ISEIS

Journal of Environmental Informatics 12(2) 140-149 (2008)

Journal of Environmental Informatics

www.iseis.org/jei

Application of Artificial Neural Network Model to Study Arsenic Contamination in Groundwater of Malda District, Eastern India

B. Purkait[1], S. S. Kadam[2], and S. K. Das[3,*]

[1]*Geological Survey of India, Publication & Information Division, 29, J.L. Nehru Road, Kolkata 700016, India*
[2]*Center for Development of Advanced Computing (C-DAC), Pune University Campus, Ganeshkhind, Pune 411007, India*
[3]*Department of Computational Fluid Dynamics, International Institute of Information Technology (I2IT), Pune 411057, India*

Received 2 Feburary 2008; revised 8 September 2008; accepted 20 October 2008; published online 1 December 2008

ABSTRACT. The extent of arsenic contamination in the groundwater has been estimated using artificial neural network (ANN) based on multi-layer perceptron (MLP) architecture. The input data to the ANN comprised samples collected from different arsenic affected blocks of Malda district, West Bengal, India. Each data sample consisted of the amount of arsenic contaminant observed in the groundwater together with other geochemical parameters believed to have some relationship with the corresponding arsenic contaminant. Here, the inputs to the ANN were observed values for pH (acidic-alkalinity ratio), specific conductivity, total dissolved solids (TDS), salinity, dissolved oxygen (DO), redox potential (Eh), and depth of tube well water, while the expected output for training the ANN was the amount of corresponding arsenic contaminant observed in the groundwater. Using the back propagation technique, the ANN model was trained with a subset of the input data. The trained ANN model was then used to estimate the arsenic contamination in groundwater beyond the specified training data. The quality of the ANN simulations was evaluated in terms of three different error measures; namely, the root mean square error, the mean absolute error, and the percent mean relative error for proper interpretation of the results. We have also used two other methods for prediction; namely, multiple linear regression and active set support vector regression. Amongst the three methods, the ANN model exhibited better prediction results for predicting the arsenic contamination in groundwater. Based on this methodology, it is possible to show that a four-layer feed-forward back propagation ANN model could be used as an acceptable prediction model for estimating the arsenic contamination in groundwater.

Keywords: arsenic contamination, back propagation neural network, groundwater arsenic estimation, active set support vector regression, multiple linear regression

1. Introduction

The demand for fresh water resources has been increasing manifold over the past several years due to urbanization, expansion of agriculture, increasing population, rapid industrialization and economic development. Potable ground water resources are already under pressure because of accelerated abstraction, especially in the arid and semi-arid regions, and are placed under a great threat by pollution resulting from natural processes and various human impacts. As water quality continues to decline, improved tools are needed to measure the water quality changes in hydrological cycle. In ground water pollution investigations, study of factors such as characterization of the water source, type of pollutants, mixing of pollutants with other water bodies, geochemical processes undergone, and the derived information on hydrodynamics of the system are very important.

Deleterious presence of certain elements in ground water especially arsenic is of great concern to mankind. Arsenic (As) is a toxic metalloid that makes up 0.00058% of the total mass of the earth's crust. Concentration of arsenic in drinking water in excess of 50 ppb is detrimental to human health. Long term ingestion of arsenic contaminated water may cause hyper-pigmentation, keratosis on hand palms and soles of feet, skin cancer, peripheral vascular disease like black foot disease, etc. Concentration of arsenic content above permissible limit (50 ppb) in water as per WHO's pre-1993 guideline (which has been reduced to 10 ppb in 1993) has been detected in eight districts of West Bengal, India. These eight districts span a length of about 400 km and a width of about 60 km along a linear stretch of the upper deltaic alluvial plain of the Ganga-Bhagirathi river system. Around 560 villages in more than 50 blocks within these districts have been affected by arsenic poisoning. The estimated population in these eight districts is around 40 million (population survey, 2006), within which the estimated population using high arsenic contaminated water (above 50 ppb) is more than 1 million, while the estimated population using moderate arsenic contaminated water (between 10 and 50 ppb) is around 1.3 million.

* Corresponding author. Tel.: +91 20 22933441; fax: +91 20 22934191.
E-mail address: samird@isquareit.ac.in (S. K. Das).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.200800132

2. Study Area

The study area forms a part of the fluviatile sedimentary area of the upper deltaic plain of the Ganga-Bhagirathi-Mahananda river system, exhibiting flat to gently undulating topography with a regional slope towards south. Geomorphologically, the area is represented by three alluvial terraces: present day flood plain terraces, relatively older terraces, and the oldest terraces. The present day flood plain terraces (comprising silver white sand, silt and clay) are followed by the older terraces (represented by meander belts, cut off channels, and swampy area), which in turn are followed up by the oldest terraces. The successive terraces are around one meter high from each other. The geological formation comprises several cycles of sand, silt, and clay with carbonaceous matters, and at some places, with carbonate and ferruginous concretions. The areas having Arsenic traces occur mostly within the older terraces whereas the areas having arsenic content beyond the permissible limit (> 50 ppb) occur within a depth of 10 to 30 meters. The wide variation or heterogeneity of arsenic concentration at the same depth in areas that are only few meters apart poses problem in delineating its distribution pattern.

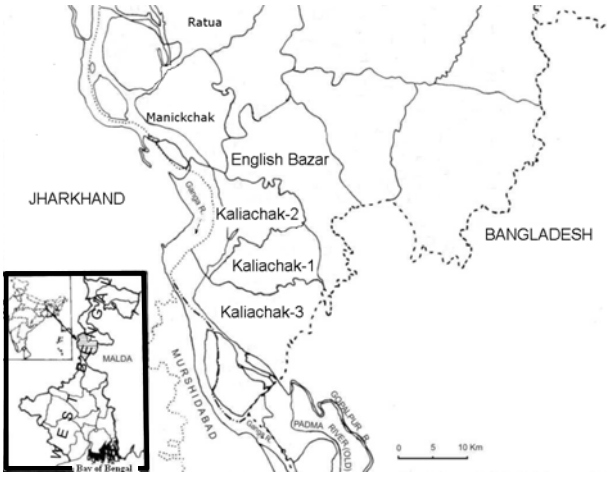


Figure 1. Map and blocks of Malda district in west Bengal, India.

