range and the extreme values of THM concentrations are higher, which makes data fit better on average with a regression line. For the same reason, the prediction capacity represented by the MSE appears to be lower for these models than for models developed with Database 3. Indeed, R^2 and MSE are more useful when comparing different THM models (with different variables, different modelling techniques, etc.) developed from the same database than when comparing THM models calibrated with different databases (which generally have different ranges of data).

ANN models

Three-layer ANN models for THM formation were developed using the back-propagation learning algorithm, considering the same predictor variables (inputs) as used in the development of MLR models, in order to compare the predictive capability of the two modelling approaches. The ANN models have, therefore, one input layer whose size depends on the number of operational and water quality parameters, the number being different depending on the database from which the model is developed and the form in which variables are considered (whether they are in their raw form or not). For example, when developing an ANN model for Database 1, there will be seven elements in the input layer if the variables are considered in their raw form and six elements if the organic matter indicators (UV-absorbance and TOC) are considered within a single transformed variable. The output layer of the ANN models consists of a single element representing concentration of THM resulting from the bench-scale chlorination experiments.

Separate ANN models were developed from Databases 1, 2 and 3 using the software Neuroshell II (Ward Systems Group 1996). The learning process of

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

Table 4 | Results for MLR model development

	Characteristic		Regression coefficients*														
Database	**of model**	**Intercept**	***pH***	***pH* 2.6**	***pH* 3.8**	***UV***	***TOC***	***DOC***	***UV · TOC***	***UV · DOC***	***D***	***t***	***T***	***Br***	***Br*+1**	**MSE**	**R^2**
Database 1	With variables in the form proposed by Amy *et al.* (1987)	0.408	n.a.	0.730	n.a.	n.a.	n.a.	n.a.	0.410	n.a.	0.392	0.261	1.045	n.a.	0.485	**7,160**	**0.933**
	With raw variables	0.306	1.533	n.a.	n.a.	0.759	– 0.070	n.a.	n.a.	n.a.	0.427	0.260	1.023	0.051	n.a.	**7,044**	**0.934**
Database 2	With variables in the form proposed by Rathbun (1996)	13.079	n.a.	n.a.	0.497	n.a.	n.a.	n.a.	n.a.	0.420	0.290	0.283	n.a.	n.a.	– 0.900	**2,345**	**0.962**
	With raw variables	8.723	1.118	n.a.	n.a.	1.021	n.a.	– 0.215	n.a.	n.a.	0.291	0.279	n.a.	0.012	n.a.	**1,947**	**0.968**
Database 3	With variables in the form proposed by Amy *et al.* (1987)	95.202	n.a.	– 0.490	n.a.	n.a.	n.a.	n.a.	0.281	n.a.	0.542	0.121	n.a.	n.a.	n.a.	**601**	**0.832**
	With variables in the form proposed by Rathbun (1996)	70.881	n.a.	n.a.	– 0.372	n.a.	n.a.	n.a.	0.281	n.a.	0.543	0.121	n.a.	n.a.	n.a.	**601**	**0.832**
	With raw variables	235.568	– 0.618	n.a.	n.a.	0.379	0.076	n.a.	n.a.	n.a.	0.553	0.123	n.a.	n.a.	n.a.	**604**	**0.831**

*Nomenclature: *UV*: absorbance UV at 254 nm (cm^{-1}); *TOC*: total organic carbon (mg/l); *DOC*: dissolved organic carbon (mg/l); *D*: chlorine dose (mg/l); *t*: contact time (h); *T*: water temperature (°C); *Br*: bromide (μg/l).

Note: All models are statistically significant ($P<0.001$). According to the form of variables, statistical *F*-value varies from 1,045 to 1,195 for models with Database 1, from 4,658 to 4,705 for models with Database 2 and from 189 to 227 for models with Database 3.

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf by guest

ANN models (the equivalent of calibration for MLR) is undertaken using the training data set, allowing for adjustment of the connection weights of the network, as well as the cross-validation data set, which ensures that after each weight adjustment the error between the observed and the calculated output by the ANN is effectively diminishing. As explained earlier, cross-validation ensures that the model generalizes instead of memorizing. For each database, an experimental modelling process was undertaken, that is, models with different topologies were implemented by varying the *learning rate*, the *momentum* term and the number of elements in the hidden layer. The verification data were used to evaluate the accuracy of all the model topologies experimented with and to select the best ANN structure based on the two performance criteria, R^2 and MSE. The aim of the experimentation process is to identify the model architecture and the learning parameters which allow for the best predictions of THM concentrations in accordance with the actual values. For each database, both the *learning rate* and the *momentum* were varied from 0.1 to 0.9. In the same way that the learning parameters with the highest performance were identified, various sizes of the hidden layer were also experimented with until one was identified which guaranteed the best THM predictions. The results of the experimentation process with ANN, varying learning parameters and sizes for the hidden layer, are illustrated in Figure 2. In this figure, it is notable that the MSE varies enormously (in some cases two-fold) and does not follow any obvious trend as to the variation of the learning parameter and the hidden layer size, except that higher learning rates are related to higher errors. This summons the conclusion that the experimental process which uses a large range of values for these parameters is justified, in order to identify the best models. For each of the three databases, the best ANN models were found to have identical learning rates (except for one case) but very different momentum terms and sizes of the hidden layer, as presented in Table 5. Also, as shown in this table, the best modelling results for all three databases were obtained with models within which the input variables are considered in their raw form (non-transformed variables).

