Anais 7º Simpósio de Geotecnologias no Pantanal, Jardim, MS, 20 a 24 de outubro 2018
Embrapa Informática Agropecuária/INPE, p. 31-39

# Genetic algorithm driven ANN model for river level in small water basins

Marcia Ferreira Cristaldo [1]
Leandro de Jesus [1]
Hevelyne Henn da Gama Viganó [1]
Celso Correia de Souza [2]
Carlos Roberto Padovani [3]
Paulo Tarso Sanches de Oliveira [4]

[1] Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso do Sul - IFMS
Rua José Tadao Arima, 222 - Ycaraí
79200-000 - Aquidauana - MS, Brasil
{marcia.cristaldo, leandro.jesus, hevelyne.vigano}@ifms.edu.br

[2] Universidade Anhanguera UNIDERP
Rua Alexandre Herculano, 1400 - Taquaral
79035-470 – Campo Grande - MS, Brasil
csouza939@gmail.com

[3] Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA
Rua 21 de Setembro, 1880 – Aeroporto
79320-900 – Campo Grande - MS, Brasil
carlos.padovani@gmail.com

[4] Universidade Federal de Mato Grosso do Sul - UFMS
Rua UFMS, 40 – Caixa Postal 549
79070-900 – Campo Grande - MS, Brasil
paulotarsoms@gmail.com

***Resumo***. A previsão de enchentes é um dos principais desafios na segurança hídrica, principalmente em regiões como o Brasil, que tem escassez de dados hidrometeorológicos observados. Uma alternativa, ainda pouco explorada no Brasil tem sido o uso de aprendizado de máquina. Neste estudo, estimamos os períodos de cheia para uma bacia do Pantanal usando dados de cota e precipitação acumulada, sendo três estações fluviométricas (cota), dados pluviométricos diários de quatro pluviômetros entre 1995 e 2014. Encontramos valores de coeficiente de Nash-Sutcliffe superiores a 0,7 para uma previsão de 5 dias e um erro menor que 1 metro.

**Palavras-chave**: Previsão, Monitoramento de Cota, Perceptron Multicamada, Rio Aquidauana.

***Abstract***. Flood prediction is one of the main challenges in water security, especially in regions such as Brazil, which has a shortage of observed hydrometeorological data. An alternative, still little explored in Brazil has been the use of machine learning. In this study, we estimated the flood periods for a Pantanal basin using cumulative rainfall and precipitation data, three fluviometric stations (quota), daily rainfall data of four pluviometers between 1995 and 2014. We found values of Nash-Sutcliffe coefficient higher than 0.7 for a 5-day forecast and an error less than 1 meter.

**Keywords**: Prediction, Quota Monitoring, Multilayer Perceptron, Aquidauana River.

## 1. Introdução

The most important issue when dealing with water crisis is the management of these valuable resources. Therefore, due to the limited water resources, proper and optimized management is the most important task of policy-makers and engineers in this field. One of the most important elements of water resources management is predicting the volume of these resources, especially predicting the flow of the rivers. In our country, most of the rivers in different geographical areas are seasonal, and many of them have experienced floods. This, as well as scarcity of water resources and the necessity of harnessing surface waters indicate the importance of identification and simulation of river behavior for long-term planning and utilization of river flows potential.

Due to the multiplicity of hydrological factors that govern rainfall phenomena in watershed basins, the reaction of most of the basins against precipitation is complex Besaw *et al*., 2010). Among the methods that have attracted attention of many researchers in recent years in analysis of complex and nonlinear phenomena, is the use of artificial neural networks.

Several models of river level forecasting, ranging from empirical to conceptual models, have been developed and applied (Jain et al., 2013). Though deficient, the conceptual models seek to portray the physical processes. Traditional empirical models such as Multiple Linear Regression (MLR), and Autoregressive Integrated Moving Average (Arima); and nonlinear models such as artificial neural networks (ANN) have been widely used because of a shorter development time and accurate prediction in real time (Tucci and Collishonn, 2012).

