Solving the Inverse Problem in Stochastic Groundwater Modelling with Artificial Neural Networks

C. Rajanayaka, S. Samarasinghe and D. Kulasiri

Centre for Advanced Computational Solutions (C-fACS), Applied Management and Computing Division
Lincoln University, Canterbury, New Zealand (rajanayc@lincoln.ac.nz)

Abstract: In this paper, prediction capability of a hybrid Artificial Neural Networks (ANN) was investigated to solve the groundwater inverse problem. Initially, a Multi Layer Perceptron (MLP) network was developed and it was found that network produced better results when the target range of the parameters is smaller. Therefore, a Self-Organising Network (SON) was used to identify the objective subrange of the parameter and then the MLP model was employed to obtain final estimates. The data for the ANN was obtained from a numerical model that was utilised to simulate the solute transport in saturated groundwater flow. The forward problem of the numerical model was solved to generate solute concentration data for range of parameters. Those input data was fed into a MLP ANN to train the network along with corresponding parameter values. Sufficiently trained ANN model was used to estimate hydraulic conductivity (single parameter), and hydraulic conductivity and longitudinal dispersion coefficient (two parameters). First, the approach was tested on synthetic data to identify its feasibility and robustness. Then an experimental dataset that was obtained from an artificial aquifer was used to validate the method. It was found that ANNs produce accurate estimates in the presence of uncertainty. However, ANN are able to produce accurate results only if the pattern of the dataset that use to estimate parameters are similar to that of the training data. Therefore, it is important to adequately simulate the aquifer system in question by a large enough training dataset. However, due to the stochastic nature of the real world heterogeneous aquifers, it is not a trivial undertaking to identify the behaviour of the aquifer. Furthermore, as ANN's extrapolation capabilities beyond its calibration range is not reliable, it is necessary to set a calibration range sufficient to meet the limits of actual data. Therefore, prior information of the system is of utmost importance to obtain reasonably accurate estimates.

Keywords: Artificial Neural Networks; Inverse Problem; Groundwater; Uncertainty; Parameter

1. INTRODUCTION

Over the past decade, Artificial Neural Networks (ANN) have become increasingly popular in many disciplines as a problem solving tool. ANN have the ability to solve extremely complex problems with highly non-linear relationships. ANN's flexible structure is capable of approximating almost any input-output relationships. Particularly ANN have been extensively used as a predicting and forecasting tool in many disciplines.

Complex and heterogeneous hydrology systems are extremely difficult to model mathematically. However, it has been proved that ANN's flexible structure can provide simple and reasonable solutions to various problems in hydrology. Since the beginning of the last decade, ANNs have been successfully employed in hydrology research, such as rainfall-runoff modeling, stream flow forecasting, precipitation forecasting, groundwater modeling, water quality and management modeling [Morshed et al., 1998; ASCE Task Committee on Application of ANN in Hydrology, 2000; Maier et al., 2000].

ANN applications in groundwater problems are limited when compared to other in hydrology. Few such applications are as follows: Ranjithan et al. [1993] successfully used ANN to simulate pumping index for hydraulic conductivity realisation to remediate groundwater under uncertainty. A similar study has been conducted by Rogers et al. [1994] to simulate a regulatory index for a multiple pumping realization containing multiple plumes at a contaminated site. Rogers et al. [1995] took another step forward to simulate regulatory index, remedial index and cost index by using ANN for groundwater remediation. Coulibaly et al. [2001] modelled water table depth

fluctuations by using three types of functionally different ANN models.

