AN ALGORITHM FOR VARIABLE SELECTION WITH AND APPLICATION TO METEOROLOGICAL TELECONNECTIONS

P1.12

Nazario D. Ramirez-Beltran*
University of Puerto Rico, Mayaguez, PR

William K.-M. Lau
NASA Goddard Space Flight Center, Greenbelt, MD

Amos Winter
University of Puerto Rico, Mayaguez, PR

Anthony Veneros
University of Puerto Rico, Mayaguez, PR

1. INTRODUCTION.

An algorithm is proposed to identify the best correlation between the variables of two data sets. The relationship among the variables could be linear or nonlinear and it can have time delays among the variables. This algorithm consists of: i) performing mathematical transformations to the original variables; ii) organize the information into small groups; iii) select the best predictors for each group; and iv) perform random search to select the winner variables.

Forty rainfall stations located in Puerto Rico (PR) with 43 years of monthly observations were used to implement and assess the proposed algorithm. Meteorological indexes based on sea level pressure, sea surface temperatures, and rainfall were used to identify teleconnections between PR rainfall processes and global meteorological indexes. It was found that the major factors that drive the PR rainfall process are the Artic Oscillation Index, Brazil Rainfall Index, North Atlantic Oscillation Index, and Sahel Rainfall Index. The winner variables from the selection algorithm were also used to design and train an artificial neural network model to predict PR rainfall process. Thirty-eight years were used for training and five years for model assessing. Results suggest that the proposed methodology is a potential tool to predict the monthly precipitation at any station, assuming that at least 40 years of monthly rainfall observations are available.

A nonlinear dynamic system was simulated and modeled using an artificial neural network model. A feedforward neural network model with the Levenberg-Marquardt algorithm was successfully implemented. The structure of the neural networks includes two layers with linear and nonlinear transfer functions. Simulation results show that a neural network model can properly represent a nonlinear dynamic system. The cross validation technique was used to select the transfer functions and the required number of the neurons in the hidden layer.

Artificial neural networks (ANN) methodology is an emerging strategy to model the multivariate inputs and outputs of an atmospheric dynamic system. ANN is especially useful for modeling nonlinear climate dynamics, since the ANN algorithm has transfer functions to model nonlinear relationships. Several researchers have reported successful applications of the neural networks methodology to atmospheric sciences. Snell (et al. 2000) pointed out that many climate studies require generating estimates of climate variable at a given location based on values from other locations. They suggest a methodology based on ANN to estimate temperatures for some locations given information from a lattice of surrounding locations. Aviolat (et al. 1998) apply an ANN to describe the creation of clouds at different layers. Two-hour period of observations was used to develop the input patterns to train an ANN. Results show evidences that the prygeometer with an ANN provide a very accurate estimate of the cloud amount at the main cloud layer.

Grecu and Krajewski (2000) proposed an efficient methodology for detecting anomalous propagation echoes in radar data. The method is based on volume scan reflectivity observations and application of neural networks for classification of the base scan radar echo into the anomalous propagation echoes or rain classes. They pointed out that the neural networks approach presents a conceptual simple yet rigorous way to address the problem of anomalous propagation echoes detection**.** Tangang (et al. 1998) applied neural networks methodology to forecast the sea surface anomaly on three regions: el Niño 4, el Niño 3.5, and el Niño 3. The inputs of the neural networks were the extended empirical orthogonal functions of the sea level pressure field that cover the Tropical Indian and Pacific Ocean.

ANN has extensively been applied to model multiple inputs and outputs of linear and nonlinear systems. Successful applications have been reported in the literature (Ramirez-Beltran, 2000a, 2000b, and 2002). The organization of this paper is as follows. Description of the utilized data is presented in section 2. Simulation of a nonlinear dynamic system to assess the capability of the artificial neural network is presented in section 3. The implemented methodology is described in section 4, and conclusions are given in section 5.

**Corresponding authors address:* Nazario D. Ramirez, University of Puerto Rico, Dpto. Industrial Engineering, Mayaguez, PR, 00680, e-mail: nazario@ece.uprm.edu

2. DATA.

The proposed algorithm was implemented to model 40 rainfall stations located in Puerto Rico (see Figure 1). The available information covers 43 years from1960 to 2002 and was provided by the Joint Institute for Caribbean Climate Studies of the University of Puerto Rico, the National Diagnostic Center located in Boulder Colorado, and the National Weather Services located in San Juan PR. It should be noted that most of the meteorological indexes were obtained by the Internet at the following address http:// tao.atmos. Washington.edu.

The rainfall records were correlated to well known sea surface temperature (SST), sea level pressure (SLP), rainfall index, and others global meteorological indexes. The studied variables are described in Table 1.