3. Methodology and Data Collection

Groundwater samples were collected during post-monsoon period (Nov. to Mar.) after ten minutes pumping of each tube well. A total of 85 water samples were collected from different blocks [Kaliachak blocks 1, 2 and 3; English (Ingraj) Bazar block; Manickchak block; Old Malda block and parts of Ratua blocks 1 and 2] covering an area of around 1000 km^2. of the Malda district in West Bengal, India (Figure 1). The arsenic content of the water was measured with the E-Merck arsenic determination kit at each tube well site. Parallel measurements were also carried out in the laboratory in order to compare the field kit measurement data with the results obtained by laboratory measurements. For laboratory measurements, each time around one quarter litre of water was collected in a cleaned polyethylene bag and acidified with supra-pure HCl (1 mL in 100 mL of water) and sent to chemical laboratory for chemical analysis.

In this research work Artificial Neural Network (ANN) technique has been applied to estimate the arsenic content in groundwater based on some geochemical parameters. For the past several years, ANNs are being popularly applied as efficient mathematical tools to represent complex relationships in many branches of hydrology (Maier et. al., 2000; Morshed et al., 1998; Poff et al., 1996; Rogers et al., 1994; Rogers et al., 1995; Smith et al., 1997; Zhu et al. 1994). ANN's flexible structure can provide good estimation to various problems in hydrology such as water quality modeling, stream flow forecasting, groundwater modeling and precipitation forecasting, etc. (Carrera et al., 1988; Clair et al., 1996; Coulibaly et al., 2001). Yeh (1986) has reported various techniques to solve inverse problem of parameter evaluation in groundwater. In particular, the back propagation algorithm, as a theoretical framework, has led to its wide application in various civil engineering problems (ASCE, 2000).

In the present study, we use data samples collected from different arsenic-affected locations of Malda district, West Bengal, India, in order to train the ANN for estimating the arsenic contaminant in ground water. The data sample for each location comprised values for seven predefined geochemical parameters, namely, pH (Acidic-alkalinity ratio), Sp. Cond. (specific conductivity), TDS (total dissolved solids), Salinity, DO (dissolved oxygen), Eh (redox potential), and Depth of the tube well, together with the amount of arsenic contaminant observed in the groundwater. Hence, the inputs to the ANN were observed data values for the seven geochemical parameters, while the expected output for the ANN (i.e., expected output while training the ANN) was the amount of corresponding arsenic contaminant observed in the groundwater. The main objective of the present study is to obtain realistic ANN simulations and to estimate the extent of arsenic contaminant at various depths using the trained ANN.

3.1. Back Propagation Algorithm in ANN

The formal algorithm to train the back propagation neural network model is illustrated below and is based on the work of Hechst-Nielsen (1990) and Simpson (1990). An implementation of this algorithm can be found in Demluth et al. (2000) and Masters (1993), while the schematic representing the architecture of the back propagation neural network model can be found in Hechst-Nielsen (1990) and Rumelhart et al. (1986).

1. Randomize the network weights in the range [-1, 1].

2. For each pattern $(X_k, Y_k) = \left\langle \left(x_1^k, x_2^k, ..., x_m^k\right), \left(Y_1^k\right) \right\rangle$ $k = 1, 2, ..., T$.

i. Present input pattern to processing elements (PEs) in the input layer.

ii. Compute new output values of PEs in the two hidden layers and one output layer using:

$$a_i = f_T\left(\sum_{h=1}^{m} u_{ih} x_h + \mu_i\right) \quad for\ i = 1,\ 2,\ ...,\ p \tag{1}$$

$$b_j = f_T\left(\sum_{i=1}^{p} v_{ji}a_i + \theta_j\right) \quad for\ j = 1,\ 2,\ ...,\ q \tag{2}$$

$$y_1^k = f_L\left(\sum_{j=1}^{q} w_{1j}b_j + \tau_1\right) \tag{3}$$

where x_h represents the output of m PEs in the input layer, a_i, b_j denote the output of p and q PEs in the two hidden layers, respectively, y_1^k represents the output of the single PE in output layer, and μ_i, θ_j and τ_1 are threshold or bias values of the PEs. The variables u_{ih}, v_{ji}, and w_{1j} represent network weights between input and hidden1, hidden1 and hidden2, and hidden2 to output layer, respectively. The functions f_L and f_T represent variations of the generic sigmoid transfer function $f(x) = 1/(1+e^{-x})$. In the two hidden layers we have used 'tan-sigmoid' transfer function (f_T), while in the output layer we have used a 'log-sigmoid' transfer function (f_L), the schematic of which is shown in Figure 2 (Demluth and Beale, 2000).

iii. Estimate the "error term" between the computed and desired output values of PEs in the output and hidden layers using:

$$d_1 = y_1^k(1 - y_1^k)(y_1^k - Y_1^k) \tag{4}$$

$$e_j = \left(1 - b_j^2\right)w_{1j}d_1 \quad for\ j = 1,\ 2,\ ...,\ q \tag{5}$$

$$f_i = (1 - a_i^2)\sum_{j=1}^{q} v_{ji}e_j \quad for\ i = 1,\ 2,\ ...,\ p \tag{6}$$

where d_1, e_j, and f_i are the errors terms in the output and hidden layer PEs, respectively. Note that the derivative of log-sigmoid function f_L is $f_L(1 - f_L)$, while the derivative of tan-sigmoid function f_T is $(1 - f_L^2)$, which are used in computing the error terms in Equations 4 to 6.

3. Update the threshold values (i.e., μ_i, θ_j, *and* τ_1 bias values at each PE):

$$\delta\tau_1 = \alpha\ d_1 \tag{7}$$

$$\delta\theta_j = \beta\ e_j \tag{8}$$

$$\delta\mu_i = \gamma\ f_i \tag{9}$$

where α, β, and γ are the learning rates.

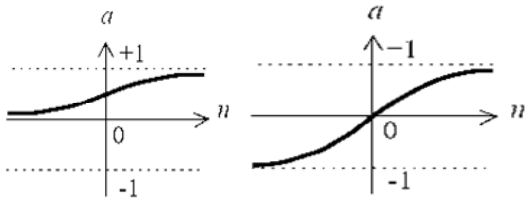
Repeat Steps 2 to 3 until the total mean-squared error (the mean of the squared differences between the target (Y_1^k) and the actual output (y_1^k) of the PE in output layer taken over all T training samples) is sufficiently low, or when the training process has reached the maximum number of epochs.

