International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Comparison of Regression and Artificial Neural Network Impact Assessment Models: A Case Study of Micro-Watershed Management in India.

Saha Narayan[a], Jha Amol Kumar[a], Pathak Krishna Kant[b],
[a]Advanced Materials and Processes Research Institute (AMPRI),
Council of Scientific and Industrial Research (CSIR),
Hoshangabad Road, Near Habibganj Naka, Bhopal-462026, Madhya Pradesh,
India(narayansaha2@yahoo.co.in, jhaamolkumar@hotmail.com); [b]National Technical
Teachers Training and Research Institute, Bhopal-462 002, India
(kkpathak1@rediffmail.com).

Abstract: Impact assessment of a micro-watershed management project has been carried out to evaluate sustainable livelihood security for local people especially, of developing countries. In general, the conventional approaches for impact assessment have been found to be time-consuming, expensive and the data generated through these studies are mostly unused in future. In order to overcome the deficiency of conventional impact assessment methods, the present study has targeted to develop suitable Regression and Artificial Neural Network (ANN) models using identified 144 randomly selected indicators data sets over nine years historical time periods, collected from a successful case study namely “Semri micro watershed, Sehore District, Madhya Pradesh, India”. Regression and ANN decision support system prediction models have been developed with eight most dominating parameters which have found most significant effect on livelihood security. The comparison study of these two models have indicated that, the statistical yield predicted through ANN models performed better than that predicted through regression models. The study has recommended the use of such models for improvement of similar degraded watershed for future reference.

***Keywords**:* Watershed impact assessment; sustainable development; regression model, artificial neural network model; model comparison.

1 INTRODUCTION

Watershed professionals always look for a suitable mathematical and computational decision support tools based on optimum information, cost effectiveness with reasonable accuracy towards measure overall sustainability [Jakeman et al 2006]. In this regard, several mathematical and soft computing models for watershed impact assessment have been developed worldwide in complex and simple approaches. They represent watershed status by applying a number of computational tools through complex numerical manipulations to understand the correlation between parameters to determine the status of resources [Gallagher et al 2007 and Pappenberger et al 2006]. Among these, three popular groups of models have been widely used. They are conceptual models, physically process based models and black-box models. Large number of data required, long time consumption, and their marginally superior results compared to the others have made them an unfavorable choice to measure sustainable watershed management. Alternatively soft computing techniques such as mathematical algorithm, mathematical simulation, stepwise regression analysis and ANN have been widely used for watershed modeling also. Due to the simplicity and forecasting capability of Regression and ANN models, they are found to be more attractive. Regression models have been used for long for watershed hydrology and water quality assessment for site-specific parameters, forecasting and model performance comparison with conventional models towards watershed management decision support system. ANN based models have been developed in last five decade but it has become popular and practiced from last decade only. The main advantage of the ANN approach over the traditional methods is that it needs less data and capable of forecasting for longer period. Also it does not require adequate knowledge on background science. The main constraint of this model is that it is based on black-box because it is generally unable to demonstrate the coefficient of determinants and trend analysis like regression model and the model is developed by trial and error approach. Despite its dependence on of black-box the models have been found more attractive since last one decade and there has a growing trend for use of ANN based watershed model as it has better performance compared to conventional models for forecasting [Sarangi et al 2000a, 2005b, Maier et al 2000, McCulloch et al 1943, Dawson et al 1998, Salas et al 2000 and Pappenberger et al 2006].

The existing mathematical and computational models used for watershed management have been developed and used mostly in site specific conditions. They have been used for advanced assessment at micro-levels, such as prediction of water quality, sediment loss, sediment transport, rainfall-runoff conditions etc. It is done either before starting a project, or during the project or after the completions of the project. Rarely, any of these models are complement to each other and they have their own limitations. In view of the above there is an immense need for less expensive, fast and reliable prediction techniques. In recent years, the focus of watershed management has broadened, incorporating more holistic approaches that deal with larger issues such as natural resources management and improving the livelihood security of the local people. Projects of this kind are becoming more common and have been implemented in micro watershed management. Although most of these projects have been found beneficial to natural resources and livelihood security of the local people living within the watershed, but it was not sustainable because enough attentions have not been paid to the monitoring, evaluation and future prediction. Saha [2010] suggested that the benefits of watershed management project is assessed in terms of increased water availability, cropping area, crop yields, improvements in village income, expenditure, savings, assets, lower migration rates, and status of below poverty level community with respect to measure sustainable livelihood security.

