Water Resources Management **15:** 299–321, 2001.

Short-Term Water Demand Forecast Modelling at IIT Kanpur Using Artificial Neural Networks

ASHU JAIN*, ASHISH KUMAR VARSHNEY and UMESH CHANDRA JOSHI
Department of Civil Engineering, Indian Institute of Technology, Kanpur, India
(author for correspondence, e-mail: ashujain@iitk.ac.in, fax: 91 512 597395)*

(Received: 20 June 2000; in final form: 27 August 2001)

Abstract. The efficient operation and management of an existing water supply system require short-term water demand forecasts as inputs. Conventionally, regression and time series analysis have been employed in modelling short-term water demand forecasts. The relatively new technique of artificial neural networks has been proposed as an efficient tool for modelling and forecasting in recent years. The primary objective of this study is to investigate the relatively new technique of artificial neural networks for use in forecasting short-term water demand at the Indian Institute of Technology, Kanpur. Other techniques investigated in this study include regression and time series analysis for comparison purposes. The secondary objective of this study is to investigate the validity of the following two hypotheses: 1) the short-term water demand process at the Indian Institute of Technology, Kanpur campus is a dynamic process mainly driven by the maximum air temperature and interrupted by rainfall occurrences, and 2) occurrence of rainfall is a more significant variable than the rainfall amount itself in modelling the short-term water demand forecasts. The data employed in this study consist of weekly water demand at the Indian Institute of Technology, Kanpur campus, and total weekly rainfall and weekly average maximum air temperature from the City of Kanpur, India. Six different artificial neural network models, five regression models, and two time series models have been developed and compared. The artificial neural network models consistently outperformed the regression and time series models developed in this study. An average absolute error in forecasting of 2.41% was achieved from the best artificial neural network model, which also showed the best correlation between the modelled and targeted water demands. It has been found that the water demand at the Indian Institute of Technology, Kanpur campus is better correlated with the rainfall occurrence rather than the amount of rainfall itself.

Key words: artificial neural networks, municipal water use modelling, regression analysis, short-term water demand forecasting, time series analysis, water resources management

1. Introduction

The United Nations recently reported that global freshwater consumption rose six-fold between 1990 and 1995, more than twice the rate of population growth (Beaudet and Roberts, 2000). As a result, more and more water systems have become over stressed in recent years. This has led to a need for better planning and design, and more efficient operation and management of the existing water systems. A key aspect in better planning, design, operation, and management of a water system is the accurate prediction of water demands. Water demand forecasting can be of two types: long-term forecasting and short-term forecasting. Long term forecasting is useful in planning and design, and making extension plans for an existing water system, while the short-term forecasting is useful in efficient operation and management of an existing water system. Some examples where short-term water demand forecasts can be put to use include operation and management of an urban water distribution system, water conservation programme evaluation during drought conditions, water pricing policy assessment, etc. As short-term water demand forecasts are needed for better performance of an urban water system, their accurate determination becomes of significant importance to the water managers. Leon *et al.* (2000) recently developed a hybrid expert system called EXPLORE to manage the water supply system of the City of Seville, Spain. They concluded their study by saying that the performance of an expert system for water resources management (such as EXPLORE) could be improved by using water demand forecasting techniques such as regression analysis, time series analysis, and artificial neural networks. The primary objective of the current study is therefore to investigate these techniques for use in short-term water demand forecasting.

Historically, water managers have adopted conventional modelling techniques such as regression analysis and time series analysis, or a combination of the two. A lot of work has been reported in literature on short-term water demand forecast modelling using regression and time series analysis. Graeser (1958) found that maximum daily demands in Dallas, Texas were significantly related to the number of previous days with maximum air temperature over 100 °F, and the number of weeks since the occurrence of a week with 1 inch of rainfall. Weeks and McMahon (1973) found in Australia that weekly pan evaporation, and average maximum daily temperature were more significant explanatory variables than rainfall in a multiple linear regression model describing weekly water demands. Howe and Linaweaver (1967) developed equations for maximum daily sprinkling water demand using maximum potential evapotranspiration, irrigable area, water price, and household value as independent variables. Some other examples of short-term water demand forecast modelling using regression analysis include: Oh and Yamauchi (1974), Hughes (1980), Anderson *et al.* (1980), Steiner and Smith (1983), and Maidment and Parzen (1984a). Maidment *et al.* (1985) developed a time series model of daily municipal water use as a function of rainfall and air temperature. This model, applied to the daily water demand data from Austin, Texas from 1975 to 1981, accounted for 97% of the variance of daily municipal water use. However, this model did not consider day of the week effects, which are generally significant in daily water demand data, and had a constant mean process for de-trending. Smith (1988) proposed a time series model of daily municipal water demand which accounted for these two inadequacies. The model proposed was a conditional auto-regressive process with randomly varying mean. Some other examples of short-term water demand forecast modelling using time series analysis include: Valdes and Sastri (1989), Miaou (1990), and Jowitt and Xu (1992). Maidment and Parzen (1984b) employed a combination of regression and time series analysis for the purpose of short-term water demand forecast modelling. They investigated the monthly water use and its relationship to climatic variables in six Texas cities. They proposed a systematic procedure for seasonal water use modelling, called cascade modelling. They recommended that the cascade modelling approach can be applied to weekly water demand data also. Later, Franklin and Maidment (1986) used the cascade modelling approach to describe weekly water demand data from Deerfield, Florida. Based on their results, they concluded that inclusion of auto-correlation term in the model considerably improved the forecast accuracy of the weekly data.

In recent years, the relatively new technique of Artificial Neural Networks (ANNs) has been proposed as an efficient tool for modelling and forecasting. ANNs have been reported to be employed in a wide range of engineering applications. Crommelynck *et al*. (1992) used back propagation ANNs, probably for the first time, to model daily and hourly water demand forecasts of some communities in Paris, France. They also compared the results from the ANN models with those obtained from some statistical models. In their article, they concluded that a comparison between statistical and ANN models, based on the tests conducted in their study, showed that the ANN models offered performance at least comparable with that of the statistical models. However, apart from the study conducted by Crommelynck *et al*. (1992) in Paris, France, efforts in the area of application of the ANNs in short-term water demand forecast modelling have been lacking. This article attempts to model short-term water demand forecasts using ANNs.

Many types of data are needed to model the water demand forecasts, most of which may be grouped into two classes: socio-economic variables and climatic variables. Socio-economic variables such as population, income, water price, and housing characteristics are responsible for the long-term effects on the water demands, while the climatic variables such as rainfall and maximum air temperature are responsible for the short-term seasonal variations in the water demands (Miaou, 1990).

The present study focuses on the modelling of short-term water demand forecasts using the climatic variables such as rainfall and maximum air temperature, in addition to the past water demands. The data derived from the Indian Institute of Technology (IIT), Kanpur were employed in this study. The data employed in this study consist of weekly water demand at IIT Kanpur campus, total weekly rainfall, and average maximum air temperature from the City of Kanpur, India. Three separate modelling techniques have been investigated to model the weekly water demands in this study: the conventional techniques of regression analysis and time series analysis, and the relatively new technique of ANNs. In all, six ANN models, five regression models, and two time series models were developed in this study. This article begins with a brief overview of the IIT Kanpur's water supply system and the data employed. Then a brief introduction to the relatively new technique of ANNs is presented before describing various model structures investigated in this study.

