A Comparative Study of Parameter Estimation in Hydrology Modelling: Artificial Neural Networks and Curve Fitting Approaches

C. Rajanayaka, D. Kulasiri and S. Samarasinghe

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

Abstract: In this paper, the performance of two estimation methods that are used to solve the inverse problem in hydrology was compared using a stochastic solute transport model (SSTM), which was presented at a previous MODSIM conference, as a test case. The first method was a hybrid Artificial Neural Network (ANN) and the second was a conventional curve fitting technique. It was found that using a smaller range for the output variables in the training model could enhance the accuracy of the estimates given by the ANN model. Self Organising Maps (SOM) were employed to cluster a larger dataset into different categories. Then the SOM model was fed with the dataset that represents the system to identify the corresponding data group. Afterwards a multi layer perceptron was employed to obtain the final estimates of the SSTM parameters. Initially the method was tested on a synthetic dataset. Results reveal that model predictions are satisfactory and the average absolute error of the estimates for a highly random data range is approximately 5.5% and outcomes are better for other ranges. Furthermore, the performance of the inverse model was robust and consistent against different sets generated using the standard Wiener processes. Then the method was tested using the data from an artificial aquifer at Lincoln University, New Zealand. The second method, the curve-fitting technique, was used to determine the hydrologic parameters of the aquifer parameters by fitting solute concentration profiles of the aquifer and the model. Comparative results of two methods are in reasonable agreement. The hybrid ANN approach is simple and easy to use but results indicate that it is a robust inverse method. However, knowledge of ANNs and prior information of the system are necessary to improve the accuracy of the estimates.

Keywords: *Hydrology Modelling; Parameter Estimation; Artificial Neural Networks; Groundwater*

1. INTRODUCTION

A sufficient knowledge of hydrogeological parameter distribution of the spatial region is one of the most important requirements in hydrology modelling. When we model the behaviour of a hydrology system, for instance groundwater flow and solute transport in porous media through differential equations, it is often necessary to assign numerical values to these parameters. These values are 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, these field scale experiments can be expensive. Conversely, often we are interested in modelling for quantities such as the depth of water table and solute concentration. This is because they are directly relevant to environmental decision-making, and we measure these variables regularly and the measuring techniques tend to be cheaper. Further we can continuously monitor these decision (output) variables in many situations. Therefore it is reasonable to assume that these observations of the output variables represent the current status of the system. If the dynamics of the system can reliably be modelled by a relevant differential equation, we can expect that the parameters estimated based on the observations may give us more reliable representative values than those obtained from laboratory tests and the literature.

There have been numerous studies dedicated to develop inverse methodologies. Some of these methods are primitive trial and error techniques, optimisation approaches and geostatistical methods. Simple optimisation techniques are not suitable to address the high random variations in real world aquifers. Whereas, the geostatistical methods take the geological uncertainty into account. However, it appears that average practitioners find it hard to understand and apply the highly theoretical geostatistical methods. Further, several reviews on geostatistical inverse methods have shown that developers of inverse methods have perhaps focused too much on validating their methods on too-simple synthetic data fields and those methods are not performing well in some heterogeneous conditions. Moreover, it may be difficult to identify the most appropriate inverse method for a given problem, as different types of heterogeneity may be prominent for the system of interest (Zimmerman et al., 1998). Over the past decade a few Artificial Neural Networks (ANN) approaches were developed to predict the hydrogeologic parameters (Aziz and Wong, 1992: Balkhair, 2002).

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. 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, ANN have been successfully employed in hydrology research, such as rainfall-runoff modelling, stream flow forecasting, precipitation forecasting, groundwater modelling, water quality and management modelling (Morshed and Kaluarachchi, 1998; Maier and Dandy, 2000).

