VOL. 5, NO. 1, JANUARY 2010 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

FILLING IN MISSING PEAKFLOW DATA USING ARTIFICIAL NEURAL NETWORKS

Steven K. Starrett[1], Shelli K. Starrett[1], Travis Heier[2], Yunsheng Su[3], Denny Tuan[3] and Mark Bandurraga[3]
[1]Starrett Engineering, LLC, Westmoreland, Kansas, USA
[2]HDR Engineering Inc., W. Highpoint Street, Springfield, Missouri, USA
[3]Ventura County, Watershed Protection District, Planning and Regulatory Division, Ventura, California, USA
E-Mail: stevestarrett@gmail.com

ABSTRACT

The objectives of this study was to: i) use Artificial Neural Networks (ANNs) to fill-in missing data from the peak annual flow rate records for the Santa Clara river watershed, and ii) compare the ANN results with linear regression. Gauging station peaks were modeled with inputs consisting of: peak flows from nearby gauging stations, precipitation data, and temporal data. Model characteristics (number of nodes and layers, transfer functions, data pre-processing methods, etc.) were also studied to optimize the ability of the ANN to learn relationships between the inputs and the peak flows. In general, the models performed well with peak flows from one to four neighboring station, maximum annual 10-d precipitation total data, and the year (representing land use changes); and it was common for testing results to be within 20% of the target. The ANN models had a sum squared error (SSE) value 2 to 400 times less than linear regression models.

Keywords: model, Santa Clara river watershed, peak flow rate, artificial neural networks, missing streamflow data.

INTRODUCTION

Missing streamflow data can be caused by mechanical failures, electronic failures and lack of funding. The ability to estimate missing water resources data is important to water resources planning and management. Llunga and Stephenson (2005) studied the use of artificial neural networks (ANNs) to fill-in missing streamflow data in Africa where water resources data are commonly limited. An extensive review of ANNs concepts and applications in hydrology was performed by the ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (Govindaraju and Rao, 2000). Many modeling methods that incorporate various artificial neural networks have been used to specifically estimate missing streamflow data (Elshorbagy, *et al.*, 2002). Current advances in estimation techniques to predict missing streamflow data continues to incorporate basic ANN concepts (Ng, *et al.*, 2009).

Artificial neural networks are being used to address many different topics in water resources. ANNs were investigated for use in determining rainfall frequencies (He and Valeo, 2009). Artificial neural networks were determined to be more accurate that conventional parametric methods for determining rainfall frequency curves for the Calgary, Canada International Airport location. ANNs modeling techniques were used to estimate missing ocean wave measurements (Makarynskyy, *et al.*, 2005). Correlation coefficients between the ANNs model results and the measured wave heights at different times ranged between 0.85-0.99. Tyagi *et al.* (2008) stated that artificial neural networks out performed regression when filling in missing groundwater quality data.

The objectives of this study were to: i) develop Artificial Neural Networks (ANNs) models to fill-in missing data from the peak annual flowrate records for the Santa Clara River (Ventura and Los Angeles Counties, California) watershed, and ii) compare the ANN results with results from linear regression.

Watershed Characteristics

Watershed information was taken from the "Santa Clara River 1994 Hydrology Study" (Ventura County Watershed Protection District formerly known as Ventura County Flood Control Department). The entire Santa Clara River Watershed is approximately 129 km in length and averages about 40 km in width (Figure-1). Approximately 40 percent of the watershed is located in Los Angeles County and the other 60 percent is located in the downstream Ventura County. The river enters Ventura County about 10 km east of Piru, then flows southward through Fillmore, Santa Paula, and Ventura to the Pacific Ocean. Approximately 90 percent of the watershed consists of high, rugged mountains of up to 2,882 m; the remainder consists of valley floor and coastal plain. The total drainage area is approximately 4,222 square km. There are four (4) principal tributaries located in this watershed. In upstream to downstream order they are: Castaic, Piru, Sespe, and Santa Paula Creeks with drainage areas of 510, 1142, 697, and 109 square km, respectively. Castaic and Piru Creeks are controlled by water storage reservoirs.

