Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mathematical and Computer Modelling 44 (2006) 469–484
MATHEMATICAL AND COMPUTER MODELLING
www.elsevier.com/locate/mcm

Forecasting chlorine residuals in a water distribution system using a general regression neural network

Gavin J. Bowden[a,*], John B. Nixon[b], Graeme C. Dandy[c], Holger R. Maier[c], Mike Holmes[b]

[a] *Optimatics Pty Ltd, 7/62 Glen Osmond Road, Parkside, SA 5063, Australia*
[b] *United Water International Pty Ltd, GPO Box 1875, Adelaide, SA 5001, Australia*
[c] *Centre for Applied Modelling in Water Engineering, School of Civil and Environmental Engineering, The University of Adelaide, Adelaide, SA 5005, Australia*

Received 6 December 2004; accepted 11 February 2005

Abstract

In a water distribution system (WDS), chlorine disinfection is important in preventing the spread of waterborne diseases. The ability to forecast chlorine residuals at strategic points in the WDS would be a significant aid to water quality managers in helping them to ensure the satisfaction and safety of their customers. In this research, general regression neural networks (GRNNs) are developed for forecasting chlorine residuals in the Myponga WDS, to the south of Adelaide, South Australia, up to 72 h in advance. A number of critical model issues are addressed including: the selection of an appropriate forecasting horizon; the division of the available data into subsets for modelling; and the determination of the inputs relevant to the chlorine forecasts. 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. Additional investigations are also performed to simulate the effects of a reduced sampling frequency, and to estimate model performance for longer lead-time forecasts. When tested on an independent validation set of data, the GRNN models are able to forecast chlorine levels to a high level of accuracy, up to 72 h in advance. The GRNN also significantly outperforms the MLR model, thereby providing evidence of the existence of nonlinear relationships in the data set.

Keywords: Artificial neural networks; Forecasting; Chlorine residual; Water distribution system; General regression neural network

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 [1].

* Corresponding author.
E-mail addresses: gavin.bowden@optimatics.com (G.J. Bowden), john.nixon@uwi.com.au (J.B. Nixon), gdandy@civeng.adelaide.edu.au (G.C. Dandy), hmaier@civeng.adelaide.edu.au (H.R. Maier), mike.holmes@uwi.com.au (M. Holmes).

0895-7177/$ - see front matter
doi:10.1016/j.mcm.2006.01.006

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 [2]. 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 and measurement points. 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 WDSs (e.g. [1,3–7]). By forecasting the chlorine residual at strategic points in the WDS, it is possible to have greater control over the chlorine dose, thereby preventing incidents of under- and over-chlorination.

The chemical kinetics of chlorine reactions within WDSs 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, a detailed hydraulic model of the system is required for accurate 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 WDSs (e.g. [1,8]). In this approach, historical data are collected on the chlorine residual at strategic points in the WDS and on any variables that are likely to influence chlorine decay. Feedforward ANNs have been shown to be capable of approximating any continuous function [9]. 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 WDS at some time in the future. The advantage of this approach is that it avoids the need for a hydraulic model of the WDS, 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 WDS. 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 (MLR) model?
- How does the ANN model perform when developed using data that have been collected at a reduced sampling frequency?
- Can the ANN model be used for long lead-time forecasting (e.g. forecasts three days in advance)?

In this paper, contributions are also made to the ANN development methodology. In Section 3.1, a procedure based on the combination of a genetic algorithm (GA) and the Kolmogorov–Smirnov (K–S) test is introduced for determining the optimal way to divide the available data into statistically representative calibration and validation subsets. In addition, a new ANN input determination algorithm is described in Section 3.2. This approach utilizes a self-organizing map (SOM) to eliminate highly correlated, redundant predictors and a general regression input determination algorithm (GRID-A) to identify statistically significant predictors.

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, South Australia, and serves a population of up to 45,000 people (Fig. 1). The source water for the

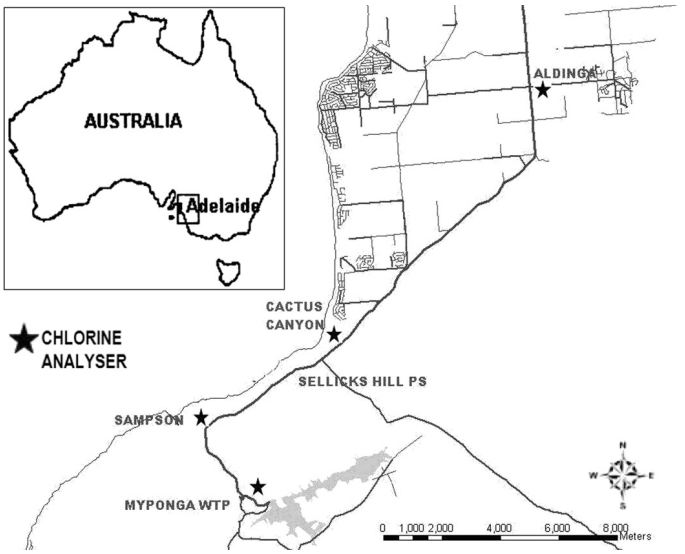

Fig. 1. Myponga trunk main, south of Adelaide, Australia, and the location of the chlorine analysers.

plant is the Myponga Reservoir and typically is high in colour and organics and low in turbidity. The plant's average daily output of drinking water during 2001/2002 was 21,000 m^3. Treatment at the plant consists of coagulation and flocculation, dissolved air flotation, filtration, pH adjustment and disinfection. 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 factors such 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 (Fig. 1). The flow, water temperature, and chlorine residual data used in this study have been collected for this section of trunk main at a number of locations by Dandy (co-author) and by [10] from March 2002 to August 2002. Free chlorine residuals and water temperature were measured using analysers at the WTP, and at the points Sampson, Cactus Canyon, and Aldinga (Fig. 1). Data from the analysers were recorded at five-minute intervals. For this study, these data were converted into hourly averaged values.

