HYDROLOGICAL PROCESSES
Hydrol. Process. **19**, 137–148 (2005)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/hyp.5763

Short-term municipal water demand forecasting

John Bougadis,[1] Kaz Adamowski[2]* and Roman Diduch[3]

[1] *Delcan Corporation, 1223 Michael St, Ottawa, Ontario K1J 7T2, Canada*
[2] *Department of Civil Engineering, University of Ottawa, Ontario K1N 6N5, Canada*
[3] *Water and Wastewater Infrastructure, City of Ottawa, Ontario K2P 2L7, Canada*

Abstract:

Water demand forecasts are needed for the design, operation and management of urban water supply systems. In this study, the relative performance of regression, time series analysis and artificial neural network (ANN) models are investigated for short-term peak water demand forecasting. The significance of climatic variables (rainfall and maximum air temperature, in addition to past water demand) on water demand management is also investigated.

Numerical analysis was performed on data from the city of Ottawa, Ontario, Canada. The existing water supply infrastructure will not be able to meet the demand for projected population growth; thus, a study is needed to determine the effect of peak water demand management on the sizing and staging of facilities for developing an expansion strategy. Three different ANNs and regression models and seven time-series models have been developed and compared. The ANN models consistently outperformed the regression and time-series models developed in this study. It has been found that water demand on a weekly basis is more significantly correlated with the rainfall amount than the occurrence of rainfall. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS water demand management; water supply; artificial neural networks; forecasting

INTRODUCTION

The projected population and employment growth of the Ottawa region in Canada, coupled with high peak water use, requires the expansion of water supply and distribution facilities. In order to have a least-cost infrastructure expansion strategy, it is necessary to develop a water demand forecast model. The data derived for the 3W pressure zone was used in this study. The 3W pressure zone encompasses most of the rapidly expanding areas of Kanata and Stittsville (West Urban Center, or WUC). This region contains a reservoir and pump station that provide the community with their water supply. It is expected that the water supply and distribution infrastructure (water mains, pumps, elevation and reservoir tanks) will not be able to provide the community with peak outdoor water demand for the projected growth of the region. The residential population is predicted to increase from 62 228 recorded in 2001 to 160 175 for the year 2021. The employment population is also expected to increase from 17 618 in 2001 to 57 951 in the year 2021.

The peak hour outdoor water use (watering lawns, sprinklers) is of particular concern in sizing and selecting the water facilities servicing the 3W region. The average day demand has increased from 17·8 Ml day^{-1} in 1993 to 28·7 Ml day^{-1} in 2002, and the maximum peak demand has also significantly increased from 67·8 Ml day^{-1} in 1993 to 109·3 Ml day^{-1} in 2002. Because of the projected growth of the 3W region, a water demand forecasting methodology is needed to provide an in depth analysis and assessment of the factors affecting peak water usage. There are different approaches to water demand forecasting, including various statistical techniques, such as regression and time-series analysis.

Water demand forecasts are used for several purposes, such as: planning new developments or system expansion; to estimate the size and operation of reservoirs, pumping stations and pipe capacities; and for

* Correspondence to: Kaz Adamowski, Department of Civil Engineering, University of Ottawa, Ottawa, Ontario K1N 6NS, Canada. E-mail: adamowsk@eng.uottawa.ca

Received 13 May 2003
Accepted 11 December 2003

urban water management issues (e.g. pricing policy, water use restrictions, etc.). Long-term forecasting is required mainly for planning and design, and short-term forecasting is useful in operation and management. The main purpose of the current study is to investigate techniques of regression and time-series analysis and an artificial neural network (ANN) for short-term forecasting of water use.

Perhaps the most frequently adopted methods for forecasting have been regression and time-series analysis (Jain *et al*., 2001). Jain *et al*. (2001) developed multiple linear regression models by using weekly maximum air temperature, weekly rainfall amount, weekly past water demand, and the occurrence and non-occurrence of rainfall as parameters for their models. The occurrence of rainfall was inputted as a binary series, where one was assigned to weeks with rainfall and zero for weeks without rainfall. Graeser (1958) used linear regression models with the number of previous days of maximum air temperature above 100 °F and the number of weeks since the occurrence of rainfall was above 1 in. Howe and Linaweaver (1967) developed a series of regression models covering domestic demands, summer sprinkler demands, and maximum sprinkler demands. Climatological variables were used in their regression models, such as summer precipitation, summer potential and maximum day potential evapotranspiration, average summer sprinkling demand, and irrigable area per household. Howe and Linaweaver (1967) also included a series of pricing factors into their regression models, such as marginal commodity charge to average summer total use, and value market of household.