The hydrologists have to use modelling tool to generate the river (water) level with the help of observed historical data to make more alternative designs for comparison or optimization rather than to make decisions based on short historical data. Thus, the problem of forecasting for quota utilizing of rainfall has been a very active area of research throughout the evolution of the subject of hydrology. As combination of many hydroclimatic factors affect quota, it becomes an extremely complex physical process. If this catchment runoff is to be determined then the methods of determining quota are either data driven or knowledge driven as explained (Rtr- Correlation coefficient for training data); (Rtst- Correlation coefficient for test data); (MAE-Mean absolute error); (ME - Mean Error); (RAE - Relative absolute error); (MARE - Mean absolute relative error); (MRE - Mean relative error); (MSRE - Mean squared relative error); (CE -Nash-Sutcliffe Coefficient of efficiency).

Some recent studies have made attempts to show that the ANNs are not purely black-box models, and it is possible to shed some light on the hydrological processes inherent in an ANN if its architectural features are explored further (Jihoon e Vasant, 2014;Sexton et al., 2012). (Jain and Indurthy, 2013) investigated the suitability of some deterministic and statistical techniques along with the artificial neural networks (ANN) technique to model an event-based rainfall-quota process. It was found that ANN models consistently outperformed conventional models, barring a few exceptions, and provided a better representation of an event-based rainfall-quota process in general, and better prediction of peak discharge and time to peak discharge, in particular.

In this context, the Genetic Algorithm (GA) can be employed to improve the performance of ANN in different ways. GA is a stochastic general search method, capable of effectively exploring large search spaces, which has been used with Back propagation (BPN) for determining the number of hidden nodes and hidden layers, select relevant feature subsets, the learning rate, the momentum, and initialize and optimize the network connection weights of BPN by Jihoon and Vasant (2014) and Sexton et al. (2012). The hybrid GA-ANN has been used in the diverse applications. GA has been used to search for optimal hidden-layer architectures, connectivity, and training parameters (learning rate and momentum parameters) for Ann by Heckerling *et al*. (2004).

## 2. Objective

The study was optimize of neural network by using Genetic Algorithm, to improve flood forecasting systems in Mato Grosso do Sul by applying ANN models which offer more advantages than the conventional regression models. Additionally, the results of this study can be applied to similar basins and further researches.

## 3. Materials and methods

The Aquidauana basin, located in the Upper Paraguay River Basin, was chosen for study for its size and quantity of data available. The basin is 20,124 km$^2$ long and the River Aquidauana, 620 km long and the River Miranda´s main tributary, is one of the major streams of the Paraguay River Basin (Fernandes *et al.*, 2012).

In Figure 1 the Aquidauana basin is shown with its rainfall and fluviatile stations monitored by the National Water Agency (ANA) (ANA, 2016).

