DEVELOPMENT OF EFFICIENT SIMULATION-OPTIMIZATION METHODOLOGY FOR IDENTIFICATION OF GROUNDWATER POLLUTION SOURCES USING META-HEURISTIC HYBRID OPTIMIZATION METHODS

Thesis submitted in partial fulfilment of the requirements
for the award of the degree of

Doctor of Philosophy

By

Leichombam Sophia
(Roll No. 126104021)

Under the Guidance of
Prof. R.K. Bhattacharjya
Professor
Department of Civil Engineering

Department of Civil Engineering
Indian Institute of Technology Guwahati
Guwahati 781039 Assam India
April 2018

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati

Guwahati – 781 039
Assam, INDIA

Phones: 0361-2582428 (O)
0361-2584428 (R)
Fax: 0361-2582440
E-mail: rkbc@iitg.ernet.in
rajibkbc@gmail.com
Web: http://www.iitg.ernet.in/rkbc

DEPARTMENT OF CIVIL ENGINEERING

Dr. Rajib Kumar Bhattacharjya
Professor

Date: 26 March 2019

Certificate

This is to certify that the thesis entitled **"Development of Efficient Simulation-Optimization Methodology for Identification of Groundwater Pollution Sources Using Meta-Heuristic Hybrid Optimization Methods"** submitted by **Leichombam Sophia** to the Department of Civil Engineering, Indian Institute of Technology Guwahati is a record bonafide research work carried under my supervision. This thesis work in my opinion has reached the requisite standard fulfilling the requirements for the award of the degree of Doctor of Philosophy. The research work contained in the thesis have not been submitted in part or full to any other University or Institute for the award of any degree or diploma.

Place: IIT Guwahati
Date:

(Prof. Rajib Kumar Bhattacharjya)
Professor
Department of Civil Engineering
Indian Institute of Technology, Guwahati

Declaration

I hereby declare that this PhD thesis entitled "Development of Efficient Simulation-Optimization Methodology for Identification of Groundwater Pollution Sources Using Meta-Heuristic Hybrid Optimization Methods" has been performed by me for the degree of Doctor of Philosophy in Civil Engineering under the guidance of Prof. Rajib Kumar Bhattacharjya, Department of Civil Engineering, IIT Guwahati.

This work has not been submitted for the award of any degree or any diploma at this institute or any institute. I have consulted some of the earlier research works, which I have clearly quoted them. Furthermore, I have acknowledged all the help which I have acquired during this course of time.

Place: **IIT Guwahati**

Date:

Leichombam Sophia
Research Scholar
Department of Civil Engineering
Indian Institute of Technology Guwahati
Assam-781039, India

Dedicated to

My Mother…

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Rajib Kumar Bhattacharjya for his painstaking guidance and valuable suggestions I received throughout the research work. I really appreciate Sir for motivating me during tough days and realizing me that hardship indeed will pay off. It was a great opportunity to pursue my Phd programme under an expertise researcher. I am grateful to the Chairman of my Doctoral Committee, Prof. Arup Kumar Sarma and other members Dr. Bimlesh Kumar and Dr. Karuna Kalita for suggesting constructive ideas which helped me refined my thesis work.

I will be always indebted to Dr. Triptimoni Borah Maa'am. She has been kind enough to extend her help and supported me during the early phase of my research work. My sincere thanks to Dr. Swapnali Barman Di for accompanying me during the initial stage of my thesis work.

I own a great debt to Mr. B.G. Rajeev Gandhi for assisting me in my crucial segment of the research work. He never hesitated to help me whenever I approached him and felt that I was being accompanied by my younger brother.

My special thanks to Mr. Bhrigumani Sharma for all the help he provided during my research work. I would further like to thank Mr. Dilip Kumar Jha Sir and Ms. Mamata Das for their support. I am grateful to all my Subsurface lab-mates Mr. Debraj Biswas, Mr. Jagadish Talukdar and Mr. Subhadip Chakraborty for sharing their ideas and thoughtful discussions during our tea breaks.

I wish to thank my good friends Ms. Ngangkham Devarani and Sanasam Sunderlal for helping me wholeheartedly in completing my thesis. Above that, they were always there to support emotionally and motivated me during the hour of difficult days. My sincere thanks to my friends Thoudam Rubina and Thokchom Bebina for the all the good days we spent together.

I am thankful to my friends Mr. Satish, Mrs Sreedevi Moharana, Ms. Rupasree Panda, Mrs. Sreya Dhar, and Mrs. Geetimukhta Mahapatra who I came across during my stay in the IIT campus.

I would cordially like to thank Ms. Jonali Saikia Maa'm for sharing her thoughtful ideas and suggestions. My sincere gratitude to Mr. Bazal Hoque and his family for being very kind and suggestive during my stay in the IIT campus.

I wish to express my sincere thanks to all the teaching and non-teaching staffs of IIT Guwahati who helped me in some way or the other way during my research work.

The working ambiance of IIT Guwahati and the beautiful environment has always motivated me to be more determined with any work I am related and work earnestly. I will always harness the moments during my stay in the IIT Guwahati campus.

I want to appreciate my colleagues of College of Food Technology, CAU Imphal Dr. Maibam Punyakishore, Prof. Maibam Somarjit, Dr. Thangjam Gopeshwor Singh, Mr. Saroj Kumar Behara and Dr. Angam Raleng for their help and support at the later stage of my research work.

My deep gratitude to my parents Mummy and Baba, my elder brother Leichombam Kenan sister in-law Mrs. Ngangom Lira and younger brother Mr. Leichomabam Prashant for their faithful support and motivation during the phase of my research work. I also thank Mr. Konsam Victor Singh for his constant support during the entire phase of my Phd work.

Leichombam Sophia

Abstract

The alarming increase in the rate of groundwater contamination has motivated the hydro-geologists to work on identification of groundwater pollution sources. The identification of groundwater pollution sources is the initial step for sustainable management of a groundwater aquifer. The groundwater pollution sources can be identified by using the inverse optimization technique. In this technique, an error function is formulated which minimizes the absolute difference between the observed and the simulated contaminant concentration at observation locations. The observed concentration can be obtained from the field whereas, the simulated concentration is obtained by using groundwater simulation model. Hence, the groundwater simulation model is required to be incorporated to the optimization model. As groundwater simulated model is linked to the optimization model, the technique is called the simulation-optimization model. The model is computationally expensive as the simulation model is repeatedly used by the optimization model. Therefore, the performance of the groundwater source identification model is related to the efficiency of the groundwater simulation model. To overcome this computational burden involved, the artificial neural network (ANN) model can be used as an approximate groundwater simulator. It has been reported that for a large aquifer system, a single ANN model is not sufficient to simulate the flow and transport processes of the aquifer and separate ANN model is required for each of the observation wells to simulate the process. However, use of large number of ANN models will increase the computational complexity of the ANN-based simulation-optimization model. Considering these aspects, an ANN-GA based methodology is proposed for identifying the groundwater pollution sources using optimal number of observation wells. This methodology could successfully identify the pollution sources for large aquifer system. The limitation of this approach is that the number of pollution sources and the locations are known to the problem. However, in real life situation, the number and the locations of the groundwater pollution sources are completely unknown. Henceforth, an iterative based approach using Groundwater Modeling System (GMS) and Genetic Algorithms (GA) has been proposed for identification of groundwater pollution sources where both the source locations and the fluxes are unknown. In this approach, the search process has been initiated considering two pollution sources in the aquifer. The number of pollution sources is then successively increased until the exact number of pollution sources are

identified. This algorithm is very efficient in identifying the number of pollution sources. However, some discrepancies have been seen in prediction of the source fluxes. This is so because the source identification problem is a mixed integer problem comprising discrete variables (source locations) and continuous variables (source fluxes). The genetic algorithm is very efficient in handling discrete variables. On the other hand, the gradient based classical algorithms are efficient in handling the continuous variable problem. Therefore, a modified GA-Gradient based algorithm with a local location search algorithm has been proposed for efficient identification of the number, location and flux of the unknown pollution sources. It has been observed that the efficiency of the algorithm is highly related to the initial solution supplied to the Genetic Algorithms. As such, a methodology has also been suggested for generation of the initial solutions using the information of the velocity field of the aquifer and the observed breakthrough curve. This has enhanced the convergence of the proposed modified GA-Gradient based local search algorithm. The performance of the developed methodologies has been evaluated using different illustrative study areas. After analysing the results, it is seen that the proposed methodologies could effectively identify the groundwater pollution sources.

Contents

List of Figures

List of tables

List of Symbols

F	Objective Function value
$C_{o,i}^{j}$	Observed concentration at j^{th} time step for i^{th} well location
$C_{s,i}^{j}$	Simulated concentration at j^{th} time step for i^{th} well location
z_i	Binary number
M	Total number of observation wells
N	Total number of time steps
M_{max}	Maximum permissible wells
M_{min}	Minimum permissible wells
P	Total no of well locations
K	Total no of time period
w_{ij}	Weights assigned
D	Large constant value assigned by user
K_{xx}	Hydraulic conductivity along x-direction
K_{yy}	Hydraulic conductivity along y-direction
K_{zz}	Hydraulic conductivity along z-direction
h	Hydraulic head
S_s	Specific storage coefficient
t	Time
W	Recharge flux per unit area
C	Dissolved concentration in the groundwater
θ	Porosity of the subsurface medium
x_i	Distance along the respective Cartesian co-ordinate axis
D_{ij}	Hydrodynamic dispersion coefficient tensor
v_i	Seepage or linear pore water velocity
q_s	Volumetric flow rate per unit volume of aquifer
C_s	Concentration of the source or sink flux
$\sum R_n$	Chemical reaction term
$I_1 \ldots I_T$	Inputs from 1 to T
I_j	Input vector
e	Calculated error
b	Bias
f	Activation function
y	Output
O	Output vector
w_1	Weight matrix between the input and the hidden layer
W_m	Weight matrix between the hidden and the output layer
t_{f1} and t_{f2}	Transfer function for the neurons in the hidden and output layer respectively
b_1 and b_2	Bias in the hidden layer and output layer respectively
Δw_{I1a}	Increase in weight between node I_1 and a

$\overline{C_{o,i}^{J}}$ $\overline{C_{s,i}^{J}}$	Mean of the observed and simulated concentration respectively
R	Coefficient of correlation
P_s	Probability of the string
p	Population size
$S1...S2$	Pollution sources
$P1...P2$	Pumping wells
$W1...W30$	Observation wells
ε	Effective porosity
α_L	Longitudinal dispersivity
α_T	Transverse dispersivity
n	Number of pollution sources
C_v	Concentration vectors of the simulated concentration
Sf	Vector of the pollutant source fluxes
X	Source location vectors
PCo_i^j	Perturbed simulated concentration value
err	Error term to be introduce
σ	Noise level
$q_{s'}$	Rate of change in transient groundwater storage
$L_1 \ldots L_2$	Orthogonal lines
P_{ij}^k	Probability
p_1	Particle
W_k	Weightage assigned at well 'k'
C_p	Concentration at the peak
$Ef_{p,r}$	Estimated source fluxes
$Af_{p,r}$	Actual source fluxes
P_1 and P_2	Parent 1 and Parent 2
V_x and V_y	Velocity field in x and y direction

List of Abbreviations

ARE	Average Absolute Relative Error
ACO	Ant Colony Optimization
ACO-LTM	Ant Colony Optimization-Long Term Monitoring
ANN	Artificial Neural Network
FEFLOW	Finite Element subsurface Flow
FFBP	Feed-Forward Back Propagation
GA	Genetic Algorithm
GA-LS	GA-Local Search
GA-LTS	Genetic Algorithm-Longitudinal Transverse Search
GA-LTS-GR	Genetic Algorithm-Longitudinal Transverse Search-Gradient search
GA-MS-GR	Genetic Algorithm-Mutation Search-Gradient search
GA-RMS-GR	Genetic Algorithm-Ripple Migration Search-Gradient search
GMS	Groundwater Modelling System
GUI	Graphical User Interphase
HDF5	Hierarchical Data Format 5
HS	Harmony Search
IGA	Improved GA
LM	Levenberg-Marquardt
LS	Local-Search
LTM	Long-Term-Monitoring
LTS	Longitudinal Transverse Search
MATLAB	Matrix Laboratory
MF2K-GWT	MODFLOW2000-Groundwater Transport
MIP	Mixed Integer Programming
MODFLOW	Modular Finite Difference Groundwater Flow Model
MRE	Minimum Relative Entropy
MS	Mutation Search
MT3DMS	Modular Three-dimensional Multi Species Transport Model
MSE	Mean Square Error
NLP	Non-Linear Programming
NLP1	Non-Linear Programming 1
NLP2	Non-Linear Programming 2

OOA	Ordinal Optimization Algorithm
OSIM1	Optimal Source Identification Model 1
OSIM2	Optimal Source Identification Model 2
PAT	Pump and Treat
PCM	Point Collocation Method
PCM-PSO	Point Collocation Method-Particle Swarm Optimization
PGA	Progressive Genetic Algorithm
PNNs	Probabilistic Neural Networks
PSVMs	Probabilistic Support Vector Machines
PSO	Particle Swarm Optimization
RE	Relative Error
RMS	Ripple Migration Search
RMSE	Root Mean Square Error
SA	Simulated Annealing
SA-MRE	Simulated Annealing- Minimum Relative Entropy
SATS-GWT	Simulated Annealing-Tabu Search-Groundwater Transport
SATSO-GWT	Simulated Annealing-Tabu Search-Ordinal optimization algorithm-Roulette-wheel-Groundwater Transport
SUTRA	Saturated Unsaturated Transport
TS	Tabu Search
USGS	United States Geological Survey
VEGA	Vector-evaluated GA

Chapter 1

Introduction

1.1 Overview

Groundwater plays a substantial role in the survival of humankind. About 2.5 billion of the population worldwide solely depends on the groundwater for its sustainability (UNESCO, 2012). About 50% of the groundwater is used for potable supplies, 40% is used in the industrial field and remaining 20% is used for agriculture purposes (UNESCO, 2003). However, the increase in the population and recent growth in the industrial as well as agricultural sectors has led to the gradual decline in the quality of groundwater. Chemicals from the industries, overuse of the pesticides in the agricultural field, leakage from the urban sewers and pipelines, improper dumping and open defecation are some of the persist examples which led to the groundwater contamination. The contamination in the aquifer may be undetected for years as the plume from these pollution sources moves at a very slow rate. This indicates that the groundwater contamination is a very slow process and life-threatening to the humankind. Considering these impacts, many hydrogeologists have been motivated to work on the identification of unknown groundwater pollution source so that necessary remedial measures can be taken up at an early stage. The concern now is to adopt accurate and cost-efficient remedial measures. Therefore, an accurate and an efficient characterization is the initial step for controlling and remediating the subsurface pollution. The characteristics of the pollution source are its location, its type (point or distributed), its duration and the magnitude of its flux.

1.2 Identification of groundwater pollution sources

The pollution source locations and the fluxes can be identified using the inverse optimization approach (Gorelick et al., 1983; Yeh, 1986; Datta et al., 1989; Bagtzoglou et al., 1992; Skaggs and Kabala, 1995Aral and Guan, 1996; Mahar and Datta, 1997, 2000, 2001; Aral et al., 2001, Datta and Chakrabarty, 2003, Mahinthakumar and Sayeed, 2005; Li et al., 2006; Bhattacahrjya and Datta, 2005, 2007, 2009; Ayvaz, 2015, 2016; Gandhi et al., 2016; Leichombam and Bhattacharjya, 2016). In inverse

optimization model, the algorithm tries to minimize the difference between the simulated and the actual contaminant concentrations at the observation locations. The actual concentrations are measured in situ. Whereas the simulated concentration is obtained using aquifer simulation model. The aquifer simulation model solves the groundwater flow and transport processes. As such, the identification of groundwater pollution source requires incorporation of the simulation model with the optimization model. The methodology is known as the simulation-optimization methodology. The simulation model is repeatedly called by the optimization model until an optimal solution is achieved. For this reason, the simulation-optimization approach relies on how effectively the simulation model can solve the groundwater flow and transport processes.

There are numerous techniques available for incorporating the simulation model with the optimization model. One of the earlier techniques that was conceived among the researchers is the response matrix approach for identifying the groundwater pollution source and is based on the principle of superposition and linearity (Gorelick, 1983). Based on the concept of linear programming, response matrix approach was applied by many in groundwater source identification problems (Gorelick and Remson, 1982; Gorelick et al., 1983), parameter estimation problems (Tyson and Weber, 1964; Kleinecke, 1971; Newman 1973) etc. The first attempt at source identification using response matrix method was performed by Gorelick and Remson (1982). In this approach, it is assumed that the aquifer behaves as a linear system and the principle of linearity can be applied. However, this approach is unsatisfactory for the highly nonlinear system. A large computational error is observed in the result when the source identification problem is solved for heterogeneous aquifers. Considering the limitation of the response matrix technique, the embedded approach came to light. In embedded approach, the governing equations of groundwater flow and transport processes are incorporated as the binding constraint to the optimization model (Aguado and Remson, 1974; Alley et al., 1976; Willis and Newman, 1977; Gorelick, 1983; Elango and Rouve, 1980; Peralta and Datta, 1990; Mahar and Datta, 2001; McPhee and Yeh, 2008; Datta et al., 2009). Though the application of embedded technique could be widely seen in the various groundwater management problems, some limitations of the algorithm have been reported. The embedded technique becomes computationally very expensive for a large-scale aquifer as the size of the constraints will be very large. To cope up with all limitations involved in these two approaches, the linked simulation-optimization came

into existence. Notable applications of linked simulation-optimization model have been reported in the field of source identification problems as well as in groundwater management problems (Aral and Guan, 1997; Aral, et al., 2001; Datta and Chakrabarty, 2003; Coppola et al., 2003; Singh et al., 2004; Bhattacharjya and Datta, 2005; Singh and Datta, 2006; Bhattacharjya et al., 2007; Ayvaz, 2010, 2015, 2016; Datta et al., 2011, Prakash and Datta, 2013, 2014a, 2015 etc.). In the linked simulation-optimization model, the aquifer simulation model is externally linked to the optimization model and a large number of calls of the simulation model is generally necessary by the optimization model to reach an optimal solution. Thus, the performance of the linked-simulation model relies on the computational efficiency of the simulation model (Bhattacharjya and Datta, 2005). Various type of groundwater simulation models such as SUTRA (Voss, 1984; Datta, et al., 2011), MODFLOW (Yeh et al., 2007; Ayvaz 2010; Jha and Datta, 2011; Datta et al., 2013; Borah and Bhattachrjya, 2014), MT3DMS (Ayvaz, 2010; Borah and Bhattacharjya, 2014; Datta, et al., 2013), FEFLOW (Diersch, 2002; Zhao, 2005; Huo et al., 2007), FEMWATER (Lin et al., 1997; Bhattacharjya, et al., 2007) etc. are available for simulating the flow and transport processes in an aquifer. These models are capable of solving the complex groundwater phenomenon and can be linked with the optimization model.

An efficient optimization algorithm is required for making the linked simulation-optimization model more productive. The gradient-based classical optimization algorithms can be applied to solve the inverse optimization model. The application of gradient-based search is reflected in some of the earlier research work carried out by Alley et al. (1976), Willis and Neuman (1977), Gorelick and Remson (1983), Peralta and Datta (1990), Peralta et al. (1991), Wang and Ahfeld (1994), Mahar and Datta (1997) and Mahar and Datta (2000), Datta et al. (2009) etc. But the gradient-based technique always yields a local optimal solution and it is not guaranteed that it will converge towards the global optimal solution. As the classical approach also depends on the initial solution, various number of simulation runs are to be carried out using different initial solutions to reach the global optimal solution. But this is not a promising technique as it is a time-consuming process and the process may not yield the global optimal solution. In order to overcome some of the limitations of the gradient-based classical optimization algorithms, many researchers have applied the non-classical approaches. Some of the non-classical approaches are Genetic Algorithm (GA), Simulated Annealing (SA), Particle Swarm Optimization (PSO) which can be

used to solve the source identification problems. Among all these, GA has been proven to be one of the robust algorithms for finding the global optimal solution of non-linear non-convex problems. However, as first-order optimality condition is not used in GA, it generally yields the near global optimal solution. As such another algorithm, preferably gradient based is needed to obtain the actual optimal solution.

The performance of the linked simulation-optimization model is related to the efficiency of the simulation model as large number of simulation calls is necessary to reach the optimal solution. To overcome this computational burden, an approximate simulation model can be used as a surrogate model in place of t he numerical simulation model. The artificial neural network (ANN) model is one of the most effective and popular models used for replacing the numerical aquifer simulation model. The basic processing unit of ANN model is the artificial neuron and it has a close analogy with the neuron present in the human brain. The capability of ANN to gather knowledge through learning from sufficient input pattern enables ANN to apply to real-world problems (Bhattacharjya and Datta., 2009). Some of the application of ANN in the field of groundwater management studies could be seen in past few decades as performed by Coppola et al. (2003); Singh et al. (2004); Singh and Datta (2007); Bhattacharjya and Datta (2005, 2009) and Borah and Bhattacharjya (2014).

1.3 Research gap from the existing work

Based on the work presented by the other researchers on similar topics, the following research gaps were identified.

i. The ANN based surrogate models were used only to predict the concentration data at different observation wells. The performance can be improved by using different ANN models for different observation well locations.
ii. The source identification problems discussed in the previous research have mostly assumed that the location of the sources are known. This is a major research gap as mostly in the field situations, the source locations are usually unknown.
iii. The Genetic Algorithms are efficient in handling the discrete variables and the gradient based minimization is efficient in handling the continuous variables. Combining these two algorithms can result in a much efficient algorithm for identifying the number, location and source fluxes.

iv. The data from the observation wells is generally used only to formulate the objective function and all the locations are given equal weightage while identifying the source locations. This is also an unexplored area which can be attempted in this thesis.

1.4 Objectives of the thesis

The objectives of the thesis can be summarized as follows:

i. Development of an aquifer simulation model using MODFLOW and MT3DMS available in Groundwater Modeling System (GMS).
ii. Development of Artificial Neural Network (ANN) model for simulating the groundwater flow and transport processes.
iii. Development of groundwater pollution source identification model using ANN-GA based linked simulation-optimization model (Model 1).
iv. Identification of the optimal number of observation wells using Model 1.
v. Development of an iterative based source identification model considering the number of sources and locations as unknown (Model 2).
vi. Development of a modified GA-Local Location Search-Gradient Search model (Model 3) for identifying the unknown pollution sources.
vii. Development of a source identification model (Model 4), in which the initial population is selected from a set of a pool consisting of most probable locations.
viii. Performance evaluation of the proposed methodologies using illustration study areas.

1.5 Summary of the thesis

As discussed earlier, for large-scale aquifer system, the performance of the ANN-based surrogate model is not satisfactory when a single ANN model is used to predict the concentration at different observation well locations (Borah and Bhattacharjya, 2014). In such a situation, the model efficiency can be enhanced by developing separate ANN model for each of the observation locations. Thus, the number of ANN models is equal to the number of observation wells in the aquifer. As a result, the complexity of the ANN-based simulation-optimization model will be related to the number of observation wells. Considering this factor, a methodology has been proposed which will only select the optimal number of observation wells required for identifying the pollution sources for different management periods. The data required for training the ANN model is

generated using the groundwater flow (MODFLOW) and transport (MT3DMS) models. A large number of input-output patterns was generated for training, testing and validating the ANN model. For evaluating the performance, an illustrative large study area (Borah and Bhattacharjya, 2014) is adopted. Observation wells were placed in the affected aquifer for observing the concentration break through curves. For each of these observation wells, an ANN model is developed for predicting the contaminant concentration at different time steps. These developed ANN models were then linked with the optimization model. As discussed above, there are numerous optimization algorithms which can be linked with the groundwater simulator. In the present study, the genetic algorithm (GA) is adopted as the optimization algorithm. As the present methodology involves in linking of the ANN-based simulator with the GA based optimization model, this approach is called ANN-GA model.

For real life scenario, the locations of the groundwater pollution sources are generally completely unknown. This increases the level of complexity of the problem. Thus, the emphasis on identifying the pollution source should be not only to source flux but also the source locations. Taking it into consideration, an iterative based inverse optimization model is proposed for identification of unknown groundwater pollution sources where both the source locations and the fluxes are unknown.

As no information about the locations and the number of pollution sources is available, the search for the optimal solution was initiated considering two sources. The number of the pollution sources has successively increased until a solution with dummy source is not achieved. The convergence towards the optimality was evaluated on the basis of objective function value.

Furthermore, it may be noted that it is not always possible to accurately measure the concentration data at the field. As such, there is a need to introduce a measurement error in the observation data. Henceforth, an analysis has been carried out to evaluate the performance of the model when there is some error in the observed data. To check the performance evaluation of this proposed methodology, a hypothetical study area is considered. The study area is similar to that considered by Mahar and Datta (2000). However, the number of sources and source locations are different. The groundwater flow and transport processes are simulated for 5 years at an interval of three months. The simulation is carried out using Groundwater Modeling System (GMS). There are four pollution sources present in the aquifer which are active for four-time steps. But it may be noted that the location and the number of the pollution sources are completely

anonymous to the optimization model. In this case, also, the optimization model is solved using the genetic algorithm.

The developed methodologies identified the source number and their locations effectively. However, the source fluxes could not be determined efficiently. So, there is a need to develop a methodology which is capable of identifying the source location as well as the source flux effectively. In groundwater source identification problem, the source locations are discrete variables and on the other hand, the source fluxes are continuous variables. The genetic algorithm is very efficient in handling discrete variable. Thus, it is efficient in determining source locations. However, the source flux being continuous variables can be effectively determined using a gradient-based search algorithm. Considering the advantages of these individual algorithms, a modified search algorithm has been proposed in this study. In this methodology, the global optimal is reached step by step. The first step is to reach a location close to the actual source by using modified GA. The second step is to improve the location based on three local location search methods proposed in this study. The third step is to find the actual flux at the source by using a gradient-based classical optimization algorithm.

The genetic algorithm is modified in such a way that two types of variable are handled differently. In order to handle both the type of variables, binary operators are used for location variables and real operators are used for flux variables. The termination criterion for the modified genetic algorithm is the stall location. The stall location is also evaluated at each iteration. This is used in calculating the percentage of mutation and crossover that will be performed to the population in the corresponding iteration. The increase in stall location increases the mutation for the locations, whereas the real crossover and mutation are unaffected by the stall location. This assures that the location is always near the optimal (as the best location will be unaffected by diverging the population) and the flux values always converge to an optimal solution. As the modified GA terminates, it assures that the location is near the optimal location. The algorithm will now proceed for three different methodologies for local location search. In the local location search, the gradient-based search is also performed so that the improved locations converge towards the actual source location with an upgraded source flux. The termination criterion for the local location search is based on first-order optimality. For this case also, the performance of the methodology is evaluated by considering the illustrative study area adopted by Mahar and Datta (2000). However, the number of pollution sources considered is three with eight number of observation

wells. It is observed that all the three algorithms could successfully identify the pollution source location and the fluxes.

The first step in modified GA is the generation of the initial population. As the initial population is randomly generated, the probability of producing an initial population with same traits does not happen every time. To overcome this problem a better methodology is proposed that reduces the likelihood of selecting the redundant locations as sources. A set of locations comprising of most likely source locations is generated using the information available in the observed breakthrough curve and the velocity vectors of the aquifer. The initial solutions picked up from this generated pool ensured a much-improved optimal solution. The performance of this methodology is checked using illustrative study areas of different number of grids and different number of pollution sources. The evaluation of the result shows that the model could successfully identify the unknown pollution sources for all the cases.

1.6 Organisation of the thesis

There are seven chapters in this thesis including the present introductory chapter. A brief review of each chapter is presented as follows:

Chapter 1: This chapter presents a concise discussion about the groundwater contamination and some of the earlier works carried out on identification of groundwater pollution sources. Based on the discussion of the earlier works, objectives of the present study have been proposed in the current chapter.

Chapter 2: This chapter discusses detailed studies on identification of groundwater pollution sources performed by researchers in the global arena. It showcases the different techniques used in identifying the groundwater pollution sources namely response matrix approach, embedded approach, and the linked simulation-optimization approach. Various methodologies adopted using these approaches have been explained in chronological order.

Chapter 3: In this chapter, the use of groundwater approximate simulator using ANN model has been discussed. An ANN-GA based model has been developed for identifying the groundwater pollution sources. To check the performance of the developed methodology, an illustrative study area from an earlier work has been adopted.

Chapter 4: In most of the groundwater identification model, the location of the pollution sources is known to the modeler. However, in real life scenario, the

information about the pollution sources is seldom available. Considering this factor, a methodology has been proposed in the present chapter which will search for the exact number and locations of the pollution source iteratively.

Chapter 5: This present chapter describes how a combination of global and local search can efficiently identify the pollution sources. The source identification problem is a mixed integer problem in which the source locations are the discrete variables whereas the source fluxes are the continuous variables. For this reason, a modified GA is developed for handling both types of variables. Solutions obtained from the modified GA are used to find the exact source locations and flux close to actual fluxes using three different local location search algorithms, namely LTS, MS, and RMS.

Chapter 6: The first step in the modified GA is the random generation of the initial population. As the initial population is randomly generated, the probability of producing near global optima does not occur every time. To overcome this problem, a methodology is developed which will select the initial solution from a generated pool comprising probable locations and assists in converging towards optimal solution effectively.

Chapter 7: In this chapter, a brief summary and conclusions of the present research work are presented. The conclusion is based on the developed methodologies and the results evaluated from each of the illustrative study areas. It further describes the future scope of the present study that can be carried out.

Chapter 2

Review of Literature

This chapter provides a brief discussion about the identification of groundwater pollution sources and the methodologies developed by researchers in the global arena. The first part of the chapter describes the overview of the earlier work carried out in the field of groundwater identification problem. The second part of the chapter explains the unknown pollution sources and the characteristics of the pollution sources. The third part describes the different approaches of simulation optimization techniques i.e. response matrix, embedded approach and linked simulation-optimization adopted in identifying the groundwater pollution sources subsequently. The various techniques of non-classical optimization algorithms used for identification of groundwater pollution sources are subsequently described. The final fifth part describes the application of hybrid optimization technique used in the field of groundwater management problems.

2.1 Overview

The demand for fresh water has been increasing at an alarming rate due to the rise of the population and substantial growth in industry and agriculture. This has resulted in the scarcity of water which has affected the livelihood of some of the underdeveloped countries significantly. At the same time, the quality of the water is also deteriorating in many parts of the world due to various human activities. Even though groundwater is considered as the purest form of water, numerous unwanted human activities like spillage, improper dumping, leakage in the septic systems, spillage from industries, overuse of pesticides from agricultural fields, unhygienic practices from individuals etc. have led to an alarming rate of groundwater contamination (Atmadja and Bagtzoglou, 2001).

The contamination in the aquifer may remain undetected for an extended period until some significant changes are noted in the quality of the extracted water (Sun et al., 2006b). Apart from this, the spreading of the plume will also accelerate the rate of contamination which ultimately affects the quality of water sources such as lakes, rivers, streams etc. The rate at which the water resources are contaminated, has grabbed the attention of the researchers worldwide. As such, an immense attempts have been

made by the researchers to develop various groundwater remediation techniques (Kleinecke, 1971; Maddock, 1972; Newman, 1973; Frind and Pinder, 1973; Nutbrown, 1975; Alley et al., 1976; Navarro, 1977; Elango and Rouve, 1980; Heidari, 1982; Gorelick, 1982; Gorelick et al., 1983; Yeh et al., 2007; Willis and Finney, 1985; Datta and Peralta, 1986; Peralta and Kowalski, 1986; Reilly and Goodman, 1987; Ahlfeld et al., 1988; Ahlfeld, 1990; Wang and Ahlfeld; 1994; Ahlfeld and Heidari, 1994; Aral and Guan, 1996; Datta and Dhiman, 1996; Mahar and Datta, 1997, 2000, 2001; Das and Datta, 2000, Mahinthakumar and Sayeed, 2005; Dhar and Datta, 2007; Chandalavada and Datta, 2008; Dhar and Datta, 2009; Ayvaz, 2010, 2016; Kollat et al., 2011; Chandalavada et al., 2011; Prakash and Datta, 2013, 2014; Jha and Datta, 2015; Gurarslan and Karahan, 2015; Zhao et al., 2015).

2.2 Unknown Groundwater pollution source

The first initial step for remediation of groundwater pollution is the accurate characterization of the pollution sources. The characteristics of the pollution sources at most of the time are unknown when the contamination is first detected. However, if the source characteristics are precisely known and the aquifer parameters are evaluated, then the nature and the transport of the source pollutant can be studied and remedial measures may be adopted. The characteristics of the pollution source include:

i. The location of the pollution source
ii. Whether the type of pollution source is a point or distributed
iii. Activity duration of the pollution sources from the time the source becomes active
iv. The magnitude of the flux released from the site of pollution source

The characteristics of the pollution sources can be estimated using the concentration of contamination at the observed well locations. Thus, the identification of groundwater pollution sources is regarded as an inverse problem. The inverse optimization technique has been successfully applied in numerous groundwater problems. Some of the application of inverse optimization problems are Gorelick et al. (1983), Skaggs and Kabala (1994), Snodgrass and Kitanidis (1997), Aral et al. (2001), Mahar and Datta (2001), Mahinthakumar and Sayeed (2005), Li et al. (2006), Bhattacharjya and Datta (2005, 2009), Bhattacharjya et al. (2007), Borah and Bhattacharjya (2014), Ayvaz, (2016), Gandhi et al. (2016), Leichombam and Bhattacharjya (2016) etc.

Detailed report on groundwater management problem solved using inverse optimization technique have been reported by Atmadja and Bagtzoglou (2001), Bagtzoglou and Atmadja (2005), Amirabdollahian and Datta (2013). In order to solve this inverse optimization technique, the objective function minimizes the difference between the simulated and the actual contaminant concentration at the observation well locations. The actual concentration can be collected from field observations. On the other hand, the simulated concentration can be obtained using the aquifer simulation models. As the aquifer simulation model has been incorporated into the optimization model, the methodology is known as the simulation-optimization methodology.

2.3 Simulation-Optimization method

In the simulation-optimization technique, the aquifer simulation model solves the partial differential equations of groundwater flow and transport processes. The groundwater flow and transport processes in an aquifer can be numerically solved using the finite difference method or the finite element method. Many of the researchers have used the finite difference and finite element methods for simulating the groundwater flow and transport processes. This has been reflected in the work of Remson et al., 1971; Wang and Anderson, 1982; Huyakorn and Pinder, 1983; Voss, 1984; Galeati and Gambolati, 1988; Bogardi et al., 1991.

The groundwater simulation model can be coupled with the optimization model using response matrix method, embedded method, and linked simulation-optimization method. The principle of response matrix approach is based on the principle of superposition and proportionality (Bhattacharjya and Datta, 2005). Response matrix approach has been adopted at the earlier stage of groundwater management studies (Heidari, 1982; Willis and Liu, 1984; Peralta and Kowalski, 1986; Yazicigil et al. 1987; Galeati and Gambolati, 1988; Skaggs and Kabala, 1994; Datta and Dhiman, 1996; Aral and Guan, 1996). Heidari (1982) developed a management model based on the linear system theory. Willis and Liu (1984) also formulated a groundwater management problem on the basis of the response equations of the groundwater system. Yazicigil et al., (1987) used response functions approach for developing groundwater development policies for a real aquifer system. They used a simulation model for representing the hydraulic response of the system which is further linked to an optimization model. Galeati and Gambolati (1988) developed a linear mixed integer programming model for designing a dewatering system. The finite element-optimization model was developed

considering the various conditions of the optimal withdrawal and other pumping constraints. Gorelick and Remson (1982) and Gorelick et al. (1983) have successfully applied the response matrix approach in solving groundwater management problems. But one of the major disadvantages was the non-satisfactory performance while solving the non-linear system. Due to this limitation, Mahar and Datta (1997) applied the embedded technique for identification of unknown groundwater sources. The embedded approach was adopted by various researchers (Aguado and Remson, 1974; Futagamiet et al., 1976; Willis and Newman, 1977; Peralta and Datta, 1990; Wang and Ahfeld, 1994; Mahar and Datta, 1997; Mahar and Datta, 2001; McPhee and Yeh, 2008; Datta et al., 2009). As reported, the algorithm has been applied successfully for identification of groundwater pollution sources and also in groundwater management model. However, the application of the algorithm is limited to a small-scale groundwater aquifer only.