The error between the target (expected output) and the actual output of the PE in output layer for the k*th* training sample is:

$$E_1^k = y_1^k - Y_1^k \tag{10}$$

The total mean-squared error taken over all T training samples is:

$$MSE = \frac{1}{T}\sum_{k=1}^{T}\left[\frac{1}{2}\cdot\left(y_1^k - Y_1^k\right)^2\right] \tag{11}$$



a=logsig(n) *a=tansig(n)*

Figure 2. Transfer functions used in the MLP.

This algorithm performs a gradient descent technique to obtain global minimum of the mean-squared error (*MSE*) along the steepest vector of the error surface defined by Equation 11. As the error surface could be hyper-paraboloid in nature, but rarely smooth, the solution space contains irregular solution vectors, which may sometimes cause the network to settle down in a local minimum. The learning procedure attempts to modify the network weights towards global minima (i.e., towards finding the minimum value of *MSE* given by Equation 11 for all *T* training samples) using built-in mathematical terms to control the speed (learning parameter) and the momentum.

A large number of individual runs were taken to determine the best possible solution, since the nature of the error space could not be determined a priori. Taking large number of runs, each starting with a different set of random weights, increases the probability of finding global minima on the error surface.

3.2. Training and Testing with ANN

In the present study, we have employed a supervised back propagation neural network model comprising four layers (7-15-15-1) simulated through the MATLAB@ neural network toolbox (Demluth and Beale, 2000). The '7-15-15-1' ANN model has seven neurons in the input layer (for the seven geochemical parameters), fifteen neurons each in the two hidden layers, and one neuron (for predicting the arsenic content) in the output layer. The two hidden layers neurons used 'tan-sigmoid' transfer function, while the output layer neuron used a 'log-sigmoid' transfer function. We have used two hidden layers in order to increase the learning capability and the generalization ability of the network. Also, it has been observed that an increase in the number of neurons in the hidden layer decreases the RMS error within the neurons (Karri and Frost, 1999).

Table 1. Summary Statistics of the Geochemical Parameters of Groundwater in Malda District (No. of Samples: 85)

Parameter	Unit	Range		Average
pH	-	6.72	8.24	7.1437647
SC*	μ S/cm	437	1450	7.1437647
TDS	mg/l	270	900	530.76471
Salinity	Ppt	0	0.5	0.1905882
DO	mg/l	1.4	9.3	2.6776471
Eh	mV	-140	286	-10.964706
Depth	Meter	9.15	45.75	21.116765
As	Ppb	0	800	187.76471

* specific conductivity

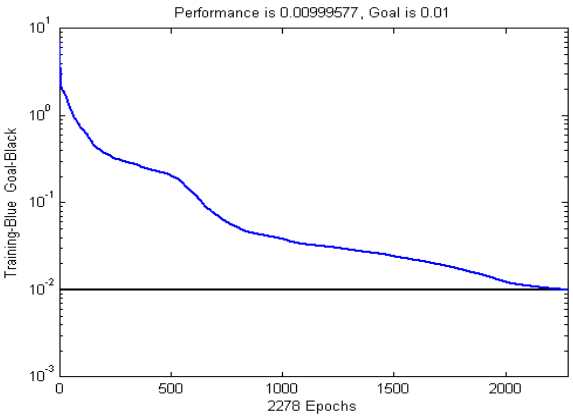


Figure 3. Convergence of the training curve for back propagation network.

In the training stage, a set of training data with input examples and their corresponding desired outputs was prepared. The network weights, which eventually store the learned patterns, were initially set to random values. During the learning process, the example inputs were given to the network, and the computed and the expected outputs were compared. An error term was then calculated, which was used to change the network weights in an iterative manner (Simpson, 1990). The processing of information was done by the log-sigmoid transfer function in the output layer. The log-sigmoid function acts as a squashing function that generates outputs between 0 and 1, irrespective of the magnitude of the neuron's input, which may go from negative to positive infinity. The advantage of using a log-sigmoid function over a threshold-type function is that the former function is continuous and differentiable, which allows the gradient of the error to be used while updating the connection weights.

The input data to the ANN consisted of data samples collected from different blocks in the Malda district. Each data sample comprised measured values for different geochemical parameters of the ground water such as pH, Specific conductivity, TDS, Salinity, DO, Eh, Depth and the arsenic content (Table 1). The data samples were divided into two sets: one set (called the training set) was used for training the ANN model, while the other set was used for testing the trained ANN. The training set comprised 60 data samples, while the testing set consisted of 25 data samples (Table 2). The data values within each parameter category were normalized to unity at their maximum values.

For measuring the error between the ANN predicted arsenic contaminant values and the corresponding values observed from different locations, we have used three different error measures, namely, the root mean square error (RMSE), the mean absolute error (MAE), and the percent mean relative error (PMRE), the expressions for which are as given below:

$$RMSE = \sqrt{\sum_{j=1}^{N}\left(y_1^k - Y_1^k\right)^2 / N}$$

$$MAE = \sum_{j=1}^{N}\left|\left(y_1^k - Y_1^k\right)\right| / N \qquad (12)$$

$$PMRE = \frac{100}{M}\cdot\sum_{j=1}^{N}\left|\left(y_1^k - Y_1^k\right)\right| / Y_1^k$$

where y_1^k represents the ANN computed arsenic contaminant value, Y_1^k represents the corresponding value observed in a particular location, N is the total number of contaminant values used in prediction, and M is the total number of non-zero ($Y_1^k \neq 0$) contaminant values used in the same prediction.

The RMSE and MAE are measured in the same units as data and are therefore easy to understand. Both represent the size of a typical error between the contaminant values predicted by the neural network and the corresponding values observed in the selected region (s). However, most researchers prefer unit-free measures for comparing methods (Armstrong and Collopy, 1992). The PMRE, unlike RMSE and MAE, is a unit-free measure that gives percent mean relative error between the computed and expected values. Moreover, the RMSE has low reliability (Chatfield, 1988), nevertheless, it has been widely used for comparing forecasting methods (Armstrong and Collopy, 1992). The three error measures were taken in order to analyze the comparative results from different perspective, take care of sensitivities due to small changes in large number of data series, and draw reliable and proper conclusions.

