Forecasting Chlorine Residuals in a Water Distribution System Using a General Regression Neural Network

G.J. Bowden[a], J.B. Nixon[b], G.C. Dandy[a], H.R. Maier[a] and M. Holmes[b]

[a]Centre for Applied Modelling in Water Engineering, School of Civil and Environmental Engineering, The University of Adelaide, Adelaide, Australia

[b]United Water International Pty Ltd, Adelaide, Australia (john.nixon@uwi.com.au)

Abstract: In a water distribution system (WDS), chlorine disinfection is important in preventing the spread of waterborne diseases. By strictly controlling residual chlorine throughout the WDS, water quality managers can ensure the satisfaction and safety of their customers. However, due to the travel time of water between the chlorine dosing point and any strategic monitoring points, water treatment plant (WTP) operators often receive information too late for their responses to be effective. Given the ability to forecast the chlorine residual at strategic points in a WDS, it would be possible to have superior control over the chlorine dose, thereby preventing incidents of under- and over-chlorination. In this research, a general regression neural network (GRNN) has been developed for forecasting chlorine residuals in the Myponga WDS to the south of Adelaide, South Australia, 24 hours in advance. A number of critical model issues are addressed including: selection of an appropriate forecasting horizon; division of the available data into subsets for modelling; and, the determination of the inputs that are relevant to the chlorine forecasts. In order to determine if the GRNN is able to capture any nonlinear relationships that may be present in the data set, a comparison is made between the GRNN model and a multiple linear regression (MLR) model. When tested on an independent validation set of data, the GRNN models were able to forecast chlorine levels to a high level of accuracy, up to 24 hours in advance. The GRNN also significantly outperformed the MLR model, thereby providing evidence for the existence of nonlinear relationships in the data set.

Keywords: *Artificial neural networks; Forecasting; Chlorine residual; Water distribution system*

1. INTRODUCTION

Providing safe drinking water to consumers, free from pathogenic and other undesirable organisms, is the primary goal of all water utilities. Disinfection is an important aspect in achieving this goal and in preventing the spread of waterborne diseases. The most commonly used disinfectant in water distribution systems worldwide is chlorine (Rodriguez and Sérodes, 1999). A properly designed chlorine disinfection system provides an immediate kill of harmful bacteria and viruses and a protective residual throughout the water distribution system (WDS), thereby preventing recontamination.

Dosing too much chlorine has a number of negative effects as it increases water treatment costs and has a deleterious effect on the taste and odour properties of the water. High chlorine levels are frequently related to consumer complaints and are commonly the largest source of customer concern for water utilities. Increased chlorine levels also raise the risk of forming disinfection by-products (DBPs), which may be harmful to human health (Milot et al., 2002). Therefore, it is important to achieve a balance between the objectives of ensuring an adequate chlorine residual for microbiological quality and preventing high chlorine residuals that impact on the aesthetic qualities of the drinking water and may also pose health problems.

Water quality managers can maintain the satisfaction and safety of their customers by strictly controlling residual chlorine throughout the WDS. At the water treatment plant (WTP), it is common practice for operators to control the chlorine dose by using information about the raw water quality and the chlorine residuals at strategic points in the WDS. However, this results in a "knee-jerk" response, as this information is subject to time delays due to the travel time of water between the dosing point and the strategic point monitoring the chlorine residual. As such, the information is often received too late for the operator's response to be effective. An understanding of this problem has

led to an increase in the number of attempts to model chlorine residuals in potable water distribution systems. By forecasting the chlorine residual at strategic points in the WDS, it is possible to have greater control of the chlorine dose, thereby preventing incidents of under- and over-chlorination.