To overcome the limitations encountered in these approaches, the linked simulation-optimization method has been developed. The linked simulation-optimization algorithm started gaining importance since the year 2000 in the field of groundwater management and source identification problems. The major advantage of using linked simulation-optimization is that the simulator is externally linked to the optimization model. As such, any type of complex simulator can be easily incorporated into the optimization model. The following sections elaborately explain the different techniques for incorporating the simulation model with the optimization model.

2.3.1 Identification of groundwater pollution source using response matrix

The response matrix approach has been considered to be one of the most effective and reliable techniques for solving unknown groundwater pollution source problem. In this approach, the behaviour of the groundwater aquifer is considered as a linear system. The response matrix is constructed using a groundwater simulation model and is calculated before executing the optimization model. In the earlier stage of groundwater studies, the concept of linear programming was considered to be an acclaimed technique. Based on the notion of linear programming, response matrix approach was applied in groundwater source identification problems (Gorelick and Remson, 1982; Gorelick et al., 1983), parameter estimation problems (Tyson and Weber, 1964; Kleinecke, 1971; Newman 1973). The first attempt in the field of source identification problem using response matrix method was performed by Gorelick and Remson (1982). The source identification problem was formulated as an optimization model where the

simulation model of the groundwater solute transports is incorporated as constraints. The objective is to identify the pollution sources by resulting a close match between the simulated concentration with the local groundwater solute concentration data. Gorelick et al. (1983) further combined the linear programming and multiple regression techniques with the aquifer simulation model for identifying the location and magnitudes of the pollution source. Two different cases were evaluated. The first case was a steady state transport which could identify the unknown pipe leak locations and magnitude of the leak using sparse and spatially distributed contaminant concentrations. The second transient case could successfully identify the annual disposal fluxes utilizing the contaminant concentrations data acquired from the observation well locations. Wagner and Gorelick (1986) presented a methodology for determining the aquifer parameters which characterized the transport of contaminants in an aquifer. The methodology used the non-linear least square regression which was combined with the contaminant transport simulation. The model was evaluated in a mountain stream in northern California and the trends, variability and the interrelationship of the parameters were analyzed.

Although some limitation started to emerge, the use of response approach could still be seen in some of the research studies. An expert system using statistical pattern recognition and stochastic dynamic programming was developed by Datta et al. (1989) in order to identify groundwater pollution source. They used response matrix method for incorporating the simulation model with the optimization model. The effect of parameter uncertainty and measurement error was studied. A preliminary screening was also carried out using the model developed by Gorelick et al. (1983). It was found that the developed methodology was very effective even under the condition of missing observed concentration data. More application of response matrix in solving source identification problem was seen in the study carried out by Wagner (1992). He adopted the maximum likelihood technique. The nonlinear maximum likelihood technique was combined with groundwater flow and transport simulation model. Then the distributed pollutant source term and the aquifer parameters are simultaneously determined. The further use of response matrix in source identification could be seen in the work of Skaggs and Kabala (1994). For recovering the release history of the groundwater contaminants in one-dimensional groundwater system, they have used Tikhonov Regularization (TR) technique. An integral equation was used for representing the linear one-dimensional groundwater transport processes. On the basis of the study, it

was concluded that if a plume is not significantly dissipated, then the release history can be effectively recovered even with moderate level of measurement error. But when the plume is more disperse, the presence of moderate measurement error lowers the accuracy of the recovered release history.

A combinatorial study was later performed by Aral and Guan (1996) using response matrix and genetic algorithms. They found that the results obtained from the model were much better than the linear programming approach. Liu and Ball (1999) used Skaags and Kabala's TR technique to study the contaminant transport at Dover Air Force Base, Delaware. From the site, they recovered the measured concentration data from low permeability aquifer contaminated with Perchloroethane and Trichloroethane chemicals. They concluded that the obtained results fall within the confidence intervals. Many other researchers adopted the response matrix technique for performing various groundwater studies (Heidari, 1982; Peralta and Kowalski, 1986; Yazicigil et al., 1987; Galeati and Gambolati, 1988; Datta and Dhiman, 1996). Even though numerous studies were carried out in the field of groundwater studies, the response matrix technique was limited to the linear system only where the aquifer is homogenous. This approach usually gives erroneous result when solved for a large-scale problem (Rosenwald and Green, 1974). Therefore, to overcome the limitations of response matrix approach, embedded optimization technique has been used.

2.3.2 Identification of groundwater pollution source using embedded approach

In embedded optimization approach, the groundwater flow and transport process are incorporated as binding constraints to the optimization problem. The governing equations in the embedded approach are in the form of finite difference or finite element approximation. One of the earliest application of the embedded technique was employed by Aguado and Remson (1974). They successfully incorporated the two and three-dimensional groundwater equations with the optimization model. Finite difference approximation was adopted for solving the steady and transient problems. Futagami et al. (1976) incorporated the finite element based groundwater equations in a linear programming management model. The objective of the management model was to maximize the optimal discharges from various types of outfalls. Willis (1976) applied the embedded approach for evaluating the disposal of the food-processing using spray irrigation. Further, Willis and Newman (1977) developed an optimal dynamic management model for a heterogeneous anisotropic porous medium. The objective of

the model was set up in order to minimize the operational cost during the planning period. Mahar and Datta (1997) applied embedded approach for identifying the groundwater pollution source. The study was carried out subsequently involving three-step methodology. In the first initial step, a preliminary identification of pollution sources was performed using an embedded optimization model. In the second step, the preliminary results were used to design an optimal monitoring network. In the last step, the concentration data from the designed monitoring network were utilized for more precise identification of groundwater pollution sources.

Mahar and Datta (2001) again utilized the embedded approach for identifying the groundwater pollution sources. They have proposed two different optimization models namely NLP1 and NLP2. The performance of the developed methodology was analyzed using an illustrative study area. It was found that NLP1 can identify the pollution sources with known aquifer parameters. Whereas NLP2 simultaneously estimated the unknown aquifer parameters and pollution sources. It is thus confirmed that embedded approach is a versatile approach for source identification problem. The performance of the developed methodology was carried out under steady-state flow and transient transport conditions. The studies revealed that the performance of the embedded technique is found to be much better than the response matrix. Detailed applications about the embedded approach in the field of groundwater management studies have been presented by Gorelick (1983). Later, Amirabdollahian and Datta (2013) also collectively presented the benefits of the embedded approach as: (i) capable of simultaneously determining the unknown pollution source as well as flow and transport parameters, (ii) can overcome the limitations of response matrix and (iii) possibility of incorporating any complex groundwater flow and transport equations as binding constraints.

However, a contrasting remark was given by Gorelick (1983). Gorelick (1983) reported that embedded approach can successfully solve the problem for a nonlinear system but when a large and highly heterogeneous aquifer system is considered, the complexity of the problem will increase. Moreover, the embedded technique will prove to be computationally very expensive for such problems when compared with the response matrix. The same remark was also given by Tung and Kolterman (1985). Even after countering these limitations, embedded approach has been applied in solving various groundwater management problems (Willis and Newman, 1977; Elango and Rouve,

1980; Peralta and Datta, 1990; Peralta et al., 1991; Wang and Ahlfeld, 1994; McPhee and Yeh, 2008; Datta et al., 2009).

Considering the persisting limitations of the two approaches, the linked simulation-optimization technique has been developed for groundwater management and source identification problems. The linked simulation-optimization approach could be effectively applied even to a large heterogeneous aquifer.

2.3.3 Identification of groundwater pollution source using linked simulation-optimization

In linked simulation-optimization approach, the simulation model is externally linked to the optimization model for solving the groundwater management problem. This approach was developed considering the limitation of response matrix and embedded approach. The linked simulation-optimization model is capable of externally linking any complex simulator with the optimization model. Therefore, it is capable of solving large, heterogeneous and nonlinear aquifer system. The groundwater simulators will simulate the groundwater flow and transport by using the set of governing equations. Here, the optimization will repetitively call the simulation model when it requires any information from the groundwater flow and transport simulators. It is an iterative process where the simulation model will provide the necessary information to the optimization model for deriving the optimal solution. Considering its various advantages, numerous studies have been conducted (Aral and Guan, 1997; Emch and Yeh, 1998; Datta and Chakrabarty, 2003; Coppola, 2003; Singh et al., 2004; Bhattacharjya and Datta, 2005, 2007, 2009; Mahinthakumar and Sayeed, 2006; Singh and Datta, 2006; Singh and Datta, 2007; Datta et al., 2009; Ayvaz, 2010; Datta et al., 2011; Prakash and Datta, 2013; Prakash and Datta, 2014; Borah and Bhattacharjya, 2014; Gurarslan and Karahan, 2015; Ayvaz, 2015, 2016).

The idea of the linked simulation-optimization approach was also seen in the earlier work of Gorelick et al. (1983). In their attempt, the source identification problem was made to run as a forward simulation. The simulation model is coupled with an optimization model with the aid of linear programming and response matrix approach. Here, two different cases of steady and transient transport cases were considered where the objective was to identify the disposal sites. Later, Aral et al. (2001) used linked simulation-optimization for a more complex problem where the source location and the release histories were considered as unknown variables. As the problem is a non-linear

optimization model, the simulation model was repeatedly required for undergoing the optimization process. Henceforth, a combinatorial technique known as Progressive Genetic Algorithm (PGA) was used. From the computational result, it signifies that the proposed PGA approach is a highly robust algorithm for the solution of source identification problem than the other non-linear problems (Aral and Guan 1996; Guan and Aral, 1999). However, the performance of the optimization model may be improved up to a certain degree by using different optimization algorithm. Datta and Chakrabarty (2003) further formulated the linked simulation-optimization model for identification of pollution sources using classical optimization method. This gradient based methodology was able to fairly solve the large simulation-optimization model. The performance evaluation showed that, even with limited missing measurement data sets, the present approach can be used. With the advantage of the groundwater simulators to solve complex groundwater phenomenon, many groundwater management and source identification studies were carried out using different groundwater simulators models such as SUTRA (Voss, 1984; Datta et al., 2011), MODFLOW (Yeh et al., 2007; Ayvaz, 2010; Borah and Bhattacharjya; 2014, Bashi-Azghadi et al., 2010; Jha and Datta, 2011; Datta et al., 2013), MT3DMS (Ayvaz, 2010; Borah and Bhattacharjya, 2014; Datta, et al., 2013), FEMWATER (Lin et al., 1997; Bhattacharjya, et al., 2007), FEFLOW (Diersch, 2002; Zhao et al. 2005; Huo et al., 2007), SEAWAT (Kourakos and Mantoglou, 2011; Rao et al., 2004) etc.

Even though the earlier studies have confirmed the wide applications of the linked simulation-optimization model in the groundwater management studies but one of the main limitations is its computational efficiency. A large number of simulation calls of the simulation model is carried out by the optimization model until an optimal solution is reached. Thus, the performance of the linked-simulation model relies on the computational efficiency of the simulation model (Bhattacharjya and Datta, 2005). For this reason, an attempt was made by many researchers for replacing the highly computationally expensive groundwater simulators. With this regard, an approximate groundwater simulators have been developed using the regression model, Artificial Neural Network (ANN), genetic programming etc. Among all these, ANN has evolved to be one of the best groundwater approximate simulators. From the earlier works on groundwater management studies, it has been confirmed that introduction of ANN enhances the computational efficiency of the linked simulation-optimization model.

Coppola et al., (2003) used ANN model for solving the groundwater management problems.

Singh et al. (2004) utilized ANN model for identifying groundwater pollution sources and also to determine the unknown hydrogeological parameters. The performance evaluation of the model showed that the ANN-based source identification model is capable of identifying the pollution sources and also for estimating the aquifer parameters. It was also found that the ANN model was highly unaffected even if there are measurement errors in the observed concentration data. More utilization of ANN model as approximate simulator was performed by Bhattacharjya and Datta (2005). They developed a saltwater intrusion management model using ANN model. Here, a linked simulation-optimization based ANN-GA model was derived for controlling saltwater intrusion in the coastal aquifer. Evaluated results showed that the developed model is capable of successfully solving the density-dependent flow and transport processes. The proposed model required less computational time compared to embedded approach.

Singh and Datta (2006) used the heuristic global search technique GA, for identification of unknown groundwater pollution sources. In the study, finite difference and Method of Characteristics (MoC) was used for solving steady-state flow equation and transient transport models respectively. The effect of measurement error was evaluated by perturbing different levels of error in the observed concentrations. The evaluation of the results showed that GA is able to handle a moderate level of error but with multiple pollution sources, error increases in source identification problem. The use of ANN as approximate simulator was carried out by Bhattacharjya et al. (2007). They used FEMWATER for generating the data to train and validate the ANN model. It has been reported that the ANN can be effectively used as an approximate simulator for three-dimensional density-dependent flow and transport processes in the coastal aquifer.

Datta et al. (2009) developed a methodology for source identification and simultaneous determination of groundwater parameter. Here, the simulation model is externally linked to a nonlinear optimization model. The performance of the methodology was evaluated using an illustrative study area. From the solution results, it was observed that the proposed methodology was more effective than the embedded approach.

Ayvaz (2009) used linked simulation-optimization based on heuristic Harmony Search (HS) algorithm. MODFLOW is used to simulate the groundwater flow process. The

performance of the model is evaluated on three different management models: (i) maximization of total pumping for a steady state aquifer (ii) minimization of total pumping cost to satisfy the given demand and (iii) minimization of pumping cost in order to satisfy the given demand for different management period. From the solution result, it was remarked that the first and second problem can be solved using HS with a reasonable number of simulations however the third problem requires a large number of simulations to find an optimum solution. He concluded that the problem with many decision variables may be solved by hybridization of the HS with the gradient-based optimization algorithm.

Datta et al. (2011) proposed linked simulation-optimization approach for solving groundwater pollution source characteristics using classical optimization model. In the study, two optimization models OSM1 and OSM2 were developed. The proposed methodology was successfully applied to a large-scale study area and could overcome the computational limitation that was faced by the embedded approach. Performance of the proposed source identification model was evaluated using illustrative study area. They concluded that the proposed model has the capability for solving fairly large study area with the multiple numbers of unknown pollutant sources. For determining the optimal source characteristic, accurate monitoring network design may be required.

Various researchers proposed groundwater monitoring network design for increasing the efficiency of source identification problem. Some of the groundwater monitoring network design includes Loaiciga (1989), Bagtzoglou et al. (1992), Wagner (1992), Skaggs and Kabala (1994) Aral and Guan (1996) Mahar and Datta (1997), Atmadja and Bagtzoglou, 2001; Singh et al. (2002); Michalak and Kitanidis, 2004; Singh and Datta, 2006, 2007; Chandalavada and Datta, 2008; Dhar and Datta, 2009 etc. The effectiveness of source identification model also depends on how effectively the monitoring well locations are placed. Different objectives were considered depending on the type of the problem such as Minimizing the total number of monitoring wells (Meyer et al., 1994; Yenigul et al., 2005; Li and Hilton, 2007); Maximization of detection possibility (Meyer and Brill, 1988 and Meyer et al., 1994); Minimizing the monitoring cost (Datta and Dhiman, 1996; Kollat and Reed, 2007; Reed and Minsker, 2004; Reed et al., 2000); Minimizing the undetected concentrations (Cieniawski et al., 1995; Datta and Dhiman, 1996; Dhar and Datta, 2007) etc.

One of the earliest monitoring network design was performed by Loaiciga (1989) using Mixed Integer Programming (MIP). The objective of the MIP network involved the minimization of the variance of estimation error subject to resource and unbiasedness constraints. The network design is performed in two steps. The first is the parameter estimation which was followed by the network optimization. The performance of the design problem was successfully tested in the buried valley aquifer in Butler County, Ohio. Bagtzoglou et al. (1992) proposed a methodology for providing the necessary information required for designing a monitoring network. The network was designed with a minimum inter-sampling distance, thus making it cost-efficient.

Dhar and Datta (2009) applied linear mixed integer formulation for designing a global optimal groundwater quality monitoring network. The objective of the proposed model was to find an appropriate set of monitoring locations which could estimate the plume approximately under budgetary limitations. Further investigation of groundwater monitoring design was carried out by Prakash and Datta (2013). A monitoring network design was developed by combining spatial interpolation of the concentration measurement and simulated annealing as the optimization algorithm. The monitoring network design is then implemented sequentially to identify the actual source locations. The optimal monitoring network design further utilized the information of the concentration gradient from the previous iteration to obtain the objective function. This feedback information is used in determining the source location and found to be effective when no such information is initially available. This new methodology was found to be capable even with very limited observation data and could successfully identify the pollution sources.

Application of linked simulation-optimization could also be seen in Jiang et al. (2013) where an almost parameter free harmony search algorithm was employed in the optimization model. Satisfactory results with irregular geometries and noisy observations were recovered. The further studies on simulation-optimization approach motivated Datta et al. (2014) to develop a software package known as GWSID. The software tool provides applicability in identifying groundwater pollution sources by integrating numerical groundwater flow-transport code with simulated annealing based optimization model.

Prakash and Datta (2014) further applied linked simulation-optimization for solving groundwater source identification problem with unknown release time history. The study was based on the notion that the starting time activity of the sources remains

unknown to the problem. As an alternative, the source identification model is used to simultaneously estimate the source flux release history and source activity starting time as explicit decision variables. The evaluation of the results showed the potential applicability of the proposed model. It has the ability to correctly estimate the unknown pollution source flux release history and source activity starting times.

Borah and Bhattacharjya (2014) proposed three linked simulation-optimization based methodologies for identification of unknown groundwater pollution sources. Initially, the groundwater modeling system (GMS) linked with the optimization model and then the optimization is solved using direct search method. However, the methodology was found to be computationally inefficient for the large study area. To overcome it, the second approaches used an artificial neural network (ANN) for simulating the groundwater flow and transport processes of the aquifer. This model is computationally efficient, but predicting capability of the model has been reported as unproductive. Thus, they have presented the third approach. This is a hybrid approach which initially used the ANN based-model for obtaining the near optimal solution of the problem. The solution obtained was further used as the initial solution to the GMS-based model. Evaluated results showed that the hybrid model is found to be more accurate than the GMS-based approach.

Jha and Datta (2015) developed a methodology based on linked simulation-optimization model where the release histories were estimated using spatially distributed fixed pollution sources. The performance was evaluated using an illustrative abandoned mine site. Gurarsian and Karahan (2015) developed a linked simulation-optimization model to determine the numbers, location, source fluxes and release histories for the pollution source. Here, MODFLOW and MT3DMS software were used to simulate the groundwater flow and transport respectively. The performance of the model was tested on two hypothetical aquifer models using real and noisy observation data. In the first model, the release histories of the pollution source were determined assuming that the source numbers, location and active stress periods of the sources are known. Whereas in the second model, the release histories were determined considering that no information is available about the source. The results obtained were found to be better than those reported in the earlier literature.

Ayvaz (2016) presented a new simulation-optimization approach for solving areal groundwater pollution source identification problem. For simulating the groundwater flow and transport processes, MODFLOW and MT3DMS models are used respectively.

The simulation model was then integrated with the optimization model. The optimization model was solved using the binary coded genetic algorithm and also by using the gradient-based method. The results showed that the proposed model is an effective technique for solving a real groundwater pollution source identification problem.

An optimization algorithm is required for solving the inverse optimization problem and is selected depending on the type of problem. The optimization model can be solved either by using the gradient-based classical technique or the non-classical approaches. Earlier, many researchers adopted the classical approach for solving groundwater management problems (Willis, 1976; Willis and Newman, 1977; Gorelick, 1983; Elango and Rouve, 1980; Datta and Peralta, 1986; Wang and Ahlfeld, 1994; Datta et al., 2009; Datta et al., 2011). Though it was very popular among the researchers, numerous limitations have been encountered during the solution of the problem.

Some of the limitations of the classical optimization approach are (i) it is not always possible to evaluate the gradient of the objective function, (ii) the classical approach is highly sensitive to the starting point, (iii) it tends to give the local optimal locations, and (iv) it is time consuming as it follows a point to point search. Considering all these limitations, the researchers started opting for the non-classical techniques (Ritzel et al., 1994; McKinney and Lin, 1994; Cieniawski et al., 1995; Aral and Guan, 1996; Aral et al., 2001; Singh and Datta, 2007; Jha and Datta, 2011). Based on the type of the problem, researchers have adopted different non-classical optimization algorithms for solving the groundwater management and pollution sources identification problems.

2.4 Groundwater source identification using non-classical

Numerous types of optimization algorithms such as Genetic Algorithms (Holland, 1975), Simulated Annealing (Kirkpatrick et al., 1983) Particle Swarm Optimization (Eberhart and Kennedy, 1995), Ant Colony (Dorigo, 1992) etc. have been developed and has proved to be much simpler in solving the linked simulation-optimization problem. Some of these non-classical optimization algorithms are based on some certain rules, it may be biological or molecular phenomenon taking place in our environment. The very famous GA is based on the natural selection process comprising biological search operators such as selection, crossover, and mutation. The SA is basically inspired by heating and cooling of solid material when a critical temperature is introduced into the body. SA being a stochastic method has a high chance of

converging towards the global minimum. Over the years, numerous studies on the source identification problem have been conducted using the genetic algorithm (GA). The application of GA was first initiated by John H. Holland (1975) at the University of Michigan, inspired by the process of natural selection. At every iteration i.e. called the generation the search gets improved (Goldberg, 1989). Based on this basic idea, the application of GA could be effectively applied in various research fields such as groundwater source identification (Aral and Guan, 1996; Yeh et al., 2007; Aral et al., 2001; Mahinthakumar and Sayeed, 2005; Singh and Datta, 2006; Borah and Bhattacharjya, 2014; Ayvaz, 2015; Leichombam and Bhattacharjya, 2016 etc.) and other groundwater management model (Ritzel et al., 1994; McKinney and Lin, 1994; Cieniawski et al., 1995; Yeh et al., 2006; Aly and Peralta, 1999; Rao et al., 2007; Bhattacharjya and Datta, 2005, 2007, 2009; Chandalavada et al., 2011).

One of the major advantages of GA is that it does not require differentiability of the objective function (Bhattacharjya and Datta, 2005). And the constraint handling capacity of GA is more effective than the gradient search based classical optimization technique (Deb, 2001). With regard to the advantages of GA, Ritzel et al., (1994) used GA for solving two objectives steady-state groundwater pollution source problem. The objective of the optimization problem is to maximize the reliability and minimize the cost for designing the pumping system. They used vector-evaluated GA (VEGA) and Pareto GA algorithm for solving the multi-objective problems. Wang and Zheng (1998) integrated the groundwater flow model (MODFLOW) and transport model (MT3DMS) with GA for groundwater remediation. Pump and treat method was used for the groundwater remediation technique in where a set of extraction wells were used. The extraction wells were used for subsequently pumping out the contaminated groundwater.

More application of GA could be seen in the groundwater management studies performed by Mckinney and Lin (1994). Three example problems were solved. The first example being the maximum yield from a homogeneous isotropic unconfined aquifer system. The second problem consists of determining minimum cost combination of wells to supply an exogenous demand. And the third problem is to minimize the cost required for the pump and treat remediation design system for removing the source from an aquifer. Aral and Guan (1996) developed an improved GA based approach for identifying the source location, leakage rate and release period of the pollution source. The proposed model is based on the iterative evolutionary process.

The objective of the evolutionary process is the search for a better population which will yield a minimum function value. This iterative search continues until a minimum value of the objective function is achieved. Aral et al. (2001) formulated a non-linear linked simulation-optimization model. They could successfully solve the identification of groundwater source location and release history using a new approach known as Progressive Genetic Algorithm (PGA). A piecewise subdomain linearization is adopted in PGA which subsequently reduces the number of repeated groundwater simulation models and enhances the computational efficiency. Bhattacharjya and Datta (2005) used GA for groundwater management studies of coastal aquifer. GA being a heuristic search technique is used for solving the optimum management model and because of its relative efficiency GA is applied for solving non-linear convex problems.

Singh and Datta (2007) proposed a methodology using ANN model to study the missing data scenario for pollutant source identification problem. Evaluation of the results showed that the model is capable of extracting the relation between the pollutant sources and their corresponding contaminant concentrations. The large-scale complex problem such as pattern recognition, nonlinear modeling, classification, association, and control can be solved using ANN from given patterns (ASCE, Task Committee, 2000).

Yeh et al. (2006) developed a multivariate geostatistical groundwater quality network design using factorial kriging and GAs for identifying groundwater quality spatial variations. The developed model has been applied in a real aquifer in Taiwan.

Later, Guan et al. (2008) developed an improved GA termed as IGA for solving the optimization problems with equality and non-equality constraints. Here, a repairing procedure was embedded in the evolution process for handling the infeasible solutions which further will make the populations satisfy all the constraints. In the same year, Chandalavada and Datta (2008) developed a methodology for dynamic monitoring network design using Genetic Algorithm. They utilized two objective functions for designing an effective monitoring network. The first objective function was to minimize the summations of all positive deviations between the simulated contaminant concentrations and a specified low threshold. The second objective function was to minimize the estimated variances of pollutant concentrations at various unmonitored locations.

Jha and Datta (2012) adopted Adaptive Simulated Annealing (ASA) based linked simulation optimization approach for identifying source characteristics. The performance of the proposed methodology was evaluated in an illustrative study area. A

comparative study was carried out using GA and ASA and it was observed that the ASA based algorithm converges much faster towards the actual source flux than the GA based algorithm.

Putting some light on the groundwater remedial measures, a Pump and Treat (PAT) groundwater remediation technique was adopted by Mategaonkar and Eldho (2012). The MeshFree Point Collocation Method (MFree PCM) was used for simulating the groundwater processes. Particle Swarm Optimization (PSO) was used for solving the optimization model. The coupled PCM-PSO was applied for remediating an unconfined aquifer near Vadodara, India. Further, when the developed PCM-PSO was compared with other grid-based technique it was observed that the performance was more effective than the later ones. They finally concluded that PSO is simple and much more effective than other optimization technique such as GA.

With further discussion of groundwater remediation technique, it may be noted groundwater long-term monitoring (LTM) is considered an important remedial procedure. But monitoring a large number of sampling locations for years may not be cost-efficient. Considering this, an Ant Colony Optimization (ACO) algorithm is adopted to fewer down the sampling locations for a given number of monitoring wells (Li and Hilton, 2007). The ACO method is based on the principle that ant colonies are able to locate the shortest route from their home colonies to the site where food is available. This ACO-LTM algorithm has been applied to a site with a total number of 30 well LTM network. When the results were evaluated, a total number of 21-27 wells gave a globally optimal solution. This shows that LTM sampling location optimization problem was successfully solved using the effective ACO algorithm.

2.5 Hybrid-optimization approach

From the earlier discussions, it may be remarked that some of the evolutionary algorithms such GAs or SA are capable of solving large-scale problems and are considered a very powerful algorithm for global search. However, such algorithms need large number of iterations for the optimal convergence. On the other hand, gradient-based classical optimization algorithms can converge in smaller number of iterations but incapable of global convergence. So, depending on the perspective of the problem, different optimization algorithms can be merged together to form an efficient hybrid model. The applications of hybrid models have been seen in many groundwater management problems and other engineering optimization problems (Heidari and

Ranjithan, 1998; Pan and Wu, 1998; Shieh and Peralta, 2005; Mahinthakumar and Sayeed, 2005; Newman et al., 2005; Espinoza et al., 2005; Yeh et al., 2007; Ayvaz, 2016).

Newman et al. (2005) applied two numerical optimization techniques viz. Simulated Annealing (SA) and Minimum Relative Entropy (MRE) for solving source identification problem. A hybrid (SA-MRE) model was developed for estimating the source magnitude and the uncertainty involved with the estimated values of source flux. It was reported that the developed model proved to be an efficient technique for obtaining the mass flux probability functions, the magnitude of source fluxes and the confidence limits. Mahinthakumar and Sayeed (2005) also adopted hybrid optimization approach which coupled the Genetic Algorithm (GA) with a Local Search (LS) approach for determining the groundwater source characteristics. The non-uniqueness in solving the inverse problem was overcome by adopting the GA-LS method. The proposed methodology was applied in a Borden site for a field experiment research (Canadian Force Base). Again, Mahinthakumar and Sayeed (2006) investigated the performance of other hybrid optimization techniques. Combination of a heuristic approach (GAs) with a gradient-based approach (conjugate gradient) was found to be very effective in solving single or multiple source identification problems. It is also capable of solving three-dimension heterogeneous flow fields.

Yeh et al. (2007) presented a heuristic hybrid approach by combining Simulated Annealing (SA) and Tabu Search (TS) and the three-dimensional groundwater flow and transport model (MODFLOW-GWT) to determine the source location, release concentration and the release period. Initially, TS was used to generate the candidate source locations and then SA for generation of release concentration and the release period at the candidate source location. Here, MODFLOW-GWT was used for simulating the three-dimensional contaminant concentrations at the monitoring wells. Six studies on a homogeneous site, two studies on a heterogeneous site and a transient problem were performed using the proposed methodology. A conclusion was drawn that the proposed approach estimates the source for homogeneous and heterogeneous aquifers. He et al., (2009) presented an optimal design for a petroleum-contaminated aquifer using a coupled simulation-optimization approach. Their model is capable of addressing the stochasticity of modeling parameters in the flow and transport simulation models. It provides a direct and rapid link between remediation strategies and the

computational cost is reduced up to a considerable range. Bashi-Azghadi (2010) presented a new methodology for determining the unknown source location and the source flux using groundwater quality monitoring data. The GA based developed methodology consists of two simulation models namely Probabilistic Support Vector Machines (PSVMs) and Probabilistic Neural Networks (PNNs). The performance of the methodology is evaluated using a real case problem. Jiang et al. (2013) adopted a metaheuristic approach by linking a transport simulation model with a heuristic harmony search algorithm for identifying the groundwater pollution sources. An almost parameter free harmony search algorithm is developed and recovered good results even under different conditions like irregular geometries, erroneous observed concentration data, and insufficient potential locations.

More application of heuristic harmony in the identification of pollution sources is seen in the work of, Ayvaz and Elci (2013). They proposed a groundwater pumping cost minimization problem using linked simulation-optimization approach. The heuristic harmony search (HS) algorithm was integrated with spreadsheet solver. The objective of the model is to solve the pumping cost minimization problem for the different number of wells considering the pumping rates and the location of the additional new wells. The performance of the developed HS-solver model was evaluated using Tahtali watershed (Izmir-Turkey). Results showed that the model is capable of efficiently identifying the optimal numbers, locations and pumping rates of the pumping wells.

A hybrid model called SATSO-GWT was proposed by Yeh et al. (2014). It was developed for solving groundwater source identification problem comprising three locations and several irregular release periods and concentrations. The SATSO-GWT consists of groundwater source identification algorithm called SATS-GWT, an Ordinal Optimization Algorithm (OOA), and roulette wheel approach. The SAT-GWT algorithm is based on the principle of SA, TS, and MD2K-GWT which is a 3D groundwater flow and transport model. The evaluated results showed that the performance of the SATSO-GWT is much superior to that of the SATS-GWT. Ayvaz (2016) presented the simulation-optimization based hybrid model where a binary genetic algorithm and a gradient approach are used. The main objective of the proposed hybrid model is to identify the spatial distributions and contaminant concentrations of the groundwater pollution sources by utilizing a limited number of monitoring well locations. The applicability of the developed model was evaluated using a hypothetical area for different scenarios. After analyzing the results, it was suggested that the

developed methodology can effectively solve groundwater pollution source identification problem.

2.6 Summary and Conclusions

From the detailed review of the literature, it can be inferred that identifying the groundwater pollution sources requires an accurate characterization of the pollution sources. However, characterization is a challenging task as it includes the identification of the location of pollution sources, the magnitude of the source flux and the activity duration of the pollution sources. Response matrix, embedded technique, and linked simulation-optimization are some of the approaches for incorporating the simulation model with the optimization model. But researchers found that linked simulation-optimization approach is the ideal technique as it is capable of solving a highly non-linear and large aquifer problem. For simulating the complex groundwater processes, there are numerous powerful groundwater simulators which can be linked to the optimization model. However, in the linked simulation-optimization model, the simulation model is repetitively called by the optimization model until an optimal solution achieved. To enhance the computational efficiency, some of the researchers have used approximate groundwater simulator. However, incorporating large number of approximate groundwater simulators such as ANN will again increase the computational efficiency. For this reason, there is a need to develop a methodology which will identify the pollution source with only optimal numbers. This presents a case where the source locations are known to the problem. Though a source identification model should be a competent enough to solve a real-life scenario. Moreover, the information about the number and the location of the pollution sources are seldom available in reality. Therefore, it further motivates the researcher to develop a methodology which will try to identify the number and the location of the pollution sources.

The next matter of concern is how effectively the sources locations and the source fluxes can be identified. It has been presented by earlier researchers that identification of pollution sources is a mixed integer problem in which the source locations are the discrete variables whereas the source fluxes are the continuous variables. From the discussions of the literature review, it is observed that genetic algorithm (GA) is very efficient is handling discrete variable. Thus, GA can be used for determining the source locations. However, the source flux being a continuous variable can be determined

using gradient-based search. A methodology can be developed which utilizes GA for determining the source location. However, as GA does not use gradient information, it generally gives the near global solution, not the exact optimal solution of the problem. Thus, a gradient-based search can be employed to converge towards the actual source fluxes. Thus, the present study will attempt to develop methodologies which are capable of identifying groundwater pollution sources efficiently in terms of quality of the solution and its computational performance. The detailed explanation about the present methodologies is subsequently described in the chapters of this thesis.

Chapter 3

Identification of Groundwater Pollution Sources using ANN-GA Model

The present chapter explains the use of Artificial Neural Network (ANN) model as an approximate simulator for determining the groundwater pollution sources. The first section briefly describes the potential applicability of ANN in identifying groundwater pollution sources. This section also explains the enhancement of computational efficiency of ANN-based linked simulation-optimization model. The second part explains the methodology for linking ANN with GA for solving the source identification model. The subsequent sections describe the optimization algorithm used in the present study along with a brief detail about the study area adopted for the present study. The last section explains the results obtained by using the ANN-GA model.

3.1 Introduction

Discussions on the earlier chapters have explained how undesirable human activities have led to the contamination of groundwater and possess a serious threat to the fresh water sources. A major challenge is laid before the hydrogeologist to accurately identify the exact location of the pollution source and the magnitude of the source flux. The review of the literature described that different simulation-optimization techniques have been developed which can be used for identifying the groundwater pollution sources. For identifying the unknown pollution sources, the groundwater simulation models have to be incorporated with the optimization model. One of the most efficient approaches for linking the simulation model with the optimization model is the linked simulation-optimization technique. In this approach, the groundwater simulator is repeatedly called by the optimization model for providing the necessary data to the optimization model. As such, the effectiveness of this approach relies on the type of groundwater simulator used.

Though there are numerous numerical groundwater simulators that may be adopted, they are found to be computationally expensive when applied to large aquifer system. To cope up such limitation encountered in identifying groundwater pollution sources,

an approximate groundwater simulator may be adopted to simulate the complex groundwater flow and transport processes. This chapter addresses the potential applicability of one of the best approximate groundwater simulators, the artificial neural network (ANN) and explains how it can be used for determining the groundwater pollution sources.

In the present chapter, a linked simulation-optimization model using ANN as the approximate groundwater simulator is proposed for solving the groundwater source identification problem. The ANN model is then linked to the GA based optimization model.