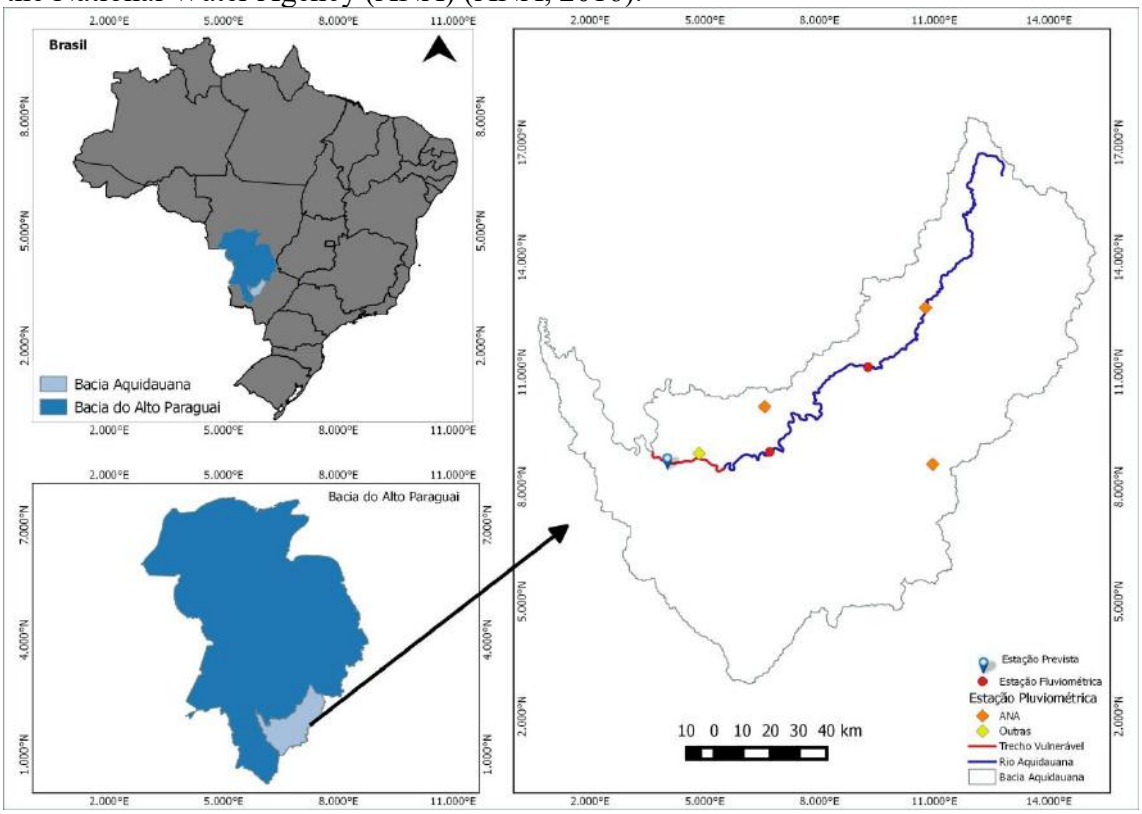



**Figura 1**. Map of the Aquidauana basin and its rainfall and fluviatile stations.

The Aquidauana basin has a well-defined climate with dry and rainy seasons. Average temperatures range between 23ºC and 25ºC. The rainfall pattern is tropical with two distinct

seasons: the dry lasting for four to five months, and the rainy one lasting for seven to eight months, with the highest concentration of rainfall from December to February. Annual rainfall rate ranges between 900 and 1,100 mm (ANA, 2016).

The data for this study were collected from the ANA stations for testing and training from 1995 to 2014, as follows: three river (water) level stations 66945000 (Aquidauana) (~1995), 66941000 (Palmeiras) (~1965) and 66926000 (Ponte do Grego) (~1982); and five rainfall stations: 01954002 (Rochedo), 01954005 (Bandeirantes), 02055002 (Palmeiras), 02055003 (Fazenda Lajeado), and 02054009 (Santa Elisa). In the stations described above, it is shown the date of installations according to ANA, maintaining the position from its installation.

Back propagation (BPN) learning works by making modifications in weight values starting at the output layer then moving backward through the hidden layers of the network. BPN uses a gradient method for finding weights and is prone to lead to troubles such as local minimum problem, slow convergence pace and convergence unsteadiness in its training procedure. Unlike many search algorithms, which perform a local, greedy search, GAs performs a global search. GA is an iterative procedure that consists of a constant-size population of individuals called chromosomes, each one represented by a finite string of symbols, known as the genome, encoding a possible solution in a given problem space. The GA can be employed to improve the performance of BPN in different ways.

*Collection Data*. For the analysis of rainfall and fluviatile stations, firstly the reading failures of quota and rainfall were observed throughout the whole period recorded in ANA's database to be used for ANNs training and testing. After checking the failures, the list of quotas and the accumulated rainfall effectively measured in chronological order were collected from the database.

*Principal Components*. PCA objective reducing the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables. Keeping most of the sample's information and useful for the compression and classification of data. This work utilized of the Principal Components Analysis (PCA) as a method to pre-process the original multivariate data, being is rewrite in a new matrix with principal components sorted by it's accumulated variance. This new matrix was utilized in input the ANN.

*Model initialization and parameter selection*. The vectors of the BPN, which used rainfall (mm) as input vectors and river level (cm) as the output vector, were normalized in the range [0,1] to eliminate the effects of dimension. Three-layer structures which included a hidden layer were selected through pre-testing and intra-checking. The transfer functions were tansig and purelin; the training function was traingd; the learning rate was 0.05; the target error was $10^{-4}$; and the maximum number of training cycles was 18,000. Thus, Figure 2 shows the configuration of ANN inputs and outputs.