Comparison of MLR and ANN models

Table 6 presents, for each database, the results for models (with variables in their raw form) applied to the verification datasets. The results are presented for the verification dataset of each database (in bold) and also for specific subsets according to the chlorine dose applied during the experiments and the range of resulting THM concentrations. Performance criteria values for the best models obtained using both modelling approaches show that for Database 1, estimation of THMs based on ANN modelling is significantly more accurate than estimation with the MLR model. Estimations with the ANN model for Database 3 were also higher than those using the MLR model, but the difference is less considerable than that observed in the previous case. In the case of Database 2, the MLR model appeared to be slightly more accurate than the ANN model. Figures 3a to 3c illustrate these findings graphically. The diagonal lines in these figures represent perfect agreement between observed and predicted values of THM. For Database 1 (Figure 3a), points representing ANN results are clearly closer to the line than points representing MLR results. A similar observation is made for Database 3 (Figure 3c) even if the agreement with lines is moderate (values of R^2 lower than 0.90). For Database 2 (Figure 3b), agreement with the lines of points representing the MLR and the ANN models appears comparable, except for a set of points representing low values of THM concentrations.

To better explain these results, a more detailed analysis is necessary, comparing model performance for the specific conditions of chlorination experiments during which THMs were generated (right side of Table 6). Information from this table confirms that, in general, ANN models have a greater capacity than MLR models to predict THM formation. For Database 1, the ANN model gave better results than the MLR model for any range of chlorination conditions and resulting THMs. Figure 4 illustrates these findings for Database 1, by showing some examples of model prediction fitting with the actual data, for bench-scale experiments carried out with both low and high chlorine doses and with low and high water temperature conditions.

For Database 2, the results of Table 6 suggest that the average performance of models is highly influenced by