There were periods of missing and erroneous values for each of the chlorine and temperature time series. Consequently, it was decided to use only 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 these time periods were available from the telemetry system operated by United Water International. Two flow variables were identified as being important: the trunk main flow and the flow at the Sellicks Hill Pump Station off-take (Fig. 1). 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 the filtered water storage tank at the WTP), filtered water turbidity, and pH. A summary of the eight variables used in this study is given in Table 1.

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 plug-flow, the optimal forecasting horizon for residual chlorine should be comparable to the average residence time in the segment being modelled [8]. 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

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

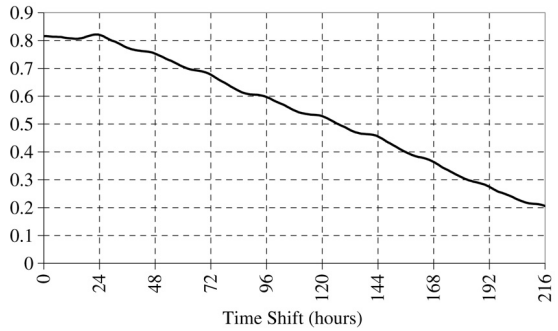

Fig. 2. Cross-correlation function between Cactus Canyon chlorine and Aldinga chlorine.

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:

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

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

In this study, the cross-correlation function was calculated between the furthest downstream chlorine input time series (i.e. Cactus Canyon) and that at Aldinga (Fig. 2). From this plot, it can be seen that the maximum correlation occurs at a time shift of approximately 24 h, suggesting that this is the average residence time between the Cactus Canyon and Aldinga sampling locations for this given period of data. Consequently, as a starting point, a value of 24 h was used for the forecasting horizon in this case study. To determine the optimal forecasting horizon, it is necessary to use a trial-and-error approach. Therefore, the effect of varying the forecasting horizon was also considered, and the results of this investigation are reported in Section 4.1.

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 [11]. GRNNs were used in this study because they are able to approximate

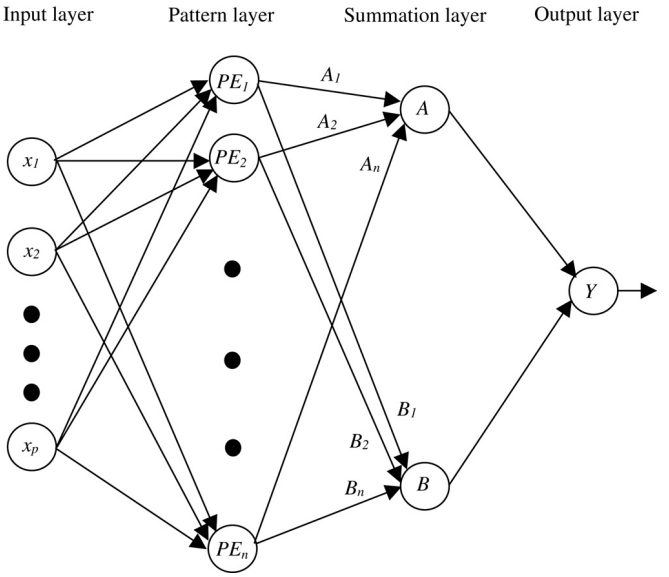

Fig. 3. The GRNN architecture.

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 optimised, and are able to model nonlinear relationships. The GRNN paradigm is briefly outlined below. Further information can be found in Specht [11].

Assume that 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 the regression of y on $\boldsymbol{X}$) by using (2):

$$E[y|\boldsymbol{X}] = \frac{\int_{-\infty}^{\infty} y.f(\boldsymbol{X}, y)\mathrm{d}y}{\int_{-\infty}^{\infty} f(\boldsymbol{X}, y)\mathrm{d}y}. \tag{2}$$

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 [11] has shown that an estimate of the conditional mean, designated $\hat{Y}(X)$, can be written as:

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

where σ is the smoothing parameter (sigma weight) and $D_i^2 = (X - X^i)^T(X - X^i)$. 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 Fig. 3. To implement (3), the A summation layer processing elements (PEs) in Fig. 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 Fig. 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 of σ are based on trial-and-error. The curve of root mean squared error (RMSE) 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 [11]. In addition, the curve is usually parabolic in shape and, because of this, a bracketing algorithm known as Brent's method [12] was used in this research to determine a near-optimal value of σ, since this method exhibits quadratic convergence near the minimum. The RMSE was used to compare each of the different models investigated in this study.

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 [13]. 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. When 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. Therefore, all of the patterns that are contained in the available data should be contained in the calibration set. Likewise, all of the patterns in the available data (and not just a subset) should be contained in the validation data, as this will provide the toughest evaluation of the generalisation ability of the model. Consequently, the genetic algorithm (GA) data division method [14] was used to divide the available 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.