Table 1. Meteorological Indexes

Number	*Code*	*Description*
1	NA	SST in the North Atlantic (5-20°N, 60-30°W)
2	SA	SST in the South Atlantic (0-20°S, 30°W-10°E)
3	TE	SST in Tropical Equatorial (10°S-10°N, 0-360°)
4	N12	SST in the equatorial Pacific: El Niño 1-2 (0-10°S, 90-80°W)
5	N3	El Niño 3 (5°N-5°S, 150-90°W)
6	N4	El Niño 4, (5°N-5°S, 160°E-150°W)
7	N34	El Niño 3-4 (5°N-5°S, 170-120°W)
8	NAO	North Atlantic Oscillation index
9	AO	Artic Oscillation Index
10	SAH	Sahel Rainfall Index (20-8N, 20W-10E)
11	BRA	North Brazil Rainfall Index
12	CT	Could Tong
13	SOI	Southern Oscillation Index
14	SR	Solar Radiation
15	SS	Sun Spots

3. SIMULATION

Simulation technique was used to assess the capabilities of the artificial neural network to model a nonlinear dynamic system.

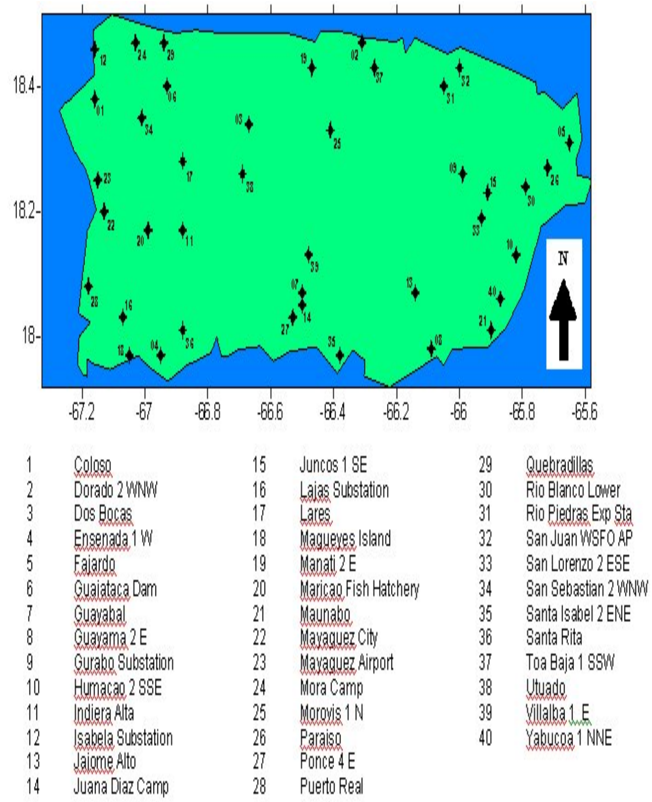

Figure 1. Location of Puerto Rico rainfall stations.

A nonlinear dynamic system can be represented by a nonlinear deferential equation, which can be approximated by a nonlinear difference equation. The difference equation is essentially a nonlinear transfer function since express the relationship between the inputs and output of a nonlinear dynamic system. The Monte Carlo simulation technique was used to mimic the behavior of a nonlinear dynamic system and ANN is used to model the relationships between inputs and outputs of the simulated system. The input variables of the system were simulated using the following expressions:

$$x_{1,t} = 200 + S + \varepsilon_{1,t} \quad (1)$$

$$x_{2,t} = 18 - 14.52 z_t + 8.70 z_t^2 - 1.02 z_t^3 + + \varepsilon_{2,t} \quad (2)$$

$$x_{3,t} = 5.72 + 1.25 w_t - 1.35 w_t^2 + 0.21 w_t^3 + \varepsilon_{3,t} \quad (3)$$

where $z_t = \ln(r_t)$ and $w_t = \ln(s_t)$; r_t is a sequence of integer numbers from *1* to *N*, s_t is also a sequence of integer numbers from *N* to *1,* where *N* is the total number of data to be generated, $\varepsilon_{i,t}$ is a random number for the i^{th} variable that follows a normal distribution with mean zero and a constant variance. The coefficients and the variances were selected such that the values resemble the behavior of a chemical process. The selected variances for $\varepsilon_{i,t}$ were 1.2, 0.24, and 0.02, respectively. $x_{1,t}$ represents the flow of amyl acetate at time *t*, $x_{2,t}$ represents the water flow at time *t*, and $x_{3,t}$ the flow of acetic acid at time *t.*

The response of the dynamic system is the pH of enriched salt at time *t* and it was assumed that it could be expressed by the following nonlinear difference equation:

$$y_t = \phi y_{t-1} + \theta \frac{\left(1 + x_{2,t-d_2-1} x_{3,t-d_3-2}^2\right) x_{1,t-d_1-1}^{0.5}}{\left(1 + x_{2,t-d_2-2}^{0.5}\right) x_{3,t-d_1-1}^{0.5}} + \varepsilon \qquad 4)$$

where ϕ and θ are designed parameters that modulate the appropriate range of the output, and d_i is the time delay applied to each variable.