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

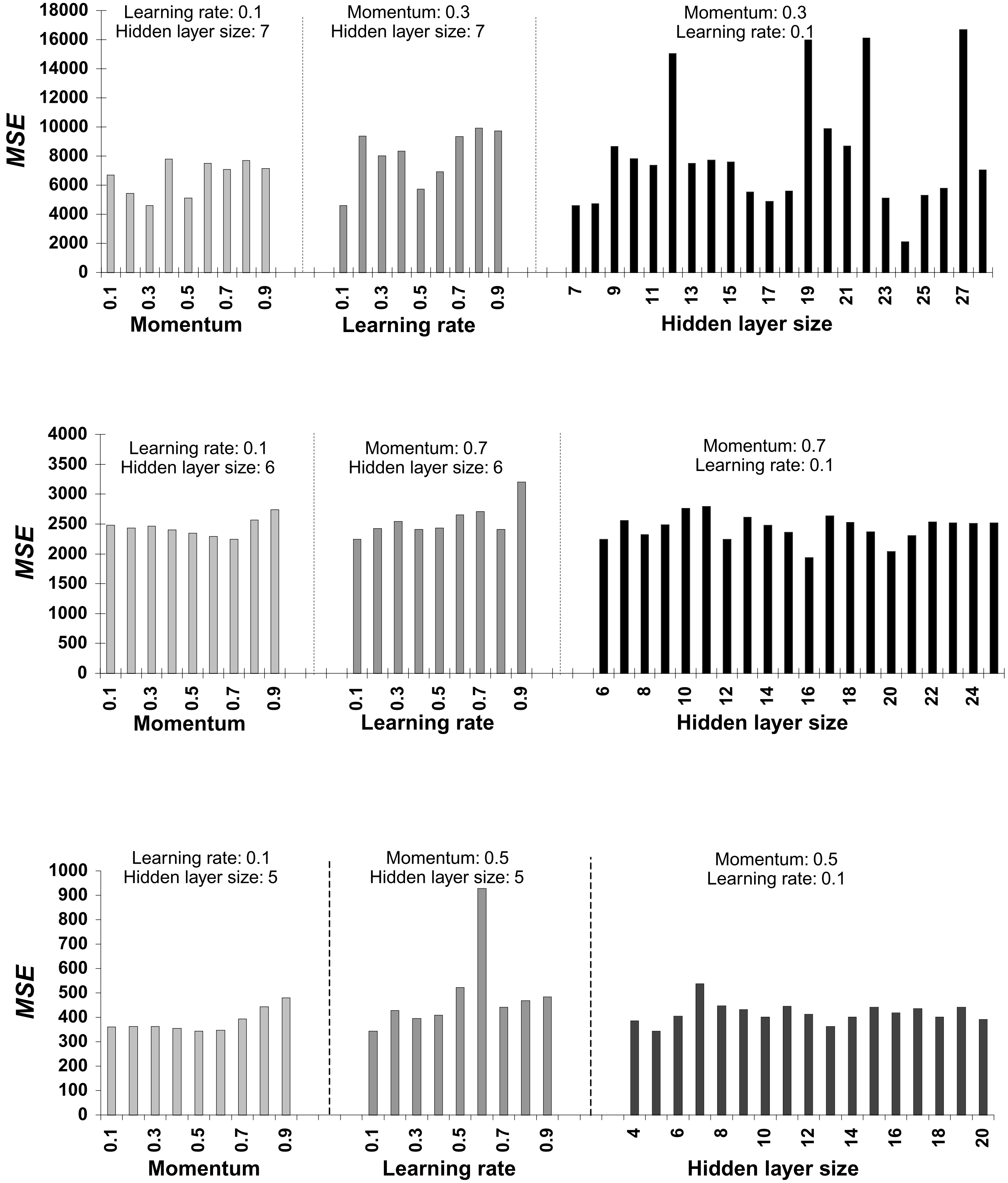

Figure 2 | Variation of ANN prediction error according to the learning parameters and the hidden layer size; (a) model with Database 1; (b) model with Database 2; (c) model with Database 3.

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

Table 5 | Results for ANN model development

	Characteristic of ANN model						
Database	Variables	Size of the input layer	Size of the hidden layer	Momentum	Learning rate	MSE	R^2
Database 1	With variables in the form proposed by Amy *et al*. (1987)	6	15	0·5	0·1	2,722	0·975
	With raw variables	7	24	0·3	0·1	1,308	0·988
Database 2	With variables in the form proposed by Rathbun (1996)	5	19	0·7	0·5	1,554	0·975
	With raw variables	6	16	0·7	0·1	1,212	0·980
Database 3	With variables in the form proposed by Amy *et al*. (1987)	4	5	0·9	0·1	427	0·881
	With variables in the form proposed by Rathbun (1996)	4	14	0·3	0·1	384	0·893
	With raw variables	5	5	0·5	0·1	324	0·909

results obtained for specific experimental conditions. Indeed, the MLR model appears to perform slightly better than the ANN model for lower chlorine doses (equal to or less than 30 mg/l) and for conditions resulting in low THM formation (less than 200 μg/l). A more careful statistical analysis of THM prediction results with Database 2 leads to the conclusion that less impressive performances of the ANN model in comparison with the MLR model were all associated with data representing instantaneous THM formation, that is, THMs formed immediately following chlorine application (considered as a contact time of some seconds). Moreover, for Database 2, measurements of instantaneous THM were carried out only in a very few experiments, and all with chlorine doses equal to or less than 30 mg/l. Also, it was noted that THMs formed at these conditions were all lower than 200 μg/l. Except for the data describing instantaneous THM formation, the ANN model gave similar or better THM predictions than the MLR model for most observations of Database 2 corresponding to experiments undertaken with a dose of 30 mg/l, as shown in the selected examples of Figure 5. Performance criteria were recalculated excluding the data for instantaneous THM formation from Database 2: for chlorine doses ranging from 15.1 to 30 mg/l, R^2 becomes 0.970 for the MLR model and 0.977 for the ANN model, whereas MSE becomes 1569 and 1201 for the MLR and ANN models, respectively. For the resulting THM levels lower than 200 μg/l, R^2 becomes 0.991 and 0.990 for the MLR and ANN models, respectively, whereas MSE becomes 415 and 473 for the MLR and ANN models, respectively. Such results are quite different than those presented in Table 6 (which includes instantaneous THMs) for the given chlorine dose and THM level ranges.

It is important to note that instantaneous THM formation was not measured during experiments leading to the creation of Databases 1 and 3. Within these two databases, initial THM formation is represented by measurements made 6 and 10 min after chlorination, respectively. At those conditions of contact times for both databases, it was found that ANN models have a better ability to predict THM formation.

There are two possible explanations for the relatively poor performance of the ANN model in comparison with the MLR model in the prediction of instantaneous THM

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

Table 6 | Comparison of model performance according to different ranges of chlorination levels and of resulting THM concentrations (with the verification data of Databases 1, 2 and 3)

