Effects of input/output parameters on artificial neural network model efficiency for breakthrough contaminant prediction

Jayashree Pal and Dibakar Chakrabarty

ABSTRACT

Groundwater quality assessment is characterized by pollution injection rates, pollution injection locations and duration of pollution injection for identifying spatial and temporal variation. In this study, spatial variations are obtained by placing observation wells in the downstream zone. Temporal variations in contaminant concentration has been simulated during the study period. Generally, simulations are carried out using various numerical models, which are subject to the availability of all required input parameters and are necessary for the proper management of contaminated aquifers. In previous publications, artificial neural networks (ANNs) are prescribed in such situations as these modeling methods focus on available input/output datasets, thus resolving the concern of obtaining all inputs that a numerical simulator usually demands. Past studies have predicted groundwater breakthrough contaminants. But the effects of input/output variations need to be discussed. This study aims to quantify the effects of a few input/output datasets in the performance of ANN models to simulate pollutant transport in groundwater systems. The combinations of input/output scenarios have rendered these ANN models sensitive to variations, thus affecting model efficiency. These outcomes can reliably be employed for contaminant estimation and provide a paradigm in data collection that will help hydrogeologists to develop more efficient prediction models.

Key words | artificial neural network, cascade-forward backpropagation, groundwater contaminant transport, groundwater quality, model performance

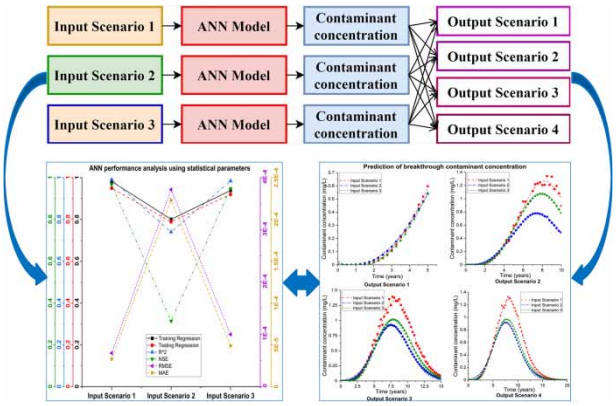
Jayashree Pal (corresponding author)
Dibakar Chakrabarty
Department of Civil Engineering,
National Institute of Technology Silchar,
Assam 788010,
India
E-mail: *jayashreepal.91@gmail.com*

HIGHLIGHTS

- A brief review on groundwater modeling using artificial neural networks.
- Effects of input and output parameters in ANN modeling.
- ANN modeling strengths and weakness in varying input/output parameters.
- The practical implication of this methodology.

doi: 10.2166/ws.2021.125

GRAPHICAL ABSTRACT



INTRODUCTION

Water resource protection and management is a vital issue for the sustenance of all living organisms on Earth. A study by Shiklomanov (1993) has documented that approximately 2.5% of the total volume of water in the hydrosphere is freshwater. A large portion of this freshwater is composed of glaciers and permanent snows. The remaining amount of the freshwater reserve is from groundwater, lakes and rivers. Groundwater constitutes about 0.76% of the total volume of water on Earth, which is 30.1% of the freshwater volume. However, the actual percentage of groundwater present in the hydrosphere has altered since 1993 due to saltwater intrusion and contamination by different human activities.

Groundwater reserves accessible by human beings are mainly in the form of aquifers. These aquifers have become vulnerable due to massive population growth, economic development, and rapid urbanization (Bierkens & Wada 2019). Consequently, literature indicating studies related to groundwater depletion (Tabari *et al.* 2012; Varni *et al.* 2013; Abiye *et al.* 2018), saltwater intrusion (Walther *et al.* 2012; Yan *et al.* 2015; Kayode *et al.* 2017; Lal & Datta 2019) and remediation (Laumann *et al.* 2013; Kazemzadeh-Parsi *et al.* 2015; Mosmeri *et al.* 2017) of various aquifer zones are immense. The methodology involves the use of various numerical models and approximation of uncertain aquifer parameters using a probability distribution.

There are some areas where the entire population is dependent on groundwater reserves. The need for sustainable usage and containment of contamination is of greater significance in these areas. Publications report work on groundwater level fluctuations for efficient utilization and future sustenance. These studies correlate flow from adjoining rivers, precipitation, runoff, temperature, evaporation, humidity and other relatable parameters for predicting groundwater level (Daliakopoulos *et al.* 2005; Khalil *et al.* 2005; Nayak *et al.* 2006; Yoon *et al.* 2011; Bisht *et al.* 2013; Gholami *et al.* 2016). This is the quantitative perspective of groundwater research, although there is also a qualitative aspect which encompasses remediation and identification of contaminated groundwater resource (Gorelick 1982; Minsker & Shoemaker 1996; Culver & Shenk 1998; Aly & Peralta 1999; Singh & Minsker 2008; Datta *et al.* 2009; Singh & Chakrabarty 2011; Chakraborty & Ghosh 2012; Milašinović *et al.* 2019).

The quantitative and qualitative studies using numerical models require information from the hydrogeological survey and man-made interventions. The hydrogeological information provides assistance in developing models analogous to field conditions. But, obtaining accurate field data as well as replicating those in numerical models becomes difficult. In such cases, these complex problem-solving numerical methods become redundant. In order to simplify such studies and derive site-specific models from predicting groundwater level and contaminant concentration at

desired locations, an artificial neural network approach is mostly considered as an alternative. However, the application of ANNs to solve various hydrological issues has been discussed in detail in ASCE Task Committee (2000b), including areas of groundwater hydrology as well. The implementation of ANNs has been reported in several case studies to predict nitrate, fluoride, arsenic, manganese and salt concentrations in groundwater (Pal *et al.* 2002; Mousavi & Amiri 2012; Sinha & Saha 2015; Wagh *et al.* 2017; McArthur *et al.* 2018). Some case studies have also reported the use of other machine learning algorithms like the random forest algorithm, support vector machine, locally weighted projection regression, relevance vector machines, etc., that consider groundwater quality index as the objective (Khalil *et al.* 2005; Podgorski *et al.* 2018). A brief survey focusing on both qualitative and quantitative groundwater modeling using ANN has been summarized in Table 1. This table includes the findings, ANN method used and input/output details of each article. These hydrogeologists have reported works that aimed at various perspectives of research. The referred studies have shown the immense utilization of ANNs in prediction research, although those involving the effects of input/output relationship on ANN model performance have remained unexplored.

NEEDS OF THE STUDY

From the survey, it was observed that input parameters play an important role in neural network prediction. Das *et al.* (2019) predicted water table depth in five different input scenarios. They used backpropagation neural network (BPNN) and Adaptive Neuro-Fuzzy Inference System (ANFIS) models for prediction, describing a quantitative study in the groundwater system. They identified and reported the appropriate input parameters for their study. Therefore researchers need to undertake such analysis of groundwater where the input and output parameters variation becomes significant. The influence of inputs and outputs in qualitative aspects of groundwater studies using ANN has been unexplored. The study reported in this paper deals with various combinations of input/output scenarios and training algorithms/transfer functions that can reliably be employed for modeling pollutant transport simulations in groundwater systems. These input and output combinations are calibrated depending on their performances, providing direction to the hydrogeologists for decision-making in data collection. The input parameters considered for this study are injection rates and injection locations. Each input parameter is tested individually to obtain contaminant concentration throughout the simulation period of 20 years. The second aim of this study is to evaluate ANN performance when the simulation period is varied. Here, the estimation of contaminant concentration in four different time spans, such as 20 years, 15 years, 10 years, and 5 years, has been reported. The effect of individual input parameters due to the reduction in prediction time has also been analyzed. ANN models are used to predict the breakthrough concentration over time and identify the significance of input/output relationships in ANN training and testing.