3.2 Methodology

The unknown groundwater pollution source can be identified using the inverse optimization method. In this method, the difference between the simulated and the observed contaminant concentrations is minimized using an optimization algorithm. The actual concentration can be collected from field observations. On the other hand, the simulated concentration is obtained using the aquifer simulation model. For this reason, the aquifer simulation model has to be linked with the optimization model. In the linked simulation-optimization model, the simulation model is repeatedly called by the optimization model until an optimal solution is achieved. The objective function of the optimization model can be written as

$$\text{Minimize } F = \sum_i^M \sum_j^N (C_{O,i}^j - C_{s,i}^j)^2 \tag{3.1}$$

Where F is the objective function of the optimization model; $C_{o,i}^j$ is the observed concentration at j^{th} time step for i^{th} well location; $C_{s,i}^j$ is the simulated concentration at j^{th} time step for i^{th} well location; M is the total number of observation wells and; N is the total number of time steps.

Initially, the candidate solutions are generated using the optimization algorithm and serve as the input parameters for the groundwater simulation model. The simulation model will further generate the spatial-temporal contaminant concentration and measures the difference between the observed contaminant concentrations at the observation well location. The difference between the simulated and observed contaminant concentration is then used to evaluate the objective function value in the

optimization model. Based on the value of the objective function, the candidate solution is modified and the whole process is repeated until the optimal solution is achieved.

A schematic representation of the linked simulation-optimization model is shown in Fig. 3.1. There are numerous advanced groundwater simulators available for simulating the complex phenomenon of the aquifer processes. These models can be externally linked to the optimization model. Some of the standard groundwater simulators are SUTRA, MODFLOW, MT3DMS, SEAWAT, FEFLOW etc. These numerical models simulate the groundwater flow and transport equations by solving the governing equations for the specified aquifer system. They are solved using either finite difference methods or by finite element methods. In all the numerical methods, the mathematical model i.e. the groundwater governing equations have to be transformed into discrete form. The discrete variables are then calculated at the respective nodes (known as a grid) of the numerical domain. The procedure involved in calculating these discrete variables is highly complex and computationally very expensive specifically when the study area is very large. As such linking of these groundwater simulators will decrease the computational efficiency of the linked simulation-optimization. To overcome this limitation, ANN model is used as the approximate groundwater simulator. Numerous studies on groundwater management have been carried out using ANN (Coppola et al., 2003; Singh et al., 2004; Bhattacharjya and Datta, 2005; Bhattacharjya and Datta, 2007; Singh and Datta, 2007; Borah and Bhattacharjya, 2014).

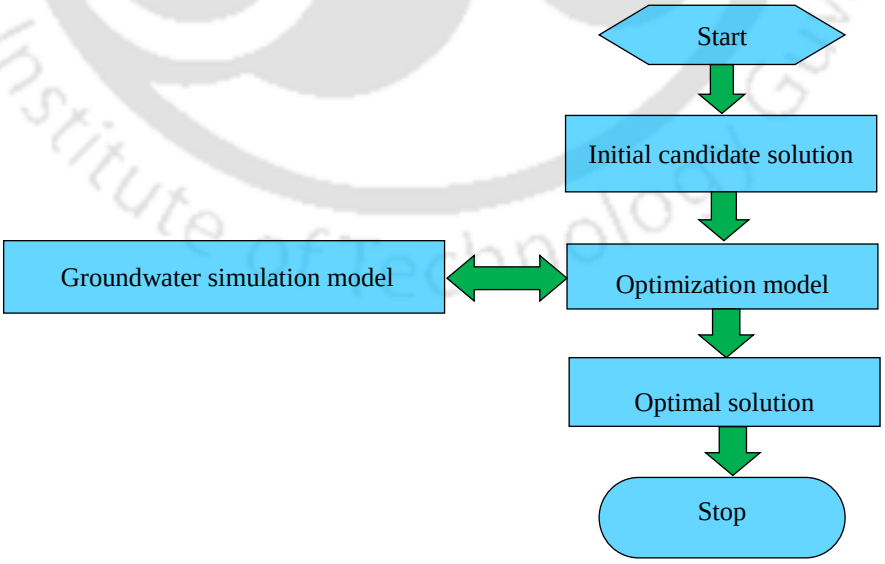


Fig. 3.1: Schematic representation of the linked simulation-optimization approach

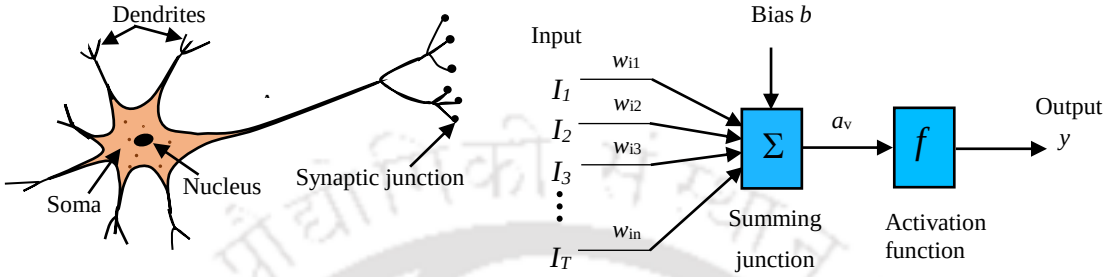
For large-scale aquifer system, the performance of the ANN-based surrogate model is not satisfactory when a single ANN model is used to predict the concentration at different observation well locations (Borah and Bhattacharjya, 2014). In such a situation, the computational efficiency of the ANN model can be enhanced by developing a separate ANN model for each of the observation locations. Thus, the number of ANN models is equal to the number of the observation wells used in the model. Adding more number of observation wells will increase the computational time of the model. At the same time, the optimal number of observation wells to be considered in the model are also not known. Considering all these aspects, this study presents a modified optimization formulation for identifying the groundwater pollution sources using an optimal number of observation well locations.

3.2.1 Artificial Neural Network

Artificial neuron model was first proposed by McCulloch and Pitts in 1943. The basic processing unit of an ANN model is the neuron. There is a close analogy between the biological neuron of the human brain and the artificial neuron. In the biological neuron, the signal comes through the nerve fibres called dendrites and the signal is passed through the axon. All the information is gathered in soma and cell activity is performed in it. The processed signals are transmitted from one neuron to another through the synaptic joint where the released transmitter will raise or lower the electric potential at the other end of the receiving neuron. As the biological neuron has information processing capability, the artificial neuron was also developed with the similar function of the biological one. Fig. 3.2 (a) and (b) show the biological neuron and the McCulloch-Pitts neuron model respectively. It can be noted that the artificial neuron has input analogous to the dendrites. Next, the function of the synaptic joint is performed by the weighs which control the effects of input signals during the entry to the neuron. The summing of the signals and the activation function is analogous to the processing of the information carried out in the soma.

For further discussion into how the different functions of this artificial neuron work let us assume that in the artificial neuron a total number of T inputs are introduced and the inputs being I_1, I_2, I_3, ... I_T. Each of these inputs is assigned individual weights denoted as w_{i1}, w_{i2}, w_{i3}, …w_{in}. A positive weight represents an excitatory connection whereas a negative value signifies that the connection is an inhibitory one. In the process of learning, the desire artificial neuron can be obtained by adjustments at synaptic weight

present following the pattern of the data set. In the next step of the artificial neuron, the weighted inputs are all summed up and biased. Finally, at the output, the summing function will pass through the activation function or transfer function which will control the amplitude of the output by applying any of the available functions such as threshold function, piecewise linear function, sigmoid or Gaussian function.



(a) (b)

Fig.3.2: (a) Biological neuron and (b) Artificial neuron

The relation for obtaining the output for a single McCulloch-Pitts neuron is given as:

$$y = f\left(\sum_{j=1}^{n} w_{ij}\, I_j + b\right) \tag{3.2}$$

Where,I_j is the input vector; w_{ij} is the weight; b is the bias; f is the activation function and y is the output.

The processing task for the artificial neuron is mainly performed by the transfer/activation function. The basic transfer function that may be applied is a step function, linear function and sigmoid function. The function may be adopted depending on the problem that the artificial neuron network will be solving. Details about transfer function have been described in the preceding section. The artificial neurons may be interconnected to form a complete network called artificial neural network (ANN). The capability of ANN to gather knowledge through learning from sufficient input pattern enables ANN to even apply to real-world problems (Bhattacharjya et al., 2009). Once the ANN model is trained according to the data pattern, subsequently it can be used to predict the output. The interconnection can be done in various ways which will result in the formation of different networks of different algorithms. No matter which method it is adopted, the main task of the network is to make the problem a cost efficient one with the weight and biases obtained through learning. Learning is a data-driven process and the input-output data pattern will help in approximating the desired system response. One of the most widely used neural networks is the feed-forward model. In feed-

forward network, signals will flow in one direction only through the connecting path. The name itself suggests that there will be no back-loops in this architecture i.e. the output will not affect the pathways.

In feed-forward network, there are three neuron layers comprising of first input layer which will receive the signal, the middle layer is known as the hidden layer and the last layer called the output layer will gives the desired output. After selecting the appropriate architecture, the network will be trained with the given data involving the input-output data sets. During the training process, the network will adjust the weight so that the error is minimized between the output and the input values. It may be noted that multiple neuron layers are interconnected as computational units in the feed-forward architecture. For this multi-layer feed-forward network, back-propagation learning algorithm is the commonly used technique for prediction, pattern recognition and non-linear function fitting. Whereas a single-layer feed-forward is the simplest form of a network where the input layer of the source will subsequently project the output. The input-output relation in the multi-layer network is given in equation (3.3).

$$O = t_{f2}[W_m.t_{f1}(w_1.I + b_1)+ b_2] \quad (3.3)$$

Where, I is the input vector, O is the output vector, w_1 is the weight matrix for the synaptic connections between input and hidden layer, W_m is the weight matrix for the synaptic connections between the hidden and the output layer, b_1 is the bias in the hidden layer, b_2 is the bias in the output layer, $t_{f1}$is the transfer function for the neurons in the hidden layer and t_{f2} is the transfer function for the neurons in the output layer.

3.2.2 Development of source identification model using ANN-GA model (Model 1)

Artificial neural network (ANN) model is one of the most effective as well as popular models used for replacing the numerical aquifer simulation model. In the present study, the ANN models are trained using the data generated by MODFLOW and MT3DMS models. The steps involved in developing the ANN model has been explained in the subsequent sections. The ANN model is linked with the optimization model for identifying the groundwater pollution sources. Then, the optimization model is solved using the genetic algorithm (GA). Hence, the proposed methodology for identifying the pollution source is named as ANN-GA model (Model 1). Fig. 3.3 explains the steps of the ANN-GA model. The ANN model is developed using the ANN toolbox available in MATLAB.

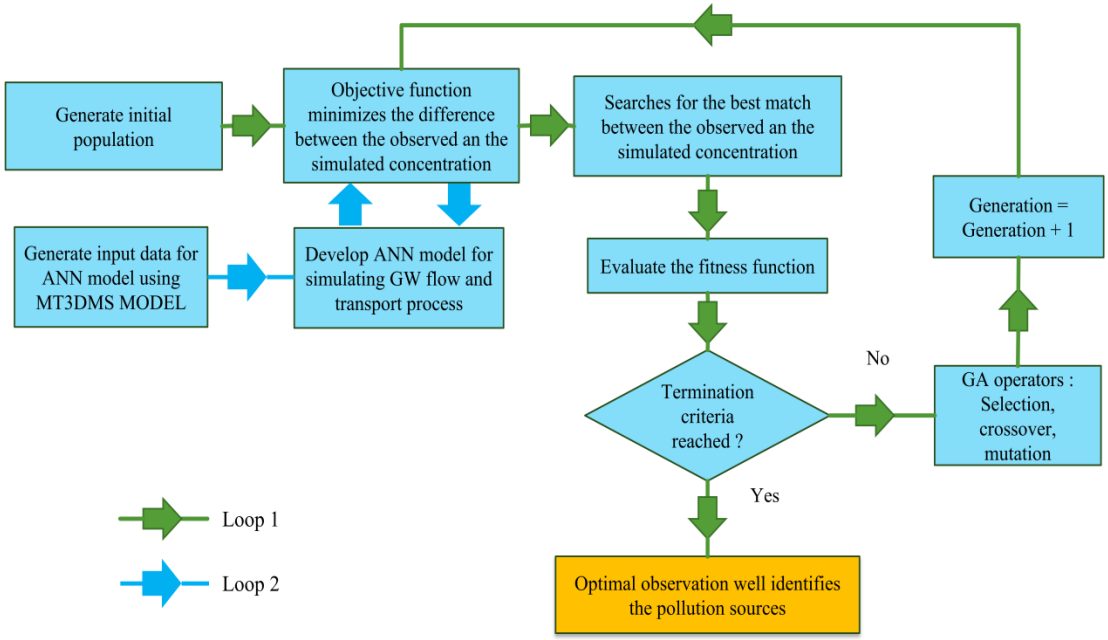


Fig. 3.3 Flowchart showing ANN-GA based linked simulation-optimization model

The first step involved in proposed methodology is the generation of an initial population which is generated randomly. These populations are sent to the objective function. The objective function calls the approximate groundwater simulator, the ANN model for supplying the simulated contaminant concentration at the observation well locations. There are two parts in the objective function, i.e. two non-conflicting objectives. The first objective minimizes the square of difference between the observed and the simulated concentration at the observation locations.

Whereas the second objective function allows the model to select those observation wells which has large concentration and effectively monitors throughout the stress period. It can be further added that the second objective function will not select those observation wells which are located very far from the pollutant sources as the contaminant concentration observed in those wells is negligible at different time steps. There are other factors also which will result in minimal concentrations; one is low strength of the contaminant concentration and other being the timing of the pollution activity. The combination of these two objectives will allow the model to identify the pollutant source effectively and will also allow selecting only those optimal well locations which will monitor large contaminant concentration. The objective function can be written as:

Minimize $$F = w_1 \sum_{i=1}^{P} \sum_{j=1}^{K} \left| C_{o,i}^{j} - C_{s,i}^{j} \right|^2 \frac{1}{\left(1000 + \sum_{i=1}^{P} \sum_{j=1}^{K} C_{o,i}^{j}\right)^2} z_i + w_2 \frac{D}{1 + \sum_{i=1}^{P} \sum_{j=1}^{K} C_{o,i}^{j}} \quad (3.4)$$

Subject to the constraints

$$\sum_{i=1}^{P} z_i \leq M_{max} \quad (3.5)$$

$$\sum_{i=1}^{P} z_i \geq M_{min} \quad (3.6)$$

Where, $C_{o,i}^{j}$ and $C_{s,i}^{j}$ represent the observed and simulated concentrations at i^{th} observation well location at time period. z_i is the binary decision variable which indicates whether an observation well will be selected or not at the location '*i*'. z_i equals to 1 if an observation well is installed or otherwise 0. M_{max} and M_{min} represent the maximum and minimum permissible wells that are to be installed for the given period. *P* and *K* are the total no of well locations and the time period that are considered, respectively. w_1 and w_2 are the weights assigned to the first and second objective functions. *D* is a very large constant value to be assigned by the user.

It is to be noted that the placing of a large number of observation wells far from the pollutant sources might not identify the pollutant source efficiently because the location of the wells being far from the source will become redundant. Hence, the optimal number of observation well locations around the pollutant sources will efficiently identify the pollutant sources. The optimization model is solved using genetic algorithms as it has a binary variable. The above objective searches for the best match between the observed and the simulated concentrations. In the subsequent step, termination criteria will be checked, if satisfies, the iteration will stop and the optimal solution is obtained. Otherwise, the population will further undergo different stages of GAs namely, selection, crossover and mutation for further refinement in the solution and the processes are repeated. The whole iterative process is known as generation and the process will continue until the termination criteria are achieved.

3.2.3 Groundwater simulation model

MODFLOW and MT3DMS models are used for simulating the groundwater flow and transport processes. The data pattern required for training the ANN model is generated using the MT3DMS model.

3.2.3.1 Groundwater flow equation

The simulation of groundwater flow and transport in an aquifer can be solved using the two partial differential equations. The groundwater flow equation for two-dimensional flow in a confined homogeneous aquifer can be written according to Bear (1972) as:

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_{zz}\frac{\partial h}{\partial z}\right)+W=S_s\frac{\partial h}{\partial t} \quad (3.7)$$

Where, K_{xx}, K_{yy} and K_{zz} are the hydraulic conductivity along the *x*, *y* and *z* directions (LT^{-1}); h is the hydraulic head (L); S_s is the specific storage coefficient; t is the time (T); W is the recharge flux per unit area (LT^{-1}).

3.2.3.2 Groundwater transport equation

The transient groundwater transport equation can be written as:

$$\frac{\partial(\theta C)}{\partial t}=\frac{\partial}{\partial x_i}\left(\theta D_{ij}\frac{\partial C}{\partial x_j}\right)-\frac{\partial(\theta v_i C)}{\partial x_j}+q_s C_s+\sum R_n \quad (3.8)$$

Where, C is the dissolved concentration in the groundwater (ML^{-3}); $\theta$is the porosity of the subsurface medium; t is the time(T); x_i is the distance along the respective Cartesian co-ordinate axis (L); D_{ij} is the hydrodynamic dispersion coefficient tensor (L^2T^{-1}); v_i is the seepage or linear pore water velocity (LT^{-1});q_s is the volumetric flow rate per unit volume of aquifer representing fluid sources (positive) and sinks (negative) (T^{-1}); C_s is the concentration of the source or sinks flux (ML^{-3}); $\sum R_n$ is the chemical reaction term ($ML^{-3}T^{-1}$).

3.2.4 Development of approximate groundwater simulator

In the proposed methodology, ANN model acts as the approximate simulator for groundwater flow and transport processes. The data required for training the ANN model is generated using MT3DMS model which is a groundwater transport package available in Groundwater Modelling System (GMS). The input to the ANN model is the number of pollution sources which are active for the specified stress period. The output from the ANN model is the concentration at the observation well location for different time steps. The study area has 30 observation wells. Out of these observation wells, an optimal number of observation wells are to be selected. The next section describes the details of the development of ANN model.

3.2.4.1 Generation of ANN pattern

For the present methodology, a single hidden layer architecture is adopted. There are no definite rules for selecting the number of hidden layers and the number of neurons in the architecture, it is generally determined on the basis of trial and error. The ANN model is trained using Levenberg-Marquardt (LM) algorithm. A unipolar sigmoidal transfer function and a purely linear transfer function are used for the neurons in the hidden layer and in the output layer of the network respectively. The input data to the ANN model are the groundwater pollutant source fluxes which are active for five-time steps at an interval of three months. The corresponding contaminant concentrations at the specified observation location constitute the output pattern for the ANN model. A total number of 2000 input-output patterns were generated using the MT3DMS model. The input and output files for the MODFLOW and MT3DMS model have been developed using GMS platform. From the total number of generated data, 60% is used for training the ANN model which subsequently will learn the input pattern for producing the desired outputs. The remaining 40% patterns are further used for testing and validating the ANN model.

3.2.4.2 Architecture of the ANN model for the study area

There are five pollutant sources in the study area and the simulation is performed for a period of five years at an interval of ninety days. As such a total number of 25 pollutant source flux is used as the input to the ANN model. The pollutant sources are active for five-time steps. The output from the ANN model is the concentration at the observation location for all the five years which will be equal to 20. The developed architecture of the ANN model can be represented as 25-40-20 which has been shown in Fig. 3.4.

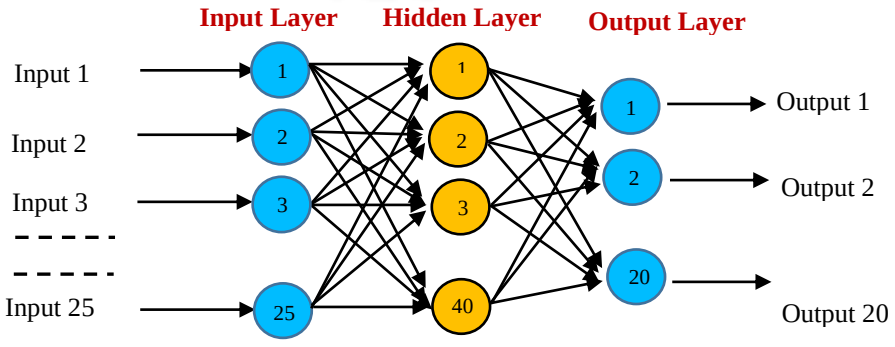


Fig. 3.4: Architecture of the developed ANN model

There are 30 observation wells in the aquifer. As such, 30 models have been developed. Each of these models is then applied for predicting the contaminant concentrations at the observation well locations.

3.3 Feed-Forward Back Propagation (FFBP)

Feed-forward back propagation (FFBP) is based on the principle of error correction learning rule and eventually updating the connection weights for reducing the error. The error correction rule in FFBP is applicable as it is a supervised learning where the network is presented with desired output values for each of the input pattern. The multi-layered neural network is trained using the back-propagation and depends on the connecting weight between the different layers of the network. The learning process of the feed-forward network using FFBP involves two processes as seen in Fig. 3.5.In the first forward process, the signals are introduced through the input layer and proceed towards the output layer passing through the hidden layers. The desired output and the actual output obtained using the network may differ up to some range. This difference may be regarded as the error term (desire output-actual output) obtained at the output layer and will be used in modifying the connection weight between the different layers.

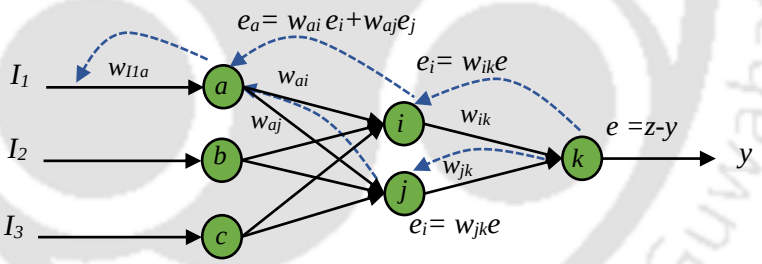


Fig. 3.5: Learning process of feed-forward back-propagation

The next step involves the propagation of the calculated error in the backward direction for each of the input samples. Once the error for each of the nodes is computed, the weights will be adjusted based on the following equation (ASCE Task committee, 2000).

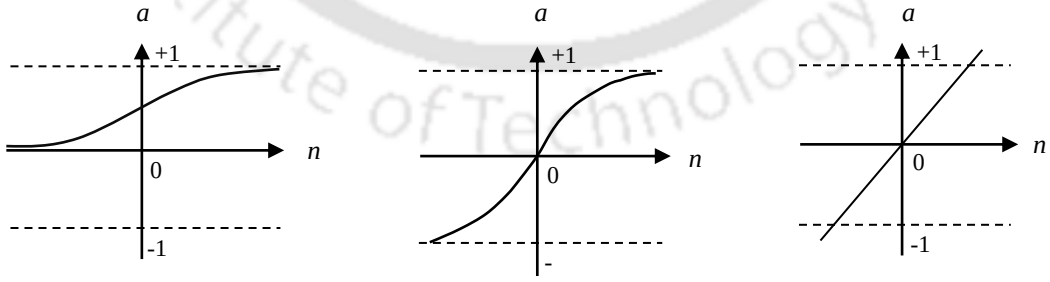
$$\Delta w_{I1a}(m) = -\beta * \frac{\partial e}{\partial w_{I1a}} + \varepsilon * \Delta w_{I1a}(m\text{-}1) \qquad (3.9)$$

Where, $\Delta w_{I1a}(m)$ and $\Delta w_{I1a}(m\text{-}1)$ are the increase in weight between node I_1 and a during the $m^{th}$and $(m\text{-}1)^{th}$ pass. β and ε are the learning and the momentum rates respectively. The momentum rate assists in speeding up the training rate and prevent the

oscillation of the weights whereas the learning rate prevents training process being trapped in the local minima rather than converging towards the global minima. In back-propagation learning, with a sufficient number of hidden layers, it can adopt any nonlinear function. Hence, this makes back-propagation learning a very good approach for training a neural network for modeling any complex system. The Levenberg-Marquardt (LM) algorithm is used for training the neural network. It is an iterative process which locates the minimum of a function that is expressed as the sum of squares of nonlinear functions.

3.3.1 Transfer Functions

A transfer function is a mathematical relation applied to the weighted input of a neuron to produce the output. The transfer function must be differentiable because it is required for computing the local gradient. The transfer function may be linear or non-linear function and some of the functions are the piecewise linear function, sigmoid or Gaussian functions (Wilby et al., 1998). Among all the functions, one of the most commonly used function is the Log-sigmoid function (*logsig)* as shown in Fig. 3.6 (a). It is also known as strictly increasing function. This function takes input ranging from plus infinity to minus infinity and generates output in the range of 0 to 1. The log-sigmoid function is frequently used in multilayer networks which are trained using the backpropagation algorithm. Alternatively, the multilayer may also use tan-sigmoid transfer function (*tansig)* which has the output in the range of -1 to +1 as observed in Fig. 3.6 (b). This function may be a very effective one when speed is considered an important criterion for the neural networks.



a = **Logsig(n)** = $\frac{1}{1+e^{-n}}$ a = **Tansig(n)** = $\frac{2}{1+e^{-2n}} - 1$ a = **Purelin (n)=n**

(a) (b) (c)

Fig. 3.6: Transfer function in feed-forward network (a) Log-Sigmoid (b) Tan-Sigmoid and (c) Purelin

Most of the discussions till now has emphasized on non-linear input/output characteristics. However, some models are there which only require minimal parameters and may be considered to possess linear characteristics. In such situation, *the purelin* transfer function is used (Fig. 3.6 (c)). This linear transfer function is capable of producing an output value. All the three-transfer functions described above are most commonly adopted transfer functions by the back propagation networks and other transfer functions may be used if preferred.

3.4 Transfer function and optimization algorithm adopted for the ANN model

There are numerous transfer functions such as *tansig, logsig, purelin* etc. that can be used in ANN. However, Borah and Bhattacharjya (2014) found that the combinations of *tansig* and *purelin* transfer functions for hidden layer and output layer gives the best performance in terms of mean square error (MSE), computational time and iteration time. Based on this report, the *tansig* and *purelin* transfer functions were used. *Tansig* gives the output in the range of -1 to +1 and is considered very effective one when speed is considered to be an important criterion for the neural networks as discussed in earlier section. In some cases, the models considered possess linear characteristics. For such situations, *purelin* transfer function is capable of producing any output value. The next important task is the selection of a suitable optimization algorithm to train the ANN network. For the present ANN model, the *trainlm* function available in MATLAB has been used to train the network. The *trainlm* function updates the weight and bias as per the principle Levenberg-Marquardt algorithm. The backpropagation algorithm is considered to be the fastest and is recommended as first choice for supervised algorithm (Sharma et al., 2008; Nooriet al., 2010; Gaur et al., 2013). Considering the computational efficiency of the transfer functions and the good convergence power of the *trainlm* function, the combinations has been used for training the network. Fig. 3.7 to Fig. 3.11 shows the performance of the ANN model using the above functions in terms of the mean squared error (MSE). It displays how MSE converges towards the best with each cycle of Epoch for training, testing and validating batches.

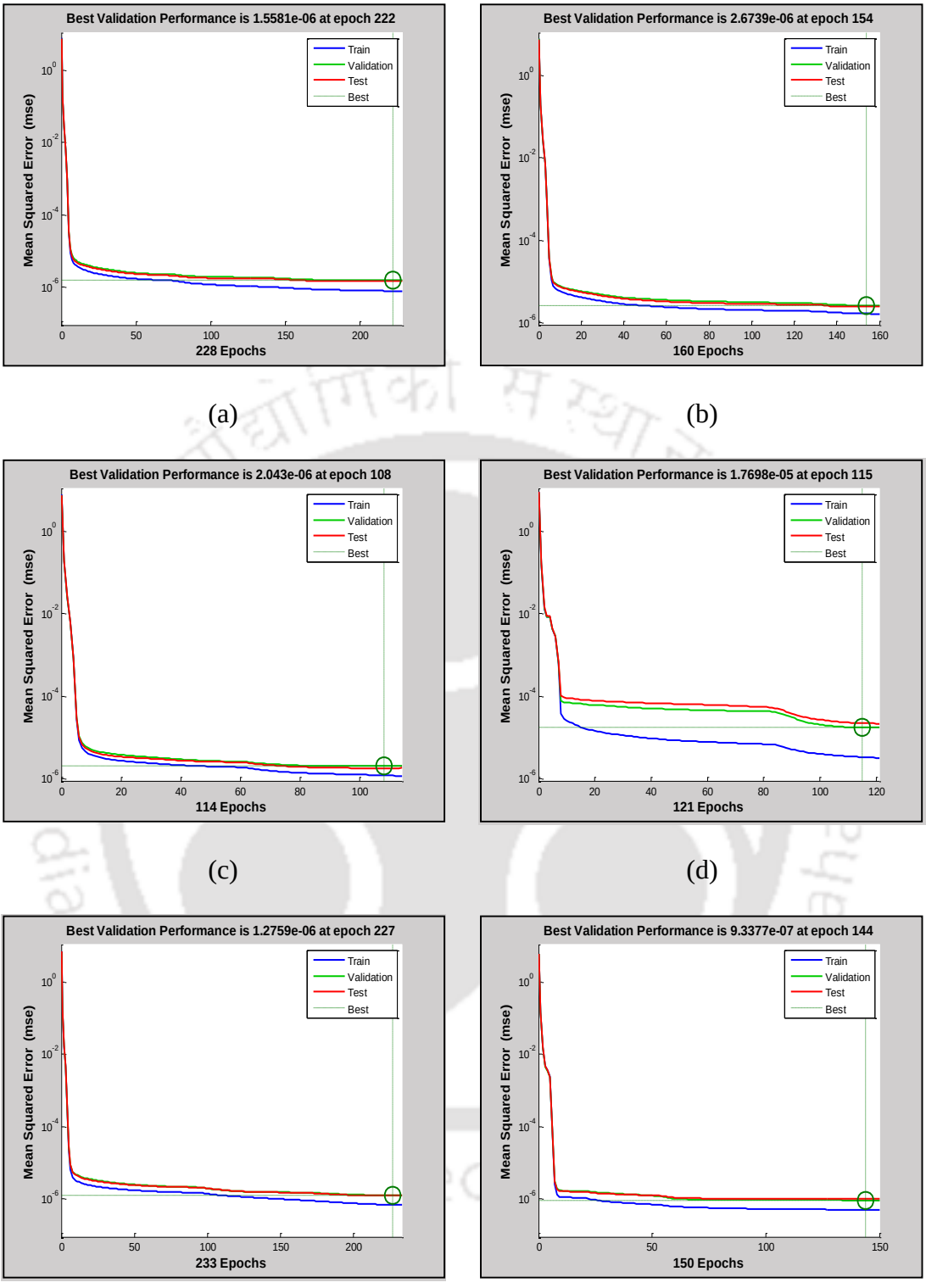


Fig. 3.7: MSE vs Epoch for (a) ANN model 1 (b) ANN model 2 (c) ANN model 3 and (d) ANN model 4 (e) ANN model 5 (f) ANN model 6

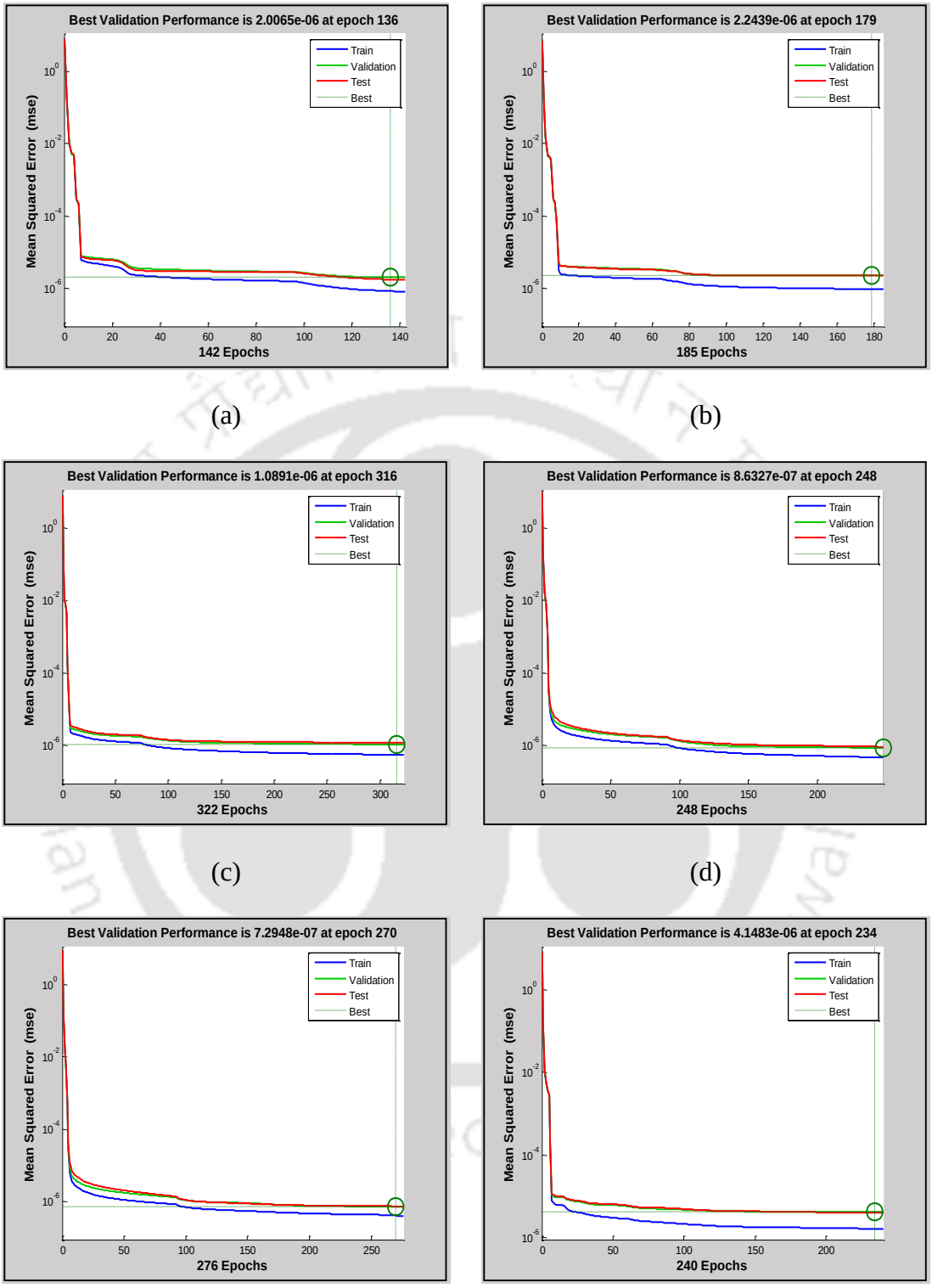


Fig. 3.8: MSE vs Epoch for (a) ANN model 7 (b) ANN model 8 (c) ANN model 9 (d) ANN model 10 (e) ANN model 11 and (f) ANN model 12