To determine the "fitness" of each data division solution in the GA method, an objective function is required. In the study conducted by Bowden et al. [14], the objective function was the minimisation of the sum of the absolute difference in mean and standard deviation values between each pair of the three subsets (i.e. training, testing and validation subsets). However, this objective function does not take into account the higher-order moments of the data distributions, such as the skewness and kurtosis. Consequently, in the present study, the Kolmogorov–Smirnov (K–S) test was used to compare the distributions of the training, testing and validation sets. The K–S test measures the maximum value of the absolute difference D between two cumulative distribution functions. Under the null hypothesis (i.e. data sets drawn from the same distribution), the distribution of the K–S statistic can be calculated. To calculate the significance, the following sum is used [12]:

$$Q_{\mathrm{KS}}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}\mathrm{e}^{-2j^2\lambda^2} \tag{4}$$

which is a monotonic function with limiting values

$$Q_{\mathrm{KS}}(0) = 1 \quad Q_{\mathrm{KS}}(\infty) = 0. \tag{5}$$

The significance level of an observed value of D is calculated (as a disproof of the null hypothesis that the distributions are the same) by using the following:

$$\text{Probability}(D > \text{observed}) = Q_{\mathrm{KS}}([\sqrt{N_e} + 0.12 + 0.11/\sqrt{N_e}]D) \tag{6}$$

where N_e is the effective number of data points, and for the case of two distributions:

$$N_e = \frac{N_1 N_2}{N_1 + N_2} \tag{7}$$

where N_1 is the number of data points in the first distribution and N_2 is the number of data points in the second distribution.

In the GA data division implementation, the log of the significance probability Q_{KS} is taken, since the probabilities can differ by orders of magnitude and, by taking the log, the values are transformed, more or less, onto the same scale. The significance probabilities are computed between the validation and training sets and the validation and testing sets for each variable. Therefore, a suitable objective (fitness) function to be used by the GA is to maximise the sum of the significance probabilities between each pair of data sets. The ANN input and output data are used in this procedure and are scaled to the range [0, 1]. Scaling to the interval [0, 1] enables penalty constraints to be included more easily (i.e. the maximum and minimum values of each variable can be identified as the zeros and ones).

In this case study, a total of 1795 data records were available, from which 1436 records (80%) were used for calibration and 359 records (20%) were used for validation. The 1436 records in the calibration set were further divided into 1149 training records (80%) and 287 testing records (20%). The statistical parameters for the training, testing and validation sets obtained are shown in Table 2. From Table 2, it can be seen that the statistics are in good

Table 2
Statistics of the Aldinga chlorine training, testing and validation data sets (data divided using Genetic Algorithm—Kolmogorov Smirnov method)

Variable and data set	Mean	Standard Dev.	Max.	Min.	Inter-quartile range (IQR)
Input variable 1: Filtered water tank outlet (WTP) chlorine (mg/L)					
Training	1.97	0.27	2.83	0.00	0.32
Testing	1.95	0.26	2.54	1.18	0.32
Validation	1.96	0.27	2.75	1.21	0.30
Input variable 2: Cactus Canyon chlorine (mg/L)					
Training	1.68	0.30	2.21	0.85	0.31
Testing	1.65	0.33	2.20	0.86	0.37
Validation	1.66	0.30	2.20	0.87	0.31
Input variable 3: Aldinga chlorine (mg/L)					
Training	0.77	0.16	1.14	0.40	0.24
Testing	0.76	0.17	1.13	0.41	0.24
Validation	0.77	0.16	1.12	0.41	0.21
Input variable 4: Filtered water tank outlet (WTP) Flow (ML/day)					
Training	17.2	8.8	36.7	5.8	17.0
Testing	16.9	8.7	36.2	6.0	16.5
Validation	17.5	8.9	36.1	5.9	17.1
Input variable 5: Sellicks Hill Pump Station off-take Flow (ML/day)					
Training	3.8	4.5	13.2	0.003	9.8
Testing	3.7	4.5	13.2	0.006	9.3
Validation	3.9	4.5	13.2	0.004	9.8
Input variable 6: Cactus Canyon temp. (°C)					
Training	16.1	3.4	22.6	10.3	6.4
Testing	15.9	3.5	22.3	10.3	6.5
Validation	16.4	3.3	22.5	10.9	6.3
Input variable 7: WTP turbidity (NTU)					
Training	0.07	0.10	2.00	0.03	0.02
Testing	0.06	0.02	0.25	0.03	0.02
Validation	0.07	0.05	0.58	0.03	0.02
Input variable 8: WTP pH					
Training	7.1	0.2	8.0	1.8	0.2
Testing	7.1	0.2	7.8	6.3	0.1
Validation	7.1	0.2	7.4	6.3	0.2
Output 1: Aldinga chlorine at $(t + 24)$ h (mg/L)					
Training	0.77	0.17	1.14	0.40	0.25
Testing	0.76	0.18	1.12	0.41	0.24
Validation	0.77	0.16	1.13	0.41	0.23

agreement and that, for each variable, the training set contains the maximum and minimum values. Hypothesis tests about the difference between the means of two samples (t-test) and about the difference in the variance of two samples (F-test) were performed. For each input and output variable, the testing and validation data sets were compared with the training sets, and a significance level of $\alpha = 0.05$ was chosen. In the t-test, it was hypothesised that there was no difference between the means of the two data sets. Likewise, in the F-test it was hypothesised that there was no difference between the standard deviations of the two data sets. The null hypotheses were accepted for all variables, except that the t-test and F-test null hypotheses were rejected at the 0.05 level for the turbidity test and validation sets. However, from Table 2, it is apparent that the turbidity time series contains a maximum value of 2.00 NTU, which may possibly be an outlier since it is more than one and a half times the inter-quartile range above the third quartile. This large value was included in the training set and was most likely the cause of the t-test and F-test being rejected for the test and validation sets, respectively. Despite this one anomaly, the GA data division method was able to produce three data sets that were representative of the same statistical population.