Another general approach to water demand forecasting is based on time-series analysis. Jain *et al*. (2001) used autoregressive models of orders 2 and 3 to forecast peak weekly water demand. Maidment *et al*. (1985) used short-term Box and Jenkins models for daily municipal water use, which were a function of rainfall and air temperature. Maidment and Miaou (1986) applied this model to the water consumption from nine cities in the USA. The coefficient of determination R^2 was used to verify the accuracy of their models and values ranged from 0·48 (Allentown, Pennsylvania) to 0·96 (Austin, Texas). Smith (1988) developed time-series models to forecast daily municipal water demand, which included day of week effects and a randomly varying mean, two factors that were not included in Maidment models. Zhou *et al*. (2000) developed time-series models for daily water consumption in Melbourne, Australia. Their models included trend, seasonality, climatic correlation and autocorrelation components.

In recent years, the technique of ANNs has been used with the back-propagation algorithm for several civil engineering applications (Lingireddy and Ormsbee, 1998). ANN models have been used to model daily and hourly water demand forecast (Crommelnyck *et al*., 1992) and for weekly peak demand (Jain *et al*., 2001). They can be formulated as a function of climatological variables (such as air temperature, volume and the occurrence of rainfall) and previous water demand. Complex ANN models have been implemented for daily water demand predictions for the city of Regina's water distribution system (Lertpalangsunti *et al*., 1999). Obeysekera *et al*. (2000) used ANN models to predict inflow volumes to Lake Okeechobee to benefit the operation of regional water management systems that include lakes and storage reservoir for storage control and water supply.

In this paper, an exploratory analysis was initially performed to quantify the relationship of climatological variables with peak demand. Three different statistical techniques were used to forecast short-term water demand. Linear and multiple linear regression models were hypothesized using weekly maximum temperature, weekly rainfall amounts, the occurrence of rainfall in a week, and the peak water demand. Time-series models were applied to the weekly peak demand series to determine an autoregressive integrated moving-average (ARIMA) model that best fit the observed data. ANN models were developed using similar input variables to those used in the regression analysis.

DATA

Numerical analysis was performed on weekly peak water demand (megalitres per day) obtained from the 3W pressure zone located in the Ottawa region. Weekly average maximum temperature (centigrade) and weekly total rainfall (millimetres) were obtained from Environment Canada. The water demand series record length ranged from 1993 to 2002. Only the summer months (1 May to 3 September) were used in the analysis, since peak demand usage occurs in summer months for a given year. The weekly peak water demand series was divided into 'training' and 'testing' data sets. The former set began in 1993 and ended in 2001, and contained 162 weeks of peak water demand, temperature and rainfall data. The latter contained 18 weeks of data for the year 2002. The performance of all statistical models was analysed by comparing the known peak water demand data in the year 2002 with predicted values obtained from different models.

METHODOLOGY

Exploratory analysis

The preliminary step in investigating which variables influence water demand is to determine if any seasonal patterns exist in the water demand series, and explore the cross-correlation coefficients between climatological variables and the water demand series.

The seasonal or periodic component can be detected by using the Fourier analysis, as defined by (Kite and Adamowski, 1973)

$$x(t) = a_0 + \sum_{m=1}^{N/2} a_m \cos\left(\frac{2\pi mt}{N}\right) + b_m \sin\left(\frac{2\pi mt}{N}\right) \tag{1}$$

where a_0 is the mean of the data set, m is the number of harmonics, N is the record length of the data set, and $t = 1, 2, \ldots, N$ is the time. The coefficients of the Fourier series (a_m and b_m) when $m \neq N/2$ are defined as

$$a_m = \frac{2}{N}\sum_{m=1}^{N} x_t \cos\left(\frac{2\pi mt}{N}\right) \qquad b_m = \frac{2}{N}\sum_{m=1}^{N} x_t \sin\left(\frac{2\pi mt}{N}\right) \tag{2}$$

whereas for $m = N/2$

$$a_m = \frac{1}{N}\sum_{m=1}^{N} x_t(-1)^t \qquad b_m = 0 \tag{3}$$

If s^2 is the total variance of the time series $x(t)$, then that part of the variance C_m^2 explained by the mth harmonic is defined by