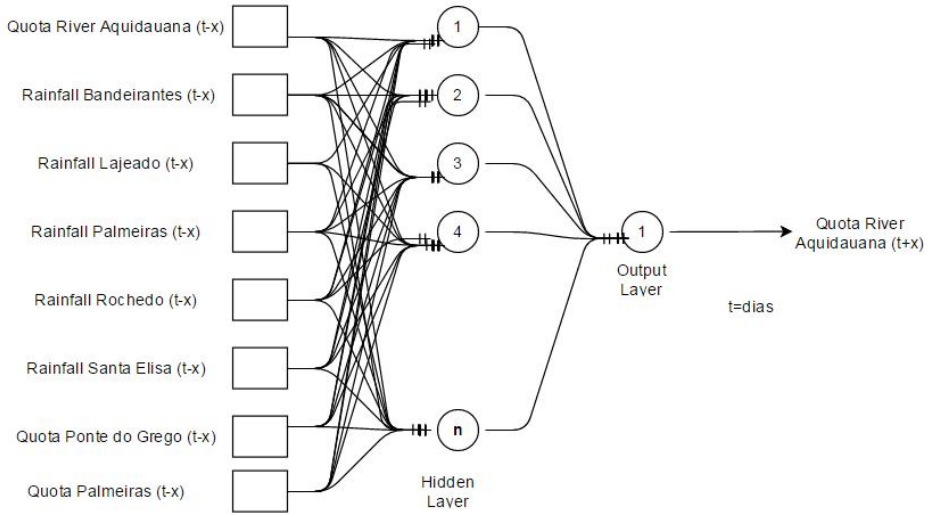



**Figure 2**. Model of an ANN for flooding prediction in the Aquidauana sub-basin.

The ANNs used in this study were built, trained and tested using the Neural Networks of the WEKA software (Mark *et al*., 2013).

*Processing of the GA-BP*. This study utilized the application of a GA to optimize the initial weight and offset of the BPN. The main process of the algorithm was divided into two stages: i) On the basis of the initialization model, the connection value was encoded and the BPN offset to compose the chromosomes of the GA, the chromosomes were optimized by the GA and the decoded chromosomes were assigned to a neural network; and ii) the network weight was further optimized and was offset by the local optimization ability of the training function of the BP. The flowchart is shown in Figure 3 and the functions in the dotted box constitute the first stage.

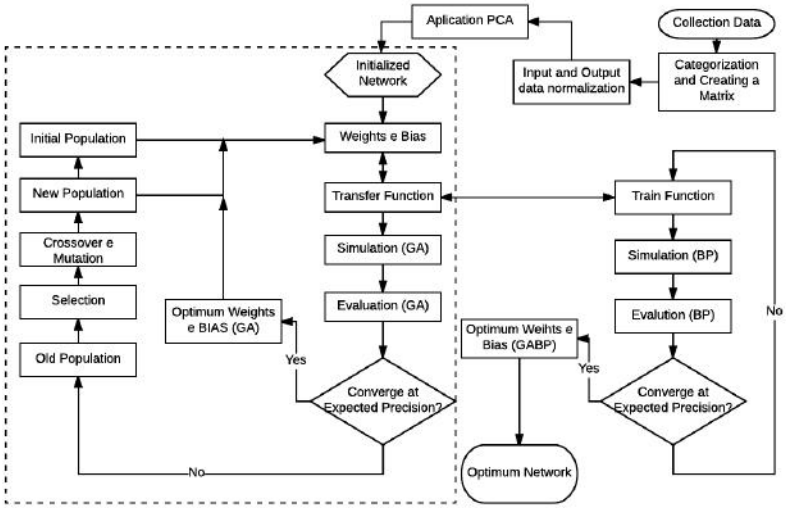



**Figure 3.** Flow chart of the GA-BP. GA-BP, genetic algorithm-back-propagation neural network.

*Training Neural Network Using Genetic Algorithms*. In the initialization, the first thing to do is to decide the coding structure. Coding for a solution is termed as chromosome in GA literature. A randomly generated population is generated which consists of 10 number of chromosomes comprising of real number encoding of size (Nmaxa+1), (Nmaxb+2) and (Nmaxc+3) which includes number of weights for both input-to-hidden layer and hidden-to-

