Chapter 14
Use of Genetic Programming Based Surrogate Models to Simulate Complex Geochemical Transport Processes in Contaminated Mine Sites

Hamed Koohpayehzadeh Esfahani and Bithin Datta

14.1 Introduction

Reactive transport of chemical species, in contaminated groundwater system, especially with multiple species, is a complex and highly non-linear process. Simulation of such complex geochemical processes using efficient numerical models is generally computationally intensive. In order to increase the model reliability for real field data, uncertainties in hydrogeological parameters and boundary conditions are needed to be considered as well. Also, often the development of an optimal contaminated aquifer management and remediation strategy requires repeated solutions of complex and nonlinear numerical flow and contamination process simulation models. To address these combination of issues, trained ensemble Genetic Programming (GP) surrogate models can be utilized as approximate simulators of these complex physical processes in the contaminated aquifer. For example, use of trained GP surrogate models can reduce the computational burden in solving linked simulation based groundwater aquifer management models

Electronic supplementary material The online version of this chapter (doi: 10.1007/978-3-319-20883-1_14) contains supplementary material, which is available to authorized users.

H.K. Esfahani (✉)
Discipline of Civil Engineering, College of Science Technology and Engineering, James Cook University, Townsville, QLD 4811, Australia
e-mail: hamed.koohpayehzadehesfahani@my.jcu.edu.au

B. Datta (✉)
Discipline of Civil Engineering, College of Science Technology and Engineering, James Cook University, Townsville, QLD 4811, Australia

CRC-CARE, Mawson Lakes, SA 5095, Australia
e-mail: bthin.datta@jcu.edu.au

A.H. Gandomi et al. (eds.), *Handbook of Genetic Programming Applications*,
DOI 10.1007/978-3-319-20883-1_14

(Sreekanth and Datta 2011a, b) by orders of magnitude. Ensemble GP models trained as surrogate models can also incorporate various uncertainties in modelling the flow and transport processes. The development and performance evaluation of ensemble GP models to serve as computationally efficient approximate simulators of complex groundwater contaminant transport process with reactive chemical species under aquifer parameters uncertainties are presented. Performance evaluation of the ensemble GP models as surrogate models for the reactive species transport in groundwater demonstrates the feasibility of its use and the associated computational advantages. In order to evolve any strategy for management and control of contamination in a groundwater aquifer system, a simulation model needs to be utilized to accurately describe the aquifer properties in terms of hydrogeochemical parameters and boundary conditions. However, the simulation of the transport processes becomes complex and extremely non-linear when the pollutants are chemically reactive. In many contaminated groundwater aquifer management scenarios, an efficient strategy is necessary for effective and reliable remediation and control of the contaminated aquifer. Also, in a hydrogeologically complex aquifer site e.g., mining site, acid mine drainage (AMD) and the reactive chemical species together with very complex geology complicates the characterization of contamination source location and pathways.

In such contamination scenarios, it becomes necessary to develop optimal source characterization models, and strategies for future remediation. Solution of optimization models either for source characterization, or optimal management strategy development requires the incorporation of the complex physical processes in the aquifer. Also, most of the developed optimization models for source characterization or remediation strategy development require repeated solution of the numerical simulation models within the optimization algorithm. This process is enormously time consuming and often restricts the computational feasibility of such optimization approaches.

In order to overcome these computational restrictions, and to ensure computational feasibility of characterizing sources and pathways of contamination it is computationally advantageous to develop surrogate models which can be trained using solutions obtained from rigorous numerical simulation models. A number of attempts have been reported by researchers to develop surrogate models for approximately simulating the physical processes. Especially the use of trained Artificial Neural Network (ANN) models has been reported by a number of researchers (Ranjithan et al. 1993). However, the architecture of an ANN model needs to be determined by extensive trial and error solutions, and may not be suitable to deal with the simulation of very complex geochemical processes in contaminated aquifer site such as mine sites. Genetic Programming (GP) based surrogate models may overcome some of the limitations of earlier reported surrogate models. Therefore, this study develops GP model to approximately simulate three-dimensional, reactive, multiple chemical species transport in contaminated aquifers.

Trained and tested GP models based surrogate models are developed using the simulated response of a complex contaminated aquifer to randomly generated source fluxes. An ensemble GP model is an extension of the GP modelling technique capable of incorporating various uncertainties in a contaminated aquifer system data.

These ensemble GP models are trained and tested utilizing transient, three dimensional groundwater flow and transport simulation models for an illustrative study area hydrogeologically representing an abandoned mine site in Australia. Performance of the developed surrogate models is also evaluated by comparing GP model solutions with solution results obtained by using a rigorous numerical simulation of the aquifer processes. The three dimensional finite element based transient flow and contaminant transport process simulator, HYDROGEOCHEM 5.0 (Sun 2004) is used for this purpose. Reactive transport processes incorporating acid mine drainage in a typical mine site is simulated. Comparison of the solutions obtained with the surrogate models and the numerical simulation model solution results show that the ensemble GP surrogate models can provide acceptable approximations of the complex transport process in contaminated groundwater aquifers, with a complex geochemical scenario.

The performance of the developed surrogate models is evaluated for an illustrative study area to establish the suitability of GP models as surrogate models for such complex geological processes. These surrogate models if suitable will ensure the computational feasibility of developing optimization based models for source characterization, and help in the development of optimum strategies for remediation of large contaminated aquifer study areas. This study will demonstrate the utility and feasibility of using trained and tested ensemble GP models as a tool for approximate simulation of the complex geochemical processes in contaminated mine sites.

Aquifer contamination by reactive chemical species is widespread especially in mining sites. Numerical simulation models incorporating both chemical and physical behaviours are essential to describe reactive chemical transport process accurately. The numerical simulation model using the chemical reactive transport processes in aquifer contamination was addressed by (Parkhurst et al. 1982) and also implemented by (Herzer and Kinzelbach 1989; Tebes-Stevensa et al. 1998; Prommer et al. 2002). Coupled physical–chemical transport processes was developed using non-reactive transport model like MT3DMS (Zheng and Wang 1999) incorporating with various reactive transport numerical models (Prommer 2002; Parkhurst and Appelo 1999; Parkhurst et al. 2004; Waddill and Widdowson 1998; Mao et al. 2006) to simulate more realistic chemical reactive transport processes.

HYDROGEOCHEM (Yeh and Tripathi 1991) as a comprehensive numerical simulation model of flow and geochemically reactive transport in saturated–unsaturated media incorporates wide range of aquatic chemical equations as well as complex physical processes effectively. Heat, reactive geochemical and biochemical transport processes along with flow equations for the subsurface (saturated and unsaturated zones) are solved by three-dimensional model, HYDROGEOCHEM 5.0 (Sun 2004). In the proposed study, HYDROGEOCHEM 5.0 (HGCH) is used to simulate groundwater flow and transport processes with chemically reactive pollutants for an illustrative subsurface study area utilizing actual hydrogeologic data and synthetic hydro-geochemical data. Trained and tested ensemble GP based surrogate models are then utilized to approximately model complex geological and geochemical processes to improve the computational efficiency as well as reasonably accurate solutions.

