Performance Evaluation of Artificial Neural Networks for Runoff Prediction

By Amin Elshorbagy,[1] S. P. Simonovic,[2] P.E., Member, ASCE, and U. S. Panu,[3] P.E., Fellow, ASCE

Abstract: Spring runoff prediction in the Red River Valley, southern Manitoba, Canada, is an important issue because of the devastating effect of the flood of 1997 in that area. Increasing the accuracy of the prediction process is a practical necessity. This study looks at the artificial neural networks (ANN) technique and compares it to linear and nonlinear regression techniques. The advantages and disadvantages of the three modeling techniques are discussed. To fill the predictive accuracy evaluation gap left by the mean squared error and the mean relative error, a modified statistic, namely, pooled mean squared error, is developed and explained. The aim of this work is to show the applicability of ANN for runoff prediction and to evaluate their performances by comparing them with traditional techniques. In this study, according to the accuracy of results, the ANN models show superiority in most of the cases. However, in some situations, the performance of the other two techniques was comparable.

INTRODUCTION

In April 1997, a major flood occurred on the Red River in southern Manitoba, Canada. To protect lives and to minimize potential damage and loss of property, evacuation was the only solution. The 1997 Red River flood emphasizes the importance of improving flood forecasting (International Joint Commission 1997). The severity of major floods contributes significantly to increased attention toward the investigation of different forecasting techniques and hence increases their prediction accuracy.

Flood forecasting follows two major modeling approaches: conceptual (phenomenological) modeling, which retains some of the physical laws in their mathematical formulation, and black-box modeling, which relies heavily on an input-output description of the conceptual models. The large amount of data required for the conceptual models, along with the costs of collecting the data, make the black-box models more attractive to hydrologists. This study compares three different techniques of the black-box approach by looking at the annual runoff records of the Red River basin. The first technique examined is the standard statistical method of multiple linear regression. Second, a multiple nonlinear regression is applied to test its applicability and usefulness in forecasting the flow of the Red River. The Water Resources Branch of the Natural Resources Department of Manitoba employs nonlinear regression for flood forecasting of the Red River. The third is the Artificial Neural Network (ANN) technique. ANN simulates biological neural systems and the human way of thinking and learning. Comparison of the techniques is important and beneficial because of the differences among the three in their underlying solution algorithms and in their use of inputs.

The last decade has witnessed many applications of ANNs in water resources. These include rainfall forecasting (French et al. 1992), multivariate modeling of water resources time series (Raman and Sunilkumar 1995), modeling of rainfall-runoff process (Hsu et al. 1995), and flow forecasting (Karunanithi et al. 1994; Zealand et al. 1999). The intent here is to contribute to the concurrent endeavors to reach a conclusion on the usefulness and applicability of ANN in the area of runoff prediction. Furthermore, any improvement in the accuracy of flood prediction (due to adoption of the ANN technique) could help those affected by the flood. Using annual input variables, this study assesses the performance of the ANN technique compared to linear and nonlinear regression in predicting the spring runoff of the Red River.

RED RIVER FLOODS

Floods in the Red River Valley typically occur during the spring as a direct consequence of the melting of the winter snowpack. Factors that contribute to severe flooding conditions are (1) high water content in the soil at freeze-up time in the fall; (2) heavy snowpack accumulation during the winter; and (3) rapid melting, possibly in combination with spring rainfall. The large size of the Red River basin (61,000 km^2) can lead to contrasting runoff patterns within the various portions of the basin due to extreme spatial variations in meteorological conditions. In this study, annual variables are used to predict spring runoff. The basic data set includes (1) the runoff (*Ro*) which is the model output (cm); (2) total precipitation (*Tot. Pcpn*) measured from November 1 to the end of the following spring's runoff (cm); (3) antecedent precipitation index (*Api*), which is weighted precipitation from May to October of antecedent year (cm); (4) melt index (*Mi*) in degree-days/day during the active snowmelt period (°C); (5) winter temperature index (*Wti*) based on mean temperature from December to February at Grand Forks, N. Dak. (°C); and (6) previous spring runoff (*Psr*) (cm), including April rainfall runoff. A yearly record of 50 years (1948–1997) is used for training (calibrating) and testing (verifying) the adopted models in different experiments. The variables are further explained in Appendix I.

METHODOLOGY

Three adopted techniques are briefly presented as follows:

- Linear Regression Analysis. Scatter diagrams of each of the predictor variables were individually analyzed against the response variable (the runoff). It was found that 83.5% of the variance in the runoff data is represented by a univariate linear regression model that relates the runoff to total precipitation. Therefore, inclusion of more predictor variables in a multiple linear regression model is worth testing in this study. A multiple regression model that de-

[1]Res. Asst., Civ. and Geological Engrg. Dept., Univ. of Manitoba, Winnipeg, MB, Canada R3T 2N2.