$$C_m^2 = \frac{a_m^2 + b_m^2}{2s^2} \tag{4}$$

Cross-correlation examines dependency between two time series. The cross-correlation coefficient can be estimated by the sample cross-correlation as follows (Box and Jenkins, 1976):

$$r_{xy} = \frac{\sum_{t=1}^{n}(x(t) - \overline{x})(y(t) - \overline{y})}{\left[\sum_{t=1}^{n}(x(t) - \overline{x})^2\right]^{1/2}\left[\sum_{t=1}^{n}(y(t) - \overline{y})^2\right]^{1/2}} \tag{5}$$

where x and y denote the time series of two different variables. The variables used to compute cross-correlation can be seen in Table I.

Linear and multiple linear regression

Regression analysis was used to investigate the linear relationship between peak water demand and climatological variables. Linear and multiple linear regression models were used.

Table I. Cross-correlation coefficients between peak water demand and climatological variables

Variable	Demand (t)
Rainfall (t)	−0·3523
Rainfall ($t-1$)	−0·3129
Rain occurrence (>25·4 mm)	0·0575
Rain occurrence (>10 mm)	0·0357
Rain occurrence (>5 mm)	0·0221
Rain occurrence	−0·0606
Temperature (t)	0·4917
Temperature ($t-1$)	0·2717
Demand ($t-1$)	0·5883

Table II. Results found from Fourier analysis

Harmonic (months)	Frequency	Period (weeks)	Explained variance C_m^2
3	0·0185	54·0	4·92
4	0·0247	40·5	6·22
5	0·0309	32·4	5·42
6	0·0370	27·0	3·83
9	0·0556	18·0	6·71
10	0·0617	16·2	4·27
16	0·0988	10·1	5·01
20	0·1235	8·1	7·83
23	0·1420	7·0	4·16
30	0·1852	5·4	3·54

Simple linear regression was used to determine the relationship between peak water demand for the current week D_t and the following variables: average weekly maximum temperature for the current week T_t, average weekly maximum temperature for the previous week T_{t-1}, weekly total rainfall for the current week R_t, weekly total rainfall for the previous week R_{t-1}, and previous weekly peak demand D_{t-1}. The coefficients of determination for variables analysed can be seen in Table II.

In this study, three multiple linear regression models were developed for weekly peak demand forecasts. The first model, called MLR-1, is a function of weekly peak demand in the previous week, weekly average maximum temperature and weekly total rainfall of the current week (Jain *et al*., 2001):

$$D_t = \beta_0 + \beta_1 D_{t-1} + \beta_2 T_t + \beta_3 R_t \tag{6}$$

The second model, denoted MLR-2, is similar to the first model, but the weekly average maximum temperature and total rainfall amount from the previous week are added to the regression model (Jain *et al*., 2001):

$$D_t = \beta_0 + \beta_1 D_{t-1} + \beta_2 T_t + \beta_3 R_t + \beta_4 T_{t-1} + \beta_5 R_{t-1} \tag{7}$$

The third model is similar to Equation (6); however, the actual rainfall amount is replaced by the occurrence or non-occurrence of rainfall for the given week, which is denoted by the β coefficient in the multilinear regression equations to follow. For this MLR-3 model, if any rain has occurred, then the coefficient $B = 1$; if no rain has occurred, then the coefficient $\beta = 0$ (Jain *et al*., 2001):

$$D_t = \beta_0 + \beta_1 D_{t-1} + \beta_2 T_t + \beta_6 B R_t \tag{8}$$

where coefficients β_0 to β_6 are found from the regression equations.

Time-series analysis

Box and Jenkins (1976) time-series models have been used to model short-term water demand. Time-series analysis investigates models governing the process of water demand. The first step in the Box–Jenkins approach is to determine whether the water demand time series is stationary (data set having a constant mean and variance) and whether it has a significant seasonality component. Non-stationary data sets can often be detected by determining the autocorrelation coefficient function (ACF) defined by (Box and Jenkins, 1976)

$$r_k = \frac{\sum(x(t)x(t+k)) - \frac{1}{n-k}\sum x(t)\sum x(t+k)}{\left[\sum x^2(t) - \frac{1}{n-k}\left(\sum x(t)\right)^2\right]^{1/2}\left[\sum x^2(t+k) - \frac{1}{n-k}\left(\sum x(t+k)\right)^2\right]^{1/2}} \tag{9}$$

where x is the variable, t is the time, and all summations are carried out from $t = 1$ to $t = n - k$. Equation (9) determines the degree of correlation between observations that are separated by k time units.