One of the most hazardous contaminants for water resources is acid mine drainage (AMD) and its related compounds spatially distributed which are the products of mining activities (Kalin et al. 2006). Generally AMD or acid rock drainage (ARD) is produced by various sulphide rocks' surface chemical weathering in presence of water, oxygen and microorganisms. Mining activities accelerate AMD production by increasing the rocks' surface as well as distributing wastewater and waste deposit of sulphide minerals such as pyrite (FeS2), pyrrhotite (Fe1-xS), chalcopyrite (CuFeS2), arsenopyrite (FeAsS), etc. in mine sites (Nordstrom and Alpers 1999). These contaminants pollute water resources widely as well as decrease the water pH which leads to increase in the concentration of other hazardous metals and heavy metals in water (Kalin et al. 2006). In this study, the transport process of sulphate, iron and copper, hazardous AMD's compounds, along with their chemical reactions through the contaminated aquifer is considered.

Recently surrogate models have been proposed as approximate replacement for numerical simulation model for developing linked simulation optimization models (Bhattacharjya and Datta 2005) for groundwater quality management. Replacing aquifer responses simulation by linear surrogate models developed using response matrix approach was initially reported (Zhou et al. 2003; Abarca 2006). Recently, Artificial Neural Network (ANN) (Ranjithan et al. 1993) and Genetic Programming (GP) based surrogate models have been proposed as efficient non-linear surrogate models (Koza 1994).

Artificial Neural Networks (ANN) has been widely used as approximate surrogate models for groundwater simulation (Aly and Peralta 1999). Rogers et al. (1995) presented one of the earliest attempts using ANN as a surrogate for a coastal groundwater flow model. They demonstrated the substantial saving in terms of computation time by using ANN and Genetic Algorithmic (GA) based meta-model (surrogate model) within a linked simulation-optimization model for evolving optimal groundwater management strategies. Replacing groundwater simulation models with ANN-base surrogate models were developed by Bhattacharjya and Datta (2005, 2009) and Bhattacharjya et al. (2007) and Dhar and Datta (2009). McPhee and Yeh (2006) used ordinary differential equation surrogates to approximating simulate of groundwater flow and transport processes. Optimizing the surrogate model parameters related on fixed initial surrogate model structure is the main concept of most of these surrogate modelling approaches to obtain the best between the explanatory and response variables. Even the most popularly used trained ANN-based surrogate modelling approach obtains the optimal model formulation by trial and error (Bhattacharjya and Datta 2005).

Bhattacharjya et al. (2007) used ANN as an approximate simulation for substitutes the three dimensional flow and transport simulation model to simulate the complex flow and transport process in a coastal aquifer. Bhattacharjya and Datta (2009) used the trained ANN-based surrogate models for approximating density depended saltwater intrusion process in coastal aquifer to predict the complex flow and transport processes. Dhar and Datta (2009) used ANN as a surrogate model for simulation of flow and transport in the multiple objective non-dominated front search process resulting in saving a huge amount of computational time.

Genetic Programming (GP), proposed by Koza (1994) is an evolutionary algorithm which is capable approximate simulation of complex models effectively using stochastic search methods. Compared to other regression techniques, the most important advantage of GP is its ability to optimize both the variables and constants of the candidate models without initial model structure definition. This approach makes GP a strong surrogate model to characterize the model structure uncertainty. Recently genetic programming has been utilized in hydrological applications in several researches (Dorado et al. 2002; Makkeasorn et al. 2008; Wang et al. 2009). Trained GP-based surrogate models has been used to substitutes the simulation models for runoff prediction, river stage and real-time wave forecasting (Whigham and Crapper 2001; Savic et al. 1999; Khu et al. 2001; Babovic and Keijzer 2002; Sheta and Mahmoud 2001; Gaur and Deo 2008). In addition, GP has been applied to approximate modelling of different geophysical processes including flow over a flexible bed (Babovic and Abbott 1997); urban fractured-rock aquifer dynamics (Hong and Rosen 2002); temperature downscaling (Coulibaly 2004); rainfall-recharge process (Hong et al. 2005); soil moisture (Makkeasorn et al. 2006); evapotranspiration (Parasuraman et al. 2007b); saturated hydraulic conductivity (Parasuraman et al. 2007a); and for modelling chemical entropy (Bagheri et al. 2012, 2013, 2014). Zechman et al. (2005) developed a trained GP-based surrogate models as an approximate simulation of groundwater flow and transport processes in a groundwater pollutant source identification problem.

Sreekanth and Datta (2010) implemented GP as meta-model to replace the flow and transport simulation of density dependent saltwater intrusion in coastal aquifers for ultimate development of optimal saltwater intrusion management strategies. Sreekanth and Datta (2011b, 2012) compared two non-linear surrogate models based on GP and ANN models, respectively and showed that the GP based models perform better in some aspects. These advantages include: simpler surrogate models, optimizing the model structure more efficiently, and parsimony of parameters. Datta et al. (2013) described the utilization of trained GP surrogate models for groundwater contamination management, and development of a monitoring network design methodology to develop optimal source characterization models. Replacing simulation groundwater model by GP-based ensemble surrogate models in linked simulation-optimization developed methodology was addressed by Datta et al. (2014) and Sreekanth and Datta (2011a) which improve the computational efficiency and obtains reasonably accurate results under aquifer hydrogeologic uncertainties.

In this study our main objectives is to develop ensemble genetic programming based surrogate models to approximately simulate the complex transport process in a complex hydrogeologic system with reactive chemical species, and to illustrate its efficiency and reliability in a contaminated aquifer resembling an abandoned mine site. The numerical model's formulations as well as using ensemble genetic programming based surrogate models are described in Sect. 14.2 and the results are presented and discussed in Sect. 14.3.

14.2 Methodology

The methodology developed includes two main components. In the first step, the simulation model for the flow and transport processes is described, and complex chemical reactive transport process is simulated by the HGCH, a three-dimensional coupled physical and chemical transport simulator, to realize the reactive contaminants behaviours is contaminated aquifers. The hydrogeochemical data and boundary conditions at the illustrative study site are similar to an abandoned mine site in Queensland, Australia. Trained ensemble GP based surrogate models are then developed to approximately obtain concentrations of the chemical contaminants at different times in specified locations while incorporating uncertainties in hydrogeological aquifer parameters like hydraulic conductivity. Comparison of the spatio-temporal concentrations obtained as solution by solving the implemented numerical three dimensional reactive contaminant transport simulation model (HGCH) and those obtained using ensemble GP models are then presented to show the potential applicability and the efficiency of using GP ensemble surrogate models under aquifer uncertainties.

14.2.1 Simulation Model of Groundwater Flow and Geochemical Transport

HYDROGEOCHEM 5.0 (HGCH), consisting of the numerical flow simulator and physio-chemical transport simulator HGCH is a computer program that numerically solves the three-dimensional groundwater flow and transport equations for a porous medium. The finite-element method is used in this simulation model.