The objectives of the present study are: (i) to develop suitable regression and ANN models based on minimum ground truth data for fast, cost effective and accurate assessment of watershed management and (ii) to compare between these two models and their performance reliability as a decision support tools with reasonable accuracy towards livelihood security assessment. This has been done with reference to a successful case study of "Semri micro-watershed in District of Sehore, Madhya Pradesh, India. Attempt has also been made to assess the benefits with equal time interval by identifying various numbers of suitable indicators data sets (144) collected over nine years period.

2 STUDY AREA

The "Semri micro-watershed" catchments fall under "Delawari milli-watershed" with geo-coded 5D 4D 8A comprising 1200 Ha of land which is located at about 80 Km from Bhopal, State Capital of Madhya Pradesh, India. The study area lies between 22^{0} $45^{/}$ to 22^{0} $55^{/}$ North Latitude and 77^{0} $25^{/}$ to 77^{0} $35^{/}$ East Longitude, under the Survey of India Topo-Sheet No. 55 F/5 and 55 F/9. The study area falls under Sukhi river Basin. Sukhi River is a tributary of River Narmada and flows in northwest direction and is fed with seasonal rainwater. The climate of the study area lies within dry deciduous semi arid region, with average annual rainfall of about 1200 mm. In general, heavy rainfall takes place during July to September. The average annual maximum and minimum temperatures varies between 42^{0}C and 6.5^{0}C, respectively. The average annual maximum and minimum humidity ranges between 87% and 27%, and annual average wind velocity ranges between 13 Km/hr and 2.7 Km/hr respectively. The map of Semri Micro Watershed is shown in Figure 1.

Before 1997 the area was identified as degraded watershed in poor socioeconomic condition. Soil and water conservation activities and augmentation of ground water recharge was carried out over past five years (1997-2002) under Rajiv Gandhi Mission for Watershed Management Programme for improving the natural resources and to provide sustainable livelihood security to the people of this region.

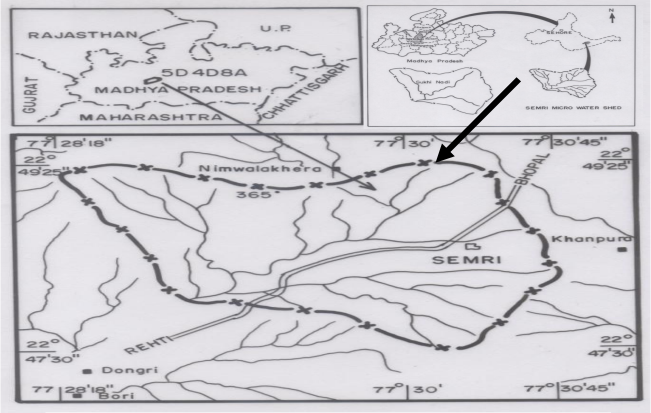

Figure 1 Semri Micro-watershed Map

3 METHODOLOGY

The present study has adopted methodology for collection of primary data based on the overall watershed resources parameters and there significant increase where the soil and water conservation treatments were undertaken as reported by the local farmers. In this regard 144 local based primary indicators data sets were identified. This data were collected through structured questionnaire schedule, by adopting 20% randomly selected village house hold survey as suggested by researchers [Bryceson 2000, Edward et al. 2001, Ellis et al. 2003, Goel et al. 2005, Bhandari 2006, Amsalu 2007]. This 144 data have been collected over nine years (1987-88 to 2006-07) periods in four time periods with three years equal intervals (1987-88, 2000-01, 2003-04 & 2006-07) from actual field work in the form of matrix as follows:

