

Comparative Application of ANN and PCA in Modeling of Groundwater

D. Sarala Thambavani¹, T.S.R. Uma Mageswari^{2*}

¹Department of Chemistry, Sri Meenakshi Government Arts College for Women (Autonomous), Madurai – 625 001, TN, India.

²Department of Chemistry, PSNA College of Engineering & Technology, Dindigul – 624 622, TN, India.

ARTICLE DETAILS

Article history:

Received 23 January 2015

Accepted 01 February 2015

Available online 05 February 2015

Keywords:

Principal Component Analysis

Neural Network

Levenberg-Marquardt Algorithm

Varimax Rotation

ABSTRACT

The main goal of this research is to develop an ANN (Artificial Neural Network) model with simple structure and ample accuracy. In the first step, an appropriate ANN model with 24 input variables is developed using back propagation neural network by Levenberg-Marquardt algorithm to optimize the network parameters for the prediction of phosphate, total hardness and turbidity concentration in Batlagundu, Tamil Nadu. Subsequently, principal component analysis (PCA) is used to reduce the number of input variables. Finally, comparison amongst the operation of ANN-PC24 and ANN-RPC models is made. Findings indicated that the ANN-RPC models have more effective results than the ANN-PC24 model.

1. Introduction

Assessment of the quality and quantity of both surface and groundwater is important in hydro-environmental management to sustain the natural systems and safe livable environment on and under the earth's surface. Groundwater and surface water are fundamentally interconnected. This interconnection should be well understood to effectively and safely manage the precious groundwater and surface-water resources while benefiting from them. During recent years, increasing pollution and losing of water sources have changed exploitation policy of water and soil resources. Groundwater is a major source of water supply in different cities around the world and therefore several studies have highlighted different aspects of groundwater such as, storage potential, hydrogeology, water quality, vulnerability and sustainability and so on [1-4]. Contamination of groundwater resources either from anthropogenic activities or from inherent aquifer material composition reduces its supply, posing a threat to development and a challenge to water managers and strategists. Agricultural activities may deteriorate the groundwater quality in small to large watersheds, especially due to uncontrolled use of fertilizer and various carcinogenic pesticides [5,6]. Variation in groundwater quality is a function of physical and chemical parameters that are greatly influenced by geological formations and anthropogenic activities as well [7].

In general, the contamination of groundwater could occur from non-point and point sources. The major contaminants linked to non-point sources are fertilizers, heavy metals and pesticides. Heavy metals contaminate groundwater from anthropogenic sources as well as natural sources. Some of the major anthropogenic sources of heavy metals are mining, fertilizers, pesticides and industrial wastes. While the effects of most chemicals commonly found in drinking water manifest themselves after long exposure, the effects of calcium and, in particular, those of magnesium on the cardiovascular system are believed to reflect recent exposures. Water quality is extremely important, because constant access to good quality water is a condition necessary for life and economy activities. Beside the human life and economy activities, as indicated by the sediment and suspended particles monitoring is essential for the sustainability of the biological resources. The turbidity of any water sample is the reduction of transparency due to the presence of particulate

matter such as clay or slit, finely divided organic matter, plankton and other microscopic organisms. Measurement of turbidity reflects the transparency of water. Therefore, it is necessary to forecast phosphate, total hardness (Ca and Mg) and turbidity. In view of this complex interaction, use of modeling techniques to solve this problem, is needed. However, the problem of obtaining models that adequately represent the dynamic behaviour of field data is not easy. Lack of good understanding and description of the phenomena involved, the availability of reliable and complete field data set and the estimation of the numerous parameters involved are the major factors contributing to this problem. Beck [8] noted that, increase in model complexity will undoubtedly increase the number of parameters, leading to the problems of identification.