The learning curve in Figure 3 shows how the total sum-squared error (SSE) converges in the process of iterative learning of the ANN. By total sum-squared error we mean $\sum_k(y_1^k - Y_1^k)$, where k ranges over all the T training patterns (Y_1^k is the expected or observed output, while y_1^k is the corresponding output computed by the network), which is similar to Equation 11 except for the constant terms. The learning process terminates either after a certain number of runs through all the training data (each cyclic run through all the training data is called an epoch), or when the total sum-squared error reaches some predefined target value or goal. In this study, we fixed the target SSE (Goal) to 0.01, and the maximum number of epochs to 5000. The number of epochs required in achieving the target SSE of 0.01 while training the ANN was 2278. Note

Table 2. Values of the Geochemical Parameters of Groundwater in Malda District (No. of Samples: 85)

Sample no.	pH	SC (μ S/cm)	TDS (mg/lit)	Salinity (ppt)	DO (mg/lit)	Eh (mV)	As (ppb)	Depth (m)
Training Samples (60)								
9 K3	7.01	792	498	0.2	2	107	210	22.875
96 P	7.03	680	421	0.1	2.2	-70	<50	21.35
91 P	7.2	616	380	0	2.9	38	100	21.35
8 K3	7.06	894	526	0.2	2.1	-26	250	21.35
89 P	6.72	1354	837	0.5	2.8	70	<50	38.125
88 P	6.88	437	270	0	2.5	-35	70	38.125
85 K3	7.06	1450	900	0.5	2.7	-129	150	21.35
84 K3	7.05	1097	679	0.3	2.6	-67	150	21.35
83 K3	7.18	832	515	0.2	2.8	-11	60	21.35
82 K3	6.91	1336	627	0.5	2.5	158	<50	9.15
81K3	7.1	1123	695	0.3	2.6	84	<50	16.775
80 K3	7.32	497	308	0	2.4	-93	50	21.35
7 K3	7.04	940	590	0.2	1.8	-55	800	21.35
79 K3	7.36	543	336	0	2.3	-100	130	15.25
78 K3	7.2	1078	667	0.2	2.1	-60	100	41.175
77 K3	7.16	819	506	0.2	2.7	-52	80	35.075
76 K3	7.17	795	492	0.2	2.2	-85	70	22.875
75 K3	7.2	609	377	0	3.1	-12	<50	15.25
74 K3	7.19	744	461	0.1	2.4	52	<50	21.35
73 K1	7.12	770	477	0.1	2	-37	200	18.3
72 K1	7.19	633	392	0.1	2.1	-140	250	25.925
71K1	7.12	940	582	0.2	2.1	-37	100	18.3
70 K1	7.15	658	414	0.1	2.5	-64	400	24.4
6 K3	7.1	701	441	0.1	2.9	47	0	27.45
69 K1	7.13	732	460	0.1	2.4	-9	<50	21.35
68 K1	7.09	995	626	0.3	2.3	-21	400	30.5
67 K1	7.02	1160	729	0.4	3.1	-84	200	21.35
66 K1	6.96	1398	881	0.5	2.8	-114	200	30.5
65 K1	7.31	724	455	0.1	2.5	-21	<50	21.35
64 K1	7.35	634	399	0.1	2.4	-35	60	15.25
63 K1	7.26	853	536	0.2	2.3	-80	130	15.25
62 K1	7.4	1554	977	0.6	2.5	-84	500	21.35
61K1	7.21	825	519	0.2	3	-113	400	21.35
60 K1	7.17	857	539	0.2	3.5	-90	250	27.45
5 K1	7.09	858	540	0.2	2.7	167	500	21.35
59 K1	7.14	738	464	0.1	2.2	-136	250	21.35
58 K1	7.05	1000	629	0.3	2.5	-123	150	21.35
57 K1	6.98	1111	699	0.3	2.6	-120	400	15.25

that many such runs were performed and the best run giving the minimum deviation in arsenic predication was taken.

The performance of the ANN method was judged by computing its learning accuracy followed by the predicting capability of the trained ANN. The graphs in Figure 4 and Figure 5, and the entries in Table 3 depict the learning accuracy of the ANN. The ANN was trained with data samples from the training set. On completion of the training process, same samples from the training set were used to predict the arsenic content in the groundwater. As seen from Figure 4, the values predicted by the trained ANN closely resemble the corresponding values observed from different ground locations. The RMSE, MAE, and MRE between the ANN predicted and the observed arsenic content values were 18.70, 4.77, and 4.27, respectively. For example, all the patterns correlating arsenic content at different location were learnt with a mean relative error (MRE) of less than 5 % (between the observed and the ANN learnt arsenic content values).

From Figure 4, the correlation coefficient R^2 is 0.9519 between the observed and ANN predicted arsenic values for the training samples, which indicates good learning capability of the ANN model. Since the ANN learned input patterns very well, increasing the number of geochemical parameters that have some correlation with the amount of arsenic content found in groundwater could further improve the prediction capability of ANN.

In order to measure the prediction capability of the ANN, the trained ANN was supplied new data samples (samples from