The outcome of the above mentioned inverse methodologies is utterly dependent on the accuracy of the model formulation. Nevertheless, most of the models, which are commonly used by the practitioners to simulate natural systems, represent linear time dependent partial differential equations mainly based on deterministic consideration. The deterministic solutions only have a single set of output values for a given set of inputs and parameters. However, real world systems such as aquifers consist of heterogeneous formation of porous media, complex multifaceted boundaries and random distribution of parameters with irregular inputs (e.g. rainfall). These complexities of groundwater systems, therefore, cannot be accurately understood by deterministic description and need to be described in a stochastic sense by using stochastic differential equations, for example Unny (1989). After the pioneering work of Freeze (1975), a large number of studies have contributed to understand the probabilistic nature of the heterogeneous formation of the underlying aquifer parameters distribution.

Kulasiri and Verwoerd (1999, 2002) developed a stochastic solute transport model (SSTM) assuming the velocity of solute as a fundamental stochastic variable;

$$v(x,t)=\overline{v}(x,t)+\xi(x,t), \tag{1}$$

where $\overline{v}(x,t)$ = average velocity described by Darcy's law, and $\xi(x,t)$ = white noise correlated in space and δ- correlated in time. The main parameters of the model are the correlation length, *b* and the variance, σ^2. Different values of these parameters regulate the statistical nature of the computational solution. This model avoids use of the Fickian assumption that gives rise to the dispersion coefficient, *D*. The *D* proved to be scale dependent (Fetter, 1999).

The objective of this paper is to estimate parameters of SSTM using a hybrid ANN approach to test the performance of the estimation method in dealing with stochastic data. Then estimates will be compared using a more conventional curve fitting technique. The ANN is implemented using the NeuroShell2 software.

2. HYBRID ANN APPROACH

Rajanayaka et al. (2002) developed a hybrid ANN method to solve the inverse problem in groundwater modelling. Initially, a Multi Layer Perceptron (MLP) network was developed and it was found that the network produced better results when the target range of the parameters is smaller. Therefore, a Self-Organising Map (SOM) (Kohonen*,* 1982) 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 a range of parameters. The input data was fed into a MLP ANN to train the network along with corresponding parameter values. A 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 is used to estimate parameters is 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.

2.1. Application to SSTM

In this section, we applied the ANN hybrid approach (Rajanayaka et al., 2002) to estimate parameters of the Stochastic Solute Transport Model (SSTM). Since, SSTM consists of two parameters; variance (σ^2) and correlation length (*b*), we estimated both parameters simultaneously.

However, Rajanayaka et al. (2002) showed that the accuracy of the estimates was inversely proportional to the extent of the objective range. Thus it is necessary to identify reasonably smaller output ranges for both parameters. Additionally Rajanayaka and Kulasiri (2002) illustrated that the larger parameter values of SSTM represent higher random flows, especially σ^2 around 0.25. For that reason, we limited the parameter range for both parameters, σ^2 and *b*, for an acceptable range between 0.0001 and 0.2.

As the first step of the implementation process, we used SSTM to simulate a one-dimensional aquifer of 10 m in length. 800 data patterns for different combinations of σ^2 and *b* were generated. Every data pattern consists of 200 inputs for 10 various spatial locations of the aquifer for 20 time intervals. The same standard Wiener process was utilized for generation of all datasets. Initial conditions of concentration value of unit 1.0 at x = 0.0 and exponentially distributed initial values for other spatial locations were considered. Throughout the simulation the same concentration (unit 1.0) was maintained at the upper end boundary. It was assumed that the mean velocity of the solute was 0.5 m/day. As mentioned above, limiting the objective range of the parameters to a smaller regime was significant to attain accurate approximations. Thus, Kohonen's SOM was employed to cluster the data set into four categories.

Since, the dataset represents the stochastic behaviour of the flow, the time needed to classify the data into separate groups was much more than in the similar case of the deterministic advection – dispersion data used in Rajanayaka et al. (2002). In the present case, it was 32 minutes and 5 seconds using a 1 GHz personal computer. Randomly selected 80% of data was used for training the network and the rest for validation. However, the performance of the ANN model can be affected by the way the dataset is divided into different groups (Maier and Dandy, 1996). This outcome is mainly caused by ANN's inability to extrapolate beyond the range of the data used for training. For that reason, the training and validation sets must be representative of the same population to obtain adequate generalization ability (Masters, 1993; Tokar and Johnson, 1999). We addressed this issue by manually adjusting the upper and lower bounds of the training model. Those limits were determined using upper and lower values of all three data samples: training, testing and validation.