		MLR		ANN			MLR		ANN			MLR		ANN	
Database	***n***	***R*²**	**MSE**	***R*²**	**MSE**	**Chlorine dose range (mg/l)**	***R*²**	**MSE**	***R*²**	**MSE**	**THM level range (μg/l)**	***R*²**	**MSE**	***R*²**	**MSE**
Database 1	205	**0.930**	**7,693**	**0.981**	**2,143**										
						0 to 16 ($n = 67$)	0.961	1,549	0.982	724	0 to 100 ($n = 62$)	0.989	572	0.991	507
						17 to 25 ($n = 74$)	0.796	6,012	0.936	1888	101 to 250 ($n = 68$)	0.840	2,340	0.912	1,290
						26 + ($n = 64$)	0.942	16,069	0.986	3,924	251 + ($n = 75$)	0.925	18,433	0.983	4,271
Database 2	116	**0·974**	**1,837**	**0·973**	**1,939**										
						0 to 15 ($n = 24$)	0.940	2,125	0.925	2,676	0 to 200 ($n = 34$)	0.994	362	0.972	1,663
						15.1 to 30 ($n = 76$)	0.975	1,424	0.971	1,657	201 to 400 ($n = 45$)	0.918	728	0.953	414
						30.1 + ($n = 16$)	0.983	3,368	0.989	2,172	401 + ($n = 37$)	0.971	4,542	0.974	4,047
Database 3	**54**	**0.869**	**351**	**0.872**	**344**										
						1 to 3.6 ($n = 19$)	0.888	199	0.950	88	16 to 60 ($n = 23$)	0.856	289	0.899	202
						3.7 to 6 ($n = 10$)	0.715	422	0.704	439	60.1 to 110 ($n = 17$)	0.446		0.610	104
						6.1 + ($n = 25$)	0.886	438	0.870	501	110.1 + ($n = 14$)	0.896	699	0.871	868

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

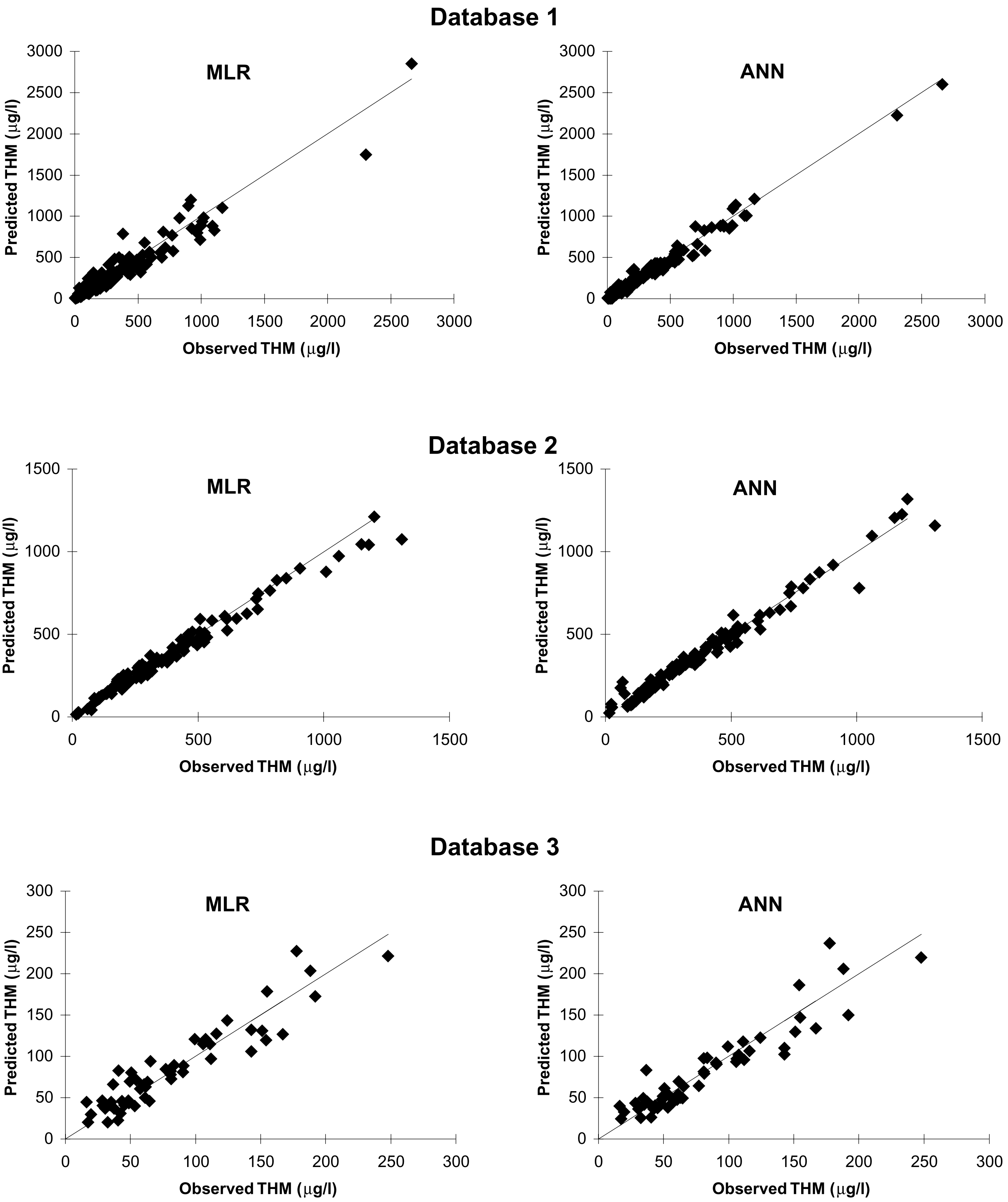

Figure 3 | Comparison of actual and predicted THM levels; (a) Database 1; (b) Database 2; (c) Database 3 (verification data of Databases 1, 2 and 3).