ANN (Artificial Neural Network) is an oversimplified simulation of the human brain and is composed of simple processing units referred to as neurons. It is able to learn and generalize from experimental data even if they are noisy, imperfect or nonlinear in nature. This ability allows this computational system to learn constitutive relationships of materials directly from the result of experiments. Unlike conventional models, it needs no prior knowledge, or any constants and/or assumptions about the deformation characteristics of the geo materials. Other powerful attributes of ANN models are their flexibility and adaptivity, which play important roles in material modeling. When a new set of experimental results cannot be reproduced by conventional models, a new constitutive model or a set of new constitutive equations needs to be developed. However, trained ANN models can be further trained with the new data set to gain the required additional information needed to reproduce the new experimental results. These features ascertain the ANN model to be an objective model that can truly represent natural neural connections among variables, rather than a subjective model, which assumes variables obeying a set of predefined relations [9]. In the literature, there are also some ANN studies aiming to predict the conditions in soil and quality of groundwater. Das [10] used computational intelligence techniques ANN and support vector machine to develop models to predict swelling pressure from the inputs: natural moisture content, dry density, liquid limit, plasticity index and clay fraction. In another study, Benerjee [11] used ANN feed forward network based ANN model as a method to predict the groundwater levels. Yesilnacar [12] developed an ANN model predicting concentration of nitrate, the most common pollutant in shallow aquifers, in groundwater of the Harran Plain. Yesilnacar [12] also developed an ANN model predicting concentration of sulfate and SAR value.

*Corresponding Author

Email Address: umamageswaritsr1980@gmail.com (T.S.R. Uma Mageswari)

In natural environment, water quality is a multivariate phenomenon, at least as reflected in the multitude of constituents which are used to characterize the quality of water body. Water quality is very difficult to model because of the different interactions between pollutants and meteorological variables. The principal component analysis (PCA) is one of the approaches to avoid this problem and has received increasing attention as an accepted method in environmental pattern recognition [13-15]. The objective of this study is to use the PCA method to classify predictor variables according to their interrelation, and to obtain parsimonious prediction model (i.e., model that depend on as few variables as necessary) for water quality parameters with other physico-chemical data as predictor variables to model the water quality of the study area. For this purpose, principal component scores of 24 physico-chemical water quality parameters were generated and selected appropriately as input variables in ANN models for predicting phosphate, total hardness and turbidity concentration.

2. Experimental Methods

2.1 Study Area and Data Analysis

The study area Batlagundu is located in Dindigul district, Tamil Nadu. It is bounded by longitude 77° 45' 33.84" E and latitude 10° 9' 55.80" N with an average elevation of 320 meters (1049 feet). The main occupation of this study area is agriculture. The sources of water supply in the area are hand pumps, bore holes and dug wells. The precipitation which is the sole source of ground water recharges in the study area is very low. The area is very humid (86%) and warm with an average temperature 22 °C. In order to achieve the research objective, samples were collected from the study area as shown in Fig. 1. Samples were collected in polythene bottles and analyzed for various water quality parameters as per standard procedures [16-18]. The water quality parameters were monitored regularly during different seasons of summer, rainy and winter over a period of two years i.e., 2012 and 2013. A total of 150 samples for 24 water quality parameters were used for the analysis. The parameters are temperature, pH, electrical conductivity, turbidity, Total dissolved solids, chloride, sulphate, nitrate, nitrite, ammonia, sodium, potassium, iron, calcium, magnesium, dissolved oxygen, biological oxygen demand, chemical oxygen demand, phosphate, total alkalinity, total hardness, bicarbonate, carbonate and fluoride.



Fig. 1 Map of the study area

2.2 Principal Component Analysis

In this work, PCA was performed on the above mentioned water quality parameters to rank their relative significance and to describe their interrelation patterns. Chosen PC scores of the 24 water quality parameters were used as input variables in ANN model to predict the phosphate, total hardness and turbidity concentration.