output layer and number of hidden nodes. The chromosomes are classified by rank-based fitness scaling. Parent selection for the next generation is accomplished in a probabilistic manner using universal stochastic sampling as done. Elitism (the best chromosomes preserved for the next generation) was used with a value 2, and the crossover operator uniform crossover used by Sexton *et al*., (2012) was taken for the combination of parent chromosomes, with a probability of 80%. The crossover operation was performed by combining the parts of the parent chromosomes that have the same length as in the succeeding sample. The mutation operator used was the Gaussian mutation as referred by Sexton *et al*., (2012), with a rate of 0.004 as finalized. Fitness function was evaluated for trials with Crossover fraction values varying from 0.1 to 0.9 in the interval of 0.1 crossover fraction of 0.8 was finalized as it provided minimum value of fitness function. Similarly for mutation rate of 0.001 to 0.01 in the interval of 0.001, fitness function was calculated and then mutation function of 0.004 was finalized corresponding to minimum value of fitness function.

Stopping criteria for the genetic algorithm determine what causes the algorithm to terminate. The optimized weights are used as a starting point for the training of BPN. Global optimization techniques are relatively inefficient for minimum local search. In this case, it is important to improve the performance of the ANNs. This strategy represents the optimization method where without changing the architecture generated by the global search, the final network produced is used as a starting point in the local search. In this way, weights obtained by global optimization are preserved. This architecture is used as the start search point for local optimization. This combination of global and local optimization techniques (hybrid training) is applied for the models.

*Training Neural Network Using Genetic Algorithms*. The MLP architecture definition depends on the choice of the number of layers and the number of hidden nodes in each of these layers. The network architecture having a single hidden layer contains N1 input nodes, N2a, N2b, N2c hidden nodes corresponding to first, second and third hidden layer resp., N3 output nodes, and BN2a, BN2b, BN2c and BN3 as the bias for the hidden and output layers, respectively. Parameter N1 is decided by input selection studies. N3 and BN3 are 1 each as runoff is the only output required but N2a, N2b, N2c and BN2a, BN2b, BN2c are to be defined in the ANN implementation which was done with the help of genetic algorithm. MLP architecture contains connections only between adjacent layers. Thus the maximum number of connections for one, two and three hidden layer network is given by equations 1, 2 and 3 resp.

$$Mmaxa = (N1 \times N2a + BN2a) + (N2a \times N3 + BN3) \quad (1)$$

$$Mmaxb = (N1 \times N2a + BN2a) + (N2a \times N2b + BN2b) + (N2b \times N3 + BN3 \quad (2)$$

$$Mmaxc = (N1 \times N2a + BN2a) + (N2a \times N2b + BN2b) + (N2b \times N2c + BN2c) + (N2c \times N3 + BN3) \quad (3)$$

For each set of input, three architectures were built using one, two and three hidden layers. Thus total five models were developed. The output node in each model is one. The models were called MLP.

**Table 1.** Initialized Architecture for GA

| Model | Input Nodes | Hidden Nodes | | | Output Nodes | Length of Chromosome |
|---|---|---|---|---|---|---|
| | N1 | N2a | N2b | N2c | N3 | Nmax |
| MLP1 | 10 | 6 | *** | *** | 1 | 79 |
| MLP2 | 10 | 6 | 4 | *** | 1 | 106 |
| MLP3 | 8 | 6 | *** | *** | 1 | 62 |
| MLP4 | 9 | 6 | 4 | *** | 1 | 95 |
| MLP5 | 9 | 3 | *** | *** | 1 | 68 |

## 4. Results

*Performance evaluation of predictive models*. Table 2 presents the error measures for the five models in which weights were optimized by Genetic Algorithm. The variable t represents time from 1 to 5 days for the forecasting.