METHODS AND MATERIALS

Data description

The numerical model of groundwater flow and transport considered for the generation of patterns is SUTRA-USGS (A Model for Saturated-Unsaturated, Variable-Density Ground-Water Flow with Solute or Energy Transport) (Voss and Provost 2010). A two-dimensional hypothetical aquifer system of dimensions $1{,}500\ \text{m} \times 1{,}400\ \text{m} \times 40\ \text{m}$ is considered for the study. The flow of groundwater occurs from the left boundary (hydraulic head = 100 m) to the right boundary (hydraulic head = 88 m), while the top and bottom boundaries are impermeable zones, as shown in Figure 1. Single pollutant species have been considered for the study, which is conservative in nature and the permissible limit of this pollutant is assumed to be 0.5 mg/L. There are three contaminant injection locations (point sources) in the aquifer at the upstream zone. The contaminant plume propagates through the aquifer towards the downstream zone. There are four observation wells that are fixed at random locations in the downstream region across the flow path. The simulation of this aquifer system

Table 1 | Brief review on groundwater modeling using artificial neural networks (ANN)

Sl. no.	Authors	Findings	ANN method	Training algorithm; transfer function	Input parameters	Output parameters
1.	Rogers (1992)	► Predict injection and pumping rates for pollution containment	Feed-forward backpropagation	Conjugate gradient Polak–Ribiere weight update rule; Sigmoidal	Pumping realizations at three remediation wells	Successful remediation, unsuccessful remediation
2.	Morshed & Kaluarachchi (1998)	► Simulate breakthrough concentration ► Compare two ANN training methods	Feed-forward backpropagation genetic algorithm	Generalized delta rule; Sigmoidal, Tangent sigmoidal	Grain size distribution index, saturated hydraulic conductivity, water flux, dispersivity, decay coefficient, Freundlich coefficient, Freundlich exponent	Breakthrough concentration curve
3.	Gümrah *et al.* (2000)	► Forecast pollutant concentrations and hydraulic heads. ► Short-term predictions proved more efficient than long-term predictions	Feed-forward backpropagation	Gradient descent; Sigmoidal	Time, concentration, head, neighbor well concentration	Chlorine concentration and head at next time step
4.	Kumar & Jain (2006)	► Estimate groundwater pollution sources from breakthrough curves data	Feed-forward backpropagation	Generalized delta rule; Sigmoidal	Breakthrough concentration curve at observation location	Groundwater pollution source
5.	Prasad & Mathur (2007)	► Identification of the uncertainty of groundwater flow and contaminant transport with imprecise parameters	ANN-GA backpropagation algorithm	Levenberg-Marquardt; Tangent sigmoidal	Seepage velocity, longitudinal dispersivity, transverse dispersivity, time	Groundwater level, concentration
6.	Banerjee *et al.* (2011)	► Prediction of safe pumping rate to prevent health hazards	Feed-forward quick propagation	Discrete pseudo-Newton method	Groundwater electrical conductivity, pumping, time, rainy period, water level	Groundwater salinity
7.	Khalil *et al.* (2014)	► Forecasting groundwater level depending on precipitation, mean temperature and tailings recharge	i. Multiple linear regression ii. Artificial neural network iii. Wavelet transform (W-MLR, W-ANN) iv. W-ensemble ANN	Levenberg–Marquardt	Tailings recharge, precipitation, mean temperature	Groundwater level
8.	Khaki *et al.* (2015)	► Simulation of decreasing trend of groundwater level	i. Feed-forward backpropagation ii. Cascade-forward backpropagation iii. ANFIS[a]	i. Levenberg-Marquardt ii. Hybrid learning iii. Algorithm for ANFIS; Tangent Sigmoidal	Rainfall, humidity, evaporation, minimum temperature, maximum temperature	Groundwater level
9.	Wagh *et al.* (2018)	► Prediction of nitrate concentration in groundwater of Kadava River Basin	i. Backpropagation ii. Backpropagation with weights iii. Resilient backpropagation with weights iv. Resilient backpropagation without weights v. Smallest absolute derivative vi. Smallest learning rate	Levenberg-Marquardt; Sigmoidal	Electrical conductivity, total dissolved solids, total hardness, magnesium, sodium, chlorine and sulphate	Groundwater nitrate concentration

(*continued*)

Table 1 | continued

Sl. no.	Authors	Findings	ANN method	Training algorithm; transfer function	Input parameters	Output parameters
10.	Das *et al.* (2019)	► Prediction of water table depth based on precipitation, runoff, temperature, humidity and evapotranspiration	i. Feed-forward ii. Backpropagation iii. ANFIS[a]	i. Gradient descent ii. Adaptive learning	Precipitation, maximum temperature, minimum temperature, average temperature, evapotranspiration losses, runoff, humidity	Water table depth
11.	Pal & Chakrabarty (2020)	► Simulate contaminant concentration based on injection rates and injection locations	i. Feed-forward backpropagation ii. Cascade-forward backpropagation	14 training algorithms like Bayesian regularization, conjugate gradient, Levenberg–Marquardt, one-step secant and so on; Pure linear, Sigmoidal, Tangent sigmoidal	Injection rate, injection location	Breakthrough curve of contaminant concentration
12.	Bedi, *et al.* (2020)	► Prediction of contamination levels using sparse data. ► Evaluation of classification performance of models. ► Assessment of class imbalance in hyperparameter tuning	i. Artificial neural networks ii. Support vector machines iii. Extreme gradient boosting		Hydrogeologic, land use and water quality	Nitrate and pesticide concentration

[a]Adaptive neuro-fuzzy inference system.

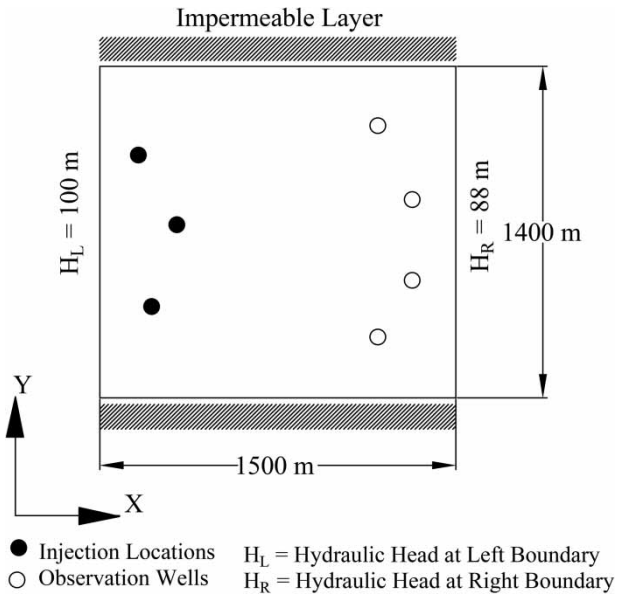


Figure 1 | Illustrative aquifer system used in the study (pollutant sources locations represent one pattern of the ANN dataset).

requires additional aquifer parameter information, which has been provided in Table S1. The numerical simulator uses information from Table S1 to generate a pattern for the ANN models.