$$\begin{array}{cc|cccc|c|cccc|c}
 & \frac{Year(j)}{Indicator(i)} & 1 & 2 & 3 & 4 & & 1 & 2 & 3 & 4 & \\
 & 1 & Y_{11} & Y_{12} & Y_{13} & Y_{14} & & 269 & 291 & 309 & 330 & \\
 & 2 & Y_{21} & Y_{22} & Y_{23} & Y_{24} & & 43 & 56 & 69 & 80 & \\
\mathbf{Y_{ij}} = & \vdots & \vdots & \vdots & \vdots & \vdots & = & \vdots & \vdots & \vdots & \vdots & (1) \\
 & 144 & Y_{1441} & Y_{1442} & Y_{1443} & Y_{1444} & & 3 & 10 & 16 & 21 &
\end{array}$$

where, i = 1,2,3,……,144 (indicator data sets), and j =1(1987-88), 2(2000-01), 3(2003-04), and 4(2006-07) (four time periods). In order to achieve the minimum but optimum data information, cost effective and less time-consumption for managing, forecasting and assessing watershed livelihood security, it was felt that there have been some most important indicators specific to the area. The primary 144 data were further refined and 20 site-specific sensitive indicators were identified. Attempts were also made to further reduce the number of indicators especially needed for instant and sustainable livelihood assessment modeling purpose. Finally 8 most dominating indicators were developed as optimum parameters by suitably merging similar type of indicators among the previously developed 20 indicators. The 8 most dominating optimum parameters were developed as displayed in Table 1:

Table 1 Development of 8 most dominating indicators

Items	1997-98	2000-01	2003-04	2006-07
Ground Water Table (m)	2.10	2.30	2.50	2.80
Cropping Area (acre)	274.48	299.38	320.40	343.99
Crop Production (Qtl.)	1302.06	1754.97	2262.10	2797.81
Per Capita Income (Rs.)	8280.76	11882.26	17534.02	23761.47
Per Capita Expenditure Rs.)	3870.10	5233.29	7122.71	8912.10
Per Capita Savings (Rs.)	4410.66	6648.96	10411.31	14849.36
Below Poverty Line Family	18.00	15.00	11.00	7.00
Persons Migrated	57.00	52.00	48.00	43.00

The developed 8 parameters were considered as useful in developing immediate and cost effective approach model to assess watershed management and were found to pose dominating effect on livelihood security assessment.

For development of model, initially 75% of the indicator data sets (1987-88, 2000-01 and 2003-04) were considered for best-fit regression model in the first approach. In order to account for the other hidden factors and speeding the assessment process, ANN model was developed as a second approach using back propagation neural network architecture of three layers to train the data. During ANN training, weight and bias functions have continuously been modified by minimizing the mean square error (MSE) using back propagation neural network to predict the output weight factor from output layer for each indicator for best-fit model. An extrapolated desired data value (Y_{id}) for each observed indicators have been developed which was found to be a highest/ lowest as observed in the (Y_i) data values over 4 time periods (X_i) in 3 equal interval years (1987-88, 2000-01, 2003-04 & 2006-07) using trial and error methods. The selection of (Y_{id}) was made based on the normalization factor (N_i) to achieve the desired training time as follows:

$$Y_{i=}[Y_1,Y_2,\ Y_3,\ Y_4,......,\ Y_{id}] \tag{2}$$

$$N_i = \frac{Y_{\max} - Y_1}{Y_{id} - Y_{av}} \tag{3}$$

where, Y_1 = Indicator data value of 1987-88, Y_2 = Indicator data value of 2000-01, Y_1 = Indicator data value of 2003-04, Y_4 = Indicator data value of 2006-07, the value *of* N_i lies between 0 to 1, Y_{max} represent highest observed value among Y_1, Y_2, Y_3 and Y_4 and Y_{av} = average value of Y_1, Y_2, Y_3 and Y_4. The values of Y_{id} against each indicator have been included in the form of matrix as follows:

$$\begin{array}{cc|cccccc|}
 & \frac{Year(j)}{Indicator(i)} & & 1 & 2 & 3 & 4 & Y_{id} \\
 & & 1 & 269 & 291 & 309 & 330 & 350 \\
 & & 2 & 43 & 56 & 69 & 80 & 90 \\
Y_{ij} = & & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
 & & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
 & & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
 & & 144 & 3 & 10 & 16 & 21 & 30
\end{array} \tag{4}$$