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

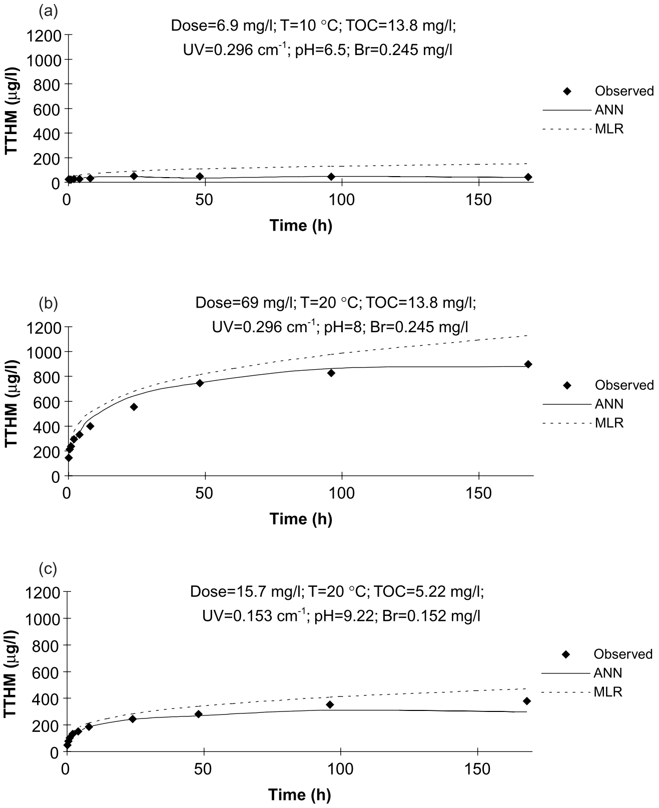

Figure 4 | Comparison of THM predictions using MLR and ANN models for specific chlorination experiments (Database 1); (a) low chlorine dose and water temperature; (b) high chlorine dose and water temperature; (c) moderate chlorine dose and high water temperature.

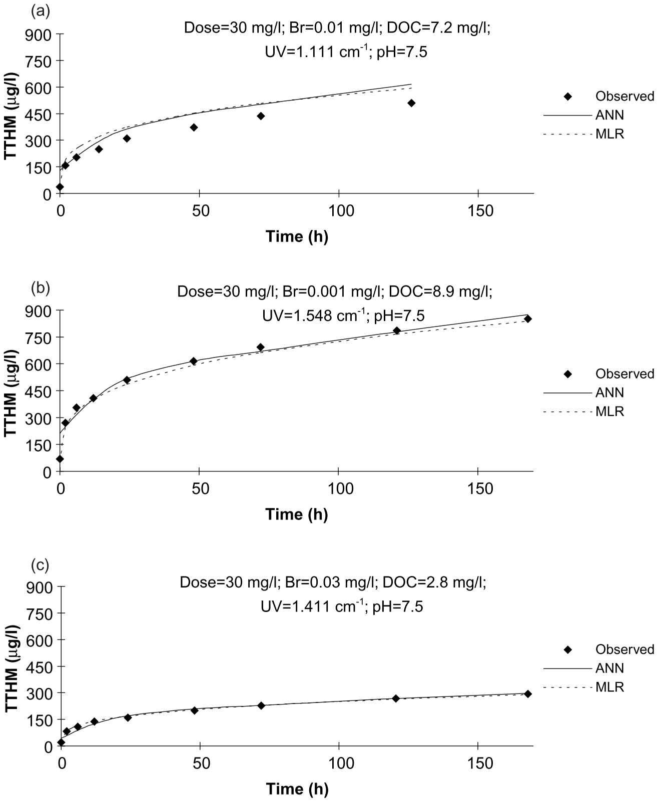

Figure 5 | Comparison of THM predictions using MLR and ANN models for experiments with chlorine dose of 30 mg/l (Database 2); (a) and (b) waters with high organic content; (c) water with low organic matter content.

formation with Database 2. The first is that instantaneous formation of THMs is a less complex phenomenon than formation over time associated with the reaction of residual chlorine with NOM, and thus can be easily modelled by a simple approach based on the OLS estimation method (as used for MLR models presented in this paper). The second reason could be that data corresponding to instantaneous formation of THMs in Database 2 being scarce (only for certain experiments with 30 mg/l as chlorine dose), there is not sufficient information to allow the ANN to learn and identify the relationship between input and output variables.