3.2. Input determination

One of the most important steps in the ANN development process is the determination of significant input variables [15]. Usually, not all of the input variables collected will be equally informative, since some may be

Table 3
Details of the eight input subsets used in this study

Variable	Location	Number of lags selected for each input set							
		1	2	3	4	5	6	7	8
Chlorine	Filtered water tank outlet (WTP)	–	4	14	–	–	4	4	–
Chlorine	Cactus Canyon	10	17	17	11	13	7	7	17
Chlorine	Aldinga	11	16	16	13	11	5	5	16
Flow	Filtered water tank outlet (WTP)	–	21	21	–	–	5	5	14
Flow	Sellicks Hill Pump Station off-take	–	–	20	–	–	3	3	–
Temperature	Cactus Canyon	–	16	16	–	–	5	6	16
Turbidity	WTP	–	–	17	–	–	–	2	–
pH	WTP	–	–	19	–	–	6	5	–
Total number of inputs		21	74	140	24	24	35	37	63

correlated, noisy, or have no significant relationship with the output variable being modelled. The inclusion of extraneous inputs only serves to increase the computational complexity of the model and to make the training process more difficult. Therefore, there are obvious advantages in using analytical procedures to select an appropriate set of inputs for ANN models [16].

In this study, the maximum lag for each input variable was set at 48 h. 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 eight 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) [17] 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 eight 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 is 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 an MSE less than 0.02 was developed. This set consisted of only 21 inputs and is referred to as Input Set #1 (Table 3). An input set consisting of the significant inputs with an MSE less than 0.03 was also compiled. This set had 74 inputs and is referred to as Input Set #2 (Table 3). 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 (Table 3). Training GRNN models using Brent's method for each input set resulted in Models 1–3, details of which are summarised in Table 4.

3.3. Additional investigations

To address some of the critical questions that may arise in the development of an optimal ANN model for forecasting chlorine concentrations in a WDS, a number of additional investigations were conducted. These included investigating the effect of varying the forecasting horizon, comparing the GRNN model with a multiple linear

Table 4
Model characteristics and forecasting errors

Model #	Input Set #	Forecasting horizon (h)	Data resolution	Model type	Training set RMSE (mg/L)	Testing set RMSE (mg/L)	Validation set RMSE (mg/L)
1	1	24	Hourly	GRNN	0.0004	0.015	0.015
2	2	24	Hourly	GRNN	0.0006	0.017	0.020
3	3	24	Hourly	GRNN	0.0014	0.025	0.029
4	4	20	Hourly	GRNN	0.0003	0.019	0.017
5	5	28	Hourly	GRNN	0.0003	0.019	0.017
6	1	24	Hourly	MLR	0.062	0.061	0.060
7	6	24	12-hourly	GRNN	0.046	0.081	0.096
8	7	72	12-hourly	GRNN	0.077	0.114	0.083
9	8	72	Hourly	GRNN	0.0002	0.024	0.033

regression (MLR) model, considering the effect on model performance of decreasing the sampling frequency of the available data and, finally, developing a model for long lead-time forecasts (i.e. 72 h). The numerical experiments conducted are outlined below and the details of each model developed are summarised in Table 4. The results of these trials are presented and discussed in Section 4.

3.3.1. Varying the forecasting horizon

As discussed in Section 2.2, the optimal forecasting horizon for residual chlorine should be comparable to the average retention time in the WDS segment being modelled. However, since there are no hydraulic models available for the system, and no tracer studies have been conducted, the average residence time is difficult to compute. Using the cross-correlation analyses, the average residence time between Cactus Canyon and Aldinga was calculated to be approximately 24 h (Section 2.2). Therefore, this was used as the initial forecasting horizon, and the effect on the GRNN's performance of decreasing and increasing this value by 4 h was investigated.

When using a different forecasting horizon, it was not necessary to re-split the data into training, testing and validation sets, because the only changed variable was the output time series of chlorine at Aldinga. The statistics obtained for the training, testing and validation sets were compared for this new output variable. It was found that the statistics (i.e. mean and standard deviation) were very similar for each set, despite the fact that the forecasting horizon was altered. However, it was necessary to re-run the input determination procedure, since changing the forecasting horizon will mean that different lags of each input variable become significant. After applying the SOM technique for dimensionality reduction, GRID-A was used to select a set of significant model predictors. Since all 140 inputs were identified as being significant (i.e. the MSE obtained using each input was less than the corresponding 95th percentile randomised input MSE) a stricter significance level was adopted to reduce the total number of inputs. In this case, only the inputs with an MSE less than 0.02 were selected. This produced Input Sets #4 and #5 (Table 3) for the models developed using forecasting horizons of 20 h and 28 h, respectively. Using these input sets, GRNNs with forecast lead times of 20 h and 28 h were trained, thereby producing Models 4 and 5, respectively (Table 4).

3.3.2. Comparison with multiple linear regression (MLR)

To determine if the GRNN was making use of any nonlinear relationships in the data set, a comparison was conducted with a MLR model. The best set of inputs identified in Section 4 was used to develop the MLR model (Model 6) for a forecasting horizon of 24 h. The MLR model was implemented using the *R* statistical package (http://www.r-project.org/).