The chemical kinetics of chlorine reactions within distribution systems are not well understood because of the complexities of the reactions involved. Consequently, simple process-based models do not always adequately represent the dynamics of chlorine decay within a WDS. A large number of process-based models have been proposed, however, the performance of these models depends on good estimation of a number of chlorine decay parameters. In addition, an accurate hydraulic model of the system is required for estimation of the residence times. More recently, data-driven methods, such as artificial neural networks (ANNs), have shown their utility in forecasting chlorine residuals within distribution systems (e.g. Rodriguez and Sérodes, 1999; Sérodes et al., 2001). In this approach, historical data are collected on the chlorine residual at strategic points in the distribution network and on any variables that are likely to influence chlorine decay. Feedforward ANNs have been shown to be capable of approximating any continuous function (Hornik, 1991). Consequently, given a sufficiently representative set of data and an appropriate training algorithm, feedforward ANNs can be used to find the relationship between a set of inputs and the concentration of chlorine at a strategic point in the distribution system, at some time in the future. The advantage of this approach is that it avoids the need for a hydraulic model of the system and the underlying physical processes governing the consumption of chlorine do not need to be known explicitly.

The objective of this study is to develop an ANN model that is capable of predicting chlorine residuals in a distribution system. The case study considered in this research involves forecasting free chlorine residuals in a WDS trunk main using general regression neural networks (GRNNs). As a secondary objective, a number of fundamental issues are also addressed, including:

- What length of forecasting horizon is most suitable?
- What inputs are relevant to the chlorine forecasts?
- Does an ANN provide significant improvement over a multiple linear regression model?

2. CASE STUDY

The Myponga WTP is managed and operated by United Water International Pty Ltd. The plant is located to the south of Adelaide and serves a population of up to 45,000 people. The Myponga WTP has a chlorinator at the plant outlet, which is flow-paced. The dose rate at the plant is set manually by the operators, using their knowledge of such factors as the raw water quality, temperature and the measured chlorine residual after the filtered water storage tank.

2.1. Available Data

The system under investigation in this study spans from the Myponga WTP to the forecasting point in the trunk main approximately 20 km downstream at Aldinga (Figure 1). Flow, water temperature, and chlorine residual data from March 2002 to August 2002 have been collected for this section of trunk main at a number of locations (Salhane, 2002). These data were also used for the present study. Free chlorine residuals and water temperature were measured using analysers at the WTP, Sampson, Cactus Canyon, and Aldinga (Figure 1). Data from the analysers were recorded at five-minute intervals. For this study, these data were converted into hourly averaged values.

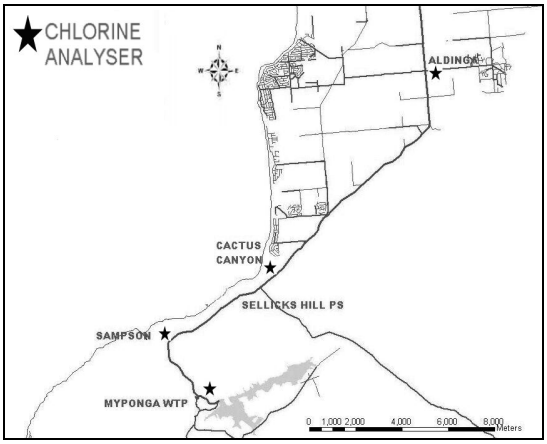

Figure 1. Myponga trunk main and the location of the chlorine analysers.

There were spans of missing and erroneous values for each of the chlorine and temperature time series. Consequently, it was decided to only use the Cactus Canyon and Aldinga chlorine time series and the Cactus Canyon water temperature time series for the periods 2:00 p.m. 26-03-2002 to 11:00 p.m. 07-05-2002 and 3:00 p.m. 19-06-2002 to 10:00 a.m. 24-07-2002, as these were the periods of reliable data. Flow data for this period of time were available from the telemetry system operated by United Water. Two flow variables were identified as important,

including: the trunk main flow, and flow at the Sellicks Hill Pump Station off-take. Data for additional variables at the Myponga WTP were also available. The additional variables identified as potential model inputs included: filtered water trunk main chlorine residual (after filtered water storage tank at the WTP), filtered water turbidity, and pH. A summary of the 8 variables used in this study is given in Table 1.

Table 1. Available data.