[2]Prof., Civ. and Geological Engrg. Dept., Univ. of Manitoba, Winnipeg, MB, Canada R3T 2N2.

[3]Prof., Civ. Engrg. Dept., Lakehead Univ., Thunder Bay, ON, Canada P7B 5E1.

Note. Discussion open until March 1, 2001. To extend the closing date one month, a written request must be filed with the ASCE Manager of Journals. The manuscript for this technical note was submitted for review and possible publication on June 17, 1998. This technical note is part of the ***Journal of Hydrologic Engineering***, Vol. 5, No. 4, October, 2000. ©ASCE, ISSN 1084-0699/00/0004-0424–0427/$8.00 + $.50 per page. Technical Note No. 18591.

scribes the relationship between the runoff and the five independent variables is given as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \quad (1)$$

where y denotes the runoff of the Red River; $x_1, \ldots, x_5$ denote the five input variables; and ε denotes the error term. The parameter estimation technique can be found in Montgomery and Peck (1992).

- Nonlinear Regression Analysis (NRA)—Any model that is not of the form given by (1) will be called a nonlinear model. Two examples of such models used in this study are

$$y = \theta_1 \cdot x_1^{\theta_2} \cdot x_2^{\theta_3} \cdot x_3^{\theta_4} \cdot x_4^{\theta_5} \cdot x_5^{\theta_6} + \varepsilon \quad (2)$$

$$y = \theta_1 \cdot e^{\theta_2 x_1 + \theta_3 x_2 + \theta_4 x_3 + \theta_5 x_4 + \theta_6 x_5} + \varepsilon \quad (3)$$

The parameters to be estimated are denoted by β in (1) and θ in (2) and (3); x_i and ε are as defined earlier. Details of the methodology can be found in Draper and Smith (1981).

- ANN Technique—The ANN technique is a computing paradigm that may have more than one mode. The feed-forward neural networks with back-propagation learning algorithm are the most widely used neural networks (Freeman and Skapura 1991; Anderson 1995). This study employs three-layer networks. The configuration of a neural network includes determining the number of hidden layers, the number of nodes in each of the hidden layers, and the connection weights. Details on the Ann and the back-propagation training algorithm (to estimate the network parameters) can be found in (Maren et al. 1990; Freeman and Skapura 1991; Anderson 1995; Dhar and Stein 1997).

INDICATORS OF MODEL PERFORMANCE

A mean squared error (MSE) is one of the most commonly used performance measures in hydrological modeling. Many researchers used MSE or its root (RMSE) as an accuracy measure (Carpenter and Barthelemy 1994; Bastarache et al. 1997; Shamseldin 1997). Others, such as Karunanithi et al. (1994), used MSE and mean relative error (MRE) to try to fill some of the gaps left by considering only MSE and stated that the squared error and the relative error provide different types of information about model predictive capabilities. Formulas for calculating MSE and MRE are given as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{D}_i - D_i)^2 \quad (4)$$

$$\text{MRE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{(\hat{D}_i - D_i)}{D_i} \right| \quad (5)$$

where D_i stands for the measured (actual) value; $\hat{D}_i$ = estimated value; and n = number of observations. Using the MRE to rank different models according to their performance may result in a different rank from that obtained using MSE. To be more comprehensive in the evaluation of different aspects of model performance, Hsu et al. (1995) used RMSE, percent error in volume, percent error in matching the maximum flow, and the correlation between the observed and simulated flows. Those who used only the MSE to assess the performance of different models failed to recognize the appropriateness of MSE as a sole measure of model performance. Meanwhile, others who used more than one measure did not provide guidelines for choosing the model when different measures favor different models. In some cases, preference of one model to another was based on a small difference in the MSE. The MSE is an absolute value and not referenced to any benchmark point, not comprehended according to a predesigned threshold to indicate the significance of the MSE value or even the significance of the difference between two calculated MSEs of different models. It is apparent that a statistical measure capable of overcoming the above discussed difficulties is required.

POOLED MEAN SQUARED ERROR (PMSE)

Assuming a common hydrological problem of predicting streamflow using two different models (A and B), a different statistic is proposed to achieve the following objectives: (1) To combine the effect of the MSE and MRE in one measure—the PMSE; and (2) to increase the effect of the significant residuals that are above a predetermined threshold and to reduce the differential effect of the insignificant residuals that are below the predetermined threshold.