The general equations for flow through saturated–unsaturated media are obtained based on following components: (1) fluid continuity, (2) solid continuity, (3) Fluid movement (Darcy's law), (4) stabilization of media, and (5) water compressibility (Yeh et al. 1994). Following governing equation is used:

$$\frac{\rho}{\rho_0} F \frac{\partial h}{\partial t} = \nabla . \left[K . \left(\nabla h + \frac{\rho}{\rho_0} \nabla z \right) \right] + \frac{\rho^*}{\rho_0} q \tag{14.1}$$

F is the generalized storage coefficient (1/L) defined as:

$$F = \alpha' \frac{\theta}{n_e} + \beta' \theta + n_e \frac{ds}{dh} \tag{14.2}$$

K is the hydraulic conductivity tensor (L/T) is:

$$K = \frac{\rho g}{\mu} k = \frac{(\rho/\rho_0)}{(\mu/\mu_0)} \frac{\rho_0 g}{\mu_0} k_s \, k_r = \frac{(\rho/\rho_0)}{(\mu/\mu_0)} K_{so} \, k_r \tag{14.3}$$

V is the Darcy's velocity (L/T) described as:

$$V = -K\left[\frac{\rho}{\rho_0}\nabla h + \nabla z\right] \tag{14.4}$$

Where:

θ: effective moisture content (L3/L3);

h: pressure head (L);

t: time (T);

z: potential head (L);

q: source or sink of fluid [(L3/L3)/T];

ρ_0: fluid density without biochemical concentration (M/L3);

ρ: fluid density with dissolved biochemical concentration (M/L3);

ρ*: fluid density of either injection (=ρ*) or withdraw (=ρ) (M/L3);

μ_0: fluid dynamic viscosity at zero biogeochemical concentration [(M/L)/T];

μ: the fluid dynamic viscosity with dissolved biogeochemical concentrations [(M/L)/T];

α′: modified compressibility of the soil matrix (1/L);

ß: modified compressibility of the liquid (1/L);

ne: effective porosity (L3/L3);

S: degree of effective saturation of water;

G: is the gravity (L/T2);

k: permeability tensor (L2);

k_s: saturated permeability tensor (L2);

K_{so}: referenced saturated hydraulic conductivity tensor (L/T);

k_r: relative permeability or relative hydraulic conductivity (dimensionless)

When combined with appropriate boundary and initial conditions, the above equations are used to simulate the temporal-spatial distributions of the hydrological variables, including pressure head, total head, effective moisture content, and Darcy's velocity in a specified study area.

The contaminant transport equations used in the HG model can be derived based on mass balance and biogeochemical reactions (Yeh 2000). The general transport equation using advection, dispersion/diffusion, source/sink, and biogeochemical reaction as the major transport processes can be written as follows:

$$\frac{D}{Dt}\int_{v}\theta C_i dv = -\int_{\Gamma} n.\,(\theta C_i)\,V_i d\Gamma - \int_{\Gamma} n.J_i d\Gamma + \int_{v}\theta r_i dv + \int_{v} M_i dv,\ i \in M \tag{14.5}$$

Where

C_i: the concentration of the ith species in mole per unit fluid volume (M/L^3);

ν: the material volume containing constant amount of media (L3);

Γ: the surface enclosing the material volume ν (L2);

n: the outward unit vector normal to the surface Γ;

Ji: the surface flux of the ith species due to dispersion and diffusion with respect to relative fluid velocity [(M/T)/L2];

θri: the production rate of the ith species per unit medium volume due to all biogeochemical reactions [(M/L3)/T];

Mi: the external source/sink rate of the ith species per unit medium volume [(M/L3)/T];

M: the number of biogeochemical species;

Vi: the transporting velocity relative to the solid of the ith biogeochemical species (L/T).

14.2.2 Genetic Programming Based Ensembles Surrogate Model

GP models are used in this study to evolve surrogate models for approximately simulating flow and transport processes in a contaminated mine site. Trained GP models are developed using the simulated response of the aquifer to randomly generated source fluxes. GP, a branch of genetic algorithms (Koza 1994), is an evolutionary algorithm-based methodology inspired by biological evolution to find computer programs that perform a user-defined task (Sreekanth and Datta 2011b). Essentially, GP is a set of instructions and a fitness function to measure how well a computer model has performed a task. The main difference between GP and genetic algorithms is the representation of the solution. GP creates computer programs in the lisp or scheme computer languages as the solution. Genetic algorithms create a string of numbers that represent the solution.

The main operators applied in genetic programming as in evolutionary algorithms are crossover and mutation. Crossover is applied on an individual by simply replacing one of the nodes with another node from another individual in the population. With a tree-based representation, replacing a node means replacing the whole branch (Fig. 14.1). This adds greater effectiveness to the crossover operator. The expressions resulting from crossover are very different from their initial parents. Mutation affects an individual in the population. It can replace a whole node in the selected individual, or it can replace just the node's information. To maintain integrity, operations must be fail-safe or the type of information the node holds must be taken into account. For example, mutation must be aware of binary operation nodes, or the operator must be able to handle missing values.

GP utilizes a set of input–output data which are generated randomly by using the flow and contaminant transport simulation models. The numerical Simulation model creates M number of out-put sets from M number of input sets, which is generated by using random Latin Hypercube sampling in defined ranges. The performance of each GP program is an evaluated formulation in terms of training, testing the validation using the set of input–output patterns. The testing data evaluates the

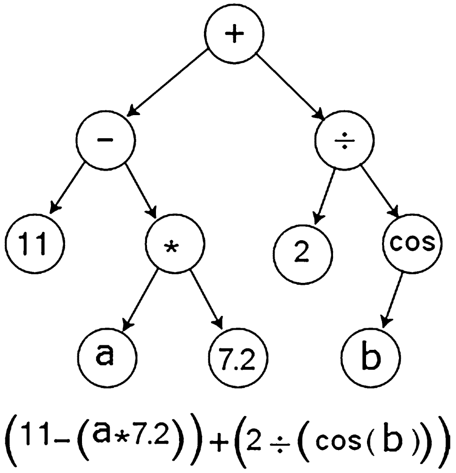

Fig. 14.1 Function represented as a tree structure

model performance for new data using the fitness function obtained in the training phase. Non-tree representations have been proposed and successfully implemented, such as linear genetic programming which suits the more traditional imperative languages (Banzhaf et al. 1998). The commercial GP software Discipulus (Francone 1998) performs better by using automatic induction of binary machine code. In the proposed methodology, Discipulus GP software is used to solve and generate GP models. Discipulus uses Linear Genetic Programming (LGP) which utilizes input variables in line-by-line approach. This objective of this program is minimizing difference in value between the output estimated by GP program on each pattern and the actual outcome. The fitness objective functions are often absolute error or minimum squared error. Almost two-thirds of the input–output data sets obtained from the numerical simulation model are utilized for training and testing the GP model. The remaining data sets are used to validate the GP models. The r-square value shows the fitness efficiency to the GP models (Sreekanth and Datta 2010).