For developing ANN model all the 144 input variables data sets of three time periods and Y_{id} were considered as input layer (Y_i). Time periods (x_i) for four input data values were represented as 2 for 1987-88, 5 for 2000-01, 8 for 2003-04 and 15 for $\boldsymbol{Y_{id}}$ respectively. These data were fed in input data file of the ANN architecture for the training of the neural network. After training, the software generated the weighted output factor (Y_{iout}) for each indicator, which was used to calculate ANN prediction (Y_{ip}) for each 144 indicators for the year 2006-07 by using the formula as represented (Equation 5).

$$Y_{ip}\ (Predicted) = (Y_{id} * Y_{iout}) \qquad (5)$$

where, ip=1, 2, 3............144. Determinates of coefficient (R^2) were calculated for different regression and ANN by using formula (Equation 6). The predicted data (Y_{ip}) values were compared with the actual observed data (Y_{io}) sets of the year 2006-07 for all the three types of data sets i.e., for 144, 20 and 8 indicators respectively. The deviation between predicted and actual value were calculated using the statistical formula to understand the performance and reliability of the models as follows (Equation 7 and 8).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_{ip} - Y_{io})^2}{\sum_{i=1}^{n}(Y_{io} - Y_{iav})^2} \qquad (6)$$

$$\%ARE = \frac{\left|Y_{io} - Y_{ip}\right|}{\left|Y_{io}\right|} \text{ x } 100\% \qquad (7)$$

$$E = 1 - ARE \qquad (8)$$

where, *%ARE* was assigned as absolute relative error, *E* represents for the model performance.

3 RESULT AND DISCUSSSIONS

For development of regression models, the study were used first 3 time periods data sets for best-fit regression model. A typical example of the regression model for indicator 1 data set (Rainy Season Cropping Area) was best fitted in regression model embedded in the software for trend analysis (Figure 2).

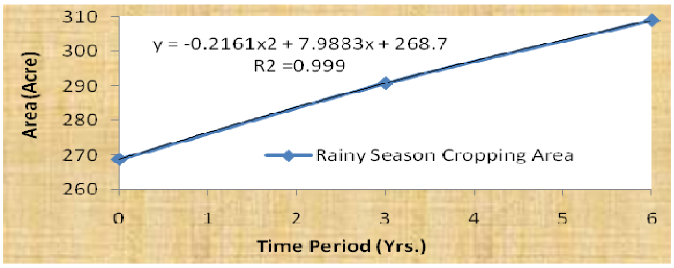

Figure 2 Typical Regression Model

Following the similar methodology various Regression models for all the 144, 20 and 8 indicators were developed respectively. It was observed that the indicators were represented as a function of time and the trends showed either quadratic equation in the form $Y= ax^2 + bx + c$ or linear equation in the form of $Y = mx + c$, where '*Y*' represents the indicator, 'x' represents as time. It was observed that in all the three types of regression models, 118 (83.22%), 20 (100%) and 8 (100%) were observed quadratic equations out of 144 primary indicators, 20 site specific indicators and 8 most dominating indicators respectively and remaining were linear equations. Predictions were calculated for all the three types of the indicators data sets for the year 2006-07 using the earlier developed regression models. The weighted output factors (Y_{iout}) for each indicator were collected after ANN training for all the three types of indicators data sets. The prediction values for each indicator were calculated by using the formula as shown in Equation 5.

The Regression and ANN predicted values of 144 random sampling indicators, 20 selected indicators and 8 most dominating indicators models were compared graphically as follows:

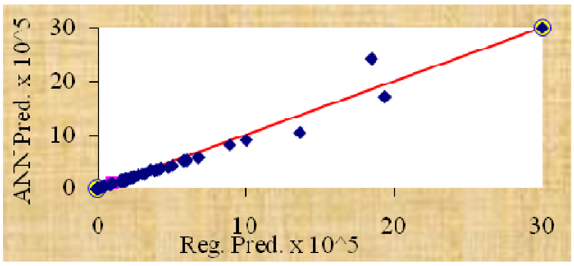

Figure 3 Prediction correlation (144 Indicators)

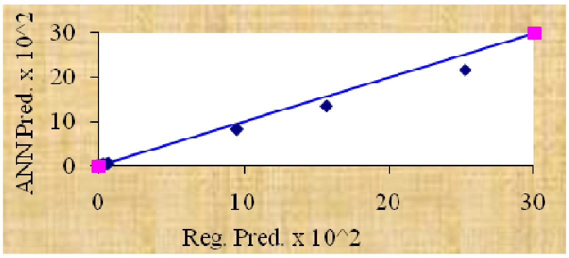

Figure 4 Prediction correlation (20 Indicators)