The principal components (PCs) can be expressed as

$$Z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + \dots + a_{im}x_{mj} \quad (1)$$

Where z is the component score, a is the component loading, x the measured value of variable, i is the component number, j is the sample number and m is the total number of variables. The PCs generated by PCA are sometimes not readily interpreted; therefore, it is advisable to rotate the PCs by varimax rotation. Varimax rotation ensures that each variable is maximally correlated with only one PC and a near zero association with the other components [19,20]. Varimax rotations applied on the PCs with eigenvalues more than 1 are considered significant where the typical criteria are 75-95% of total variance [21]. The rotations were carried out, in order to obtain new groups of variables. Variables with communality greater than 0.7 are considered, having significant factor loadings.

2.3 Artificial Neural Network

Artificial neural network models (Fig. 2) were specified by the network topology, training and/or learning rules. These aspects have primarily affected the network performance; with three different layers in the network topology can be distinguished as:

- 1) An input layer: connecting the input information to the network.
- 2) Hidden layer: acting as an intermediate computational layer.
- 3) Output layer: producing the desired outputs.

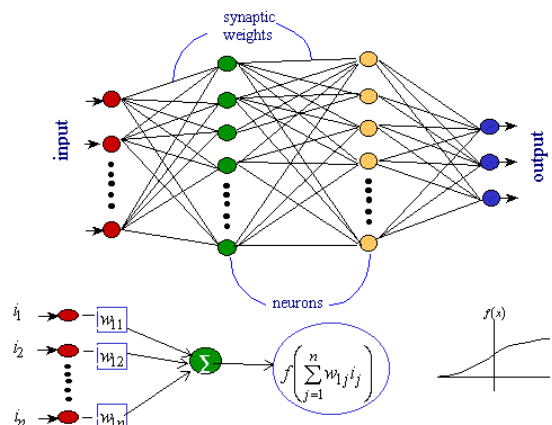


Fig. 2 Architecture of neural network model

The ANN architecture for water quality prediction were decided by the performance of the respecting networks. The training algorithm for the whole networks utilizes the Levenberg - Marquardt Back Propagation algorithm and three layer back-propagation ANN is used in this study. The ANNs were trained for both type of inputs and different number of hidden neurons, sequentially selected to find the best performing network in the training data. An early stopping approach also applied to the network training. The best performance network on the training data was selected as trained network, and featured in the water quality predictor network. The number of input and output neurons is determined by the nature of the problem under study. In this study, the networks were trained, tested and validated with one hidden layer and 1 to 10 hidden neurons.

Two different types of ANN models were developed. In the first type, prediction was performed based on the original PCs. In the second type, scores of rotated (varimax rotation) PCs (ANN-RPCs) with eigen values greater than 1 were selected as input. For this model, prediction of phosphate, total hardness and turbidity concentration was performed using two to six rotated principal components separately.

The original PCs and rotated PCs (RPCs) data sets consist of 150 observations (150 rows) and are divided into training, testing and validating phases for prediction. The input data matrix consists of 24 water quality variables (column) and 150 observations (rows) [24×150]. Table 1 describes the data structure. The validation data is at least 10% of the whole data set, with 75% training set and 25% testing set data [22].

Table 1 The data structure for ANN Prediction model

No. of Observations	Input Parameters					Output
	Input 1	Input 2	Input 3		Input 24	
1	Obs 1,1	Obs 1,2	Obs 1,3		Obs 1,24	O _{1,1}
2	Obs 2,1	Obs 2,2	Obs 2,3		Obs 2,24	O _{2,1}
.....						
.....						
.....						
150	Obs 150,1	Obs 150,2	Obs 150,3		Obs 150,24	O _{150,1}

2.4 Determination of Model Performance

After the training is complete, the performance of the ANN is determined. The verification statistics are useful in evaluating the modeling results. The performances of the back-propagation neural network models were based on the correlation coefficient, root mean square error. The root-mean square is a measure of the prediction error and is used to summarize the overall quality of the model. Lower numbers of the root-mean square error indicate good performance of the model. The correlation coefficient statistic evaluates the linear correlation between the measured and computed values.

$$R = \frac{\sum (Q_o - M_o)(Q_p - M_p)}{\sqrt{\sum (Q_o - M_o)^2 \sum (Q_p - M_p)^2}} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Q_o - Q_p)^2}{n}} \quad (3)$$

Where Q_o and Q_p are observed and predicted value. M_o and M_p are mean of observed and predicted values. ANN models and statistical analyses were carried out using MATLAB 10.0.