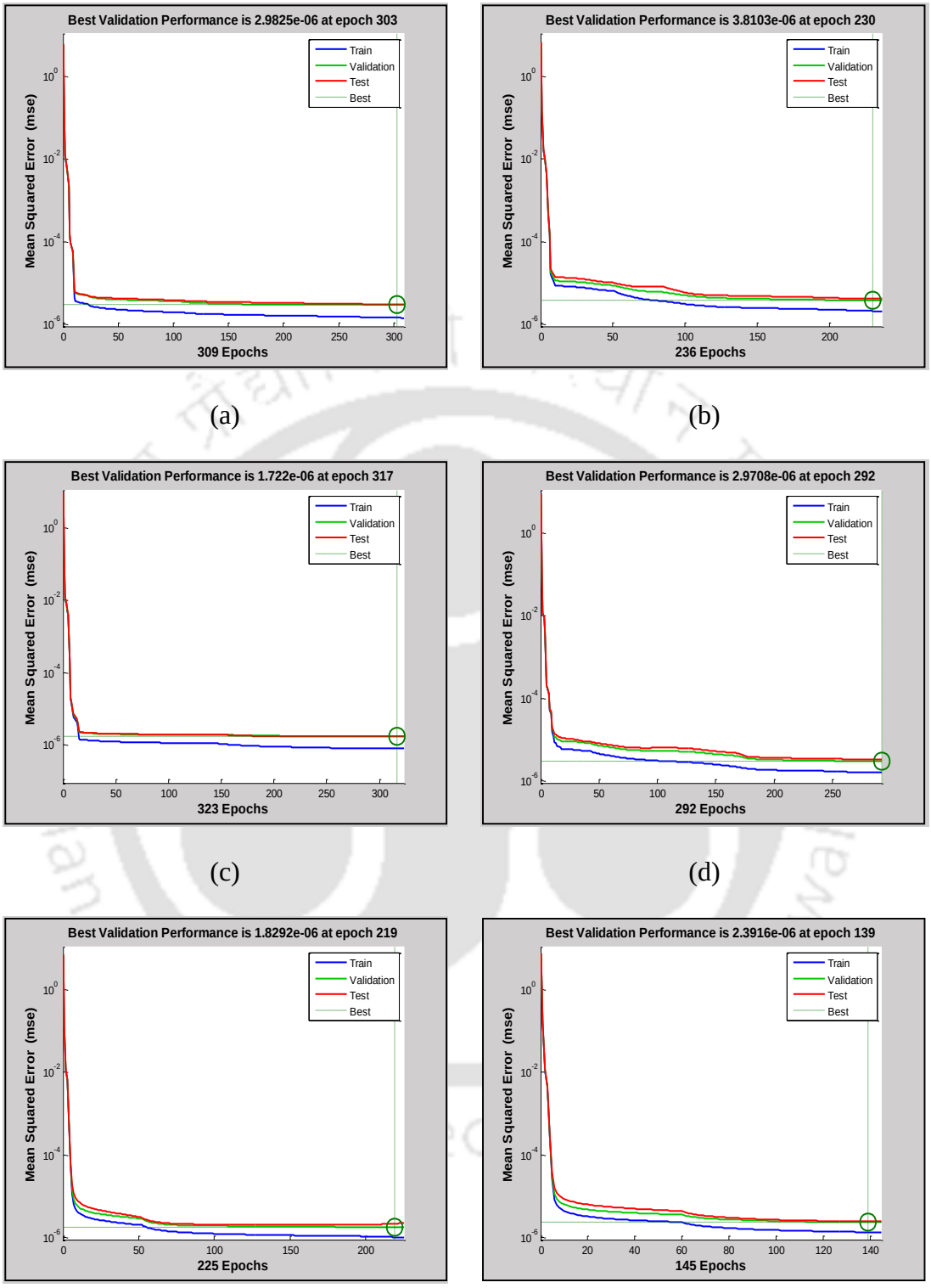


Fig.3.9: MSE vs Epoch for (a) ANN model 13 (b) ANN model 14 (c) ANN model 15 (d) ANN model 16 (e) ANN model 17 and (f) ANN model 18

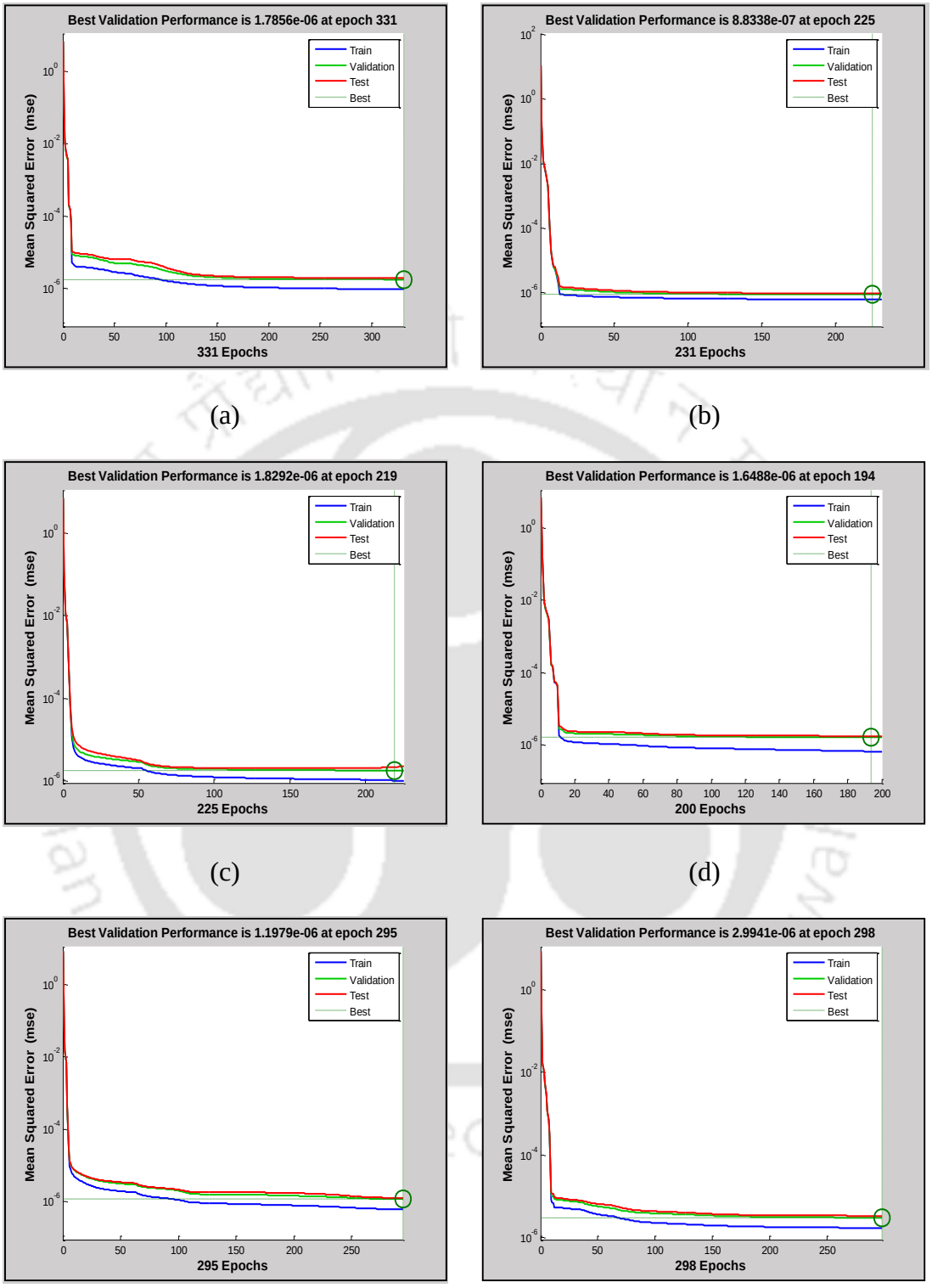


Fig. 3.10: MSE vs Epoch for (a) ANN model 19 (b) ANN model 20 (c) ANN model 21 (d) ANN model 22 (e) ANN model 23 and (f) ANN model 24

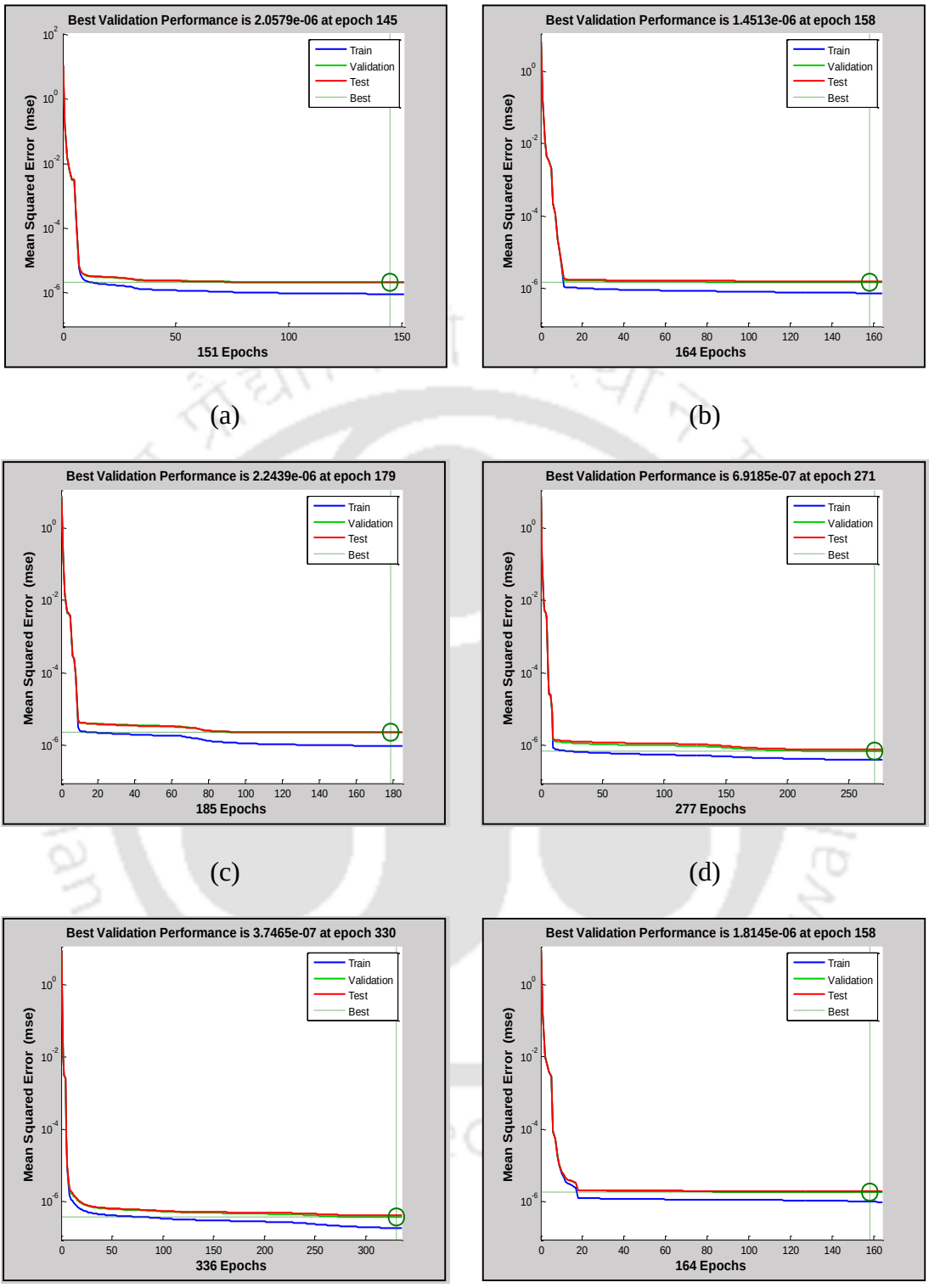


Fig. 3.11: MSE vs Epoch for (a) ANN model 25 (b) ANN model 26 (c) ANN model 27 (d) ANN model 28 (e) ANN model 29 and (f) ANN model 30

3.5 Optimization algorithm used in the study

In the present study, Genetic Algorithm (GA) is used to solve the inverse optimization model. The optimization toolbox available in MATLAB is used for implementing the GA.

3.5.1 Genetic Algorithms (GAs)

GA is a stochastic global search optimization algorithm which mimics the natural biological evolution. The GA was first proposed by John H. Holland (1975) at the University of Michigan. This algorithm is based on Darwin's theory of natural evolution which in turn follows the principle of "survival of the fittest". This is considered one of the most robust approaches as it is capable of converging towards a better solution with each generation. Considering the benefits of GAs, wide range of application has been carried in various groundwater related problems (Ritzel et al., 1994; McKinney and Lin, 1994; Cieniawski, 1995; Aral and Guan, 1996; Mahinthakumar and Sayeed, 2005; Yeh et al., 2006; Aly and Peralta, 1999; Bhattacharjya and Datta, 2005, 2007, 2009; Chandalavada, 2011; Borah and Bhattacahrjya, 2014; Ayvaz, 2015). The earlier studies carried out using GA has revealed that the performance of GA is found to be more efficient than the other traditional optimization algorithms when applied to the non-linear non-convex problem. Some of the primary difference of GA from the other classical techniques are: it is very effective for solving large complex non-linear model, GA does not require derivative information, it rather deals with the objective function value, one important factor about GA is that it searches with a set of points and not a single point and it uses probabilistic transition rules, not the deterministic ones. At each generation, a set of solutions, i.e. population are selected from the domain according to their fitness level. The selected solutions are improved subsequently to produce a better offspring following the rule of genetic operators. The genetic operators are selection, crossover and mutation as seen in Fig. 3.12.

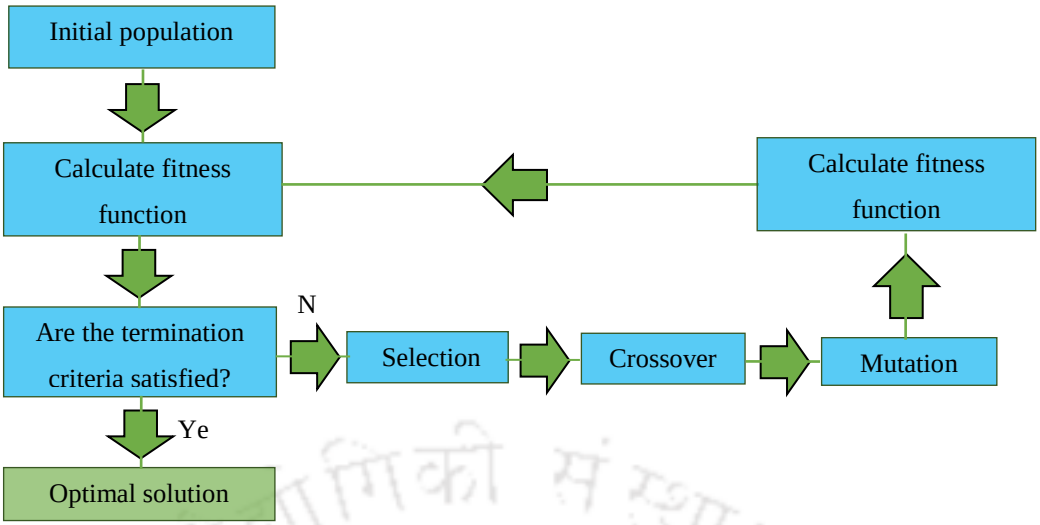


Fig. 3.12: Schematic representation of Genetic Algorithm

3.5.1.1 Representation

The representation of the chromosomes in GA are commonly denoted as binary strings. It is an important step as a proper representation makes the problem domain more suitable. Here, the set of the decision variables are encoded as a binary string. This set of decision variables or the solutions forms the population. The decision variable can be coded to binary string within the specified bounds. The performance of the individual members of the population is evaluated on the basis of the fitness function derived from the objective function. Thus, the objective function will depict the performance of the individual and its survival capability in the problem domain. As we have applied a string representation for the particular problem, the genetic operators can be applied to the strings for a better solution. The genetic operators as subsequently described below.

3.5.1.2 Reproduction (Selection)

This operator emphasizes on eliminating the bad string from the population according to the fitness value and forms an improved mating pool. As the good solutions having better fitness function are sorted out, there is a possibility of contributing better offspring in the next generation which is preferable for the problem. Any string in the population is selected with probability proportional to the string's fitness. The sum of all the string in the population is equal to one. Therefore, the probability for selecting the string is given by

$$P_s = \frac{f(x_i)}{\sum_{i=1}^{p} f(x_i)} \tag{3.10}$$

Here, $f(x_i)$is the fitness of the string and p is the population size. There are numerous ways of selecting the strings. Some of the methods are tournament selection, ranking selection, roulette wheel, proportionate selection etc.

3.5.1.3 Crossover

Crossover operator is applied to the mating pool. The individuals in the mating pool are used in generating the new offspring for the next generation. Therefore, it is desirable for the mating pool to have good individuals. Any two individuals (parent) are randomly selected from the mating pool and recombined by exchanging information among the strings to form better strings (children). The crossover can take place at a single point or multipoint. The simplest form of crossover is the single point and is explained in Fig. 3.13. In single point crossover, a random site is selected for both the strings and are cut at the points. Each of the strings exchanged among themselves at the cut points as seen in Fig. 3.13 With random sites, there is a chance that the children strings may or may not be a good one. It relies on whether the crossing site of the parent is an appropriate one or not. But this does not affect much to the crossover operator because if good strings are created by the crossover then more copies of good strings will be there for the next generation. Even if bad children strings are created by the crossover, then the bad copies will not survive in the subsequent generations as reproduction operator will not select the bad strings. In two-point crossover operator, two random sites are selected.

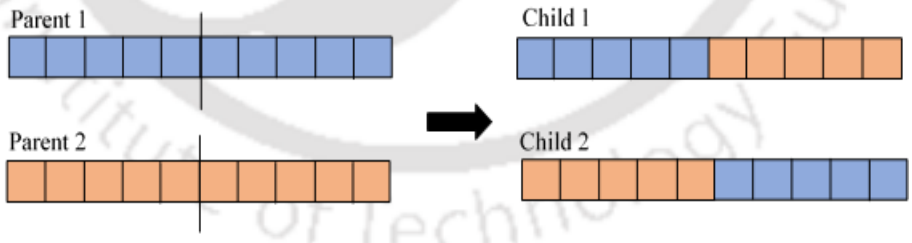


Fig. 3.13: Single point crossover

3.5.1.4 Mutation

After selection and crossover operators, some of the newly recovered population are directly copied producing a similar population. So, in order to ensure that genetic diversity is maintained among the population, a small change is performed on the strings. Mutation is generally performed at low scale, if carried out at the high scale it

may increase the search space and prevent the population to converge towards the optimal solution. However, it should be noted that very small mutation scale will lead to premature converges and gives a local optimal solution instead of a global optimal solution. In the binary coded string, mutation changes 1 to 0 and vice versa as seen in Fig. 3.14.

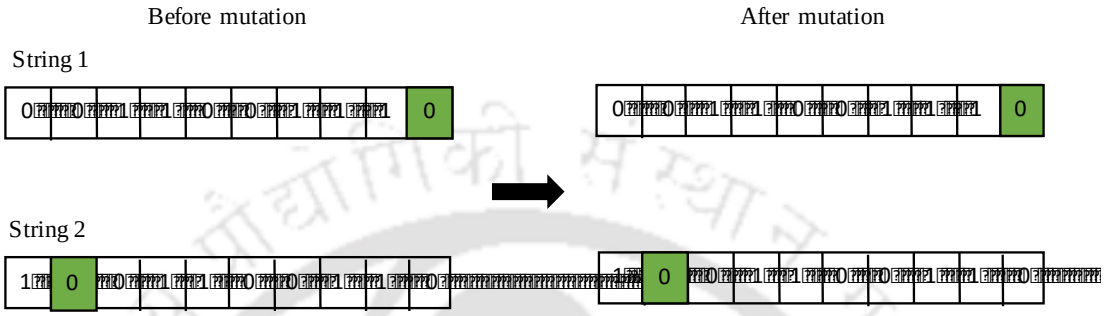


Fig. 3.14: Mutation carrying out at strings

3.6 Performance evaluation using statistical criteria

For evaluating the performance of the ANN model, three statistical criteria are adopted. They are Average Absolute Relative Error (AARE), Root Mean Square Error (RMSE) and Coefficient of Correlation (R). The mathematical expressions for the three statistical criteria are explained below.

3.6.1 Average Absolute Relative Error (AARE)

The AARE is the average value of the magnitude of the difference between the observed and the simulated contaminant concentration divided by the simulated concentration.

AARE can be calculated using the following expression as

$$\text{AARE} = \frac{1}{n_c}\sum_{i=1}^{n_c}\left|\frac{C_{o,i}^{j}-C_{s,i}^{j}}{C_{o,i}^{j}}\right| \text{X } 100 \qquad (3.11)$$

3.6.2 Root Mean Square Error (RMSE)

RMSE is defined as the square root of the average value of the squares of the difference between the observed and simulated concentration data. RMSE can be represented as

$$\text{RMSE} = \sqrt{\frac{1}{n_c}\sum_{i=1}^{n_c}(C_{o,i}^{j} - C_{s,i}^{j})^2} \tag{3.12}$$

3.6.3 Coefficient of Correlation (R)

The coefficient of correlation is the sum of the products of the deviation of each data from its respective mean, divided by the product of the number of set and the standard deviations. It can be expressed as

$$\text{R} = \frac{\sum_{1}^{n_c}\left(C_{o,i}^{j} - \overline{C_{o,\iota}^{J}}\right)\left(C_{s,i}^{j} - \overline{C_{s,\iota}^{J}}\right)}{\sqrt{\sum_{i=1}^{n_c}\left(C_{o,i}^{j} - \overline{C_{o,\iota}^{J}}\right)^2\left(C_{s,i}^{j} - \overline{C_{s,\iota}^{J}}\right)^2}} \tag{3.13}$$

Where, $C_{o,i}^{j}$ is the observed contaminant concentration at well location i for j^{th} time steps obtained using MT3DMS model;$C_{s,i}^{j}$ is the simulated contaminant concentration at well location i for j^{th}time steps recovered using ANN model; $C_{s,i}^{j}$ is the mean of the simulated concentration; $\overline{C_{o,\iota}^{J}}$ is the mean of the observed concentration; n_c is the total number of concentration data. Lower values of AARE, RMSE and R^2 signifies a good performance of the ANN model. If the value of R^2 is close to 1, then there is a close relationship between the observed and the simulated concentration.

3.7 Study Area

The hypothetical problem considered by (Borah and Bhattacharjya, 2014) is considered to evaluate the performance of the proposed model. The area is bounded by two rivers on the western and southern direction as shown in Fig. 3.15. The coverage area of the confined aquifer is approximately found to be 17.35 km^2. The presence of two rivers on the western and southern sides of the aquifer defines the boundary condition to be a constant head, whereas no flow boundaries exist on the remaining north-east directions. The hydrogeological parameters for the study area are given in Table 3.1. The area comprises of five pollutant sources, designated as S1, S2, S3, S4 and S5. The pollutant sources are active for five-time steps and the magnitude of source fluxes are shown in Table 3.2. There are four pumping wells located in the aquifer marked as P1, P2, P3, and P4. The pumping rates of the pumping well locations are shown in Table 3.3. For the present study area, a total number of thirty observation wells are considered (W1, W2…W30). The breakthrough curve for the thirty observation wells are shown in Fig. 3.16 to Fig. 3.20. The groundwater flow and transport processes are simulated for a

period of 5 years at a time step of three months. The genetic algorithm parameters used in in the present methodology is presented in Table 3.4.

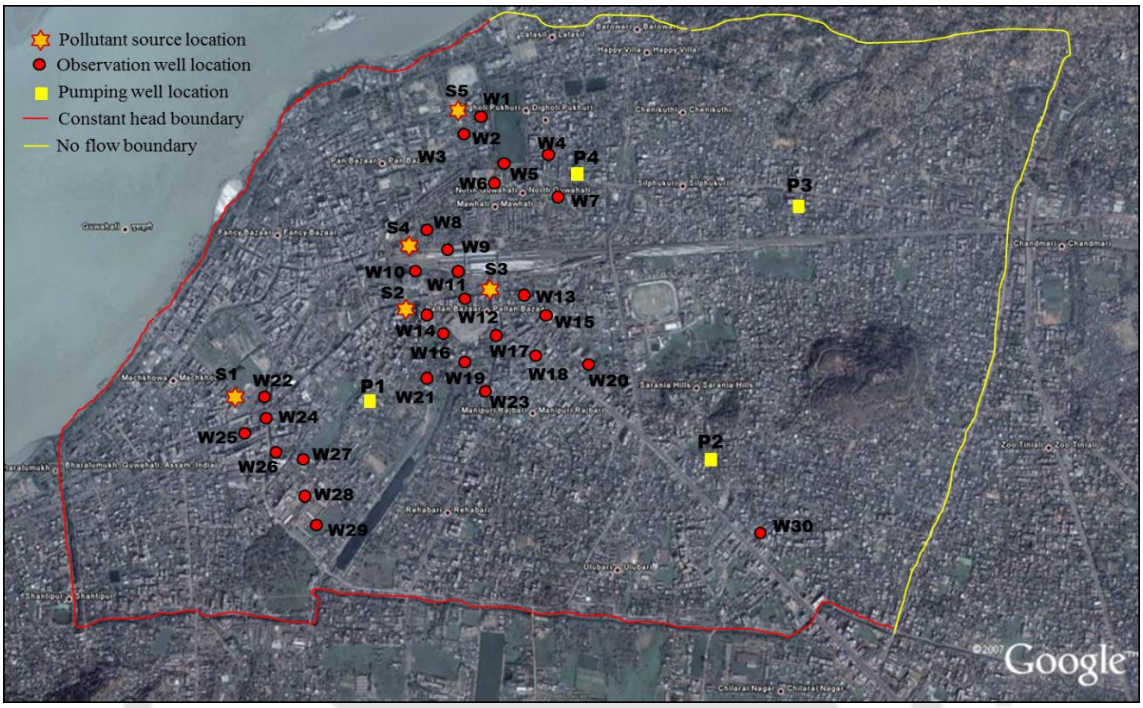


Fig. 3.15: Map of the study area showing pollutant sources, observation well locations and pumping well

Table 3.1: Hydrological parameters used in the study area

Parameters	Values
Hydraulic conductivity in x-direction, $\boldsymbol{K_{xx}}$ $(\mathbf{m/s})$	0.0002
Hydraulic conductivity in y-direction, $\boldsymbol{K_{yy}}$ $(\mathbf{m/s})$	0.0002
Effective porosity, $\boldsymbol{\varepsilon}$	0.25
Time Steps, $\Delta \boldsymbol{t}$ (months)	3
Longitudinal dispersivity, $\boldsymbol{\alpha_L}$ (m)	40
Transverse dispersivity, $\boldsymbol{\alpha_T}$ (m)	9.6

Table 3.2: Source fluxes for different time steps (g/s)

Sources	Time Step 1	Time Step 2	Time Step 3	Time Step 4	Time Step 5
S1	908.42	1130.50	653.35	902.13	721.25
S2	644.02	1023.87	1139.88	781.09	889.77
S3	0	0	0	0	0
S4	0	1024.16	652.05	1117.45	889.77
S5	987.08	0	0	1104.82	639.93

Table 3.3: Pumping rates of the wells at the pumping well location of the aquifer

Time Step	Pumping Rates (m^3/day)	Time Step	Pumping Rates (m^3/day)
1	327.024	11	272.52
2	163.512	12	218.016
3	218.016	13	327.024
4	318.528	14	163.512
5	109.008	15	381.528
6	327.024	16	217.72
7	272.520	17	272.520
8	163.512	18	218.010
9	381.528	19	327.024
10	109.008	20	272.520

Table 3.4: Genetic Algorithm parameters used in the present methodology

Parameter	Adopted value	Function Parameter	Adopted
Population size	200	Scaling function	Rank
Generations	2000	Selection function	Stochastic uniform
Crossover fraction	0.8	Mutation function	Constraint dependent
Elite count	0.5	Crossover function	Scattered

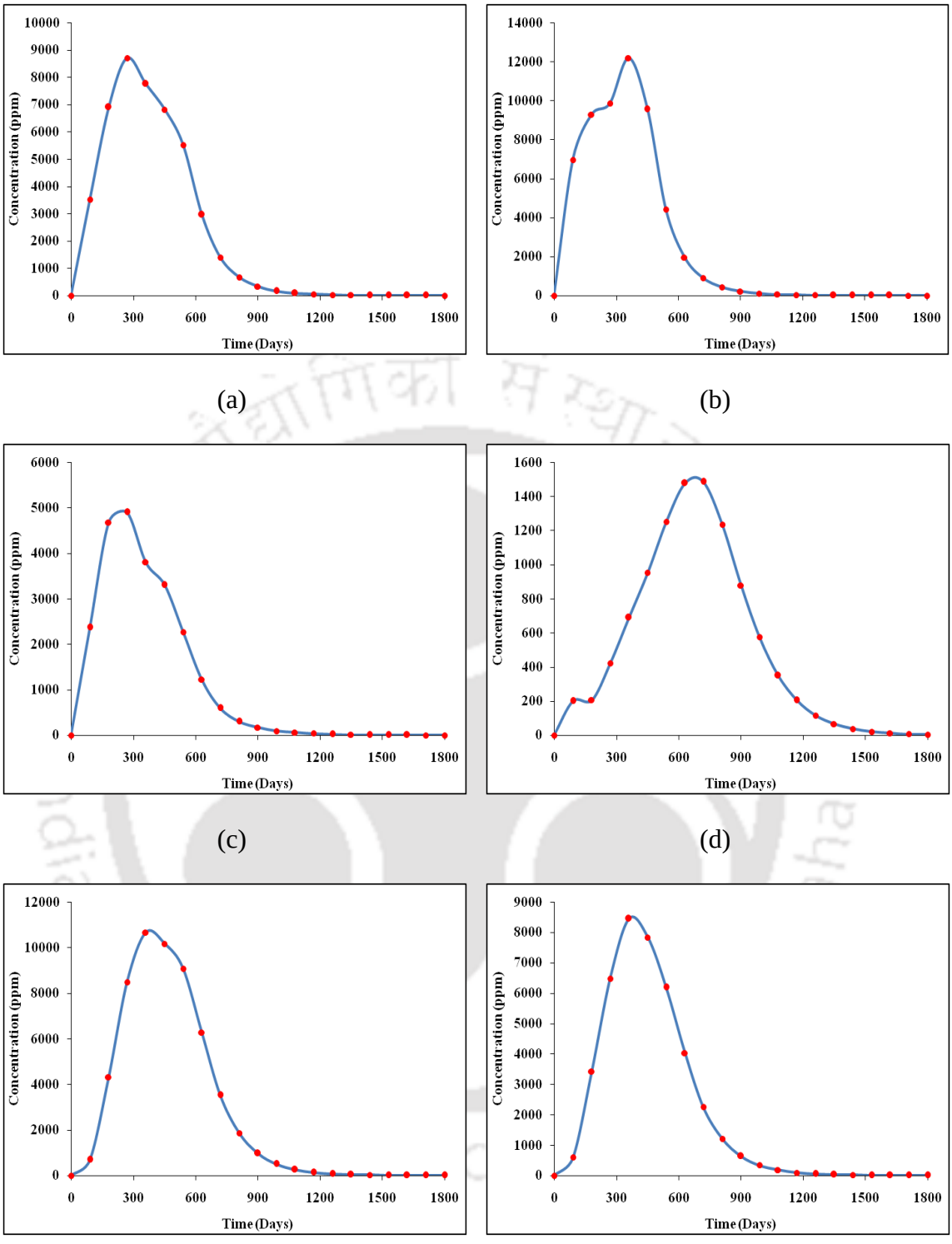


Fig. 3.16: Breakthrough curve for (a) Well 1 (b) Well 2 (c) Well 3 (d) Well 4 (e) Well 5 and (f) Well 6

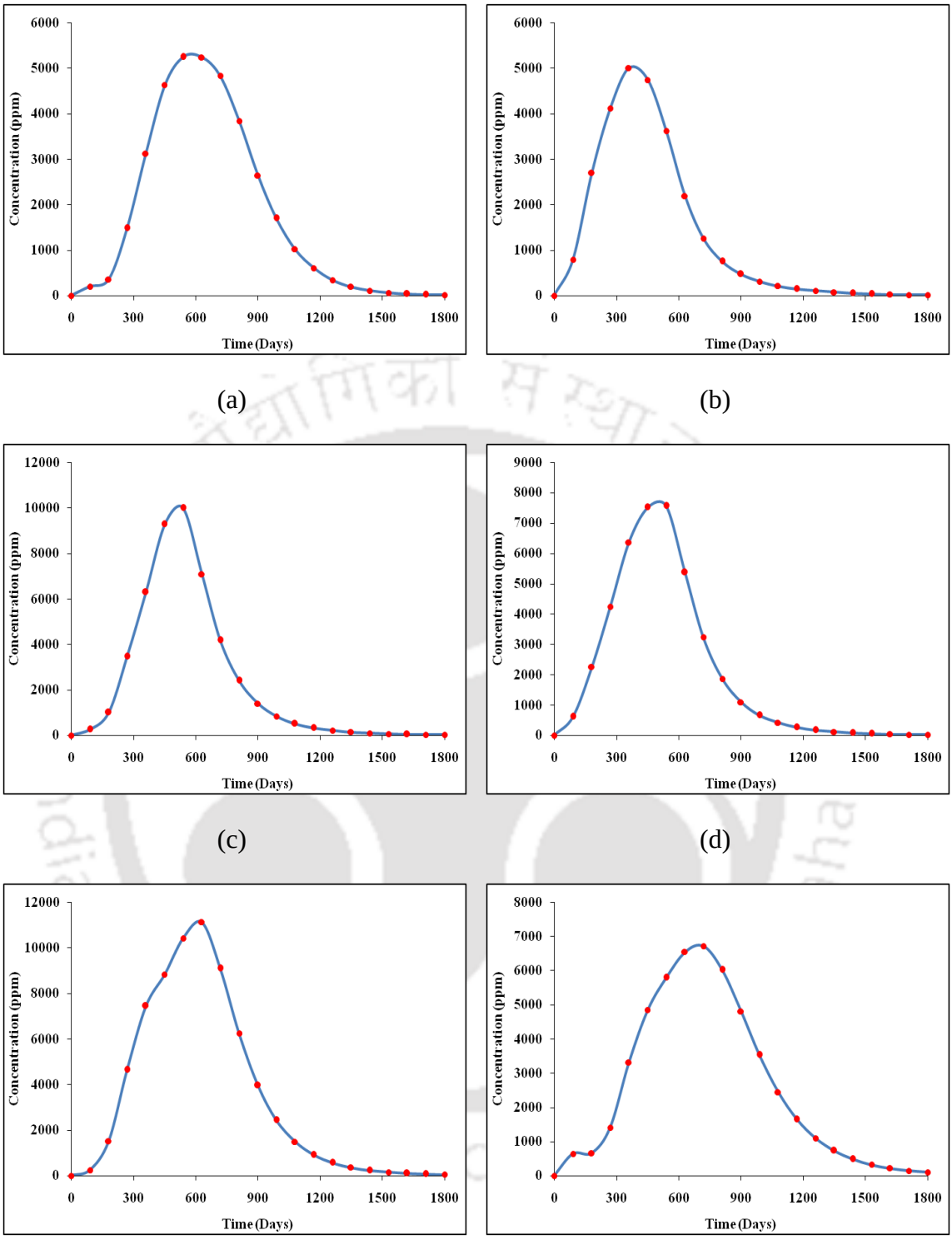


Fig. 3.17: Breakthrough curve for (a) Well 7 (b) Well 8 (c) Well 9 (d) Well 10 (e) Well 11 and (f) Well 12