If there is a strong dependence in the ACF plot, meaning no quick decay to zero, then the data set is not stationary. To make a data set stationary, Box and Jenkins (1976) recommend a differencing approach, which is defined by

$$w_j = D_j - D_{j-1} \tag{10}$$

where w_j is a new series that is stationary and D_j represents the peak demand at time j. Seasonality in the data set can also be seen by a periodogram, which has been discussed in the previous section. If significant seasonality is detected, then a seasonality term can be added to the stochastic model.

Once stationary conditions have been satisfied, the ACF and partial ACF (PACF) plots are used to identify the stochastic model. If the data are stationary without differencing, then autoregressive moving average (ARMA) models are used. These ARMA(p,q) models are defined by autoregressive components of order p and moving-average components of order q. The equation describing ARMA(p,q) models is (Box and Jenkins, 1976)

$$D_t = \mu + \sum_{j=1}^{p}\alpha_j(D_{t-j} - \mu) - \sum_{j=1}^{q}(\theta_j\varepsilon_{t-j}) + \varepsilon_t \tag{11}$$

where p represents autoregressive parameters $\alpha_1, \ldots, \alpha_p$, and q moving-average parameters $\theta_1, \ldots, \theta_q$. The μ term is the mean of the peak water demand series, and ε_t is the uncorrelated normal random variable (also referred to as white noise) with a mean of zero and constant variance.

If the data are transformed into a stationary model, then the ARIMA model of order p, d, and q is used, where d is the number of times the data are differenced. Usually, one difference of the data set is required to transform the series into a stationary process. The equation defining an ARIMA(p,d,q) process is

$$D_t = D_{t-1} + \sum_{j=1}^{p}\alpha_j(w_{j-1}) - \sum_{j=1}^{q}(\theta_j\varepsilon_{t-j}) + \varepsilon_t \tag{12}$$

where all the variables are as previously defined.

Various tests exist to determine whether the model chosen accurately describes the observed time series. In this paper, the Akaike information criterion (AIC) was used to verify the model (Box and Jenkins, 1976):

$$\mathrm{AIC} = N\ \ln(\text{MLE of residual variance}) + 2(p+q) \tag{13}$$

where MLE stands for the maximum likelihood estimate of the residual variance.

ANNs

ANNs are a class of mathematical models that function similar to biological processes of the brain. ANN models are comprised of user-defined inputs (rainfall, temperature, etc.) and desired output (prediction of peak demand) that are connected by a set of highly interconnected nodes arranged in a series of layers. These nodes are connected to the user-defined inputs and to the desired output. Figure 1 illustrates an ANN used in this paper with one input layer, one hidden layer and one output layer (Lingireddy amd Ormsbee, 1998).

The main advantage of using ANN models is the capability of the network to self-learn. Knowing the inputs and desired output(s), the ANN model will try to reproduce the observed outputs through a series of iterations. The most common ANN network is the feed-forward network, which uses the back-propagation algorithm for training. In this process, the selected inputs are directed to the nodes in the input layer and are propagated forward to compute the output vector using randomly selected initial weights in the hidden layer and the use of a non-linear function called an activation function, which often is assumed to be a continuous differentiable sigmoidal logistical function (Lippman, 1987):

$$f(x) = \frac{1}{1 + e^{-x}} \tag{14}$$

The error from the output vector computed by the ANN model is calculated knowing the observed output. This computed error is then back-propagated through the network and the initial randomly selected weights are updated using the following equation (Lippman, 1987):

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_j y_j + \alpha(w_{ij}(t) - w_{ij}(t-1)) \tag{15}$$

where $w_{ij}(t+1)$ is the weight from hidden node i or from an input to node j at time t, η is the learning coefficient, y_j is the actual output at node j, α is the momentum correction factor (which speeds up the convergence), and δ_j is the error term defined as (Lippman, 1987)

$$\delta_j = y_j(1 - y_j)(d_j - y_j) \tag{16}$$

where d_j is the desired demand at node j. Equation (16) is used if node j is an output node; however, if node j is a hidden node then the error term becomes (Lippman, 1987)