Variable	Location
Chlorine	Filtered Water Tank Outlet (WTP)
Chlorine	Cactus Canyon
Chlorine	Aldinga
Flow	Filtered Water Tank Outlet (WTP)
Flow	Sellicks Hill Pump Station Off-take
Temperature	Cactus Canyon
Turbidity	WTP
pH	WTP

2.2. Forecasting Horizon

An important consideration in modelling chlorine residuals in a WDS is the selection of a suitable forecasting horizon. The residence time between two points in the WDS will fluctuate depending on network demand. However, since water flowing in a pipe can be considered as a plug-flow, the optimal forecasting horizon for residual chlorine should be comparable to the average residence time in the segment being modelled (Sérodes et al., 2001). Since no hydraulic models of the Myponga WDS were available and there have been no tracer studies conducted on this segment of the WDS, an alternative method was used to determine the average residence time. In this method the cross-correlation function is computed between the time series of chlorine residual at an upstream measurement location and the time series of chlorine at the downstream forecasting location. The time shift at which the correlation between the two series is a maximum is an approximation of the average residence time between these two points for the period considered. Even though chlorine residual is a non-conservative constituent and will decay between the two points in the WDS, the fluctuations in the time series will be preserved as damped fluctuations downstream and hence, will be most highly correlated at a shift equal to the average residence time between the two points in the WDS. To compute the cross-correlation r between the two series x_i and y_i, (1) is used:

$$r_d = \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})(y_{i-d} - \bar{y})\right]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_{i-d} - \bar{y})^2}} \qquad (1)$$

where $\bar{x}$ and $\bar{y}$ are the means of the corresponding series and the cross-correlation function is computed for all shifts d = 0, 1, 2, …, n-1.

In this study, the cross-correlation function was calculated between the furthest downstream chlorine input time series (i.e. Cactus Canyon) and the chlorine at Aldinga (Figure 2). In this plot the maximum correlation occurs at a time shift of approximately 24 hours, suggesting that this is the average residence time between the Cactus Canyon and Aldinga sampling locations for this given period of data. Consequently, a forecasting horizon of 24 hours was used in this case study.

Figure 2. Cross-correlation function between Cactus Canyon chlorine and Aldinga chlorine.

3. MODEL DEVELOPMENT

The type of ANN model investigated was the general regression neural network (GRNN). The GRNN is a feedforward ANN developed by Specht (1991). GRNNs were used in this study because they are able to approximate continuous functions, only have one parameter (weight) that needs to be optimised, are very fast to train, have a fixed network architecture that does not need to be determined, and are able to model nonlinear relationships. The GRNN paradigm is briefly outlined below. Further information can be found in Specht (1991).

Assume a vector $\boldsymbol{x}$ of p independent, random variables is used to predict a dependent scalar random variable y. Let $\boldsymbol{X}$ be a particular measured value of the random variable $\boldsymbol{x}$. If the joint density $f(\boldsymbol{X}, y)$ is known, then it is possible to compute the conditional mean of y given $\boldsymbol{X}$ (or regression of y on $\boldsymbol{X}$) by using (2):

$$E[y \mid \boldsymbol{X}] = \frac{\int_{-\infty}^{\infty} y.f(\boldsymbol{X}, y)\, dy}{\int_{-\infty}^{\infty} f(\boldsymbol{X}, y)\, dy} \qquad (2)$$

If, however, the joint density $f(\boldsymbol{X}, y)$ is not known, then an estimate $\hat{f}(\boldsymbol{X}, y)$ based on a sample of observations of $\boldsymbol{x}$ and y must be used. The GRNN utilises a class of consistent nonparametric estimators known as Parzen window estimators. Using Parzen window density estimation, Specht (1991) has shown that an estimate of the conditional mean, designated $\hat{Y}(\boldsymbol{X})$, can be written as:

$$\hat{Y}(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} Y^i \cdot \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{D_i^2}{2\sigma^2}\right)} \qquad (3)$$