Box *et al.* (2015) has discussed time-series forecasting models. In this study, two input parameters have been used for evaluating time-series breakthrough concentrations (BTC) at downstream water supply wells. The first parameter is injection rate and the second is injection location. The BTC estimated contaminant concentrations at observation wells constitute ANN output. This problem is functional to protect the population dependent on those water supply wells from contaminated water. The knowledge of contaminant concentration values over a time span will enable site engineers to undertake necessary remediation or containment plans. This study compares the efficiency of individual input attributes in the contaminant prediction process under varying simulation time of the study. Thus, the effects of input/output parameters are considered by incorporating a number of combinations of input/output parameters, training algorithms and activation functions in the performance analyses of the developed ANN models.

The contaminant injection occurs for the first 5 consecutive years at three source locations. The injection rate changes yearly for all the sources during those 5 years. Beyond this period, no contaminant injection was done. The injection rates are uniformly distributed random

values ranging between 20 and 70 L/s. The injection locations are integer random numbers that range between 200 and 1,200 m. Each set of injection rate and injection location represents a pattern/case for ANN modeling. The number of cases generated from different combinations in pollutant injection rate and respective injection locations used in this study is 120. This dataset has been divided into two parts, that is, 84 patterns (70% of data) for ANN training and the remaining 36 patterns (30% of data) for ANN testing. The contaminant plume movement has been monitored from injection time (initial time) up to 20 years. The plume concentration data are recorded at 3 month intervals for 20 years, that is, 80 data points from each observation well. These data represent the breakthrough curves at each observation well. This time-series dataset of four observation wells is placed consecutively in ANN output. The input and output data have been normalized by mapping row minimum and maximum method in the interval [−1, 1] for standardization before training and testing to enhance modeling speed and performance of ANN using the following equation:

$$\hat{X} = \{(R_{max} - R_{min}) * (\mathrm{X} - \mathrm{X}_{\min})/(\mathrm{X}_{\max} - \mathrm{X}_{\min})\} + R_{min}$$

where $\hat{X}$ = normalized value; X = input/output value to be normalized; X_{min} = minimum input/output of the dataset; X_{max} = maximum input/output of the dataset, R_{min} = minimum range of the normalization interval, and R_{max} = maximum range of the normalization interval. The input (I) and output (O) vectors are represented as:

$$I = \{IR_{11}, \ldots, IR_{15}, IL_{1X}, IL_{1Y}, IR_{21}, \ldots, IR_{25}, IL_{2X}, IL_{2Y}, IR_{31}, \ldots, IR_{35}, IL_{3X}, IL_{3Y}\}$$

$$O = \{1O_1, \ldots, 1O_{20}, 2O_1, \ldots, 2O_{20}, \ldots, 3O_1, \ldots, 3O_{20}, \ldots, 4O_1, \ldots, 4O_{20}\}$$

where IR_{15} = injection rate at source 1 in the 5th year; IL_{2X} = injection location at source 2 in x-direction; and $3O_{20}$ = contaminant concentration of monitoring well number 3 at 20 years.

Artificial neural network

The artificial neural network (ANN) is a very sophisticated information-processing paradigm that imitates the functioning of the human nervous system to identify patterns within a dataset (McCulloch & Pitts 1943). The learning process of ANNs to solve a problem is analogous to the central nervous system as it is capable of developing a memory of a large number of associated input/output patterns and provides outputs for unknown input patterns. This helps the ANN model to interpret a wide range of complex problems such as nonlinear modeling, prediction, control, pattern recognition and classification (ASCE Task Committee 2000a). There are two broader classifications of ANN model based on the dependence of the learning function/training algorithm. The formation of the ANN model guided by a training algorithm is referred to as a supervised neural network, and that not guided by a training algorithm is known as an unsupervised neural network (Haykin 1999). This study involves a supervised neural network model. The supervised ANN model constitutes three components such as modeling method, training algorithm and activation function. The following section discusses the modeling method and its components.

Cascade-forward backpropagation neural network

The present study employs a cascade-forward backpropagation neural network as the ANN modeling method for developing pollutant transport models in groundwater systems. As many as five better performing ANN methods (as detailed in Table 2), with different combination of training algorithms and transfer functions, were taken from the literature (Pal & Chakrabarty 2020) to study the effects of different input/target data sets on the ANN model performances. This supervised multi-layer perceptron (MLP) has self-adaptive network parameters regulated by the combined influence of training vector and reverse error signal. The backpropagation algorithm (BPA) applied for the study is a gradient descent with momentum method, which assists neural networks to minimize errors caused by a mismatch between the target value of data (in this case, it is SUTRA data) and the output value produced by the network. The errors obtained during the first iteration are propagated

Table 2 | Details of neural network training algorithm and transfer function

Cascade-forward backpropagation neural network (C)

Training algorithm		Activation function		
Name	**MATLAB function**	**Name**	**MATLAB function**	**Abbreviation**
Conjugate gradient Fletcher–Reeves updates (**CGF**)	traincgf	Hyperbolic tangent sigmoid	tansig (**T**)	**CCGFT**
One-step secant (**OSS**)	trainoss			**COSST**
Gradient descent with adaptive learning (**GDA**)	traingda	Sigmoid	logsig (**L**)	**CGDAL**
Resilient (**RP**)	trainrp			**CRPL**
Conjugate gradient Powell–Beale restarts (**CGP**)	traincgp			**CCGPL**

backwards through each node, followed by modification in corresponding connection weights and bias of the network. The network trains again in the forward direction using a training algorithm (stated in the following section) to evaluate a new output set of data for evaluation of errors in the second iteration. The network development process continues until the desirable minimum error is reached. In the cascade-forward modeling method, there is an additional network connection from the input layer directly to the output layer. A representative neural network structure is shown in Figure 2. This property of CFBNN makes it more sensitive compared with a feed-forward backpropagation neural network. For ANN model development, Neural Network Toolbox in MATLAB version 2020a has been used. The system configuration for neural network training is Intel® Core(TM) i3-6006 U (64 bit), 12 GB RAM and 2.00 GHz processing speed.

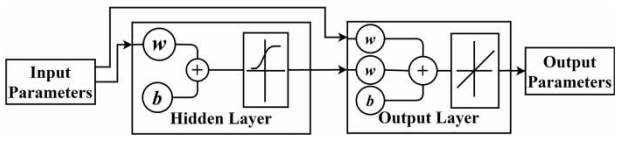


Figure 2 | Cascade-forward backpropagation neural network using a tangent sigmoidal transfer function.

Gradient descent with momentum (GDM) algorithm used for backpropagation of errors aids the network to act upon the local gradient as well as focuses on variations in error surface. The momentum factor prevents the network from getting stuck in a shallow local minimum, thus ignoring small features in error surface. GDM has two vital controlling parameters known as the learning rate (*lr*) and the momentum constant (*mc*). Learning rate is negative of the gradient, which evaluates the change in the weights and bias. The larger the learning rate, the faster is the convergence and vice versa. But too large a value can also lead to model instability. The momentum constant can be explained as the amount of impetus gained for model convergence. A momentum value of zero signifies no momentum, while unity signifies high momentum. The formulation of GDM is shown as:

$$\Delta X = mc * \Delta \mathrm{X}_{\mathrm{prev}} + \mathrm{lr} * (1 - \mathrm{mc}) * \frac{\partial \mathrm{E}}{\partial \mathrm{X}}$$

where ΔX = change in weight/bias; ΔX_{prev} = previous change in weight/bias; E = mean squared error of the network, which is computed using the following equation:

$$E = \frac{1}{Nn} \sum_{N} \sum_{n} (\boldsymbol{Y} - \boldsymbol{T})^2$$

where $\boldsymbol{Y}$ = output vector (y_1, y_2, y_3,…, y_n) obtained from forward training of the neural network; $\boldsymbol{T}$ = target vector (t_1, t_2, t_3,…, t_n), that is, actual output data provided during training; n = number of nodes in output layer; N = number of training patterns.