3.3.3. Decreasing the sampling frequency

To investigate the effect on the GRNN's performance of using data collected on a coarser time-step, the sampling frequency was reduced significantly. This was simulated by using only the data recorded at 12:00 a.m. and 12:00 p.m. on each day, resulting in a data set with a time-step of 12 h. Consequently, the 1795 data records that were available on an hourly time-step were reduced to 151 data records on a 12-hourly time-step. The forecasting horizon remained the same (i.e. 24 h).

In this investigation, the GA data division technique (Section 3.1) was used to divide the data into representative subsets. The 151 data records were divided into 121 records (80%) for calibration and 30 records (20%) for validation. The 121 records in the calibration set were further divided into 97 training records (80%) and 24 testing records (20%).

Using the GA data division technique, the statistical parameters for the training, testing and validation sets were found to be in good agreement.

The maximum lag considered for each of the eight input variables remained at 48 h. After applying the SOM and GRID-A input determination techniques, a final subset of 35 significant inputs was selected (Input Set #6, Table 3). A GRNN model trained using Brent's method was developed using the 35 significant inputs for a forecasting horizon of 24 h (Model 7, Table 4).

3.3.4. Long lead-time forecasting

As an extension of the decreased sampling frequency trial, an investigation was performed to determine the ability of the GRNN model to predict chlorine concentrations at Aldinga, 72 h in advance, while using the 12-hourly time-step data (Model 8, Table 4). This longer lead-time was identified as being more useful for longer-term forecasts of chlorine residuals, as it coincides with the current planning timeframe used by WTP managers. The data were re-split using the GA data division method and the significant inputs were determined using GRID-A. This resulted in Input Set #7 (Table 3), which included 37 significant inputs. Since the forecast horizon exceeds the average travel time in the trunk main for this period of data, the model needed to make use of the information contained in the temporal evolution of the input time series (i.e. serial dependence) and, hence, make an extrapolation in time. This is because cross-correlation with upstream chlorine time series becomes less important. However, even at a time shift of 72 h, the cross-correlation between the Cactus Canyon chlorine time series and the Aldinga chlorine time series was still relatively high at 0.68 (Fig. 2). This correlation is due in part to diurnal variations in the chlorine residuals and also because the travel time of water from Cactus Canyon to Aldinga may approach 72 h during low flow conditions, even though 24 h was the average travel time.

A trial was also conducted using the data set on an hourly time-step and a forecasting horizon of 72 h. For this investigation it was also necessary to re-run the input determination procedure. However, it should be noted that it was not necessary to conduct the SOM analysis again, since this is an unsupervised technique used to remove redundant, highly correlated inputs and, consequently, is not dependent on the output time series. To determine the significant inputs, GRID-A was applied to the 140 inputs obtained after the SOM analysis (see Section 3.2). To reduce the number of inputs, only inputs with an MSE < 0.02 were selected in the final subset. This yielded a final set of 63 selected inputs (Input Set #8, Table 3). A GRNN model trained using Brent's method was developed using the 63 significant inputs for a forecasting horizon of 72 h (Model 9).

4. Results and discussion

GRNN models trained using Brent's method were developed using the three input subsets obtained in Section 3.2. The training, testing and validation results for each of these models are given in Table 4 (Models 1–3). 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.

In Fig. 4, a plot is shown of the validation set 24 h 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 (Table 3). This is not surprising, given the high correlation between chlorine at Cactus Canyon and chlorine at Aldinga (Fig. 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.

A time series plot of the training, testing and validation forecasts produced by Model 1 is shown in Fig. 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 (Fig. 4), it is evident that this model is capable of predicting the concentration of chlorine at Aldinga, 24 h in advance.

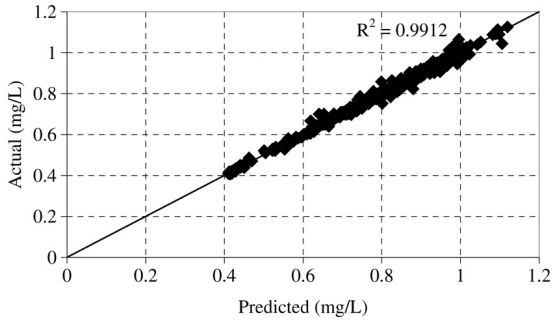

Fig. 4. Validation set 24 h forecasts for the model developed using Input Set #1 (Model 1).

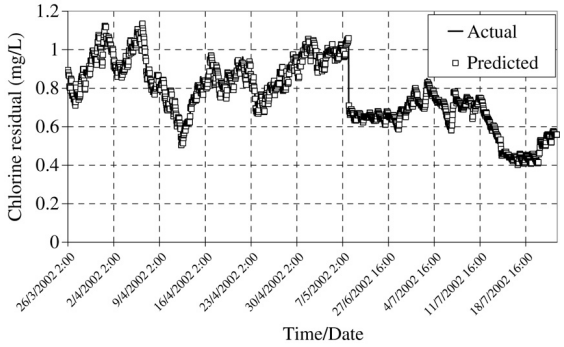

Fig. 5. Training, testing and validation set forecasts of chlorine at Aldinga, 24 h in advance (Model 1).

4.1. Varying the forecasting horizon

The training, testing and validation set forecasting errors obtained for the two additional forecasting horizons (i.e. 20 h and 28 h) investigated in this trial are presented in Table 4 (Models 4 and 5). It is apparent that increasing and decreasing the forecasting horizon caused a slight reduction in model performance when measured using the testing set RMSE. The best performance was obtained when using a forecasting horizon of 24 h (i.e. Model 1), which is in agreement with the results obtained from the cross-correlation analyses conducted in Section 2.2. Consequently, 24 h was found to be the optimal forecasting horizon for this study.