Distribution of the large dataset into four categories by using a SOM is shown in Figure 1. Comparing the performance of the deterministic data distribution given by Rajanayaka et al. (2002) (800 data into four categories of 201, 200, 197 and 202), the results presented here failed to reach the same accuracy. However, notwithstanding the random nature of the current dataset, the SOM has clustered the data to an adequate degree of accuracy that may be sufficient for the problem at hand.

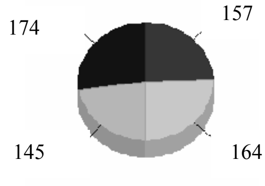


Figure 1. Distribution of SSTM data for σ^2 and *b* range between 0.0001 and 0.2 into four categories.

Having substantiated that the SOM could be successfully used to cluster a large dataset that represents heterogeneous system data such as that given by SSTM, we created four separate specialized network models for smaller parameter regimes of both parameters; (i) 0.0001 – 0.05 (ii) 0.05 – 0.1 (iii) 0.1 – 0.15 (iv) 0.15 – 0.2. Each dataset comprises of 441 data patterns of different combinations of σ^2 and *b* at intervals of 0.0025. Same SSTM model that was used above for SOM cluster distribution was utilized for data generation. Nevertheless, in this case each data pattern contains not only 200 inputs but also 2 corresponding output parameters. The number of training patterns can considerably influence the performance of the ANN model (Flood and Kartam, 1994). Increasing the number of data patterns provides more information about the shape of the solution surface and thus improves the accuracy of the model prediction. However, in most real world applications, numerous logistical issues impose limitations on the amount of data available and consequently the size of the training set. Hence, in developing a method for practical applications, it is important to test the robustness of the methodology for such data limitations. For that reason, we limited the parameter range of each model to 21 values (between 0.0001 and 0.2 at intervals of 0.0025).

Maier and Dandy (2000) denoted that it is important to select a suitable network architecture and model validation method in the development of ANN models to achieve optimum results. Besides, it may be necessary to select the most suitable model in the case of handling highly random data such as SSTM data. Therefore, we conducted a few trial and error exercise to choose the appropriate model structure, and training and testing procedure. The standard multi layer, parallel and jump connections with different architectures available in NeuroShell2 were considered. After numerous attempts, it was found that the network model with 'five layer standard connections' (input, 3 hidden, and output layers) could produce the best trained model. Each hidden layer consisted of 30 neurons. Activation functions of linear <0, 1>, logistic, tanh, Gaussian and logistic were used for layers of input, hidden (3 layers) and output, respectively. The default network parameters of NeuroShell2 were employed; learning rate = 0.1, momentum = 0.1, initial weight = 0.3. The stopping criterion was set to a minimum error of 0.000001. 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. Further, that procedure ensures that overtraining did not occur. All four networks reached the stopping condition in about 30 minutes in a 1 GHz personal computer with performance measurements shown in Table 1.

Table 1. Performance measurements of trained ANN model for four different parameter ranges.

Parameter range	R^2		Mean absolute error	
	σ^2	b	σ^2	b
0.0001– 0.05	0.991	0.987	0.0	0.0
0.05 – 0.1	0.991	0.987	0.0	0.0
0.1 – 0.15	0.989	0.987	0.0	0.0
0.15 – 0.2	0.972	0.977	0.0001	0.0001

After the completion of successful training of each model, separate datasets were generated to test the prediction capability of the models. The same SSTM was employed to produce another dataset of 441 data patterns for each parameter range. However, different standard Wiener process increments were used. In addition, initial and boundary conditions were adjusted up to ±5% by adding random values. Input data values of each dataset were then fed into the corresponding trained network and processed to obtain model predictions. Figure 2 illustrates the absolute error of estimated parameter σ^2, that ranges from 0.0001 to 0.05. It shows that ANN model prediction is extremely satisfactory and that the average absolute error is approximately 0.04%.