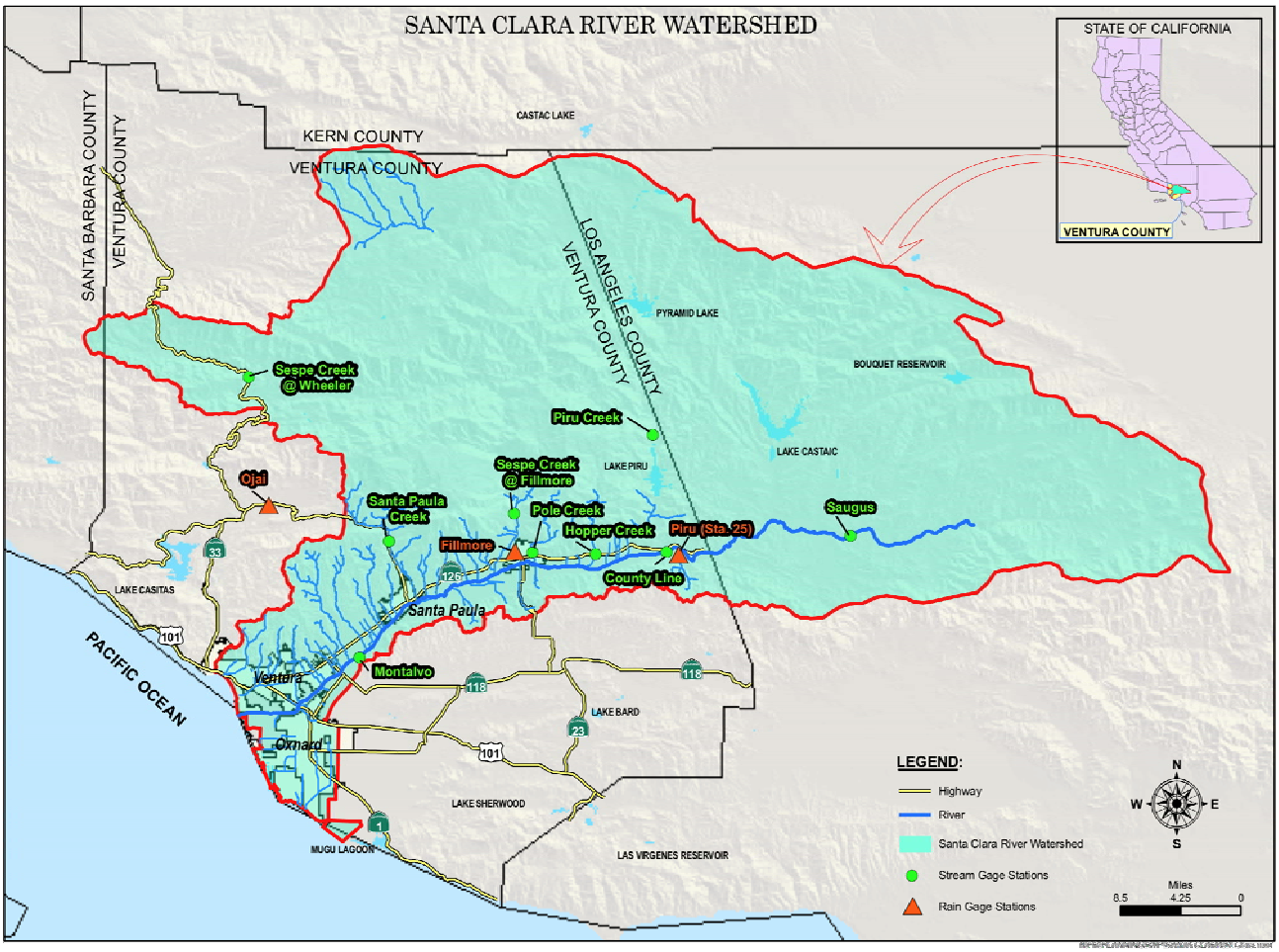

Figure-1. Santa Clara river watershed, LA and Ventura counties, California (Ventura County, CA).

ANN Background

ANNs are a form of artificial intelligence based on imitating neurons in the brain and have been used in many engineering applications (Starrett, 1997 and 1998). Groups of mathematical neurons are interconnected into networks and trained to perform desired functions. In this application, a feed-forward ANN is used to find a mapping between the various inputs (nearby flowrates, precipitation data, year, month, etc.) and the desired peak annual flowrate for a given watershed. A feed-forward network consists of an input layer, one or more hidden layers, and an output layer. Each hidden and output layer neuron is a mathematical element that performs a sum and then passes the result through a transfer function before passing the result on to the next layer (Figure-2). Each interconnection has an associated weight multiplier, and each output and hidden neuron also has a bias parameter.

