

Artificial neural networks: development and application in groundwater pollution remediation design

T. D. KROM

ELSAMPROJEKT, Kraftværksvej 53, DK-700 Fredericia, Denmark

D. ROSBJERG

Groundwater Research Centre, Denmark's Technical University, DK-2800 Lyngby, Denmark
e-mail: dam@isva.dtu.dk

Abstract Artificial neural networks (ANNs) are investigated as a tool for the simulation of contaminant loss and recovery in three-dimensional (3-D) heterogeneous groundwater flow and contaminant transport modelling. These methods have useful applications in expert system development, knowledge base development and optimization of groundwater pollution remediation. Conventional numerical model runs are used to develop the ANNs. ANNs have been analysed with the goal of estimating objectives that normally require the use of traditional flow and transport codes such as recovered mass, unrecovered mass and remediation failure. The inputs to the ANNs are variable pumping withdrawal rates at fairly unconstrained 3-D locations. A forward-feed backwards error propagation ANN architecture is used. The significance of the size of the optimization data set, network architecture and network weight optimization algorithm, with respect to the estimation accuracy and objective are shown to be important. Finally, cross-validation techniques quantify the quality of the weight optimization for strongly under described systems.

INTRODUCTION

The application of numerical solutions for differential equations describing liquid and contaminant transport is very computationally intensive. In seeking a methodology where limiting problem scope can be traded for large savings in computing time, artificial neural networks (ANNs) have been investigated. Others have shown that ANNs can be used advantageously to analyse groundwater problems (Rogers, 1992; Johnson & Rogers, 1995; Ranjithan *et al.*, 1993; Rizzo & Dougherty, 1994; Dowd & Sarac, 1994).

We generalize previous work to constraining well location in a three-dimensional (3-D) model, to a zone within the model domain (downstream from the centre of the contaminant source and not a boundary node), and well rates must be negative (i.e. not injecting).

The goal here is three fold: first to show that neural networks can be used to solve more general contaminant transport problems than have been solved in the past. Secondly, to show the impact of problem generality on the requirements for the size of the optimization data set, net architecture size and the optimization approach with respect to the desired estimation accuracy. Lastly, to demonstrate some basic techniques which can aide in improving confidence and reliability in the ANN prediction.

ARTIFICIAL NEURAL NETWORK THEORY

All of the ANNs used in this work are fully connected error-back propagation forward-feed neural networks (supervised learning with continuous i/o) (Bishop, 1995). Figure 1 shows an example of a forward-feed network of the type used in this work.

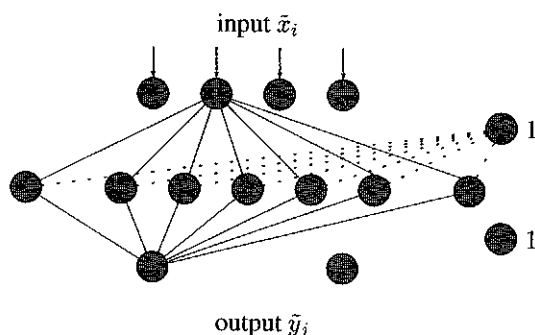


Fig. 1 A schematic of an error-back propagation forward-feed ANN, showing only the connections from one input and one bias correction node (bias nodes are shown as 1) for clarity.

An ANN is an ordered assemblage of nodes where an input signal (x) to each node is mapped $y = F(x)$. The result from the entire assemblage of nodes, as ANN, is a non-parametric mapping: $X \rightarrow Y$. Equation (1) gives the mapping function that is used in this work, where x_i is an input signal (e.g. well coordinates and pumping rate), μ_{ij} is a weight vector and y_j is the mapped output (e.g. contaminant recovered). $\mu_{bias,j}$ is a bias correction required during weight vector optimization:

$$y_j = \tanh \left(\sum_i x_i \mu_{ij} + \mu_{bias,j} \right) \quad (1)$$

In order to improve the performance of the optimization, raw input x , and output y signals, are normalized by a simple algorithm:

$$\begin{aligned} \tilde{x}_i &= x_i \sigma_{x,i} + m_{x,i} : \text{inputs} \\ \tilde{y}_j &= y_j \sigma_{y,j} + m_{y,j} : \text{outputs} \end{aligned} \quad (2)$$

where σ is the experimental standard deviation and m is the experimental mean of the raw data sets and the normalized values are substituted into the optimization equation (1). The experimental m and σ are based only on the part of the raw data used directly in the optimization; the information used in cross-validation and testing is excluded.