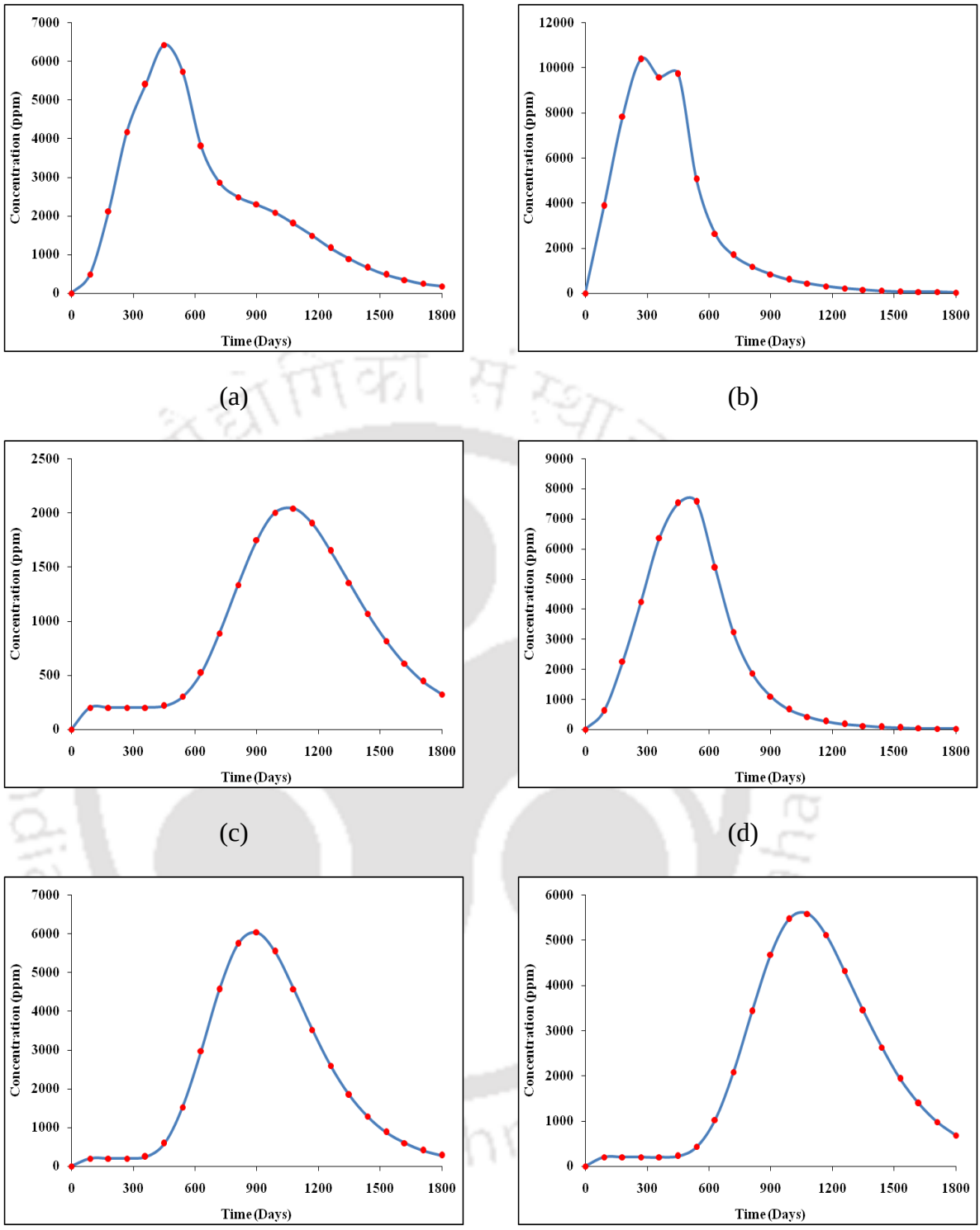


Fig. 3.18: Breakthrough curve for (a) Well 13 (b) Well 14 (c) Well 15 (d) Well 16 (e) Well 17 and (f) Well 18

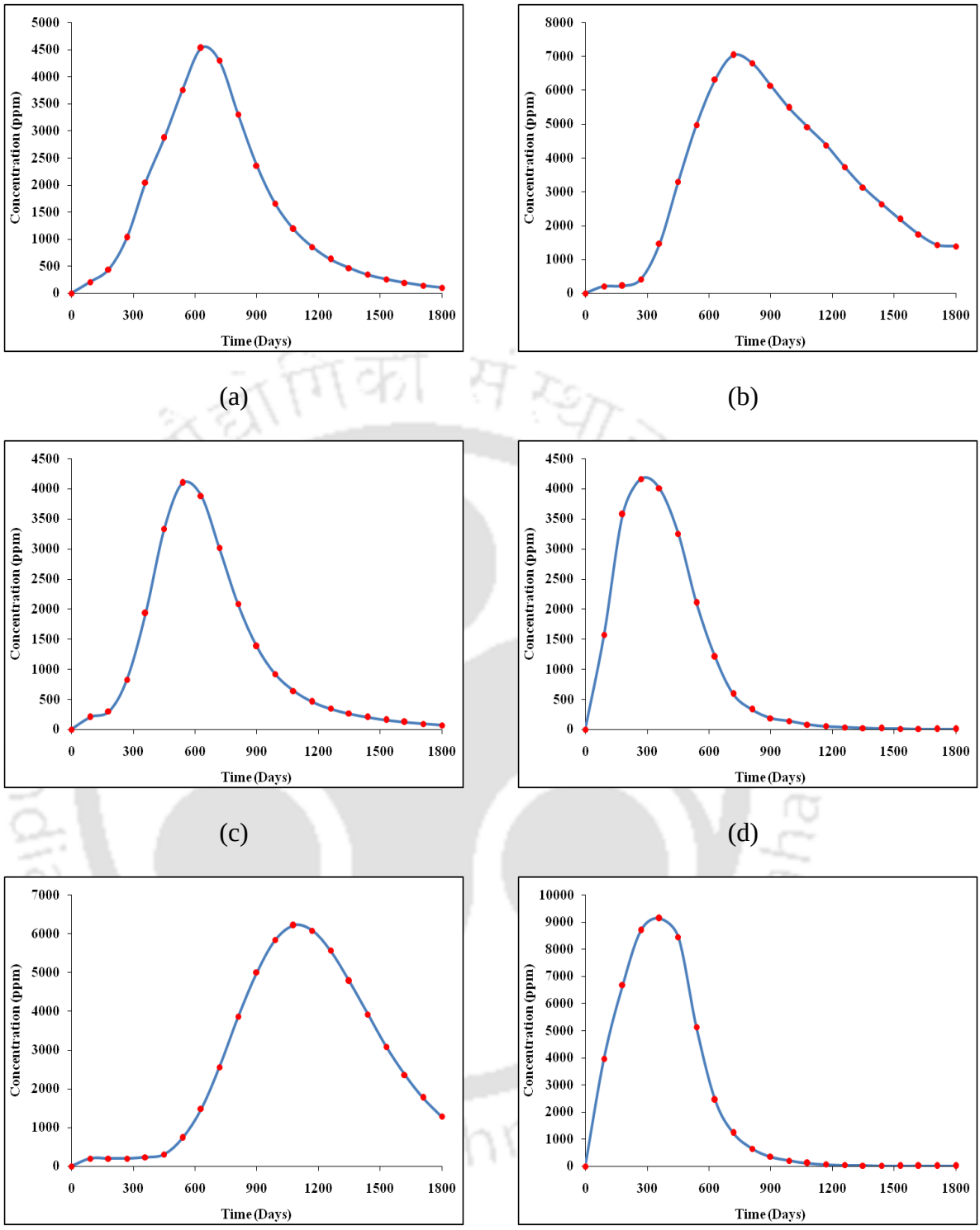


Fig. 3.19: Breakthrough curve for (a) Well 19 (b) Well 20 (c) Well 21 (d) Well 22 (e) Well 23 and (f) Well 24

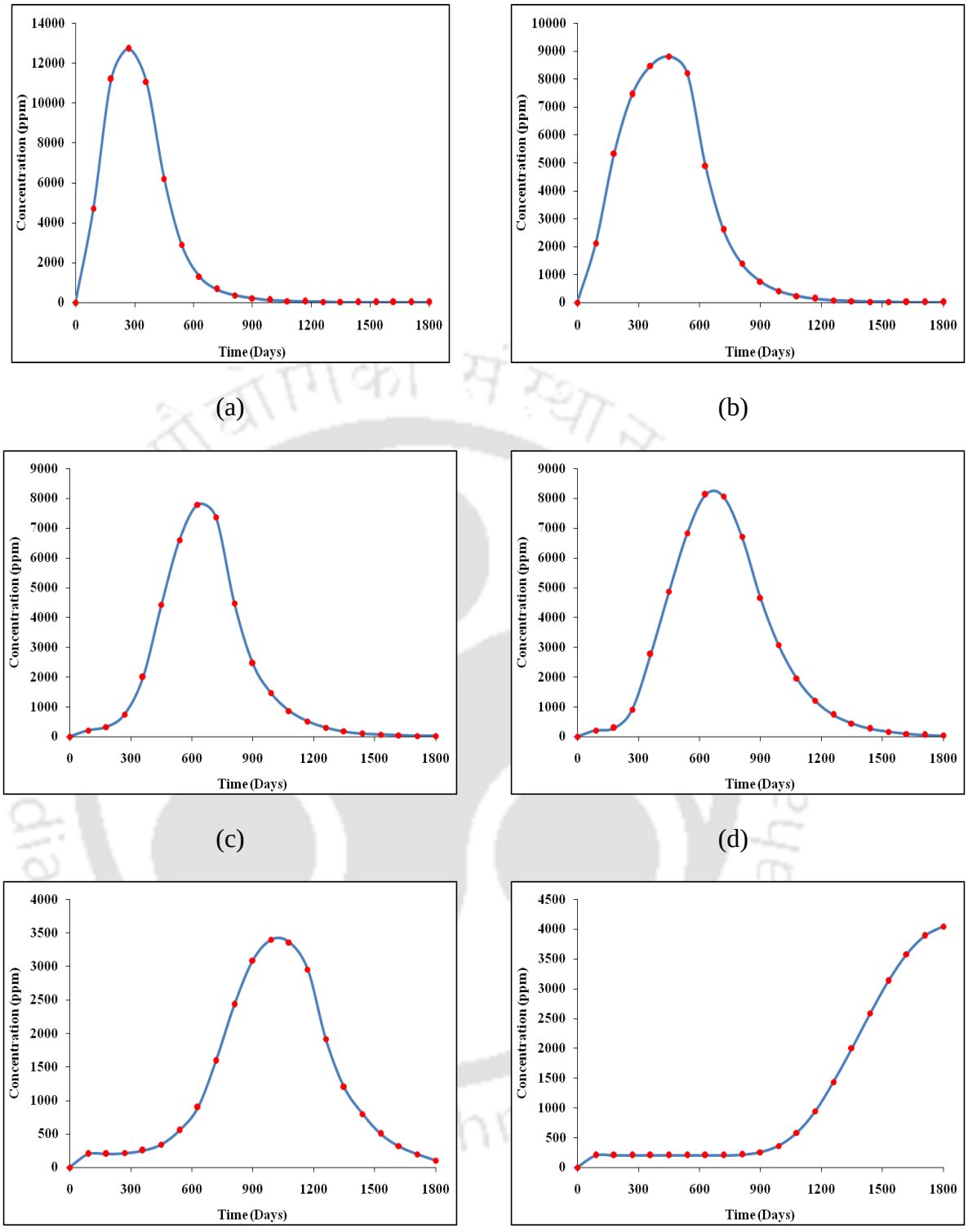


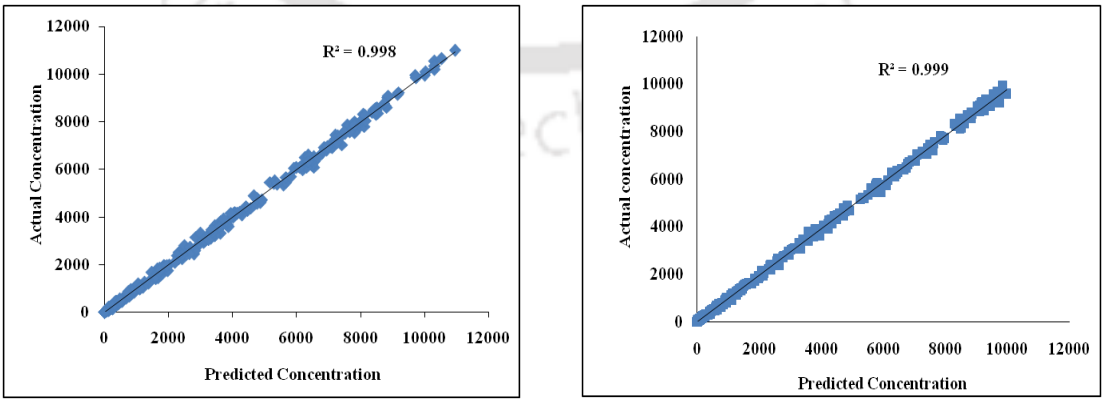
Fig. 3.20: Breakthrough curve for (a) Well 25 (b) Well 26 (c) Well 27 (d) Well 28 (e) Well 29 and (f) Well 30

3.8 Results and Discussion

The performance of the developed ANN model as the approximate groundwater simulator using the thirty well locations have been analyzed below. Furthermore, the proposed source identification model is validated using the hypothetical study area and the performance is discussed accordingly.

3.8.1 Performance of the ANN model

Fig. 3.21 to Fig. 3.26 shows the scatter plot between the actual and predicted normalized contaminant concentration for all the thirty observation wells. The scatter plots are plotted for five-time steps at an interval of 90 days at the respective observation wells placed at random locations on the adopted study area. The actual concentrations are acquired using the groundwater transport model MT3DMS whereas the predicted concentrations are obtained using the ANN model. It is observed that the best and the worst values of coefficient of correlation (R^2) among all the observation well are equal to 0.99991 and 0.99678 respectively. As the R^2values are very close to 1, it suggests that there is high correlation between the actual and the predicted contaminant concentrations. This clearly suggests that the developed ANN model has the competence to act as an approximate simulator for groundwater transport processes. The further performance of the ANN model for training the data using the two statistical criteria viz. average absolute relative error (AARE) and root mean square error (RMSE) for all the 30 developed ANN models are represented in Table 3.5**.** The calculated error terms are found to be very negligible and suggest that the performance of the ANN models is acceptable.



(a) (b)

Fig. 3.21: Scatter plot for (a) ANN model 1 (b) ANN model 2

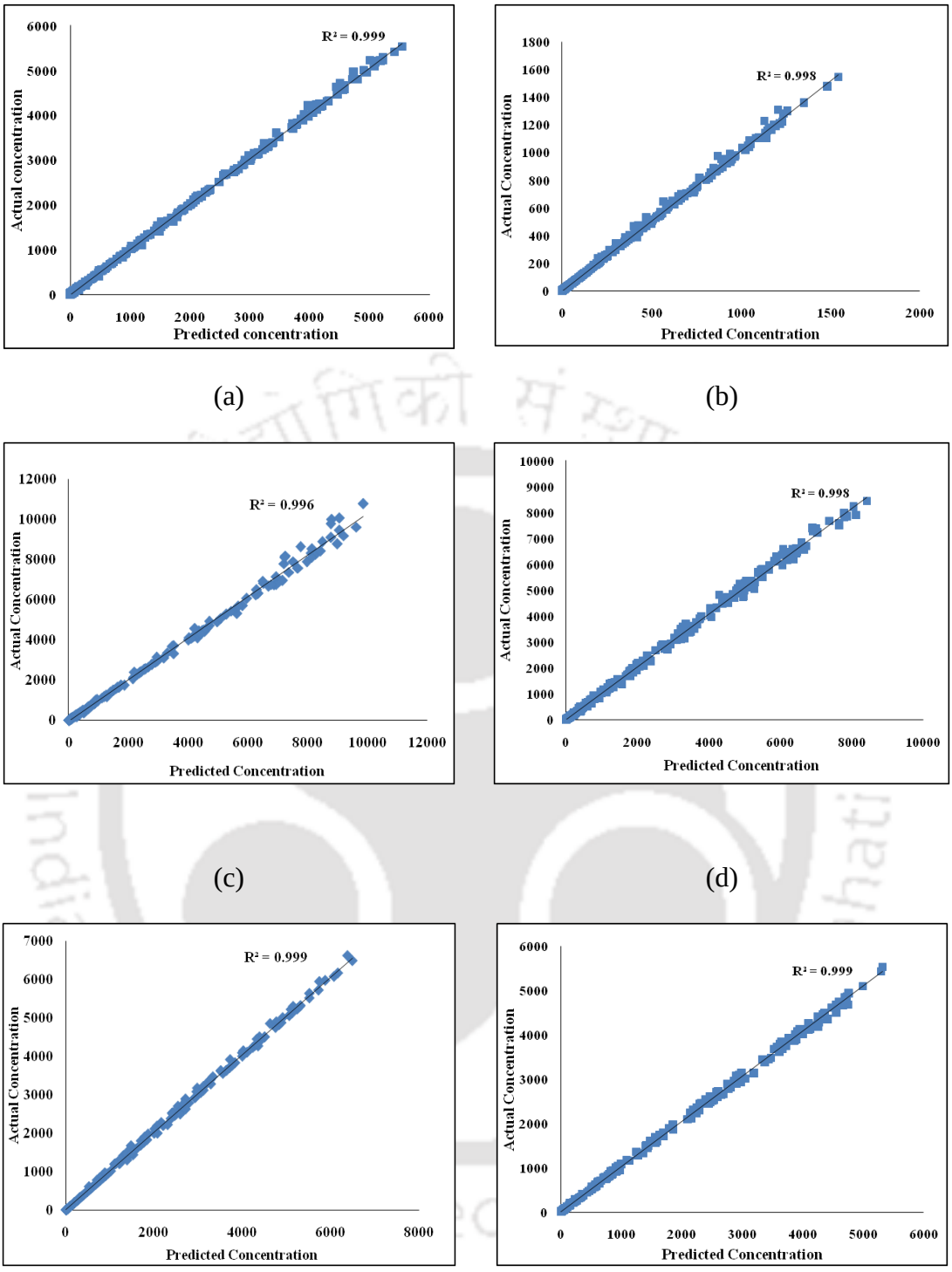


Fig. 3.22: Scatter plot for (a) ANN model 3 (b) ANN model 4 (c) ANN model 5 (d) ANN model 6 (e) ANN model 7 and (f) ANN model 8

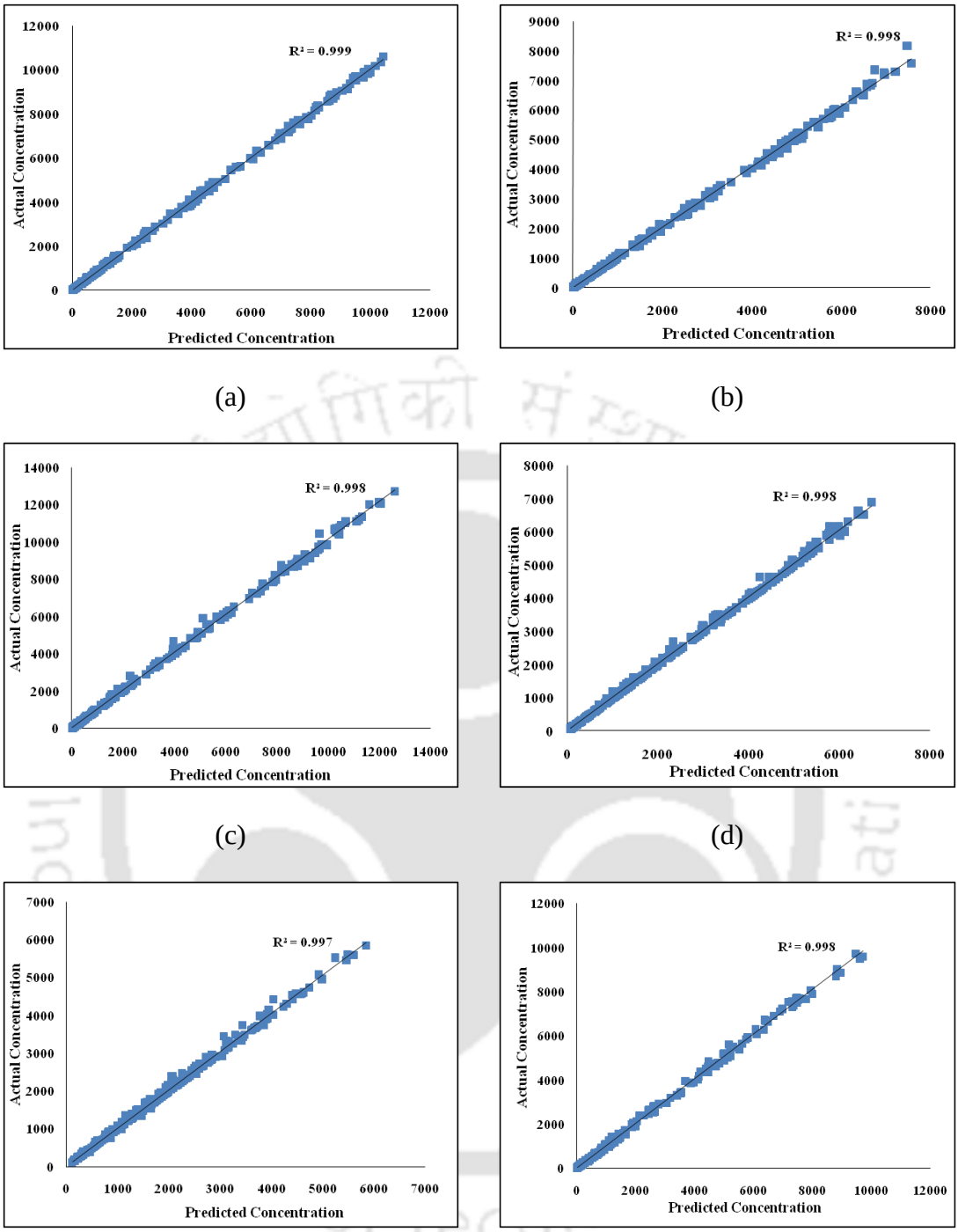


Fig. 3.23: Scatter plot for (a) ANN model 9 (b) ANN model 10 (c) ANN model 11 (d) ANN model 12 (e) ANN model 13 and (f) ANN model 14

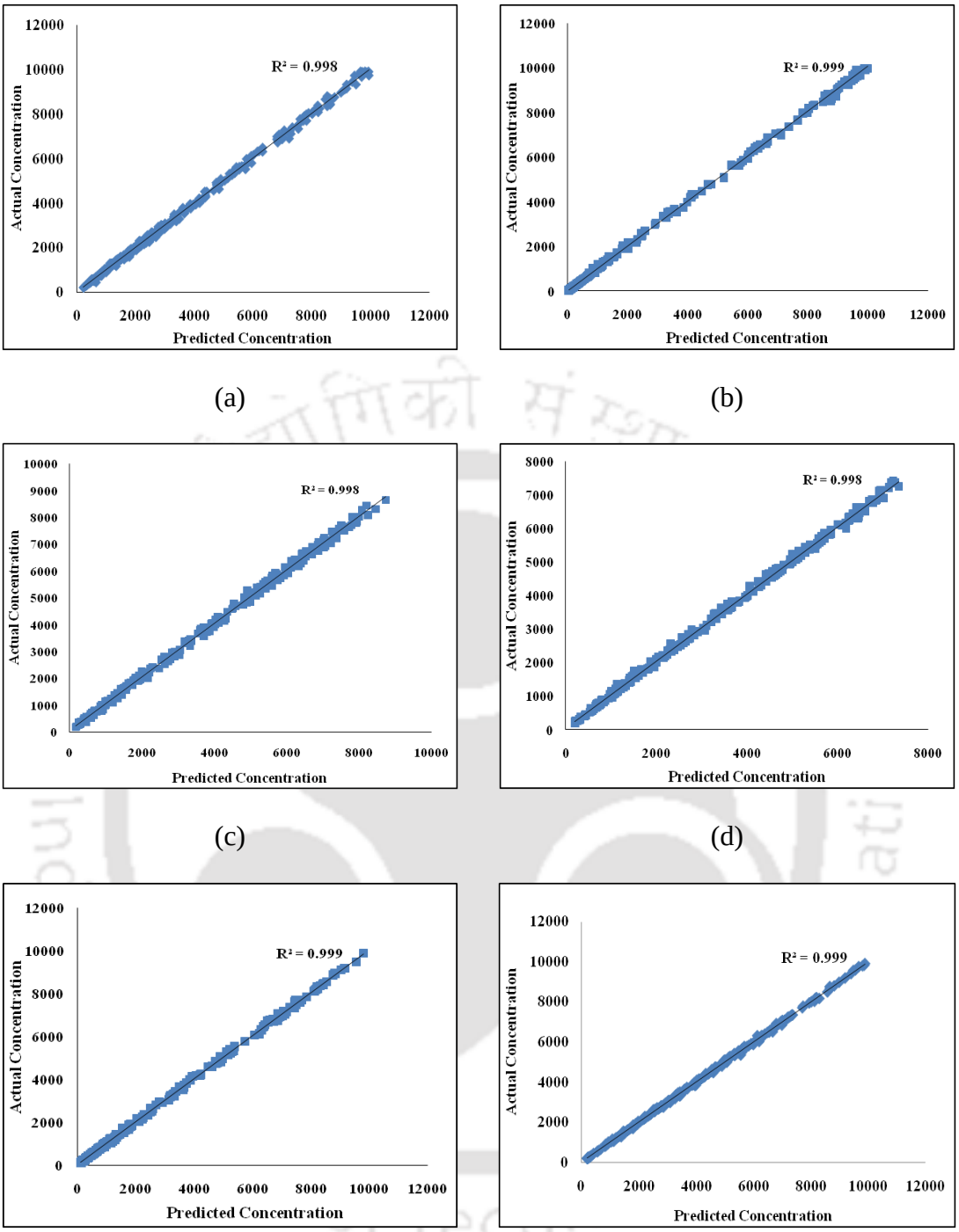


Fig. 3.24: Scatter plot for (a) ANN model 15 (b) ANN model 16 (c) ANN model 17 (d) ANN model 18 (e) ANN model 19 and (f) ANN model 20

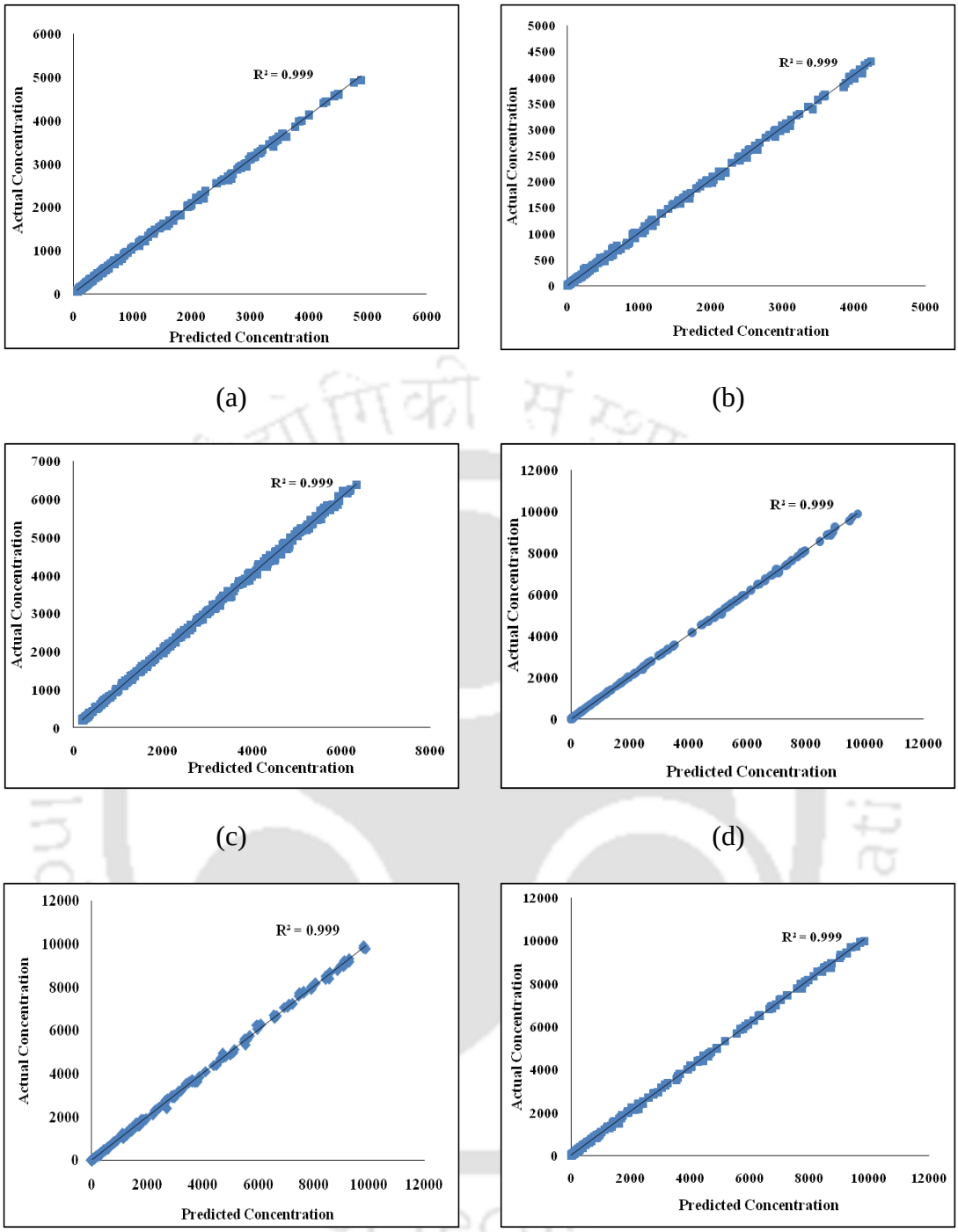


Fig. 3.25: Scatter plot for (a) ANN model 21 (b) ANN model 22 (c) ANN model 23 (d) ANN model 24 (e) ANN model 25 and (f) ANN model 26

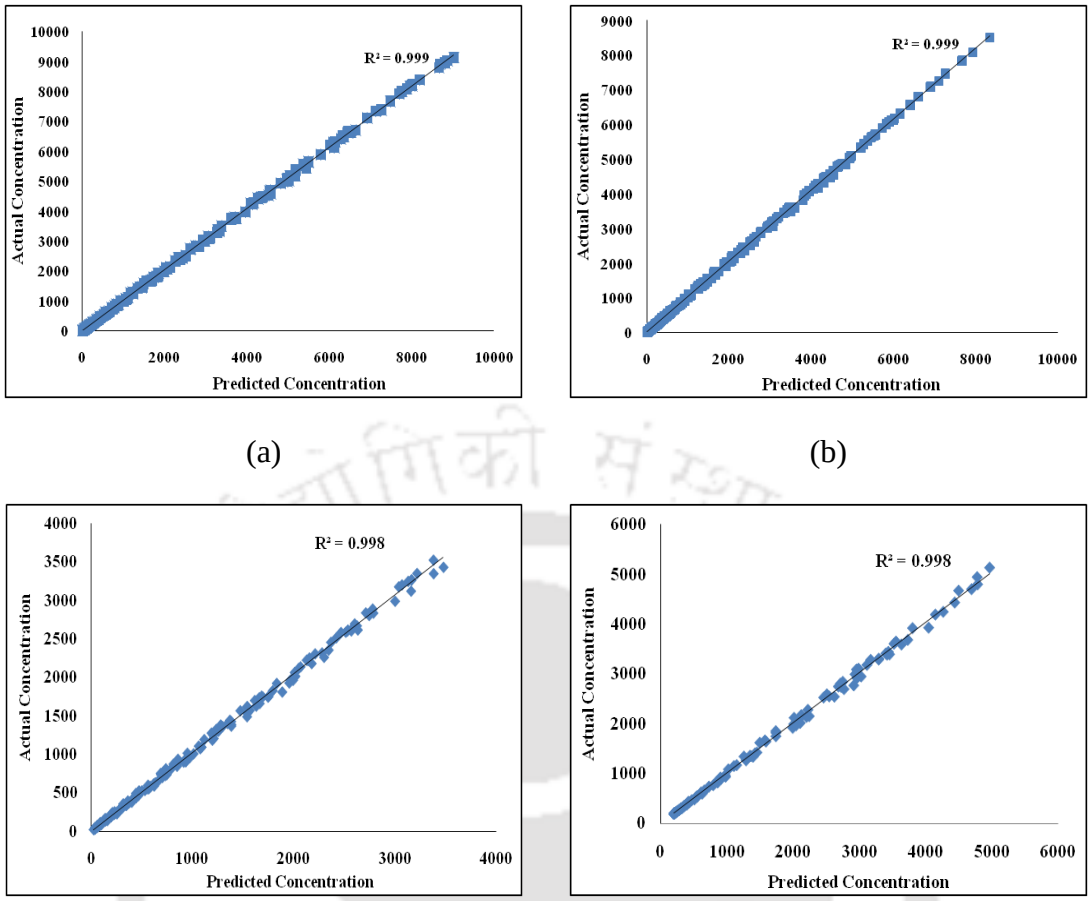


Fig. 3.26: Scatter plot for (a) ANN model 27 (b) ANN model 28 (c) ANN model 29 (d) ANN model 30

Table 3.5: Performance of the trained ANN model using AARE and RMSE

ANN MODEL	AARE	RMSE	ANN MODEL	AARE	RMSE
ANN model 1	0.72	0.31	**ANN model 16**	0.91	0.40
ANN model 2	0.88	0.72	**ANN model 17**	0.76	0.52
ANN model 3	0.88	0.18	**ANN model 18**	0.83	0.16
ANN model 4	4.98	0.34	**ANN model 19**	1.86	0.21
ANN model 5	0.72	0.32	**ANN model 20**	0.93	0.23
ANN model 6	1.89	0.31	**ANN model 21**	0.98	0.40
ANN model 7	1.23	0.42	**ANN model 22**	0.76	0.11
ANN model 8	1.01	0.65	**ANN model 23**	3.17	0.52
ANN model 9	0.87	0.26	**ANN model 24**	0.85	0.18
ANN model 10	0.71	0.14	**ANN model 25**	1.26	0.22
ANN model 11	0.45	0.16	**ANN model 26**	1.04	0.15
ANN model 12	0.84	0.17	**ANN model 27**	0.58	0.34
ANN model 13	0.62	0.11	**ANN model 28**	0.51	0.26
ANN model 14	1.19	0.14	**ANN model 29**	2.55	0.49
ANN model 15	0.75	0.42	**ANN model 30**	3.11	0.23

3.8.2 Performance of the ANN-GA model in identifying the optimal wells

As presented above, the developed ANN model is linked with the optimization model. The optimization problem is solved using Genetic Algorithms. In every generation of Genetic Algorithms, the ANN models have to be run to calculate the simulated concentration at observation well locations. Since there are thirty ANN models developed for each observation well, the model need to run all the available ANN model in every generation. However, this will lead to be computationally very expensive. For this reason, the proposed optimization model will actually run the ANN model of only those locations where the z_i value is one. The selected well locations with z_i equal to one imply that the contaminant concentration is observed at the respective well locations.

Furthermore, a constraint has been set up for the model to select the ANN model within a specified range. Thus, the number of selected ANN model that runs in every generation is always less than the maximum number of observation wells. Now, we can say that the selected ANN models represent the optimal number of observation wells. It may be mentioned that the rate at which plume is moving in groundwater is a very slow process, and may even take in terms of years for moving a few meters. Deliberating this idea, the selection of an optimal number of observation well locations is carried out for different number of years. The following explains the selection of the optimal number of observation well at the end of a specified time period.

3.8.2.1 Selection of optimal observation wells for different number of years

Even though the movement of the pollutant sources in the groundwater is considered to be a slow process but considering the dynamic nature of the pollutant sources, the location for an optimal number of observation wells will be selected at the end of the first year, third year and fifth year. The model selects the optimal number of wells for each time period from a total number of 30 observation wells locations.

3.8.2.1.1 Selection of optimal wells at the end of the first year

In the first year, the optimal well locations selected by the model are close to the source locations. It can be seen that the location of the optimal well follows the path of the plume and a total number of 15 wells are selected ranging between 10 and 20 which are the maximum and the minimum number of optimal wells allowed by the model (Table 3.6).

Table 3.6: Optimal wells selected by the ANN-GA model at the end of first year

Location of the optimal observation wells *(i,j)*	W1 (30,56)	W3 (32,53)	W4 (63,52)	W5 (66,49)	W6 (70,61)
	W9 (72,49)	W10 (75,50)	W14 (82,53)	W16 (86,32)	W19 (87,32)
	W21 (89,31)	W22 (86,32)	W24 (87,32)	W25 (89,31)	W26 (92,32)

A total number of five optimal wells are detected close to the source S5 (with *i= 28, j=53*). The index '*i*' and '*j*' refers to the location of cells in terms of rows and column respectively. For instance, W1 refers to the selected optimal well with row, *i* = 30 and column, *j* =56. The model selects a sum of six optimal wells around the sources S4 (*i=60, j=48*) and S3 (*i=71, j=59*). However, it is observed that S2 (*i= 73, j=48*) being a dummy source, no optimal wells were selected around it by the model. The model manages to select four optimal wells for the pollutant source S1 (*i=8, j=30*). The source locations along with the optimal observation well locations at the end of the first-time period are shown in Fig. 3.27.