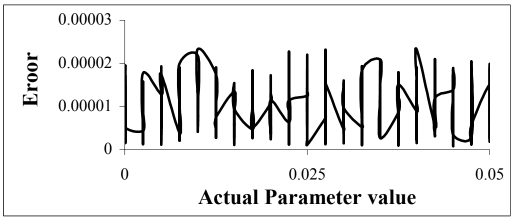


Figure 2. Absolute error of estimated parameter σ^2, for the range of 0.0001 – 0.05.

A similar approach was applied to other parameter ranges as well. The precision of the estimates given by ANN models shrinks with highly heterogeneous data. As larger values of parameters indicate excessive stochastic flows, we can expect the accuracy of the prediction to diminish for highly stochastic flows. Nonetheless, the average absolute error for the estimates for a range of 0.15 to 0.2 was approximately 5.5%, which may be acceptable for most of practical applications.

The above prediction accuracy analyses of the ANN models were based on similar ranges for both parameters. However, in most practical circumstances we may have to associate with different values of parameters for σ^2 and *b*. Thus the robustness of the ANN methodology for different values of parameter regimes for two parameters was assessed. We generated two separate datasets for two extreme cases; (i) smaller σ^2 ranges from 0.0001 to 0.05, and higher *b* ranges from 0.15 to 0.2 (ii) higher σ^2 ranges from 0.15 to 0.2, and smaller *b* ranges from 0.0001 to 0.05. A similar approach to that which was used for earlier investigations was employed to gauge the capability of the ANN model. Estimates reveal that the trained network has predicted the estimates with reasonable precision. In both cases the percentage average absolute error is approximately 4%.

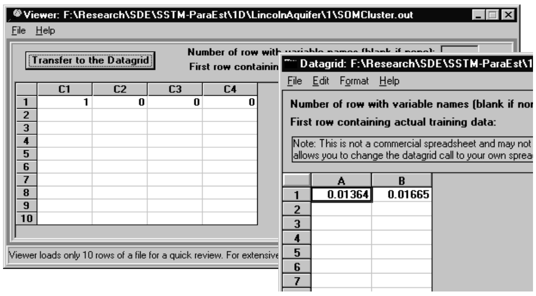
3. CASE STUDY

In this section, we applied the hybrid ANN inverse methodology with contaminant transport tests conducted at a large, confined, artificial aquifer at Lincoln University, Canterbury, New Zealand. A detailed description of the aquifer and transport experiments can be found in Kulasiri and Verwoerd (2002).

Firstly, we utilized the known conditions such as initial and boundary conditions and hydraulic gradient to simulate the aquifer using SSTM. The initial concentration at $x = 0$ was 1.0 unit and it was reduced exponentially with time. Initial values of other spatial points were considered as zero. 1681 data patterns were generated for different combinations of parameters of σ^2 and *b*. Single standard Wiener process increments were retained for every simulation run. Both parameters were varied between 0.0001 and 0.2.

Generated data patterns were fed into Kohonen's self-organizing map architecture to cluster them into four different groups. After classifying them with reasonable accuracy aquifer datasets were fed into the SOM model to identify relevant groups that the data resemble the most. We constructed 20 different 1-D datasets to represent five different sets of sample wells along the aquifer at four levels. Concentration values of the aquifer were normalized to be able to weigh them against the

SSTM data that was produced with unit initial concentration. In addition, we made an assumption that transverse dispersion of the aquifer was compensated by consideration of the stochastic flow. Initially a dataset closer to the middle of the aquifer was chosen for the estimation of parameters. The aquifer dataset had to be interpolated to produce missing data, and to fabricate uniform spatial and time grids. Having constructed an exact number of data for similar spatial and time intervals as for the original ANN model, the aquifer dataset was fed into the trained model. The processed results, as shown in Figure 3, illustrated that the aquifer data fits into the first group of the original dataset (by numeric 1, and others 0). Then, the selected dataset was separated from the original larger set and trained a model for the smaller parameter range. Based on the above model selection, 'five layer standard connections' was utilized with the same activation functions, initial weights, momentum and learning rates, which were used in the previous section.