Table 2. Continued

Sample no.	pH	SC (μ S/cm)	TDS (mg/lit)	Salinity (ppt)	DO (mg/lit)	Eh (mV)	As (ppb)	Depth (m)
56 K2	7.07	579	364	0	2.6	-101	210	21.35
55 K2	7.09	760	478	0.1	2.8	-101	250	21.35
54 K2	7.1	829	521	0.2	2.6	-67	300	15.25
53 K2	7.02	986	620	0.3	2.4	-116	200	18.3
52 K2	7.05	1153	725	0.4	2.5	-65	450	21.35
51K2	7.15	825	517	0.2	3.4	-119	200	15.25
50K2	7.14	927	582	0.2	2.3	-37	80	15.25
4 K1	7.13	945	595	0.2	2.3	99	250	15.25
49 K2	7.13	791	498	0.2	2.6	-80	350	21.35
48 K2	7.36	670	421	0.1	2.8	-3	<50	21.35
47K1	7.25	815	512	0.3	2.7	12	<50	27.45
46 K1	7.06	1063	669	0.3	2.4	-80	70	15.25
45 K1	7.1	1004	632	0.3	2.3	-120	200	18.3
44 K1	7.36	1019	641	0.3	2.5	-54	60	27.45
43 E	8.24	494	309	0	9.3	184	100	15.25
42 R2	7.38	503	316	0	2.1	-53	<50	24.4
41 R1	7.02	735	462	0.1	2.3	-120	80	18.3
40 R1	6.9	1410	886	0.5	2.2	64	<50	15.25
3 K1	6.91	1326	835	0.5	3	286	300	18.3
39 M	7.38	769	483	0.1	3.1	124	<50	21.35
38 M	7.04	1358	854	0.5	3.5	182	<50	21.35
37 M	6.98	952	598	0.2	2.6	160	<50	24.4
Testing Samples (25)								
35 E	7.13	546	343	0	2.5	-114	130	15.25
34 E	6.97	903	568	0.2	2.4	55	<50	15.25
33 E	7.12	782	492	0.1	2.7	25	<50	21.35
32 E	6.97	938	589	0.2	2.9	106	<50	30.5
31 E	6.92	866	545	0.2	2.1	160	<50	21.35
29M	7.08	783	493	0.1	3	-82	500	21.35
28 M	6.99	1356	853	0.5	2.3	-54	350	15.25
27 M	7.2	686	431	0.1	2	-122	500	21.35
26M	7.13	744	468	0.1	2.5	-114	500	21.35
25 M	7.27	599	376	0	2	-70	100	21.35
24 M	7.45	676	424	0.1	3.6	165	200	21.35
23 M	7.07	837	526	0.2	3.2	-117	600	27.45
22 M	7.2	772	454	0.1	1.9	-83	650	15.25
21 M	7.19	703	442	0.1	2.4	160	60	45.75
20 M	7.2	635	400	0.1	3.2	-14	500	21.35
19 M	7.1	784	493	0.1	2	-71	90	15.25
18 E	7.09	770	485	0.1	1.4	-87	800	21.35
17 E	7.56	553	347	0	5.7	197	60	15.25
16 E	7.16	949	596	0.2	2.4	146	70	21.35
15 E	7.12	833	521	0.2	4.1	153	60	30.5
14 E	7.38	564	354	0	5.5	161	80	27.45
13 E	6.92	696	437	0.1	1.6	57	<50	36.6
12 E	7.13	785	493	0.1	2.9	-67	200	21.35
11 E	7.22	789	497	0.1	1.7	19	100	15.25
10 E	7.11	992	624	0.3	2.2	44	200	21.35

the test data set that it had not seen before) for computing the arsenic content at different locations. The graphs in Figure 6 and Figure 7, and the entries in Table 4 depict the prediction accuracy of the trained ANN model. The line graphs in Figure 6 depict arsenic values computed by the trained ANN and those obtained during actual field trial. Although both curves do not match as closely as those in Figure 4, the trained ANN does show a close trend towards correct predicting of the arsenic values. Table 4 displays the actual values of the arsenic content as predicted by the trained ANN and those observed during the field trial. Note that some entries in Table 4 depict values of arsenic content to be less that 50 ppb (< 50) at certain locations. The ANN was trained with an average value of 25 ppb for all such entries that were present in the training set. From Table 4, it is observed that around 68% arsenic content values were predicted reasonably well with less that 50% error. The RMSE, MAE, and MRE between the ANN predicted and the observed arsenic content values were 184.11, 114.29, and 42.90, respectively, with a reasonably good agreement having correlation coefficient R^2 as 0.6672 (Figure 7).

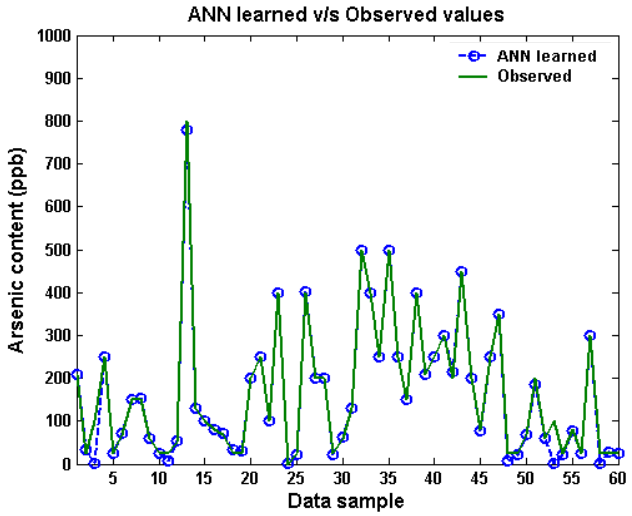


Figure 4. Patterns correlating arsenic content at different locations learned by the ANN.

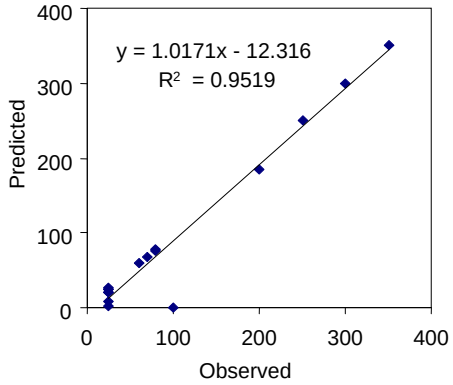


Figure 5. Relation between observed & ANN predicted arsenic values for the training data.