where σ is the smoothing parameter (sigma weight) and $D_i^2 = (\boldsymbol{X} - \boldsymbol{X^i})^T (\boldsymbol{X} - \boldsymbol{X^i})$. The regression in (3) is directly applicable to numerical data and can be easily implemented via four layers of parallel ANN architecture as shown in Figure 3. To implement (3), the A summation layer processing elements (PEs) in Figure 3 have their weights set to the actual output values i.e. $A^i = Y^i$; $i = 1, \ldots n$ and the B summation layer PEs in Figure 3 have their weights set to unity i.e. $B^i = 1$; $i = 1, \ldots n$. For the network to generalize well, the optimal sigma weight σ must be found empirically. The most common methods for determining a suitable value for the sigma weight are based on trial-and-error. The curve of mean squared error (MSE) versus σ typically exhibits a wide range of values near the minimum, and hence, it is not difficult to select a good value for σ (Specht, 1991). In addition, the curve is usually parabolic in shape, and because of this, a bracketing algorithm known as Brent's method (Press et al., 1992) was used in this research to determine a near-optimal value of σ, since this method exhibits quadratic convergence near the minimum.

3.1. Data Division

The way in which the available data are divided into subsets can have a significant influence on an ANN's performance. This is because ANNs (like other statistical and empirical models) are typically unreliable when extrapolating beyond the range of the data used for training (Bowden et al., 2002). For adequate generalisation ability, given the available data, all of the patterns that are contained in the data need to be represented in the calibration set. By choosing calibration and validation data arbitrarily, without any knowledge of which types of patterns have been included in either, the quality of the model developed, and hence the performance of the model on the validation data, has a large random component associated with it. It follows that if all of the patterns that are contained in the available data should be contained in the calibration set, then the toughest evaluation of the generalisation ability of the model is if all of the patterns (and not just a subset) are contained in the validation data. Consequently, the genetic algorithm (GA) data division method (Bowden et al., 2002) was used to divide the data into statistically representative subsets. This technique helps to ensure that the training, testing, and validation data sets are statistically representative of the same population so that a fair comparison of the models developed can be made.

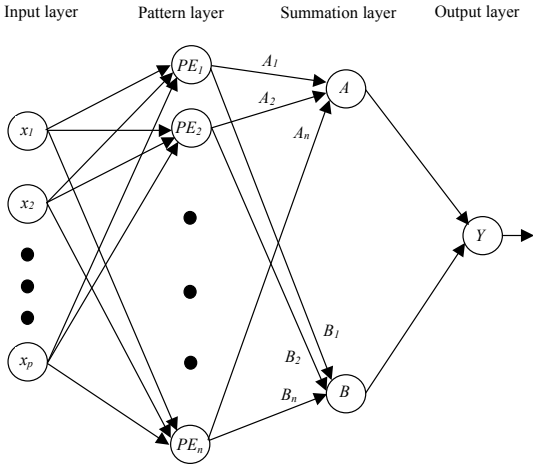

Figure 3. The GRNN architecture.

3.2. Input Determination

In this study, the maximum lag for each input variable was set at 48 hours. Given the hourly time step of the data, this was considered a sufficiently large lagging window to capture the dynamics of the system under investigation. Since there are 8 input variables (Table 1), lagging the variables resulted in a total of 384 potential model inputs. An unsupervised technique known as the self-organizing map (SOM) (Kohonen, 1982) was used to reduce the number of lags for each input variable and ensure that the remaining lags were approximately independent. In this approach, the SOM was used to cluster the lags of each input variable into groups of similar lags. By then sampling one lag from each cluster, it was possible to remove highly correlated, redundant lags from the original input set. This procedure was repeated for each of the 8 variables and reduced the total number of potential inputs to 140.