3. Results and Discussion

Post PCA, out of the 24 principal components generated, only six PCs with eigenvalues higher than 1 (Table 2) were selected for the ANN input parameters. Selected PCs explained 88.55 % of the total variation. Furthermore, communality values were high for the selected PCs, for example, the values are 99% for Total dissolved solids, electrical conductivity and chloride, 95% for total hardness, calcium, magnesium and sodium, 92% for potassium and total alkalinity (Table 3). These results further confirm the choice of the selected number of PCs (Stevens, 1986). For the first six rotated PCs (RPCs), the loadings from PCA are given in Table 3. The highest correlations between variables are noted in bold. Significant variables in RPC1 are Total dissolved solids, electrical conductivity, total hardness, calcium, magnesium, sodium, potassium, chloride and sulphate; RPC2 are Total alkalinity, bicarbonate, carbonate, total hardness, calcium, magnesium; RPC3 are nitrate, nitrite, iron and ammonia and RPC4 dissolved oxygen, biological oxygen demand, chemical oxygen demand and pH. The only meaningful loads in RPC5 are Temperature and nitrate and in RPC6 are Turbidity and Fluoride.

Using the original principal component scores as inputs, the best architecture consist of a three layer network with 24 input neurons, 5 neurons in the hidden layer and one neuron in the output layer. Considering RPC scores as inputs, the best architectures were achieved with almost the same number of hidden neurons. Training was carried out for a maximum 10000 iterations. Selection of the network was performed at maximum correlation coefficient (r) and low RMSE.

Table 2 Descriptive statistics of selected original PCs with eigen values more than 1

Components	PC1	PC2	PC3	PC4	PC5	PC6
Eigen value	10.089	3.582	2.661	1.587	1.249	1.156
Variability(%)	42.039	14.927	11.088	6.614	5.203	4.816
Cumulative %	42.039	56.966	68.054	74.668	79.870	88.557

Table 3 Rotated factor loadings using six PCs

Variables	RPC1	RPC2	RPC3	RPC4	RPC5	RPC6	Communalities
DO	-0.225	0.007	-0.105	-0.782	0.343	-0.125	0.806
BOD	0.055	0.460	0.214	0.649	0.059	0.070	0.689
COD	0.016	0.500	0.136	0.621	0.247	0.029	0.715
Phosphate	0.431	0.525	0.470	0.087	0.371	-0.148	0.850
Temp	0.186	0.031	-0.053	0.050	-0.905	-0.011	0.860
Turbidity	0.117	-0.265	0.222	-0.031	0.313	0.734	0.771
TDS	0.938	0.236	0.188	0.087	-0.002	0.120	0.993
EC	0.940	0.221	0.194	0.086	-0.005	0.121	0.992
pH	0.210	-0.265	0.114	0.563	-0.069	-0.120	0.679
TA	0.084	0.928	0.206	0.105	-0.476	-0.004	0.927
TH	0.696	0.634	-0.107	0.039	0.064	0.228	0.955
Calcium	0.671	0.662	-0.131	0.022	0.070	0.209	0.955
Magnesium	0.720	0.592	-0.076	0.064	0.060	0.253	0.947
Sodium	0.908	0.039	0.324	0.149	-0.027	-0.021	0.954
potassium	0.825	-0.270	0.385	0.110	-0.066	-0.058	0.922
Chloride	0.949	0.191	0.154	0.076	0.038	0.141	0.989
Sulphate	0.873	-0.033	0.159	0.077	-0.111	-0.096	0.815
Bicarbonate	0.076	0.931	0.202	0.100	-0.081	0.002	0.929
Carbonate	0.116	0.929	0.220	0.060	-0.050	0.000	0.930
Nitrate	0.146	0.074	0.589	0.087	0.563	0.196	0.737
Nitrite	0.221	0.440	0.663	-0.035	0.044	-0.025	0.686
Fluoride	-0.098	-0.254	-0.075	-0.124	0.175	-0.668	0.572
Iron	0.225	0.218	0.804	0.177	0.011	0.132	0.794
Ammonia	0.328	0.090	0.785	0.295	0.094	0.177	0.858