4.2. Comparison with multiple linear regression

To provide a basis for comparison with the best GRNN model (Model 1), an MLR model was developed, the performance of which is given in Table 4 (Model 6). 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 nonlinearities in the data set. Chlorine decay in a pipeline is a complex phenomenon, therefore it is not surprising that the GRNN was able to provide better predictions for this case study when compared with a linear regression model. The validation set forecasts produced by the MLR are shown in Fig. 6. The R^2 obtained by the MLR was 0.86, which is significantly lower than that achieved by the GRNN in Fig. 4 (i.e. $R^2 = 0.99$). The magnitude of the over- and under-predictions was also considerably greater.

To further investigate why the MLR performed comparatively worse, the regression statistics obtained in the analysis were examined. It was evident that there were only four inputs (independent variables) that were found to be significant at the $\alpha = 0.05$ significance level. These included: Cactus Canyon chlorine at lag 1, lag 24 and lag 27, and Aldinga chlorine at lag 1. The latter variable was identified as the most significant, as indicated by the large coefficient (0.749) and t-value (18.905). Weighting this variable so highly results in a type of naïve model that tended to lag the actual concentration of chlorine at Aldinga by approximately 24 h. Although the MLR used upstream inputs

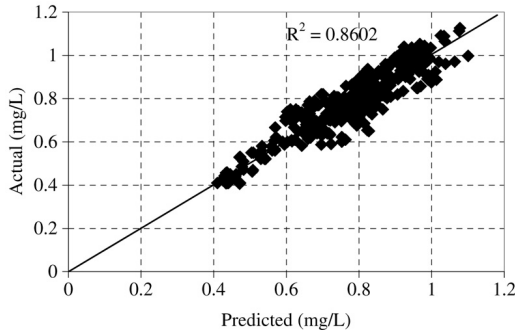

Fig. 6. Validation set 24 h forecasts for the multiple linear regression model developed using Input Set #1 (Model 6).

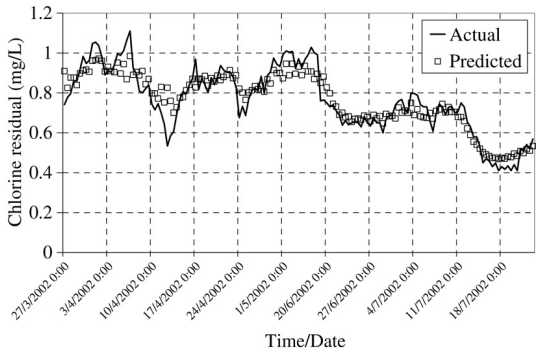

Fig. 7. Training, testing and validation set forecasts of chlorine at Aldinga, 24 h in advance using 12-hourly time-step data (Model 7).

of chlorine residual at Cactus Canyon, it was not able to use this information to model the decay that occurs by the time the water reaches Aldinga. The GRNN was able to make better use of this information and nonlinearly transform these upstream chlorine residuals to provide suitable forecasts of chlorine at Aldinga.

4.3. Decreasing the sampling frequency

The training, testing and validation results for the model developed using data collected on a 12-hourly time-step for a forecasting horizon of 24 h are given in Table 4 (Model 7). It is apparent that the results for this model were not as good as the results obtained when using a GRNN model trained with the hourly time-step data (Model 1), however the results were still of an acceptable level of accuracy. The reduced performance of this model, when compared with the model developed using hourly data, is expected since the decreased data resolution only allows for the major trends in the Aldinga chlorine time series to be modelled. This was particularly evident in the time series plot of the training, testing and validation data forecasts for Model 7 (Fig. 7), which showed that most of the major trends were accounted for, while the smaller fluctuations were not able to be modelled as well.

The independent validation set forecasts for Model 7 are shown in Fig. 8. Due to the uncertainties introduced in the sampling and measurement of chlorine residuals in a WDS, chlorine forecasts that are within ± 0.1 mg/L of the actual values are generally considered to be useful from an operational perspective. Such forecasts would enable operators to have sufficient control over the system. Consequently, this value provides a useful benchmark for assessing model performance. While the independent validation set forecasts produced by Model 7 appear to be useful from an operational perspective, there was a number of large over- and under-predictions that had an absolute error greater than 0.1 mg/L. The small data set used in this part of the study means that it is not possible to reach a firm conclusion regarding the efficacy of the GRNN model when using data on a 12-hourly time-step. The validation data set comprised only 30 observations and, hence, a larger data set would be required to confirm the viability of this approach. However, these results are promising and indicate that this approach needs to be investigated further using a larger data set.

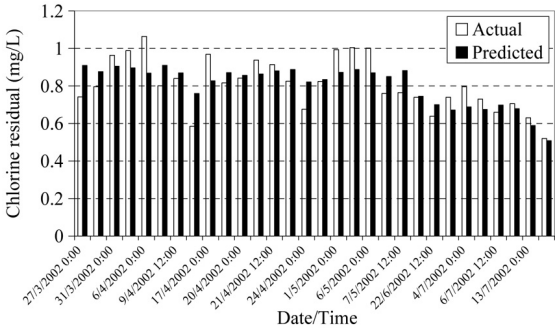

Fig. 8. Validation set forecasts of chlorine at Aldinga, 24 h in advance using 12-hourly time-step data (Model 7).