All the weights (μ), are optimized so that the best possible estimation of y is obtained. This is done by optimization in the form of repeated presentation of a set of input-output pairs coupled with an optimization procedure. The quality of the optimization is evaluated by “testing” against a second set of input-output pairs (cross-validation). The estimate that results is the posterior probability, or the value with the greatest posterior probability, for the estimation objective.

The forward-feed error back propagation optimization algorithm shown in equation (3) is used. This contains a gain factor, ψ , and momentum term α

(a smoothing parameter over each optimization cycle k). The error expression ε , is shown in equation (4). $N_{optimize}$ varies between 1 and $N_{training}$ (maximum available number of input-output pairs). $N_{optimize} = 1$ should be best for noise insensitive and $N_{optimize} = N_{training}$ for noisy optimization problems.

$$\mu_{k+1} \leftarrow \mu_k - \psi \nabla \varepsilon_k - \alpha \nabla \varepsilon_{k-1} \quad (3)$$

where k is an iteration counter, and ε is given in equation (4):

$$\varepsilon = \frac{1}{N_{optimize}} \sum_{t=1}^{N_{optimize}} (\hat{y}_t - \tilde{y}_t)^2 \quad (4)$$

where $N_{optimize} \subset N_{training}$ and \hat{y}_t is the estimate for \tilde{y}_t .

The groundwater transport estimation problem is under-described and thus a measure for optimization reliability is developed. A global optimum is impossible to test for. Cross-validation allows the evaluation of learning to be judged upon an ANN's ability to generalize. This is done by developing 20 ANNs using different starting points in the weight space. Then statistics are calculated on the error ε and cross-validation error; thus the reliability of the optimization can be evaluated.

METHODOLOGY

The method is fairly straightforward. One needs realizations for a parameter set that can be implemented in a conventional numerical model so that optimization, cross-validation and testing sets can be developed. A parameter set has been generated via geostatistical simulation, while MODFLOW (McDonald & Harbaugh, 1988; Harbaugh, 1992) and MT3D (Zheng, 1990) are used for the flow and transport simulations, in which the variables sampled are well location and pumping rates.

Problem formulation

The example case focuses on estimating the performance of one remediation well in a $20 \times 20 \times 5$ ($1000 \times 1000 \times 35$ m) numerical model with no-flow boundaries on four sides (east–west–south–bottom), constant flux on top (infiltration is 300 mm year^{-1}), and a constant head boundary condition on the north side (Fig. 2). The model is started at steady state conditions and then the well is turned on. The well remains on during the rest of the model run. The contaminant source is in layers 2, 3 and 4. The hydraulic conductivity varies over five orders of magnitude and is based on an actual field site.

The zone in the model where wells can be placed is limited to the downstream half of the model relative to the source, and to the lower three layers (Fig. 2). The source is in layers 2, 3 and 4. This provides 648 possible well sites with variable extraction rates of between 0 and $90 \text{ m}^3 \text{ day}^{-1}$. One extraction rate per well site was used to develop the ANNs. The extraction rates came via random draws from an exponential probability density function. If a well placement coincides with a low permeability formation the extraction rate is reduced by a constant factor of 0.1 relative to the random draw. Thirty seven percent of the model is low K_{SAT} formations.

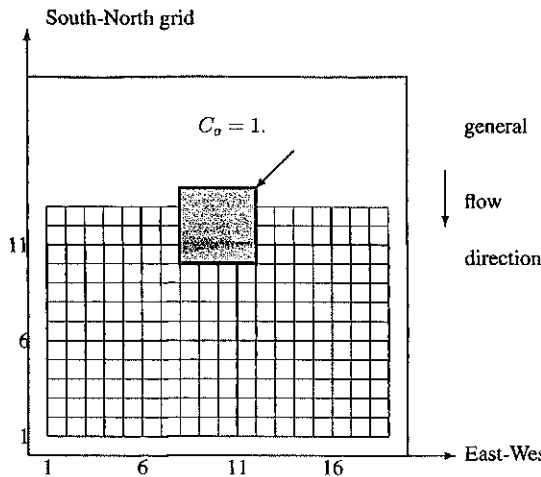


Fig. 2 The initial concentration conditions, and allowable zone (grid) for well placement. The source is in layers 2, 3 and 4.