14.2.2.1 Performance Evaluation

The trained ensemble GP surrogate models are evaluated to verify the performance of the surrogate models approximating flow and transport processes simulation with reactive chemical species, under hydrogeological uncertainties. Input data sets are generated randomly by Latin Hypercube sampling in defined ranges. The aquifer hydrogeological uncertainties include uncertainties in estimating hydraulic conductivity, water content and constant groundwater label in boundary conditions.

14.2.3 Performance Evaluation of Developed Methodology

In order to evaluate the performance of the proposed methodology, ensemble GP based surrogate models are utilized for an illustrative study area shown in Fig. 14.2. The specified hydrogeologic conditions resemble a homogeneous and isotropic aquifer. In order to evaluate the methodology, the ensemble GP surrogate models are first trained using the sets of solution results obtained using the 3-D finite element based flow and reactive transport simulation model. Once trained and tested, the GP models are utilized for simulating the transport process in the study site. Then the surrogate model solution results are compared with the actual numerical simulation solution results.

The areal extent of the specified study area is 10,000 m^2 with complex pollutant sources including a point source and a distributed source. The spatial concentrations are assumed to measure at different times at ten arbitrary observation well locations. The thickness of the aquifer is specified as 50 m with anisotropic hydraulic conductivity in the three directions. The boundaries of the study area are no-flow for top and bottom sides while left and right sides of the aquifer have constant head boundaries with specified hydraulic head values. The total head decreases from top to bottom and left to right gradually. The aquifer system is shown in Fig. 14.2.

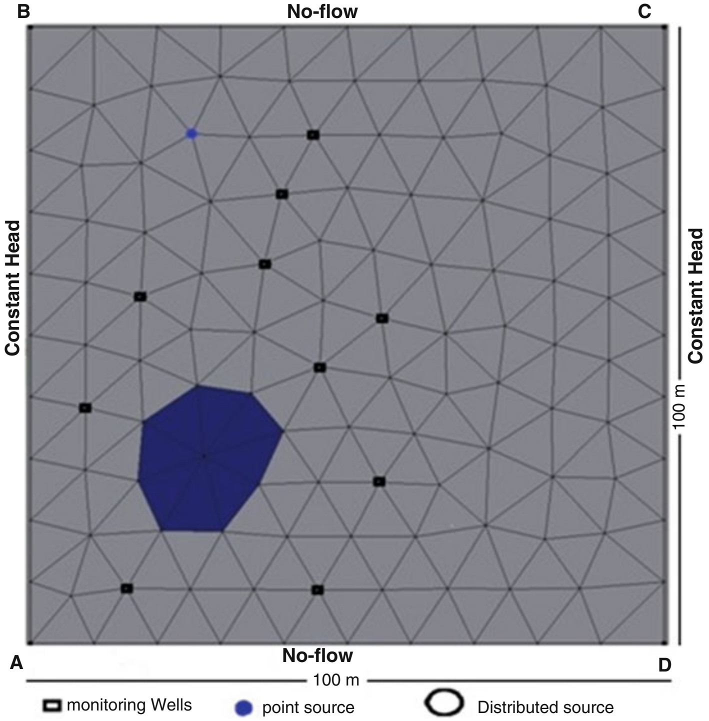

Fig. 14.2 Illustrative study area (total head: A = 37 m, B = 40 m, C = 33 m, D = 30 m)

As shown in Fig. 14.2, the dark blue area represent the contaminant sources S(i) which include distributed and point sources. Concentration data from monitoring well locations, shown as black rectangular points, are used to train, test and validate the GP model formulations.

Table 14.1 shows dimensions, hydrogeological properties, and boundary conditions of the study area which are utilized for numerical models to simulate groundwater flow and chemical reactive transport processes. The synthetic concentration measurement data used for the specified polluted aquifer facilitates evaluation of the developed methodology. These synthetic concentration measurement data at specified observation locations are obtained by solving the numerical simulation model with known pollution sources, boundary conditions, initial conditions, and hydrogeologic as well as geochemical parameter values. In the incorporated scenario, copper (Cu^{2+}), Iron (Fe^{2+}) and sulphate ($SO4^{2-}$) are specified as the chemical species in the pollutant sources. The associated chemical reactions are listed in Table 14.2.

Nine different scenarios are defined based on different hydraulic conductivity and boundary conditions with maximum 10 % differences between maximum

Table 14.1 Aquifer's properties

Aquifer parameter	Unit	Value
Dimensions (length * width * thickness) study area	m * m * m	100 * 100 * 50
Number of nodes		387
Number of elements		1432
Hydraulic conductivity, Kx, Ky, Kz	m/d	10.0, 5.0, 3.0
Effective porosity, $\ominus$		0.3
Longitudinal dispersivity, αL	m/d	10.0
Transverse dispersivity, αT	m/d	6.0
Horizontal anisotropy		1
Initial contaminant concentration	Mole/lit	0–5
Diffusion coefficient		0

Table 14.2 Typical chemical reactions during the contaminant transport process

Chemical reaction equations	Constant rate (Log k)[a]
Equilibrium reactions	
(1) $Cu^{2+} + H_2O \leftrightarrow Cu(OH)^+ + H^+$	−9.19
(4) $Cu^{2+} + SO_4{}^{2-} \leftrightarrow CuSO_4$	2.36
(7) $Fe^{2+} + SO_4{}^{2-} \leftrightarrow FeSO_4$	2.39
(9) $4Fe^{2+} + 4H^+ \leftrightarrow 4Fe^{3+} + 2H_2O$	8.5
(14) $Fe^{3+} + SO_4{}^{2-} \leftrightarrow FeSO_4{}^+$	4.05
(15) $Fe^{3+} + SO_4{}^{2-} + H^+ \leftrightarrow FeHSO_4{}^{2+}$	2.77
Kinetic reactions	
(17) $FeOOH_{(s)} + 3H^+ \leftrightarrow Fe^{3+} + 3H_2O$	$K_f = 0.07$

[a] Constant rates are taken from Ball and Nordstrom (1992)

and minimum values, and with the mean value assumed as the actual value for simulating the synthetic concentration observation (N11, N12, …, N21, … N33). First digit indicates an index for hydraulic conductivity values and second one represents an index for the hydraulic head as boundary condition. 1 illustrates parameters with 5 % less than the actual definition as well as 2 and 3 shows the exact data and 5 % more than actual parameters in illustrative aquifer respectively.

14.2.3.1 Generation of Training and Testing Patterns for the Ensemble GP Models

The total time of source activities is specified as 800 days, subdivided into eight similar time intervals of 100 days each. The actual pollutant concentration from each of the sources is presumed to be constant over each stress period. The pollutant concentration of copper, iron as well as sulphate in the pit is represented as Cpit(i), Fepit(i) and Spit(i) respectively, where i indicates the stress period number, and also C(i), Fe(i) and S(i) represent copper, iron and sulphate concentrations in the point sources, respectively at different time steps.