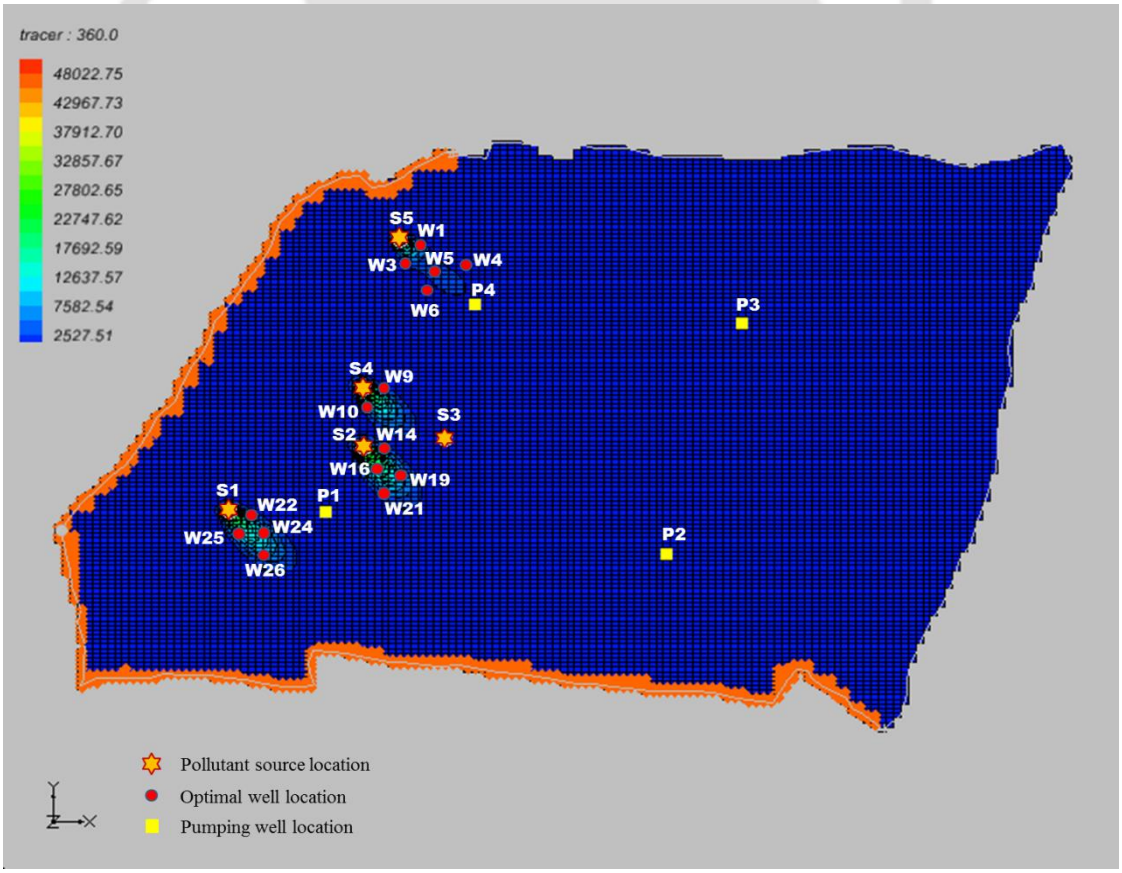


Fig. 3.27: Study area showing the optimal observation well locations at the end of the first year

3.8.2.1.2 Selection of optimal wells at the end of third year

The number and the locations of the optimal observation well locations selected at the end of the third year are given in Table 3.7.

Table 3.7: Optimal wells selected by the ANN-GA model at the end of third year

Location of the optimal observation wells (*i,j*)	W7 (38,62)	W10 (66,49)	W11 (68,53)	W12 (69,57)	W19 (82,53)
	W21 (82,50)	W23 (86,61)	W26 (92,32)	W27 (93,36)	W28 (99,36)

A total number of ten observation well are selected by the model. Fig. 3.28 shows the optimal well locations selected at the end of the third year. It is found that the selected well locations follow the path of the plume and are moving slowly away from the pollution source locations.

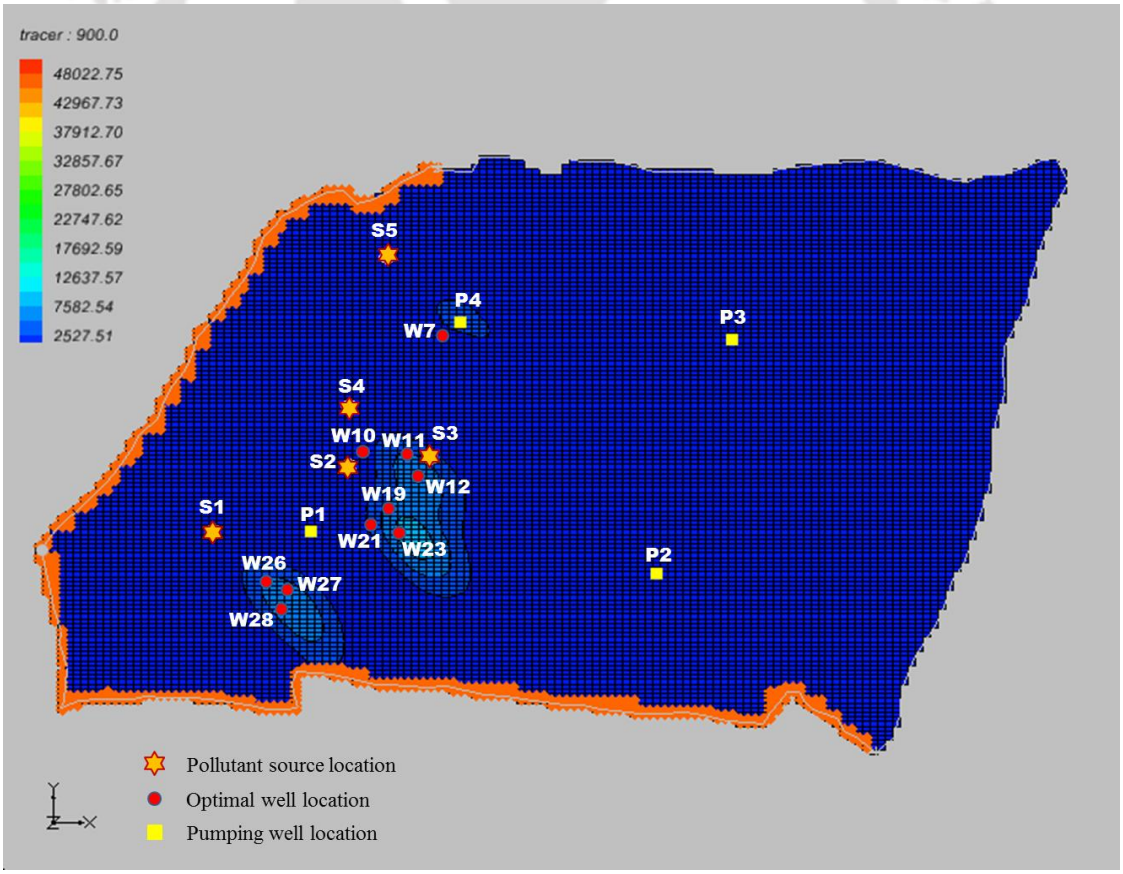


Fig. 3.28: Study area showing optimal observation well locations at the end of third year

3.8.2.1.3 Selection of optimal wells at the end of fifth year

Table 3.8 shows the number and the location of the selected well locations at the end of the fifth year. A total number of ten observation wells are selected for this phase. Fig. 3.29 gives a well-defined presentation of the selected well locations.

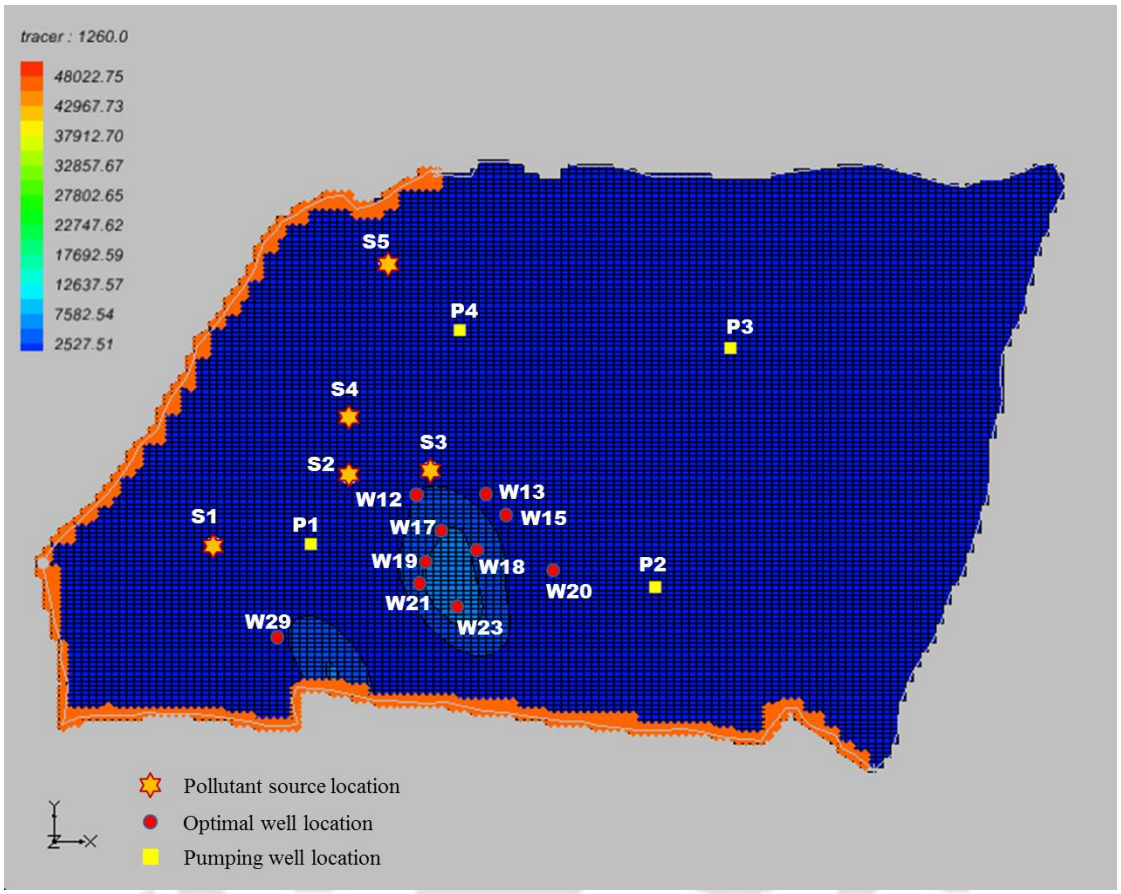


Fig. 3.29: Study area showing optimal observation well locations at the end of fifth year

It is seen that some of the wells fall far away from the outer contour of the contaminant plume. This is because the wells detected will monitor for two long years and the concentration of the wells being dynamic in nature induces the wells to change their location with time. It can also be noted that the observation well (W30) located very far from the pollutant sources is not selected by the model as the well is not within the reach of contaminant plume. Hence, with the passage of time, the observation wells are moving along with the direction of the plume.

Table 3.8: Optimal wells selected by the ANN-GA model at the end of fifth year

Location of the optimal observation wells (*i,j*)	W12 (69,57)	W13 (70,61)	W15 (74,63)	W17 (75,59)	W18 (80,61)
	W19 (82,53)	W20 (82,66)	W21 (83,50)	W23 (86,61)	W29 (103,36)

Most of the selected optimal observation wells for all the three-time period were found to be very close to the pollutant sources. It is due to the fact that the pollutant sources are active for five-time steps only. In the remaining time steps, the pollutant source becomes inactive resulting in the decrease of the pollutant concentration for the remaining time steps. The application of these identified optimal observation wells is seen in the subsequent section in identifying the groundwater pollution source

3.8.3 Identification of groundwater pollution sources using ANN-GA model

By using the selected optimal observation wells, the ANN-GA based model simultaneously identifies the groundwater pollution sources. Fig. 3.30 shows the comparison between the estimated pollution source fluxes with the actual source fluxes for the active time steps. The bar diagram indicates a close match between the predicted and the actual source fluxes.

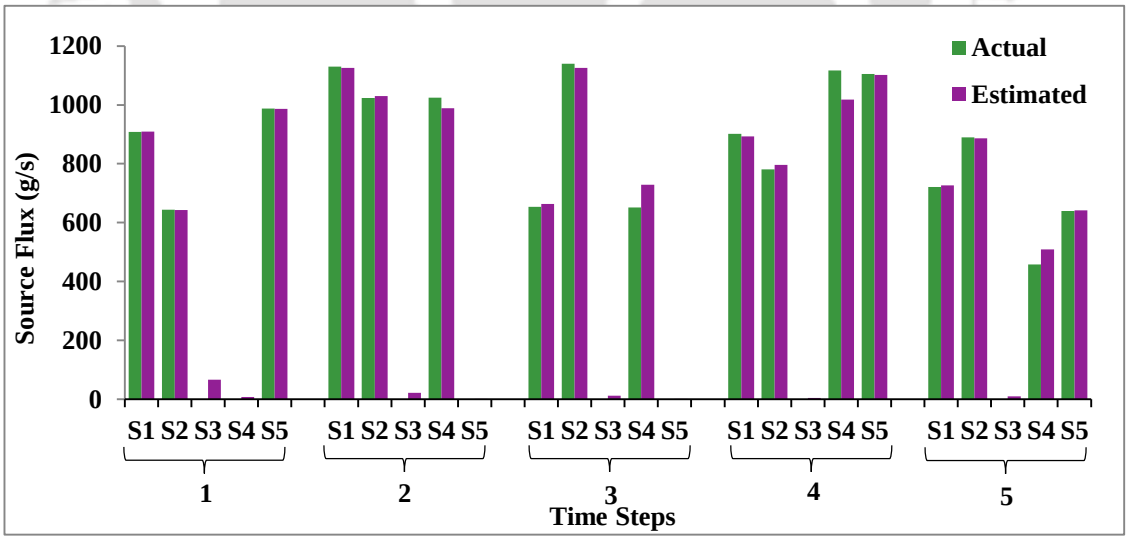


Fig. 3.30: Comparison between the actual fluxes and the estimated source fluxes for five stress periods

Table 3.9 shows the comparison between the source fluxes estimated by the ANN-GA model with the actual source flux. There are altogether four active sources (S1, S2, S4 and S5) and one dummy source (S3).

Table 3.9: Absolute relative error between actual sources and estimated sources

Time Steps	Source Locations	Actual Sources (g/s)	Estimated Sources (g/s)	Absolute Relative Error (%)
1	S1	908.42	909.54	0.12
	S2	644.02	643.20	0.12
	S3	0	66.6	-
	S4	0	7.73	-
	S5	987.08	986.203	0.08
2	S1	1130.5	1125.20	0.46
	S2	1023.87	1029.73	0.57
	S3	0	22.16	-
	S4	1024.16	988.78	3.45
	S5	0	0.38	-
3	S1	653.35	663.18	1.50
	S2	1139.88	1125.71	1.24
	S3	0	12.34	-
	S4	652.05	728.73	11.76
	S5	0	0.48	-
4	S1	902.15	892.55	1.06
	S2	781.09	796.56	1.98
	S3	0	4.73	-
	S4	1117.45	1017.66	8.92
	S5	1104.82	1101.45	0.30
5	S1	721.24	726.87	0.77
	S2	889.77	886.64	0.35
	S3	0	10.23	-
	S4	457.91	509.22	11.20
	S5	639.93	642.24	0.36

The dummy source is a source with negligible contaminant concentration hence it is an inactive source. Whereas the four pollution sources are active for five-time steps i.e. 15 months only. Relative efficiency of the ANN-GA model in predicting the source flux is also checked by evaluating the relative error with respect to the actual source flux. The estimated source flux for S1, S2, S4 and S5 in the first-time step are 909.54 g/s, 643.20

g/s, 7.73 g/s and 986.20 g/s respectively whereas, the respective actual flux being 908.42 g/s, 644.02 g/s, 0 g/s and 987.08 g/s. The close resemblance implies a good prediction capability of the model. This is further reflected in the relative error of S1, S2 and S5 sources as 0.12%, 0.12% and 0.08% respectively. An equivalent result is also observed for other time steps. However, some deviation is witnessed in estimated source flux for S4 in later time steps. The actual source flux for S4 at third, fourth and fifth time steps are 652.05 g/s, 1117.45 g/s and 457.91 g/s respectively while the estimated being 728.73 g/s, 1017.66 g/s and 509.22 g/s. Subsequently, the relative errors are estimated as 11.76%, 8.92% and 11.20%. But these values of relative error are not very high and are considered within the acceptable range. Even though the source pollutant S3 is a dummy source, some source concentration has been identified due to the effect of other pollutant sources.

The above discussion supports the competency of the present ANN-GA model in identifying the pollution sources for a large study area. The computational efficiency of the model is also analyzed considering the number of function count and the objective function value. The number of function evaluations was found to be 80201. This signifies the number of calls performed by the objective function to the simulation model. The function count for the present source identification problem can be regarded a good one because the performance evaluation of the present methodology is carried out in a large study area. The objective function value for the ANN-GA model in convergence towards the optimal solution is found to be 2.06E-05. Thus, it can be concluded that the present ANN-GA based methodology is computationally efficient in predicting the pollution sources also.

3.9 Summary and Conclusions

This chapter has proposed an ANN-GA based linked simulation-optimization methodology for identifying the unknown groundwater pollution sources. The performance of the linked simulation-optimization is computationally very expensive as a large simulation call is carried out by the optimization model to achieve the optimal solution. For reducing the computational time of the groundwater flow and transport processes, ANN model has been used as an approximate groundwater simulator. The developed methodology has been applied to a large illustrative study area. A total number of 30 observation wells was used on the large study area. The ANN model is used to predict the concentration of all the observation well and is developed for each of

the thirty observation wells. The computational efficiency will decrease if all these observation wells are utilized. Therefore, the ANN-GA model selects an optimal number of observation wells for identifying the groundwater pollution sources. It is seen that the ANN-GA model could successfully identify the pollution sources with the optimal number of observation wells for a different time period. A major drawback of this methodology is that the information such as location and number of pollution source should be always to known for developing the ANN model. But this case is not applicable in real life scenario where very little information about the source location and number of sources are available. So, the next chapter will discuss the situation where the information about the number and the location of the pollution sources are completely unknown to the problem.

Chapter 4

Identification of Pollution Sources Considering the Number, Source Location and Fluxes as Unknown

In the present chapter, a methodology has been proposed to identify the exact number and locations of the pollution source iteratively. The first part gives a brief introduction of the groundwater source identification problem when the number and the source locations are completely anonymous. The second section explains the iterative search technique for identifying the exact number of pollution sources and their locations. The third section describes the effect of measurement error on the observed concentration data. The other sections explain the development of groundwater simulation model, study area adopted and the discussion on results of the present methodology.

4.1 Introduction

The previous chapter presented an ANN-GA based linked simulation-optimization model for identifying the groundwater pollution sources. Taking into consideration the computational efficiency, the ANN model is used as an approximate groundwater simulator. The ANN-GA model successfully identified the pollution sources with an optimal number of observation wells. However, a major drawback of this approach is that the information about the number of sources present in the affected aquifer and the source location are known as there involves a prior training of the ANN model. Such scenario restrains the linked simulation-optimization to deal with a more realistic situation. For such reason, the present chapter has proposed an iterative based inverse optimization methodology where the information about the number of pollution sources and the source locations will be completely anonymous to the source identification problem.

4.2 Methodology

The earlier chapters have displayed that the inverse optimization method is usually adopted for solving the source identification problem and it is also revealed that the linked simulation-optimization approach is one of the most effective techniques for

incorporating the groundwater simulation model with the optimization model. For such reason, the present methodology utilizes the linked simulation-optimization approach by linking the MT3DMS and MODFLOW packages of GMS to the optimization model. The present methodology is proposed considering the realistic case of identification of the groundwater pollution sources. In real case scenario, the number, and the source characteristics in most of time are not known. With this concern, a methodology is proposed for identifying the number of pollution sources, their locations and fluxes available in the contaminated aquifer. The search for the exact number of pollution sources is carried out in an iterative manner and the iteration is initiated from two pollutant sources. The search is continued with successive number of pollutant source until all the sources are identified and the termination criteria are achieved. The detailed explanation of the present methodology will be explained subsequently.

4.2.1 Identification of pollution sources considering the number of pollution sources as unknown

In the proposed methodology, it is assumed that no information is available about the pollution sources. Fig. 4.1 shows the schematic representation of the iterative based unknown source identification problem (Model 2). The search is initiated considering that there are two pollution sources in the aquifer. Thus, initially *n* is set equal to 2, where n is the number of pollution sources. In the optimization model, the source locations, number and the magnitude of the source fluxes are treated as the unknown decision variables. The Genetic Algorithms has been adopted for solving the optimization model. The initial solution to start the genetic algorithms has been generated randomly. The solutions comprising of source locations and the fluxes are utilized in the groundwater simulation model as input.

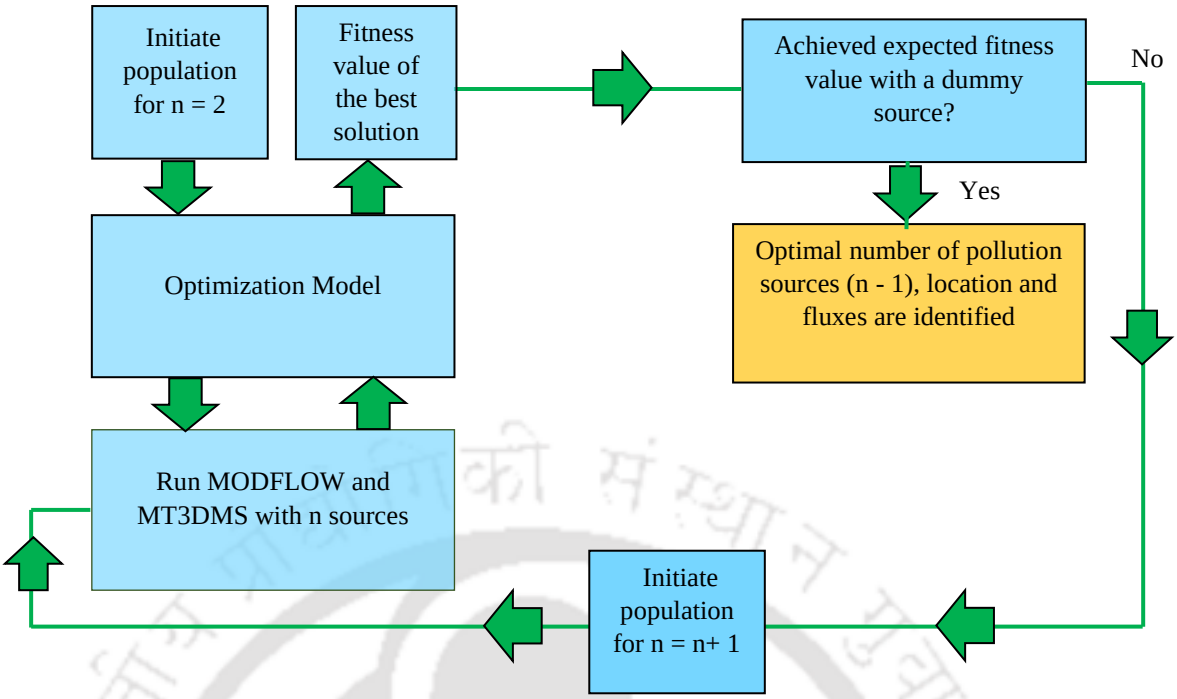


Fig. 4.1: Methodology of the iterative search model

The MT3DMS and the MODFLOW packages of the Groundwater Modeling System (GMS) are used for simulating the groundwater flow and transport processes. These models simulate the groundwater flow and transport processes using the assumed number of pollution sources. The simulation model provides the simulated contaminant concentrations to the optimization model for calculation of fitness values of each of the solution in the population. The optimization model then minimizes the difference between the observed contaminant concentration to obtain the location and the magnitude of the pollution sources.

The iterative search for the exact number of pollution sources is performed based on the fitness function value (F). A large value of fitness function means that the magnitude of the actual concentration and the model achieved concentration at the pollution source does not match with each other. The recovered source location and the source flux are bound to be erroneous values as the concentration values at the respective observation wells are giving inaccurate results. Henceforth, it clearly shows the presence of more pollution sources in the aquifer. Eventually, another pollution source is added to the existing one as $n = n+1$. Now the groundwater flow and transport processes will be again simulated with $n = 3$. It further goes to the optimization model and the whole process is repeated again.

If the fitness function value converges to a minimum value for n = 3, it corresponds to the best match between the observed and simulated contaminant concentration. Therefore, for further confirmation, another source is added ($n = n+1$) to the prevailing one and now with $n = n+1$, the present methodology is repeated as the earlier one. After repeating the whole process, if the fitness value converges to a minimum value with a dummy source (sources with negligible concentrations), it signifies that the actual number of pollution sources have been attained. The confirmation for the exact number of pollution sources is based on the dummy source. It is because the retrieved source being a redundant source (dummy), it does not have contaminant concentration. It suggests that there is no further presence of the pollution source in the aquifer. Hence, the dummy source can be discarded and the exact number of pollution source in the aquifer is given by $n = n-1$.

The above steps are repeatedly carried out for the search of the exact number of pollution sources and continued as n = 4, 5, 6… until the termination criteria are satisfied.

4.2.2 Development of unknown source identification model

The groundwater pollution source identification problem can be solved using the linked simulation-optimization model which minimizes the difference between the simulated concentration and the observed concentration at the well locations. The observed concentration is the concentration measured at different observation locations at different time steps. The simulated concentrations are obtained by solving the groundwater flow and transport simulation model. In the present methodology, the modules MODFLOW and MT3DMS present in GMS is used in simulating the groundwater flow and transport processes respectively.

For minimizing the square of difference between the observed and the simulated concentration in space and time, an optimization model is required. For the present scenario, genetic algorithms (GAs) are used for solving the optimization problem. The exact pollution source location will be identified when the observed and the simulated contaminant concentrations perfectly match with each other at different time steps. The decision variables of the present optimization model are the pollution source location (X, Y) and the source flux (Sf). The objective function of the optimization model can be written as

$$\text{Minimize } F_n = \sum_i^M \sum_j^N (C_{o,i}^j - C_{s,i}^j)^2 \quad (4.1)$$

Subject to

$$C_v = f(X, Y, Sf) \quad (4.2)$$

$$Sf_{min} \leq Sf \leq Sf_{max} \quad (4.3)$$

$$X_{min} \leq X \leq X_{max} \quad (4.4)$$

$$Y_{min} \leq Y \leq Y_{max} \quad (4.5)$$

Here, F_n is the objective function for the present optimization model with *n* number of pollution sources; $C_{o,i}^j$is the observed concentration at j^{th} time step for i^{th} well location; $C_{s,i}^j$ is the simulated concentration at j^{th} time step for i^{th} well location; *M* is the total number of observation wells and; *N* is the total number of time steps; C_v is the concentration vectors of the simulated concentration; *Sf* is the vector of the pollutant source fluxes such that $Sf = [Sf_1, Sf_2, Sf_{3,\ldots,} Sf_n]^T$; *X* and *Y* are the source location vectors such that $X = [x_1, x_2, x_{3,\ldots,} x_n]^T$ and $Y = [y_1, y_2, y_{3,\ldots,} y_n]^T$; Sf_{min} and Sf_{max} are the lower and upper bounds of the source flux; X_{min}, X_{max}, Y_{min} and Y_{max} are the lower and upper bounds where the source locations are expected. It can be noted that the bounds for source fluxes are taken as Sf_{min} = 0 g/s and Sf_{max}= 100 g/s. It may be noted that in MT3DMS model, the location of the pollutant sources should coincide with the centre of the discretized grid blocks (Ayvaz, 2010). Therefore, the lower and upper bounds for the source locations are provided on the basis of the maximum number of grids discretized for the study area.

4.3 Measurement error of observed data

As seen in the source identification model, the pollution source can be efficiently identified by minimizing the error function between the simulated and observed contaminant concentration at the observation well location. The observed contaminant concentrations are obtained from the field so, some measurement errors are bound to occur in observed contaminant concentration for real life scenario. As such as study has been carried out to evaluate the effect of measurement error in identifying the pollution sources. The extent of measurement error is not known therefore, different level of error has been introduced to the observed concentrations for evaluating the impact of measurement error. The effect of measurement error in the source identification problem was performed by Mahar and Datta 2001; Singh and Datta 2004; Ayaz 2010;

Datta et al. 2010; Prakash and Datta 2015, however not much emphasis was given on the impact it might have while determining the pollution source locations and the magnitude of source fluxes. The measurement error in the observed contaminant concentration can be generated randomly using the following equation (Mahar and Datta, 2001).

$$PCo_i^j = Co_i^j(1 + err) \tag{4.6}$$

Where, PCo_i^j is the perturbed simulated concentration value; Co_i^j is the simulated concentration value; *err* is the error term to be introduced. The error term (err) is calculated using the normal distribution with sigma (σ) and zero mean (μ= 0). Here, σ signifies different level of error magnitude ranging from 0.05 to 0.2 (Singh and Datta, 2007). When $\sigma < 0.1$, it signifies a low noise level, $0.1 \leq \sigma \leq 0.15$ denotes moderate level of noise and $\sigma \geq 0.15$ signifies high level of noise (Singh and Datta, 2007). If the *err* term is not introduced to the observed concentrations, then it may be considered that there is zero noise in the measured contaminant concentration. With the assumption that there is no measurement error in the observed concentration, it signifies that the source identification model will give the best possible solution. However, it does not support the realistic case as explained above. Therefore, introducing the error term will be regarded as one of the best approaches to check the effectiveness of the model with a real situation. Henceforth, the different noise levels are incorporated in the observed concentrations and are used to identify the pollution source locations and source fluxes using the proposed methodology.

4.4 Performance evaluation criteria

The performance of the proposed simulation-optimization model has been evaluated using relative error (*RE*) (Borah and Bhattacharjya, 2014).

The relative error (RE) criterion to evaluate each of the source flux can be computed as

$$RE = \frac{|Ef_{p,r} - Af_{p,r}|}{Af_{p,r}} \text{ x } 100 \tag{4.7}$$

Where $Ef_{p,r}$ and $Af_{p,r}$ are the estimated and the actual source fluxes respectively at r^{th} location and p^{th}stress period.

4.5 Development of groundwater simulation model

The groundwater flow and transport processes have been simulated by using the Groundwater Modeling System (GMS). GMS has a graphical user interface (GUI) for preparing the input files for the different models such as MODFLOW, MT3DMS, RT3D, MODPATH, FEMWATER, SEEP2D, SEAM3D etc. In the present study, MODFLOW and MT3DMS are used for simulating the groundwater flow and transport process. The required input files for MODFLOW and MT3DMS simulation are prepared using the graphical user interface of GMS. Subsequently, the simulation is performed by executing the 'exe' files of MODFLOW and MT3DMS. The following section presents a detailed explanation of how the simulation is performed by using MODFLOW and MT3DMS models.

4.5.1 MODFLOW

MODFLOW (McDonald and Harbaugh, 1989) is an executable program written in FORTRAN. It is based on finite difference method that can numerically solve the groundwater flow equation on a porous medium. There are two approaches available in GMS for MODFLOW simulation. They are grid approach and conceptual approach. In the grid approach, the model is developed using 3D grid module. Here, the editing of the parameters of the model is carried out on cell-by-cell basis. On the other hand, the conceptual approach uses map module for developing a model of the study area. With the development of the conceptual model, a grid is automatically generated that will fit the domain. In the present study, the grid approach is used for simulating the groundwater processes. The study area adopted is a simple rectangular hypothetical study area which does not require a map module. For this reason, the grid approach has been preferably used.

The first step involved in developing a MODFLOW model is the creation of a 3D/2D cell centered grid using the required command available in the main menu. The next important task is the preparation for the MODFLOW simulation using different packages. The required packages are available in *packages* dialog for specifying the various hydrogeological characteristics of an aquifer. The division of various packages enables the user to select particular hydrologic conditions of the aquifer accordingly. Some of the important packages included in MODFLOW are boundary head and flow packages, layer property flow package, strongly implicit procedure packages, source/sink section etc. First of all the packages required for the simulation is turned on.

The next step involves the development of boundary layers using *IBOUND* dialog. In the next step, the layer property flow (LPF) package is selected up for setting up the conductance between the each of the cell. The well in the cell is assigned by using the *source/sink* command of the well package.

After all the necessary input files from the respective packages are prepared, the MODFLOW simulates the flow processes following the time domain and is known as 'stress periods'. For transient state case, the stress periods are further divided into time steps. Whereas in case of steady state there is no change in the time domain, hence there is a single stress period with a single time step. After simulating the groundwater flow processes using MODFLOW, different HDF5 files of the used packages get saved in the hard drive. These files are then used by MT3DMS in simulating the groundwater transport process. The groundwater flow equation used in MODFLOW can be written as

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_{zz}\frac{\partial h}{\partial z}\right)+W=S_s\frac{\partial h}{\partial t} \quad (4.8)$$

Where, K_{xx}, K_{yy} and K_{zz} are the hydraulic conductivity along the *x*, *y* and *z* directions (LT^{-1}); h is the hydraulic head (L); S_s is the specific storage coefficient; t is the time (T); W is the recharge flux per unit area (LT^{-1}).

4.5.2 MT3DMS

MT3DMS has various packages similar to MODFLOW that simulate the groundwater transport processes. MT3DMS works in conjunction with MODFLOW. MODFLOW compute the hydraulic head value at the center of each cell during the groundwater flow simulation and are written in formatted files. Later MT3DMS read these files as the flow fields which are further used in transport model simulation. Therefore, for performing MT3DMS simulation, one has to run the MODFLOW to obtain the flow fields. Likewise the MODFLOW model, the MT3DMS model uses the grid approach for the simulation. It may be noted that MODFLOW and MT3DMS should have the same number of stress periods because the flow field and the source/sink data will be required in the transport model. The graphical interface present in GMS can be used for performing the MT3DMS transport simulation by undergoing pre-processing and post-processing steps. The inputs required for MT3DMS are generated in GMS, and then the *exe* files are executed. For setting up the MT3DMS transport model, it uses a series of packages available in the main menu. It follows a similar sequence as that MODFLOW

for preparation of the input data. The additional important packages in the MT3DMS model are the concentration package for defining the species, the advection package and the dispersion package for the setting up the details of the transport processes. In the present study, the previously created MODFLOW solution is being read for simulating the transporting model. With the saved data, the necessary transport packages are used for preparing the input files. Once the input files are ready, the simulation is performed. The simulation result shows the contaminant concentrations computed by the MT3DMS. The mass of the contaminant in the aquifer at various time steps can also be calculated.

For post-processing the data, the output files are exported from the MT3DMS. The *exe* files are exported to a different environment for modifying the post-processed output data. In the present study, a MATLAB code is written for exporting and transforming the input files of MT3DMS. The MT3DMS model is then linked to the optimization model for identification of unknown groundwater sources as discussed in earlier sections. Fig. 4.2 shows the steps involved in pre-processing and post-processing of the input and output files in the GMS and MATLAB environment.