1.1 Groundwater Inverse Problem

Subsurface contamination by an endless variety of organic compounds is widespread and it has been the subject of numerous studies. In these studies, we simulate or represent the interested system by a mathematical model (by excitation and response relationship) for forecasting and management problems. In the process of developing the models, we introduce the parameters, which we consider attributes or properties of the system. These values of the parameters are generally obtained from laboratory experiments and/or field scale experiments. However, such values may not represent the often complex patterns across a large geographic area, hence limiting the effectiveness of the model. In addition, such field scale experiments can be expensive. However, often we are interested in modelling quantities such as the depth of watertable and solute concentration. This is because they are directly relevant to environmental decision making, and we measure these variables regularly and relatively more cheaply. Further, we can continuously monitor these decision (output) variables in many situations. If the dynamics of the system can reliably be modelled, we can expect the parameters estimated based on the observations may give us more reliable representative values than those obtained from laboratory tests and literature. The reason for the inaccuracy of the laboratory tests happen due to the scale of the heterogeneity of the porous media is, most of the time, thinner than the scale of the flow and the transport model, hence, the parameter values obtained from laboratory tests are not directly usable in models, and generally need to be upscaled using difficult and often subjective techniques. However, there are a number of methods that have been developed for groundwater parameter estimation (see reviews such as Yeh, 1986). They range from primitive trial and error techniques that are very time consuming and whose solution strongly depends on the skills of the practitioner, to advance mathematical and geostatistical methods, such as the linearised cokriging approach [Kitanidis at el., 1983]. However, usage of the superior methodologies is limited due to their highly theoretical nature. Therefore, reliable, robust and easy to use methodologies need to be developed to be able to use by general practitioners to deal with detrimental groundwater contamination problems.

In this paper we used distributed contaminant concentration values of saturated groundwater flow to inversely estimate two hydraulic parameters namely hydraulic conductivity (K, m/day) and longitudinal dispersion coefficient (D_L, m^2/day) by means of ANN. First, we employed ANN to estimate a single parameter, K. Then extended the procedure to estimate two parameters. We used two types of dataset in this study. First dataset was generated synthetically by using a computer software package to simulate a 2-D groundwater aquifer. The second dataset was obtained from an artificial experimental aquifer. A hybrid approach that consists of a supervised Multi Layer Perceptron (MLP) ANN and self-organising network (SON) was employed to limit the permissible parameter range and to enhance the accuracy of estimates.

2. ESTIMATION OF PARAMETERS

We used a two-dimensional groundwater transport model to solve the inverse problem. In the first part of the study, we employed a deterministic 2-D advection-dispersion transport numerical model to generate synthetic data. Afterwards, ANN were trained to learn the complex excitation and response relationship of generated data. This was done by training the network sufficiently to minimise the error between the actual and network response while retaining generalising capabilities of the network. Then we estimated the associated parameters by using noisy concentration data that represents real world aquifer systems. We also tested the ability of the model to estimate hydraulic conductivity of an artificial experimental aquifer.

Before describing in detail the specific methodologies that were employed in the study, we briefly discuss the two dimensional advection-dispersion solute transport model that was used as the governing equation for this project. It may be important to mention that other possible phenomenon that can present in solute transport such as adsorption, the occurrence of short circuits were neglected in the governing equation on the assumption that the introduction of noise into the solute concentration values that were used to estimate parameters would compensate for them.

Two-dimensional deterministic advection–dispersion equation with the flow parallel to the x-axis can be written as [Fetter, 1999],

$$\frac{\partial C}{\partial t} = D_L\left(\frac{\partial^2 C}{\partial x^2}\right) + D_T\left(\frac{\partial^2 C}{\partial y^2}\right) - v_x\left(\frac{\partial C}{\partial x}\right), \qquad (1)$$

where C = solute concentration (M/L^3), t = time (T),

D_L = longitudinal dispersion coefficient (L^2/T),

D_T = transverse dispersion coefficient (L^2/T),

$v_x = K h_x / n_e$ = steady state average linear velocity in homogeneous media (L/T),

K = hydraulic conductivity (L/T),
h_x = hydraulic gradient and n_e = effective porosity.

The main tools used in this study were an ANN software package called NeuroShell2 and C++ programming language. In the first part of the study, C++ was used to model (1) to generate synthetic deterministic concentration values for the required spatial and temporal distribution for a given parameter. Then NeuroShell2 was used to train a network. After sufficient training we used the ANN model to estimate parameters from a noisy dataset obtained by adding noise to a dataset generated by (1) to simulate the real world randomness.