Estimation objective

The ANN output is the posterior estimate for the transformed variable \tilde{y} , which is conditional on the optimization information: \tilde{x}_i, \tilde{y}_j pairs. One variable estimated is the amount of contaminant crossing a control surface within the model: contaminant loss. At the same time, contaminant recovery, the contaminant recovered by the well, is estimated. The latter expresses the efficiency of the remediation design. Finally, failure is estimated, as a 0/1 problem where failure is given by the situation where the amount of contaminant recovered is less than some arbitrary value.

NEURAL NETWORK PERFORMANCE

The development of reliable ANNs was studied with regard to their complexity, the size of the neural network, the size of the optimization data set and the optimization approach.

Experimental methodology

All the ANNs developed in this work used momentum learning (equation (3)), with cross validation used as the measure of optimization performance over 500 optimization iterations. Based upon the cross-validation analysis the best set of weights was stored and used in post-optimization testing.

Optimization data set sizes were 50, 100, 150, 200 and 300 members chosen at random from an exhaustive set of 648 locations. Fifty members (the same for all experiments) were used for cross-validation; 250 members (the same for all experiments) were used for testing generalization accuracy. Twenty ANNs were developed for each net size to allow statistics to be calculated on cross-validation results, shown as confidence intervals.

ANNs with 4, 8 and 16 hidden nodes were investigated for estimating mass recovered and loss. Estimating remediation failure was carried out using ANNs with 16, 24 and 32 hidden nodes.

ANN complexity

ANN optimization and testing performance for estimating the contaminant loss and recovery objectives is shown in Fig. 3. The ANNs with only four nodes in the hidden layer never yield satisfactory estimates for both objectives, and testing results are inconsistent despite increasing information. Increasing the net size results in improvement, while testing shows that an ANN with 16 hidden nodes is needed to obtain reliable estimation.

For the remediation failure problem, the likelihood of an incorrect response is reduced markedly by increasing net size from 16 to 24 hidden nodes, while there is no improvement by increasing again to 32. Thus the complexity of the net is a function of the problem to be solved.

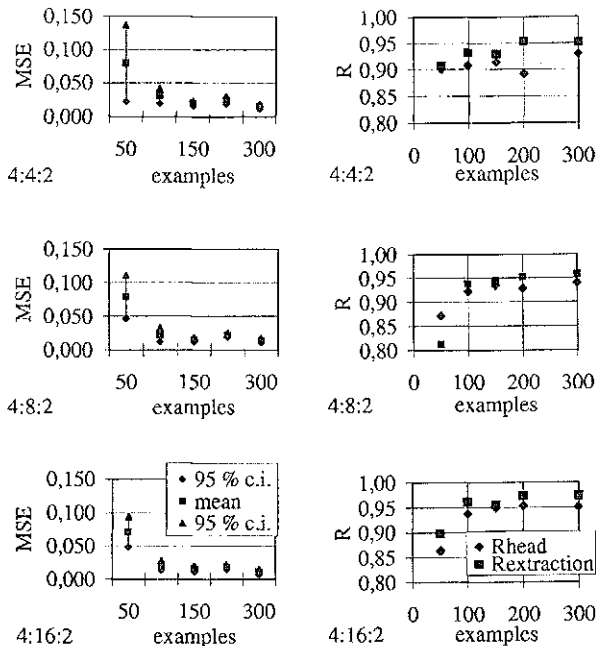


Fig. 3 ANN optimization results for contaminant loss and recovery for varying optimization data set size (shown on the abscissa) and ANN size (lower left corner). The left side is mean squared error (*mse*) of cross-validation, and the right side is test results as the regression coefficient *R* between ANN and numerical model results.

Training set size

For the mass recovery and loss problem, 50 members in the optimization data set results in very poor performance. However, a net with only four hidden nodes out

performs an ANN with more nodes at this information level, though that result is unreliable. In general the accuracy in finding reproducible minima (in a cross-validation sense) was good at and above the 100 member level.

Performance improvement and reliability gains stop above the 150 member optimization set level for both the mass recovery/loss and remediation failure problems. This is rather surprising given the complexity of the problems and shows that problem type does not affect the number of required optimization data examples.

$N_{optimize} = N_{training}$ is the best optimization scheme for estimating the contaminant loss and recovery. However, for the success-failure objective (0/1 problem) $N_{optimize} = 1$ is more successful. This is seen by comparing Figs 4(a) and 4(b).