Finally, concerning model results for Database 3, differences in the average performance of models are principally due to the significantly higher performance of ANN in comparison with MLR in conditions of low chlorine doses (less than 4 mg/l) and low THM formation (less than 100 µg/l). Figure 6 illustrates this result for selected chlorination experiments.

SUMMARY

In this paper, MLR and ANN models for the prediction of THM formation during bench-scale chlorination experiments were developed and compared through the independent use of three different databases. Two of these databases were built by other researchers in the US following chlorination experiments on natural surface waters, whereas the third was implemented by the authors following chlorination of natural surface and treated waters in Quebec, Canada.

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

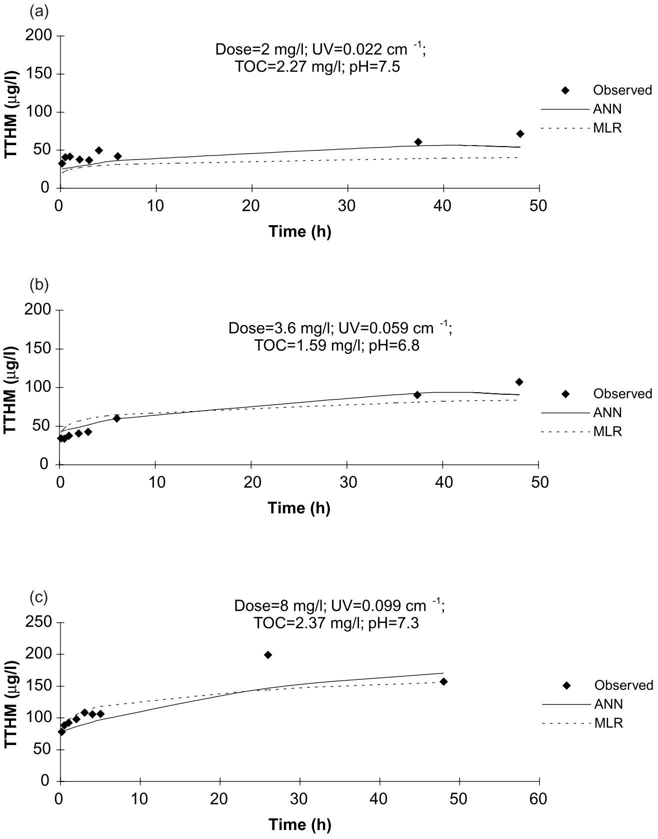

Figure 6 | Comparison of THM predictions using MLR and ANN models for specific chlorination experiments (Database 3); (a) and (b) low chlorine dose; (c) high chlorine dose.

Even if MLR models have already been developed by other researchers for predicting THM formation in bench-scale conditions with relative success, the results of this paper demonstrate that performance of such models could be significantly improved by using ANN as an alternative modelling approach. The present investigation shows that, given the three databases through which the models were developed, ANN can significantly reduce the calculated MSE between measured and predicted values. This is an important issue considering that THM formation models have several potential applications as decision-making tools for drinking water management.

Most of the models published in the literature on THM formation in bench-scale conditions are based on the MLR modelling approach. In such models, independent and dependent variables are generally submitted to ln transformation before linear estimation of parameters by the OLS method. The estimation of parameters for the MLR models presented in this paper was also carried out with the OLS method following such a transformation of variables. Even if variables influencing THM formation are previously transformed using a non-linear function (in this case, ln), such a function is imposed on all variables, thus it does not represent necessarily the best variable transformation for this problem. Also, the OLS parameter estimate provides a straight-line fit unlikely to be able to represent the optimal non-linear relationship between each water quality/ operational variable and the resulting THM levels. The obtained results demonstrate that the developed ANN models provide a very good fitted function. Moreover, the results also demonstrate the flexibility of the ANN modelling approach for THM formation in the way that input variables (factors influencing THM formation) and output variables (THM concentrations) can be considered in the learning process of ANNs in the same form in which they are measured, without any transformation or manipulation of variables before parameter estimation.

One theoretical disadvantage of ANNs in comparison with other techniques is that they are close to black-boxes, and thus the interpretation of estimated parameters (that is the connection weight matrix) as well as of the individual effect of independent variables (inputs) on the variability of dependent variables (outputs) is difficult. However, once developed, a sensitivity analysis of the ANN (carried out by varying the values of the input variables one at a time) allows for an understanding of the significance of each water quality and operational parameter at issue in the occurrence of THMs. Therefore, once ANNs are developed, it becomes possible to carry out a sensitivity analysis with these models.

According to this investigation, development of ANN models for THM prediction is a more time-consuming process compared to development of MLR models due to the experimental process required to determine the network topology and the learning parameters which minimize the errors. However, as explained in the previous paragraph, once developed, their use is easier than the use of a MLR equation.