An overall of sixteen concentration values for each contaminant are considered as explicit variables in the simulation model. The concentration measurements are simulated for a time horizon of 800 days since the start of the simulation. The pollutant concentration are assumed to be the resulting concentrations at the observation wells at every 100 days interval and this process is continued at all the observation locations till t = 800 days. Only for this methodology evaluation purpose, these concentration measurements are not obtained from field data, but are synthetically obtained by solving the numerical simulation model for specified initial conditions, boundary conditions and parameter values. In actual application these measurement data need to be simulated using a calibrated flow and contaminant simulation model. However, using field observations for calibration, and then for evaluation of a proposed methodology results in uncertain evaluation results as the quality of the available measurement data cannot be quantified most of the time. Therefore as often practiced, synthetic aquifer data is used for this evaluation of the methodology proposed.

The comprehensive three-dimensional numerical simulation model was used to simulate the aquifer flow and chemical reactive transport processes due to complex pollutant sources in this study area. Different random contaminant source fluxes as well as different realization of boundary conditions and hydraulic conductivities were generated using Latin hypercube sampling. For random generation purpose, 10 % initial aquifer properties are considered as Maximum error for the uncertainties of aquifer parameters. HGCH was utilized to obtain the concentrations resulting from each of these concentration patterns. The simulated concentration measured data at monitoring network and the corresponding concentration of contaminants at sources form the input–output pattern. Totally, 8000 concentration patterns for all the ten concentration observation locations were used in this evaluation. Eight input–output patterns were defined based on different time steps.

Genetic programming models were obtained using each of these data sets to create ensemble GP based surrogate models. Each data set was split into halves for training and testing the genetic programming-based surrogate models.

Surrogate models were developed for simulating pollutant concentrations at the observation locations at different times resulting from the specified pollutant sources at different times under hydrogeological uncertainties. All the GP models used a population size of 1000, and mutation frequency of 95. The Discipulus, commercial Genetic Programming software, was used to develop the surrogate models. The model was developed using default parameters values of Discipulus. The GP fitness function was the squared deviation between GP model generated and actually simulated concentration values at the observation locations and times.

14.3 Evaluation Results and Discussion

The flow and concentration simulation results for the study area obtained using the numerical HGCH simulation model are shown in Figs. 14.3, 14.4, 14.5, and 14.6. The flow movement, total head contours in top layer and also velocity vectors are shown in Figs. 14.3, 14.4 and 14.5 respectively. Figures 14.3, 14.4 and 14.5 show the hydraulic heads for flow. The contours show a gentle slope from point B towards D. Figure 14.6 shows the copper concentration distribution in the study area which shows the complex transport processes with reactive chemical species.

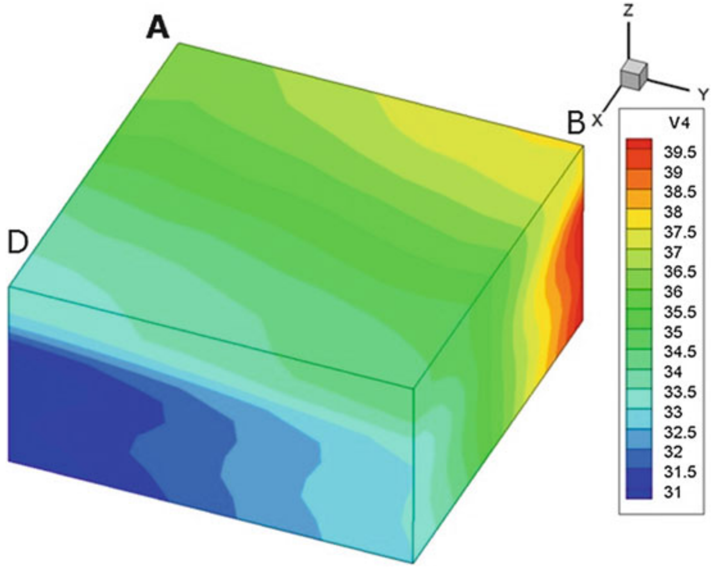

Fig. 14.3 3-D view of hydraulic head distribution

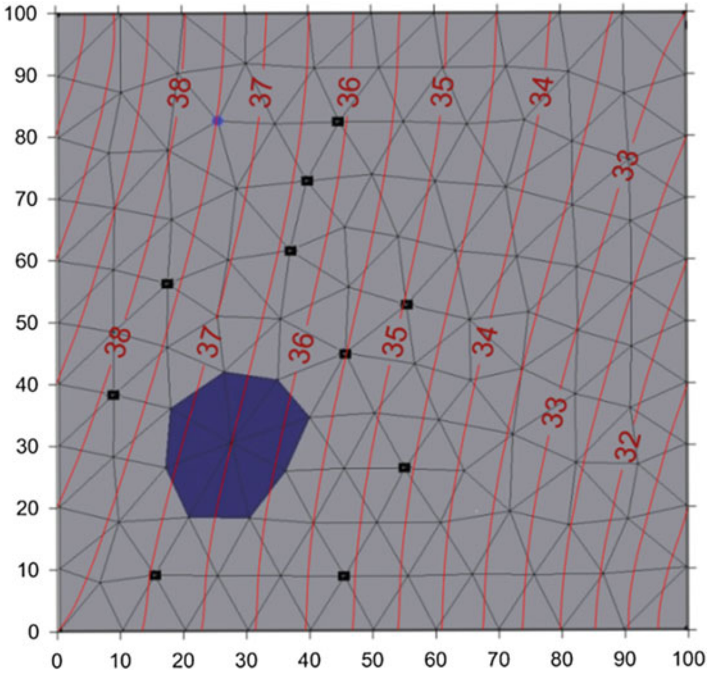

Fig. 14.4 Hydraulic head contours (m)