A feedforward neural network with two layers was selected to model the dynamic system. The training patterns at time *t* are represented by the $\mathbf{P}_t$ input matrix and the $\mathbf{T}_t$ target vector. The input matrix contains the values of the *x*'s and the elements of the target vector are the pH values. Thus, the training patterns can be written as follows:

$$\mathbf{P}_t = \begin{bmatrix} x_{1,1} & \cdots & x_{1,t-1-d_1} & x_{1,t-d_1} \\ x_{2,1} & \cdots & x_{2,t-1-d_2} & x_{2,t-d_2} \\ x_{3,1} & \cdots & x_{3,t-1-d_3} & x_{3,t-d_3} \end{bmatrix} \qquad (5)$$

$$\mathbf{T}_t = \begin{bmatrix} y_1 & \cdots & y_{t-1} & y_t \end{bmatrix} \qquad (6)$$

The pH of the enriched salt was simulated during 150 hours. The first 130 values were used to train the ANN model and the last 20 hours were used to validate the model. The training patterns were used to identify the best transfer functions and the optimal number of neurons in the hidden layer. The training patterns suggest the following structure: the logsig and linear functions should be used in the first and second layers respectively. The optimal numbers of neurons were: two and one in the first and second layers, respectively. The first 130 values were used to fit the neural network model and the level of fitting was R^2 =.90. Thus, the ANN model explains 90% of the variability of the pH and the average absolute prediction error was 0.43. These measurements indicate that the ANN model successfully represents the underlying dynamic system. Figure 2 shows the output of the simulated pH and the predicted values from the neural network model. This figure confirms the prediction capability of the ANN model.

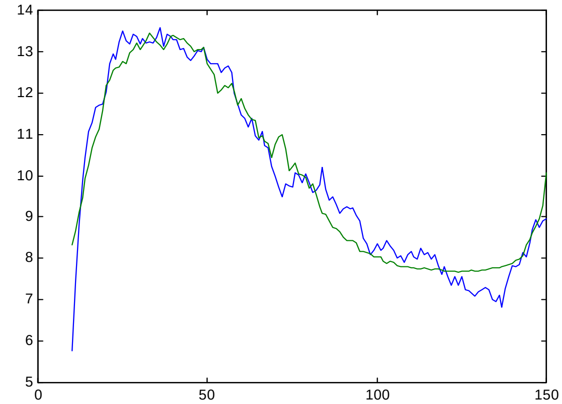

Figure 2. Simulation of a nonlinear dynamic system.

4. METHODOLOGY

The proposed methodology includes two major steps: (1) Regression and variable selection algorithm were used to identify the meteorological variables that are highly correlated with rainfall processes. (2) An ANN model was designed and trained to predict the monthly rainfall processes.

4.1. *Identifying the best predictors*.

An ANN algorithm will provide predictions with reasonable skillfulness if the training patterns contain the appropriate information; otherwise, the ANN will generate misleading predictions. Thus, an algorithm is introduced here to identify the best predictors given that the number of variables is greater than the number of observations.

The algorithm consists of six major steps: (1) implement time lags and mathematical transformations to the original variables; (2) organize the information into small groups, (3) select the best predictors for each group; (4) reorganized the best predictors; (5) repeat three times steps 1 to 4, and (6) select the winner variables.

(1) *Mathematical transformation.* A mathematical transformation is performed to be able to identify linear and nonlinear relationships between the inputs and outputs of the given system. Lag transformation is also implemented to each variable. The lead-time corresponds to the smallest lagged variable, and consequently the prediction will be constructed based on the passed values.

(2) *Small groups*. The transformed variables will be organized into groups with small amount of variables. The number of variables in a group should be 25% or less than the number of observations contained in the response variable. This rule was implemented to avoid bias on regression estimators, which occurs

when the number of variables exceeds the number of observations. The members of each group will be randomly selected. This randomly selection will produce a robust variable identification and random search.

(3) *Best predictors in each group*. The stepwise algorithm is used to select the predictors that are highly correlated with the dependent variable in each group. A diagnostic test of multicollinearity was implemented to avoid the presence of this problem and to ensure robust prediction (Montgomery, et al 2001).

(4) *Reorganize the best predictors*. The winner variables for each group will be included into a new data set and the elements of the new set will be organized into small groups. Steps 2 and 3 will be repeated until the number of variables is equal or less than the designed group size. At the end of this step the winner variables are saved.

(5) *Repeat random selection*. Repeat three times steps 1 throughout 4, and collect the winner variables at each iteration.

(6) *Select winner variables*. Prediction errors will be computed at each group and the final group will be the one that provides the smaller prediction error without multicollinearity problem.

The described algorithm was used to identify which meteorological indexes explain best the PR rainfall processes. Forty rainfall stations with 43 years of monthly precipitation data were used to integrate the available information. To minimize calculations only lags of order one and two were implemented. The number of observations for each variable is 43 and the number of variables varies depending on the month to be predicted. For instance, fifteen meteorological indexes were used to predict January i.e., (24*15) 360 predictors were created, for February were (24*15+15) 375, and (24*15+15*15) 525 for December. If a mathematical transformation is implemented then the number of variables will be duplicated. For instance if January will be predicted and the implemented transformation is natural log, then the number of variables will be 720. The response variable and predictors were standardized to obtain regression coefficients that express the proportional contribution of each variable under the same scale. Thus, the sum of the regression coefficients reveals the importance of each winner variable for every month and for every rainfall station. Thus based on the studied information, the most important variables to predict the rainfall process in PR are exhibited in Figure 3. It can be shown that the best predictors to predict one month ahead the rainfall in Puerto Rico are: Artic Oscillation Index, Brazil Rainfall Index, North Atlantic Oscillation, and Sahel Rainfall Index.