Future research must focus on the assessment of ANN modelling techniques for other THM data, especially

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

bench-scale data generated from experiments using chlorine doses closer to those applied in real water utilities and data collected in the field at full-scale distribution systems (at different locations with variable water residence times). Further research is also necessary to confirm the hypothesis by which instantaneous THM formation is easily predicted by MLR. Research on ANN modelling is also required to evaluate multi-parameter modelling (for example simultaneous simulation of residual chlorine depletion, microbial regrowth and THM occurrence), THM speciation modelling (for example estimation of bromine incorporation factors in natural and treated waters) and modelling of other DBPs, such as haloacetic acids, the second most important group of CBPs in drinking water. In the perspective of the modelling approaches, it will also be important to investigate the use of alternative ANNs (with different learning algorithms and procedures). Finally the potential for integration of ANN models for THMs within expert system control for water treatment plants has to be assessed. To accomplish all this, it will be also necessary to generate further CBP data.

CONCLUSIONS

Results presented in this paper show that ANNs constitute a very promising alternative to MLRs for modelling THM formation in experiments conducted at bench-scale conditions. Model development and analysis with the different databases (with small and high numbers of observations) show that, in general, the ability of ANN models to estimate THM formation is better than that of MLR models. ANN models gave more accurate THM predictions than MLR models for databases resulting from experiments conducted under conditions of THM formation potential tests (relatively high chlorine doses) and under typical chlorination conditions in water utilities (relatively low chlorine doses). The performance of the two modelling approaches was, however, comparable for the database resulting from experiments conducted under tests for THM potential formation with very high chlorine doses (in particular for early THM formation).

ACKNOWLEDGEMENTS

The authors would like to thank the personnel of the six Quebec water utilities who participated in this research project as well as to thank Dr G. Amy and Dr R. Rathbun from the University of Colorado and the U.S. Geological Survey respectively, for providing bench-scale THM data. We also thank Martin Villeneuve, Marie Montixi, Michel Morin and Geneviève Sansoucy for technical field and laboratory support. This project is part of research programmes supported by the *Fonds pour la Formation de Chercheurs et l'Aide à la Recherche* (FCAR) of Quebec and the National Science and Engineering Research Council of Canada (NSERC).

REFERENCES

Amy, G. L., Chadik, P. A. & Chowdhury, Z. K. 1987 Developing models for predicting trihalomethane formation potential and kinetics. *J. Am. Wat. Wks Assoc.* **79**(7), 89–97.

APHA, AWWA & WPCF 1995 *Standard Methods for the Examination of Water and Wastewater*. 19th edition, Washington, DC.

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. *J. Wat. Suppl.: Res. & Technol. – AQUA* **48**, 129–136.

Brion, G. M., Neelakantan, T. R. & Lingireddy, S. 2001 Using neural networks to predict peak *Cryptosporidium* concentrations. *J. Am. Wat. Wks Assoc.* **93**(1), 99–105.

Cantor, K. P., Hoover, R.. Hartge, P., Mason, T. J., Silverman, D. T., Altman, R., Austin, D. F., Child, M. A., Key, C. R., Marret, L. D., Myers, M. H., Narayana, A. S., Levin, L. I., Sullivan, J. W., Swanson, G. M., Thomas, D. B. & West, D. W. 1987 Bladder cancer, drinking water source, and tap water consumption: a case-control study. *J. National Cancer Inst.* **79**(6), 1269–1279.

Caudill, M. 1991 Neural network training tips and techniques. *AI Expert* (January), 56–61.

Clark, R. M. & Sivaganesan, M. 1998 Predicting chlorine residuals and formation of TTHMs in drinking water. *J. Environ. Engrg* **124**(12), 1203–1210.

Collins, A. G., Ellis, G. W., Ford, C. & Bristol, L. E. 1991 Coupling expert systems to databases for water treatment plant control. In: *The American Water Works Association Annual Conference*. Philadelphia, PA.

Conseil Européen 1998 Directive 98/83/CE relative à la qualité des eaux destinées à la consommation humaine. *Journal Officiel* JO L330, 05/12/98, 32–52.

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest

Cook, D. F. & Wolfe, M. L. 1991 A back-propagation neural network to predict average air temperatures. *AI Applications* **5**(1), 40–46.

Crommelynck, V., Duquesne, C. & Miniussi, C. 1992 Precision tools for the daily and hourly forecasting of water consumption by a network of neurones. In: *The American Water Works Association Computer Conference*, Nashville, TN.

Engerholm, B. A. & Amy, G. L. 1983 A predictive model for chloroform formation from humic acid. *J. Am. Wat. Wks Assoc.* **75**(8), 418–423.

Hammerstrom, D. 1993 Neural networks at work. *IEEE Spectrum* (June), 26–32.

Health Canada 1996 *Guidelines for Canadian Drinking Water Quality*. Sixth edition.

Heller, M. & Singh Thind, H. 1994 Forecasting with cascade correlation: an application to potable water demand. In: *Artificial Neural Networks in Engineering Conference* (ANNIE '94). St. Louis, MO, pp. 115–1160.