A back-propagation training method is used to set the parameters of the networks. The network is initialized with a random set of weights and biases. Back-propagation training consists of using sets of sample input and target output data to adjust the network parameters by applying a sample input, finding the network output, calculating the error measured by comparing the ANN output to the target output, and propagating the error backwards through the network. At each layer, the associated error is used to adjust the network parameters (weights and biases) to reduce the error. This process of adjusting the ANN parameters is done interactively either one sample at a time or in batches of samples all processed at once. This is done repeatedly with each training cycle being called an epoch. By training a network with appropriate samples of input and output targets, a mapping function is formed between the input and output. Once the network is satisfactorily trained, it can be used to predict outputs given new sets of inputs.

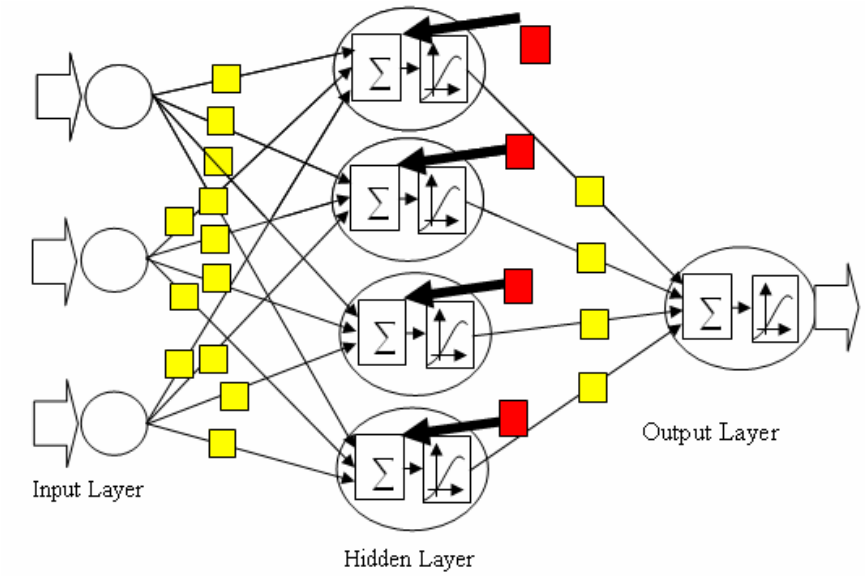

Figure-2. A typical feed-forward ANN.

Because the network weights and biases are set using samples of input and output data, having appropriate data is critical to the development of an ANN. Likewise, it is also important to have the appropriate input data available for predicting desired outputs. ANNs can find mappings in cases where no analytical or empirical equations exist or are too complex to utilize. The ANN training can find the functional connections between inputs and outputs, thus inputs that are related to the output but in an unknown way can be used for an ANN application. To be useful, the inputs must vary over the data set.

In using ANNs for a specific application, the selection of network output is an early consideration. In this case, it was decided that each sub-watershed would be modeled by a separate ANN. This allows each network to be more straightforward, and increases the accuracy of the results. Using multiple, more specialized, networks instead of one larger more generic network avoids the need to determine parameters that do not vary for the smaller application, but could be needed in a more generic application. In this application, watershed parameters such as drainage area and curve number are fairly constant and not useful inputs for an ANN specialized on a single sub-watershed. If a larger, more generic ANN were to be developed, these parameters, in addition to those needed for the specialized ANNs, could be needed.

In applications such as this where training data are limited, overfitting (or poor generalization) by a network is a concern. Poor generalization means that the network mapping function fits the training data very well, but between the sample data points the network function varies in a too complex way that it doesn't fit new data appropriately. Several strategies were used to avoid overfitting in this application: 1) one hidden layer with less than 10 neurons, 2) less than 200 epochs training, and 3) use of Bayesian Regularization training algorithm. Each of these work in different way to "smooth" the mapping and help ensure better fits. Bayesian regularization utilizes a training method that limits the sizes of the network parameters to help produce a smoother mapping [see Matlab ANN manual for more detail].

Sum Squared Error (SSE) was used to measure the success of the ANNs in this application. To test ANNs, a percentage of the training set is removed from the set. The testing data is run through the trained network and the SSE is calculated for this data as well. For this study around ten percent of the data was held out as testing data. The various runs were compared in terms of total SSE (SSE of the whole set = training SSE plus testing SSE), and in terms on the separate training and testing SSEs.