Training algorithm

The role of the training algorithm in neural network modeling is to optimize connection weights and bias for approximating target vectors and enhance network performance. The training algorithms are primarily optimizing technique which supervises training at each epoch in supervised neural networks. Hence, these can be conventional,

heuristic as well as meta-heuristic in nature. Here, only conventional optimization techniques have been used for modeling. The training algorithms adopted for this study based on the performance are conjugate gradient and Fletcher–Reeves updates (CGF), one-step secant (OSS), gradient descent with adaptive learning (GDA), resilient (RP) and conjugate gradient and Powell–Beale restarts (CGB).

Activation function

The activation function controls the activation or de-activation of a neuron depending upon network weights and bias. The mathematical relation of the total input signal entering a neuron to obtain relevant output in CFBNN (Warsito *et al.* 2018) is represented as:

$$y^{n+1}(k) = \sum_{j=1}^{nh} w^{n}(1,j)x^{n}(j) + \sum_{j=1}^{nh} w^{n+1}(i,j)x^{n}(j) + b^{n+1}(i)$$

where, n = total number of hidden layer; nh = total number of neurons in a hidden layer; $x^n(j)$ = input attributes of nth layer to jth hidden neuron, $w^{n+1}(i,j)$ = network connection weight matrix of $(n+1)$th hidden layer from ith input attribute to jth hidden neuron; $b^{n+1}(i)$ = bias of the network to ith input attribute; $y^{n+1}(k)$ = kth output attribute from $(n+1)$th hidden layer. The optimal number of hidden neurons used for training CCGFT, COSST, CRPL, CGDAL and CCGBL models are 60, 70, 78, 30 and 40, respectively. The types of activation functions used in this study are sigmoid and hyperbolic tangent sigmoid, which are denoted respectively as:

$$f(y) = 1/(1+e^{-y})$$

$$f(y) = (e^{y} - e^{-y})/(e^{y} + e^{-y})$$

Performance analysis of ANN model

The developed ANN models are capable of simulating the transport of contaminant concentration, which is BTC curves of 20 years under varying contaminant source locations and contaminant injection rates. Therefore, contaminant locations and contaminant injection rates constitute the input attributes. The output attributes include the contaminant concentration at the monitoring well locations for 20 years. These outputs from the prescribed ANN models are compared with the outputs of groundwater numerical model results using statistical parameters.

The statistical measures used to determine ANN performance are as follows (Nie *et al.* 2017):

(a) coefficient of determination (R^2):

$$R^2 = \frac{\sum_{i=1}^{N}(O_i - \bar{O})(P_i - \bar{P})^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2 \sum_{i=1}^{N}(P_i - \bar{P})^2}$$

(b) Nash–Sutcliffe efficiency coefficient (NSE):

$$NSE = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2}$$

(c) root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2}$$

(d) mean absolute error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i|$$

where, P_i = predicted concentration at time t, O_i = estimated concentrations at time t, $\bar{P}$ = mean of predicted concentrations, $\bar{O}$ = mean of estimated concentrations, N = total number of output attributes. The best fit neural network model is the one that has R^2 and NSE values tending to one, while RMSE and MAE are approaching zero.

ANN model information

The aim of this study is to identify the effects of various combinations of inputs and outputs in estimating the

breakthrough curve at four observation wells. ANN model classification and respective ANN model identifiers have been summarized in Figure 3. There are 12 combinations formed out of different input/output parameters. The abbreviation of the ANN model accompanied by the model number shown in the rectangular box of each link in the figure constitutes the model identifier. These model identifiers are used as a reference for depicting model performances depending on statistical analysis, which has been reported in results and discussion.

RESULTS AND DISCUSSION

Effect of input parameters on ANN performance

The 12 models shown in Figure 3 have been evaluated based on statistical parameters. In order to assess ANN model performance, the testing dataset has been used. The input parameters of the testing dataset have been provided to these ANN models. The predicted output is generated by these ANN models for respective input patterns are obtained. The predicted output and the estimated target values are compared to identify the effect on performance due to the variation in input parameters. The output parameters are considered to be 20 years BTCs. Figure 4(a) provides a vivid idea of ANN performance under different input scenarios. It has been observed that input scenario 1 has proved to perform better than the other two scenarios. Input scenario 2 has shown considerable reduction in model efficiency, thus leading to low regression, R^2 and NSE, and high errors. Input scenario 3 has shown good performance; that is, the correlation of output and target dataset is more than 90%. However, this correlation value is not as good as in the case of input scenario 1, which is above 97% for five ANN methods. The model training time for the majority of cases ranges from few seconds to five minutes. Only CCGFT9 has exhibited greater central processing unit (CPU) time of approximately 19 minutes, as mentioned in Table S2.

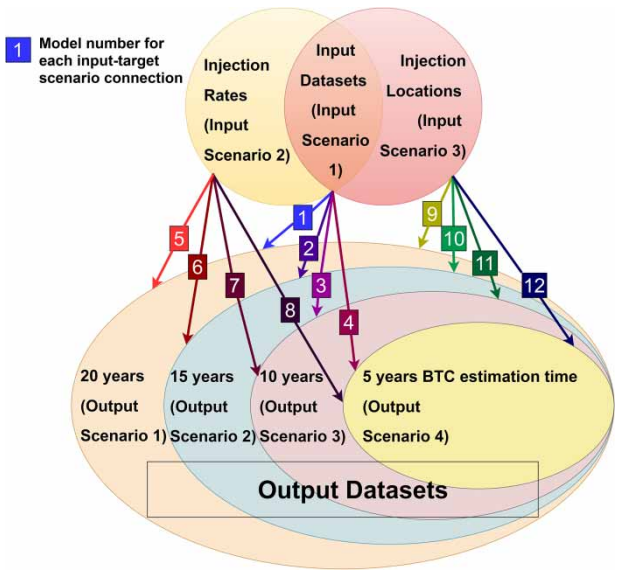


Figure 3 | ANN model combination and model identifiers.

Apart from the factors like ANN modeling method and input scenarios, there are some termination criteria during model training that contribute to model efficiency. These termination criteria are user-defined on a trial-and-error basis. The ANN models trained in this study consist of the following termination criteria: epoch = 1.0e6; performance/mean squared error limit = 1.0e-04; gradient = 1.0e-10; validation checks = 10,000; and step size = 1.0e-6. The attainment of any one of the termination criteria is necessary to stop the training process. Often, modification in termination criteria is performed done to minimize the slight overfitting of ANN data during training and enhance the quality of testing data.

Effects of output parameters on ANN performance

In the previous section, model performance has been interpreted under different input scenarios. So, this section focuses on the performances of ANN models for varying contaminant concentration estimation periods. Each input scenario is linked with four different time-series predictions. The performance of models for 20 years contaminant prediction, that is, output scenario 1, has already been discussed. The performance analyses of the remaining three output scenarios are reviewed as follows.