2.1. Estimating One Parameter

Deterministic solute concentration values were generated for 10 m x 5 m 2-D aquifer by using (1). 800 data examples (patterns) were generated for different hydraulic conductivity, *K*, values that ranged from 40 to 240 m/day. It was assumed that all other parameters, control variables and subsidiary conditions are fixed. Initial concentration value of 100 ppm was considered as a point source at middle of the header boundary of the aquifer and the same source was maintained at the boundary throughout the 10 day time period considered. Exponentially distributed point source concentration values along the longitudinal and lateral directions were considered as the initial conditions of other spatial coordinates. We gathered 50 input values for each example. Those input values represent solute concentration values at 10 spatial locations at 5 different time intervals; t = 1, t = 3, t = 5, t = 7, t = 10 day. We examined the possibility of amalgamating the time as an independent variable into concentration input data. However, it was difficult to meaningfully integrate them into presently available ANN architectures and innovative model structures need to be developed.

A simple 3 layer MLP network was utilised to train the network to find the complex relationship of output, *K*, and the associated concentration values. The dataset was divided into two categories, 80% of them were used for training and the rest was utilised for testing. The maximum and minimum values of the training network prediction range was set by selecting the values from both training (and testing) and estimating dataset, to prevent the ANN from extrapolating beyond its range. We applied scale functions of none, logistic and logistic for input, hidden and output layers, respectively. The default network parameters were used; learning rate = 0.1, momentum = 0.1, initial weight = 0.3. After a number of trial and error tests, it was found that the optimum results can be achieved by 20 hidden neurons. The network reached the stopping criterion of average error on test set, fixed at 0.000002, in less than 2 min in a 1GHz personal computer with performance measurements of the coefficient of multiple determination, R^2 = 0.9999 and the square of the correlation coefficient, r^2 = 0.9999. The network that produces best results on the test set is the one most capable of generalising and this was saved as the best network.

Having completed the successful training, another dataset was employed to test the network prediction of the estimating parameter. We made use of the same model to generate 800 new data values, however, initial concentration was randomly changed by up to ±5% and up to ±5% noise was arbitrarily added to all concentration input values. The reason for adding the noise is to simulate the real world problem of erratic behaviour of aquifers. The estimation error of each *K* value is given in Figure 1, which shows that the error increases with *K*.

Table 1 illustrates that mean squares error (MSE) percentage error (AAPE) is 5.63% and maximum error is 22.45 m/day, which may not be acceptable in most practical cases. Since the objective range of parameters are fairly large (40 –240 m/day), the accuracy of the approximation tend to decrease (Figure 1). Therefore, we conducted the same estimation procedure with four smaller permissible

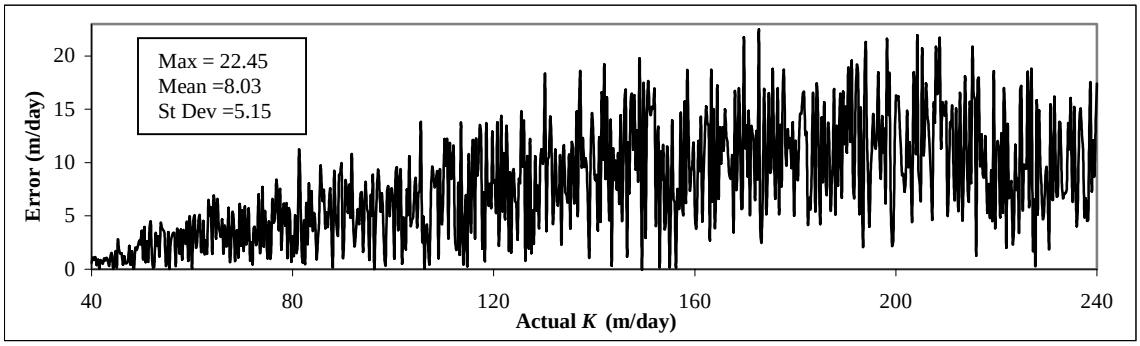

Figure 1. Error of estimated parameter *K*, when consider the whole range, 40 – 240 m/day.

parameter regimes of *K*; (i) 40-90, (ii) 90-140, (iii) 140-190 and (iv) 190-240 m/day. Table 1 shows that accuracy of the estimates improved considerably.