Thus the output variability for 0/1 problems is more complex than for the continuous output from the loss-recovery problem, and to capture that the $N_{optimize} = 1$ approach is better. $N_{optimize} = 1$ is more CPU intensive as there are more optimization calculations for any given number of iterations through the optimization data set.

Contaminant loss and recovery are co-estimated and recovery is almost always estimated more precisely than the amount of contaminant crossing a boundary. This difference is independent of the amount of optimization information.

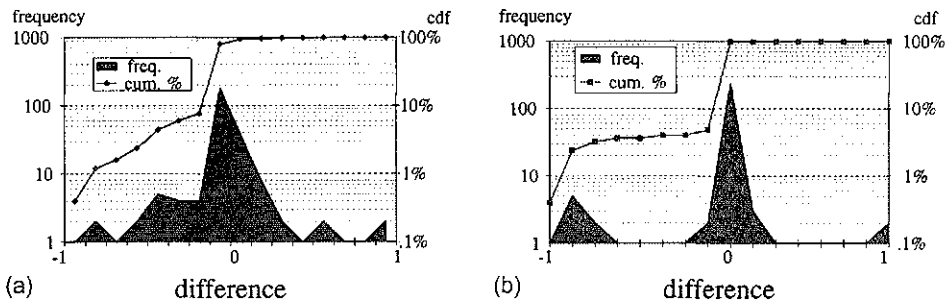


Fig. 4 Histograms of test results for the ANNs estimating the remediation success or failure: (a) is an ANN developed with $N_{optimize} = N_{trainings}$, (b) is developed with $N_{optimize} = 1$.

DISCUSSION AND CONCLUSIONS

A key aspect in developing and implementing ANNs is the need for a quantified measure for performance outside the optimization data set. We have shown that developing a statistically significant number of ANNs and calculating statistics and optimization performance in a cross validation sense illustrates the reliability of an ANN.

It has been shown that an ANN can perform at accuracy levels comparable to the numerical models used to develop the optimization data sets for two different problems. Furthermore, as an advance to what has been shown previously, ANNs can accurately determine contaminant recovery using comparatively unconstrained well placement and moderately constrained extraction rates for highly heterogeneous dynamic 3-D groundwater flow and contaminant transport systems.

At a more detailed level it is demonstrated that the choice of weight optimization scheme and ANN size is dependent on the problem to be solved. Surprisingly, the

amount of information needed to develop reliable ANNs does not seem to be dependent on the problem type, but could be controlled by the system (flow and transport regime) to be described.

REFERENCES

- Bishop, C. M. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, UK.
- Dowd, P. & Sarac, C. (1994) A neural network approach to geostatistical simulation. *Math. Geol.* 26(4), 491–503.
- Harbaugh, A. (1992) A generalized finite-difference formulation for the US Geological Survey modular three-dimensional finite difference ground-water flow model. *US Geological Survey Open-File Report 91-494. Technical Report, US Geological Survey, Reston, Virginia, USA.*
- Johnson, V. & Rogers, L. (1995) Location analysis in ground-water remediation using neural networks. *Ground Water* 33(5), 749–758.
- McDonald, M. & Harbaugh, A. (1988) A modular three-dimensional finite-difference ground-water flow model. *US Geological Survey Techniques of Water-Resources Investigations Book 6, Chapter A1. Technical Report, US Geological Survey, Reston, Virginia, USA.*
- Ranjithan, S., Eheart, J. & Garrett, J. J. H. (1993) Neural network-based screening for groundwater reclamation under uncertainty. *Wat. Resour. Res.* 29(3), 563–574.
- Rizzo, D. & Dougherty, D. (1994) Characterization of aquifer properties using artificial neural networks: neural kriging. *Wat. Resour. Res.* 30(2), 483–497.
- Rogers, L. L. (1992) Optimal groundwater remediation using artificial neural networks and the genetic algorithm. PhD Thesis, Stanford University, California, USA. UCRL-LR-114124, available from NTIS.
- Zheng, C. (1990) MT3D a modular three-dimensional transport model for simulation for advection, dispersion and chemical reactions of contaminants in groundwater systems. Technical Report prepared for US Environmental Protection Agency. S. S. Papadopoulos & Associates, Inc., Rockville, Maryland, USA.