Table I. Details of tube wells at IIT Kanpur

S. No.	Tube well	Pump capacity (HP)	Yield (lps)
1	STW1 (near NCC)	15	12.63
2	STW2 (near DTW4)	41	22.73
3	STW3	33	25.20
4	STW4	15	10.10
5	STW5	15	22.73
6	DTW2	41	15.20
7	DTW3	41	14.20
Totals		201	122.79

2. IIT Kanpur's Water Supply System

The IIT Kanpur is located in the western part of the Northern State of Uttar Pradesh in India. The IIT Kanpur, a premier teaching and research institute of the nation, is a place of residence for approximately 12 000 people consisting of mainly students, faculty, staff, and their families. Though the river Ganges flows nearly 8 km north of the campus, the major source of water for the IIT Kanpur residents is ground water. The IIT Kanpur campus has its own water supply system consisting of many Shallow Tube Wells (STW) and Deep Tube Wells (DTW) maintained and operated by the Institute Works Department (IWD) on campus. The pumps operate, on an average, for about sixteen hours per day. The capacity of the pumps at each tube well and the associated yields are shown in Table I. It can be noted from the Table I that the total capacity of all the pumps and the total yield of the whole water supply system of the IIT Kanpur is 201 HP and 122.8 L per second (lps), respectively.

The present water consumption at IIT Kanpur campus is about 570 L per capita per day (lpcd). As per the Indian Standards Code of Basic Requirements of Water Supply, Drainage, and Sanitation (IS:1172-1993), the average per capita water consumption of 200 (lpcd) has been specified (MWH, 1996). Clearly, the per capita water consumption at IIT Kanpur is approximately three times of that specified by the Indian Standards. The high per capita water consumption at IIT Kanpur campus may be attributed to the high standards of living and excessive use of water for lawns and gardening. The IIT Kanpur housing was constructed in the early 1960s with most of the housing units containing lawns and kitchen gardens. The IIT Kanpur also maintains its own nursery catering to the horticultural needs of the campus residents.

The authors believe that a major fraction of the water consumption at IIT Kanpur can be attributed to watering of the lawns and gardens. The water demand process in such situations is mainly driven by the maximum air temperature with

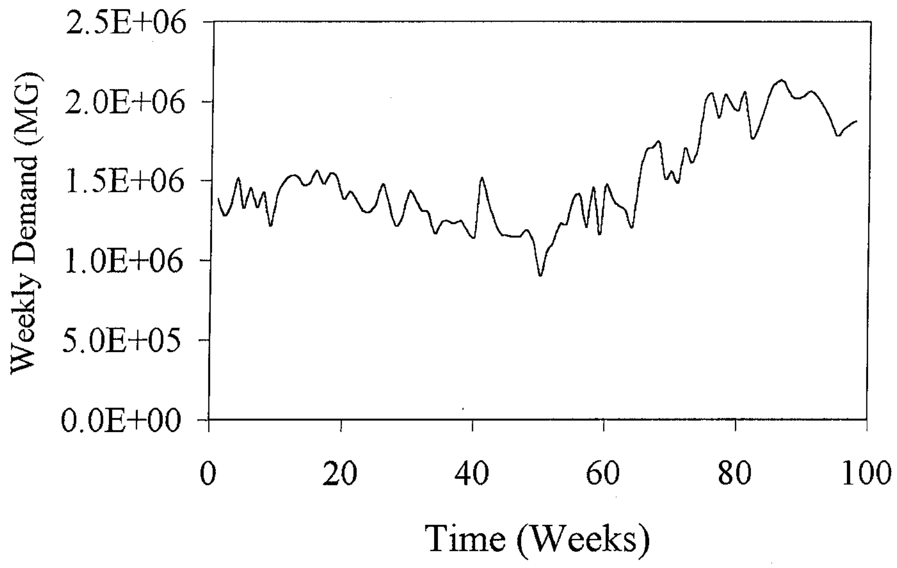

Figure 1. Weekly water demand time series.

the rainfall occurrences interrupting the process to cause transient drops in the water use (Valdes and Sastri, 1989). In other words, the water consumption can be expected to be high on dry days with high temperatures and low on the rainy days. Moreover, the water demand may not depend on the amount of rainfall, rather it may be a function of the occurrence of the rainfall. This can be attributed to the fact that the people may not want to water their lawns/gardens on a rainy day regardless of the amount of rainfall. However, this needs to be investigated. Therefore, a secondary objectives of this study is to test the following two hypotheses: (i) the short-term water demand process at IIT Kanpur is mainly driven by maximum air temperature and is interrupted by the occurrence of rainfall, and (ii) the occurrence of rainfall is a more significant explanatory variable than the amount of rainfall itself in modelling short-term water demand process. In order to test the second hypothesis, the rainfall was employed as a binary input in the regression and ANN models. This was accomplished by representing the non-occurrence of rainfall as zero and the occurrence of rainfall by one in the models, irrespective of the amount of rainfall.

3. Model Data

The data employed in this study consist of weekly water demand in million gallons (MG) at IIT Kanpur campus, and weekly total rainfall (mm) and weekly average maximum air temperature (°C) from the City of Kanpur, India. The water demand data were obtained from the Institute Works Department (IWD 1998), IIT Kanpur, and the climatic data were obtained from the India Meteorological Department

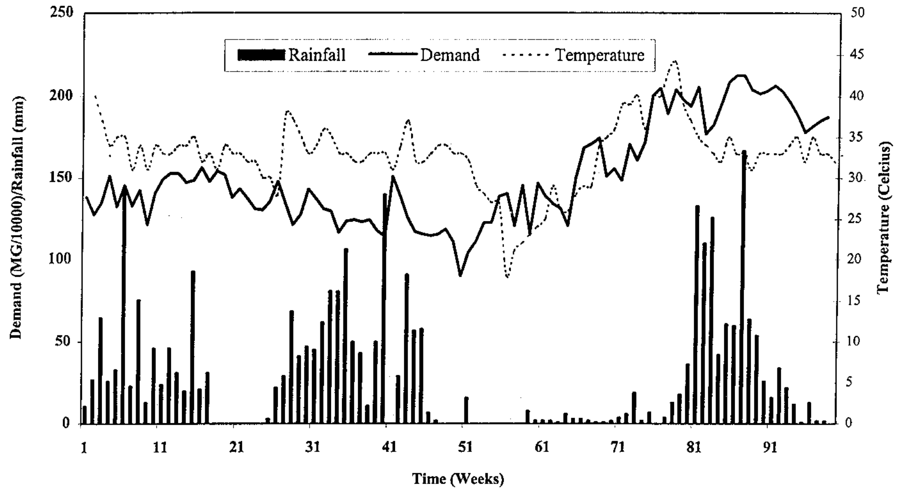

Figure 2. Weekly water demand, temperature, and rainfall series.