A hypothetical numerical example is provided to show the steps followed in the development of the PMSE. Table 1 shows the actual, estimated flows, residuals, and relative errors for the hypothetical example. The steps are as follows:

1. A threshold is determined, and only the residuals above its value are considered significant. A threshold can be any value, but this study considers 5% of the actual values.
2. Relative error of the outputs from the two models are pooled together to form one set of size N(N = 20 in this hypothetical example, $n_1 = n_2 = N/2$).
3. If the relative error is $\leq$5%, the predicted value is given a constant rank. In our example, predicted values with residuals $\leq$5% of the actual values are ranked 1.
4. Predicted values with relative errors higher than the threshold 5% are ranked so that the highest residual is given the rank N, the next highest is given the rank N-1, and so on.
5. The PMSE is calculated using the following relationship:

$$\text{PMSE} = \frac{\sum_{i=1}^{N/2} (\hat{D}_i - D_i)^2 \cdot k_i}{\sum_{i=1}^{N} k_i} \quad (6)$$

where k_i = rank of the residual error. The ideal value of MSE, MRE, and PMSE is 0.0. In this study's example, MSE of Model A = 25 and MRE(A) equals 0.048, whereas MSE(B) = 20.8 and MRE(B) = 0.0281. One can favor Model B over Model A, and the values in Table 2 clearly show that all predicted values of Model A can be

TABLE 1. Actual and Predicted Flows, Residuals, and Relative Errors (Hypothetical Case)

	Model A			Model B		
Actual flow (1)	Predicted flow (2)	Residual values (3)	Relative error (4)	Predicted flow (5)	Residual values (6)	Relative error (7)
100	105	5	0.05 (1)	101	1	0.01 (1)
110	115	5	0.045 (1)	109	−1	0.009 (1)
90	85	−5	0.056 (18)	100	10	0.111 (20)
100	105	5	0.05 (1)	101	1	0.01 (1)
105	110	5	0.048 (1)	104	−1	0.0095 (1)
113	108	−5	0.044 (1)	114	1	0.0088 (1)
95	100	5	0.053 (17)	96	1	0.0105 (1)
120	125	5	0.042 (1)	119	−1	0.0083 (1)
105	100	−5	0.048 (1)	115	10	0.095 (19)
117	122	5	0.043 (1)	118	1	0.0085 (1)

Note: Values in parenthesis are pooled rank.

TABLE 2. Performance of Different Models on Testing Data

Experiment (1)	ANN MSE (2)	ANN MRE (3)	LRA MSE (4)	LRA MRE (5)	LRA R^2 (6)	NRA Model 1 MSE (7)	NRA Model 1 MRE (8)	NRA Model 1 R^2 (9)	NRA Model 2 MSE (10)	NRA Model 2 MRE (11)
1	0.0808	0.4357	0.1839	0.5665	0.896	0.0495	0.473	0.92	0.070	0.675
2	0.1467	0.447	0.2939	0.5777	0.877	0.0481	0.543	0.877	0.647	0.599
3	0.0556	0.5549	0.1454	0.5090	0.911	0.1279	0.418	0.965	0.705	0.731
4	0.0777	0.3155	0.1196	0.4447	0.844	0.111	0.455	0.875	0.436	0.707
5	0.073	0.5243	0.1258	0.3709	0.78	0.0566	0.623	0.792	NA	NA
6	0.1286	0.4708	0.1895	1.1624	0.969	1.544	0.707	0.987	1.543	2.84

Note: NA = not available.

accepted because they lie within the predefined threshold. Two values out of 10 (20%) predicted by Model B are in clear violation of the threshold, which indicates low reliability of Model B. It can be concluded that neither of the two regular measures is able to interpret or explain the results of this study's example. On the other hand, calculating the PMSE results in values of 9.72 and 43.42 for Models A and B, respectively. These values are more reasonable according to the concept of satisfying or violating the threshold. It should be noted that other authors established measures of errors for different purposes. The closest example is Sorooshian et al. (1983, 1993), who used a maximum likelihood estimator function for the heteroscedastic error case. The main point of their approach weighs the squared residuals by the actual observations. In other words, they combine the squared error and the relative error.

DESIGN OF TEST EXPERIMENTS

The three techniques adopted in this study (LRA, NRA, and ANN) are evaluated using six testing experiments to compare the results in terms of robustness of accuracy, convenience, and ease of use. The difference among various test experiments is mainly in the input structure for the purpose of investigating its impact on the output accuracy and determining the most appropriate one for the case of the Red River.