Table 3. Learned Sample Prediction by the ANN

Observed	ANN	AE*	% RE
210.00	210.10	0.10	0.05
25.00	32.68	7.68	30.70
100.00	0.11	99.89	99.89
250.00	250.08	0.08	0.03
25.00	24.84	0.16	0.63
70.00	70.20	0.20	0.28
150.00	151.26	1.26	0.84
150.00	152.39	2.39	1.59
60.00	58.91	1.09	1.82
25.00	24.54	0.46	1.83
25.00	6.20	18.80	75.19
50.00	55.13	5.13	10.26
800.00	778.26	21.74	2.72
130.00	128.72	1.28	0.98
100.00	100.84	0.84	0.84
80.00	80.00	0.00	0.00
70.00	73.00	3.00	4.28
25.00	33.53	8.53	34.11
25.00	29.65	4.65	18.60
200.00	199.60	0.40	0.20
250.00	249.74	0.26	0.10
100.00	100.92	0.92	0.92
400.00	398.99	1.01	0.25
0	0.24	0.24	0
25.00	23.24	1.76	7.05
400.00	400.63	0.63	0.16
200.00	200.33	0.33	0.16
25.00	23.17	1.83	7.33
60.00	61.90	1.90	3.17
130.00	129.95	0.05	0.04
500.00	499.94	0.06	0.01
400.00	400.17	0.17	0.04
250.00	250.12	0.12	0.05
500.00	499.69	0.31	0.06
250.00	249.88	0.12	0.05
150.00	149.93	0.07	0.05
400.00	398.18	1.82	0.46
210.00	209.18	0.82	0.39
250.00	250.25	0.25	0.10
300.00	300.27	0.27	0.09
200.00	215.46	15.46	7.73
450.00	450.09	0.09	0.02
200.00	199.97	0.03	0.02
80.00	76.61	3.39	4.24
250.00	251.19	1.19	0.48
350.00	349.88	0.12	0.04
25.00	8.62	16.38	65.54
25.00	20.84	4.16	16.64
70.00	68.37	1.63	2.32
200.00	185.27	14.73	7.36
60.00	60.00	0.00	0.01
100.00	0.01	99.99	99.99
25.00	20.57	4.43	17.73
80.00	76.98	3.02	3.78
25.00	23.60	1.40	5.59
300.00	299.93	0.07	0.02
25.00	1.74	23.26	93.03
25.00	27.37	2.37	9.50
25.00	25.44	0.44	1.76

AE: Absolute error; RE: Relative error.

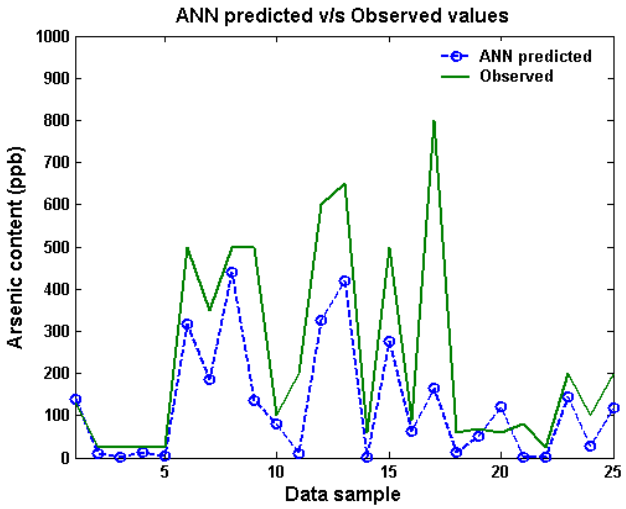


Figure 6. The arsenic content at new locations as predicted by the trained ANN.

Table 4. Observed Versus Predicted Arsenic Content (ppb) for 25 Test Data Samples

Observed	ANN	AE*	% RE**
130.00	138.47	8.47	6.51
(< 50) 25.00	(< 50) 9.26	0.0	0.0
(< 50) 25.00	(< 50) 2.05	0.0	0.0
(< 50) 25.00	(< 50) 13.44	0.0	0.0
(< 50) 25.00	(< 50) 5.83	0.0	0.0
500.00	317.18	182.82	36.56
350.00	186.51	163.49	46.71
500.00	438.72	61.28	12.26
500.00	135.20	364.80	72.96
100.00	79.46	20.54	20.54
200.00	11.21	188.79	94.39
600.00	325.18	274.82	45.80
650.00	420.39	229.61	35.33
60.00	1.63	58.37	97.29
500.00	275.89	224.11	44.82
90.00	63.21	26.79	29.77
800.00	164.63	635.37	79.42
60.00	12.30	47.70	79.50
70.00	52.27	17.73	25.33
60.00	122.68	62.68	104.47
80.00	0.78	79.22	99.03
(< 50) 25.00	(< 50) 1.36	0.0	0.0
200.00	145.26	54.74	27.37
100.00	26.94	73.06	73.06
200.00	117.14	82.86	41.43
Mean	-	114.29	42.90

*Absolute error; **Relative error.

4. Arsenic Contaminant Prediction through Other Prediction Methods

In this section we present two other methods for arsenic content prediction, namely, multiple linear regression and active set support vector regression and compare their prediction results with those of the ANN model.

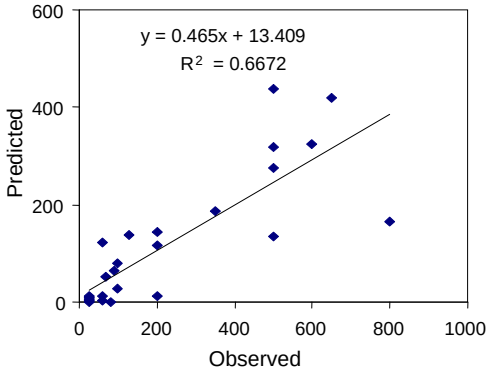


Figure 7. Relation between observed & ANN predicted arsenic values for the test data.

4.1. Prediction using Multiple Linear Regression (MLR)

Multiple linear regression attempts to model the relationship between two or more independent variables (x_i) and a dependent variable (y) by fitting a linear equation to the observed data (Kutner et. al, 2004). In this research work, the independent variable y represents the arsenic content, while the dependant variables x_i (i = 1, …, 7), represent the seven geochemical parameters as stated above. The multiple linear regression equation computed using the same data samples as in the training set of the ANN model is as follows:

$$y = -9.07\,x_1 - 0.58\,x_2 + 0.93\,x_3 + 245.88\,x_4 - 1.48\,x_5 - 0.33x_6 + 0.09\,x_7 + 172.37 \quad (13)$$

where, the independent variable y represents the arsenic content, and the dependant variables x_i, i = 1, …, 7, represent the seven geochemical parameters such as pH, specific conductivity, total dissolved solids (TDS), salinity, dissolved oxygen (DO), redox potential (Eh), and depth of the tube well water, respectively. Using Equation 13 and the same test data samples as in the ANN model, we predict the arsenic contaminant values for the test data. The predicted values of the arsenic content using multiple linear regression model are displayed in Table 5.