$$\delta_j = y_j(1 - y_j)\sum \delta_k w_{jk} \tag{17}$$

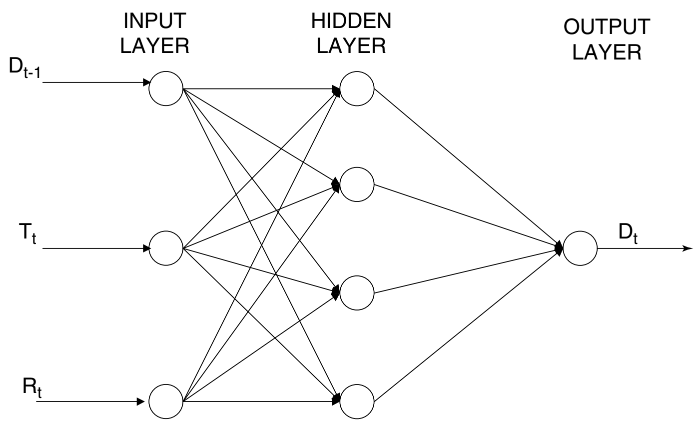

Figure 1. The ANN network architecture

The output vector is recalculated based on the adjusted weights, and the error between the observed data and model predictions is recalculated. The weights are then adjusted again, computing a new set of output vectors. This process is repeated until an acceptable error is achieved.

In this paper, simple ANN networks were developed consisting of an input layer with three or four input nodes, one single hidden layer composed of four nodes, and one output layer consisting of one node denoting the predicted peak water demand. This architecture was found on a trial- and -error basis, as was the optimum learning coefficient (assumed to lie within the range of 0 to 0·2). The correction factor was fixed. Regression Equations (6) and (7) were used for the input variables in the ANN-1 and ANN-2 models respectively. They consisted of previous weekly peak demand, temperature and rainfall volume variables. Regression Equation (8) was used for the input variables for the ANN-3 model. The occurrence or non-occurrence of rainfall instead of the actual rainfall amount was used.

Model performance

As an indication of goodness of fit between the observed and predicted values, the coefficient of determination R^2, the average absolute relative error (AARE), and the maximum absolute relative error (Max ARE) were calculated. AARE is defined by

$$\mathrm{AARE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{O_i - D_i}{O_i}\right| \tag{18}$$

where O_i is the observed peak water demand and D_i is the predicted peak water demand found from regression, time series, and stochastic models. The smaller the value of AARE the better is the performance of the model. The maximum of the absolute relative error among all the model predictions is a measure of the robustness of the model, whereby the smaller it's value, the better the performance of the model. The coefficient of determination measures the degree of correlation among the observed and predicted values. The results of these statistics are found in Table VII.

RESULTS AND DISCUSSION

Exploratory analysis

Based on the cross-correlations (Table I), it can be seen that the peak water demand series at time t is strongly correlated with the peak demand from the previous week ($t - 1$) with a positive correlation value of 0·5883. Temperature at time t is also correlated with peak demand, with a correlation of 0·4917. For rainfall volume, the strongest link was found between peak demand and rainfall at the current time interval with a value of −0·3523, meaning that the weekly demand decreases in magnitude whenever there is increasing rainfall totals. However, the occurrence of rain with or without threshold did not appear to be correlated with the peak demand series, with the highest correlation of −0·06 found between peak demand and the occurrence of rainfall without threshold.

Table II gives the results from Fourier analysis for the peak demand series. It can be observed that there is no predominant single periodicity component in the demand series. The components explaining the most variance are at 18 weeks (corresponding to a 1 year period) and at 8 weeks (corresponding to a 2 month period), although they account for only 14% of explained variance.

Linear and multiple linear regression

The results from linear regression analysis are shown in Table III; these confirm the cross-correlation results, that the previous peak water demand $t - 1$ (with $R^2 = 0{\cdot}346$) and temperature at the current time interval t (with $R^2 = 0{\cdot}2418$) are the variables that describe the peak demand series the best.

Table III. Coefficient of determination R^2 for simple linear regression models

Peak Demand (t) vs	Slope	R^2
Peak demand ($t-1$)	0·598	0·3460
Temperature (t)	11·508	0·2418
Temperature ($t-1$)	5·622	0·0738
Rainfall (t)	−1·645	0·1241
Rainfall ($t-1$)	−1·459	0·0979

Table IV. Multiple linear regression models

Parameter	Coefficient	Model		
		MLR-1	MLR-2	MLR-3
D_{t-1}	β_1	0·488	0·555	0·632
T_t	β_2	7·157	12·801	14·159
R_t	β_3	−1·415	−1·176	
T_{t-1}	β_4		−7·730	−9·438
R_{t-1}	β_5		−0·629	
BR_t	β_6			−11·842
Intercept	β_0	0·139	34·846	−1·886