METHODS

In order to find the missing peak annual flowrate data, an analysis of the available inputs was performed. The peak flowrates at nearby stations had been used previously in a regression method (Ventura County, 1994), and they were also used in this study. Water year rain totals, 10-d maximum rain totals, day of peak annual flowrate, and month of peak annual flow precipitation variables were considered. In addition, the year of interest and the month of peak annual flowrate were also studied. The usefulness of the year and month variables illustrates the ability of ANNs to extract trends in data from the applied inputs. Because development of the watershed increases slowly from year to year, the degree of land use is related to the year. In addition, it is feasible that long term weather trends and vegetative cover (i.e. fire burn) patterns may be captured by using the year as a variable. The year was entered in a year format (i.e. 1933 = 1933). In the same manner, the month of peak flow can be related to variables such as: temperature, evapotranspiration, soil moisture conditions, and vegetative cover which can affect flowrates. The month was entered as a number (i.e. January = 1).

To determine the usefulness of the various input variables, multiple ANNs were built and tested. Multiple training runs were done with each ANN model (selected inputs, number of neurons, input/output normalization, training algorithm, maximum number of epochs).

Another critical factor in the processing of the input data was the selection of data with consistent dates of peak annual flowrates. Because of the size and variations of topography in the watershed, peak flowrates of the Santa Clara River and its various tributaries don't always happen during the same storm. This inconsistency could lead to difficulty in finding the mapping of inputs to outputs. Thus, for this work, years with inconsistent dates (as much as could be determined) were not used for training. In addition, when using a flowrate as an input, the years for which the input data flowrates are missing must also be removed.

Three primary sets of training data were developed by removing the appropriate data. Years with inconsistent dates were removed from all sets. Set A eliminated missing data for all six base stations. This set was used to train ANNs when Montalvo station flowrate was used as an input. A second set (Set B) which kept the years with missing Montalvo station data, but eliminated those with missing data from the other base stations was used for comparison of most of the cases. A third set (Set C) was used when both Montalvo station and Saugus station flowrates were not used. Set C was largest of the three and was used when Saugus station and Montalvo station flowrates were not found to be in the cases with the best SSE. In addition, for each sub-watershed ANN model, those years for which the data was missing were also eliminated from the training data.

Scaling of Inputs and Outputs

Before sets of input and output samples are applied to the network, they are normalized to make them all comparable. In this project, inputs were scaled by subtracting the mean value of the data set and then dividing by the standard deviation. This produces a normalized set of inputs with a mean value of zero. The outputs were scaled in the same manner and then an additional scaling was done to produce normalized targets with a minimum value near zero and a maximum near one.

When new inputs are applied to the network they are first scaled using the parameters of the training set. In a similar manner, the network outputs are reverse-normalized using the inverse of the scaling function used in the training. These scaling steps are included in the ANN model to make it more user-friendly.

Size of Networks

As stated earlier, smaller networks are desirable to avoid overfitting. In this study, many runs were done with different numbers of neurons to determine suitable sizes for each of the networks. Typically, runs with 5, 10, and 15 neurons were done for each case in determining which inputs to use. Then, once a set of inputs was selected for a particular site, a range of cases with 4 to 10 neurons was run and a suitable number was chosen.

Selection of a Final Model

Over 1000 different models were tried and studied. Once the configuration of the network for each site was chosen, a large number of runs (typically approximately 300) were done to find the five best runs. From this set, a single run was chosen by looking at the predictions for the missing data. Size of the flowrates, consistency of results, and total, training, and testing SSE were used in making the final selections.

RESULTS

The ANNs developed to fill-in missing data at eight stations varied in inputs used and number of hidden neurons in the network (Table-1). The SSE given is for the combined training and testing sets for the corresponding case. The cases have differing numbers of points as discussed earlier, and stations with larger flowrate values have larger errors. Developed ANN models performed well (Figure-3) in estimating missing peakflow data. In general, the developed model for Hopper Creek over estimated when flows were very low, and performed well during high flow events.

Comparing ANNs and Linear Regression

Linear regression fitting was done using the same training data as used for the ANN and using Matlab (Figure-4, Table-2). The first two data columns show the SSE results for the training/testing data sets for the ANN and regression models. The whole data set contains numerous years of data with non-matching dates. The models were developed for the reduced set and selection of inputs was in some cases based on data available for the missing data years for a specific station. One important point to note is that the regression models give negative flowrate values on several points in the large data set. It should also be noted that when the data are missing, the SSE calculation has a zero measured value and this will lead to larger SSEs. The ANN models had orders of magnitude less SSE when compared with the linear models. Others have also shown ANN models are superior over linear regression models in estimating missing data because of the chaos and nonlinearity of streamflow (Elshorbagy, *et al*., 2001).