Fifteen years contaminant concentration prediction (output scenario 2)

The reduction in the estimation period from 20 years to 15 years has shown some evident difference, as shown in

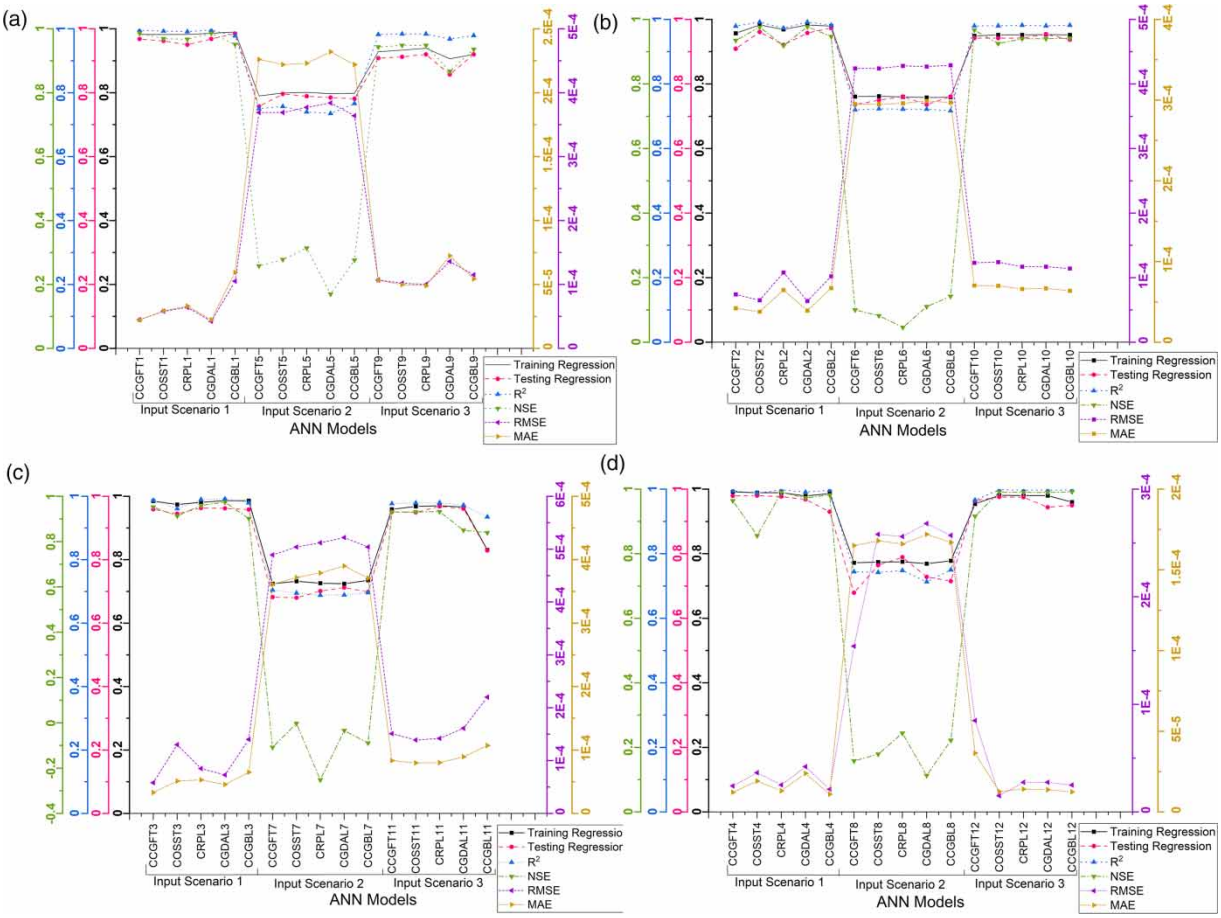


Figure 4 | Performance analysis of ANN model (a) Output scenario 1, (b) Output scenario 2, (c) Output scenario 3 and (d) Output scenario 4.

Figure 4(b). The selected ANN models have lowered performance in input scenario 1. The training regression, testing regression, and coefficient of determination have reduced by 2–3%. Hence, the error parameters like RMSE and MAE have increased in output scenario 2. The trend observed in input scenario 2 is the same as input scenario 1. But, input scenario 3 has performed better with output scenario 2, which contrasts with output scenario 1. The model development time is the same as earlier.

Ten years contaminant concentration prediction (output scenario 3)

While estimating the 10 years contaminant concentration dataset, ANN performance has shown some peculiarities (Figure 4(c)). In both the input parameters, training regression is nearly the same as output scenario 1. In contrast, the remaining statistical parameters do not represent better performance. For input scenario 2, the overall performance of ANN models has deteriorated compared with output scenario 1 with input scenarios 1 and 2. In the case of input scenario 3, all ANN models have performed moderately.

Five years contaminant concentration prediction (output scenario 4)

For five years of contaminant estimation, all the input scenarios have out-performed compared with the previously mentioned models when analyzed respectively to each input scenario (Figure 4(d)). But, the performance rate within input scenarios shows a similar trend as discussed in output scenario 1. The reduction of contaminant concentration estimation by 75% of the actual prediction has enhanced model performance.

Illustrative example

A pattern from ANN testing dataset has been considered to represent the breakthrough curve generated using these ANN models. The details of ANN input parameters have been provided in Table 3. There are four observation wells in the downstream region of the confined aquifer. Each of these observation locations has respective BTC curves obtained from the numerical model SUTRA for this illustrative example. The inputs of this example are provided to each of the ANN models as per input scenarios to predict 20 years, 15 years, 10 years and 5 years outputs.

There are four BTC contaminant concentrations from each of the developed ANN models. Therefore, a total of 48 BTC has been estimated from 12 ANN models and four observation wells. Only a few representative breakthrough curves for all the input/output scenarios have been shown. For output scenario 1, the models predict 20 years contaminant concentrations and the corresponding BTC from OBS 3 has been denoted in Figure 5(a). Figure 5(b) shows 15 years prediction (output scenario 2) of BTC at OBS 2. The

Table 3 | ANN inputs of illustrative example

ANN attributes		Source 1	Source 2	Source 3
Contaminant concentration (L/s)	Year 1	37.3	56.5	54.3
	Year 2	68.5	22.3	45.9
	Year 3	22.9	68.3	60.4
	Year 4	28.3	49.6	69.3
	Year 5	21.2	68.8	58.7
X co-ordinate (m)		500	300	300
Y co-ordinate (m)		250	300	650