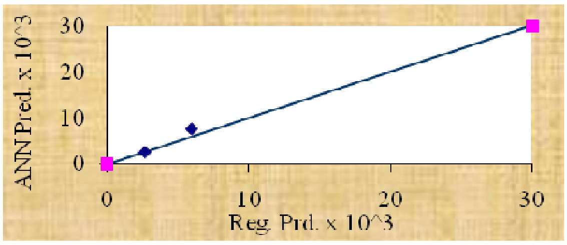

Figure 5 Prediction correlation (8 Indicators)

The statistical yield of both Regression and ANN models were calculated by using formula (Equation 6 to 8) for 144 data, 20 data and 8 parameters. They were summarized to understand the model acceptability and data reliability for holistic watershed impact assessment. It is evident from the Regression validations that statistical yield values for 144 indicators were as follows: $R^2 = 0.99$, E = 0.93, %ARE = 8.09% respectively. It is revealing that out of 144 indicators 114 (79.17%) are within the standard acceptable range i.e. <10% ARE (Average: 3.30%) and 30 (20.83%) are beyond standard acceptable range i.e. >10% (Average: 24.77%). Statistical yields values for 20 selected indicators were as of $R^2 = 0.9974$, E = 0.9677 and %ARE = 7.35%. The results of 20 indicators showed 14 (70%) are within standard acceptable range i.e. <10%ARE (Average: 4.79%) and 6 (30%) are beyond standard acceptable range i.e. >10% ARE (Average:13.15%). Similarly for 8 most important parameters the value observed are as follows: $R^2 = 0.998$, E = 0.994 and %ARE = 5.54% respectively. It is clear that out of 8 indicators 7 (87.5%) indicators are within standard acceptable range <10% ARE (Average: 4.28%) and 1 (12.5%) is beyond standard acceptable range >10% ARE (Average: 14.33%) respectively.

It was observed that in case of ANN validation statistical yields for 144 indicators the values are $R^2 = 0.92$, E = 0.94, %ARE = 6.44% and out of 144 indicators 138 (95.83%) indicators are within standard acceptable range i.e. <10% ARE (Average: 6.21%) and 6 (4.17%) are beyond standard acceptable range i.e. >10% ARE (Average: 12.75%). For 20 selected indicators the values are indicated as follows: $R^2 = 0.93$, E - 0.94, %ARE = 6.26%. In this case 20(100%) indicators are observed within standard acceptable range i.e. <10% ARE (Average: 6.26%) and for 8 most important parameters the values are as follows: $R^2 = 0.93$, E = 0.97, %ARE - 5.62% respectively. It is evident that the 8 (100%) indicators are observed within standard acceptable range i.e. <10% ARE (Average: 5.62%).

The performance reliability (i.e. %ARE) are calculated by using formula (Equation 6) for both regression and ANN models and their comparison are represented graphically, which indicates that ANN model performance for all the three criterion: 144 indicators, 20 indicators and 8 indicators respectively are much more reliable compared to Regression models (Figure 6).

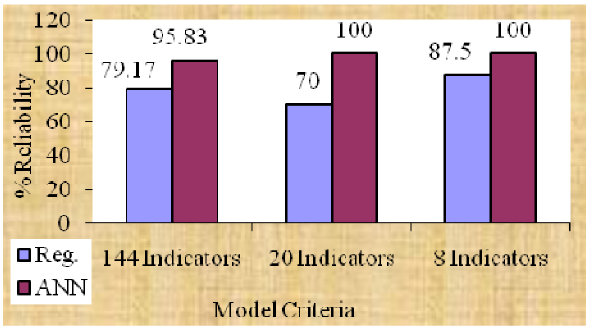

Figure 6 Reg. & ANN Models Performance Reliability

From the above discussions it is evident that the ANN predictions are more reliable compared to regression predictions. It is also revealing that 8 most dominating indicators are sufficient for assessment of watershed management with reasonable accuracy for decision support system and managing watershed management.

