Prediction of Surface Water Supply Sources for the District of Columbia Using Neural Networks Methods

Francisco Lourenco, Nian Zhang and Sasan Haghani
Department of Electrical and Computer Engineering, University of the District of Columbia, 4200 Connecticut Ave NW, Washington, DC, 20008

Abstract

Water availability for municipal and industrial use, irrigation, navigation support, hydropower, and environmental flows is a significant concern in regions throughout the United States. Hence, there is a need to develop analytical tools to assess and help in the long-term planning of the availability of water supply sources.

In this paper, we studied the long-term prediction of surface water resources at the Potomac Watershed in the District of Columbia. A predictive model, based on recurrent neural networks, trained with the Levenberg-Marquardt backpropagation learning algorithm is proposed to forecast the runoff discharge using the past runoff discharge and gage height. Using this computational intelligence modeling tool, the impact of discharge and gage height to the long-run discharge forecast accuracy was studied. Our experimental results indicate that the proposed learning algorithm can successfully train the recurrent neural network for the runoff prediction.

Keywords

Runoff prediction, water quantity prediction, time series prediction, neural networks, backpropagation learning algorithm.

1. Introduction

A wide range of evidence indicates that the earth has been warming over the past century. This warming is causing the melting of mountain glaciers and sea ice in many parts of the world, a rise in sea levels, and changes in patterns of precipitation. Most scientists agree that these trends are likely to continue, and to accelerate largely due to increasing levels of carbon dioxide and other "greenhouse" gases in our atmosphere. Changes in temperatures and precipitation may impact the availability, use, and management of water resources. Since the publication of Intergovernmental Panel on Climate Change (IPCC) reports in 2007, many federal agencies have been developing guidelines for planning, design, operation of water resources systems that include the potential impacts of climate change.

A 2009 report by the United States Geological Survey entitled "Climate change and Water Resources Management: A Federal Perspective" identified that climate change could affect all sectors of water resources management, since it may require changed design and operational assumptions about resource supplies, system demands or performance requirements, and operational constraints [1]. Water availability for municipal and industrial use, irrigation, navigation support, hydropower, and environmental flows is a significant concern in regions throughout the United States. Thus, there is a need to develop analytical tools to assess, and help in long-term planning of the availability of water supply sources.

Climate change may have a significant impact on current Washington metropolitan area (WMA) water supplies such as Potomac River watershed. Though it is uncertain whether precipitation will increase or decrease in our region, study results indicate that higher temperatures may raise rates of evaporation and evapotranspiration to a significant degree [2]. The higher evapotranspiration rates are predicted to reduce the amount of water available to recharge basin aquifers and to decrease flows in the Potomac River and in streams that fill and replenish system reservoirs. Study simulations produced a wide range of effects. However, under the assumption that no changes are made to the WMA system, results indicate that in a basin altered by climate change a moderate drought occurring in the year 2040 may cause the imposition of emergency water use restrictions, nearly empty reservoirs, which can lead to water supply shortages.

To ensure that water supply systems continue to meet demands and satisfy environmental flow requirements, it is important that water supply planners keep abreast of developments in climate science and regularly review and assess local trends and projections of how hydrologic conditions might change in the coming decades. Under the set of cooperative agreements which govern water supply planning and management in the Washington Metropolitan Area, the area's three main water suppliers are committed to conducting regular forecasts of future demands and resources.

Climate change will likely add additional stress to a system facing the challenge of future population growth. The region's suppliers are also committed by cooperative agreements to increase water availability if assessments determine a need to do so. This could be done by funding structural solutions and/or other means of ensuring a reliable supply. To this end, studies on alternatives to increase water supplies have been conducted. These alternative options include use of the Potomac and Occoquan estuaries as supplies, and retired quarries as storage facilities. Other measures that could improve system performance under climate change include:

1) increased flexibility in shifting between the system's Potomac and off-Potomac resources,

2) improved stream flow forecasts to inform reservoir release decisions, and

3) earlier and stricter water use restrictions.

In order to study the long-term predication of surface water sources at the Potomac Watershed, this paper proposes the development of predication models based on computer intelligence methods. The study area is focused on the free-flowing portion of the Potomac River, the primary water supply source for the Washington D.C. metropolitan area. The Potomac River is one of the least dam-regulated large river systems in the eastern United States [3]. The Potomac River has the highest level of nitrogen and the third highest level of phosphorus loading of all the major rivers in the Chesapeake Bay watershed. These nutrients can limit the growth of submerged aquatic vegetation, cause low oxygen conditions and create dead zones.