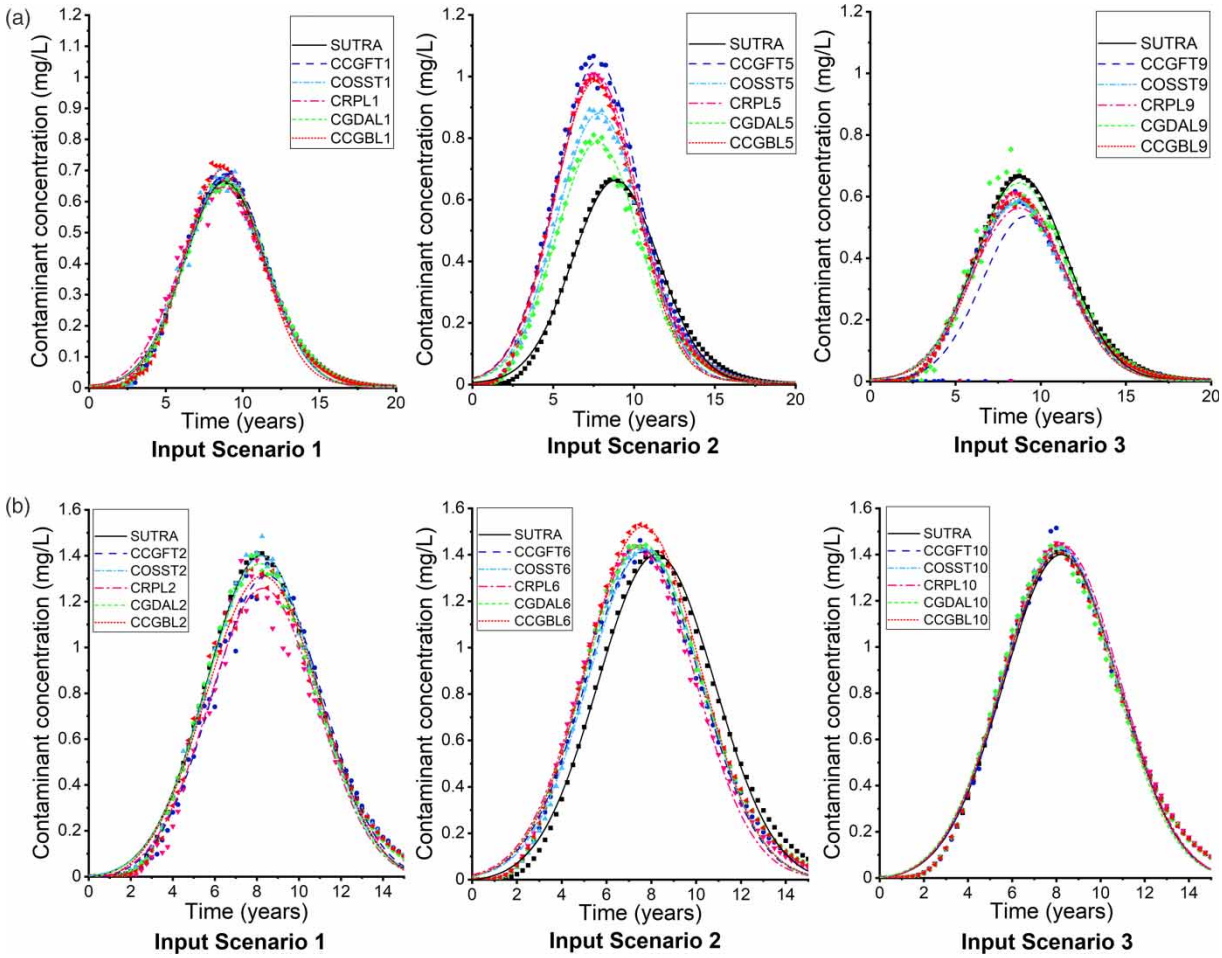


Figure 5 | Breakthrough contaminant concentration of (a) 20 years at OBS 3, (b) 15 years at OBS 2, (c) 10 years at OBS 4 and (d) 5 years at OBS 1. (*continued*).

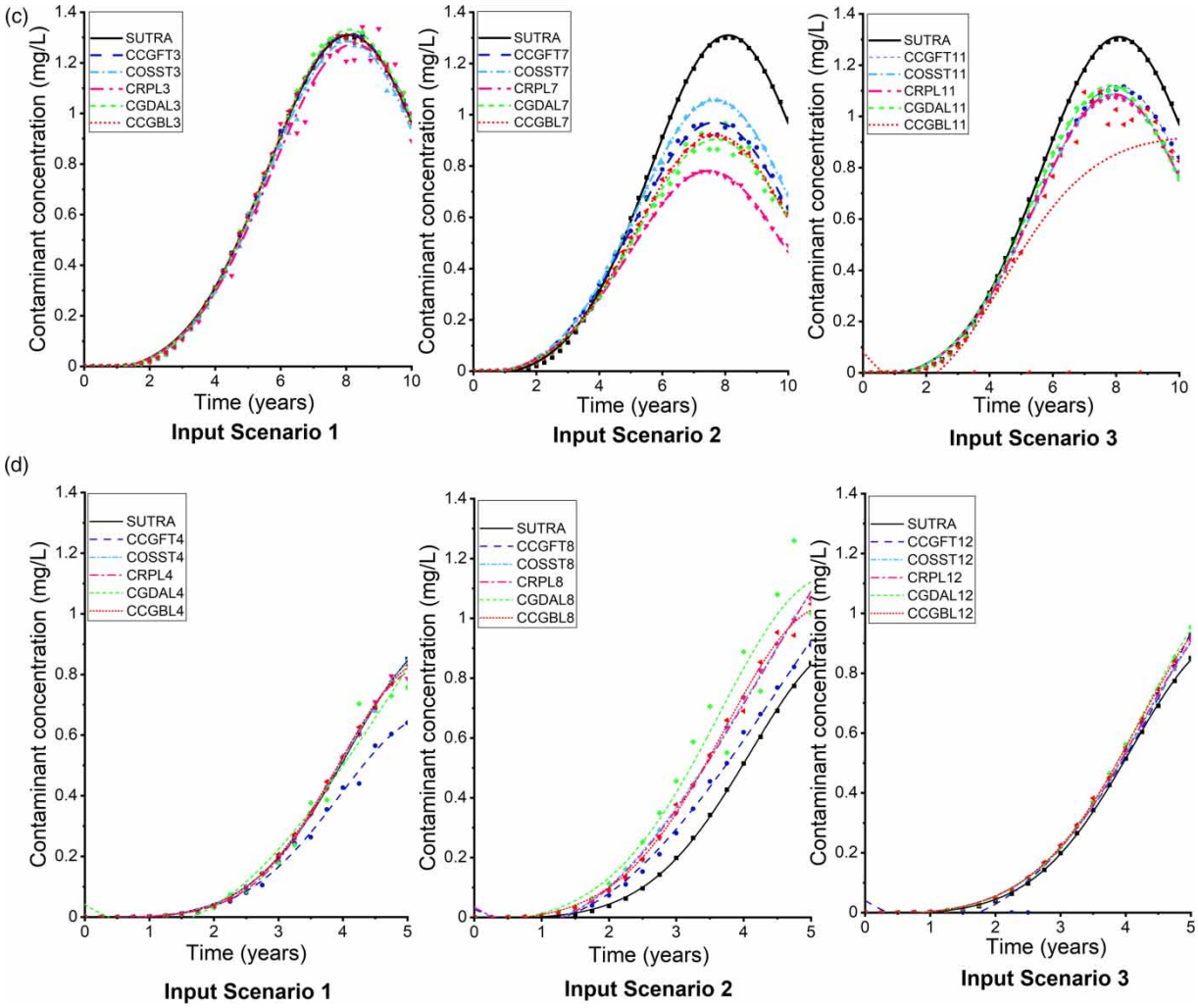


Figure 5 | continued.

BTC estimated for 10 years (output scenario 3) at OBS 4 has been depicted in Figure 5(c). The result obtained from output scenario 4 at OBS 1 is plotted in Figure 5(d). All the ANN predicted BTCs had been compared with SUTRA concentrations to identify the difference between them. Moreover, the input scenarios are also shown in Figure 5 to visualize their effect on ANN model development.

Outcome of the study

The performance evaluation for different input/output scenarios has established the potential applicability of this methodology. For input scenario 1, both training and testing regression in combination with all the output scenarios are maxima ranging between [0.9, 0.98] which shows a similar trend but different value in different output scenarios. The outcome of training and testing regression for input scenario 2 by varying output scenarios is minimum ranging between [0.7, 0.8]. The regression range for input scenario 3 is less than input scenario 1, ranging between [0.85, 0.95]. The other two statistical parameters that represent goodness-of-fit are coefficient of determination (R^2) and Nash–Sutcliffe efficiency coefficient (NSE). The significance of R^2 is the variability of ANN output from the mean, where the mean denotes the fitted regression line between ANN output and SUTRA output (Ross 2014). At the same time, NSE segregates ANN output into three categories on a normalized scale (Knoben *et al.* 2019). NSE = 1 indicates ANN output is similar to

SUTRA target, NSE = 0 indicates that ANN output is equivalent to the mean of target dataset and NSE < 0 denotes that the model is the worst predictor. Thus, NSE value above zero implies a good model and below zero implies a bad model. Input scenarios 1 and 3 prove to be good predictor models for all target scenarios. However, input scenario 2 in combination with output scenario 3 is a poor predictor model as NSE value is negative and R^2 value is less than 0.7. For other output scenario combinations with input scenario 2 also suffer from performing better. Depending on the model performance discussed so far, the error parameters show corresponding outcomes for R^2 and NSE. The poor predictor models have higher error values, whereas the good models have a minimum error.