In PR there are two rainy seasons characterized by having excesses of rainfall, the first occurs on May and the second one on September. Thus, it is expected that the mechanisms that generates these two rainfall seasons are different. Figure 4 shows the predictors that best explain the rainfall occurrences in May and they are: Solar Radiation, Sunspots, Artic Oscillation Index, North Atlantic Oscillation Index, and Sahel Rainfall Index. The horizontal axis shows the codes of the meteorological indexes, which were defined in Table 1, and the vertical axis reveals the proportion of importance of the regression coefficients. Figure 5 shows the identified lags to predict precipitation on May in each variable. The identified lags are given in the horizontal axis while the regression coefficients are given on the vertical axis. For instance, to predict PR precipitation on May for a given year it is required to use the predictor values for February of the current year, and also predictor values for May of the previous two years. Thus, the selection algorithm not only identifies the variables but also the optimal lags involved in the process.

Figure 6 shows the predictors that best explain the rainfall occurrences on September and they are: Sahel Rainfall Index, North Atlantic Oscillation Index, Brazil Rainfall Index, and Artic Oscillation Index. The horizontal axis shows the codes of the meteorological indexes, and the vertical axis reveals the proportion of importance of the regression coefficients. Figure 7 shows the identified lags to predict precipitation on September. The identified lags are given in the horizontal axis while the regression coefficients are given on the vertical axis. For instance to predict PR rainfall on September of a given year it is required to use the predictor values for January and April of the current year and previous years, as well as the predictor values for June of the previous two years.

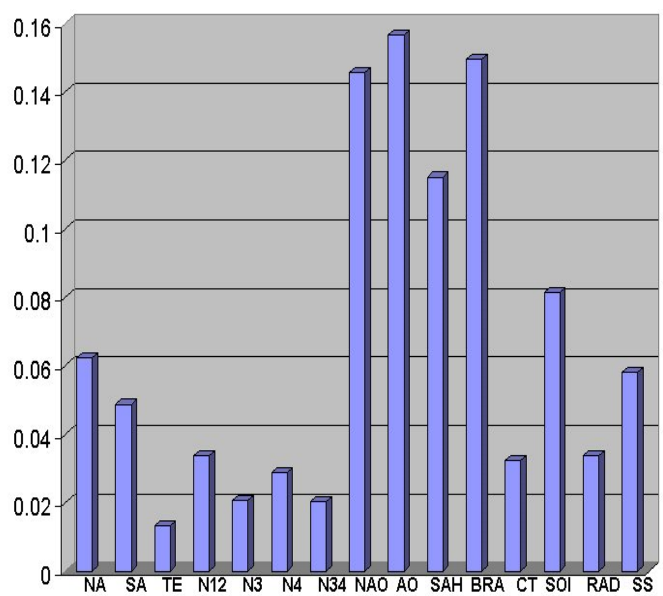

Figure 3. The best predictors that explain the monthly rainfall variability in PR (40 stations, 1960 – 2002).

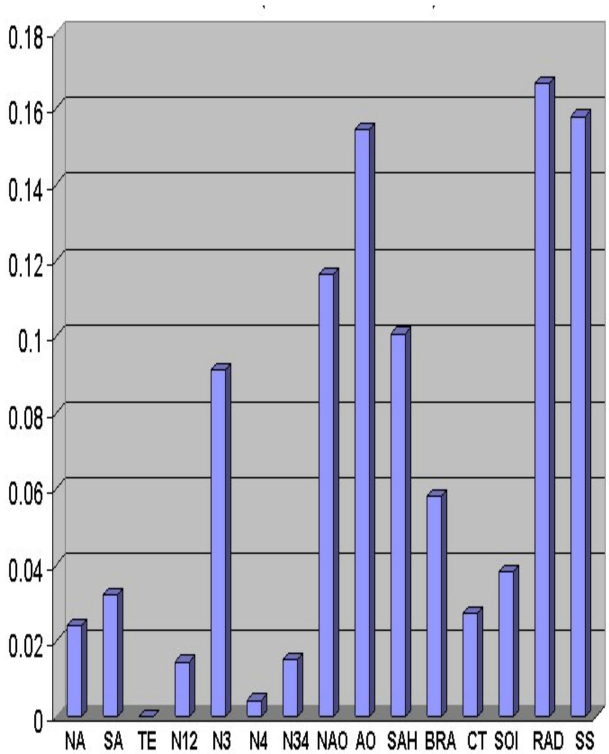

Figure 4. The best predictors that explain the rainfall in May.