Approximately 90% DC area drinking water comes from Potomac River. The Washington Aqueduct is right beside the Potomac River, as shown in Fig. 1. It produces drinking water for approximately one million citizens living, working, or visiting in the District of Columbia, Arlington County, Virginia, and the City of Falls Church, Virginia, and its service area [4]. In the last three decades, many areas in the watershed have seen their population more than double. A growing population alters and stresses the natural state of its land and water. The Potomac watershed is expected to add more than 1 million people to its population over the next 20 years [5]. The most densely populated area in the watershed is the Middle Potomac, including Washington, DC, which is home to 3.72 million or about 70% of the watershed's population. In the next 20 years, the population of the Potomac watershed is expected to grow 10% each decade, adding 1 million inhabitants to reach a population of 6.25 million. The Potomac River delivers the largest amount of sediment to the Chesapeake Bay each year which can limit the growth and submerged aquatic vegetation and affect populations of all fish, shellfish and birds that depend on this vegetation as a source of food or shelter.

Given the existing flow conditions of Potomac River, reliable estimation of stream flows for the District of Columbia is very important. Water resources professionals and regulatory authorities need this streamflow information for planning, analysis, design and operation & maintenance of water resources systems (e.g., water supply systems, dams and hydraulic structures).

The study area will focus on the Four Mile Run at Alexandria, VA. The Four Mile Run is 9.2 miles long, and is a direct tributary of the Potomac River, which ultimately carries the water flowing from Four Mile Run to the Chesapeake Bay, as shown in Fig. 1.

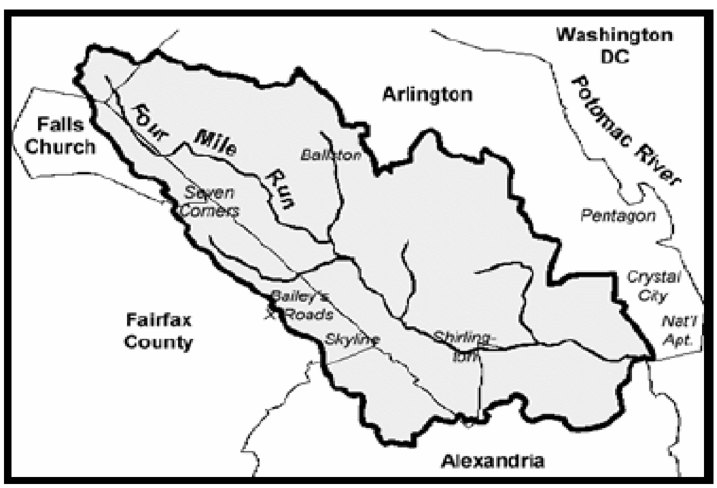

Fig. 1. Four Mile Run at Alexandria, VA is a nine-mile long stream located in a highly urbanized area in Northern Virginia. It is a direct tributary of the Potomac River, which ultimately carries the water flowing from Four Mile Run to the Chesapeake Bay.

The stream passes from the Piedmont through the fall line to the Atlantic Coastal Plain, and eventually empties out into the Potomac River. Potomac River was determined to be one of the most polluted water bodies in the nation mainly due to the CSOs and stormwater discharges and wastewater treatment plant discharges. In addition, because of the highly urbanized nature of the Four Mile Run watershed, the neighborhoods and businesses adjacent to this portion of the run were subjected to repeated flooding, beginning in the 1940s. Therefore, the flood-control solutions are the major concern. Runoff prediction would provide a promising solution for flood-control.

Many research studies have been performed to forecast the runoff. They benefit substantially from the progress of computational intelligence techniques [6]. The techniques include feedforward neural networks [7], radial basis function (RBF) neural network [8][9][10], fuzzy logic [11], evolutionary algorithm [12], support vector machine [13], particle swarm optimization [14], or the combination of them [15][16]. Comparatively, various runoff forecast models based on neural networks perform much better in accuracy than many conventional prediction models.

However, a fact could not be neglected that most of the existing computational intelligence based models have not yet satisfied researchers in forecast precision, and the generalization capability of these networks needs further improving. In addition, none of computational intelligence methods is used for the urban runoff prediction in the District of Columbia and the suburbs, although a few runoff quality analysis tools of urban catchments with probabilistic models were developed [17].