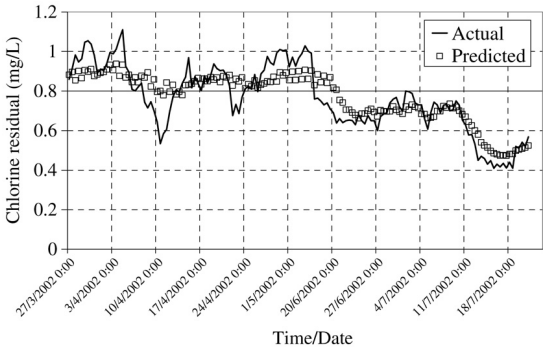

Fig. 9. Training, testing and validation set forecasts of chlorine at Aldinga, 72 h in advance using 12-hourly time-step data (Model 8).

4.4. Long lead-time forecasting

The training, testing and validation results for the model developed using data on a 12-hourly time-step for a 72 h forecasting horizon are given in Table 4 (Model 8). It can be seen that the training and testing set RMSEs for Model 8 were larger than those obtained for Model 7. This is to be expected, since the forecasting horizon for Model 8 is significantly larger than the travel time between Cactus Canyon and Aldinga, as discussed above. Therefore, as expected, the reduced correlation between Cactus Canyon and Aldinga did result in a decrease in model performance. Interestingly, the independent validation set RMSE for Model 8 was actually lower than that obtained for Model 7, however this result is most likely an anomaly due to the very small size of the validation set. Visual observation of the time series plot of the training, testing and validation data forecasts for Model 8 (Fig. 9) confirms that the forecasts were poorer than those achieved by Model 7 (Fig. 7). The major trend in the Aldinga chlorine time series was well approximated in the last four and a half weeks, but the first six weeks were not well predicted. The independent validation set forecasts for Model 8 are shown in Fig. 10. Like the validation set forecasts obtained by Model 7, the validation forecasts for Model 8 appear to be useful from an operational perspective, however there were a number of large over- and under-predictions that had an absolute error greater than 0.1 mg/L. Due to the small nature of the data set, it is also difficult to reach a firm conclusion regarding the efficacy of this GRNN model.

The training, testing and validation results for the model developed using data on an hourly time-step for a 72 h forecasting horizon are given in Table 4 (Model 9). The testing and validation set RMSEs were only slightly higher than those obtained by Model 1, which used a forecasting horizon of 24 h (Table 4). A plot of the independent validation set forecasts is shown in Fig. 11. From this figure, it is evident that Model 9 was able to produce excellent 72 h forecasts, as indicated by the high R^2 value obtained (i.e. 0.96). Also shown on this plot are ± 0.1 mg/L error lines, which represent an approximate benchmark for a suitable operational model, as discussed previously. Only a small proportion (2%) of the forecasts in the validation set exceeded these error lines, thus indicating that such a model may be quite useful if deployed in an operational environment.

A time series plot of the training, testing and validation forecasts produced by Model 9 is shown in Fig. 12. It is evident that the chlorine forecasts were good for this period, however there was a large over-prediction that occurred

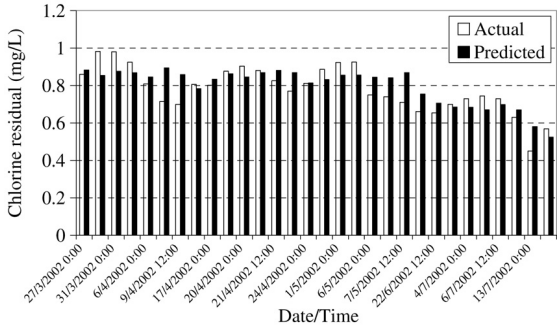

Fig. 10. Validation set forecasts of chlorine at Aldinga, 72 h in advance using 12-hourly time-step data (Model 8).

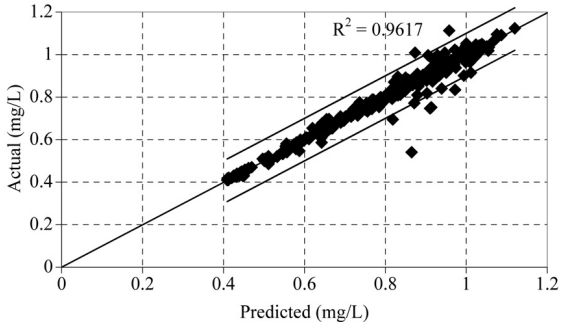

Fig. 11. Validation set forecasts of chlorine at Aldinga, 72 h in advance using hourly time-step data (Model 9).

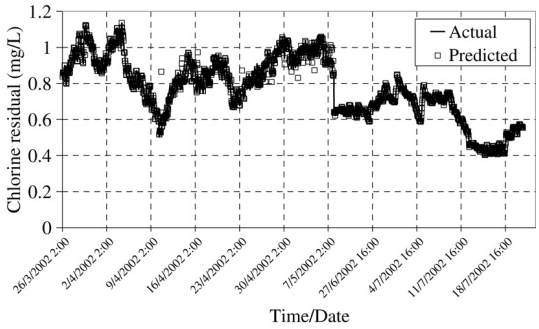

Fig. 12. Training, testing and validation set forecasts of chlorine at Aldinga, 72 h in advance using hourly time-step data (Model 9).

on 10-04-2002. With the exception of this one large over-prediction, the short- and long-term fluctuations in the time series were predicted to a high level of accuracy. This result indicates that, for this case study, it is possible to forecast the chlorine residual at Aldinga 72 h in advance, while still maintaining an acceptable level of accuracy for use as an operational model. This finding has important ramifications for WTP operators, since such a long lead-time would be very useful in managing the system and in adjusting the chlorine dose rate.