- Experiment 1—Five independent variables are used as inputs with the first 10 years (1948–1957) for testing the ANN and verifying the regression models and the last 40 years (1958–1997) for training the ANN and calibrating the regression models.
- Experiment 2—Same as Experiment 1 with the extreme observation of year 1997 removed from the training data set. This means that the four models of the three techniques will not see the extreme value during training while another extreme (year 1950) is still part of the testing/verification data set.
- Experiment 3—All odd years (1949, 1951, . . . , 1997) are considered for training the models, whereas all even years (1948, 1950, . . . , 1996) are kept for testing the models. In this case the length of training and testing data sets is equal—25 years. This experimental design captures the effect of any land use change in the basin in training and testing data sets.
- Experiment 4—Two variables only, total precipitation and antecedent precipitation index, are considered as inputs to the different models. Training and testing data set sizes are the same as in Experiment 1.
- Experiment 5—Same as Experiment 4 with the exception of using only total precipitation as the input.
- Experiment 6—Ten years (1948–1957) are considered for training, and the remaining 40 years (1958–1997) are considered for testing the models. The last experiment investigates the sensitivity of the different techniques to shorter training data sets.

Although the experiments suggested here should handle a wide range of possible input alternatives, they are not exhaustive.

RESULTS AND DISCUSSION

The performance of the identified models is first evaluated based on the results of verification (testing) using MSE and MRE. The results are summarized in Table 2 and discussed below.

Table 2 shows that Model 2 of the NRA is dominated by other models in all experiments. Throughout the six experiments, it seems that for the problem under consideration the multiplicative NRA model in (2) performs better than the model in (3). Therefore, the NRA model in (3) is dropped from any further analysis or discussion. The discussion on the rank of the performance of the three models in different experiments based on MSE and MRE is condensed and depicted in Table 3. This table shows the contradictory rank of different models according to MSE and MRE in different experiments. Complexity of the ranking process can be reduced by combining the MSE and the MRE in one statistic, PMSE. Table 4 presents the values of PMSE for the identified models in each of the six experiments. It should be noted that each experimental test data set results in a unique configuration of the ANN model. Table 5 also provides the rank of the different models according to PMSE. For each experiment only one

TABLE 3. Rank of Different Models Using MSE and MRE

Experiment (1)	Rank According to MSE 1 (2)	Rank According to MSE 2 (3)	Rank According to MSE 3 (4)	Rank According to MRE 1 (5)	Rank According to MRE 2 (6)	Rank According to MRE 3 (7)
1	NRA	ANN	LRA	ANN	NRA	LRA
2	NRA	ANN	LRA	ANN	NRA	LRA
3	ANN	NRA	LRA	NRA	LRA	ANN
4	ANN	NRA	LRA	ANN	LRA	NRA
5	NRA	ANN	LRA	LRA	ANN	NRA
6	ANN	LRA	NRA	ANN	NRA	LRA

TABLE 4. Performance of Different Models According to PMSE

Experiment (1)	LRA (2)	NRA Model 1 (3)	ANN (4)	ANN configuration (5)
1	0.042	0.016	0.011	(5-1-1)
2	0.125	0.018	0.086	(5-3-1)
3	0.069	0.074	0.036	(5-5-1)
4	0.032	0.029	0.022	(2-1-1)
5	0.035	0.022	0.02	(1-1-1)
6	0.087	0.833	0.055	(5-1-1)

TABLE 5. Rank of Different Models Using PMSE

Experiment (1)	Rank According to PMSE 1 (2)	2 (3)	3 (4)
1	ANN	NRA	LRA
2	NRA	ANN	LRA
3	ANN	LRA	NRA
4	ANN	NRA	LRA
5	ANN	NRA	LRA
6	ANN	LRA	NRA

rank is provided. In five out of six experiments, the ANN models show better performance than NRA and LRA. Thus, PMSE helps achieve more conclusive results.

SUMMARY AND CONCLUSIONS

A variety of models based on the concepts and techniques of ANN, LRA, and NRA have been developed and tested on six experiments using the Red River data. The essential difference among various experimental test data sets is mainly in the model input structures. The results of most of the six experimental tests indicate that the performance of ANN-based models is better but dependent on the data input structure. During the training phase, if the data set does not encompass entire data patterns (i.e., questionable quality), then one needs to exercise caution in the use of ANN-based models. In such cases, the use of other models such as NRA may prove more fruitful. However, the ANN-based models show better prediction ability than the NRA models for cases where unsufficient amount of data is encountered during the training phase. In cases of small training data sets, even the LRA models (with their large ability to generalize) may prove suitable candidates for consideration.