To resolve the above problems, it is extremely important to investigate state-of-the-art computational intelligence with the potential for higher rates for urban runoff forecast. Based on the fact that neural networks [18], genetic regulatory network [19], echo state network [20], particle swarm optimization [21][22], and a number of computational intelligence methods [23][24] have very successfully applications on the time series prediction problems, and because time series prediction is a generalized form of runoff quantity and quality prediction, we expect these methods will also work the best for the runoff prediction problem.

This paper is organized as follows. In Section 2, the data and design methods are presented and the study area and the runoff data are briefly introduced. The neural network architecture and the learning algorithm are also illustrated In Section 2. In Section 3, experimental results are provided. Finally, the conclusions are given in Section 4.

2. DATA and METHODS

Study Area and Stormwater Runoff Data

Real-time stormwater runoff data are obtained from the U.S. Geological Survey (USGS). The Four Mile Run station provides both the discharge data and gage height data, which is useful for investigating their impact to the long-run discharge forecast. The runoff data was retrieved for 120 days between August 28, 2010 and December 4, 2010. The runoff discharge (cubic feet per second) data is plotted in Fig. 2.

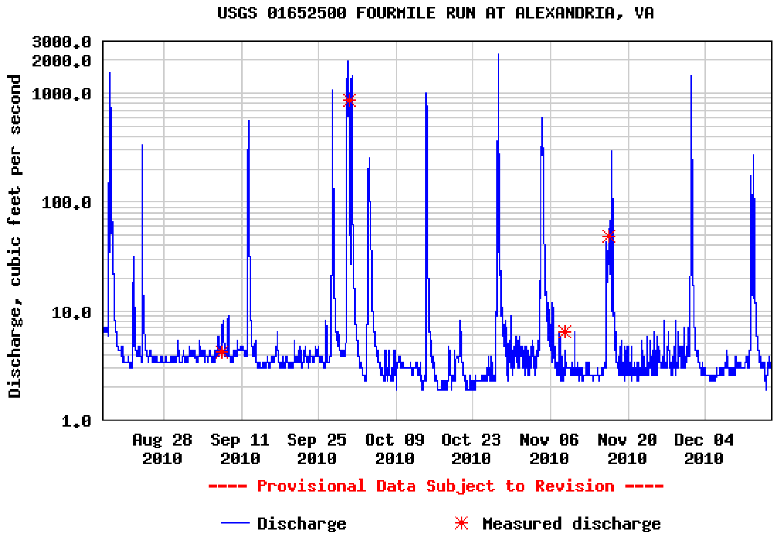

Fig. 2. The runoff discharge data (cubic feet per second) collected at the Four Mile Run site at Alexandria, VA during August 28, 2010 to December 4, 2010.

Real-time data typically are recorded at 15- to 60-minute intervals. Therefore, each figure plots 34721 data during the 120 days. The input, gage height is a 34721x1 matrix, representing dynamic data, i.e. 34721 time steps of 1 element. The target, discharge is of the same length as the input, which represents 34721 time steps of 1 element dynamic data. The target data are randomly divided up into 34721 time steps. 70% of the data is used for training. They are presented to the network during training, and the network is adjusted according to its error. 15% will be used to validate that the network is generalizing and to stop training before overfitting. The last 15% will be used as a completely independent test of network generalization.

Neural Network Architecture

Since previous values of discharges are needed, a recurrent neural network based predictive models is to be developed to predict future values of runoff discharge, based on the previous runoff discharge. The predictive model can be represented mathematically by predicting future values of the discharges time series $y(t)$ from past values of that time series. This form of prediction can be written as follows:

$$y(t) = f(y(t-1), \ldots, y(t-d))$$

The proposed neural network model is a two-layer feedforward network, with a sigmoid transfer function in the hidden layer and a linear transfer function in the output layer, as shown in Fig. 3. W is the weight matrix, and b is the bias. This network also uses tapped delay lines to store previous values of the $y(t)$ sequence.