In addition, the prediction of this model based on the illustrative example also exhibits a similar trend in model performance. The BTCs of observation wells at 3 month intervals have been fitted to clearly present ANN and SUTRA outputs. These BTC plots identify both strengths and vulnerabilities of ANN models under various input/output scenarios. Figure 5 is representative of the model performances at 20 years, 15 years, 10 years, and 5 years prediction periods. The optimum prediction period is 15 years in which the performance of five ANN modeling methods has been observed maximum with minimum error values for all input scenarios. In output scenario 3 (prediction period = 10 years), the predicted concentration differs from SUTRA outputs to a greater extent; thus, MAE and RMSE are greater than output scenarios 1, 2, and 4. The 10-year prediction model shows RMSE and MAE values more than 5E-5 and 2.5E-5, respectively, which is greater than three other prediction models when two input parameters are considered. Over-estimation of concentration values has been observed in input scenario 2 in combination with output scenarios 1, 2 and 4, with an exception in the case of output scenario 3.

In Morshed & Kaluarachchi (1998), both flow and transport parameters are considered for ANN modeling. Among several parameters, few parameters determining flow and transport in groundwater system have been selected by them for developing ANN models in order to predict contaminant breakthrough concentrations. This study involves only transport parameters as the flow parameters are constant for the developed training patterns. Das *et al.* (2019) have reported that precipitation, average temperature and humidity are the three significant input parameters for predicting water table out of the five input parameters. In this groundwater quality study, it can be inferred from the statistical analysis of ANN models that both of the input parameters, that is, injection rate and injection location, are essential for solute concentration prediction. The changes observed in model efficiency are minor due to variation in output parameters.

CONCLUSIONS

The limited study reveals that there are multiple ANN models encompassing combinations of different input/output datasets and different training algorithms/transfer functions expected to perform reliably in estimating pollutant transport in groundwater systems. A total of 60 ANN models are generated from the mentioned combinations that show satisfactory results based on statistical parameters like the coefficient of determination, Nash–Sutcliffe coefficient, RMSE and MAE. However, it has been identified that five ANN methods have performed reasonably well under input scenarios 1 and 3, whereas model performance for input scenario 2 is comparatively inferior. In addition, the study highlights that even though out of the five ANN models- CCGFT has been reported to perform better than the rest of them in the literature, inconsistency in its performance has been observed upon varying input/output parameters. While predicted breakthrough concentrations by COSST (cascade-forward backpropagation, OSS, tangent sigmoid) and CGDAL (cascade-forward backpropagation, GDA, tangent sigmoid) models are found to be equivalent to SUTRA outputs. There are some limitations of these ANN models as efficiency can reduce when injection rates and injection locations are beyond the considered range and due to erroneous data. Despite few limitations, this study shows that variation in input/output parameters has significant impacts on model efficiency indeed. This study also reveals that ANN inputs form the determining factor for contaminant prediction, whereas the output parameter has proved to have a very nominal effect. Therefore, the need to recognize crucial input parameters has a high influence on these prediction models, thus proving

this work to be inevitable prior to ANN application. Hence, the significance of input/output variations can assist the hydrogeologists to consider performing such analyses beforehand in order to build more efficient prediction models.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Abiye, T., Masindi, K., Mengistu, H. & Demlie, M. 2018 Understanding the groundwater-level fluctuations for better management of groundwater resource: a case in the Johannesburg region. *Groundwater for Sustainable Development* **7**, 1–7.

Aly, A. & Peralta, R. C. 1999 Optimal design of aquifer cleanup systems under uncertainty using a neural network and a genetic algorithm. *Water Resources Research* **35** (8), 2523–2532.

ASCE Task Committee 2000a Artificial neural networks in hydrology. I: preliminary concepts. *Journal of Hydrologic Engineering* **5** (2), 115–123.

ASCE Task Committee 2000b Artificial neural networks in hydrology. II: hydrologic applications. *Journal of Hydrologic Engineering* **5** (2), 124–137.

Banerjee, P., Singh, V. S., Chatttopadhyay, K., Chandra, P. C. & Singh, B. 2011 Artificial neural network model as a potential alternative for groundwater salinity forecasting. *Journal of Hydrology* **398** (3–4), 212–220.

Bedi, S., Samal, A., Ray, C. & Snow, D. 2020 Comparative evaluation of machine learning models for groundwater quality assessment. *Environmental Monitoring and Assessment* **192** (2), 776.

Bierkens, M. F. P. & Wada, Y. 2019 Non-renewable groundwater use and groundwater depletion: a review. *Environmental Research Letters* **14** (6), 063002.

Bisht, D., Jain, S. & Raju, M. M. 2013 Prediction of water table elevation fluctuation through fuzzy logic & artificial neural networks. *International Journal of Advanced Science and Technology* **51**, 107–120.

Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. 2015 *Time Series Analysis: Forecasting and Control*, 5th edn. John Wiley & Sons, Hoboken, New Jersey.

Chakraborty, R. & Ghosh, A. 2012 Analysis of 1D contaminant migration through saturated soil media underlying aquifer using FDM. *Journal of Hazardous, Toxic, and Radioactive Waste* **16** (3), 229–242.

Culver, T. & Shenk, G. 1998 Dynamic optimal ground water remediation by granular activated carbon. *Journal of Water Resources Planning and Management* **124** (1), 59–64.

Daliakopoulos, I. N., Coulibaly, P. & Tsanis, I. K. 2005 Groundwater level forecasting using artificial neural networks. *Journal of Hydrology* **309** (1–4), 229–240.

Das, U. K., Roy, P. & Ghose, D. K. 2019 Modeling water table depth using adaptive neuro-fuzzy inference system. *ISH Journal of Hydraulic Engineering* **25** (3), 291–297.

Datta, B., Vennalakanti, H. & Dhar, A. 2009 Modeling and control of saltwater intrusion in a coastal aquifer of Andhra Pradesh, India. *Journal of Hydro-Environment Research* **3** (3), 148–159.

Gholami, V., Sebghati, M. & Yousefi, Z. 2016 Integration of artificial neural network and geographic information system applications in simulating groundwater quality. *Kerman University of Medical Sciences* **3** (4), 173–182.

Gorelick, S. M. 1982 A model for managing sources of groundwater pollution. *Water Resources Research* **18** (4), 773–781.

Gümrah, F., Öz, B., Güler, B. & Evin, S. 2000 The application of artificial neural networks for the prediction of water quality of polluted aquifer. *Water, Air, and Soil Pollution* **119** (1–4), 275–294.

Haykin, S. 1999 *Neural Networks and Learning Machines, McMaster University*, 3rd edn. Pearson Prentice Hall, Hamilton, Ontario, Canada, pp. 34–36.

Kayode, O., Odukoya, A. & Adagunodo, T. 2017 Saline water intrusion: its management and control. *Journal of Informatics and Mathematical Sciences* **9** (2), 493–499.