Table 1. Statistics of estimated error for different ranges of *K* with up to ±5% difference in initial value and up to ±5% noise in observations.

Error	*K* Range (m/day)				
	40-240	40–90	90–140	140–190	190-240
Max	22.45	1.88	2.23	2.99	2.98
Mean	8.03	0.27	0.38	0.36	0.39
StDev	5.15	0.32	0.41	0.49	0.47
MSE	45.25	0.11	0.14	0.20	0.19
AAPE(%)	5.63	0.11	0.12	0.18	0.18

Maximum error of 190-240 range has been reduced by about 90% (Figure 1 and Figure 2). Therefore, it is reasonable to assume that if we can gather prior information about the system in consideration, it is possible to obtain more accurate estimates. However, in real world problems the prior knowledge of the system is limited. In section 2.2, we specified a method to identify the range of parameters by using SON. However, before using SON, we explored the robustness of the ANN estimation models.

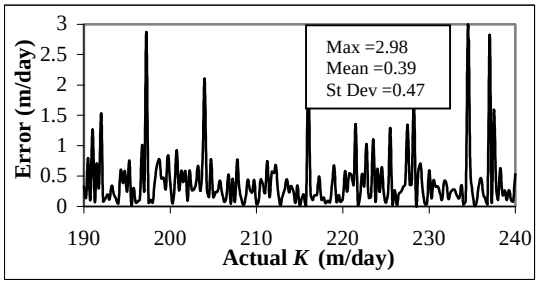

Figure 2. Error of estimated parameter *K*, only for the range 190 – 240 m/day.

Real world aquifer systems are subject to numerous random effects. One of them may be initial value problem. First, we investigated in the range of *K* between 190 – 240 m/day for the stability of the model for different initial values. The point source value of the initial concentration as well as all other resulting initial values of the system were changed from –50% to 50% (Table 2) and the parameter was estimated accordingly. Further, to illustrate the heterogeneity of the aquifers, up to ±5% extra noise was added to the concentration values. Table 2 shows the statistics of the estimates. Estimates exhibit direct relationship to the noise; however, most of the results are dependable even at higher noise levels. Random boundary conditions and irregular porous structure can result in erratic distribution of flow paths. Therefore, solute concentration spreads could be highly stochastic. We addressed this issue by extending the investigation of robustness by adding different level of randomness to the concentration values. First, data was generated by using the deterministic solutions of (1) for each case and then noise was added randomly to each deterministic concentration value to generate a noisy dataset. For example, to generate up to ±10% noise component to a deterministic value, *d*, two random functions are used as follows,

random function 1 → generate a random number between 0-1 (say *n*)

random function 2 → generate either +or -.

Therefore, noisy data = $d(1 \pm 10\% * n)$.

Table 3 demonstrates the statistics of the estimates obtained for noisy concentration data. Estimates

Table 3. Statistics of estimates for noisy data for *K* range 190-240 m/day.

± % added noise	Error of Estimate *K* (m/day)				
	Max	Mean	StDev	MSE	AAPE
10	2.29	0.85	1.10	1.98	0.68
20	3.54	1.05	1.19	2.04	0.76
30	5.46	1.74	1.22	2.25	0.81
40	5.88	1.96	1.35	2.31	0.92
50	6.00	1.99	1.39	2.39	0.93

Table 2. Statistics of estimated error for different initial values with up to ±5% error for *K* range 190-240.

Initial condition		Error of *K* (m/day)				
Point C value	Noise	Maximum	Mean	Stdev	MSE	AAPE (%)
50	-50%	9.72	3.48	1.68	12.66	0.95
60	-40%	7.55	2.06	1.57	3.36	0.94
70	-30%	6.18	2.01	1.42	3.01	0.92
80	-20%	4.36	1.59	0.97	2.58	0.77
90	-10%	2.84	0.97	0.72	1.46	0.43
Trained value (100)	0%	1.64	0.34	0.44	0.17	0.17
110	+10%	2.92	1.04	0.86	1.46	0.44
120	+20%	4.57	1.68	1.05	2.95	0.84
130	+30%	6.49	2.11	1.48	3.26	1.06
140	+40%	7.58	2.08	1.57	3.47	1.07
150	+50%	10.14	3.67	1.73	13.81	1.07

show that ANN model is stable even for highly stochastic systems.