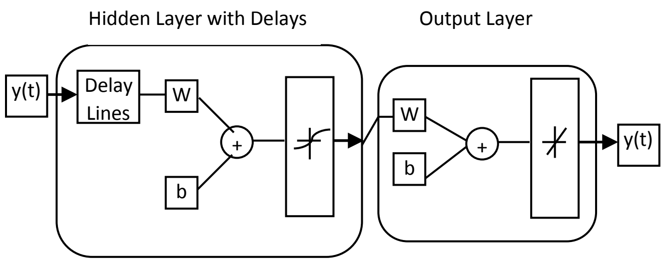

Fig. 3. Neural network architecture for the predictive model. The network is a two-layer feedforward network, with a sigmoid transfer function in the hidden layer and a linear transfer function in the output layer. This network also uses tapped delay lines to store previous values of y(t) sequences. W is the weight matrix, and b is the bias.

The above neural network is trained in open loop form. Open loop (single-step) is more efficient than closed loop (multi-step) training. Open loop allows us to supply the network with correct past outputs as we train it to produce the correct current outputs. When the feedback loop is open, it is performing a one-step-ahead prediction. It is predicting the next value of $y(t)$ from previous values of $y(t)$. With the feedback loop closed, it can be used to perform multi-step-ahead predictions. This is because predictions of $y(t)$ will be used in place of actual future values of $y(t)$. After training, the network may be converted to closed loop form that the application requires.

Learning algorithm

The neural network is trained using Levenberg-Marquardt backpropagation algorithm. It is a network training function that updates weight and bias values according to Levenberg-Marquardt optimization. It is often the fastest backpropagation algorithm for training moderate-sized feedforward neural networks (up to several hundred weights), although it does require more memory than other algorithms.

Like the quasi-Newton methods, the Levenberg-Marquardt algorithm was designed to approach second-order training speed without having to compute the Hessian matrix. When the performance function has the form of a sum of squares (as is typical in training feedforward networks), then the Hessian matrix can be approximated as

$$H = J^T J$$

and the gradient can be computed as

$$g = J^T e$$

where J is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights and biases, and e is a vector of network errors. The Jacobian matrix can be computed through a standard backpropagation technique that is much less complex than computing the Hessian matrix.

The Levenberg-Marquardt algorithm uses this approximation to the Hessian matrix in the following Newton-like update [25]:

$$x_{k+1} = x_k - [J^T J + \mu I]^{-1} J^T e$$

When the scalar μ is zero, this is just Newton's method, using the approximate Hessian matrix. When μ is large, this becomes gradient descent with a small step size. Newton's method is faster and more accurate near an error minimum, so the aim is to shift toward Newton's method as quickly as possible. Thus, μ is decreased after each successful step (reduction in performance function) and is increased only when a tentative step would increase the performance function. In this way, the performance function is always reduced at each iteration of the algorithm.

Training stops when any of these conditions occur:

- The maximum number of epochs (repetitions) is reached.
- The maximum amount of time is exceeded.
- Performance is minimized to the goal.
- The performance gradient falls below the minimum gradient.
- μ exceeds its maximum set value
.

3. EXPERIMENTAL RESULTS

Number of Hidden Neurons and Delays

Increasing the number of neurons and the number of delays requires more computation, and this has a tendency to overfit the data when the numbers are set too high, but it allows the network to solve more complicated problems. The goal of these simulations was to ascertain which would be the best combination between the number of delays and the number of hidden neurons layers in the proposed neural network model.

We continuously increase both the number of neurons in the hidden layer and the number of delays in the tapped delay lines until the network performed well in terms of the mean squared error (MSE) and the error autocorrelation function. The mean squared error is the mean squared normalized error performance function. The error is the difference between the output and the target. After several trials, the best number of hidden neurons is determined to be 10, and the best number of delays in the tapped delay lines is 2.

After testing different number of delays and hidden neurons layers, with the help of Matlab Neural Network Time Series Prediction, we got to a final set of delays, 2, 10, 20 and 50, and a set of hidden neurons layers, 8, 10, 12 and 15. As shown in Fig.4, with this final set, for the data points applied, the best performance was achieved with the tapped delay equal to 2 and 10 hidden neurons layers.