4 CONCLUSIONS AND RECOMMENDATIONS

The regression and ANN models developed in present study revealed that the most of these 144 indicators consideration are relevant for overall watershed impact assessment. The comparison of the two models have predicted with actual ground truth data are revealed that the ANN models performance is much reliable and fast compared to conventional regression models and capable to extrapolate even for forecasting of the futuristic impact of any watershed management project based on minimum use of data. It has also been revealed that the decision support system prediction models as developed with the most eight important watershed indicators have got significant positive impacts in nine years assessment period after the completion of the conservation work only and there parameters are sufficient for fast and accurate watershed impact assessment which have dominating effect for assesing livelihood security. The proposed models can prove to be very powerful and handy tools for the researchers, implementers, planners and sponsors working in the field for watershed management and its impact assessment. The study has recommended implementing the model for improvement of similar degraded watershed for future reference.

ACKNOWLEDGEMENTS

The authors are thankful to the Director Advanced Materials and Processes Research Institute (Council for Scientific and Industrial Research), Bhopal, India and Vice-chancellor, Rajiv Gandhi Technical University of Madhya Pradesh State, Bhopal, India for facilitating to conduct the study and to publish the work. The author is also thankful to Dr. D.P. Mondal, Principal Scientist for his valuable guidance to make success of the study.

REFERENCES

Amsalu, A., Graaff, J. D., Determinants of adoption and continued use of stone terraces for soil and water conservation in an Ethiopian highland, *Ecological Economics,* 294-302, 2007.

Bhandari. B.S., Grant, M., Analysis of livelihood security: A case study in the Kali-Khola watershed of Nepal. *Journal of Environmental Management*, 85(1), 17-26, 2006

Bryceson, D., Rural Africa at the crossroads: livelihood practices and policies. *Natural Resources Perspective,* 52, London, 2000.

Dawson, C.W., Wilby, R., 1998, An artificial neural network approach to rainfall-runoff modeling", *Journal of Hydrology,* 43 (1), 47-66, 1998.

Edward, H.A., Ellis, F.,The livelihood approach and management of small-scale sherry, *Marine Policy*, 25, 377-388, 2001.

Ellis, F., Mdoe, N., Livelihoods and poverty reduction in Tanzania, *World Development, 31* (8), 1367-1384, 2003.

Gallagher. M., and Doherty. J., Parameter estimation and uncertainty analysis for a watershed model, *Environmental Modelling & Software*, 22, 1000-1020, 2007.

Goel., A.K., Kumar, R., Economic analysis of water harvesting in a mountainous watershed in India, *Agricultural Water Management* , 71: 257-266, 2005.

Jakeman, A. J., Letcher, R. A., Norton, J. P., Ten iterative steps in development and evaluation of environmental models, *Environmental Modelling & Software*, 21, 602-614, 2006.

Maier, H.R., Dandy, G.C., Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, *Environmental Modeling and Software,* 15 (1), 101–124, 2000.

McCulloch, W.S., Pitts, W.H., A logical calculus of the ideas immanent in neural nets, *Bull. Math. Biophys,* 5, 115-133, 1943.

Pappenberger, F., Beven, K.J., Ignorance is bliss: or seven reasons not to use uncertainty analysis, *Water Resources Research,* 42(5), WO5302, doi: 10.1029/2005WR004820, 2006.

Saha, N, Mathematical and Computational Models for Economic and Environmental Impact Assessment –A case study of Micro Watershed in Sehore, *PhD thesis,* Rajiv Gandhi Technical University of Madhya Pradesh, Bhopal, 2010.

Salas, J.D., Markus, M., Tokar, A.S., Streamflow forecasting based on Artificial Neural Networks In: Govindraju, R.S., Ramachandra Rao, A. (Eds.). *Artificial Neural Networks in Hydrology,* Kluwer Publishers, London, 23-51, 2000.

Sarangi, A., Bhattacharya, A.K., Use of geomorphological parameters for sediment yield prediction from watersheds, *J. Soil Water Conservation,* 44(1-2), 99-106, 2000a

Sarangi, A., Bhattacharya, A.K., Comparison of Artificial Neural Network and regression models for sediment loss prediction from Banha watershed in India, *Agricultural Water Management,* 78, 195-208, 2005b.