2.2 Self Organising Networks (SON)

As shown in section 2.1, the ANN model gives more accurate estimates when the parameter range is small. However, in real world heterogeneous aquifers, it may be a difficult task to identify the accurate parameter range without reliable prior information. We developed a methodology by using SON [Kohonen, 1982] to identify the parameter range for given solute concentration values. SON has the ability to cluster the data of similar attributes into lower dimensions. We employed SON to cluster 800 x 50 dimension noisy dataset (used in section 2.1) with parameter range of 40 –240 m/day into four different categories. The "Supervised Kohonen" network architecture of NeroShell2 successfully categorised four different groups with 201, 200, 197 and 202 data patterns in each cluster respectively. SON put data into categories with high accuracy with few exceptions, which can be expected with noisy data, at the boundaries of the parameter ranges. Then we created and fed 10 different test datasets with the same number of input variables (50) into the trained SON and it accurately identified the correct parameter range for all the datasets.

2.3 Estimating Two Parameters

We extended the hybrid methodology to solve the groundwater inverse problem in the case of two unknown system parameters. We simulated the same aquifer that we used above. Our two parameters to be estimated are hydraulic conductivity, K (m/day) and longitudinal dispersion coefficient, D_L (m^2/day). We fed 50 concentration values and two actual outputs (K and D_L) to train the network.

In line with earlier work, we used a simple 3 layer network and it produced R^2 = 0.9999 and r^2 = 0.9999 for both outputs in 2 min and 50 sec. Then we fed a different dataset, which has not been seen by the trained network before. The new dataset consisted of randomly varying (up to ±5%) initial conditions and added noise to replicate a natural system. We explored two different levels of noise; up to ±5% and ±50%. The parameter ranges are; K between 190 – 240 m/day, D_L between 0.03 – 0.08 m^2/day. ANN model produced reasonable estimates for both parameters and the summary of estimates is given in Table 4.

Table 4. Statistics of estimates for 2 parameter case

Parameter	Actual Range	± % noise	Error of Estimate			
			Max	Mean	MSE	AAPE
K	190-240	5	2.48	0.99	2.65	0.81
		50	6.78	2.35	3.18	1.12
D_L	0.03-0.08	5	0.00341	0.00092	0.0014	0.0005
		50	0.00875	0.00247	0.0029	0.0010

3. CASE STUDY

In this section, we applied the hybrid inverse approach presented in the section 2.2 to estimate parameters of an artificial aquifer. We obtained the data for this investigation from a large, confined, artificial aquifer which is used for contaminant transport tests at Lincoln University, New Zealand. This aquifer is 9.49 m long, 4.66 m wide and 2.6 m deep, and porous media is sand. Although, initial conditions, other parameters and the subsidiary conditions are somewhat known, we had to conduct a fairly tiresome, "trial and error" exercise to replicate the aquifer. 800 data patterns were generated for the hydraulic conductivity range of 80 to 280 m/day. Each pattern consisted of 100 concentration input variables for 10 distinct spatial locations for 10 different time intervals. Then we used Kohonen's SON (80% data for training and 20% for testing) to classify the input values into 4 clusters as shown in Figure 3. Then we fed the actual aquifer data into the trained network and the selected subrange is shown in Figure 4. Here, the trained network determines which cluster most resembles the input vector by numeric 1 (others 0). It was determined that the aquifer parameter should be within the second cluster (130 – 180 m/day). Based on this information we generated a separate dataset for the specified range and trained an MLP network with associated K values.

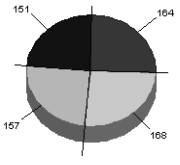

Figure 3. Distribution of clusters by SON.

The estimate given by the trained ANN was 152.86 m/day. The experimental value of hydraulic conductivity, K, was found to be 137 m/day, which was calculated by calibration tests conducted by aquifer testing staff. In these experiments, they have assumed that the aquifer is homogeneous.