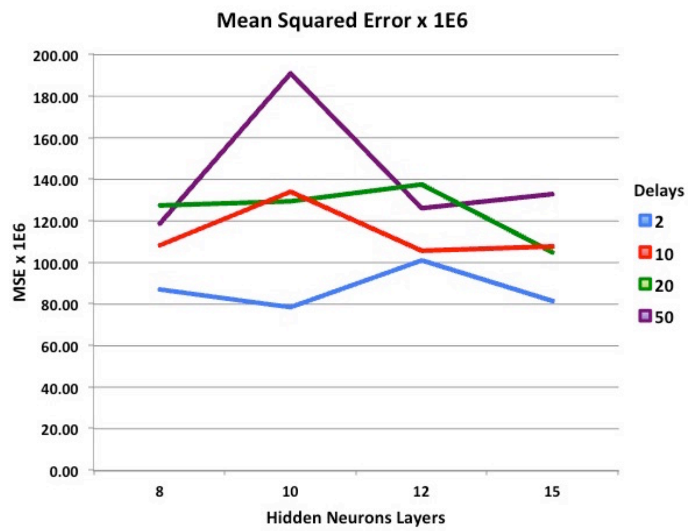

Fig. 4. Mean squared error for different number of delays and hidden neurons layers.

Fig. 4 shows that the best result is obtained with only 2 delays, given the same number of hidden neurons. This came about because the original discharge data had an average value of 4.2397, with a standard deviation of 0.2538. Only 5.7% of data points were out of this standard deviation. In Fig. 5, the actual discharge and the predicted value are shown in the top portion of the figure, while the bottom portion of the figure shows the difference between the actual and the predicted values. Note that the predicted values corresponding to the smallest MSE (equal to 7.85e-5) is plotted in Fig. 5. One can observe from Fig. 5 that the actual discharge and the predicted values overlap, showing that the prediction is very accurate. Fig. 5 also shows that the peaks in the prediction error occur whenever there is a peak in the actual discharge.

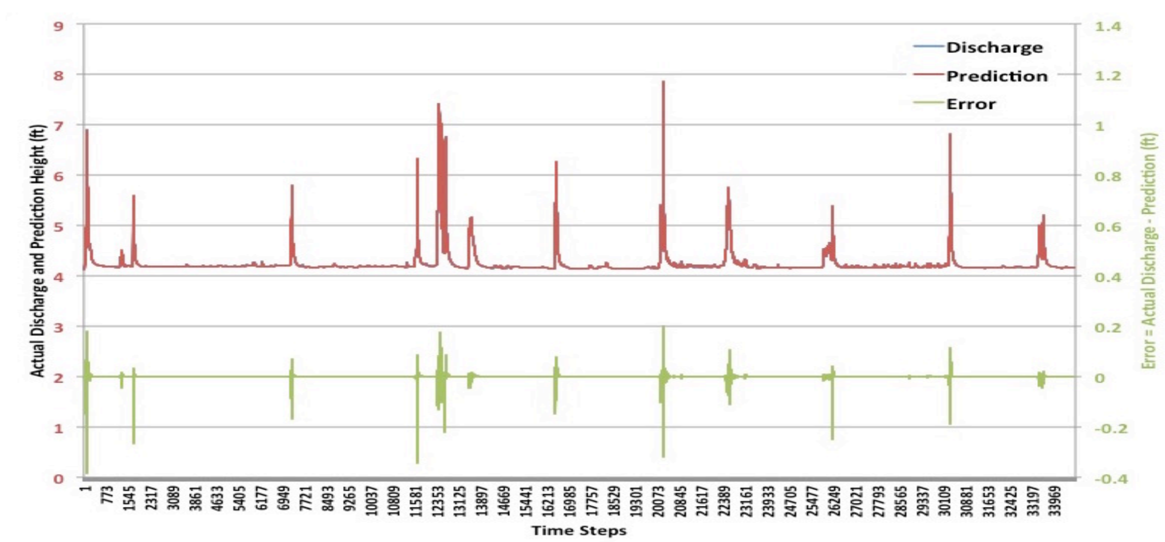

Fig. 5. Actual Discharge and predicted values based on the best number of hidden neurons and tapped delays. Bottom figure shows the difference between the actual discharge and the predicted value.

Time Series Prediction Performance

Validation vectors are used to stop training early if the network performance on the validation vectors fails to improve or remains the same, as indicated by an increase in the mean squared error of the validation samples. Test vectors are used as a further check that the network is generalizing well, but do not have any effect on training. The best validation performance is 0.000194670 at epoch 28 when the input to the neural network is the discharge height, as shown in Fig. 6. It demonstrates that training, validation and testing errors decreased to 1.96052e-4, 1.94670e-4, and 1.68724e-4, respectively until iteration 28. It does not appear that any over fitting has occurred, since neither testing or validation error increased before iteration 28.