The ANN-based models with only one hidden layer are easier to configure and require less effort than the other corresponding NRA-based models. The necessity of preformulation of the NRA-based models limits the practical utility of such models. ANN models are not recommended as a single approach for the runoff prediction. As a reference for comparison, use of other traditional techniques is necessary.

The inadequacy of the MSE and the MRE as measures of the level of model performance has been demonstrated. To overcome the inadequacy of MSE and MRE, a new statistical measure, PMSE, is developed. The relative efficacy of the PMSE, as a measure of model performance, is demonstrated.

APPENDIX I. DATA USED FOR ANALYSIS

Runoff (*Ro*): Spring runoff volume expressed in centimeters over the drainage area of 61,000 km^2.

Precipitation (*Tot. Pcpn*): Basin precipitation in centimeters from November 1 of the preceding year through the spring runoff period of the current year.

Antecedent precipitation index (*Api*): Computed as the sum of weighted monthly basin precipitations for May to October of the preceding year using the following weighting factors: May—0.07, June—0.08, July—0.12, August—0.18, September—0.25, and October—0.30.

Melt index (*Mi*): Expressed as the average number of degree-days per day, in degrees Celsius above a base of −2, at Grand Forks, N. Dak., during the active snowmelt period.

Winter temperature index (*Wti*): Computed as the mean of average monthly temperatures, in degrees Celsius, for December, January, and February at Grand Forks, N. Dak.

ACKNOWLEDGMENTS

This work was partially supported through a grant to U. S. Panu by NSERC. The data used in this paper were provided by A. Warkeintin of Water Resources Branch, Manitoba.

APPENDIX II. REFERENCES

Anderson, J. A. (1995). *An introduction to neural networks*, MIT Press, Cambridge, Mass.

Bastarache, D., El-Jabi, N., Turkkan, N., and Clair, T. A. (1997). ''Predicting conductivity and acidity for small streams using neural networks.'' *Can. J. Civ. Engrg.*, Ottawa, 24, 1030–1039.

Carpenter, W. C., and Barthelemy, J. F. (1994). ''Common misconceptions about neural networks as approximators.'' *J. Comp. in Civ. Engrg.*, ASCE, 8(3), 345–358.

Dhar, V., and Stein, R. (1997). *Intelligent decision support methods*, Prentice-Hall, Englewood Cliffs, N.J.

Draper, N. R., and Smith, H. (1981). *Applied regression analysis*, Wiley, New York.

Freeman, J. A., and Skapura, D. M. (1991). *Neural networks. Algorithms, applications, and programming techniques*, Addison-Wesley, Reading, Mass.

French, M. N., Krajewski, W. F., and Cuykendall, R. R. (1992). ''Rainfall forecasting in space and time using a neural network.'' *J. Hydro.*, Amsterdam, 137, 1–31.

Hsu, K. L., Gupta, H. V., and Sorooshian, S. (1995). ''Artificial neural network modeling of the rainfall-runoff process.'' *Water Resour. Res.*, 31(10), 2517–2530.

International Joint Commission (IJC). (1997). ''Red River flooding. Short-term measures.'' *Interim Rep. of the International Red River Basin Task Force*, Ottawa-Washington.

Karunanithi, N., Grenney, W. J., Whitley, D., and Bovee, K. (1994). ''Neural networks for river flow prediction.'' *J. Comp. in Civ. Engrg.*, ASCE, 8(2), 201–220.

Maren, A., Harston, C., and Pap, R. (1990). *Handbook of neural computing applications*, Academic, San Diego.

Montgomery, D. C., and Peck, E. A. (1992). *Introduction to linear regression analysis*, Wiley, New York.

Raman, H., and Sunilkumar, N. (1995). ''Multivariate modeling of water resources time series using artificial neural networks.'' *Hydrological Sci. J.*, 40(2), 145–163.

Shamseldin, A. Y. (1997). ''Application of a neural network technique to rainfall-runoff modeling.'' *J. Hydro.*, Amsterdam, 199, 272–294.

Sorooshian, S., Duan, Q., and Gupta, V. K. (1993). ''Calibration of rainfall-runoff models: Application of global optimization to the Sacramento soil moisture accounting model.'' *Water Resour. Res.*, 29(4), 1185–1194.

Sorooshian, S., Gupta, V. K., and Fulton, J. L. (1983). ''Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility.'' *Water Resour. Res.*, 19(1), 252–259.

Zealand, C. M., Burn, D. H., and Simonovic, S. P. (1999). ''Short term streamflow forecasting using artificial neural networks.'' *J. Hydro.*, Amsterdam, 214, 32–48.