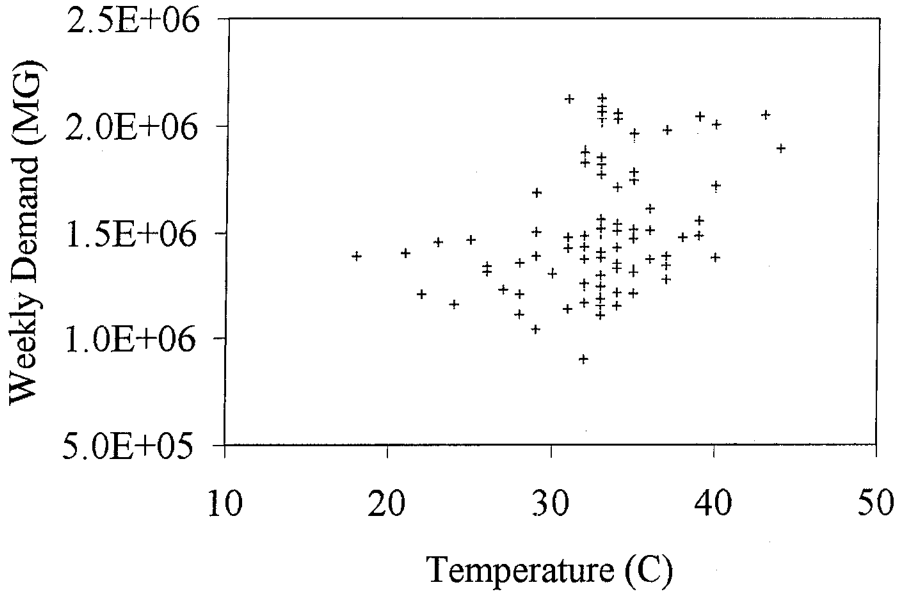

Figure 3. Water demand versus temperature.

(IMD, 1998). The water demand data were available for a period from 1989 to 1998; however, the data were scattered with a lot of missing records. As a result, the water demand data for the most recent 98 weeks were considered for model development and testing in this study. All the data were divided into two sets: a training or calibration set consisting of first 78 weeks of data, and a testing set consisting of the remaining 20 weeks of data. The training data set was used to train all the ANN models and calibrate all statistical models, while the testing data

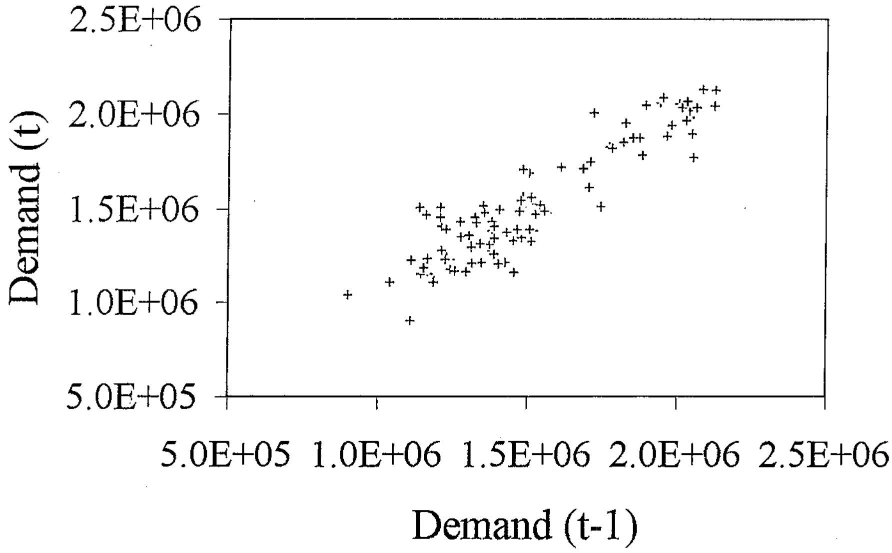

Figure 4. Water demand versus past demand.

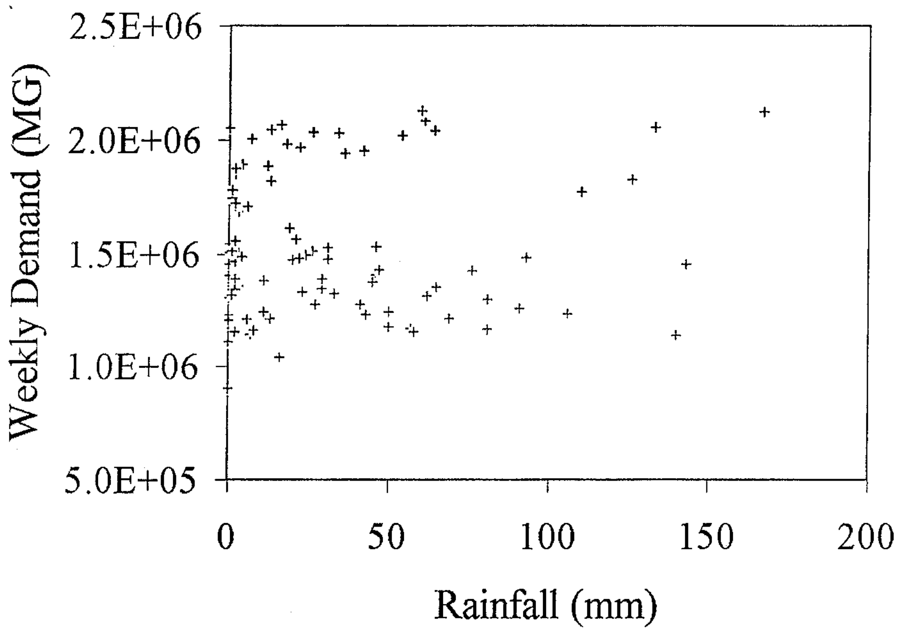

Figure 5. Water demand versus rainfall.

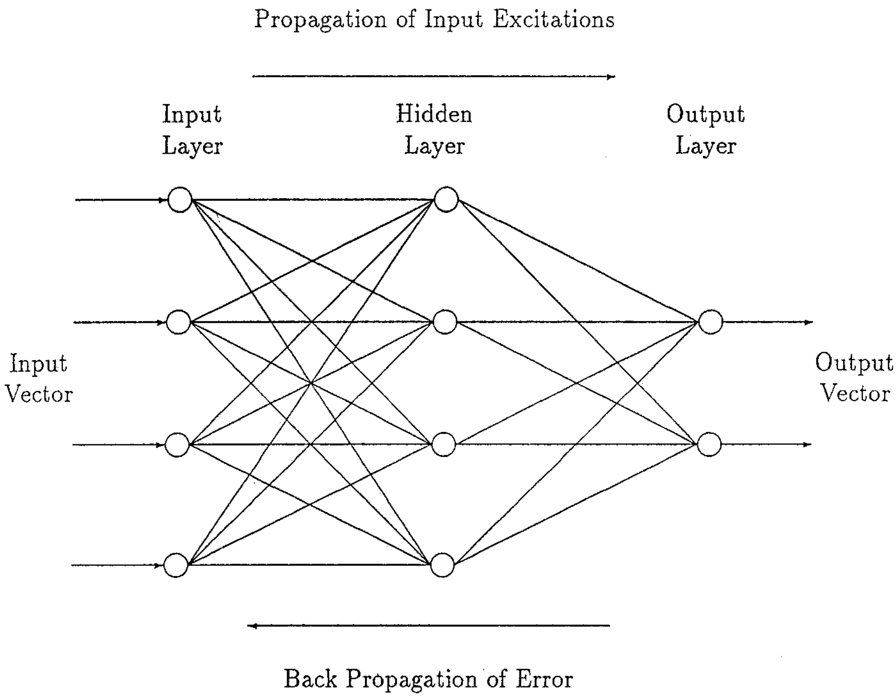

Figure 6. Structure of a back propagation ANN.

set was used to test the performance of all the models in terms of various standard statistical measures considered in this study. The weekly water demand time series at IIT Kanpur is shown in Figure 1, while Figure 2 shows the weekly water demand, temperature, and rainfall data for the same time period on the same diagram. The variation of weekly water demand with temperature, previous week's demand, and rainfall is shown in Figures 3 to 5, respectively.

4. Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are mathematical models of theorized mind and brain activity which attempt to exploit the massively parallel local processing and distributed storage properties believed to exist in the human brain (Zurada, 1992). Lately, the ANN technique, also called parallel distributed processing, has received a great deal of attention as a tool of computation by many researchers and scientists. One of the most important characteristics of an ANN is its ability to learn from the facts or input data and the associated output data. While developing an ANN model, one presents an ANN with input patterns and their associated outputs. The network then internally organizes itself to reconstruct the complex relationships among inputs and outputs.

An ANN is a highly interconnected network of many simple processing units called ‘neurons’ or ‘neurodes’. Neurons having similar characteristics are grouped in one single layer. For example, the neurons in an input layer receive input from an external source, compute the output, and transmit the computed output to a neuron in an adjacent layer, which could either be a hidden layer or an output layer. Each neuron in an ANN is also capable of comparing an input to a threshold value. The input vector presented to an ANN should be normalized between 0 and 1. The ANN stores the information captured from the input vector as the ‘strengths of the connections’ between the neurons. The most commonly used ANN in engineering applications is a back propagation ANN (see Figure 6). In this figure, each neuron is represented by a circle and each connection by a line. The ANN shown in Figure 6 consists of three layers: an input layer consisting of four neurons, a hidden layer also consisting of four neurons, and an output layer consisting of two neurons. In a back propagation ANN, the input presented to the neurons in an input layer are propagated in a forward direction and the output vector is calculated through the use of a non-linear function called activation function. The activation function should be continuous, differentiable, and bounded from above and below. Then, knowing the output, the error at the output layer from the ANN model can be computed. The computed error is then back propagated through the network and the ‘connection strengths’ are updated using some training mechanism such as ‘generalised delta rule’ (Rumelhart *et al.*, 1986). This process of feed-forward calculations and error back propagation is repeated until an acceptable level of convergence is reached. This whole process is known as training of the ANN.

Before starting the training of an ANN, the weight matrix (say w_{ij}) is initially randomized between suitable limits. The weights can be updated using the generalized rule according to the following equation:

$$w_{ij}(n+1) = w_{ij}(n) + \Delta w_{ij}(n)$$

$$\Delta w_{ij}(n) = \eta \delta_j O_i + \alpha \Delta w_{ij}(n-1)$$

where w_{ij} $(n+1)$ is the value of a weight at iteration number $(n+1)$, Δw_{ij} (n) is the change in the value of a weight at iteration number n, η is the learning coefficient, α is the momentum correction factor, δ_j is the error signal, and O_i is the output from the output neuron i, n is an index representing iteration, and i and j are indices representing a neuron in a particular layer. The values of the learning coefficient and the momentum correction factor normally range between 0 and 1. The error signal (δ_j) is computed in two different manners depending upon whether the index j represents an output layer or a hidden layer. If j represents an output layer then the error signal (δ_j) can be computed using the following equation:

$$\delta_j = O_j(1 - O_j)(t_j - O_j)$$

and if j represents a hidden layer then the error signal (δ_j) can be computed using the following equation:

$$\delta_j = O_j(1 - O_j)\sum_{k=1}^{n_2} \delta_k w_{jk}$$

where n_2 is the number of neurons in a layer next to the one represented by j, k is an index representing the neuron in a layer next to the one represented by j, t_j is the target output from output neuron j, and the expression in the summation is computed for a layer next to the one represented by j. Thus these equations are used at each iteration to update each of the weight in the weight matrix. This weight updating is continued until an acceptable level of convergence in terms of total error at the output is achieved. At this point of time the network is said to have trained. Once the network has been trained, it can be used for validation and prediction.

5. Model Development

Three types of modelling techniques: ANN technique, time-series analysis, and regression analysis were investigated to model the weekly water demand forecasts in this study. The first step in modelling any physical system is to identify the physical variables affecting the process. In modelling the weekly water demand forecasts, the physical variables considered were weekly average maximum air temperature (°C) and total weekly rainfall (mm), in addition to the water demand (MG) observed in the past. The water demand at IIT Kanpur seems to be related with temperature and the past water demands but the same could not be said for the rainfall. However, it was decided to explore the rainfall as one of the explanatory variables to model the weekly water demand forecasts at IIT Kanpur. The rainfall was employed as a binary input also in the regression and ANN models. This was accomplished by representing the non-occurrence of rainfall as zero and representing occurrence of rainfall as one in the models, irrespective of the rainfall amount.

5.1. ANN MODEL DEVELOPMENT

A back propagation ANN was employed with the ‘generalized delta rule’ as the training algorithm in this study for the development of all ANN models for modelling the weekly water demand forecasts. Two different types of ANN models were developed in this study: 1) simple ANN models consisting of only one hidden layer, and 2) complex ANN models consisting of two hidden layers. While developing an ANN model, the primary objective is to arrive at the optimum architecture of the ANN to capture the relationship among the various input and output variables. The task of identifying the number of neurons in the input and output layers is normally

Table II. Optimum architectures of simple ANN models

Model	Parameters	Optimum learning coefficient	Optimum architecture
SANN-1	D_{t-1} and T_t	0.06	2-3-1
SANN-2	D_{t-1}, T_t, and R_t	0.02	3-4-1
SANN-3	D_{t-1}, T_t, and R_{t-1}	0.015	3-4-1
SANN-4	D_{t-1}, T_t, and BR_t	0.18	3-4-1
SANN-5	D_{t-1}, T_t, and T_{t-1}	0.10	3-4-1

simple as it is dictated by the input and output variables considered to model the physical process. However, the number of neurons in the hidden layer(s) have to be optimized using the available data through the use a trial and error procedure. Another important task is to find the optimum value of some ANN parameters such as learning coefficient and momentum correction factor that will result in an optimum architecture of the ANN. In this study, the momentum correction factor was fixed and the learning coefficient was optimized for each model using the average error from the model as the objective function. The optimum number of neurons in the hidden layer(s) were determined using a trial and error procedure in this study. Six separate ANN models were developed in this study including both simple and complex type. These different ANN models resulted from different combinations of input variables considered in each model. The development of various ANN models developed in this study is described next.