To further refine the set of candidate inputs, a new input determination technique known as the general regression input determination algorithm (GRID-A) was developed in this research. The approach proceeds as follows:

1. Identify the set of variables that could be useful predictors of the system being modelled. Denote this variable set as z_{in}.
2. Select input x_i from candidate set z_{in} and train a single-sigma GRNN model using x_i and the dependent (output) variable y. Denote the mean squared error obtained from this model as MSE_i.
3. Force independence between x_i and y by randomising x_i. Repeat for 100 bootstrap replicates of x_i.
4. Train a single-sigma GRNN model using each bootstrap and y. Compute the MSE for each model.
5. Estimate the 95th percentile randomised input MSE denoted $rand\_MSE_{95}$
6. If MSE_i is lower than $rand\_MSE_{95}$ of step 5, include the variable in the predictor set z, else discard from z_{in}.
7. Repeat steps 2-6 for all d candidate inputs in z_{in}.

GRID-A was applied to the 140 inputs derived from the SOM analysis. All 140 inputs were identified as significant since the MSE obtained using each input was less than the corresponding 95th percentile randomised input MSE. Even though all inputs were found to be significant, their relevance to the forecast can be obtained by looking at their respective MSE. Since 140 is a large number of inputs to include in a model, an input set consisting of the significant inputs with a MSE less than 0.02 was developed. This set consisted of only 21 inputs and is referred to as Input Set #1. An input set consisting of the significant inputs with a MSE less than 0.03 was also compiled. This set had 74 inputs and is referred to as Input Set #2. Finally, an input set consisting of all significant inputs was also compiled. This set had 140 inputs and is referred to as Input Set #3.

4. RESULTS AND DISCUSSION

GRNN models trained using Brent's method were developed using the 3 input subsets obtained in Section 3.2. The root mean squared error (RMSE) was used to compare the different models. The training, testing and validation results for each of these models are given in Table 2. By virtue of the GRNN's architecture (i.e. a separate pattern layer node for each training sample), the training set can be predicted to a high level of accuracy. This was evident in the low training set forecasting errors that were obtained for each model. The testing and validation sets provide a better representation of the model's generalisation ability. It can be seen that all three models exhibited good testing and validation set performance. Based on the test set performance, Model 1 produced the lowest RMSEs. Model 1 was developed using Input Set #1, which only contained 21 inputs.

Table 2. Forecasting errors for the GRNN models developed using each of the input sets.

Data Set	Model	Training Set RMSE (mg/L)	Testing Set RMSE (mg/L)	Validation Set RMSE (mg/L)
Input Set #1	1	0.0004	0.015	0.015
Input Set #2	2	0.0006	0.017	0.020
Input Set #3	3	0.0014	0.025	0.029

In Figure 4 a plot is shown of the validation set 24-hour forecasts for the model developed using Input Subset #1. It can be seen that this model produced good forecasts for the independent validation set despite the fact that it only used chlorine at Cactus Canyon and previous lags of chlorine at Aldinga as inputs. This is not surprising, given the high correlation between chlorine at Cactus Canyon and chlorine at Aldinga (Figure 2). Additional water quality parameters (e.g. turbidity and pH) were not important for the forecasts. The influence of these parameters would inherently be contained in the temporal evolution of the chlorine time series, and hence, this may explain why these parameters were not needed. Flow and temperature inputs were also not needed to produce good forecasts, for similar reasons.

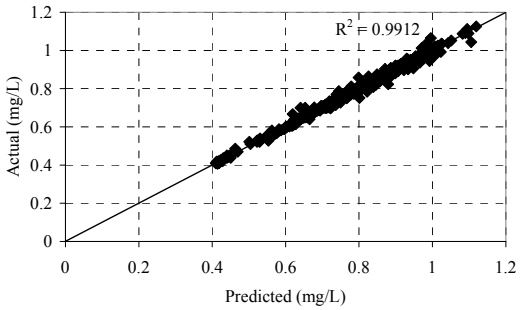

Figure 4. Validation set 24-hour forecasts for the model developed using input set #1 (Model 1).

A time series plot of the training, testing and validation forecasts produced by Model 1 is shown in Figure 5. It is evident that the chlorine forecasts were very good for this period, however, it must be noted that this plot also contains the training and testing data points, which were used

in calibrating the model. The validation set was independent of the model calibration process, and since the forecasts for this set were also good (Figure 4), it is evident that this model is capable of predicting the concentration of chlorine at Aldinga, 24 hours in advance.