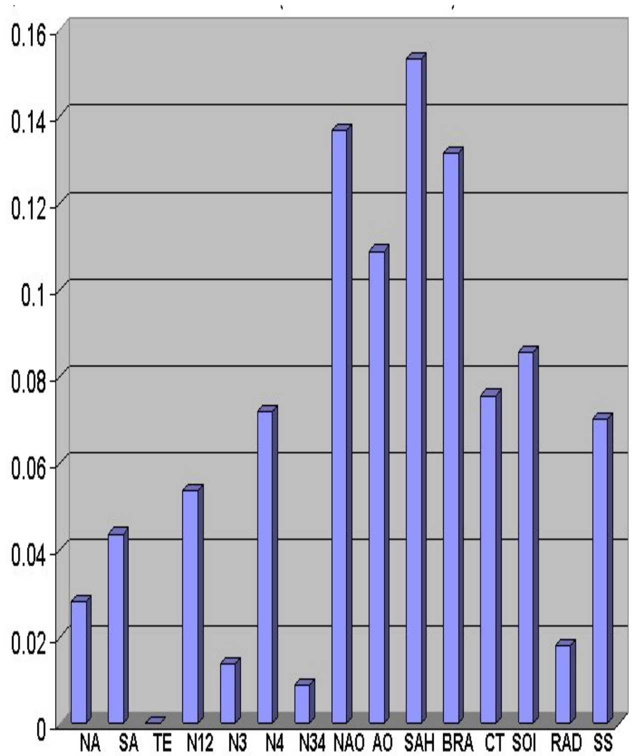

Figure 6. The best predictors that explain the rainfall on September.

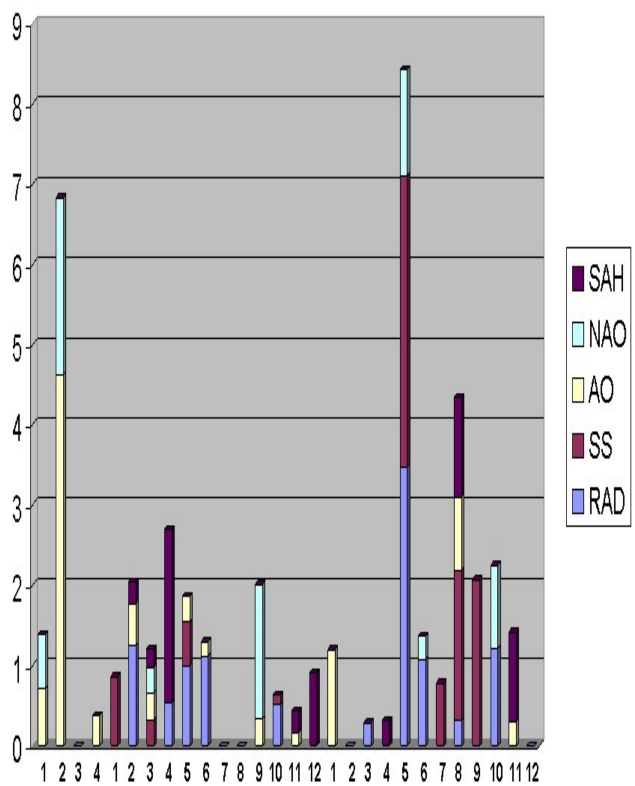

Figure 5. Optimal lags to predict PR rainfall on May

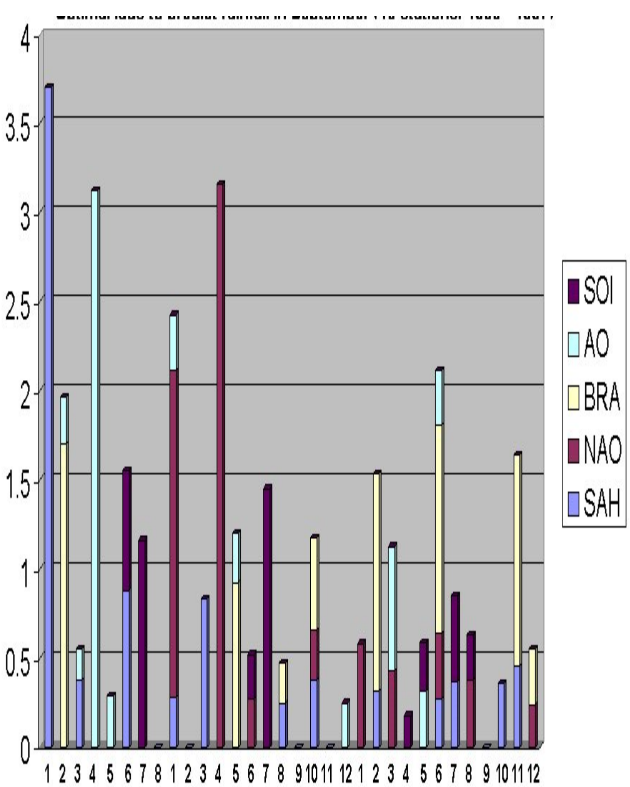

Figure 7. Optimal lags to predict PR rainfall on Sept.