**Table 2**. Comparative analysis of model performances.

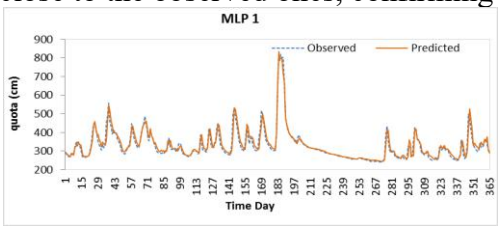
| ***t*** | **Rtr** | Rtst | RMSE (cm) | MAE | ME | RAE | MARE | MRE | MSRE | CE |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **0.95** | 0.93 | **30** | 0.71 | 0.0958 | 0.6328 | 0.016 | 0.014 | 0.00658 | 0.43 |
| 2 | 0.89 | 0.87 | 45 | 0.66 | 0.0825 | 0.4356 | 0.016 | 0.014 | 0.000557 | 0.51 |
| 3 | 0.79 | 0.76 | 57 | 0.60 | **0.0236** | 0.3796 | 0.015 | 0.013 | 0.000574 | 0.68 |
| 4 | 0.69 | 0.67 | 65 | 0.57 | 0.1424 | **0.339** | 0.013 | 0.013 | 0.000563 | **0.69** |
| 5 | 0.56 | 0.54 | 69 | 0.63 | 0.1339 | 0.301 | **0.012** | 0.001 | **0.000407** | 0.67 |

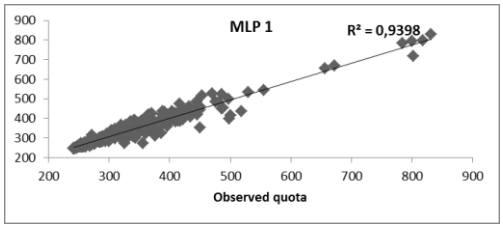
Those values highlighted in bold in this table indicate the "best" model out of the four when assessed using each particular evaluation metric. It can be seen from table 2 that no one model is consistently "best" in terms of the numerous evaluation metrics, although some models appear to be "better" than others, and various tradeoffs exist. Take, as an example, two popular evaluation metrics Mean Squared Relative Error (MSRE) and Root Mean Squared Error (RMSE).

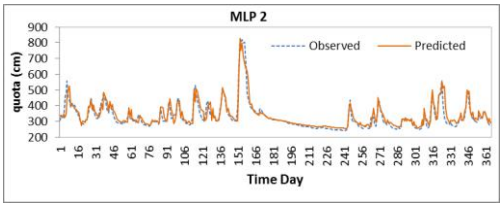
MSRE measures relative performance and is thus more critical of errors that occur at low flows. For five day predicting has the lowest MSRE value of 0.000407. It also predicts the flood peak most accurately. For three day predicting, has conversely over estimates all of the observed values and has the "worst" MSRE (0.00574). RMSE, on the other hand, measures overall performance across the entire range of the dataset. It is sensitive to small differences in model performance and being a squared measure exhibits marked sensitivities to the larger errors that occur at higher magnitudes. In this case, because for one day forecasting follows the hydrograph across the full range of flow events, it has the smallest RMSE. This simple example illustrates the dangers of relying on one measure alone to evaluate and select between different models.

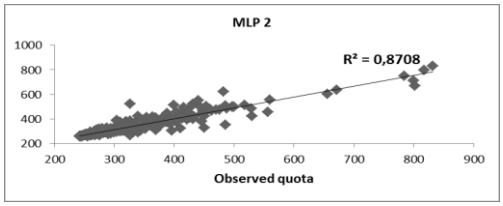
Among the predictions, the best performance was for one day, Rtr of 0.95 and RMSE of 30 cm. Another very satisfactory correlation value, according to Smith (1987), is a two-day prediction. On the other hand, for 5 day forecasting produces a high RMSE and does not satisfactorily meet the model for the region.

Figure 4 displays the estimates of river level provided by the prediction model for MLP for one to five days. The figure shows hydrographs and scatter plots, with the predicted values close to the observed ones, confirming the RMSE statistics in Table 3.