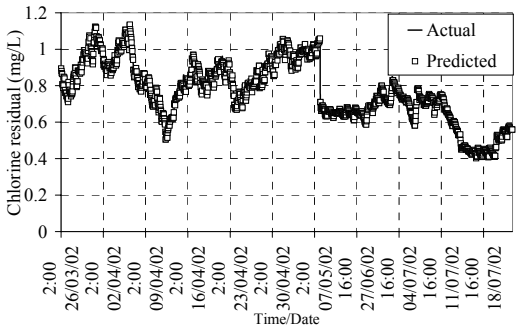

Figure 5. Training, testing, and validation set forecasts of chlorine at Aldinga, 24 hours in advance (Model 1).

To determine if the GRNN was making use of any nonlinear relationships in the data set, a comparison was conducted with a multiple linear regression (MLR) model. The best set of inputs identified in Table 2 (i.e. Input Set #1) was used for both the GRNN and MLR models. The MLR model was implemented using the R statistical package and the results of the comparison are given in Table 3. The GRNN achieved a significantly lower error for the training, testing and validation sets when compared to the MLR model. The improved performance exhibited by the GRNN model indicates that this model was able to make use of additional nonlinearities in the data set. Chlorine decay in a pipeline is a complex phenomenon, it is not surprising therefore that the GRNN was able to provide better predictions for this case study when compared to a linear regression model.

Table 3. Comparison of the GRNN with a Multiple Linear Regression model.

Data Set	Training Set RMSE (mg/L)	Testing Set RMSE (mg/L)	Validation Set RMSE (mg/L)
GRNN	0.0004	0.015	0.015
MLR	0.062	0.061	0.060

5. CONCLUSIONS

From the results obtained in this study, GRNN models were found to be useful tools for forecasting chlorine residuals in a WDS. One difficulty in applying ANNs to this type of problem is that the forecasting horizon is fixed. In this study, a method based on cross-correlation analysis was used to determine the average residence time between two points in the WDS for which chlorine time series were available. In the absence of additional information (e.g. hydraulic models or tracer studies) this is a useful approach for selecting the forecasting horizon and yielded good results for this case study.

An input determination algorithm (GRID-A) was devised in this study and was successful in determining inputs that had a significant relationship with the output variable. However, applying a tighter significance level (i.e. only selecting inputs with a MSE < 0.02) helped to further reduce input dimensionality and improved the model's performance. Only upstream chlorine levels and previous chlorine levels at the forecasting site were required to develop a successful model. GRNN models were found to significantly outperform MLR models, suggesting that they were able to make use of the nonlinear relationships in the data set.

6. REFERENCES

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

Hornik, K., Approximation capabilities of multilayer feedforward networks, *Neural Networks*, 4, 2151-2157, 1991.

Kohonen, T., Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, 43, 59-69, 1982.

Milot, J., Rodriguez, M. J., and Sérodes, J. B., Contribution of neural networks for modelling trihalomethanes occurrence in drinking water, *Journal of Water Resources Planning and Management*, September/October, 370-376, 2002.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., Numerical Recipes: The Art of Scientific Computing, Cambridge University Press, 933 pp., Cambridge, U.K., 1992.

Rodriguez, M. J., and Sérodes, J. B., Assessing empirical linear and non-linear modelling of residual chlorine in urban drinking water systems, *Environmental Modelling and Software*, 14, 93-102, 1999.

Salhane, L., Monitoring and Modelling Chlorine Levels along the Myponga Trunk Main, Masters Thesis, The University of Adelaide, 158 pp., Adelaide, 2002.

Sérodes, J. B., Rodriguez, M. J., and Ponton, A., Chlorcast(c): a methodology for developing decision-making tools for chlorine disinfection control, *Environmental Modelling and Software*, 16, 53-62, 2001.

Specht, D. F., A general regression neural network, *IEEE Transactions on Neural Networks*, 2(6), 568-576, 1991.