3.1 ANN Phosphate Modeling

Table 4 illustrates the prediction performances of phosphate concentration of ANN models using different combinations of PC scores as input variables. ANN using 6 inputs performs very well as far as accuracy

is concerned for all the data sets. It is observed that the prediction performance of the training phase is best compared to the validation and testing phases. It is important to point out that for this model, the cumulative percentage in explaining the variance given by these six RPCs is 88.55%. Based on the results, it is apparent that the forecasting performance of phosphate concentration increases with the increase in number of input variables. The highest accuracy in predicting phosphate concentration is given by model ANN-RPC6, which contains six RPCs with 88.5% variation explained, giving an correlation coefficient value of 0.91 (training), 0.80 (testing), and 0.84 (validation) respectively. The results are shown in Fig.3. From table 4, it can be observed that the prediction performance of the ANN model using original PCs (24 input PC scores) is not significantly different from the RPC models. However, as RPC models use fewer variables and is far less complex, the advantage over the ANN-PC24 model is obvious.

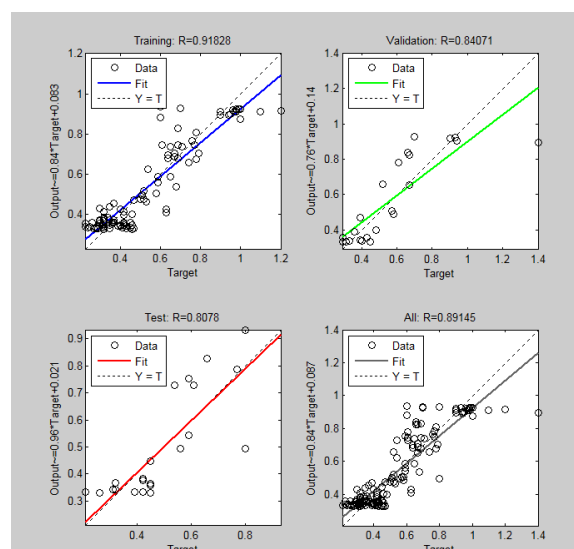


Fig. 3 ANN-RPC 6 model for phosphate prediction

3.2 ANN Total Hardness Modeling

Table 5 illustrates the prediction performances of total hardness concentration of ANN models using different combinations of PC scores as input variables. ANN using the first 2 PCs (PC1 and PC2) performs very well as far as accuracy is concerned for all the training, testing and validation phases. In contrast to phosphate prediction, the above result shows that the forecasting performance of total hardness concentration increases with the decrease in number of input variables. The highest accuracy in predicting total hardness concentration is given by model ANN-RPC2, giving an correlation coefficient value of 0.9982 (training), 0.9971 (testing), and 0.9973 (validation) respectively (Fig.4). Prediction performance of the ANN model using original PCs (24 input PC scores) shows significant variation from the RPC models. Accuracy and Excellency is viewed in the case of varimax rotated PCs compared to unrotated original PCs with a complex network structure.