Viewer: F:\Research\SDE\SSTM-ParaEst\1D\LincolnAquifer\1\SOMCluster.out
File
Help
Transfer to the Datagrid
Datagrid: F:\Research\SDE\SSTM-ParaEst\1
File
Edit
Format
Help

Number of row with variable names (blank if none):
First row containin

	C1	C2	C3	C4
1	1	0	0	0
2				
3				
4				
5				
6				
7				
8				
9				
10				

Viewer loads only 10 rows of a file for a quick review. For extensive

Number of row with variable names (blank if nor

First row containing actual training data:

Note: This is not a commercial spreadsheet and may not
allows you to change the datagrid call to your own sprea

	A	B
1	0.01364	0.01665
2		
3		
4		
5		
6		
7		

Figure 3. Classification of aquifer dataset and estimates produced by trained ANN model.

Following sufficient training (minimum error of the test set = 0.000001), the artificial aquifer dataset was fed into the ANN model to estimate parameters. The estimates that were produced by the model are given in Figure 3; $\sigma^2 = 0.0136$ and $b = 0.0166$. We conducted a similar procedure for all the flow paths of the artificial aquifer to further test the robustness of the methodology. The estimates indicate that although the artificial aquifer is reasonably heterogeneous, the ANN model estimates the parameters with a reasonable accuracy with average values of $\sigma^2 = 0.01256$ and $b = 0.0141$.

4. CURVE FITTING

In this section we used a more conventional technique, namely curve fitting, to alternatively estimate parameters of the Lincoln University artificial aquifer. Even though the curve fitting technique is to some extent a primitive and time consuming approach, it gives accurate estimates provided a correct procedure is followed. Since, the present stochastic model is a one-dimensional model, we experimented in directly relating to one-dimension solute concentration profiles of the aquifer. However, as one can assume, the actual aquifer is subjected to transverse dispersion. Therefore, consideration of mere one-dimensional deterministic flow is not sufficiently accurate. Hence, we employed the following methodology to approximate the aquifer parameters.

First, we selected spatial coordinates closer to the middle of the aquifer. Then, we developed a two-dimensional deterministic advection-dispersion transport model and obtained corresponding concentration values of the model that are similar to selected spatial locations of the aquifer. As given by Fetter (1999), based on plausible arguments we assumed that the transverse dispersion coefficient is 10% of the longitudinal dispersion. A mean velocity of 0.5 m/day was considered. Afterward, the profiles of both the aquifer and the deterministic model were plotted on a one axis system to compare their similarity. This curve fitting technique was carried out in association with a trial and error exercise to determine the most suitable fitting of the curves by changing dispersion coefficients of the deterministic model.

After investigating many combinations of parameters by trial and error, it was found that the closest fit could be obtained by a longitudinal dispersion coefficient of 0.15 m^2/day (Figure 4). For simplicity, concentration values of the aquifer were normalized.

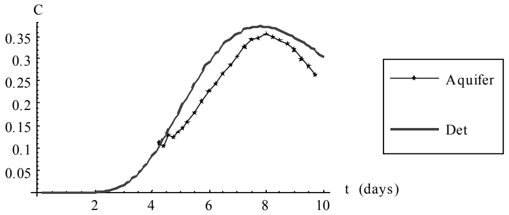


Figure 4. Concentration profile of trial and error curve fit for D = 0.15 m^2/day of advection dispersion model of aquifer data.