Jones, W. P. & Hoskins, J. 1987 Back-propagation: A generalized delta learning rule. *Byte* (October), 155–162.

Joo, D. S., Choi, D. J. & Park, H. 2000 Determination of optimal coagulant dosing rate using an artificial neural network. *J. Wat. Suppl.: Res. & Technol.—AQUA* **49**, 49–55.

Lewis-Beck, M. S. 1980 *Applied Regression: An Introduction*. Sage Publications Inc., Newbury Park, CA.

Lippmann, R. P. 1987 An introduction to computing with neural nets. *IEEE ASSP Magazine* **4**, 4–22.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and application. *Environ. Modelling & Software* **15**, 101–124.

Menard, S. 1995 *Applied Logistic Regression Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-106, Thousand Oaks, CA.

Milot, J., Rodriguez, M. J. & Sérodes, J. B. 2000 Modeling the susceptibility of drinking water utilities to form high concentrations of trihalomethanes. *J. Environ. Mgmt* **60**(2), 155–171.

Montgomery Watson 1993 *Mathematical Modeling of the Formation of THMs and HAAs in Chlorinated Natural Waters*. American Water Works Association (AWWA), Denver, Colorado.

Neter, J., Wasserman, W. & Kutner, M. 1990 *Applied Linear Statistical Models*, 3rd edition. Irwin, Homewood, IL.

Racaud, P. & Rauzy, S. 1994 Étude de la cinétique de formation des principaux sous-produits de chloration. *TSM* **89**(5), 243–249.

Rathbun, R. E. 1996 Regression equations for disinfection by-products for the Mississippi, Missouri and Ohio rivers. *Sci. Total Environ.* **191**, 235–244.

Rodriguez, M. J. & Sérodes, J. B. 1994 Development of neural net-based models for water quality management and prediction in distribution systems. In: *Computers in Water Industry,* American Water Works Association, Los Angeles, CA, pp. 563–581.

Rodriguez, M. J., Sérodes, J. B. & Morin, M. 2000 Estimation of water utility compliance with trihalomethane regulations using a modelling approach. *J. Wat. Suppl.: Res. & Technol.-AQUA* **49**(2), 57–73.

Rodriguez, M. J. & Sérodes, J. B. 2001 Spatial and temporal evolution of trihalomethanes in three water distribution systems. *Wat. Res.* **35**(6), 1572–1586.

Rook, J. J. 1974 Formation of haloforms during chlorination of naturals waters. *Wat. Treat. & Examin.* **23**, 234–243.

Rumelhart, D. E., Widrow, B. & Lehr, M. A. 1994 The basic ideas in neural networks. *Communications of the ACM* **37**(3), 87–92.

Sérodes, J. B. &. Rodriguez, M. J. 1996 Predicting residual chlorine evolution in storage tanks within distribution systems: application of a neural network approach. *J. Wat Suppl.: Res. & Technol.-AQUA* **45**(2), 57–66.

Sérodes, J. B., Rodriguez, M. J & Ponton, A. 2001 Chlorcast©: a methodology for developing decision-making tools for chlorine disinfection control. *Environ. Modelling & Software* **16**, 53–62.

Sharfenaker, M. A. 2001 USEPA offers first glimpse of stage 2 D/DBPR. *J. Am. Wat. Wks Assoc.* **93**(12), 20–34.

Simpson, P. K. 1992 Foundations of neural networks. In: *Artificial Neural Networks: Paradigms, Applications and Hardware Implementations* (ed. E. Sanchez-Sinencio & C. Lau)., pp. 3–24. IEEE Press, New York.

SPSS France 1997 *SPSS Base 7.5 pour Windows, Manuel d'Utilisation*. Boulogne, France.

Symons, J. M., Bellar, T. A., Carswell, J. K., DeMarco, J., Krapp, K. L., Robeck, G. G., Seeger, D. R., Sloccum, C. J., Smith, B. L. & Steevens, A. A. 1975 National organics reconnaissance survey for halogenated organics. *J. Am. Wat. Wks Assoc.* **67**(11), 634–648.

USEPA 1994 *National Primary Drinking Water Regulations; Disinfectants and Disinfection Byproducts; Proposed Rule*. Federal register, 59: 145: 38667.

Ward Systems Group 1996 *NeuroShell 2 User's Manual*, 4th edition. Frederick, MD.

Wen, C. G. & Lee, C. S. 1998 A neural network approach to multiobjective optimization for water quality management in a river basin. *Wat. Res.* **34**(3), 427–436.

Zhang, Q. & Stanley, S. J. 1997 Forecasting raw-water quality parameters for the north Saskatchewan river by neural network modeling. *Wat. Res.* **31**(9), 2340–2350.

First received 6 March 2002; accepted in revised form 26 June 2002

Downloaded from https://iwaponline.com/aqua/article-pdf/52/3/199/402309/199.pdf
by guest