Table IV displays the results found by using multiple linear regression analysis. The model that performed the best in training was MLR-2 ($R^2 = 0{\cdot}6199$), which was a function of maximum average weekly temperature at times t and $t-1$, weekly total rainfall at times t and $t-1$ and the previous weeks peak water demand. However, the model performing the best in the testing data set was MLR-1 ($R^2 = 0{\cdot}5334$), which was a function of previous weekly peak demand, weekly average maximum temperature and total rainfall amount of the current week, and is given by the following equation:

$$D_t = 0{\cdot}488D_{t-1} + 7{\cdot}157T_t - 1{\cdot}415R_t + 0{\cdot}139 \tag{19}$$

It appears for this time scale that the peak demand series is, in fact, better described with the use of the actual rainfall amount than the occurrence or non-occurrence of rainfall.

Time series

The first stage in developing time-series models is to determine whether the peak water demand series is a stationary process. Therefore, an ACF plot of the peak demand series was performed; this is shown in Figure 2a, and illustrates a strong dependence in the series for lags up to 10. This implies that the data series is not stationary, violating the assumption required to perform this analysis. Therefore, the data were differenced using Equation (10), and the ACF plot was performed on the differenced series (Figure 2b). The new plot shows that the ACF contains two significant lags, and the remaining lags are not significantly correlated; therefore, ARIMA models were identified to predict water demand.

A total of seven ARIMA models were selected to fit the trained peak water demand series. These are shown in Table V, along with their coefficients and AIC values. It appears that the ARIMA(3,1,1) has the best fit with the observed data, which is shown by the lowest AIC value. In the prediction stage, all seven models were used to determine which model has the best agreement with the training and testing data sets, as is seen in Table V. All models had relatively low R^2 values, ranging from 0·30 to 0·34, with the ARIMA(2,1,2) and

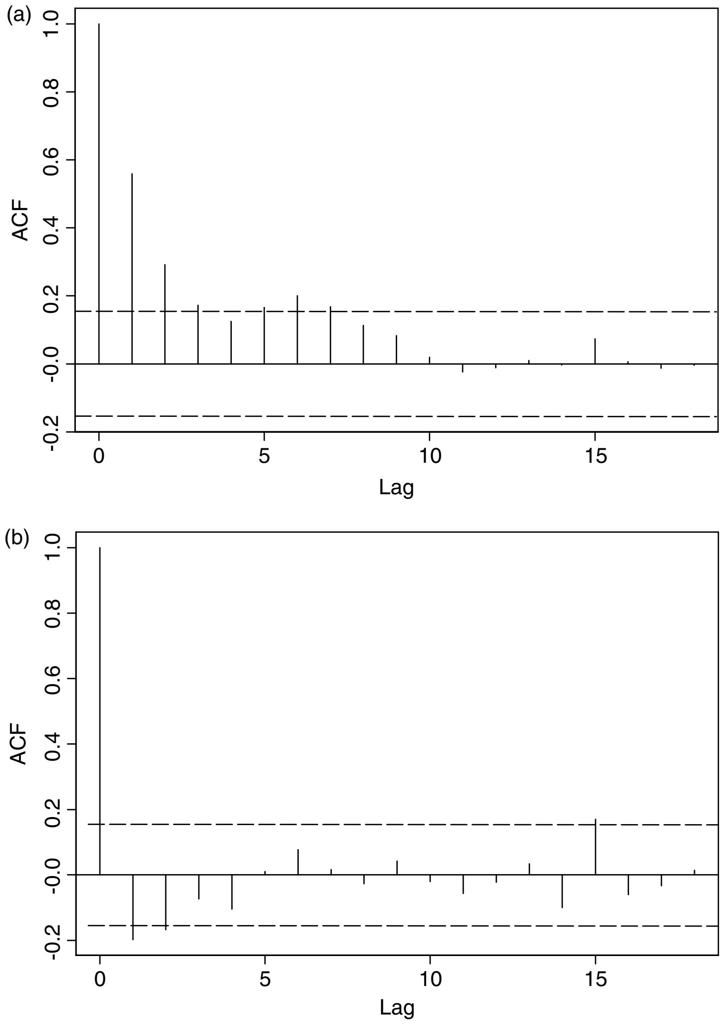

Figure 2. Autocorrelation function of the peak demand for (a) original demand series and (b) differenced demand series