5.1.1. *Simple ANN Model Development*

The simple ANN models developed in this study consisted of three layers: an input layer consisting of either two or three neurons, a hidden layer, and an output layer consisting of one neuron. The neurons in the input layer represented different combinations of various physical variables considered. The number of neurons in the hidden layer is, in fact, responsible for capturing (or mapping) the relationship among various input and output variables considered in developing an ANN. The number of neurons in the hidden layer were determined using a trial and error procedure and the learning coefficient was optimized for each model. Five separate simple ANN models were developed depending upon different combination of input variables considered in each model. The configuaration of each of the simple ANN model in terms of the physical variables considered is presented in Table II. It must be pointed out that the fourth simple ANN model was developed in an attempt to investigate the relationship of *occurrence* of rainfall rather than the *amount* of rainfall with the water demand in the current week and test the validity of the second hypothesis being investigated in this study mentioned earlier. This third neuron in the input layer could take on only two values: either 1 or 0, depending on whether it rained or did not rain in the current week, respectively. Hence, the three neurons in the input layer of the fourth ANN model represented water demand in the previous week (D_{t-1}), average maximum air temperature in the current week (T_t), and the *occurrence* of rainfall in the current week as a binary input (BR_t). The only neuron in the output layer of all the five simple ANN models represented the weekly water demand to be modelled (D_t). The number of neurons in the hidden layer were determined using a trial and error procedure. The final architecture and the values of optimum learning coefficients corresponding to each of the simple ANN model structure developed in this study are presented in Table II.

5.1.2. *Complex ANN Model Development*

An ANN model with three layers is able to capture the relationship among the input and output variables reasonably well. However, when the relationship among the variables involved is complex, then a simple ANN model may not be able to capture the relationship very accurately. It may be possible to model the physical process in a better way by increasing the number of hidden layers. However, caution must be exercised at this step because increasing the number of layers in an ANN model will also increase the number of connections or the total number of parameters of the model. And based on the principle of parsimony, a simple model giving a reasonable performance should be preferred over a more complex model giving a similar performance. However, the principle of parsimony can be followed by keeping the number of hidden neurons equal, or close if greater, in both the models. Specifically, it can be ensured that the sum of neurons in the hidden layers of a complex ANN is equal to or close to the number of neurons in the single hidden layer of a simple ANN. This approach was used to explore the possibility of developing a complex ANN to model the weekly water demand forecasts in this study.

The complex ANN models developed in this study consisted of four layers: an input layer consisting of three neurons, two hidden layers, and an output layer consisting of one neuron. The neurons in the input layer of the complex models represented the same variables as those represented by the input neurons of the best simple ANN model i.e. the SANN-4 model (see the Section Results and Discussions later on). This was because of the fact that the set of combination of various variables employed in SANN-4 model had already proved to be the best for modelling the weekly water demand forecasts for the data considered in this study. Hence, it was decided to take this combination of the input variables and investigate complex ANN models in an attempt to improve the performance of the best simple ANN model. Various complex ANN models were investigated consisting of three neurons in the input layer and one neuron in the output layer. The neurons in the input layer represented water demand in the previous week (D_{t-1}), average maximum air temperature in the current week (T_t), and the *occurrence* of rainfall in the current week as a binary input (BR_t). The sum of number of neurons in the two hidden layers was kept to a maximum of eight, keeping the principle of parsimony in mind. The number of neurons in each of the hidden layers were varied from three to five resulting in various complex ANN models. All of the complex ANN models were first trained using the data in the training set to obtain the optimized set of connection strengths. Then all the models were tested using the testing data set and various standard statistical measures were calculated. Then the best complex ANN model was selected based upon the performance of all the models in terms of various standard statistical measures considered in this study.

5.2. REGRESSION MODELS

The regression analysis technique was also investigated for developing short-term water demand forecasts in this study. Two types of regression models were developed in this study: linear multiple regression models and non-linear multiple regression models. These are described in the following paragraphs.

5.2.1. *Linear Multiple Regression Models*

In developing a linear multiple regression model for weekly water demand forecasts, the water demand in a week can be regressed against the water demand in the past week, and total rainfall and average maximum air temperature in the same and/or past weeks. In the present study three separate linear multiple regression models were developed for weekly water demand forecasts. The first linear multiple regression model, called LMRM-1 model, considered weekly water demand in the previous week (D_{t-1}), average maximum air temperature in the same week (T_t), and the total amount of rainfall in the same week (R_t), as the explanatory variables to describe the weekly water demand (D_t). The second linear multiple regression model, called LMRM-2 model, used average maximum air temperature in the previous week (T_{t-1}) and total amount of rainfall in the previous week (R_{t-1}) also, in addition to using the explanatory variables employed by LMRM-1 model, to describe the weekly water demand (D_t). The third linear multiple regression model, called LMRM-3, was similar to LMRM-1 model except that the rainfall in the current week in the LMRM-3 model was expressed as a binary variable. That is, the rainfall in the current week could take on only two values i.e. one for occurrence of rainfall, and zero for non-occurrence of rainfall. This model was developed to test the validity of the second hypothesis being investigated in this study mentioned earlier. The structures of the three linear multiple regression models can be described by the following three equations, respectively:

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

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

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

where β's are the regression coefficients to be determined, BR_t represents rainfall in week t expressed as binary input, and other notations are as explained above. The regression coefficients were determined using the same training data set used to train all ANN models. The calculated values of regression coefficients of all the three linear multiple regression models are presented in Table VII.

5.2.2. *Non-Linear Multiple Regression Models*

The water demand process is a dynamic process with a complex relationship among the various input and output variables involved. It is only appropriate to investigate for a non-linear relationship among the dependent and explanatory variables using a non-linear regression model. Two different non-linear multiple regression models were developed in this study. The explanatory variables involved in the two non-linear multiple regression models were same as those employed in the first two linear multiple regression models. The structure of the two non-linear multiple regression models can be expressed by the following two equations, respectively.

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

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

where β's are the regression coefficients to be determined and other notations are as explained earlier. The regression coefficients were determined using the same training data set used to train all ANN models and calibrate linear multiple regression models. The calculated values of regression coefficients of both of the non-linear multiple regression models are presented in Table VII.

5.3. TIME SERIES MODELS

The regression models are useful in developing relationships among various input and output variables involved in a physical process reasonably well. However, they do not illuminate the inherent auto-correlation structure of a water use pattern over time (Maidment and Parzen, 1984b). A time series analysis is required to reveal the auto-correlation structure of a short-term water demand pattern over time such as the one under investigation in the current study. As a result, the time series modelling technique was also investigated for use in weekly water demand forecasting in this study. The structures investigated in developing time series models for forecasting weekly water demands included auto-regressive models only. The structure of an auto-regressive model can be described by the following equation:

$$D_t = \sum_{i=1}^{p} \alpha_i D_{t-1} + \varepsilon_t$$

where α's are the auto-regressive parameters to be determined, ε_t is a random process, p is the order of the auto-regressive process, and i is an index representing the order of the auto-regressive process. Auto-regressive models of orders up to two have been found to be adequate for modelling short-term water demand forecasts. In this study, two separate auto-regressive time series models were developed to model the weekly water demand forecasts. The first model, called AR(1) model, had only one auto-regressive parameter. The second auto-regressive model, called AR(2) model, had two auto-regressive parameters. The auto-regressive parameters of both the models were determined using the same training data set used to train all ANN models and calculate regression coefficients of all regression models. The final model structures of the two AR models can be described by the following two equations, respectively.

$$D_t = 0.8008 D_{t-1} + \varepsilon_t$$

$$D_t = 0.6809 D_{t-1} + 0.1388 D_{t-2} + \varepsilon_t$$

6. Model Performance

Once all of the model structures were calibrated using the calibration/training data set, their performance was evaluated in terms of various standard statistical parameters. The performance of a mathematical model can be quantified using many standard statistical parameters describing the errors associated with the model forecasts. Four different statistical parameters were used to test the performance of each model investigated in this study. The first statistical parameter, called the average absolute relative error (AARE), is a quantitative measure of the average error in one step ahead forecasts from a particular model. It can be computed using the following equations:

$$RE_t = \frac{DO_t - DF_t}{DO_t} \times 100\%$$

$$\text{AARE} = \frac{1}{N} \sum_{t=1}^{N} |RE_t|$$

where RE_t is the relative error in forecasting water demand at time t expressed as a percentage, DF_t is the forecasted water demand at time t, DO_t is the observed water demand at time t, and N is the total number of water demands forecasted. Clearly, smaller is the value of the AARE, better is the performance of the model.