Table-1. ANN training results for Santa Clara River gauging stations.

Sub-watershed station	ANN Inputs	# Hidden neurons	# Samples	SSE for set**
Santa Clara R. at Montalvo (USGS #11114000)	Flowrates: Sespe-Fillmore, Santa Paula Precipitation: Ojai 10d, Sta. 25 10d Other: Year, month of max flowrate	8	32	2.37×10^8
Piru Creek above Lake Piru* (USGS #11109600)	Flowrates: Sespe-Fillmore, Santa Paula, Hopper Creek Precipitation: Ojai Water Yr, Fillmore Water Yr Other: Year	7	41	4.9×10^5
Santa Paula Creek near Santa Paula (USGS #11113500)	Flowrates: Santa Clara-Montalvo, Piru above Lake Precipitation:Ojai 10d, Fillmore 10d Other: Year	7	35	32×10^5
Hopper Creek (USGS #11110500)	*Flowrates*: Santa Paula, Sespe-Fillmore Precipitation: Sta. 25 10 d Other: Year, month of max flowrate	8	39	16×10^5
Santa Clara at Saugus (Old Highway) (USGS #11108000)	Flowrates: Santa Clara-Montalvo, Piru above Lake Precipitation: Station 25 10 d Other: Year, month of max flowrate	8	29	80.1×10^5
Pole Creek (VCWPD #713)	Flowrates: Hopper Creek, Piru above Lake Precipitation: Fillmore 10 d Other: Year, month of max flowrate	7	18	50.3
Sespe Creek near Wheeler (USGS #11111500)	Flowrates: Sespe-Fillmore, Santa Paula, Hopper Creek Precipitation: Fillmore 10 d, Station 25 10 d Other: Year	4	30	51×10^5
Santa Clara at County Line/Piru (USGS #11108500)	Flowrates: Sespe-Fillmore, Santa Paula, Santa Clara-Montalvo, Piru above Lake Precipitation: Ojai 10d, Fillmore 10d Other: Year, month of max flowrate	5	31	23.9×10^5

* Data for Piru above lake for 1933-1955 was found by scaling the retired station Piru at Piru (#11109700) for the smaller watershed size of the new, after lake construction upstream gauging station location.

** SSE between sub-watersheds is not comparable because of flowrate magnitude differences and differing numbers of data points.

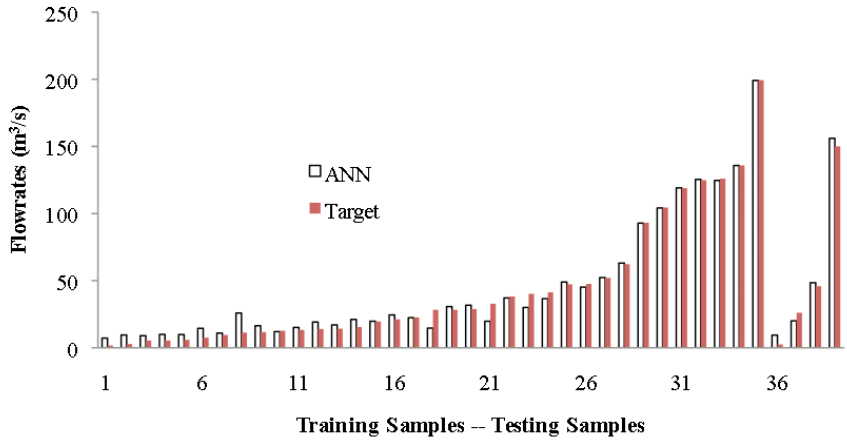

Figure-3. Example of ANN model training and testing prediction data versus targets (Hopper Creek). The first set of ranked values is from the training process, the second set of ranked data (36-40) are from the testing of the developed model.