ARIMA(2,1,1) having the highest coefficients of determination in the training data set. However, it was found that the ARIMA(2,1,0) model performed best ($R^2 = 0{\cdot}35$) and is shown by the following equation:

$$D_t = D_{t-1} - 0{\cdot}2409(D_{t-1} - D_{t-2}) - 0{\cdot}2168(D_{t-2} - D_{t-3}) \tag{20}$$

ANNs

The results of the ANN models are shown in Table VI. The learning coefficients were found to produce the lowest root-mean-square error between the observed and predicted water demand in the training session. The optimized learning coefficients ranged from 0·06 to 0·08 for the four models hypothesized.

The results indicate that the best results are a function of previous weekly peak demand, temperature of the current week and the total rainfall amount of the current week with an R^2 value of 0·7078 in training and a value of 0·8102 in testing. This ANN model considerably outperformed the linear, multiple linear, and time-series models. As found in the multiple linear regression models, the occurrence or non-occurrence of

Table V. ARIMA models hypothesized for forecasting

Parameter	Coefficient	ARIMA (p,d,q) Model						
		(1,1,0)	(2,1,0)	(3,1,0)	(1,1,1)	(2,1,1)	(2,1,2)	(3,1,1)
AR(1)	α_1	−0·20	−0·24	−0·28	0·49	0·52	0·10	0·52
AR(2)	α_2		−0·22	−0·27		−0·08	0·14	−0·07
AR(3)	α_3			−0·19				−0·02
MA(1)	θ_1				0·96	0·96	0·54	0·95
MA(2)	θ_2						0·40	
AIC		1873·72	1857·40	1843·46	1844·08	1834·57	1836·32	1825·89

Table VI. ANN models developed to predict peak water demand

Model	Parameter	Learning coefficient
ANN-1	D_{t-1}, T_t, R_t	0·08
ANN-2	D_t, T_t, T_{t-1},	0·06
ANN-3	D_t, T_t, T_{t-1}, BR_t	0·07

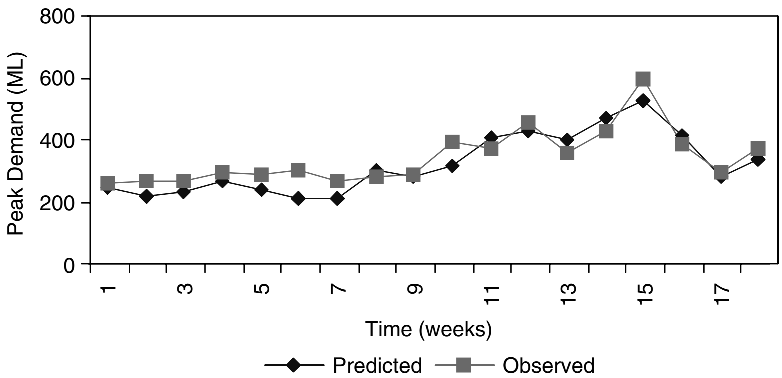

Figure 3. Observed and predicted water demand from the ANN-1 model

rainfall models did not perform as well as models containing the actual rainfall total. The ANN-3 model did well during the training session ($R^2 = 0{\cdot}6848$); however, the R^2 during testing dropped considerably to 0·53. Figure 3 shows the predicted and observed water demand for the year 2002.

Comparative analysis

The results of the comparative analysis are given in Table VII. It can be observed that the ANN-1 model performed best, with the highest R^2 value of 0·8102, lowest Max ARE and the second lowest AARE statistic. The model with the lowest AARE was the ANN-2 model. The time-series models performed better than the regression models, with the AARE statistic ranging from 13·09 to 14·61, and the regression models recorded values ranging from 17·29 to 19·25. The regression models recorded a lower range for the Max ARE statistic, with values ranging from 37·45 to 43·18, and the time-series value ranged from 35·40 to 46·43.