4.2. Prediction using Active Set Support Vector Regression (ASVR)

In recent years, the use of support vector machines (SVMs) on various classification and regression problems has become increasingly popular. SVMs can be applied to both classification (Cristianini and Taylor, 2000) and regression problems (Kutner et al., 2004). In the classification case, we try to find an optimal hyper plane that separates two classes. In regression, the goal is to estimate an unknown continuous-valued function based on a finite number set of noisy samples. The SVM algorithm is a nonlinear generalization of the generalized portrait algorithm (Vapnik and Lerner, 1963). Active set support vector regression is an active set strategy to solve a reformulation of the standard support vector regression pro-

Table 5. Observed Versus Predicted Arsenic Contaminant Values for 25 Test Data Samples Using Three Methods

Observation (ppb)	ANN prediction (ppb)	MLR prediction (ppb)	ASVR prediction (ppb)	Absolute error (ppb)			Relative error (%)		
				ANN	MLR	ASVR	ANN	MLR	ASVR
130.00	138.47	147.85	148.43	8.47	17.85	18.43	6.51	13.73	14.17
25.00 (< 50)	9.26	145.60	146.26	0.0	120.60	121.26	0.0	482.39	485.02
25.00 (< 50)	2.05	130.44	140.94	0.0	105.44	115.94	0.0	421.75	463.76
25.00 (< 50)	13.44	132.21	140.89	0.0	107.21	115.89	0.0	428.83	463.57
25.00 (< 50)	5.83	114.30	110.64	0.0	89.30	85.64	0.0	357.21	342.56
500.00	317.18	165.54	175.71	182.82	334.46	324.29	36.56	66.89	64.86
350.00	186.51	256.85	250.49	163.49	93.15	99.51	46.71	26.62	28.43
500.00	438.72	177.60	176.65	61.28	322.40	323.35	12.26	64.48	64.67
500.00	135.20	175.63	181.41	364.80	324.37	318.59	72.96	64.87	63.72
100.00	79.46	134.81	145.81	20.54	34.81	45.81	20.54	34.81	45.81
200.00	11.21	78.85	79.06	188.79	121.15	120.95	94.39	60.57	60.47
600.00	325.18	202.55	198.50	274.82	397.45	401.50	45.80	66.24	66.92
650.00	420.39	134.61	144.99	229.61	515.39	505.01	35.33	79.29	77.69
60.00	1.63	93.25	99.51	58.37	33.25	39.51	97.29	55.42	65.86
500.00	275.89	141.46	133.16	224.11	358.54	366.84	44.82	71.71	73.37
90.00	63.21	160.77	170.21	26.79	70.77	80.21	29.77	78.64	89.12
800.00	164.63	169.54	178.50	635.37	630.46	621.50	79.42	78.81	77.69
60.00	12.30	37.66	43.58	47.70	22.34	16.42	79.50	37.23	27.36
70.00	52.27	111.69	119.54	17.73	41.69	49.54	25.33	59.56	70.76
60.00	122.68	111.50	107.75	62.68	51.50	47.75	104.47	85.83	79.58
80.00	0.78	55.22	63.14	79.22	24.78	16.86	99.03	30.97	21.07
25.00 (< 50)	1.36	126.98	127.04	0.0	101.98	102.04	0.0	407.92	408.16
200.00	145.26	159.19	170.34	54.74	40.81	29.66	27.37	20.41	14.83
100.00	26.94	131.65	143.24	73.06	31.65	43.24	73.06	31.65	43.24
200.00	117.14	175.09	168.22	82.86	24.91	31.78	41.43	12.45	15.89
Mean				114.29	160.65	161.66	42.90	125.53	129.14

blem. The algorithm consists of solving a finite number of linear equations with a dimensionality equal to the number of training data samples to be approximated (Musicant, 2004). We have used the ASVR program as given in (Musicant and Feinberg, 2002) to approximate the training data with regression surface and used this to predict the arsenic content. The predicted values of the arsenic content using ASVR model are displayed in Table 5.

4.3. Comparison of the Three Predictions Methods

The entries in Table 5 depict predicted arsenic contaminant values and the corresponding prediction errors for the three methods, while Figure 8 displays the comparative line graphs of the prediction values. The RMSE, MAE, and MRE between the predicted and the observed arsenic contaminant values computed by the three methods are displayed in Table 6. Both MLR and ASVR show identical prediction values, while the prediction results of the ANN are relatively better since the ANN model can model the non-linearity in the data better than the other two models. Note that the prediction curves of both MLR and ASVR prediction methods in Figure 8 almost overlap each other. The RMSE, MAE, and MRE values for the MLR and ASVR methods are also identical, while for the ANN model they have relatively lower values.

Although it has been claimed that the support vector machines regression (SVR) technique provides a very attractive alternative to the current optimization methods (such as artificial neural networks) used in the inversion problems (Durbha et al., 2007), the ASVR, a variation of the SVR, did not provide expected prediction results in our study.

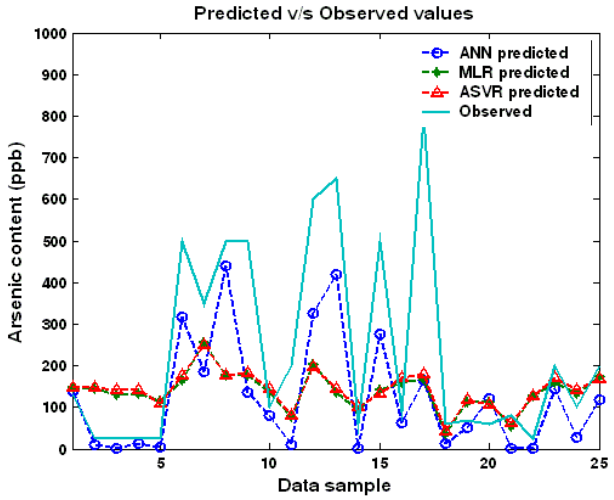


Figure 8. Arsenic contaminant values predicted by the three methods.