Subsequently, we developed a one-dimensional deterministic advection-dispersion model by using the longitudinal dispersion coefficient obtained from the two-dimensional comparison. Then we used a similar curve fitting technique to that used above, with the 1-D deterministic model and 1-D stochastic model. Investigation of curve fitting for different parameter combinations was conducted for the same Wiener process. After numerous attempts it was found that the parameter combination of $\sigma^2 = 0.01$ and $b = 0.01$ closely represents the aquifer (Figure 5). Having determined the appropriate parameters of SSTM that simulate the Lincoln University aquifer at the

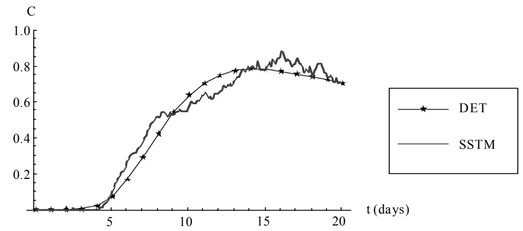


Figure 5. Concentration profiles of deterministic advection-dispersion model (D = 0.15 m^2/day and SSTM with σ^2 = 0.01 and b = 0.01.

spatial location considered we investigated the robustness of the model for different Wiener processes. It was found that the model is reasonably stable for eight different standard Wiener increments. Even though the above results show that the parameter combination of σ^2 = 0.01 and b = 0.01 is a fairly accurate representation of the experimental aquifer for the given spatial point, we extended the validation process for other spatial locations. The results show that the same parameter combination is reasonably valid for other spatial points as well.

5. CONCLUSION

In this paper, the performance of two estimation methods that are used to solve the inverse problem in hydrology was compared using a stochastic solute transport model (SSTM), as a test case. The first method is a hybrid ANN approach that consists of a supervised multi layer perceptron and a self-organising map. Initially the method was applied to a synthetic dataset. Then it was utilised to estimate parameters of an artificial aquifer. The estimates were compared using a more conventional curve fitting technique. Estimates given in both approaches are in reasonable agreement and they reveal that the ANN approach provides credible estimates even for highly stochastic data.

6. REFERENCE

Aziz, A.R.A., and K.V. Wong, A neural-network approach to the determination of aquifer parameters*, Ground Water*, 30(2), 164-166, 1992.

Balkhair, K. S., Aquifer parameters determination for large diameter wells using neural network approach, *Journal of Hydrology*, 265(1-4), 118-128, 2002.

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

Flood, I., and N. Kartam, Neural networks in civil engineering. I: Principles and understanding, *Journal of Computing in Civil Engineering*, 8(2), 131-148, 1994.

Freeze, R. A., A stochastic-conceptual analysis of one dimensional groundwater flow in a non-uniform homogeneous media, *Water Resources Research*, 11(5), 725-74, 1975.

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

Kulasiri, D., and W.S. Verwoerd, A stochastic model for solute transport in porous media: mathematical basis and computational solution, Proc. MODSIM 1999, Hamilton, New Zealand, 31-36, 1999.

Kulasiri, D., and Verwoerd, W. *Stochastic Dynamics: Modeling Solute Transport in Porous Media, North-Holland Series in Applied Mathematics and Mechanics, vol 44*. Amsterdam: Elsevier Science, 2002.

Maier, H.R., and G.C. Dandy, The use of artificial neural networks for the prediction of water quality parameters, *Water Resources Research*, 32(4), 1013-1022, 1996.

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.

Masters, T., *Practical Neural Network Recipes in C++*, Academic Press, San Diego, 1993.

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

Rajanayaka, C., S. Samarasinghe and D. Kulasiri, Solving the inverse problem in stochastic groundwater modelling with artificial neural networks, Proc. of IEMSS 2002, Lugano, Switzerland: vol 2, 154-159, 2002.

Rajanayaka, C., & Kulasiri, D. Exploration of the behaviour of a stochastic solute transport model using computational experiments. In *Proc. IEMSS 2002*, Lugano, Switzerland: vol 2, 301- 306, 2002.

Tokar, A. S., and P. A. Johnson, Rainfall-runoff modeling using artificial neural networks, *Journal of Hydrologic Engineering*, 4(3), 232-239, 1999.

Unny, T.E., Stochastic partial differential equations in groundwater hydrology – Part 1: *Journal Hydrology and Hydraulics*, 3,135-153, 1989.

Zimmerman, D. A., G. de. Marsily, C. A. Gotway, M.G. Marietta, et al., A comparison of seven geostatistical based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow, *Water Resources Research*, 34(6): 1373-1413, 1998.