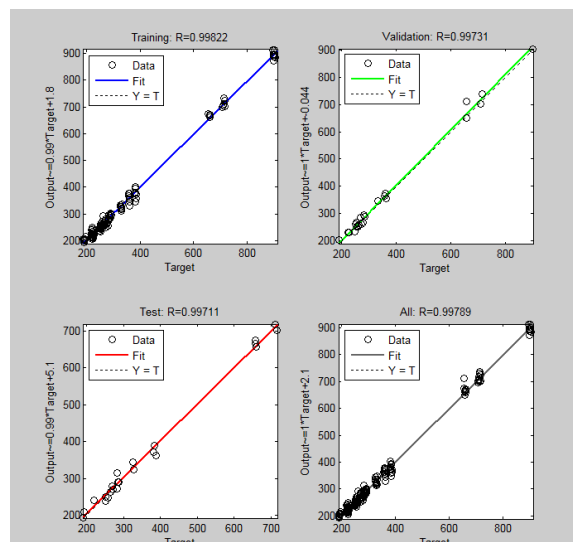


Fig. 4 ANN-RPC 2 model for total hardness prediction

Table 4 Different ANN models for phosphate prediction

Model	No. of PC	Correlation coefficient (r)				RMSE			
		Training	Testing	Validation	All	Training	Testing	Validation	All
ANN-RPC2 (2 Inputs)	2	0.8427	0.7687	0.6733	0.8169	0.0132	0.0473	0.1252	0.0231
ANN-RPC3 (3 Inputs)	3	0.8944	0.5400	0.7480	0.8261	0.1083	0.1411	0.2012	0.0198
ANN-RPC4 (4 Inputs)	4	0.8723	0.7836	0.8647	0.8587	0.1414	0.1510	0.1682	0.0176
ANN-RPC5 (5 Inputs)	5	0.8920	0.8593	0.8792	0.8818	0.1225	0.1374	0.1175	0.0157
ANN-RPC6 (6 Inputs)	6	0.9182	0.8078	0.8407	0.8915	0.0933	0.0447	0.0358	0.0134
ANN-PC24 (24 original Inputs)	24	0.9030	0.8228	0.8866	0.8876	0.0316	0.0447	0.0268	0.0165

Table 5 Different ANN models for total hardness prediction

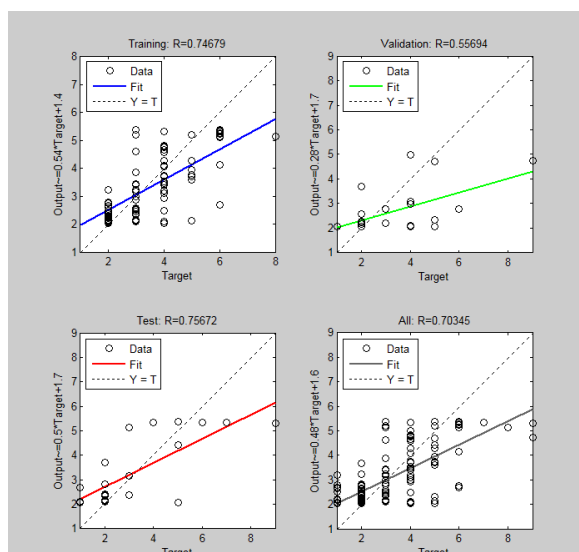
Model	No. of PC	Correlation coefficient (r)				RMSE			
		Training	Testing	Validation	All	Training	Testing	Validation	All
ANN-RPC2 (2 Inputs)	2	0.9982	0.9971	0.9973	0.9979	0.0151	0.0148	0.0163	0.0154
ANN-RPC3 (3 Inputs)	3	0.9977	0.9775	0.9607	0.9856	0.1049	0.1319	0.1183	0.0168
ANN-RPC4 (4 Inputs)	4	0.8448	0.7003	0.9486	0.8361	0.2883	0.1204	0.4849	0.2908
ANN-RPC5 (5 Inputs)	5	0.7426	0.5102	0.8087	0.7055	0.4575	0.3034	0.7503	0.3106
ANN-RPC6 (6 Inputs)	6	0.6139	0.5513	0.5395	0.5905	0.5586	0.7492	0.8205	0.6534
ANN-PC24 (24 original Inputs)	24	0.8489	0.8850	0.8286	0.8464	0.1789	0.1676	0.1162	0.1567