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 investigation was performed to verify that this method gave the most suitable estimate

of the average travel time in the section being modelled (i.e. forecasting horizon). In this investigation, the forecast lead-time was perturbed by ±4 h and the model development process was repeated in each case. It was found that the original forecasting horizon determined by the cross-correlation analysis (i.e. 24 h) was the most appropriate for the case study under investigation.

A genetic algorithm (GA) data division technique was developed and tested for dividing the available data set into representative modelling subsets (i.e. training, testing and validation sets). The fitness function in this technique was derived from the Kolmogorov–Smirnov (K–S) test. This approach was found to be successful at dividing the available data into three representative subsets. Examination of the statistical properties of each subset revealed that they were in good agreement, which verified that they were representative of the same statistical population.

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 an MSE < 0.02) helped to reduce input dimensionality further 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.

The GRNN models developed in this study were found to outperform MLR models significantly, suggesting that they were able to make use of the nonlinear relationships in the data set. Unlike the MLR model, the GRNN was better able to make use of the upstream chlorine information at Cactus Canyon to ensure that the timing of the predictions was accurate. The MLR placed a large emphasis on the most recent Aldinga chlorine input (i.e. lag 1) and consequently acted as a type of naïve model that simply lagged the actual values by approximately 24 h.

To investigate the effect of decreased data resolution, the original hourly data set was reduced to a 12-hourly set. It was found that using the 12-hourly data, a GRNN model could be developed that was capable of providing suitable 24 h chlorine forecasts at Aldinga. As expected, day-to-day fluctuations were not modelled well, however longer-term variation was well predicted. Repeating the model development procedure using the 12-hourly data set and forecasting chlorine at Aldinga 72 h in advance revealed that the forecasts were not acceptable from an operational perspective. However, when the process was repeated using the hourly data, the 72 h forecasts improved significantly. When this model was tested on an independent validation set, a coefficient of determination (R^2) equal to 0.96 was obtained, indicating that this model was suitable from an operational perspective.

In this study, ANNs have been found to be capable of forecasting chlorine residuals in the Myponga trunk main at Aldinga, up to 72 h in advance. The next level of complexity in this research is to repeat the model development procedure using a larger data set that also includes data for all seasons. A model deployment trial should be investigated whereby different retraining scenarios are tested to determine how the model would perform when deployed in a real-time operational environment. The modelling approach also needs to be extended to the forecasting of chlorine at other strategic points in the WDS. This can only be achieved by placing chlorine analysers at these locations to collect the relevant data needed to develop the model. Finally, the model development process described here should be applied to other WDS case studies to validate its efficacy further.

References

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

[2] J. Milot, M.J. Rodriguez, J.B. Serodes, Contribution of neural networks for modelling trihalomethanes occurrence in drinking water, Journal of Water Resources Planning and Management (September/October) (2002) 370–376.

[3] J.J. Vasconcelos, W.M. Grayman, L. Kiene, O. Wable, P. Biswas, A. Bhari, L.A. Rossman, R.M. Clark, J.A. Goodrich, Characterization and Modeling of Chlorine Decay in Distribution Systems, AWWA Research Foundation, Denver, CO, 1996.

[4] R.M. Clark, J.A. Coyle, Measuring and modeling variations in distribution system water quality, Journal AWWA (August) (1990) 46–53.

[5] M.J. Rodriguez, J.B. Sérodes, P.A. Cote, Advanced chlorination control in drinking water systems using artificial neural networks, Water Supply 15 (2) (1997) 159–168.

[6] M.J. Rodriguez, J.R. West, J. Powell, J.B. Sérodes, Application of two approaches to model chlorine residuals in Severn Trent Water Ltd (STW) distribution systems, Water Science and Technology 36 (5) (1997) 317–324.

[7] V.K. Chambers, J.D. Creasey, J.S. Joy, Modelling free and total chlorine decay in potable water distribution systems, Journal of Water Supply Research and Technology-Aqua 44 (1995) 60–69.

[8] J.B. Sérodes, M.J. Rodriguez, A. Ponton, Chlorcast©: A methodology for developing decision-making tools for chlorine disinfection control, Environmental Modelling and Software 16 (2001) 53–62.

[9] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Networks 4 (1991) 2151–2157.

[10] L. Salhane, Monitoring and Modelling Chlorine Levels along the Myponga Trunk Main, School of Civil and Environmental Engineering, The University of Adelaide, Adelaide, 2002, p. 158.
[11] D.F. Specht, A general regression neural network, IEEE Transactions on Neural Networks 2 (6) (1991) 568–576.
[12] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes: The Art of Scientific Computing, Cambridge University Press, Cambridge, UK, 1992.
[13] I. Flood, N. Kartam, Neural networks in civil engineering. I: Principles and understanding, Journal of Computing in Civil Engineering 8 (2) (1994) 131–148.
[14] G.J. Bowden, H.R. Maier, G.C. Dandy, Optimal division of data for neural network models in water resources applications, Water Resources Research 38 (2) (2002) 2-1–2-11 (10.1029/2001WR000266).
[15] H.R. Maier, G.C. Dandy, Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications, Environmental Modelling and Software 15 (2000) 101–124.
[16] G.J. Bowden, G.C. Dandy, H.R. Maier, Input determination for neural network models in water resources applications: Part 1—Background and methodology, Journal of Hydrology 301 (1–4) (2005) 75–92. http://www.sciencedirect.com/science/journal/00221694.
[17] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43 (1982) 59–69.