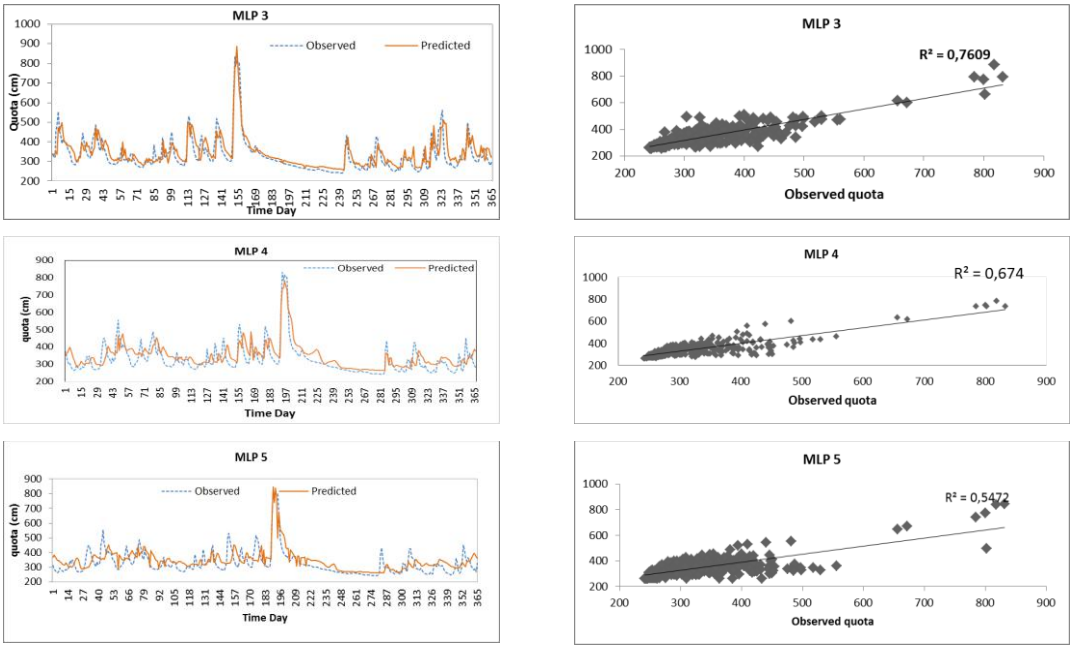



**Figure 4.** Comparison of the proposed quota and the values observed for one to five days in advance.

As shown in the result, model MLP 1 yielded more accurate quota estimates (cm) ($r^2$=0.93 and RMSE = 30 cm).

*Sensitivity Analysis Results*. While training a network, the effect that each of the network inputs is having on the network output should be studied. This provides feedback as to which input channels are the most significant, based on which we may decide to prune the input space by removing the insignificant parameters. This will reduce the size of the network, which in turn reduces the complexity and the training time. Sensitivity analysis is a method for extracting the cause and effect relationship between the inputs and outputs of the network.

The sensitivity analysis consists of disturbing a given model, varying one of its input parameters, and identifying the variation that occurred in the results of interest.

To evaluate the performance of models, two indices RMSE and $R^2$ were used. Table 3 expresses the summarized ANN results of $R^2$ and RMSE for five days quota forecasting. There is a consistency in the performance of models, where ANN model is quite stable.

**Table 3.** Performance statistics for sensitivity analysis.

| Model Index | MLP 1 | MLP 2 | MLP 3 | MLP 4 | MLP 5 |
|---|---|---|---|---|---|
| Training (1995-2013) | | | | | |
| RMSE (cm) | 33 | 49 | 60 | 68 | 75 |
| R2 | 0,93 | 0,85 | 0,72 | 0,67 | 0,52 |
| Testing (2014) | | | | | |
| RMSE (cm) | 31 | 47 | 58 | 66 | 72 |
| R2 | 0,92 | 0,83 | 0,71 | 0,65 | 0,50 |