4.2. *ANN model to predict rainfall process.*

An ANN model was designed and trained with the best predictors identified by applying the selection algorithm described in the previous section. It was noted that on the average 50 % of the rainfall variability was explained by using only the fifteen meteorological indexes. Thus, with the purpose of increasing the prediction capabilities of the neural networks the following predictors were added to the meteorological indexes: lagged variables of the considered rainfall process, the first three principal components and its lagged values of the following variables: rainfall variables, maximum air temperature, minimum air temperature from the underlying 40 stations. In summary the total number of predictors varies depending of the month to be predicted, the obtained predictors are as follows: 600 for January, 624 for February, and 831 for December. The explained variability increased from 50% to about 65% after adding the described variables.

The LM algorithm was used as the learning rule to speed up the convergence process. The LM algorithm has the advantage of computing automatically the learning rate, depending on the roughness of the optimization surface. Thus, for instance the learning rate is small when the slope of the optimization surface is changing very fast. However, the learning rate is large when the optimization surface is very smooth. The constants that control the magnitude of the learning rate is called the Marquardt constant and this constant is selected in such a way that Jacobian matrix is always positive definite and consequently the algorithm ensure convergence. The major disadvantage is that it requires large memory to update the weights and the biases of the ANN. The conventional backpropagation algorithm updates the weights by using only the current information, while the LM requires using all the training sets at the same time (Hagan et al. 1996; Hagan and Menhaj, 1994).

The proposed variable selection algorithm was implemented with a group size of 10 and the best predictors usually fluctuated between 4 to 8 variables. The training patterns can be written in the following form:

$$\mathbf{P}_t = \begin{bmatrix} z_{1,1} & \cdots & z_{1,t-1-d_1} & z_{2,t-d_1} \\ z_{2,1} & \cdots & z_{2,t-1-d_2} & z_{2,t-d_2} \\ \vdots & \vdots & \vdots & \vdots \\ z_{k,1} & \cdots & z_{k,t-1-d_k} & z_{k,t-d_k} \end{bmatrix} \qquad (7)$$

$$\mathbf{T}_t = \begin{bmatrix} w_1 & \cdots & w_{t-1} & w_t \end{bmatrix} \qquad (8)$$

where $z_{i,t-d_i}$ is the i^{th} predictor at time *t* with delay d_i, w_t is the rainfall observed at time *t* for a given station.

An ANN model was designed for every month and for every station. Thus, to predict a single year requires developing 480 different models and every model requires applying variable selection, training and performing prediction. Thirty-eight years (1960 – 1998) of monthly data were used to train the ANN and the last five years (1998 – 2002) were used to validate the prediction algorithm. Figure 8 shows the model fitting performances on station 1 during 38 years for the month of September, this figure shows the observed and fitted values. The number of the year is given in the horizontal axis while the amount of precipitation in inches is exhibited in the vertical axis.

The cross validation results are shown in Figures 9, 10 and 11. Figure 9 shows the comparison of 40 stations between the observed and predicted values for year 1998. The blue bar is the observed value and the red bar is the predicted value. The horizontal axis of this figure is the number of station and the vertical axis is the amount of rainfall in inches. This particular year and month was selected with the purpose of evaluating the model prediction capabilities during the presence of the hurricane Gorges, which made a landfall in PR on September 22, 1998. The accumulated rainfall in the 40 stations for the year 1998 was 591.63 inches and the predicted value was 527.15 inches (see Figure 10). It should be noted that the average, which was based on 43 years of accumulates rainfall for the month of September, is 338.80 inches. Therefore, the neural networks model was able to predict the excess of rainfall generated by the hurricane Gorges. Figure 10 shows the observed and predicted accumulated rainfall in the 40 stations. The blue bar is the observed value and the red bar is the predicted value. The horizontal axis shows the years and the vertical axis the accumulated rainfall in the 40 stations. This figure shows that the observed and predicted values are almost in agreement; indicating that the ANN model is able to predict rainfall either in dray or in wet years. Figure 11 shows the contours of observed and predicted rainfall for September 1998. These contours reveal a good agreement between the observed and predicted values. This figure shows that heavy rainfall was recorded along the hurricane track. The largest amount of rainfall recorded during that month was 36 inches on Maricao station (station 20).

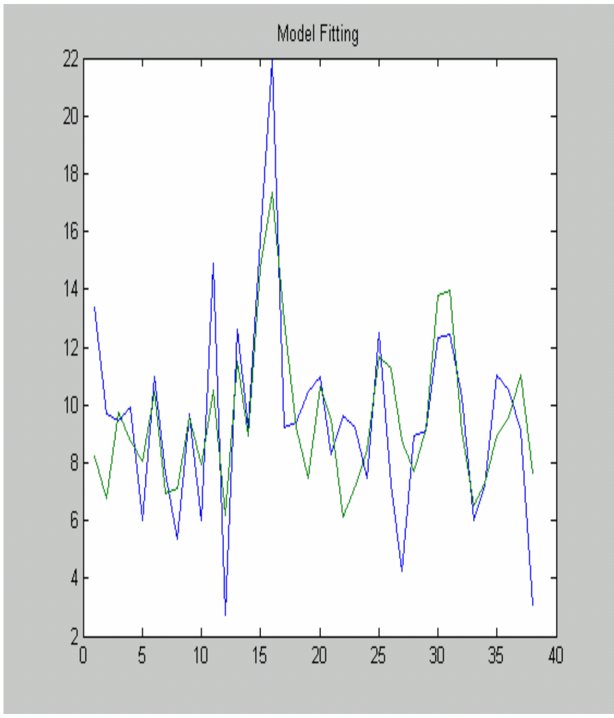

Figure 8. Model fitting performances station 1 (September, 1960 – 1997).