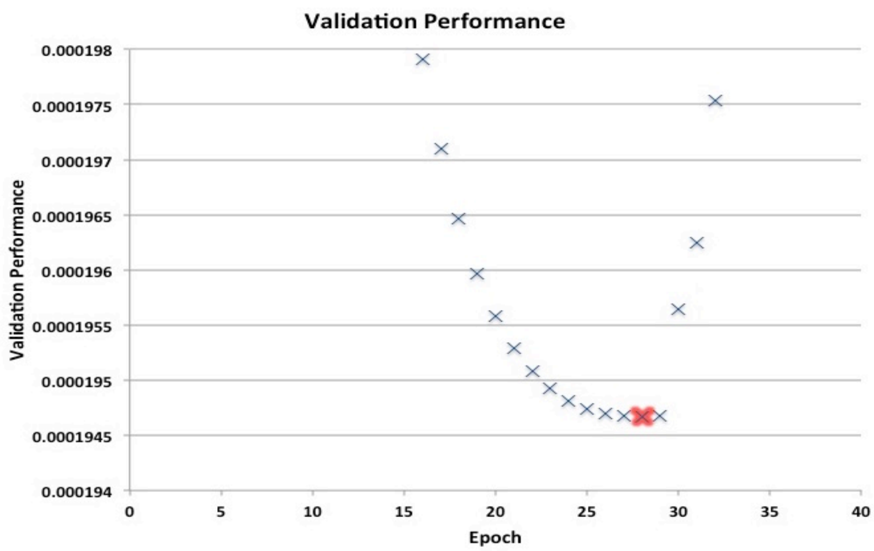

Fig. 6. The best validation performance is 0.000194670 at epoch 28 when the input is discharge height.

4. CONCLUSIONS

We proposed a predictive model based on recurrent neural networks trained with the Levenberg-Marquardt back-propagation learning algorithm to forecast the runoff discharge using the past runoff discharge. A two-layer feed-forward network, with a sigmoid transfer function in the hidden layer and a linear transfer function in the output layer was developed. The runoff at the Four Mile Run station was studied, because of its impact to the Chesapeake Bay and Potomac River near Washington D.C. The input data are discharge and gage height with 120 days duration from August 28, 2010 and December 4, 2010. Our simulations showed that the best number of hidden neurons is 10, and the best number of delays in the tapped delay lines is 2. After testing the input data of discharge and gage height with several different combinations of hidden neurons and delays, it was found that this number of hidden neurons layers and delays generated the least mean squared error. Our experimental results show that the proposed learning algorithm is successful in training the recurrent neural network for the runoff prediction. Through this paper a graduate student was trained in neural networks theory and modeling. The student is now trained to apply neural networks learning algorithm to engineering applications on solar radiation prediction, climate change, stock market analysis and process and quality control.

ACKNOWLEDGEMENT

This work was supported in part by the U.S. National Science Foundation under Grants HRD #1435947, HRD #1505509, and USGS Grant under DC WRRI Grant #2015DC174B.

References

[1] L.D. Brekke, J.E. Kiang, J.R. Olsen, R.S. Pulwarty, D.A. Raff, D.P. Turnipseed, R.S. Webb, and K.D. White, "Climate change and water resources management—A federal perspective," U.S. Geological Survey Circular 1331, 65 p. (Also available online at http://pubs.usgs.gov/circ/1331/.)

[2] S.N. Ahmed, K.R. Bencala, and C.L. Schultz, 2010 Washington Metropolitan Area Water Supply Reliability Study, 2013. http://www.potomacriver.org/publicationspdf/ICPRB13-07.pdf.

[3] A Snapshot of Potomac Watershed Health, Potomac Conservancy, http://www.potomac.org/site/snapshot-watershed/index.php.

[4] Potomac Conservancy, State of the Nation's River, Potomac Watershed. 2007. Available: http://www.potomac.org.

[5] US Army Corps of Engineers, http://washingtonaqueduct.nab.usace.army.mil/.

[6] S. Haykins, *Neural Networks: A Comprehensive Foundation* (Second Edition, Prentice Hall, 1998).

[7] K. Solaimani, "Rainfall-runoff Prediction Based on Artificial Neural Network (A Case Study: Jarahi Watershed)," *American-Eurasian J. Agric. & Environ. Sci.*, *5*(6), 2009, 856-865.