## 5. Conclusion

For the study catchment, data driven model using Artificial Neural Networks and Genetic algorithm is developed. Input parameters are selected by three selection criterion. For Neural Network architecture, hidden layers and nodes are fixed by GA's. The training of network used optimized weights given by Genetic algorithm. Five model performing criterions are estimated to judge the relative and absolute errors. From a visual inspection of the hydrographs it is relatively easy to categorize the different types of model, but when we require more objective

measures of model performance, dimensionless coefficients that contrast model performance with accepted norms or standards are used. The criterion of RMSE and coefficient of efficiency for test data are found to be highest and lowest. The evaluation results indicate that the best prediction is that for one day, though it is acceptable for up to three days. Model MLP 1 yielded the best result with a coefficient of determination and a mean squared error of 0.95 and 30 (cm), respectively. The analysis sensitivity gave exactly the same BP result but the contributions were not sufficiently expressed. The methodology used with data that always converge avoids stochastic training and is widely applicable to the small watersheds not yet studied.

## 9. Referências

ANA. Agência Nacional da Água. Available: < http://www.ana.gov.br/>. Access in: 20 junho 2016.

Asati, R. S. Comparative Study of Stream Flow Prediction Models. **International Journal of Life Sciences Biotechnology and Pharma Research**, Tirupati, v. 1, n. 2, p. 139-151, 2012.

EMBRAPA. Empresa Brasileira de Pesquisa Agropecuária. **Sistema brasileiro de classificação de solos**. 2 ed. EMBRAPA: Rio de Janeiro, 2006. 306p.

Jones, D. R.; Sshonlau, M.; Welchh, W. J. Efficient global optimization of expensive black box functions. **Journal of Global Optimization**, Ontario, v. 13, n. 1, p. 455–492, 1998.

Hutter, F; Hoos, H. H.; Leyton-Brown, K. Sequential model-based optimization for general algorithm configuration. In: Learning and Intelligent Optimization: 5th International Conference, 507-523, Roma. **Proceedings…** Roma: LNCS, 2013.

Heckerling, P. S.; Gerber, B.S.; Tape, T.G.; Wigtons, R.S. Use of genetic algorithms for neural networks to predict community-acquired pneumonia. **Artificial Intelligence in Medicine**, New York, v. 30, n. 1, p.71-84, 2004.

IBGE. Instituto Brasileiro de Geografia e Estatística. Available in:< **http://cidades.ibge.gov.br/xtras/perfil.php?codmun=500110**>. Access in: 15 fev. 2015.

Jain, A.; Indurthy, S.K.V.P. Comparative Analysis of Event based Rainfall-Runoff Modeling Techniques-Deterministic, statistical and Artificial Neural Networks. **Journal of Hydrologic Engineering**, New York, v. 8, n. 2, p. 1-6, 2013.

Jihoon, Y; Vasant, G. H. Feature Subset Selection Using a Genetic Algorithm. **Journal IEEE Intelligent Systems**, Arizona, v. 13, n. 2, p. 111-126, 2014.

Mark, H.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, H. I. The Weka Data Mining Software: An Update. **SIGKDD Explorations**, San Francisco, v. 11, n. 8, 2013.

Sexton, R. S; Dorsey, R. E; Johnson, J. D. Toward global optimization of neural networks: A comparison of the genetic algorithm and back propagation. **Decision Support System**, New York, v. 22, n.3, p. 171-185, 2012.

Suryanarayana, C.; Sudheer, C.; Mahammood, V.; PanigrahI, B. K. An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India. **Neurocomputing**, New York, v. 8, n. 145, 324–335, 2014.

Tucci, C. E. M.; Collishonn, W. Coupled Hydrologic-Hydraulic Modeling of the Upper Paraguay River Basin. **Journal of Hydrologic Engineering**, New York, v. 17, p. 635-646. 2012.

Van Liew, M. W.; Veith, T. L.; Bosch, D. D.; Arnold, J. G. Suitability of SWAT for the conservation effects assessment project: A comparison on USDA-ARS experimental watersheds. **Journal of Hydrologic Engineering**, New York, v. 12, n. 2, p. 173-189, 2007.

Yao, X. Evolutionary artificial neural networks. **Encyclopedia of computer science and techonology**, New York n. 5, v. 33, p. 137-170, 2014.