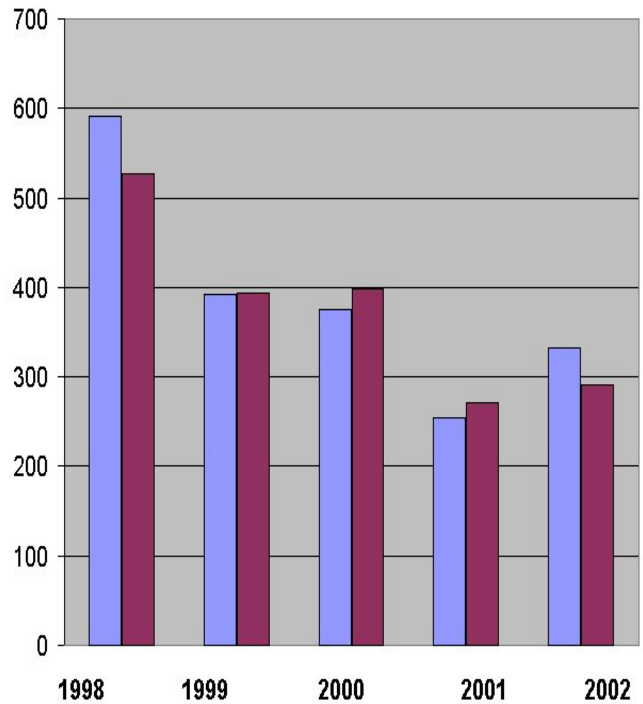

Figure 10. Accumulated rainfall in September including 40 stations

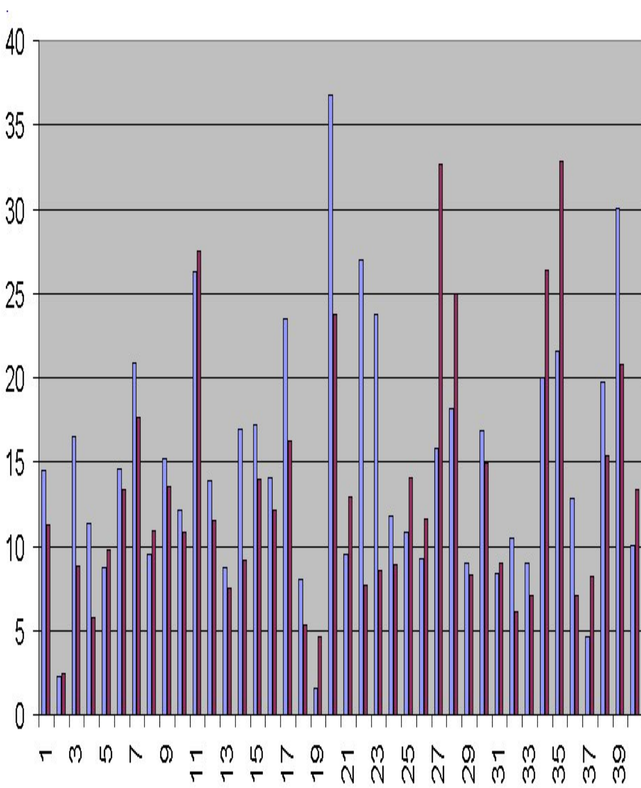

Figure 9. Observed versus predicted values September 1998

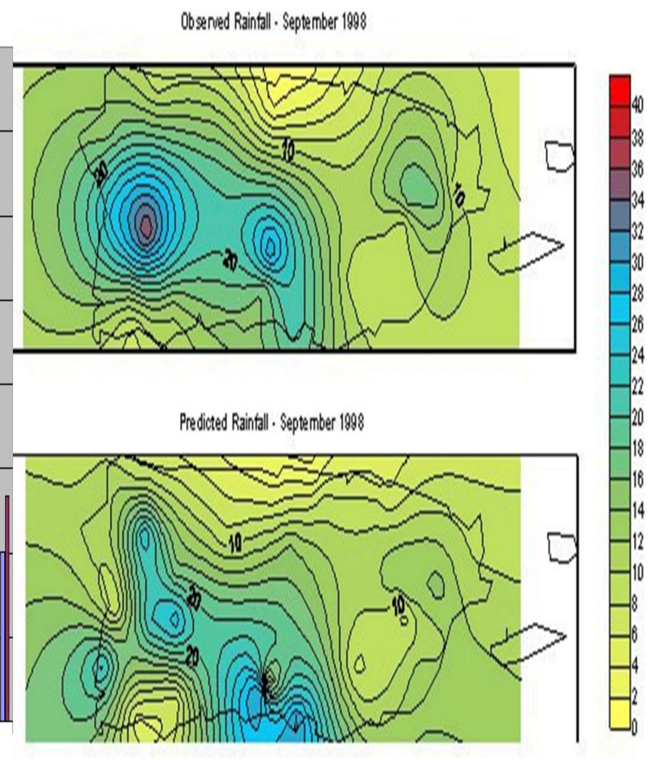

Figure 11. The observed and predicted contours are given on the top and bottom plots, respectively.