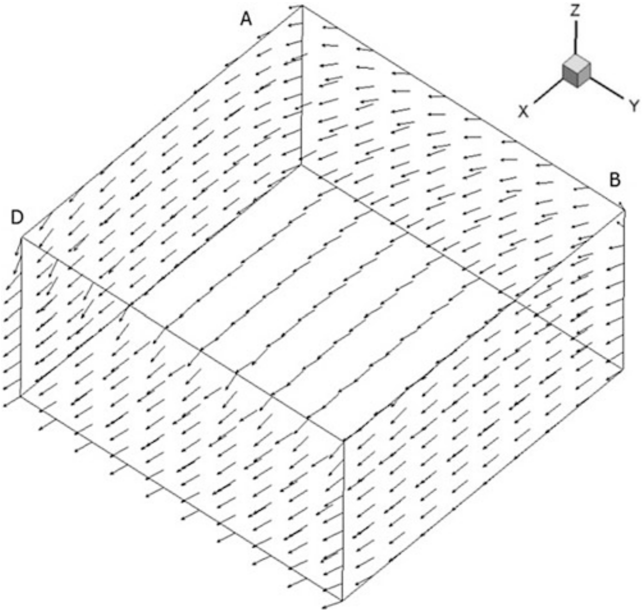

Fig. 14.5 Velocity vectors of groundwater movement

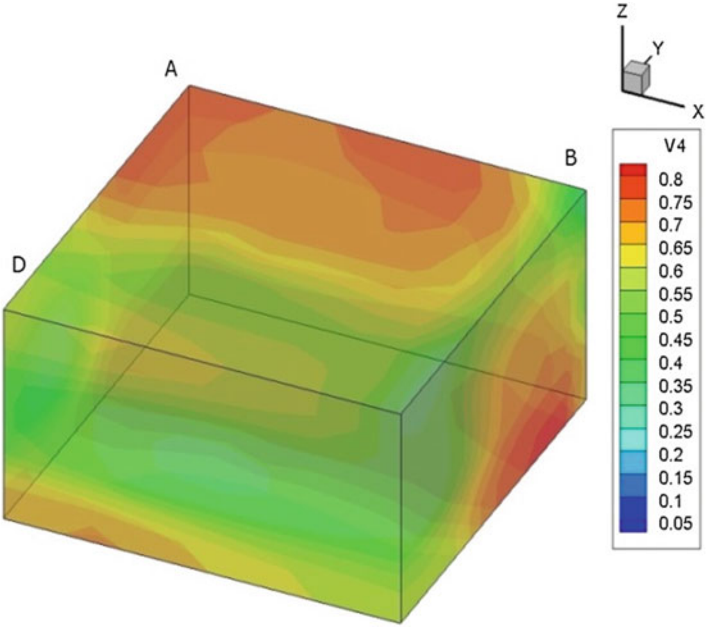

Fig. 14.6 Copper concentration (mole/lit) distribution in the study area

The concentration of sulphate remains almost the same while iron concentration is lower in groundwater. Based on pH changes the iron can react and cease to be in solute phase, thus removed form groundwater.

The results obtained using the developed ensemble genetic programming based surrogate models for approximate simulation of pollutant concentrations are compared with the numerical simulation results obtained using the HGCH. Nine different scenarios are considered. These nine scenarios are characterized by different hydraulic conductivity value realizations and hydraulic head boundary conditions. Each randomized within 10 (±5)% errors in the mean values (assumed same as the actual values) for hydrogeological parameters and boundary conditions. Incorporation of these scenarios together with the Latin Hypercube based randomization to achieve the efficiency of ensemble GP based surrogate models. The uncertainties in the parameter values of the scenarios are within the range of input data which are used to create the ensemble GP models. Figure 14.7a–c illustrate these comparison results in which one scenario for one particular hydraulic conductivity is selected for obtaining simulated output data from HGCH model at each monitoring networks. Each time step is marked on the x-axis. Each of the bars corresponds to contaminant concentration in each well, obtained by HGCH and ensemble GP models.

Figure 14.7 shows that the results obtained from the ensemble GP based surrogate models are very close to the simulated results obtained using the numerical

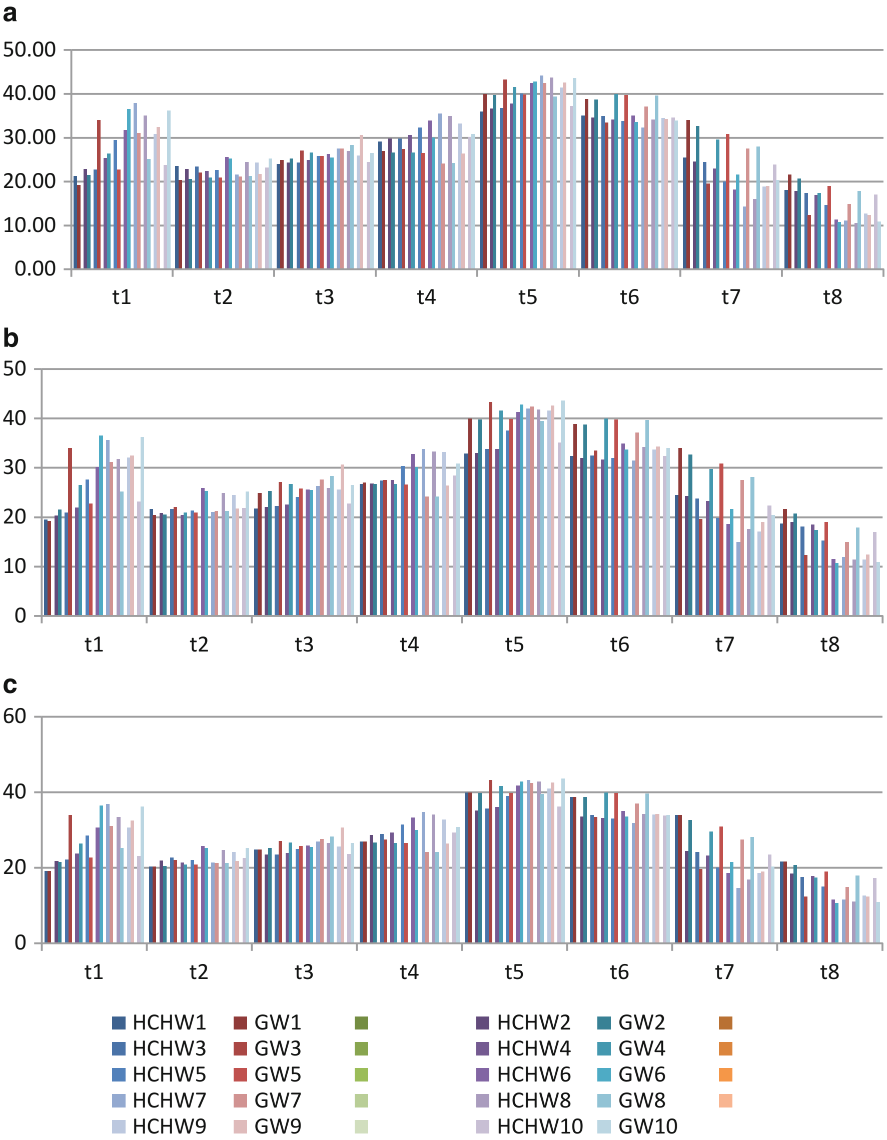

Fig. 14.7 Comparison of ensemble GP model solutions with HGCH simulation results for specified parameter values defined by (**a**) lower bound on uncertain aquifer parameter values, (**b**) actual or mean parameters values and (**c**) upper bound on aquifer parameter values (GW1: concentration data at well number 1 based on GP formulation, HCHW1: concentration data at well number 1 based on HGCH simulation)