The second statistical parameter, called 'threshold statistic for a level of $x\%$' is a measure of the consistency in forecasting errors from a particular model. The threshold statistic is represented as TS_x and is expressed as a percentage. The threshold statistic can be expressed for different levels of absolute relative error from the model. The threshold statistic for any level of absolute relative error (say $x\%$) can be computed using the following equation:

$$TS_x = \frac{Y_x}{N} \times 100\%$$

where TS_x is the threshold statistic for a level of $x\%$ from the model, and Y_x is the number of data points (out of N) for which the absolute relative error is less than $x\%$ from the model. Clearly, larger is the value of the threshold statistic, better is the performance of the model. The threshold statistics were calculated for levels of absolute relative errors of 1, 3, 5, 10, and 15% in the present study.

The third statistical parameter calculated in this study, called Max ARE, is the maximum of the absolute relative errors among all of the forecasted data points. The Max ARE from a model is a measure of the robustness of the model. Obviously, smaller is the value of the Max ARE, better is the performance of the model. The fourth statistical parameter, called coefficient of correlation (R^2), is a measure of the strength of the model in developing a relationship among various input and output variables. This can be calculated using the standard equations available for its determination. The coefficients of correlation were calculated for both training and testing data sets from various models developed in this study.

7. Results and Discussions

The data analysis results are shown graphically in Figures 1 through 5. Figure 1 shows the weekly water demand series. It can be observed that the weekly water demand at IIT Kanpur is on the rise in recent times. Figure 2 shows the variation of weekly water demand, weekly average of the daily maximum air temperature, and weekly total rainfalls as a function of time. Figure 3 shows the variation of weekly demand as a function of temperature. The correlation between weekly water demand and temperature, as indicated in Figure 3, was confirmed by various model structures developed in this study. Figure 4 shows the variation of weekly water demand as a function of its value in the previous week. The strong correlation between weekly water demand at two successive weeks, as suggested by Figure 4, was confirmed by various conventional and ANN models developed in this study. Figure 5 shows the variation of weekly water demand as a function of the magnitude of the weekly total rainfall. There does not seem to be any correlation between weekly water demand and the magnitude of the weekly total rainfall. This fact was confirmed by various model structures developed in this study.

The calibration results in terms of regression coefficients from various linear and non-linear multiple regression models are presented in Table VII. It can be noted from Table VII that the coefficients corresponding to temperature are positive and large and those corresponding to rainfall are negative and small. This indicates an increase in weekly water demand with an increase in temperature values. On the other hand, a decrease in weekly water demand is indicated with increase in

Table III. AARE and threshold statistics during testing from ANN models

Model	AARE	TS1	TS3	TS5	TS10	TS15	Max ARE
SANN-1	5.45	5.26	20.00	50.00	95.00	100.00	10.49
SANN-2	6.10	0.00	15.00	40.00	85.00	100.00	11.12
SANN-3	6.72	0.00	10.00	40.00	75.00	100.00	12.62
SANN-4	3.74	21.11	52.63	63.16	95.00	100.00	12.08
SANN-5	6.02	0.00	10.00	45.00	90.00	100.00	11.09
CANN	2.41	31.58	68.42	84.21	100.00	100.00	6.49

Table IV. AARE and threshold statistics during testing from conventional models

Model	AARE	TS1	TS3	TS5	TS10	TS15	Max ARE
LMRM-1	7.38	5.26	15.79	42.10	68.42	89.47	17.18
LMRM-2	7.57	5.26	15.79	36.84	68.42	89.47	17.17
LMRM-3	5.16	5.26	26.32	42.10	89.47	100.00	10.39
NLMRM-1	6.90	5.26	10.52	21.05	84.21	100.00	12.45
NLMRM-2	5.59	5.26	15.79	42.10	84.21	100.00	10.52
AR(1)	19.64	0.00	0.00	0.00	5.26	5.26	25.03
AR(2)	17.68	0.00	0.00	0.00	5.26	21.05	24.07

rainfall or occurrence of rainfall. This means that the weekly water demand process at IIT Kanpur is a dynamic process driven by the temperature and interrupted by the occurrence of rainfall.

Table V. Coefficient of correlation from all models

Model	Training	Testing
SANN-4	0.9634	0.6397
CANN	0.9921	0.8724
LMRM-1	0.7156	0.4059
LMRM-2	0.7156	0.4106
LMRM-3	0.6842	0.6415
NLMRM-1	0.6478	0.6261
NLMRM-2	0.6672	0.6017
AR(1)	0.8261	0.6470
AR(2)	0.8261	0.6236

The results in terms of all the standard statistical parameters discussed above are presented in Tables III to V in this section. Specifically, the AARE, threshold statistics for the levels of 1, 3, 5, 10, and 15%, and Max ARE from all ANN models are presented in Table III, and those from all conventional models are presented in Table IV. All of these statistics shown in Tables III and IV are for the testing data set. The coefficients of correlation (R^2) statistic for both training and testing data sets from the best ANN models and all conventional models are presented in Table V.

It can be observed from Table III that the SANN-4 model performed the best among the simple ANN models in terms of AARE and threshold statistics, while SANN-1 model performed the best among simple ANN models in terms of Max ARE. However, the complex ANN model (CANN) performed the best among all models developed in this study in terms of all the standard statistical measures considered. The CANN model produced the least AARE of 2.41% and the least Max ARE of 6.49%. It can be observed from Table III that nearly 32% of the forecasted water demands from CANN model and nearly 21% from the SANN-4 model had absolute relative errors less than 1% (see TS1 in Table III); whereas, the same statistic from other simple ANN models was either 0.00% or 5.26% only. Further, all of the forecasted water demands from CANN model had absolute relative errors less than 10%, while 95% of the forecasted water demand from the SANN-4 model had absolute relative errors less than 10% (see TS10 in Table III). It must be pointed here that the best simple ANN model (SANN-4) and the best ANN model (CANN) employed rainfall as a binary input as opposed to the amount of rainfall in other models. This suggests that the water demand process at IIT Kanpur is better correlated with the occurrence of rainfall and not with the amount of rainfall. This finding helps in validating the second hypothesis being investigated in this study.

Moving on to the conventional models, it is clear from the Table IV that all regression models performed better than all time series models developed in this study. Among the regression models, non-linear multiple regression models performed marginally better than the linear multiple regression models, with the exception of LMRM-3 model. The LMRM-3 model performed the best among all conventional models developed in this study in terms of AARE, threshold statistics for all levels, and the Max ARE. The LMRM-3 model produced the least AARE of 5.16% and least Max ARE of 10.39% considering all conventional models together. Further, nearly 26% of the forecasted water demands had absolute relative errors less than 3%, and nearly 90% of the forecasted water demands had absolute relative errors less than 10%, from the LMRM-3 model (see TS10 in Table IV). The LMRM-3 model performed the best in terms of other threshold statistics also. It must be pointed out here again that the best model among all conventional models (LMRM-3) also incorporated rainfall as binary input as opposed to the amount of rainfall in other models. This further strengthens the finding that the water demand process at IIT Kanpur is better correlated with the occurrence of rainfall rather than its amount.