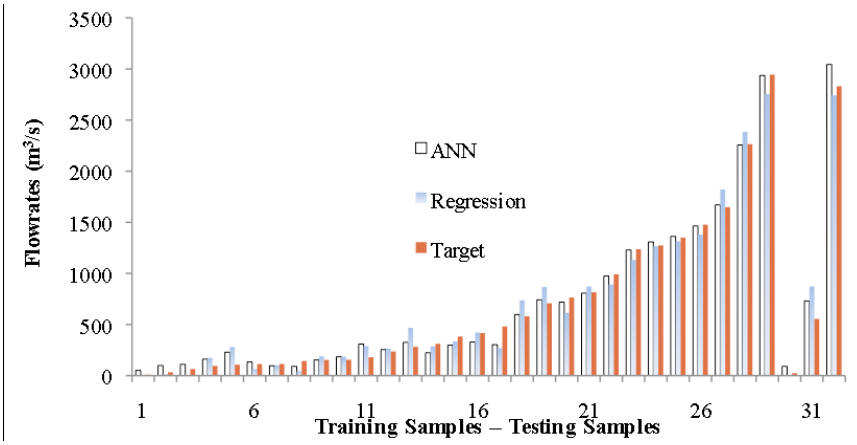

Figure-4. Comparison of ANN, linear regression predictions with measured data at Montalvo Station, Santa Clara River. First series is data used in training and the second series is data used for testing (testing data not used in training of ANN model or regression model).

Table-2. Comparison of ANN and linear regression models.

Station	**ANN test/train SSE**	**Regression test/train SSE**	**ANN better** (≈X times)
Santa Clara- Montalvo	2.38E+008	5.67E+008	2
Piru above Lake	4.86E+005	1.88E+008	387
Santa Paula	3.19E+006	1.16E+008	36
Hopper Creek	1.61E+006	2.25E+007	14
Santa Clara-Saugus	8.01E+006	2.55E+008	32
Pole Creek	5.03E+001	1.66E+006	33002
Sespe-Wheeler	5.07E+006	2.20E+007	4
Santa Clara-County Line	2.39E+006	1.69E+008	71

CONCLUSIONS

ANN developed models have proven themselves to be able to simulate very complex situations and problems. There are complex relationships between peakflow, and watershed and precipitation characteristics. ANNs tend to represent these relationships more accurately than linear regression models. Our findings support the use of ANN technology in estimating missing peakflow data.

REFERENCES

Elshorbagy A., Panu U.S. and Simonovic S.P. 2001. Analysis of cross-correlated chaotic streamflows. Hydrol. Sci. J. 46(5): 781-792.

Elshorbagy A., Simonovic S. P. and Panu U. S. 2002. Estimation of missing streamflow data using principles of chaos theory. J. Hydrol. 255: 123-133.

Govindaraju R.S. and Rao A. Ramachandra. 2000. Artificial Neural Networks in Hydrology. Kluwer Academic Publishers, Dordrecht, Nertherlands.

He Jianxun and Valeo Caterina. 2009. Comparative study of ANNs versus parametric methods in rainfall frequency analysis. J. Hydrologic Engg. 14(2): 172-184.

Llunga M. and Stephenson D. 2005. Infilling streamflow data using feed-forward back-propogation (BP) artificial neural networks: Application of standard BP and pseudo Mac Laurin power series BP techniques. Water SA. 31(2): 171-176.

Makarynskyy O., Makarynska D., Rusu E. and Gavrilov A. 2005. Filling gaps in wave records with artificial neural networks. In: Maritime Transportation and Exploitation of Ocean and Coastal Resources. Guedes Soares, Garbatov and Fonseca (eds). Taylor and Francis Group, London. 2: 1085-1091.

Ng W.W., Panu U.S. and Lennox W.C. 2009. Comparative studies in problems of missing extreme daily streamflow records. J. Hydrologic Engg. 14: 91-100.

Starrett Steven K., Starrett Shelli K., Najjar Y., Adams G. and Hill J. 1998. Modeling pesticide leaching from golf courses using artificial neural networks. Comm. in Soil Sci. and Plant Anal. 29: 3093-3106.

Starrett Steven K., Starrett Shelli K. and Adams G.L. 1997. Using Artificial Neural Networks and Regression to Predict Percentage of Applied Nitrogen Leached under Turfgrass. Comm. in Soil Sci. and Plant Anal. 28: 497-507.

Tyagi P., Chandramouli V., Lingireddy Srinivasa and Buddhi D. 2008. Relative performance of artificial neural networks and regression models in predicting missing water quality data. Environ. Engg. Sci. 5(25): 657-668.

Ventura County Watershed Protection District formerly known as Ventura County Flood Control Department. 1994. Flood Control Department, October 27th. Santa Clara River 1994 Hydrology Study. Ventura County, CA.