Table 6 Different ANN models for turbidity prediction

Model	No. of PC	Correlation coefficient (r)				RMSE			
		Training	Testing	Validation	All	Training	Testing	Validation	All
ANN-RPC2 (2 Inputs)	2	0.5009	0.3032	0.5723	0.4698	1.5297	1.6912	1.2961	1.3745
ANN-RPC3 (3 Inputs)	3	0.6634	0.5357	0.3799	0.6102	1.2884	1.1662	1.6031	1.5734
ANN-RPC4 (4 Inputs)	4	0.7339	0.4815	0.6657	0.6822	1.1747	1.4870	1.1662	1.2312
ANN-RPC5 (5 Inputs)	5	0.7217	0.6058	0.5051	0.6698	1.1576	1.2369	1.1662	1.0956
ANN-RPC6 (6 Inputs)	6	0.7468	0.7567	0.5569	0.7034	1.0724	1.1269	1.1313	1.1178
ANN-PC24 (24 original Inputs)	24	0.6549	0.7479	0.7437	0.6814	1.0247	1.3491	1.6156	1.5649

3.3 ANN Turbidity Modeling

Table 6 illustrates the prediction performances turbidity concentration of ANN models using different combinations of PC scores as input variables. ANN using the first 2 PCs (PC1 and PC2) does not perform very well as far as accuracy is concerned for all the training, testing and validation phases. It is observed that the prediction performance of the testing phase is slightly worse compared to the training and validation phases. It is important to point out that for this model, the cumulative percentage in explaining the variance given by these two RPCs is only 56.9%. From the table, it clearly shows that turbidity prediction performance increases with the increase in number of input variables. The highest accuracy in predicting turbidity is given by model ANN-RPC6, which contains six RPCs with 88.5% variation. Fig.5 shows the correlation coefficient value of 0.74 (training), 0.75 (testing), and 0.55 (validation) respectively. From table 6, it can be observed that ANN –RPC6 model have more effective results than the ANN-PC24 model.

**Fig. 5** ANN-RPC6 model for turbidity prediction

4. Conclusion

In this work, a combination of PCA and ANN is used to predict phosphate, total hardness and turbidity based on 24 historical water quality parameters. To obtain the latent variables as inputs into the ANN, two different approaches were used; one based on un-rotated original PCs and the other based on varimax rotated PCs. Using six PCs, significant variables in RPC1 are Total dissolved solids, electrical conductivity, total hardness, calcium, magnesium, sodium, potassium, chloride and sulphate; RPC2 are Total alkalinity, bicarbonate, carbonate, total hardness, calcium, magnesium; RPC3 are nitrate, nitrite, iron and ammonia and RPC4 dissolved oxygen, biological oxygen demand, chemical oxygen demand and pH. The only meaningful loads in RPC5 are Temperature and nitrate and in RPC6 are Turbidity and Fluoride. ANN models based on these 6 PC scores can predict 3 parameters with acceptable accuracy (within 95% confidence limit). Moreover, the ANN model using the 24 original PCs as input, do not render the prediction more accurate, even with a complex network structure. The use of rotated PC scores based models is clearly more effective and efficient due to the elimination of collinearity and reduction of predictor variables without losing important information.