Kazemzadeh-Parsi, M. J., Daneshmand, F., Ahmadfard, M. A., Adamowski, J. & Martel, R. 2015 Optimal groundwater remediation design of pump and treat systems via a simulation-optimization approach and firefly algorithm. *Engineering Optimization* **47** (1), 1–17.

Khalil, A., Almasri, M. N., Mckee, M. & Kaluarachchi, J. J. 2005 Applicability of statistical learning algorithms in groundwater quality modeling. *Water Resources Research* **41**, W05010.

Khalil, B., Broda, S., Adamowski, J., Ozga-Zielinski, B. & Donohoe, A. 2014 Short-term forecasting of groundwater levels under conditions of mine-tailings recharge using wavelet ensemble neural network models. *Hydrogeology Journal* **23** (1), 121–141.

Khaki, M., Yusoff, I. & Islami, N. 2015 Simulation of groundwater level through artificial intelligence system. *Environmental Earth Sciences* **73**, 8357–8367.

Knoben, W. J. M., Freer, J. E. & Woods, R. A. 2019 Technical note: inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta. *Hydrology and Earth System Sciences* **23** (10), 4323–4331.

Kumar, J. & Jain, A. 2006 Neural network based solutions for locating groundwater pollution sources. *Hydrology Journal* **29** (1–2), 55–66.

Lal, A. & Datta, B. 2019 Multi-objective groundwater management strategy under uncertainties for sustainable control of saltwater intrusion: solution for an island country in the South Pacific. *Journal of Environmental Management* **234**, 115–130.

Laumann, S., Micić, V., Lowry, G. V. & Hofmann, T. 2013 Carbonate minerals in porous media decrease mobility of polyacrylic acid modified zero-valent iron nanoparticles used for groundwater remediation. *Environmental Pollution* **179**, 53–60.

McArthur, J. M., Sikdar, P. K., Leng, M. J., Ghosal, U. & Sen, I. 2018 Groundwater quality beneath an Asian megacity on a delta: Kolkata's (Calcutta's) disappearing arsenic and present manganese. *Environmental Science and Technology* **52** (9), 5161–5172.

McCulloch, W. S. & Pitts, W. 1943 A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics* **5** (4), 115–133.

Milašinović, M., Ranđelović, A., Jaćimović, N. & Prodanović, D. 2019 Coupled groundwater hydrodynamic and pollution transport modelling using Cellular Automata approach. *Journal of Hydrology* **576**, 652–666.

Minsker, S. & Shoemaker, A. 1996 Differentiating a finite element biodegradation simulation model for optimal control. *Water Resources Research* **32** (1), 187–192.

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

Mosmeri, H., Alaie, E., Shavandi, M., Dastgheib, S. M. M. & Tasharrofi, S. 2017 Benzene-contaminated groundwater remediation using calcium peroxide nanoparticles: synthesis and process optimization. *Environmental Monitoring and Assessment* **189** (9), 1–14.

Mousavi, S. F. & Amiri, M. J. 2012 Modelling nitrate concentration of groundwater using adaptive neural-based fuzzy inference system. *Soil and Water Research* **2**, 73–83.

Nayak, P. C., Rao, Y. R. S. & Sudheer, K. P. 2006 Groundwater level forecasting in a shallow aquifer using artificial neural network approach. *Water Resources Management* **20**, 77–90.

Nie, S., Bian, J., Wan, H., Sun, X. & Zhang, B. 2017 Simulation and uncertainty analysis for groundwater levels using radial basis function neural network and support vector machine models. *Journal of Water Supply: Research and Technology – Aqua* **66** (1), 15–24.

Pal, J. & Chakrabarty, D. 2020 Assessment of artificial neural network models based on the simulation of groundwater contaminant transport. *Hydrogeology Journal* **28** (6), 2039–2055.

Pal, T., Mukherjee, P. K. & Sengupta, S. 2002 Nature of arsenic pollutants in groundwater of Bengal basin – a case study from Baruipur area, West Bengal, India. *Current Science* **82** (5), 554–561.

Podgorski, J. E., Labhasetwar, P., Saha, D. & Berg, M. 2018 Prediction modeling and mapping of groundwater fluoride contamination throughout India. *Environmental Science and Technology* **52**, 9889–9898.

Prasad, R. K. & Mathur, S. 2007 Groundwater flow and contaminant transport simulation with imprecise parameters. *Journal of Irrigation and Drainage Engineering* **133** (1), 61–70.

Rogers, L. L. 1992 *Optimal Groundwater Remediation Using Artificial Neural Networks and the Genetic Algorithm. PhD Thesis*, Lawrence Livermore National Laboratory, University of California, Livermore, California.

Ross, S. M. 2014 *Introduction to Probability and Statistics for Engineers and Scientists, University of Southern California*, 5th edn. Elsevier, New York, pp. 383–384.

Shiklomanov, I. A. 1993 *Water in Crisis : A Guide to the World's Fresh Water Resources*. Oxford University Press, New York.

Singh, A. & Minsker, B. S. 2008 Uncertainty-based multiobjective optimization of groundwater remediation design. *Water Resources Research* **44**, 1–20.

Singh, T. S. & Chakrabarty, D. 2011 Chance-constrained multi-objective programming for optimal multi-layer aquifer remediation design. *Engineering Optimization* **43** (4), 417–432.

Sinha, K. & Saha, P. 2015 Assessment of water quality index using cluster analysis and artificial neural network modeling: a case study of the Hooghly River basin, West Bengal, India. *Desalination and Water Treatment* **54** (1), 28–36.

Tabari, H., Nikbakht, J. & Shifteh Some'e, B. 2012 Investigation of groundwater level fluctuations in the north of Iran. *Environmental Earth Sciences* **66** (1), 231–243.

Varni, M., Comas, R., Weinzettel, P. & Dietrich, S. 2013 Journal of irrigation and drainage engineering. *Hydrological Sciences Journal* **58** (7), 1445–1455.

Voss, C. I. & Provost, A. M. 2010 SUTRA: a model for saturated-unsaturated, variable-density ground-water flow with solute or energy transport (Version 2.2). *United States Geological Survey* Water Resource Investigation Report 02-4231.

Wagh, V. M., Panaskar, D. B. & Muley, A. A. 2017 Estimation of nitrate concentration in groundwater of Kandava river basin-Nashik district, Maharastra, India by using artificial neural network model. *Modeling of Earth Systems and Environment* **3**, 36.

Wagh, V. M., Panaskar, D. B. & Muley, A. A. 2018 Neural network modelling for nitrate concentration in groundwater of Kadava River basin, Nashik, Maharashtra, India. *Groundwater for Sustainable Development* **7**, 436–445.

Walther, M., Delfs, J. O., Grundmann, J., Kolditz, O. & Liedl, R. 2012 Saltwater intrusion modeling: verification and application to an agricultural coastal arid region in Oman. *Journal of Computational and Applied Mathematics* **236** (18), 4798–4809.

Waristo, B., Santoso, R., Suparti & Yasin, H. 2018 Cascade forward neural network for time series prediction. *IOP Conference Series: Journal of Physics* **1025**, 012097.

Yan, S. F., Yu, S. E., Wu, Y. B., Pan, D. F., She, D. L. & Ji, J. 2015 Seasonal variations in groundwater level and salinity in coastal plain of eastern China influenced by climate. *Journal of Chemistry* **2015**, 1–8.

Yoon, H., Jun, S. C., Hyun, Y., Bae, G. O. & Lee, K. K. 2011 A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *Journal of Hydrology* **396** (1–2), 128–138.

First received 15 January 2021; accepted in revised form 18 April 2021. Available online 30 April 2021