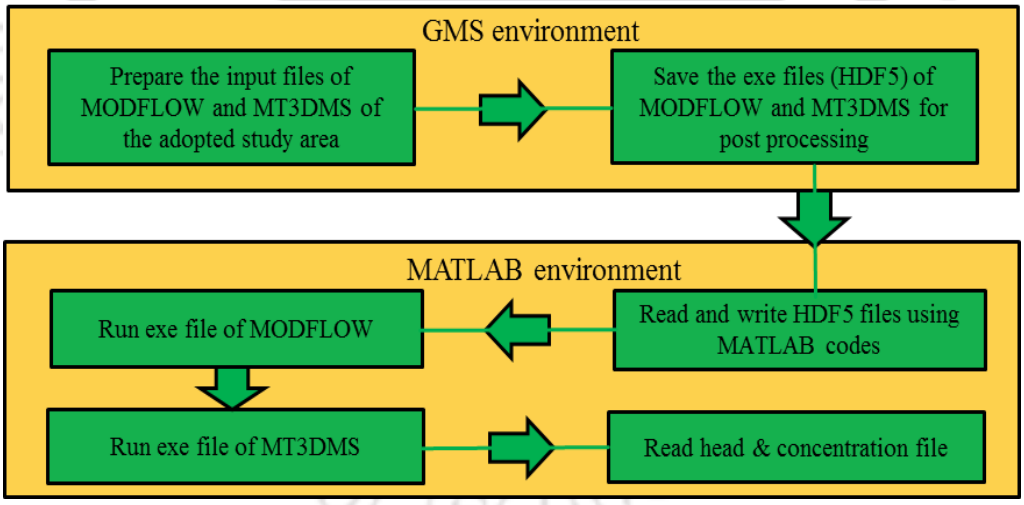


Fig. 4.2: Pre-processing and post-processing of input-output files in GMS and MATLAB environments

4.6 Study Area

An illustrative study area with an area of 1.04 km^2 (Fig. 4.3) is adopted for evaluating the performance of the proposed model. The boundary conditions and the geometry of the study area have been considered as proposed by Mahar and Datta (2001). It is a homogeneous and isotropic aquifer with constant head boundary on the left and right-hand boundaries. The hydraulic head varies from 100 m to 99.58 m on the right-hand

side of the aquifer. On the other end of the left end, the hydraulic head is varying from 88.72 m to 88.00 m. The no-flow boundary condition has been considered on the upper and lower boundaries of the aquifer.

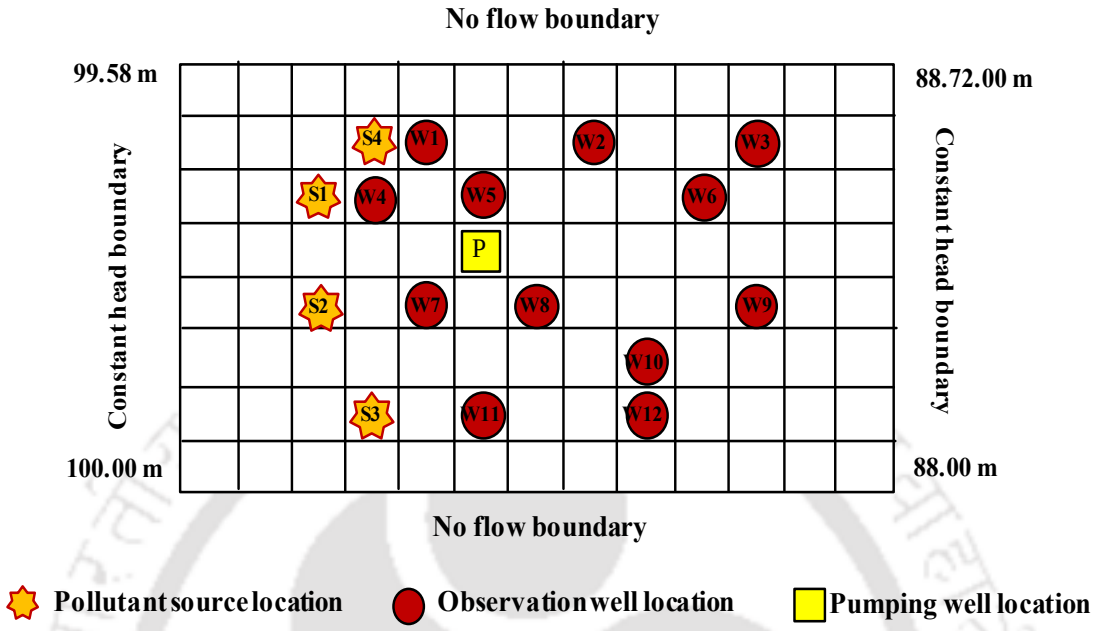


Fig. 4.3: Illustrative study area showing the actual source locations, observation well locations and pumping wells

The 2D study area has been discretized in 13 columns and 8 rows. The cell size is 100m x 100m. A total number of four pollutant sources are present in the aquifer but the location and number of source pollutant are completely unknown and have to be identified using the present methodology. It is assumed the pollution sources are active for 4-time steps i.e. the source releases pollutant for one year. The groundwater flow and transport processes are simulated for 5 years (5 stress periods) at an interval of three months. A total number of 12 observation wells are placed randomly around the suspected area of the aquifer.

Table 4.1: Hydrological parameters used in the study area

Parameters	Values
Hydraulic conductivity in x-direction, $\boldsymbol{K_{xx}}$(m/s)	0.0002
Hydraulic conductivity in y-direction, $\boldsymbol{K_{yy}}$ (m/s)	0.0002
Porosity, $\boldsymbol{\varepsilon}$	0.25
Thickness of the aquifer, *b* (m)	30.5
Longitudinal dispersivity, $\boldsymbol{\alpha_L}$(m)	40
Transverse dispersivity, $\boldsymbol{\alpha_T}$ (m)	9.6
Time steps, $\Delta \boldsymbol{t}$ (months)	3

Table 4.2: Source fluxes for different time steps (g/s)

Sources	Time Step 1	Time Step 2	Time Step 3	Time Step 4
S1	30	58.5	0	35
S2	47	15	37	0
S3	41.26	0	14.40	16.88
S4	21.7	0	29.68	0

Table 4.3: Pumping rates of the well at the pumping location of the aquifer (m^3/d)

Time Step	Rate	Time Step	Rate
1	273.02	11	163.29
2	163.29	12	327.45
3	327.45	13	273.02
4	163.29	14	163.29
5	273.02	15	381.02
6	327.45	16	217.72
7	163.29	17	273.02
8	273.02	18	163.29
9	381.02	19	327.45
10	217.72	20	217.72

The hydrological parameters used in the study area are shown in Table 4.1. The observation wells are designated as W1, W2…W12. The adopted study area can be regarded as a complex model due to the presence of a large number of pollutant sources in the aquifer.The magnitudes of the pollutant sources are shown in Table 4.2. There is a pumping well in the aquifer and the pumping rates for the twenty-stress periods are given in Table 4.3.

4.7 Results and Discussion

The present section will display the results obtained considering different number of pollution sources. The effect of incorporation of different noise level in the observed contaminant concentration is subsequently presented.

4.7.1 Different number of pollution sources

It can be noted that the present linked simulation-optimization model will search for the number of pollution sources and source flux starting from two number of pollution sources. With no information available about the number of pollution sources in the aquifer, the pollution sources are successively increased in the simulation-optimization model until the optimal solution is achieved.

4.7.1.1 Number of sources n = 2

In the first trial, it is assumed that only two pollutant sources are present in the affected aquifer. Table 4.4 shows the estimated source locations, fluxes and the final objective function value (*F*) for two pollution sources. Here, it is assumed that there is no measurement error in the observed concentration. Therefore, the identification of the source location, magnitude of the flux and number of pollution source are basically based on zero noise level.

Table 4.4: Estimated source location, fluxes and final objective function for $n = 2$

No. of sources	Number of decision variable	Estimated source location (*i,j*)	Estimated source flux (g/s)				Final function value *F*
			Time Step 1	Time Step 2	Time Step 3	Time Step 4	
2	10	(3,2)	59.38	30.00	5.01	0.00	*F2* = 5.38
		(4,4)	69.73	58.99	76.89	20.70	

For this scenario, the candidate solution for two sources has 10 decision variables. Eight of these decision variables (4-time steps x 2 sources) represent the source fluxes while the remaining two variables represents the source location (coordinate points of the grids).

The linked simulation-optimization model is applied and the estimated source locations are found to be (3,2) and (4,4) with 59.38 g/s, 30 g/s, 5.01 g/s, 0 g/s and 69.73 g/s, 58.99 g/s, 76.89 g/s, 20.70 g/s as source fluxes respectively. The function value *F2* for the two pollution sources is found to be 5.38. This value of the objective function emerges to be the best function value for $n = 2$ as no improvement is seen in it and ultimately the function tolerance (stopping criteria) compels the simulation to end. The high value of *F2* suggests that a third pollution source is available in the aquifer.

4.7.1.2 Number of sources n = 3

Table 4.5 shows the estimated source location, fluxes and the final objective function values (*F*) for three pollution sources ($n = 3$). As there is three pollution sources, there are 15 decision variables. Twelve of the decision variables (4-time steps x 3 sources) denotes source fluxes and the remaining three being locations of the sources. With three pollution sources, the iterative search model identified the source locations as (3, 3), (5, 3) and (7, 4). In the case, the *F3* reduces to 3.94. Even though the final function values are reduced to some extent, this does not represent the minimum value. Therefore, it further suggests that more pollution sources are present there in the aquifer.

Table 4.5: Estimated source location, fluxes and final objective function for n = 3

No. of sources	Number of decision variable	Estimated source location (*i,j*)	Estimated source flux (g/s)				Final function value *F*
			Time Step 1	Time Step 2	Time Step 3	Time Step 4	
3	15	(3,3)	43.89	72.49	8.99	20.22	*F*3= 3.94
		(5,3)	70.54	28.01	10.45	0.07	
		(7,4)	68.96	47.95	16.03	8.05	

4.7.1.3 Number of sources n = 4

The next iterative process is performed repeatedly with four pollution sources. The present iterative search methodology determined the source location as (3, 3), (5, 3), (7, 4) and (2, 4). The estimated source location, fluxes and the final objective function values (F) for the number of pollution sources n = 4 are given in Table 4.6. The final function value (*F4*) for four sources remarkably reduced to 0.072. This smaller objective function value indicates that the actual number of pollution sources may have attained. Therefore, it can be concluded that n = 4 represents the actual number of the pollution sources. But for further confirmation, the model is run again with five number of pollution sources.

Table 4.6: Estimated source location, fluxes and final objective function for n = 4

No. of sources	Number of decision variable	Estimated source location (*i,j*)	Estimated source flux (g/s)				Final function value *F*
			Time Step 1	Time Step 2	Time Step 3	Time Step 4	
4	20	(3,3)	30.18	57.91	1.48	33.98	*F4* = 0.072
		(5,3)	46.6	16.61	34.59	1.18	
		(7,4)	40.7	0.14	15.59	15.94	
		(2,4)	20.92	2.16	27.10	1.41	

4.7.1.4 Number of sources n = 5

Table 4.7 shows the estimated location and flux and fitness function for n = 5. Again, the final *F5* value is found to be 0.0201, closely matches with the fitness function for n = 4. The magnitudes of source fluxes for the fifth source are found to be very negligible as compared with the other sources and can be regarded as a dummy source. This suggested that the solution converges to an optimal solution with a dummy source. Thus, the number of pollution sources is equal to four, the fifth being a dummy source.

Table 4.7: Estimated source location, source flux and Final fitness (F) for n = 5

No. of sources	Number of decision variable	Estimated source location (i,j)	Estimated source flux (g/s)				Final function value F
			Time Step 1	Time Step 2	Time Step 3	Time Step 4	
5	25	(3,3)	30.58	56.81	2.03	33.36	F5 = 0.0201
		(5,3)	46.36	16.33	34.05	1.24	
		(7,4)	41.16	2.42	9.29	15.53	
		(2,4)	21.70	0.13	29.23	0.34	
		(1,10)	3.82	2.90	4.38	6.12	

Furthermore, when the estimated source locations and the source fluxes were checked for n = 4 and n = 5, it is seen that they closely match with each other. The estimated source locations are (3, 3), (5, 3), (7, 4) and (2, 4).It is also observed that the fitness function value slowly decreases with the increase in the number of pollution sources as seen in Fig. 4.4. The fitness value converges to a minimum value when the number of pollution sources reaches four. As the exact number of pollution sources has been attained, the fitness function value for n =5 almost match with n =4.

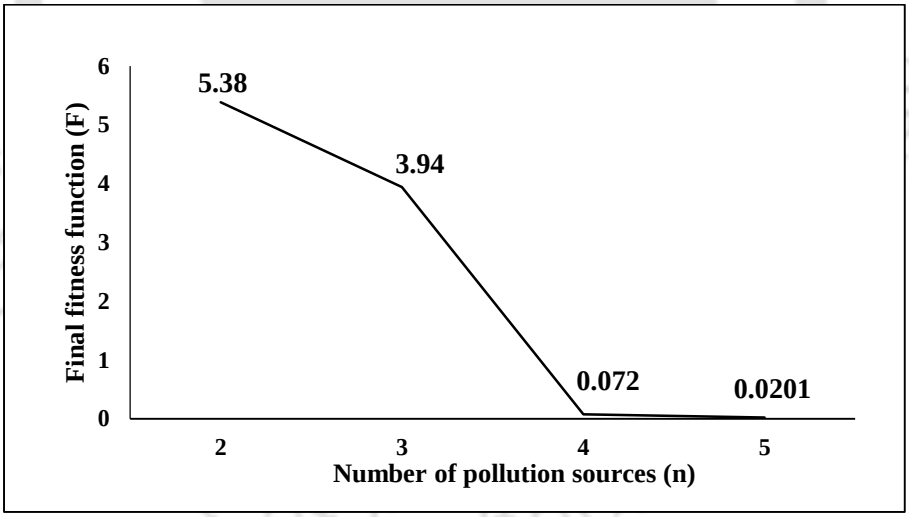


Fig. 4.4: Final fitness function for different number of pollution sources

With this confirmation that there is four pollution sources, a comparison between the estimated pollution sources and the actual pollution sources is performed and showed a close proximity as seen in Table 4.8. Some slight discrepancies are observed in the estimated source fluxes. This can be reflected in the estimated source fluxes for S1 (3,3) and are found to be 30.58 g/s, 56.91 g/s, 1.48 g/s and 33.98 g/s. The respective actual source fluxes being 30.00 g/s, 58.80 g/s, 0 g/s and 35 g/s respectively.

For a better understanding of the differences between the estimated and the actual source fluxes, the relative errors of the estimated source fluxes are evaluated with respect to the actual fluxes. The highest relative error among all the source fluxes is found to be 8.86 % whereas the lowest is 0.00 %. These values of relative error are within acceptable ranges and showed a good prediction capability of the model. The comparison between the estimated and the actual source fluxes are further represented using the bar diagram (Fig. 4.5). It clearly signifies the close resemblance of the source fluxes when no measurement error has been considered. Presence of source fluxes could be seen on some of the inactive time steps (S23, S24, S31, and S42) but as the value of the recovered source fluxes is very small, it can be neglected.

Table 4.8: Comparison between the estimated and the actual pollution sources using relative error

Actual source location (*i,j*)	Estimated source location (*i,j*)	Actual source flux (g/s)				Estimated source flux (g/s)				Relative error (%)			
		Time Step 1	Time Step 2	Time Step 3	Time Step 4	Time Step 1	Time Step 2	Time Step 3	Time Step 4	Time Step 1	Time Step 2	Time Step 3	Time Step 4
(3,3)	(3,3)	30.00	58.80	0	35	30.58	56.81	2.03	33.36	1.93	3.38	-	4.68
(5,3)	(5,3)	47.00	15.00	37	0.00	46.36	16.33	34.05	1.24	1.36	8.86	7.97	-
(7,4)	(7,4)	41.26	0	14.40	16.88	41.16	2.42	10.29	15.53	0.24	-	7.71	7.10
(2,4)	(2,4)	21.70	0.00	29.68	0	21.70	0.13	29.23	0.34	0.00	-	1.52	-

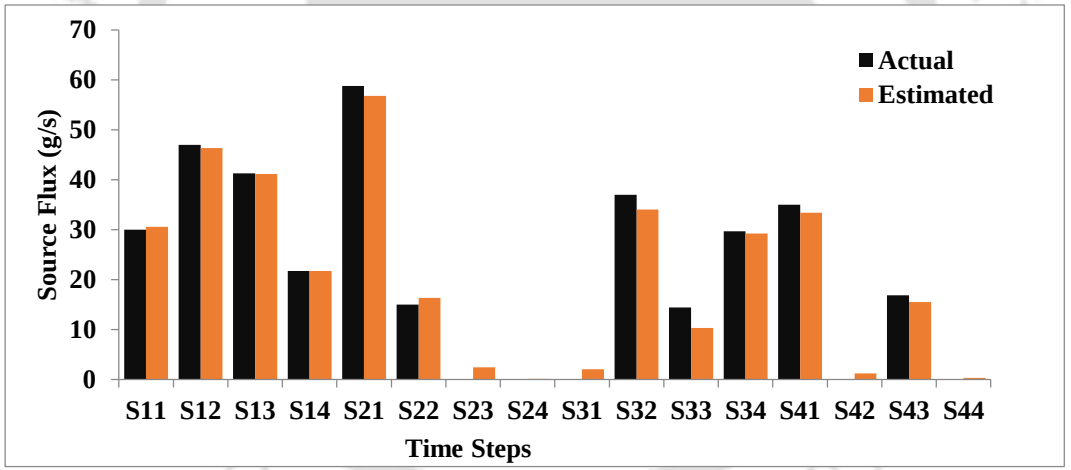


Fig. 4.5: Comparison of source flux between actual and estimated source flux

4.7.2 Effect of different noise level on source location and fluxes

It is not always possible to accurately measure the concentration data from the field. As such an analysis has been carried out to evaluate the performance of the model when there is some error in the observed data. In this case, a measurement error is introduced to the observed concentration values. The error is calculated using the normal distribution with sigma (σ) and zero mean (μ= 0). The random error is added to the observed concentrations. The different level of noises (σ) introduced are 0.05, 0.1, 0.15 and 0.2.

4.7.2.1 Effect of noise level for n = 2

Table 4.9 shows the coordinates of the estimated source location for n =2 considering noise levels σ = 0.05, 0.1, 0.15 and 0.2. The estimated source locations at different noise levels are found to be (3,2), (3,3), (8,4), (4,4) (5,5) and (6,3). The actual coordinates identified considering the present iterative technique are found to be (3,3), (5,3), (7,4) and (2,4). A close comparison shows that the identified source locations are found to be located close the actual source locations. It is observed that the source location (3,3) is exactly identified in most of different noise levels.

Table 4.9: Identified source location for n = 2 at different noise level

Number of pollution sources	**Actual pollution source location (*i, j*)**	**Estimated source location (*i,j*)**				
		At zero noise level	**At noise level σ = 0.05**	**At noise level σ = 0.1**	**At noise level σ = 0.15**	**At noise level σ = 0.2**
2	(3,3) (5,3)	(3,2)	(3,3)	(8,4)	(3,3)	(6,3)
	(7,4) (2,4)	(4,4)	(4,4)	(3,3)	(5,5)	(3,3)

The prediction capability of the present model is further analysed by evaluating the source fluxes (for n = 2) at the different noise level (Table 4.10). It may be noted that when the noise level is zero, the identified source fluxes are found to closely resemble the actual source fluxes. However, the identified source fluxes for different noise levels showed some variations from the actual source fluxes. It can be clearly observed from Fig. 4.6 that the variation in the source fluxes does not follow a trend with the increase in the noise level from σ = 0.05, σ = 0.1, σ =0.15 and σ =0.2.

Table 4.10: Estimated source fluxes for n = 2 at different noise level

Time Steps	Source locations (*i*,*j*)	Actual source fluxes	Magnitude of estimated source fluxes (g/s)				
			At zero noise level	At noise level σ = 0.05	At noise level σ = 0.1	At noise level σ = 0.15	At noise level σ = 0.2
1	S1	30.00	30	25.95	99.96	51.64	89.74
	S2	47.00	46.67	79.89	40.19	43.86	32.08
2	S1	58.80	57.91	46.58	65.20	28.52	53.89
	S2	15.00	16.61	78.78	97.94	37.31	78.67
3	S1	0.00	1.49	0.01	8.36	53.50	0.44
	S2	37.00	34.59	19.99	9.03	12.20	0.00
4	S1	35.00	33.98	19.96	2.05	9.61	0.06
	S2	0.00	1.185	7.34	25.47	20.41	39.74

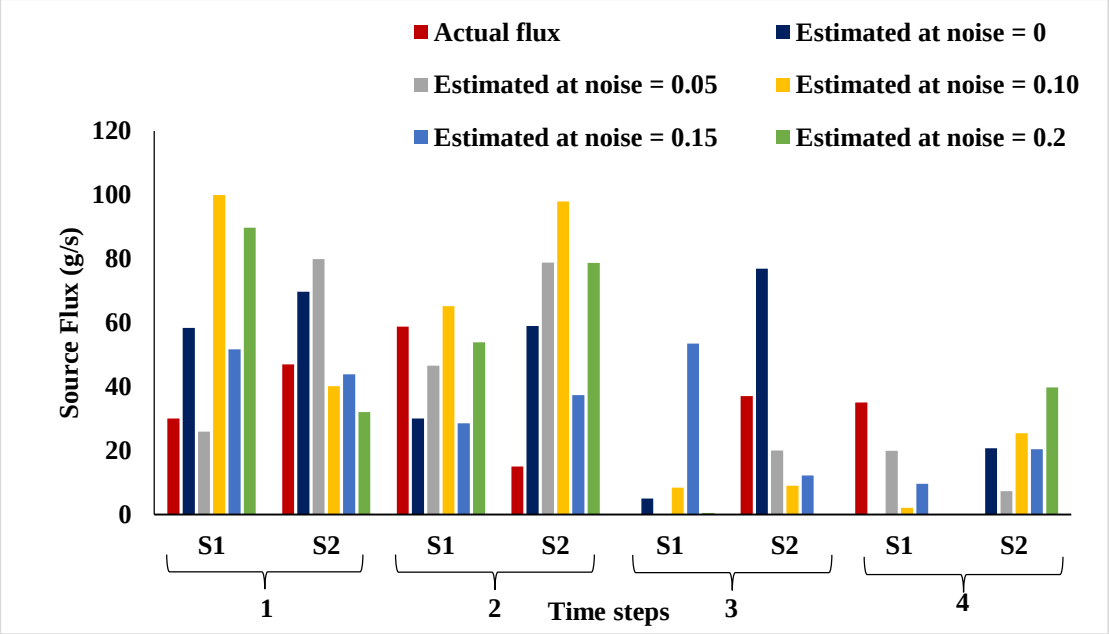


Fig. 4.6: Comparison of the source fluxes at different noise level for n = 2

4.7.2.2 Effect of noise level for n = 3

In case three number of pollution sources, the decision variables are also increased and a slight variation can be noted in the predicted source location. It is observed that at zero noise level, the source locations i.e. (3,3), (5,3) and (7,4) are exactly identified as observed in Table 4.11. The effect of noise levels σ = 0.05 and σ = 0.1 could not obstruct the present model in identifying the exact location (3,3) however, for other source locations (5,3) and (7,4), the impact of low noise levels compels the model to identify only nearby source locations. It is observed that the exact source locations are

not exactly identified for other noise levels, although it is seen that the estimated source locations (6,3), (4,4) and (1,4) are repeatedly identified by the model. It may be noted that they are very close to the actual locations. The estimated source fluxes of the identified source locations showed a large fluctuation with an increase in the noise level as seen in Table 4.12.

Table 4.11: Identified source location for n = 3 at different noise level

Number of pollution sources	**Actual pollution source location (*i*,*j*)**	**Estimated source location (*i*,*j*)**				
		At zero noise level	**At noise level σ = 0.05**	**At noise level σ = 0.1**	**At noise level σ = 0.15**	**At noise level σ = 0.2**
3	(3,3) (5,3) (7,4) (2,4)	(3,3)	(3,3)	(3,3)	(1,4)	(6,3)
		(5,3)	(6,3)	(1,4)	(6,3)	(3,3)
		(7,4)	(4,4)	(6,3)	(3,3)	(1,3)

Table 4.12: Estimated source fluxes for n = 3 at different noise level

Time Steps	**Source locations (*i*,*j*)**	**Actual source fluxes**	**Magnitude of estimated source fluxes (g/s)**				
			At zero noise level	**At noise level σ = 0.05**	**At noise level σ = 0.1**	**At noise level σ = 0.15**	**At noise level σ = 0.2**
1	S1	30.00	43.89	33.05	35.56	25.39	29.06
	S2	47.00	50.54	77.82	79.99	59.35	32.08
	S3	41.26	68.69	58.95	79.98	49.83	79.84
2	S1	58.80	72.49	57.83	62.68	67.41	69.81
	S2	15.00	28.01	22.71	19.30	65.92	78.67
	S3	0.00	47.95	34.06	64.35	52.24	52.08
3	S1	0.00	8.99	11.53	4.72	30.52	17.93
	S2	37.00	10.45	15.52	0.36	10.96	0.00
	S3	14.40	16.03	14.26	0.56	0.28	5.47
4	S1	35.00	20.22	27.34	34.28	10.05	20.53
	S2	0.00	0.07	0.03	0.07	6.48	39.74
	S3	16.88	8.05	18.45	0.67	11.89	0.06

It is also prominently depicted in Fig. 4.7 that none of the estimated source fluxes closely matches with the actual one. This shows the poor prediction capability of the model in identifying the source fluxes, under the influence of different noise levels.

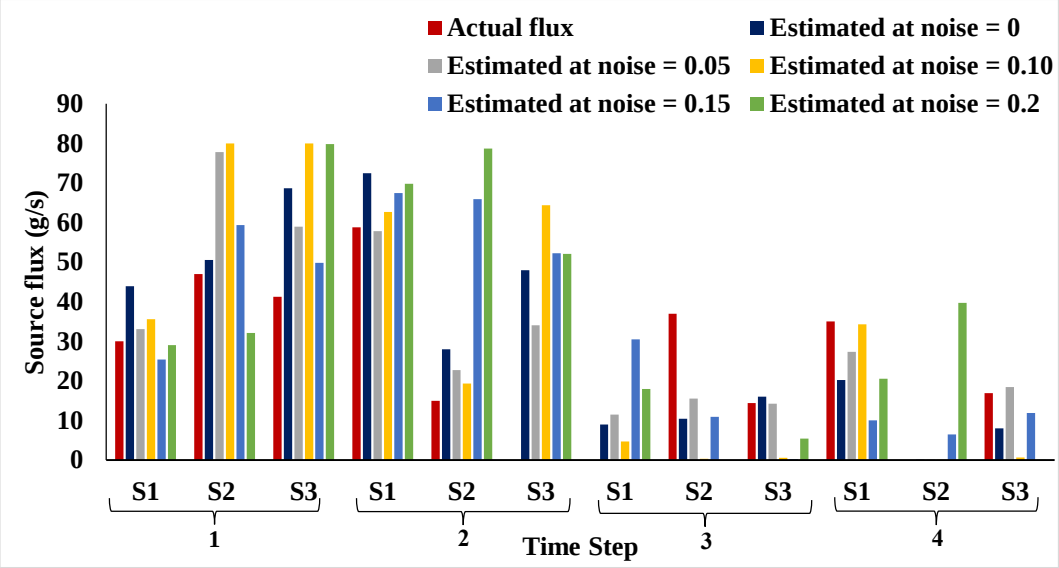


Fig. 4.7: Comparison of the source fluxes at different noise level for $n = 3$

4.7.2.3 Effect of noise level for n = 4

In case of four number of pollution sources, the present model was able to predict the exact pollution source locations i.e. (3,3), (5,3), (7,4) and (2,4) very precisely except for the location (5,3) for noise level $\sigma = 0.2$ (Table 4.13).

Table 4.13: Identified source location for $n = 4$ at different noise level

Number of pollution sources	Actual pollution source location (*i*,*j*)	Estimated source location (*i*,*j*)				
		At zero noise level	At noise level σ = 0.05	At noise level σ = 0.1	At noise level σ = 0.15	At noise level σ = 0.2
4	(3,3) (5,3) (7,4) (2,4)	(3,3)	(3,3)	(3,3)	(3,3)	(3,3)
		(5,3)	(5,3)	(5,3)	(5,3)	(6,3)
		(7,4)	(7,4)	(7,4)	(7,4)	(7,4)
		(2,4)	(2,4)	(2,4)	(2,4)	(2,4)

As there are four number of pollution sources and the impact of noise level on the exact number of pollution sources showed the least influence in identifying the source locations. This is so because the concentrations of the exact number of pollution sources exactly match with the observed concentrations. Hence, the different noise levels have less impact when the exact number of pollution sources has attained.

The present model effectively identified the source locations when the iterative search reaches the exact value of the number of pollution sources and has no impact with the addition of various noise level. However, a contrasting result is obtained in case of

source fluxes. At zero noise level, the estimated source fluxes closely resemble with the actual ones whereas some divergence is observed with the increase in noise level. Referring to Table 4.14, it is noted that the actual flux for S2 in the first-time step is 47 g/s, but the estimated source fluxes with noise level varied from 11.52 g/s to 59.99 g/s. It is also observed that for S3 source, the actual flux is 0 g/s at the second-time step, but after incorporation of the noise level, the estimated value of source flux even rose to 13.29 g/s, 34.54 g/s, 37.61 g/s and 17.47 g/s at 0.05, 0.10, 0.15 and 0.20 respectively.

Table 4.14: Estimated source fluxes for n = 4 at different noise level

Time Steps	**Source locations (*i*,*j*)**	**Actual source fluxes**	**Magnitude of estimated source fluxes (g/s)**				
			At zero noise level	**At noise level σ = 0.05**	**At noise level σ = 0.1**	**At noise level σ = 0.15**	**At noise level σ = 0.2**
1	S1	30.00	30.18	26.45	32.19	13.79	22.01
	S2	47.00	46.67	36.88	59.99	31.70	11.52
	S3	41.26	40.78	55.68	18.07	32.58	14.25
	S4	21.70	20.92	11.99	33.10	16.29	10.35
2	S1	58.80	57.91	42.28	42.74	44.59	18.24
	S2	15.00	16.61	19.99	16.81	25.98	21.44
	S3	0.00	0.14	13.29	34.54	37.61	17.47
	S4	0.00	2.16	14.33	1.67	15.93	4.87
3	S1	0.00	1.49	59.99	39.76	60.38	69.99
	S2	37.00	34.59	19.99	14.23	50.60	70.01
	S3	14.40	15.60	21.99	10.17	5.02	7.77
	S4	29.68	27.10	11.54	17.10	17.54	11.84
4	S1	35.00	33.98	39.99	23.02	45.64	59.97
	S2	0.00	1.185	11.88	1.85	7.85	68.06
	S3	16.88	15.94	28.60	27.67	32.03	33.38
	S4	0.00	1.41	0.06	0.16	39.34	0.037

The bar graph, showing the comparison between different source fluxes at different noise levels reflects the variation in fluxes with the increase in noise level (Fig. 4.8). It is observed that not much difference is seen in case of 0.05 noise level.

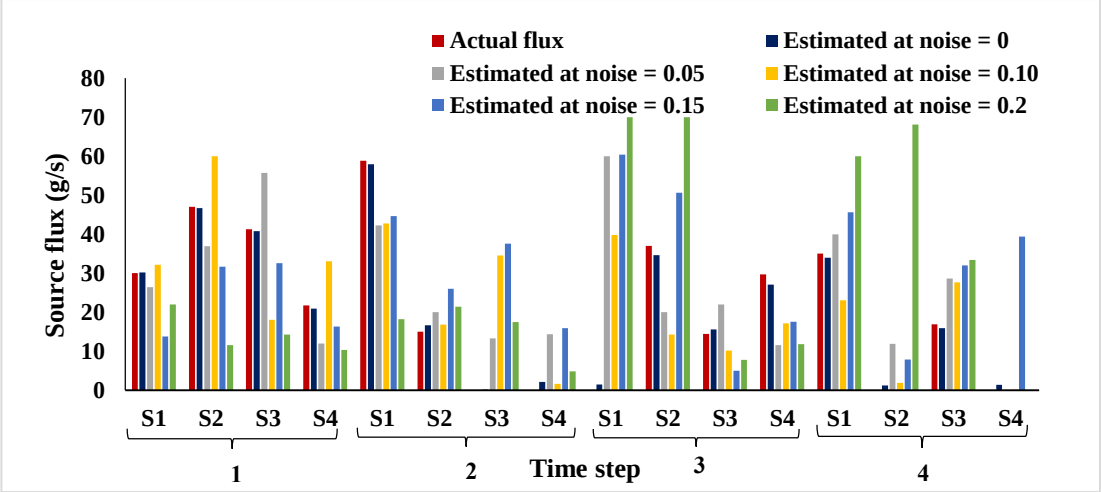


Fig. 4.8: Comparison of the source fluxes at different noise level for n = 4

4.7.2.4 Effect of noise level for n = 5

For n = 5, the identified source locations are found to be exactly same as the actual location at all the noise levels as observed in Table 4.15. It has been found that the 5th source is a dummy source and does not contribute any concentration to it. So, the 5thsource location identified by the model is a fictitious one. Furthermore, it can be concluded that the predicted source locations (1,10) and (1,11) has negligible plumes. Hence, the present simulation-optimization model has the potential of identifying the pollution sources at that location where the effect of plumes is very high.

Table 4.15: Identified source location for n = 5 at different noise level

Number of pollution sources	Actual pollution source location (*i,j*)	Estimated source location (*i,j*)				
		At zero noise level	At noise level σ = 0.05	At noise level σ = 0.1	At noise level σ = 0.15	At noise level σ = 0.2
5	(3,3) (5,3) (7,4) (2,4) (x)	(3,3)	3,3)	(3,3)	(3,3)	(3,3)
		(5,3)	(5,3)	(5,3)	(5,3)	(5,3)
		(7,4)	(7,4)	(7,4)	(7,4)	(7,4)
		(2,4)	(2,4)	(2,4)	(2,4)	(2,4)
		(1,10)	(1,11)	(1,10)	(1,11)	(1,11)

It has been discussed in previous sections that in the present iterative model, the dummy source plays an important role in deciding the exact number of pollution sources present in the aquifer. Moreover, the position of the source location identified by the model at different noise further confirms the non-availability of any pollution sources.

Table 4.16 shows the estimated source fluxes for five pollution source under different noise levels. At zero noise level, the predicted source fluxes showed some similarity with the actual fluxes however some contradictory flux values are also observed. The actual magnitude of source flux for S1, S2 and S3 at second and third-time steps are 58.80 g/s, 37 g/s and 14.40 g/s but the model predicted as 56.81g/s, 34.05 g/s and 9.29 g/s respectively. However, these differences do not show any major impact on the other source fluxes of the respective locations. S5 being a dummy source, a negligible amount of source fluxes is predicted by the model. The reason for the presence of some source fluxes in dummy source is due to the effect of contaminant concentration from other pollution sources on it. With the rise in noise level, the magnitudes of source fluxes in the dummy source gradually increase but still, it can be regarded as a dummy as lesser values of contaminant concentrations are obtained at all the time steps.