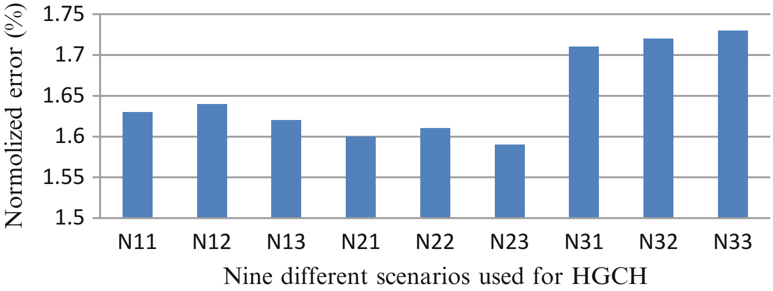

Fig. 14.8 Normalized error for all scenarios under uncertainties

simulation model, and also incorporates der uncertainties. Figure 14.8 shows the summation of normalized error at each of the observation locations for each monitoring network averaged over the 8 time periods. It is noted that, the ensemble GP models provide relatively accurate results for concentrations at observation locations. Although the boundary conditions are different, the normalized errors for all the three scenarios with same hydraulic conductivity are almost the same. The most important advantage of using the developed GP models is that the numerical simulation model requires long computational time usually several hours for a typical study area, while ensemble genetic programming surrogate models deliver the solution results in typically fraction of a second. Also the ensemble GP models directly incorporate hydrogeologic uncertainties in the modelled system. Therefore the computational advantage of using the ensemble GP for approximate simulation of complex reactive transport processes in aquifers is enormous if the errors in simulation are within acceptable range. Especially, this computational time saving could be critical in development and solution of linked simulation-optimization models (Datta et al. 2014) for management of contaminated aquifers.

14.4 Conclusion

Although surrogate models are widely used in solving groundwater management problems replacing the actual complex numerical models, often the main issue is the accuracy and reliability of surrogate model predictions under input data uncertainties. This study developed a methodology based on ensemble GP surrogate models to substitute numerical simulation for approximate simulation of the chemically reactive multiple species transport process in a contaminated aquifer resembling the geochemical characteristics of an abandoned mine site. The evaluation results show the applicability of this methodology to approximating the complex reactive transport process in an aquifer. The developed ensemble GP models result in increasing the computation efficiency and computational feasibility, while providing acceptable results.

The linked simulation-optimization approach is an effective method to identify source characterization and monitoring network design under uncertainties in complex real life scenarios which important for robust remediation strategies and groundwater management. The main difficulty with linked simulation-optimization models generally is the required huge computation time, due to iterative repeated solution of the numerical flow and transport simulation models. To address this, ensemble GP based surrogate models may be used to approximate the numerical simulation model under uncertainties, in the linked simulation-optimization model. Ensemble GP based surrogate models can increase efficiency and feasibility of developing optimal management strategies for groundwater management in geochemically complex contaminated aquifers such as mine sites, while at the same time incorporating uncertainties in defining the hydrogeologic system. The evaluations results show that it is feasible to use ensemble GP models as approximate simulators of complex hydrogeologic and geochemical processes in a contaminated groundwater aquifer incorporating uncertainties in describing the physical system.

References

Abarca, E., Vazquez-Sune, E., Carrera, J., Capino, B., Gamez, D., Battle, F., (2006) *Optimal design of measures to correct seawater intrusion.* Water Resource Research., 42.

Aly AH, and Peralta, R. C. (1999) Optimal design of aquifer cleanup systems under uncertainty using a neural network and a genetic algorithm. Water Resour Res 35 (8):10

Babovic V and Abbott M B (1997) Evolution of equation from hydraulic data. part I: Theory. J Hydraul Res 35:14

Babovic V, and Keijzer, M., (2002) Rainfall runoff modelling based on genetic programming. Nord Hydrol 33 (5):15

Bagheri M, Borhani T.N.G., Gandomi A.H., and Manan Z.A., (2014) A simple modelling approach for prediction of standard state real gas entropy of pure materials. SAR and QSAR in Environmental Research, 25 (9): 695–710

Bagheri M., Gandomi A.H., Bagheri M., and Shahbaznezhad M., (2013) Multi-expression programming based model for prediction of formation enthalpies of nitro-energetic materials. Expert Systems, 30 (1): 66

Bagheri M., Bagheri M., Gandomi A.H., Golbraikhc A., (2012) Simple yet accurate prediction method for sublimation enthalpies of organic contaminants using their molecular structure. Thermochimica Acta 543: 96–106.

Ball JW, Nordstrom, D.K., (1992) User's manual for WATEQ4F, with revised thermodynamic database and test cases for calculating speciation of major, trace and redox elements in natural waters. US Geol Surv Open-File Rep 91-183 (Revised and reprinted August 1992)

Banzhaf W, Nordin, P., Keller, R. E., and Francone, F. D. (1998) Genetic Programming: An Introduction. Morgan Kaufmann, Inc, San Francisco, USA

Bhattacharjya R, and Datta, B., (2005) Optimal management of coastal aquifers using linked simulation optimization approach. Water Resour Manage 19 (3):25

Bhattacharjya R, K., Datta, B., and Satish, M., (2007) Artificial neural networks approximation of density dependent saltwater intrusion process in coastal aquifers. Journal of Hydrologic Engineering 12 (3):10

Bhattacharjya RK, and Datta, B., (2009) ANN-GA-based model for multiple objective management of coastal aquifers. Water Resour Planning Manage 135 (5):8

Coulibaly P (2004) Downscaling daily extreme temperatures with genetic programming, Geophys. Res Lett 31 (L16203)

Datta B, Prakash, O., Campbell, S., Escalada, G., (2013) Efficient Identification of Unknown Groundwater Pollution Sources Using Linked Simulation-Optimization Incorporating Monitoring Location Impact Factor and Frequency Factor. Water Resour Manage 27:18

Datta B, Prakash, O., Sreekanth, J (2014) Application of genetic programming models incorporated in optimization models for contaminated groundwater systems management. EVOLVE - A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation V Advances in Intelligent Systems and Computing 288:16

Dhar A, Datta, B. (2009) Saltwater Intrusion Management of Coastal Aquifers. I: Linked Simulation-Optimization. Journal Hydrology Engineering 14:9

Dorado J, Rabunal, J. R., Puertas, J., Santos, A., and Rivero, D., (2002) Prediction and modelling of the flow of a typical urban basin through genetic programming. Appl Artifici Intel 17 (4):14

Whigham PA, Craper, P. (2001) Modelling rainfall-runoff using genetic programming. Math Comput Modell 33:14

Francone FD (1998) Discipulus Software Owner's Manual, version 3.0 draft. Machine Learning Technologies Inc, Littleton, CO, USA

G MJaYW (2006) Experimental design for groundwater modeling and management. Water Resources Research 42 (2):1-13

Gaur S, and Deo, M. C., (2008) Real-time wave forecasting using genetic programming. Ocean Eng 35 (11-12):7

Herzer j, Kinzelbach, W., (1989) Coupling of Transport and Chemical Processes in Numerical Transport Models. Geoderrna 44:13

Hong Y S aRMR (2002) Identification of an urban fractured rock aquifer dynamics using an evolutionary self-organizing modelling. J Hydrol 259:15

Hong, Y.S., White, P. A., Scott, D. M. (2005) *Automatic rainfall recharge model induction by evolutionary computational intelligence.* Water Resour. Res., 41.