Table VI. Results of the comparative analysis

Statistic	ANN	Regression	Time series
AARE	5.07	6.52	18.66
TS1	9.62	5.26	0.00
TS3	29.34	16.84	0.00
TS5	56.37	36.84	0.00
TS10	90.00	78.95	5.26
TS15	100.00	95.79	13.16
Max ARE	10.65	13.54	24.55
R (Training)	0.9778	0.6861	0.8265
R (Testing)	0.7561	0.5372	0.6353

Table VII. Regression coefficients of various linear and non-linear regression models

Model	β_0	β_1	β_2	β_3	β_4	β_5
LMRM-1	140510.0	0.7916	5837.1	–1292.6	–	–
LMRM-2	142552.3	0.7908	6042.0	–1282.0	–220.9	–24.8
LMRM-3	160330.7	0.8357	2326.4	749.5	–	–
NLMRM-1	2.4869	0.8126	0.0489	–0.0025	–	–
NLMRM-2	2.1479	0.8471	–0.0487	–0.0083	0.0536	0.0092

It can be observed from Table V that the complex ANN model (CANN) performed the best both during training and testing in terms of the coefficient of correlation (R^2). The best coefficient of correlation values, both during training (0.9921) and during testing (0.8724), were obtained from the CANN model. Hence, it can be concluded that the CANN model was able to capture the relationship among various input and output variables the most efficiently among all the models developed in this study. Among the conventional models, even though the time series models had shown better correlation coefficients during training, they were not able to perform well during testing. This may be because of the fact that the time series models were the only models in this study which did not consider climatic variables during the modelling process. This demonstrates the validity of the first hypothesis investigated in this study that the water demand process at IIT Kanpur is mainly a function of maximum air temperature and is interrupted by the occurrences of rainfall.

Finally, a comparative analysis was performed to evaluate the suitability of each modelling technique investigated in this study to model the short-term water demand forecasts. In this comparative analysis, average values of the standard statistical measures were calculated from all the models employing a particular

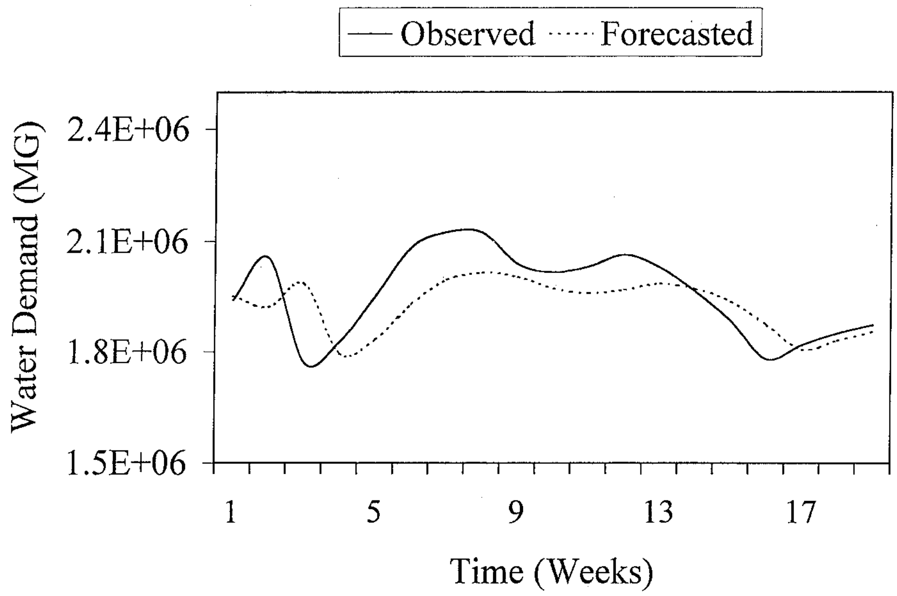

Figure 7. Observed and forecasted water demands from SANN-4 model.

technique. The results of this comparative analysis for all the three techniques investigated in this study; regression analysis, time series analysis, and ANN technique in terms of all statistical measures are presented in Table VI.

It can be observed from Table VI, that the models developed using ANN technique have consistently outperformed the models using regression and time series analysis in this study. The least AARE of 5.07% and the least Max ARE of 10.65 % were obtained from the models employing ANN technique. Further, nearly 90% of the water demands forecasted from the models employing ANN technique had absolute relative errors less than 10% as compared to nearly 79% from the regression models and only 5.26% from the time series models (see TS10 in Table VI). Also, the maximum value of coefficient of correlation during training and testing of 0.9778 and 0.7561, respectively, were obtained from the models using ANN technique. Hence, based on the results obtained in this study, it can be said that the ANN technique is certainly more suitable and superior to the conventional techniques of regression and time series analysis for modelling short-term water demand forecasts. The results in terms of observed and forecasted water demands are depicted in graphical form in Figures 7 and 8 from the SANN-4 model and CANN model, respectively. These figures are prepared for testing data set only. It can be noticed from these figures that the water demands forecasted from the two ANN models match with the observed water demands exceedingly well.

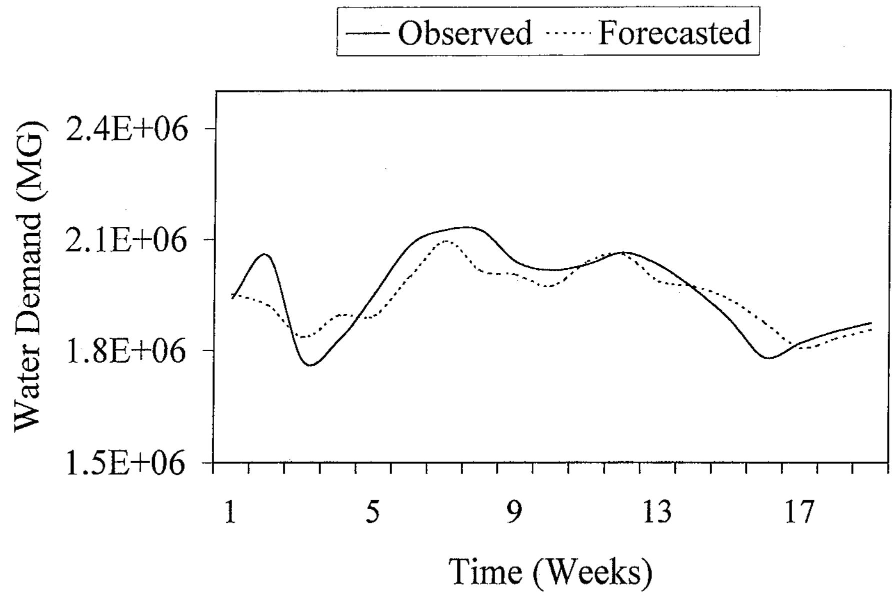

Figure 8. Observed and forecasted water demands from CANN model.