Table 6. Error Measures between the Predicted and the Observed Arsenic Contaminant Values Computed by the Three Methods

Method	RMSE	MAE	MRE
ANN	184.11	114.29	42.90
MLR	233.72	160.75	125.53
ASVR	232.75	161.66	129.14

5. Conclusions

An ANN-based methodology to determine the amount of arsenic contaminant in groundwater was presented and tested using the actual field data. The methodology is based on the idea that although the amount of arsenic contaminant spreads across different locations changes over time due to various factors, the variation of contaminant values is interrelated and can be well estimated through ANN simulations. A four-layer (7-15-15-1) feed-forward back propagation ANN with a nonlinear differentiable tan-sigmoid and log-sigmoid transfer function in the hidden layers and the output layer, respectively, and a variable learning rate has proved to be useful than the traditional modeling of arsenic contaminant in groundwater. The ANN model learned the patterns used for predicting the arsenic content very well. It could accurately compute the amount of arsenic content for the data samples that it had learnt. For new locations, the prediction of arsenic content using the ANN model and the amount of arsenic actually observed at such locations during the field trial showed acceptable agreement. The paper also presented two other methods for arsenic content prediction; namely, multiple linear regression and active set support vector regression. Amongst the three methods, the ANN model exhibited better prediction results for predicting the arsenic contamination in groundwater.

Acknowledgments. The first author is grateful to the Deputy Director General, Eastern Region, Geological Survey of India (GSI) for providing the infra-structural facilities during field investigation work as per approved field season program of the GSI, and for permission to publish this research work. Authors are also very much thankful to the referees for theirs valuable suggestions and comments.

References

Armstrong, J.S., and Collopy, F. (1992). Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons, *Int. J. Forecast.*, 8, 69-80.

ASCE. (2000). Task Committee on Application of Artificial Neural Networks in Hydrology, Artificial neural networks in hydrology, II: Hydrologic applications, *J. Hydrol. Eng.*, 5, 124-137.

Carrera, J. (1988). *State of the art of the inverse problem applied to the flow and solute transport equations*, Ground water flow and quality modeling, D. Reidel Publishing Corporation.

Chatfield, C., (1988). Apples, oranges and mean square error, *Int. J. Forecast.*, 4, 515-518.

Clair, T.A., and Ehrman, M. (1996). Variations in discharge and dissolved organic carbon and nitrogen export from terrestrial basins with changes in climate: A neural network approach, *Limnol. Oceanogr.*, 41, 921-927.

Coulibaly, P., F. Anctil, R. Aravena, and Bobee, B. (2001). Artificial neural network modeling of water table depth fluctuations, *Water Resour. Res.*, 37, 885-896.

Cristianini N., Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press.

Demluth, H., and Beale, M. (2000). *Neural network toolbox for use with MATLAB*, Users Guide Version 3, The MathWorks, Inc, Natic, Maine.

Durbha, S.S., King, R.L., and Younan, N.H. (2007). Support vector machines regression for retrieval of leaf area index from multi-angle imaging spectroradiometer, *Remote Sens. Environ.*, 107, 348-361, doi:10.1016/j.rse.2006.09.031.

Hechst-Nielsen, R. (1990). *Neurocomputing*, Addison-Wesley Publishing Company, Boston, MA, USA.

Karri, V., and Frost, F. (1999). Optimum back propagation network conditions with respect to computation time and output accuracy, *in the Proceedings of third international conference on Computational Intelligence and Multimedia Applications (ICCIMA '99)*, New Delhi, India, 50-54.

Kutner, M.H., Neter, J., Nachtsheim, C., Nachtsheim, C.J., and Neter, J. (2004). *Applied Linear Regression Models*, McGraw-Hill Higher Education

Masters, T. (1993). *Practical neural network recipes in C++*, Academic Press, New York.

Maier, H.R., and Dandy, G.C. (2000). Neural networks for the prediction and forecasting water resources variables: a review of modeling issues and applications, *Environmental Modelling and Software*, 15, 101-124, doi:10.1016/S1364-8152(99)00007-9.

Morshed, J., and Kaluarachchi, J.J. (1998). Application of artificial neural network and generic algorithm in flow and transport simulations, *Adv. Water Resour.*, 22, 145-158, doi:10.1016/S0309-1708 (98)00002-5.

Musicant, D.R. (2004). Active set support vector regression, *IEEE Transactions on Neural Networks*, 15 (2), 268-275.

Musicant, D.R., and Feinberg, A. (2002). Active set support vector regression software, http://www.cs.carleton.edu/faculty/dmusican/asvr/.

Poff, N.L., Tokar, S., and Johnson, P. (1996). Stream hydrological and ecological responses to climate change assessed with an artificial neural network, *Limnol. Oceanogr.*, 41, 857-863.

Rogers, L.L., and Dowla, F.U. (1994). Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling, *Water Resour. Res.*, 30, 457-481.

Rogers, L.L., Dowla, F.U., and Johnson, V.M. (1995). Optimal field-scale groundwater remediation using neural networks and the genetic algorithm, *Environ. Sci. Technol.*, 29, 1145-1155.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning Internal Representations by Error Propagation, *in Parallel Distributed Processing*, Vol. 1, ed. Rumelhart, D. E. and J. L. McClelland, Cambridge, MA: MIT Press, pp. 318-362.

Simpson, P.K., (1990). *Artificial Neural Systems: foundations, paradigms, applications, and implementations*, Pergamon Press Inc.

Smith, J., and Eli, R.N. (1997). Neural-network models of rainfall-runoff processes, *J. Water Resour. Plann. Manage.*, 121, 499-508, doi:10.1061/(ASCE)0733-9496(1995)121:6(499).

Vapnik, V., and Lerner, A. (1963). Pattern recognition using generalized portrait method, *Automation and Remote Control*, 24, 774-780.

Yeh, W.W.G. (1986). Review of parameter identification procedures in groundwater hydrology: The inverse problem, *Water Resour. Res.*, 22, 95-108.

Zhu, M.L., Fujita, M., and Hashimoto, N. (1994). *Application of neural networks to runoff prediction, in Stochastic and statistical methods in hydrology and environmental engineering*, Kluwer Academic Publishers, Mass.