Kalin M, Fyson, A., Wheeler, W. N., (2006) Review The chemistry of conventional and alternative treatment systems for the neutralization of acid mine drainage Science of the Total Environment 366:14

Khu S T, Liongs S Y, Babovic V, Madsen H, and Muttil N, (2001) Genetic programming and its application in real-time runoff forecasting. J Am Water Resour Assoc 8:20

Koza JR (1994) Genetic programming as a means for programming computers by natural-selection. Statistics and computing 4:26

Makkeasorn A, Chang, N. B., and Zhou, X. (2008) Short-term streamflow forecasting with global climate change implications—A comparative study between genetic programming and neural network models. J Hydrol 352 (3-4):19

Makkeasorn A CNB, Beaman M, Wyatt C and Slater C (2006) Soil moisture estimation in semiarid watershed using RADARSAT-1 satellite imagery and genetic programming. Water Resour Res 42 (W09401)

Mao X, Prommer, H., Barry, D.A., Langevin, C.D., Panteleit, B., Li, L. (2006) Three-dimensional model for multi-component reactive transport with variable density groundwater flow. Environmental Modelling & Software 21 ((5)):14

McPhee, J., Yeh, W. G. (2006) Experimental design for groundwater modeling and management. WATER RESOURCES RESEARCH. 42. 2. P:1–13

Nordstrom D.K., Alpers C.N., (1999) Geochemistry of acid mine waters, in Plumlee, G.S., and Logsdon, M.J., eds., The environmental geochemistry of mineral deposits, Part A: Processes, techniques, and health issues, Reviews Economic Geology, 6A:28

Parasuraman K, Elshorbagy, A, and Carey S. K., (2007b) Modelling the dynamics of evapotranspiration process using genetic programming. Hydrol Sci J 52:15

Parasuraman K., Elshorbagy A., and Si, B. C., (2007a) Estimating saturated hydraulic conductivity using genetic programming. Soil Sci Soc Am J 71:9

Parkhurst DL, Appelo, C.A.J. (1999) User's guide to PHREEQC—A computer program for speciation, reaction-path, 1D-transport, and inverse geochemical calculations. Technical Report 99-4259, US Geol Survey Water-Resources Investigations Report

Parkhurst DL, Kipp, K.L., Engesgaard, P., Charlton, S.R. (2004) PHAST e A Program for Simulating Ground-water Flow, Solute transport, and Multicomponent Geochemical Reactions. US Geological Survey, Denver, Colorado

Parkhurst DL, Thorstenson, D.C. and Plummet, L.N., (1982) PHREEQE - A computer program for geochemical calculations. Water Resour Invest 210:16

Prommer H (2002) PHT3D—A reactive multi-component transport model for saturated porous media. Version 10 User's Manual, Technical report, Contaminated Land Assessment and Remediation Research Centre, The University of Edinburgh

Prommer H, Barry, D.A., Davis, G.B. (2002) Modelling of physical and reactive processes during biodegradation of a hydrocarbon plume under transient groundwater flow conditions. Journal of Contaminant Hydrology 59:19

Ranjithan S, Eheart, J. W., and Garrett, J. H. (1993) Neural network based screening for groundwater reclamation under uncertainty. Water Resour Res 29 (3):12

Rogers LL, Dowla, F. U. and Johnson, V. M. (1995) Optimal field-scale groundwater remediation using neural networks and the genetic algorithm. Environmental Science & Technology 29:11

Savic D A WGAaDJW (1999) Genetic programming approach to rainfall-runoff modelling. Water Resour Manage 13:12

Sheta AF, and Mahmoud, A., (2001) Forecasting using genetic programming. 33rd Southeastern Symposium on System Theory: 5

Sreekanth J, Datta, B., (2010) Multi-objective management models for optimal and sustainable use of coastal aquifers. Journal of Hydrology 393:11

Sreekanth J, Datta, B., (2011a) Comparative Evaluation of Genetic Programming and Neural Network as Potential Surrogate Models for Coastal Aquifer Management Water Resour Manage (25):18

Sreekanth J, Datta, B., (2011b) Coupled simulation-optimization model for coastal aquifer management using genetic programming-based ensemble surrogate models and multiple-realization optimization. Water resource management 47 (W04516):17

Sreekanth J, Datta, B., (2012) Genetic programming: efficient modelling tool in hydrology and groundwater management. Genetic Programming - New Approaches and Successful Applications, Dr Sebastian Ventura Soto (Ed), InTech 01/2012

Sun J (2004) A Three-Dimensional Model of Fluid Flow, Thermal Transport, and Hydrogeochemical Transport through Variably Saturated Conditions. M S Thesis Department of Civil and Environmental Engineering, University of Central Florida, Orlando, FL 32816

Tebes-Stevensa C, Valocchia, A. J., VanBriesenb J. M., Rittmannb, B. E. (1998) Multicomponent transport with coupled geochemical and microbiological reactions: model description and example simulations. Journal of Hydrology 209:16

Waddill DW, Widdowson, M.A. (1998) SEAM3D: A numerical model for three-dimensional solute transport and sequential electron acceptor-based bioremediation in groundwater. Technical report, Virginia Tech, Blacksburg, Virginia

Wang WC, K. W. Chau, Cheng, C. T., and Qiu, L., (2009) A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. J Hydrol 374 (3–4):12

Yeh GT, Cheng, J.R. and Lin, H.C. (1994) 3DFEMFAT: User's Manual of a 3-Dimensional Finite Element Model of Density Dependent Flow and Transport through Variably saturated Media. Technical Report submitted to WES, US Corps of Engineers, Vicksburg, Mississippi Department of Civil and Environmental Engineering, Penn State University, University Park, PA 16802

Yeh GT, Tripathi, V. S (1991) HYDROGEOCHEM: A coupled model of hydrologic transport and geochemical multicomponent equilibria in reactive systems. Environmental Science Division Publication No. 3170, Oak Ridge National Laboratory, Oak Ridge, TN

Yeh, G.T., Cheng, J.R., Lin, H.C.,(2000) *Computational Subsurface Hydrology Reactions, Transport, and Fate of Chemicals and icrobes.* Kluwer Academic Publishers.

Zechman E, Baha, M., Mahinthakumar, G., and Ranjithan, S. R., (2005) A genetic programming based surrogate model development and its application to a groundwater source identification problem. ASCE Conf 173 (341)

Zheng C, Wang, P.P. (1999) MT3DMS: A modular three-dimensional multispecies model for simulation of advection, dispersion and chemical reactions of contaminants in groundwater systems; Documentation and User's Guide. Contract Report SERDP-99-1, US Army Engineer Research and Development Center, Vicksburg, MS

Zhou, X., Chen, M., Liang, C., (2003) *Optimal schemes of groundwater exploitation for prevention of seawater intrusion in the Leizhou Peninsula in southern China.* Environmental Geology, 43: p. 8.