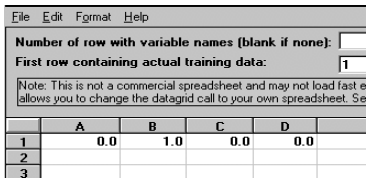

Figure 4. Classification output of parameter range.

The difference between two estimates is only 10.37%. Considering the assumptions of homogeneity made by the aquifer researchers and other possible human errors, it is fair to state that the estimate from ANN model is reasonable and acceptable.

4. CONCLUSIONS

This paper presented a hybrid approach using a combination of two types of ANN models to solve the inverse problem in groundwater modelling. Supervised Multi Layer Perceptron (MLP) ANN and Self-Organising Network (SON) was amalgamated to estimate parameters reasonably accurately by using solute concentration observations that were obtained from a numerical model. The prediction error of the estimate of K from the MLP network increased with the actual K value in the range from 40-240 m/day. However, after subdividing the original range into four smaller ranges and trained a separate network for each range, the original error was reduced by 90%. Furthermore, it was observed that the ANN provide reliable parameters even under uncertain and noisy conditions. Extension of the prediction to two parameters (K and D_L) also provided reliable estimates with absolute percentage errors of 1.1% and 0.001%, respectively, for a highly noisy system. A SON was developed to identify the range of K represented by a particular data set, which then allows the development an appropriate MLP network for prediction. The hybrid (SON-MLP) model was applied to an artificial experimental aquifer and the estimate of K was found to be quite accurate with 10.37% error.

Our investigations emphasised the importance of modelling a sufficiently true representation of the physical system and subsidiary conditions to obtain accurate parameter values. As Minns et al. [1996] pointed out, ANN is susceptible to becoming "a prisoner of its training data". Therefore, prior information, such as type of contaminant source, boundary conditions and subsidiary conditions is crucial to modelling the system accurately. If we could gain such prior information and model the system with ANN, it would be capable of solving the inverse problem with greater accuracy even with highly noisy data as well as different system input values.

5. REFERENCE

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, Artificial neural networks in hydrology. II: Hydrologic applications, *Journal of Hydrologic Engineering, ASCE*, *5*(2), 124-137, 2000.

Coulibaly. P., F.Anctil, R. Aravena and B. Bobee, Artificial neural network modeling of water table depth fluctuations. *Water Resources Research*, *37*(4): 885-896, 2001.

Fetter. C. W. - *Contaminant Hydrogeology*, Prentice-Hall Inc., New Jersey, 1999.

Kitanidis, P and E.G. Vomvoris, E.G. A geostatistical approach to the problem of groundwater modelling (steady state) and one-dimensional simulation. *Water Resources Research*, *19*(3): 677-690, 1983.

Kohonen, T., Self-organized formation of topologically correct feature maps, *Biological Cybernetics*, *43*, 59-69, 1982.

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

Minns, A. W., and M. J. Hall, Artificial neural networks as rainfall-runoff models, *Hydrological Sciences Journal*, *41*(3), 399-417, 1996.

Morshed. J. and J.J.Kaluarachchi, Application of artificial neural network and generic algorithm in flow and transport simulations. *Advances in Water Resouces*, *22*(2):145-158, 1998.

Ranjithan. S., J.W.Eheart and Jr.J.H.Garrett, Neural network-based screening for groundwater reclamation under uncertainty. *Water Resources Research*, *29*(3):563-574, 1993.

Rogers. L.L and F.U.Dowla, Optimization of groundwater remdiation using artificial neural networks with parallel solute transport modeling. *Water Resources Research*, *30*(2):457-481, 1994.

Rogers. L.L , F.U.Dowla and V.M.Johnson, Optimal field-scale groundwater remediation using neural networks and the genetic algorithm. *Environmental Science and Technology*, *29*(5): 1145-1155, 1995.

Yeh. W.W-G, Review of parameter identification procedures in groundwater hydrology: The inverse problem. *Water Resources Research*, *22*(2): 95-108, 1986.