Therefore, it is quite clear that the ANN models are far superior than regression and time-series models, recording the lowest AARE and Max ARE statistics, while recording the highest R^2 values in training and

Table VII. Performance statistics found during testing of all statistical models

Model	AARE	Max ARE	R^2	
			Training	Testing
MLR-1	17·29	41·72	0·5394	0·5334
MLR-2	19·25	43·18	0·6199	0·4446
MLR-3	17·76	37·45	0·5413	0·3849
ARIMA(1,1,0)	13·43	35·40	0·3024	0·3271
ARIMA(2,1,0)	14·31	36·28	0·3003	0·3520
ARIMA(3,1,0)	13·86	35·93	0·3001	0·3158
ARIMA(1,1,1)	13·25	36·41	0·3326	0·2865
ARIMA(2,1,1)	14·61	46·43	0·3369	0·2706
ARIMA(2,1,2)	13·09	40·50	0·3364	0·2909
ARIMA(3,1,1)	14·00	42·69	0·3349	0·2537
ANN-1	12·26	30·05	0·7078	0·8102
ANN-2	10·12	31·72	0·7797	0·7555
ANN-3	14·19	38·11	0·6848	0·5300
Regression[a]	18·01	40·78	0·5669	0·4543
Time series[a]	13·79	39·09	0·3205	0·2995
ANN[a]	12·19	33·28	0·7241	0·6986

[a] Average values for the three classes of models.

testing. The regression models had a better correlation between the peak and observed series than the time-series models; however, they recorded the highest relative error. The time-series models had a poor correlation with peak and observed time series and robustness (Max ARE); however, the AARE values recorded for these models are comparable with the ANN models.

CONCLUSIONS

The short-term water demand forecasting models of regression analysis, time-series analysis, and the ANNs technique have been investigated in this study. The weekly water demand data, along with rainfall and maximum air temperature from Ottawa, were used to develop and test these models. A split sample methodology was employed to evaluate the performance of each technique.

Based on the results of this study, it is concluded that the ANN technique substantially outperformed regression and time-series methods in terms of accuracy of forecasting. It is also concluded that the best results are obtained when employing previous weekly demand along with the current week's rainfall and temperature. Furthermore, the amount of rainfall is more significant than the rainfall occurrence.

ACKNOWLEDGMENTS

Financial support provided by the National Science and Engineering Council of Canada and data provided by the City of Ottawa is acknowledged.

REFERENCES

Box GEP, Jenkins GM. 1976. *Time Series Analysis: Forecasting and Control*, 2nd edition. Holden-Day: San Francisco.

Crommelnyck V, Duquesne C, Mercier M, Miniussi C. 1992. Daily and hourly water consumption forecasting tools using neural networks. In *Proceedings of the AWWA's Annual Speciality Conference*, Nashville, Tennessee; 665–676.

Graeser Jr HJ. 1958. Meter records in system planning. *Journal of American Water Works Association* **50**(11): 1395–1402.

Howe CW, Linaweaver FP. 1967. The impact of price on residential water demands and its relation to system design and price structure. *Water Resources Research* **3**(1): 13–32.

Jain A, Varshney AK, Joshi UC. 2001. Short-term water demand forecast modeling at IIT Kanpur using artificial neural networks. *Water Resources Management* **15**: 299–321.

Kite GW, Adamowski K. 1973. Stochastic analysis of Lake Superior elevations for computation of relative crustal movement. *Journal of Hydrology* **18**: 163–175.

Lertpalangsunti N, Chan CW, Mason R, Tontiwachwuthikul P. 1999. A tool set for construction of hybrid intelligent forecasting systems: application for water demand prediction. *Artificial Intelligence in Engineering* **13**: 21–42.

Lingireddy S, Ormsbee LE. 1998. Neural networks in optimal calibration of water distribution systems. In *Artificial Neural Networks for Civil Engineers: Advanced Features and Applications*, Flood I, Kartam N (eds). American Society of Civil Engineers.

Lippman RP. 1987. An introduction to computing with neural nets. *IEEE ASSP Magazine*; 4–22.

Maidment DR, Miaou SP. 1986. Daily water use in nine cities. *Water Resources Research* **22**(6): 845–851.

Maidment DR, Miaou SP, Crawford MM. 1985. Transfer function models of daily urban water use. *Water Resources Research* **21**(4): 425–432.

Obeysekera J, Trimble PJ, Cadavid L, White C. 2000. Use of climate outlook for water management in south Florida, U.S.A. In *Proceedings of the Engineering Jubilee Conference*, University of Peradeniya, Sri Lanka.

Smith JA. 1988. A model of daily municipal water use for short-term forecasting. *Water Resources Research* **24**(2): 201–206.

Zhou SL, McMahon TA, Walton A, Lewis J. 2000. Forecasting daily urban water demand: a case study of Melbourne. *Journal of Hydrology* **236**(3): 153–164.