5. CONCLUSIONS.

An algorithm to perform variable selection is proposed. The algorithm has the capability of identifying the predictors that explain a response variable. The algorithm has the capability of determining the linear and nonlinear relationship as well as the time lags among the inputs and outputs of a dynamic system. The selection algorithm uses the stepwise method with a random search to select the best predictors. The limitation of the proposed algorithm is the required computational time. Expensive calculation occurs (in a PC Pentium IV) when the number of predictors is more than 1,500. For moderate amount of variables the calculation is obtained in a reasonable computational time.

An ANN model was designed and trained to predict the PR monthly rainfall processes. The proposed prediction scheme uses the outputs from the variable selection algorithm to increase prediction capabilities, i.e., the best predictors are used to train the ANN model. The LM algorithm was used as the learning rule to ensure convergence. The cross validation technique is used to identify the best transfer functions and the number of neurons in the hidden layer. The neural networks model was able to predict the excess of rainfall that occurred on September 1998. The excess of rainfall was caused by the landfall of the hurricane Gorges. In summary the proposed prediction scheme is a potential tool to predict the monthly rainfall process in almost any rainfall station as long as rainfall records is available for at least during 40 years.

Meteorological indexes based on sea level pressure, sea surface temperature, and rainfall were used to identify teleconnections between PR rainfall processes and global meteorological indexes. It has been shown that meteorological indexes that best explain the variability of the rainfall in PR are: Artic Oscillation Index, Brazil Rainfall Index, North Atlantic Oscillation, and Sahel Rainfall Index. PR has two rainy seasons characterized by having two rainfall peaks, one occurs on May and the second one on September. It has been shown that the variables that drive the rainfall patterns in May are: Solar Radiation, Sunspots, Artic Oscillation Index, North Atlantic Oscillation Index, and Sahel Rainfall Index. The variables that best explain the rainfall variability in September are: Sahel Rainfall Index, North Atlantic Oscillation Index, Brazil Rainfall Index, and Artic Oscillation Index.

6. ACKNOWLEDDGEMENTS

This research has been supported by NASA-EPSCoR grant NCC5-595 and also by the University of Puerto Rico. Authors want to recognize the technical support provided by the National Diagnostic Center located in Boulder Colorado and the National Weather Services located in San Juan PR., especially to Dr. Robert Webb and Israel Matos for their technical assistances.

7. REFERENCES.

Aviolat, F., T. Cornu, and D. Cattani, 1998: Automatic Clouds Observation Improved by an Artificial Neural Network., *Journal of Atmospheric and Oceanic Technology*, Vol. 15, 114-126.

Grecu, M., and W. Krajewski, 2000: An Efficient Methodology for Detection of Anomalous Propagation Echoes in Radar Reflectivity Data Using Neural Networks. *Journal of Atmospheric and Ocean Technology*, Vol. 17, 121-129.

Hagan, M.T., H.B. Demuth, and M. Beale, 1996: *Neural Network Design*, PWS Publishing Co.

Hagan, M.T., and M.B., Menhaj 1994: Training Feedforward Networks with the Marquardt Algorithm, *IEEE Transactions on Neural Networks*, Vol. 5, No. 6, 989-993.

Montgomery, D.C., E.A. Peck, and G.G. Vining, 2001: *Introduction to Linear Regression Analysis*, John Wiley & Sons, Inc., New York.

Ramirez-Beltran, N.D. and H. Rodriguez, 2000a: A Neural Network Approach to Estimate the Chemical Activity Coefficients. *Proceedings of Group Technology and Cellular Manufacturing*, World Symposium 349-354.

Ramirez-Beltran, N.D., and L. Olivares, 2000b: Neural Networks and Accelerated Tests to Predict Shelf Life of Drug Products. *Proceedings of Group Technology and Cellular Manufacturing*, World Symposium 355-360.

Ramirez-Beltran, N.D., and J. Montes, 2002: Neural Networks to Model Dynamic Systems with Time Delay. *IIE Transition Operations Engineering,* Vol. 34, 313-327.

Snell, S.E., Gopal, S. and R.K Kaufmann, 2000: Spatial Interpolation of Surface Air Temperature Using Atificial Neural Networks: Evaluating their use for Downscaling GCMs. *Journal of Climate,* Vol. 13, 886-895.

Tangang, F.T., B. Tang, A. Monahan, and W.W. Hsieh, 1998: Forecasting ENSO Events: A Neural Network-Extended EOF Approach., *Journal of Climate*. Vol. 11, 29-41.