8. Conclusions

Three modelling techniques: regression analysis, time series analysis, and artificial neural networks (ANNs) technique, have been investigated in this study to model the short-term water demand forecasts. Specifically, six ANN models, five regression models, and two time series models were developed in this study. The weekly water demand data derived from the Indian Institute of Technology (IIT), Kanpur, and the climatic data in terms of rainfall and maximum air temperature from the City of Kanpur, India were employed to develop and test all the models investigated in this study. Certain standard statistical parameters were used to evaluate the performance of all the models developed in this study.

A comparative analysis was carried out to evaluate the suitability of each technique investigated to model the short-term water demand forecasts. Based upon the results obtained in this study and the consequent comparative analysis, it has been found that the models employing the ANN technique have consistently outperformed the models using conventional techniques of regression and time series analysis. This clearly establishes the suitability and superiority of the relatively new technique of ANNs over the conventional techniques of regression and time series analysis for use in short-term water demand forecast modelling. Further, the models which employed rainfall occurrence as opposed to the rainfall amount as one of the explanatory variables performed the best among the respective category. For example, the linear multiple regression model, LMRM-3, performed the best among all conventional models, SANN-4 model performed the best among all simple ANN models, and the CANN model performed the best among all complex

ANN models developed in this study. In fact, the complex ANN model (CANN) performed the best among all the models developed in this study. Also, the time series models, which did not consider the dynamic effects of climatic variables such as rainfall and maximum air temperature performed the worst among all the models developed in this study. Hence, based upon the results obtained in this study, it can concluded that 1) the short-term water demand process at IIT Kanpur is a dynamic process mainly driven by the maximum air temperature and interrupted by the occurrences of rainfall, and 2) the occurrence of rainfall is a more significant explanatory variable than the amount of rainfall in modelling short-term water demand process. As a result, the two hypotheses under investigation were found to be valid in case of the data considered in this study.

The authors believe that no research effort is ever complete and there is always room for some improvement. In light of the present research effort, the following limitation/improvements are cited. A back propagation ANN with generalized delta rule as the training algorithm was employed for training all the ANN models developed in this study. It may be possible to further improve the performance of the ANN models by employing some other training algorithms or ANNs such as radial basis functions, genetic algorithms, or self-organizing networks, etc. Further, the data considered in this study were scattered with a lot of missing records. A large record of continuous data set may provide a large spectrum for the models to obtain information from in order to develop better relationship among various input and output variables. The time series models developed in this study did not consider the dynamic effects of climatic variables such as rainfall and maximum air temperature on the short-term water demands. The performance of the time series models may be improved by regressing the residuals against the climatic variables. Finally, it is realised that all the models developed in this study need forecasted information on the climatic variables. In the presented work, perfect information has been used while computing various performance statistics. How well the models developed in this study will perform when presented with the forecasted information on the climatic variables, remains to be seen. It is hoped that future research efforts will concentrate in overcoming some of the limitations cited here. So that the performance of the short-term forecast models may be improved for the betterment of planning, design, operation, and management of the water resources systems.

References

Anderson, R. C., Miller, T. A. and Washburn, M. C.: 1980, 'Water Savings from Lawn Watering Restrictions During a Drought Year in Fort Collins, Colorado', *Water Resour. Bull.*, **16**(4), 642–645.

Beaudet, B. and Roberts, R. L.: 2000, 'A Perspective on the Global Water Marketplace', *J. Amer. Water Works Assoc.* **92**(4), 10–12.

Crommelynck, V., Duquesne, C., Mercier, M. and Miniussi, C.: 1992, 'Daily and Hourly Water Consumption Forecasting Tools Using Neural Networks', *Proc. of the AWWA's Annual Computer Specialty Conference*, Nashville, Tennessee, pp. 665–676.

Franklin, S. L. and Maidment, D. R.: 1986, 'An Evaluation of Weekly and Monthly Time Series Forecasts of Municipal Water Use', *Water Resour. Bull.* **22**(4), 611–621.

Graeser Jr., H. J.: 1958, 'Meter Records in Systems Planning', *J. Amer. Water Works Assoc.* **50**(11), 1395–1402.

Howe, C. W. and Linaweaver, F. P.: 1967, 'The Impact of Price on Residential Water Demand and Its Relation to Systems Design', *Water Resour. Res.* **3**(1), 13–22.

Hughes, T. C.: 1980, 'Peak Period Design Standards for Small Western U.S. Water Supply', *Water Resour. Bull.* **16**(4), 661–667.

IMD: 1998, *Weekly Weather Reports*, India Meteorological Department, India, 1989–1998.

IWD: 1998, *Daily Water Supply Record*, Institute Work Department, IIT Kanpur, India, 1989–1998.

Jowitt, P. W. and Xu, C.: 1992, 'Demand Forecasting for Water Distribution Systems', *Civil Engg. Sys.* **9**, 105–121.

Leon, C., Martin, S., Jose, M. E. and Luque, J.: 2000, 'EXPLORE-Hybrid Expert System for Water Networks Management', *J. Water Resour. Plng. and Mgmt.* **126**(2), 65–74.

MWH: 1996, *Manual on Water Supply and Treatment*, Central Public Health and Environmental Engineering Organization, Ministry of Works and Housing, New Delhi, India.

Maidment, D. R. and Parzen, E.: 1984a, 'Monthly Water Use and its Relationship to Climatic Variables in Texas', *Water Resour. Bull.* **19**(8), 409–418.

Maidment, D. R. and Parzen, E.: 1984b, 'Time Patterns of Water Use in Six Texas Cities', *J. Water Resour. Plng and Mgmt.* **110**(1), 90–106.

Maidment, D. R., Miaou, S. P. and Crawford, M. M.: 1985, 'Transfer Function Models of Daily Urban Water Use', *Water Resour. Res.* **21**(4), 425–432.

Miaou, S. P.: 1990, 'A Class of Time Series Urban Water Demand Models with Non-Linear Climatic Effects', *Water Resour. Res.* **26**(2), 169–178.

Oh, H. S. and Yamauchi, H.: 1974, 'An Economic Analysis of the Patterns and Trends in Water Consumption within the Service Area of the Honolulu Board of Water Supply', *Report No. 84*, Hawaii Water Resources Research Center, University of Honolulu, Honolulu, pp. 647–666.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: 1986, 'Learning Internal Representations by Error Back Propagation', in D. E. Rumelhart and J. L. McClelland (eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, Foundations, The MIT Press, Ch 8.

Smith, J. A.: 1988, 'A Model of Daily Municipal Water Use for Short-Term Forecasting', *Water Resour. Res.* **24**(2), 201–206.

Steiner, R. C. and Smith J. A.: 1983, *Short-Term Municipal Water Use Forecasting*, Paper Presented at the ASCE's National Specialty Conference, Tampa, Florida.

Valdes, J. B. and Sastri, T.: 1989, 'Rainfall Intervention Analysis for On-Line Applications', *J. Water Resour. Plng. and Mgmt.* **115**(4), 397–415.

Weeks, C. R. and McMahon, T. A.: 1973, 'A Comparison of Water Utilities in Australia and the U.S.', *J. American Water Works Assoc.* **65**(4), 232–241.

Zurada, M. J.: 1992, *An Introduction to Artificial Neural Systems*, PWS Publishing Company, Mumbai, India.