References

- [1] V.P. Pandey, F. Kazama, Hydrogeologic characteristics of groundwater aquifers in Kathmandu Valley, Nepal, *Environ. Earth Sci.* 62 (2011) 1723–1732.
- [2] V.P. Pandey, S. Shrestha, F. Kazama, Groundwater in the Kathmandu Valley: development dynamics, consequences and prospects for sustainable management, *European Water* 37 (2012) 3–14.
- [3] V.P. Pandey, S. Shrestha, S.K. Chapagain, F. Kazama, A framework for measuring groundwater sustainability, *Environ. Sci. Policy* 14 (2011) 396–407.
- [4] S.K. Chapagain, V.P. Pandey, S. Shrestha, T. Nakamura, F. Kazama, Assessment of deep groundwater quality in Kathmandu Valley using multivariate statistical techniques, *Water Air Soil Pollut.* 210 (2010) 277–288.
- [5] M.N. Almasri, J.J. Kaluarachchi, Neural networks to predict the nitrate distribution in ground water using the on ground nitrogen loading and recharge data, *Environ. Model. Soft.* 7 (2005) 851–871.
- [6] M.I. Yesilnacar, E. Sahinkaya, Artificial neural network prediction of sulfate and SAR in an unconfined aquifer in southeastern Turkey, *Environ. Earth Sci.* 67 (2012) 1111–1119.
- [7] T. Subramani, L. Elango, S.R. Damodarasamy, Groundwater quality and its suitability for drinking and agricultural use in Chithar River Basin, Tamil Nadu, India. *Environ. Geol.* 47 (2005) 1099–1110.

- [8] M.B. Beck, Identification, estimation and control of biological waste-water treatment processes, IEEE Proceeding 133 (1986) 254-264.
- [9] M. Banimahd, S.S. Yasrobi, P.K. Woodward, Artificial neural network for stress strain behavior of sandy soils: Knowledge based verification, Comput. Geotech. 32 (2005) 377–386.
- [10] S.K. Das, P. Samui, A.K. Sabat, T.G. Sitharam, Prediction of swelling pressure of soil using artificial intelligence techniques, Environ. Earth Sci. 61 (2010) 393–403.
- [11] P. Benerjee, , R.K. Prasad, V.S. Singh, Forecasting of groundwater level in hard rock region using artificial neural network, Environ. Geol. 58 (2009) 1239–1246.
- [12] M.I. Yesilnacar, E. Sahinkaya, M. Naz, B. Ozkaya, Neural network prediction of nitrate in groundwater of Harran Plain, Turkey, Environ. Geol. 56 (2008) 19–25,
- [13] V. Simeonov, J.A. Stratis, C. Samara, G. Zachariadis, D. Voutsas, A. Anthemidis, M. Sofoniou, Th. Kouimtzis, Assessment of the surface water quality in Northern Greece, Water Res. 37 (2003) 4119–4124,
- [14] B. Helena, R. Pardo, M. Vega, E. Barrado, J.M. Fernandez, L. Fernandez, Temporal evaluation of ground water composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis, Water Res. 34 (2000) 807–816.
- [15] K. Loska, D. Wiechula, Application of principal component analysis for the estimation of source of heavy metal contamination in surface sediments from the Rybnik Reservoir, Chemosphere 51 (2003) 723–733.
- [16] APHA, AWWA and WPCF, *Standard Methods for the Examination of Water and Wastewater*, American Public Health Association, Washington D.C. 14 (1995).
- [17] D.K. Trivedy, P.K. Goel, Chemical and biological methods for water pollution studies, Environment publication, Karad, India, 1984.
- [18] NEERI, Manual on water and waste water Analysis, National Environment Engineering Research Institute, Nagpur, India. 1986.
- [19] S.A. Abdul-Wahab, C.S. Bakheit, S.M. Al-Alawi, Principal component and multiple regression analysis in modeling of ground-level ozone and factors affecting its concentrations, Environ. Model. Soft. 20 (2005) 1263–1271.
- [20] S.I.V. Sousa, F.G. Martins, M.C.M. Alvim-Ferraz, M.C. Pereira, Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations, Environ. Model. Soft. 22 (2007) 97–103,
- [21] Q. Chen, A.E. Mynett, Integration of data mining techniques and heuristic knowledge in fuzzy logic modeling of eutrophication in Taihu Lake, Ecol. Model. 162 (2003) 55–67.
- [22] J.T. Kuo, M.H. Hsieh, W.S. Lung, N. She, Using artificial neural network for reservoir eutrophication prediction, Ecol. Model. 200 (2007) 171–177.