[8] J. Zhang, C. Zhu, Prediction of Reservior Runoff Using RBF Neural Network-Grey System United Model, *IITA International Conference on Control, Automation and Systems Engineering (CASE)*, 2009, 43 - 46.

[9] S. Hui, L. Xinxia, Multi-scale RBF Prediction Model of Runoff Based on EMD Method, *2010 Third International Conference on Information and Computing (ICIC)*, 2010, 296 - 299.

[10] W. Guo, H. Wang, J. Xu, Y. Zhang, RBF Neural Network Model Based on Improved PSO for Predicting River Runoff, *2010 International Conference on Intelligent Computation Technology and Automation (ICICTA)*, 2010, 968 - 971.

[11] W. Wang, L. Qiu, Prediction of Annual Runoff Using Adaptive Network Based Fuzzy Inference System, *2010 Seventh International Conference onFuzzy Systems and Knowledge Discovery (FSKD)*, *3*, 2010, 1324 - 1327.

[12] P. Bai, X. Song, J. Wang, W. Shi, Q. Wang, A Hillslope Infiltration and Runoff Prediction Model of Neural Networks Optimized by Genetic Algorithm, *2010 International Conference on Mechanic Automation and Control Engineering (MACE)*, 2010, 1256 – 1259.

[13] J. Liu, B. Wang, Parameters Selection for SVR Based on the SCEM-UA Algorithm and its Application on Monthly Runoff Prediction, *2007 International Conference on Computational Intelligence and Security*, 2007, 48 - 51.

[14] W. Guo, H. Wang, J. Xu, Y. Zhang, BF Neural Network Model Based on Improved PSO for Predicting River Runoff, *2010 International Conference on Intelligent Computation Technology and Automation (ICICTA)*, *2*, 2010, 968 – 971.

[15] H. Bo, X. Dong, X. Deng, Application of GA-ANN Hybrid Algorithms in Runoff Prediction, *2010 International Conference on Electrical and Control Engineering (ICECE)*, 2010, 5039 – 5042.

[16] Y. Liu, Y. Chen, J. Hu, Q. Huang, Y. Wang, Long-term Prediction for Autumn Flood Season in Danjiangkou Reservoir Basin Based on OSR-BP Neural Network, *2010 Sixth International Conference on Natural Computation (ICNC)*, *4*, 2010, 1717 - 1720.

[17] P. Behera, J. Li, and B. Adams, Runoff Quality Analysis of Urban Catchments with Analytical Probabilistic Models", Journal of Water Resources Planning and Management, ASCE, 2006.

[18] D. K.. Barrow, S.F. Crone, N. Kourentzes, An Evaluation of Neural Network Ensembles and Model Selection for Time Series Prediction, *2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, 1-8.

[19] N.A. Barker, C.J. Myers, H. Kuwahara, Learning Genetic Regulatory Network Connectivity from Time Series Data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *8*(1), 2011, 152 - 165.

[20] A. Rodan, P. Tino, Minimum Complexity Echo State Network, *IEEE Transactions on Neural Networks*, *22* (1), 2011, 131 - 144.

[21] X. Cai, N. Zhang,G. K. Venayagamoorthy, and D. C. Wunsch, II, Time Series Prediction with Recurrent Neural Networks Trained by a Hybrid PSO-EA Algorithm, *Neurocomputing*, *70*(13-15), 2007, 2342-2353.

[22] X. Cai, N. Zhang, G.K. Venayagamoorthy, D.C. Wunsch II, Time Series Prediction with Recurrent Neural Networks Using Hybrid PSO-EA Algorithm, *INNS-IEEE International Joint Conference on Neural Networks (IJCNN)*, *2*, Budapest, Hungary, 2004, 1647-1652.

[23] Artificial Neural Network & Computational Intelligence Forecasting Competition in 2006, 2008, and 2010. http://www.neural-forecasting-competition.com/

[24] Lendasse, E. Oja and O. Simula, M. Verleysen, Time Series Prediction Competition: The CATS Benchmark, *INNS-IEEE International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 2004, 1615-1620.

[25] M.T. Hagan and M. Menhaj, Training Feed-Forward Networks with the Marquardt Algorithm, *IEEE Transactions on Neural Networks*, *5*(6), 1994, 989-993.