Table 4.16: Estimated source fluxes for n = 5 at different noise level

Time Steps	**Source locations (*i,j*)**	**Actual source fluxes**	**Magnitude of estimated source fluxes (g/s)**				
			At zero noise level	**At noise level σ = 0.05**	**At noise level σ = 0.1**	**At noise level σ = 0.15**	**At noise level σ = 0.2**
1	S1	30.00	30.58	30.75	28.41	27.79	29.82
	S2	47.00	46.36	45.61	46.83	45.35	43.19
	S3	41.26	41.16	40.62	43.32	39.65	37.59
	S4	21.70	21.70	24.62	24.33	24.78	27.29
	S5	x	3.82	1.17	3.20	7.83	7.42
2	S1	58.80	56.81	56.17	63.89	61.80	64.94
	S2	15.00	16.33	18.83	17.43	17.83	21.36
	S3	0.00	2.42	1.63	4.32	2.42	3.76
	S4	0.00	0.13	2.69	2.76	3.87	7.78
	S5	x	2.90	5.73	1.89	6.90	7.01
3	S1	0.00	2.03	4.18	6.19	3.02	5.82
	S2	37.00	34.05	32.92	29.92	29.34	23.05
	S3	14.40	9.29	13.63	6.73	9.85	14.56
	S4	29.68	29.23	29.16	25.91	19.76	20.90
	S5	x	4.38	1.39	3.60	5.37	10.01
4	S1	35.00	33.36	31.68	29.70	28.76	27.73
	S2	0.00	1.24	1.59	2.13	4.76	8.80
	S3	16.88	15.53	16.69	17.56	19.21	13.85
	S4	0.00	0.34	1.17	1.03	1.39	3.36
	S5	x	6.12	1.53	5.96	4.31	6.05

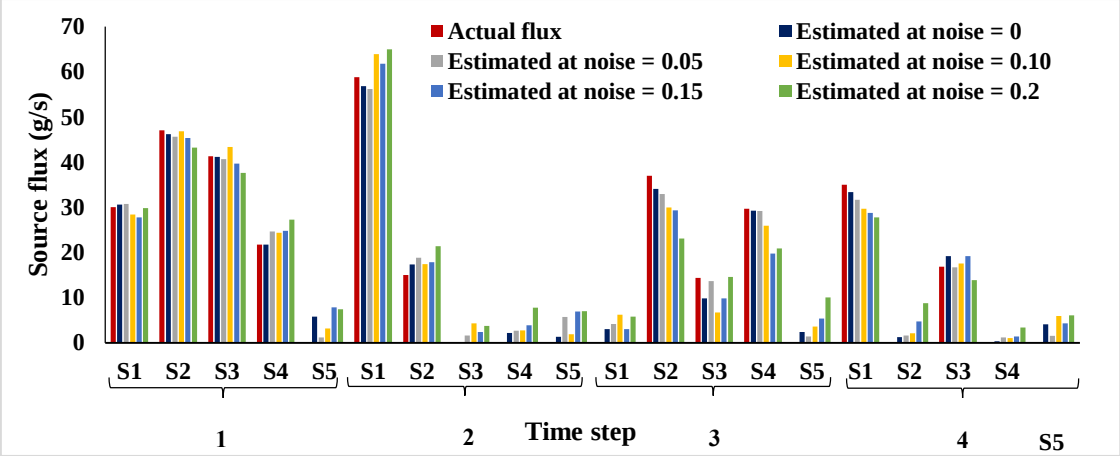


Fig. 4.9: Comparison of the source fluxes at different noise level for n = 5

Fig. 4.9 presents a comparison between the actual source fluxes and estimated source fluxes at different noise levels and shows a close match between the actual and the estimated source fluxes at zero noise level. But the fluctuation increases with the rise in noise level. These clearly signify that affect measurement error in the observed concentration greatly hampers the present methodology in effectively identifying the source fluxes under different noise levels.

4.7.3 Fitness function for different number of sources

The present section describes the effect of various noise level on the objective function values for each number of pollution sources. In the previous section, it has been explained that identification of pollution sources is carried out by minimizing the difference between the observed and the simulated concentration of the objective function. With the incorporation of the measurement error on the observed concentration, the chance of minimizing the error function between the observed and simulated concentration is reduced and hence the objective function rises with the noise level. For each number of pollution sources (n = 2, 3, 4 and 5), the objective function increases with the increase in noise level. Table 4.17 shows the evaluated value of fitness function at the different noise level. For n =2, a gradual rise in the objective function is observed with the rise in noise level ranging from 5.18 to 6.78. However, for n = 3, n = 4 and n = 5, the objective function values reduces and does not show much difference with the increase in noise level. For n = 3, the function varies from 3.94 at zero noise level and gradually rises to 2.98, 3.78, 3.91 and 4.34 at 0.05 0.10 0.15 and 0.20 respectively. The objective function value for n = 4 and n =5 are significantly reduced 0.072 and 0.020 respectively at zero noise level. With the rise in noise level,

the function value varies from 0.072 to 0.084 for n = 4 whereas for n =5 the function value varies from 0.020 to 0.075. Not much difference could be seen in the objective function for these sources. This is because at n = 4 and n = 5, the iterative based model has reached the exact number of pollution sources and most of the identified pollution sources match with the actual pollution sources.

Table 4.17: Fitness function values for different number of pollution sources at different noise level

Noise level σ	**Objective Function (*F*)**			
	For n =2	**For n = 3**	**For n= 4**	**For n =5**
0	5.38	3.94	0.072	0.020
0.05	5.86	2.98	0.058	0.013
0.10	6.17	3.78	0.034	0.045
0.15	6.76	3.91	0.065	0.084
0.20	6.78	4.34	0.084	0.075

Fig.4.10 presents the bar graph showing objective function values for the different number of pollution sources considering the different level of noises and its impact on the function value. The effect of the various noise levels (i.e. zero noise level, moderate noise level and high noise level) does not show much variation on the objective function value for each number of different number of pollution sources. It can be remarked from the bar graph that the final fitness function values are converging towards the best minimum value as it proceeds towards the actual number of pollution sources. The final fitness value with four and five number of pollutant sources almost matches with a minimum value of fitness function and thus confirming that no more pollution sources are available in the affected aquifer. This implies that the presence of noise in the observed data has a less significant impact in determining the optimal number of pollution sources.

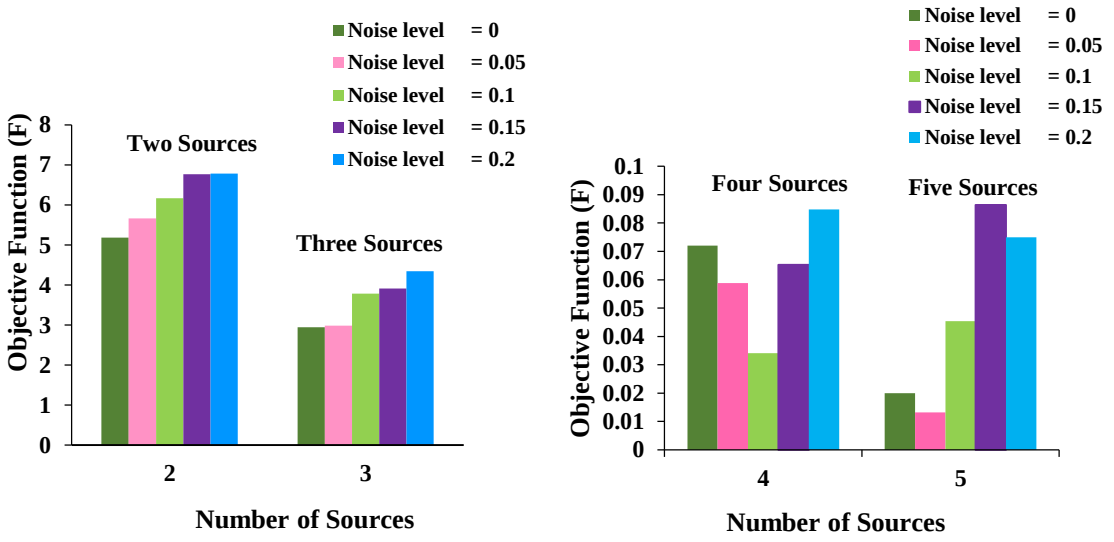


Fig. 4.10 (a) and 4.10 (b): Objective function values for the different number of pollution sources

4.8 Summary and Conclusions

The present methodology has been proposed considering the realistic scenario of groundwater pollution sources. It is assumed that no information is available about the pollution sources and hence an iterative search technique is adapted for identifying the exact number of pollution sources and the source locations. The identification of the unknown groundwater pollution sources is carried out using inverse optimization technique by minimizing the difference between the simulated and observed contaminant concentration. The simulated concentration is obtained using MODFLOW and MT3DMS which solves the groundwater flow and transport processes. The observed concentration is the concentration measured at the observation wells. To evaluate the proposed methodology, an illustrative study area has been taken. As no information is available regarding the number of pollution sources, the search for the optimal solution is started by initiating the simulation-optimization model with two number of pollution sources. The number of pollutant sources is subsequently increased until the best result is achieved. When the number of pollution sources reached four, the objective function reduced to a minimum value and it signified the convergence to the optimal solution. When the present model is further simulated with five number of sources, a dummy source is obtained with negligible concentrations. This confirms that

the present model has four number of sources and the 5th source is an unreal one. During the measurement of the observation concentration, some measurement errors always exist due to improper field measurement or the laboratory test (Singh and Datta, 2006). So, taking this into consideration, some noise is added to the observed concentrations. With each of these noise levels, the characteristics of the unknown pollution sources are identified using the present methodology. The location of the unknown pollution sources is efficiently identified when there is no measurement error (zero noise level) in the observed concentration. Even after the introduction of various level of noise, the locations are exactly identified. Speaking about the magnitude of source flux, a contrasting result could be seen. Even at the low noise level, the magnitude of identified source flux is immensely affected. So, it indicates that the magnitudes of source flux are most susceptible to give an erroneous result with a mere noise level. However, it has been observed that the model is capable of converging towards the location of the pollutant sources even at the high noise level. A contradictory result is obtained between the source location and the source flux after considering various noise levels. But it may be remembered that our concern is to determine the location of the sources which has been successfully achieved by the present approach. Therefore, the model proves to be quite a robust one in identifying the number of pollution sources and their location of the pollution source.

The next chapter will present a methodology on how source fluxes can be effectively identified along with the source locations. The source location being discrete variable can be effectively identified by GA whereas the source fluxes being continuous variables can be identified using classical optimization approach. Thus, a methodology is proposed combining these two approaches. This may be regarded a global-local search technique.

Chapter 5

Identification of Unknown Groundwater Pollution Sources using Hybrid Optimization Methodology

In the present chapter, the first section emphasised on the combined approach of GA and classical optimization technique in the source identification problem. The next section explains the methodology using the combinatorial approach for solving the groundwater source identification problem. The present section elaborately describes about three proposed model named as Genetic Algorithm-Longitudinal-Transverse Search-Gradient Search (GA-LTS-GR), Genetic algorithm-Mutation Search-Gradient Search (GA-MS-GR) and Genetic Algorithm Search-Ripple Migration Search-Gradient Search (GA-RMS-GR). The subsequent section further explains about the simulation model adopted in the present study followed by the study area. The last section explains the result obtained using the present methodology.

5.1 Introduction

In the earlier chapter, a methodology has been presented considering the real case of groundwater source identification problem where the number of pollution source locations and the source flux are completely unknown. The developed model could effectively identify the actual source numbers and their locations in the aquifer under various level of measurement errors. However, a contradictory result is observed in estimating the source fluxes. The model could not estimate the actual fluxes even under small measurement error. The Genetic Algorithms (GA) has been used to solve the optimization problem which has been emerged as one of the most effective approaches for solving non-linear non-convex optimization problems. The Genetic Algorithms can handle the problem having discontinuous function and integer variables effectively. The source location being discontinuous variable could be effectively identified by GA. However, it is not very efficient in determining the magnitude of the sources flux and may obtain the near optimal solution. As GA is efficient in determining the discrete variable, the source fluxes being continuous variables can be obtained by using gradient based classical optimization technique. As such some researchers have also combined

these two different approaches for solving groundwater problem having discontinuous variable and integer variable.

Considering the advantages of GA and the gradient based classical technique, a combinatorial methodology is proposed for effectively identifying the groundwater pollution source locations and the source fluxes. The review of literatures reveals that GA is efficient in obtaining the solution of the optimization problem having discrete variables. On the other hand, the gradient-based classical optimization algorithms are better for solving the problems with continuous variables. Another advantage of GA is that it has the potential to explore large search space and thus can explore the area having the global optimal solution of the problem. The solution obtained by using GA can then be further used as the initial points for the gradient search to get the exact optimal solution of the problem. Taking this into consideration, the source identification problem can be solved by using GA for identifying the source locations and the near optimal value of source fluxes. The gradient-based search technique can then be used for determining the exact value of the source fluxes. However, many a time it is not guaranteed that this combinatorial approach would yield the exact optimal solution. As such some local fine-tuning of the solution is necessary to obtain the optimal solution of the problem. In this study, three local location search algorithms have been proposed for fine tuning the solutions obtained using GA. The gradient based classical optimization technique is then applied along with the local location search algorithms to obtain the exact optimal value of the locations and source fluxes. The efficiency of the algorithms has been evaluated by solving the problem considered by Mahar and Datta (2001).

5.2 Methodology

As discussed earlier, the source identification problem is a mixed-integer problem. As such the Genetic Algorithms has been modified to handle the location variable and flux variable differently. As GA may not give the actual optimal solution, the solution obtained by using the modified GA has been used as the initial solution for local location search and gradient based search algorithm. Three local location search algorithms viz. Longitudinal-Transverse Search (LTS), Mutation Search (MS) and Ripple-Migration Search (RMS) have been proposed. As GA provides the initial solution for these three-local location search and gradient based search, the three-employed models are named as Genetic Algorithm-Longitudinal-Transverse Search-Gradient search (GA-LTS-GR), Genetic Algorithm-Mutation Search-Gradient search (GA-MS-GR) and Genetic

Algorithm-Ripple Migration Search-Gradient search (GA-RMS-GR). The present methodology is explained in the subsequent sections.

5.2.1 Source Identification Model

The earlier chapter have described how groundwater pollution sources can be identified using the linked simulation-optimization model. Here, the model minimizes the difference between the simulated concentration and the observed concentration at the observed well locations. The observed concentration is the concentration measured at different observation wells of the aquifer at different time steps. The simulated concentration is obtained by solving the groundwater flow and transport processes. The exact pollution source locations and fluxes will be identified when the observed contaminant concentration and the simulated concentration at the observed locations perfectly match with each other at different time steps. The objective function of the optimization model can be written as:

$$\text{Minimize } F = \sum_i^M \sum_j^N (C_{o,i}^j - C_{s,i}^j)^2 \tag{5.1}$$

Here, F is the objective function for the present optimization model with n number of pollution source; $C_{o,i}^j$is the observed concentration at j^{th} time step for i^{th} well location; $C_{s,i}^j$ is the simulated concentration at j^{th} time step for i^{th} well location; M is the total number of observation wells and; N is the total number of time steps.

This optimization model has been solved using the proposed Modified GA-Local Location Search-Gradient approach. The model is described in the following section.

5.2.2 Modified GA-Local Location-Gradient approach

The main objective of the optimization algorithms is to minimize the objective function using an algorithm that takes a minimum number of function evaluations. As stated earlier, the nature of the variables in source identification problem is mixed-integer type having integer variables for the locations and continuous variables for the source flux (Fig. 5.1). The prime reason of using GA in the identification of the source locations is its capability of determining the discrete variables which are the locations of the sources. The source fluxes being continuous variables may not be effectively identified by GA but can be successfully identified by using the gradient-based techniques, once the locations are known. As such, GA is initially used to obtain the source locations as well as the approximate value of the source fluxes.

The algorithm proposed to solve the problem is explained in detail in Fig. 5.2. The model is started by randomly generating the initial population. Then it calls the simulation model for obtaining the simulated concentrations. In the next step, the fitness function is evaluated on the basis of the stall location tolerance. If the tolerance is not satisfied, then the population goes through a set of modified genetic operators for improving the population. If the tolerance is reached, the solution is obtained using the modified GA. The solution from modified GA then goes through the gradient-based search and Local location search algorithms until the termination criterion is satisfied and optimal solution is reached. The present methodology focusses on three main objectives. The first objective is to obtain the actual source location and the approximate source strength using modified GA. The genetic algorithm is modified to handle both the types of variables differently, the location variables are handled using binary coded crossover and mutation operators. The flux variables are handled by using real coded crossover and mutation operators. The termination criterion of the modified genetic algorithm is based on stall location. Further, the flux obtained from modified GA is fine-tuned using gradient-based search at the same location as obtained by modified GA. At this solution, it is sure that the locations are very near to the optimal location of the sources.

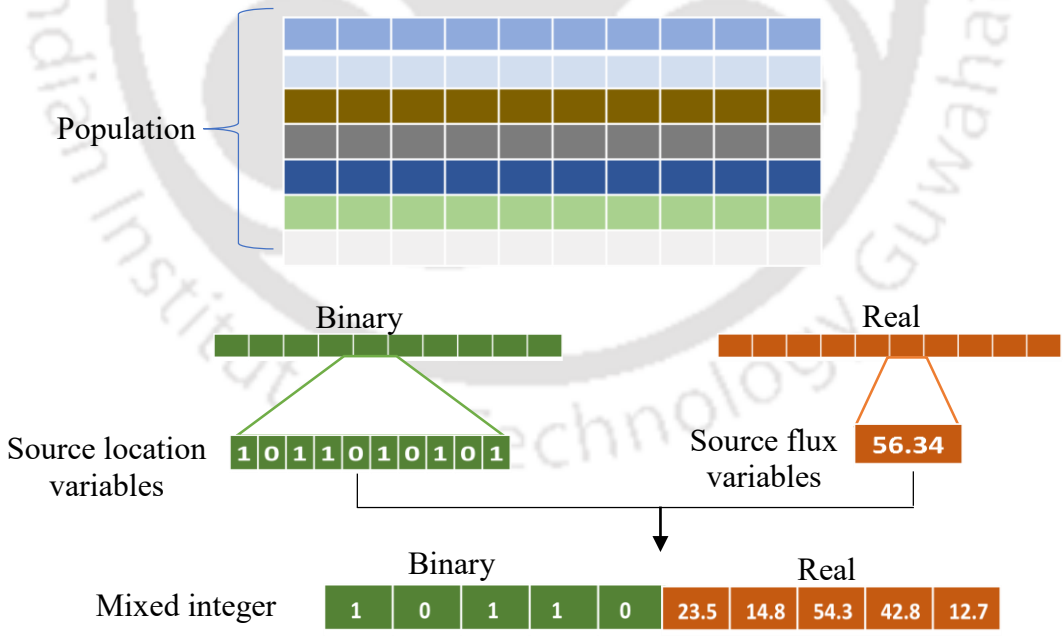


Fig. 5.1: Source identification problem showing mixed integer variables

Now, the second objective have been set up to obtain the actual source locations using local location search. Here the algorithm proceeds to local location search for which three different methodologies have been proposed to improvise the location to an optimal

solution. The local location methodologies proposed here are Longitudinal-Transverse Search (LTS), Mutation Search (MS) and Ripple Migration Search (RMS). At each of the improved location, the gradient-based search is performed to find the optimal value of the source fluxes. The termination criterion for this case is based on the first-order optimality. After termination from local location search, the third and final objective is to obtain the actual source locations with exact value of source fluxes using the gradient based search. Thus, the algorithm is called Modified GA-Local Location-Gradient Based Algorithm.

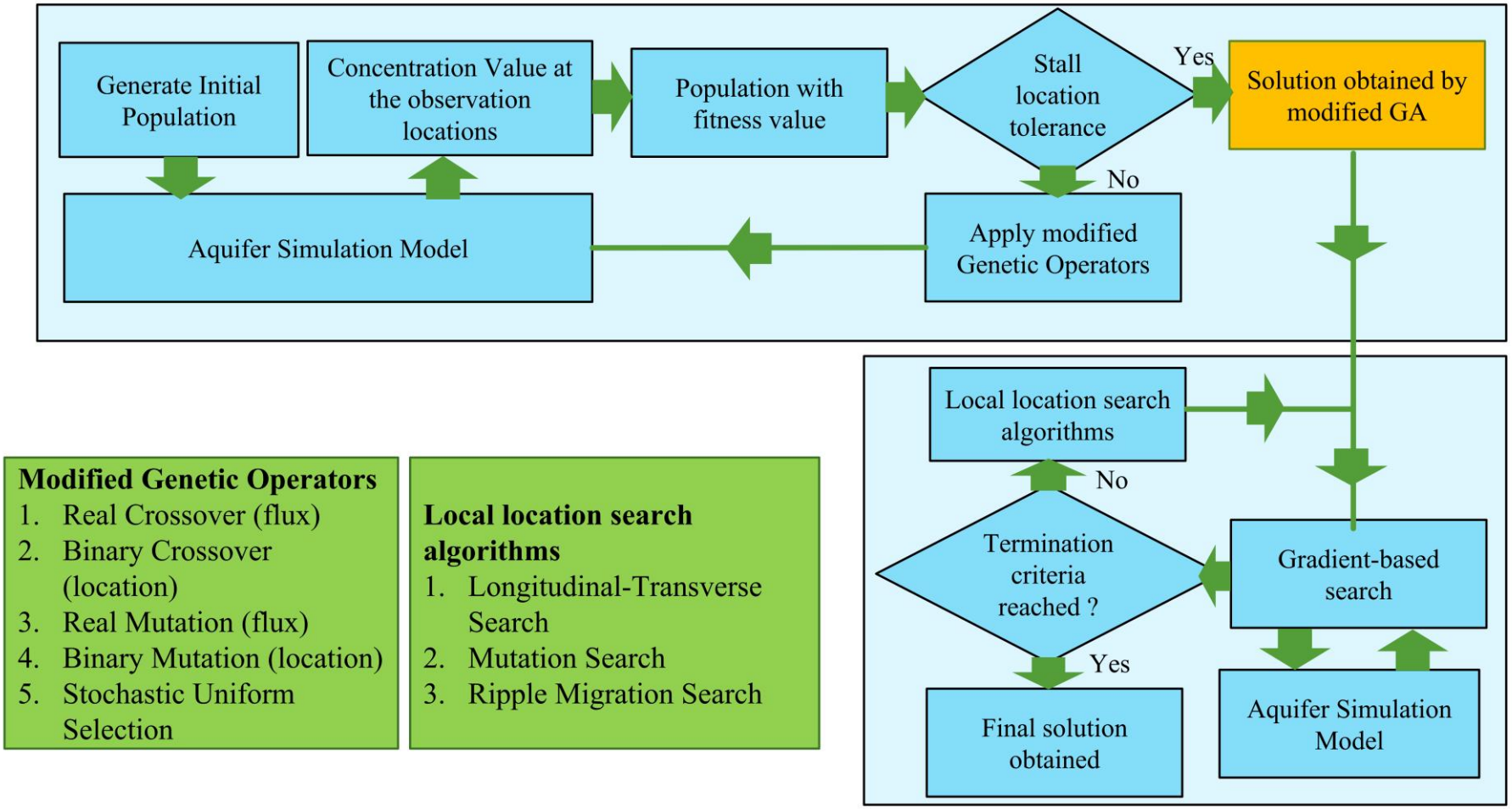


Fig. 5.2: Overview of the Modified-GA-Local-Location-Gradient based algorithm

5.2.2.1 Modified GA

The aim of this modified genetic algorithm is to obtain the exact location of the pollution sources. Here, the integer variable and continuous variable are handled differently. For the integer part of the string, the binary coded crossover and mutation operators have been used. On the other hand, for the continuous part, real coded crossover and mutation operators have been used. It has been observed that many a time GA does not provide the exact locations, rather it gives a solution very close to the optimal location. Thus, the solution obtained by the modified GA can be taken as initial guess to local location search algorithm. As such, the algorithm is designed in such a way that it provides a solution near to the global optimal solution and a tolerance is introduced based on the stall location. Thus, if the location is not changing for a specific number of iterations (30 in this case), the modified GA will terminate. The other modifications to GA are done in such a way that it handles the source identification problem efficiently and gives a solution as per the aim in this step. Fig. 5.3 describes the algorithm in a flowchart.

The modified GA is started with randomly generated initial solutions called the population. The selection operator is the first step where the chance of fitter individual to be selected is high. Selection of individuals from the population is based on fitness and can be performed using various algorithms such as roulette wheel, stochastic uniform selection, tournament selection etc. (Goldberg et al., 1992). In the present case, stochastic uniform selection has been adopted. the mutation is increased undergoes crossover operator which results in producing an offspring with combined traits of its parents. There is a possibility that some good traits may have been lost in the selection process, so elitism operator has been applied to retain the best individuals. During the search process, candidate solution of the population will be improved with the application of various genetic operators such as selection, binary crossover and binary mutation for the location variables. The real crossover and real mutation are applied to the continuous variable (flux variables) portion of the chromosome.

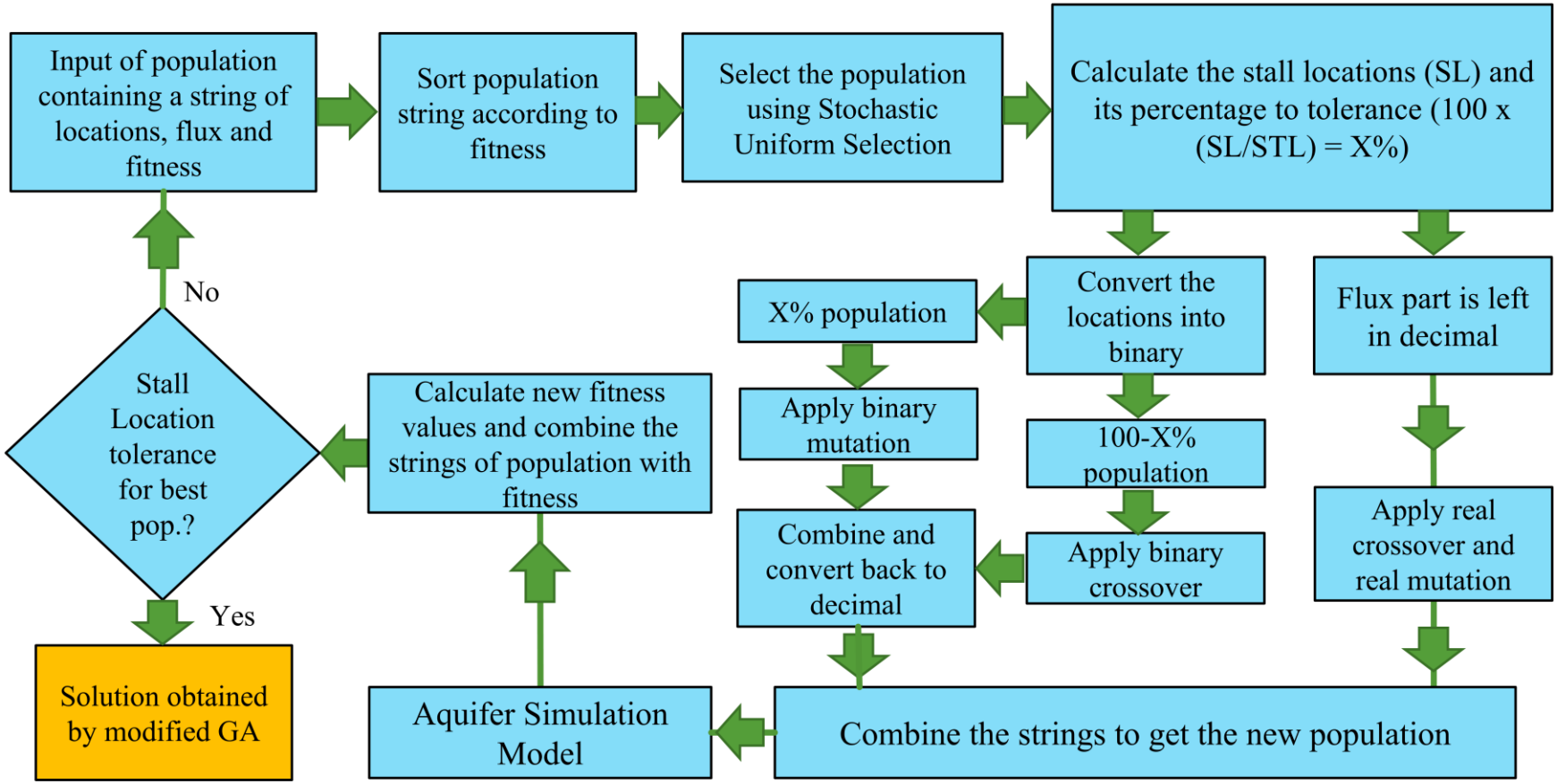


Fig. 5.3: Flowchart describing the modified GA algorithm

For the process of binary crossover, a random crossover site is generated and the selection of genes from parent 1 (P_1) and parent 2 (P_2) is performed based on the positions of crossover site. This is called the scattered crossover. Further, the mutation is performed randomly on any position of the binary string. However, it may be noted that ultimately the binary outputs are to be converted back into decimal terms. The source flux being continuous variables would be preferred with real coded technique. So, the implementation of the genetic operators will be on real values. Real crossover and real mutation are carried out for improving the candidate solution of the source flux. The real crossover is carried out using simulated binary crossover based on the function described by Deb and Agarwal (1995). The real mutations are carried out using polynomial mutation based on the functions described by Deb and Goyal (1996). These complete steps of the genetic operator will constitute one generation or iteration. As fitter individuals have the chance of being selected, the frequency of preserving good traits increases with additional generations.

In normal GA with successive generation, the diversity of population decreases and the population will converge towards the solution (Mahinthakumar and Sayeed, 2005). But the present modified GA uses the convergence only for the flux variables, i.e. the population of the flux variables converges and the population of the location variables diverges as the generation increases. This is due to a very specific reason to obtain the global optimal location which is explained in Fig. 5.4.

If the location variable of the constrained problem does not change in the successive generations, there are two possibilities. One of the possibility is that the identified locations are the optimal or near optimal locations, or the solution is local optimal solution. To confirm whether the locations are the global optimal locations, the mutation for the locations increases corresponding to the stall location level. This usually diverges the population more. Then if the solution is a near optimal solution globally, the best will not change and the same stall location continues. But, if it is a local optimal solution, the best will change accordingly towards the optimal locations and the stall location will be reset to zero. This continues until it reaches the global or near global optimal locations. Once the given stall locations are satisfied, the modified GA algorithm terminates and the classical optimization and the local location search algorithms are applied to fine tune the solution.

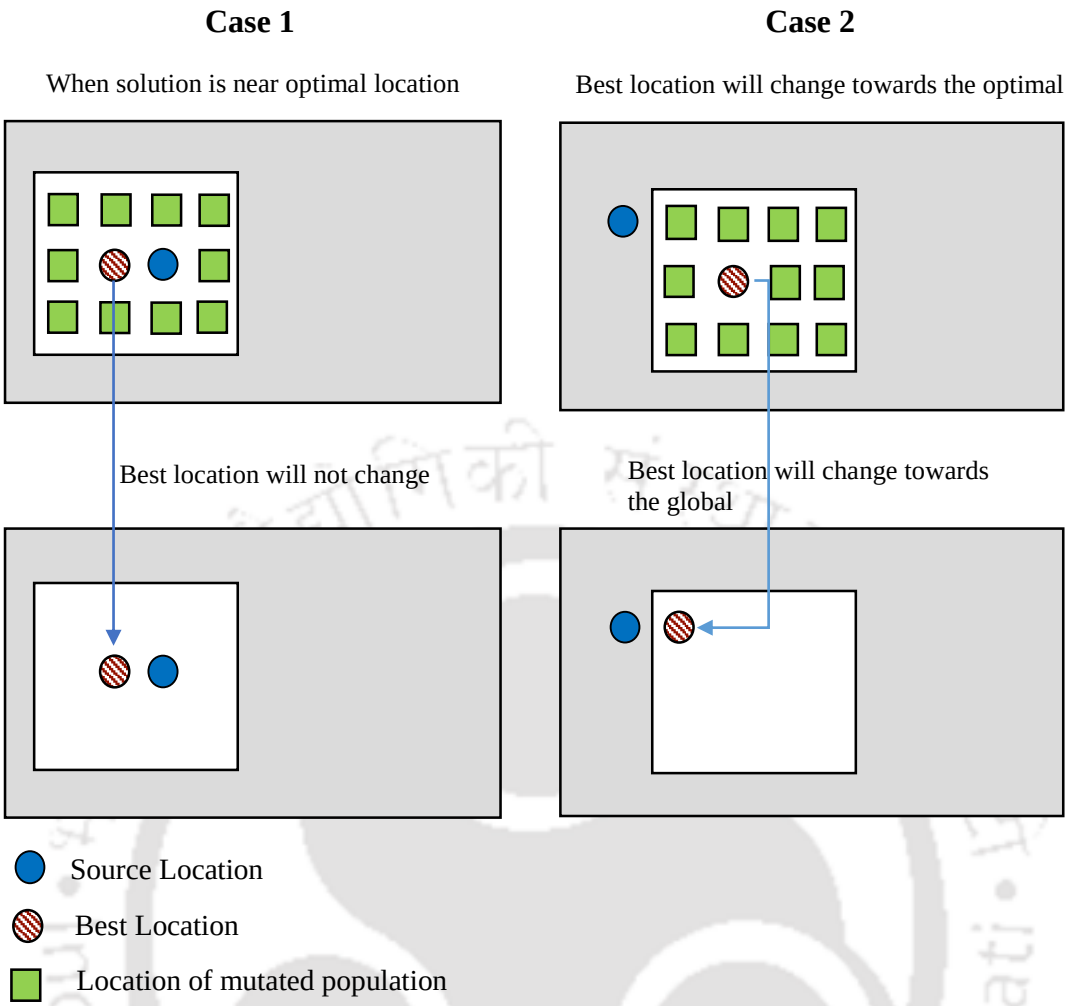


Fig. 5.4: Cases for increasing the mutation as the flux converges

5.2.2.2 Local-Location search

It has been reported that the combination of global-local optimization approach has proved to be a better alternative as compared to the stand-alone optimization approaches (Mahinthakumar and Sayeed, 2005). However, it is not always guaranteed that this approach will give the best solution with lesser number of function evaluations. So, for further fine-tuning the solutions and to reduce the computationally expensive function evaluation, an additional algorithm has been introduced. Here, the pollution source location identified by the modified GA and the source flux obtained from the classical are the inputs to the local location search algorithm. As modified GA tends to give global solution, the best source locations are bound to be around the locations obtained by the modified GA.

Now the search space has drastically reduced and hence becomes much easier for the Local-Location search to obtain the global optimal solution. In order to skip all the

expensive fitness function evaluations, the present Local-Location Search along with gradient based optimization method will only utilize the best global solution obtained by the modified GA and neglect the rest of the population. With the advantage that the worst individual has been discarded and the remaining being the near-best solution, the concept of Local-Location search can be applied effectively. Three different algorithms namely Longitudinal-Transverse Search (LTS), Mutation Search (MS) and Ripple Search (RS) used for local location search are described in the subsequent sections.

5.2.2.2.1 Longitudinal-Transverse Search (LTS)

In groundwater contaminant transport, advection is more dominant over dispersion for the contaminant to get transported. This advection has a natural direction according to the groundwater flow. The dominant direction is taken as the transverse direction and the orthogonal direction to that is the longitudinal direction. Taking into consideration the effect of longitudinal and transverse directional variations in the function value, this reasoning of this algorithm has been introduced.

Let the location identified from modified GA is *i,j* and this being the best solution will be always be around the exact location. For this reason, travelling one step location at a time will not require large function evaluation. Initially, the longitudinal location search takes place with the option that it can either go in (*i, j+1*) or (*i, j-1*) direction. If the identified location tends towards (*i, j+1),* it will proceed towards the actual location. As the identified location has been upgraded to a new location, the source flux also gets updated from this longitudinal location.

In the next step, the local location search being transverse tries to travel either in (*i+1, j)* or (*i-1, j)* direction.

$$i,j = \begin{cases} i, j+1 \text{ or } \mathrm{i}, j-1, & \text{for longitudinal location search} \\ i+1, j \text{ or } i-1, \mathrm{j}, & \text{for transverse location search} \end{cases} \tag{5.2}$$

Each of the solutions is checked for the termination criteria. All the combination of sources in longitudinal direction are checked first. The termination of this search is based on first order optimality tolerance (10^{-1}). If the tolerance is satisfied, then exact pollution source location and the flux closer to actual are identified. If not the whole search cycle continues with the modified location as the input until the exact location is identified.

Fig. 5.5 shows the steps involved in LTS algorithm. For a single pollution source, the location search will be in either of the transverse and longitudinal direction. If there are

three number of pollution sources a total number of six combinations will be formed for the alternative transverse-longitudinal search. It may be noted that once the actual location is identified the source flux is also modified simultaneously. The schematic representation of LTS algorithm is shown in Fig. 5.6.

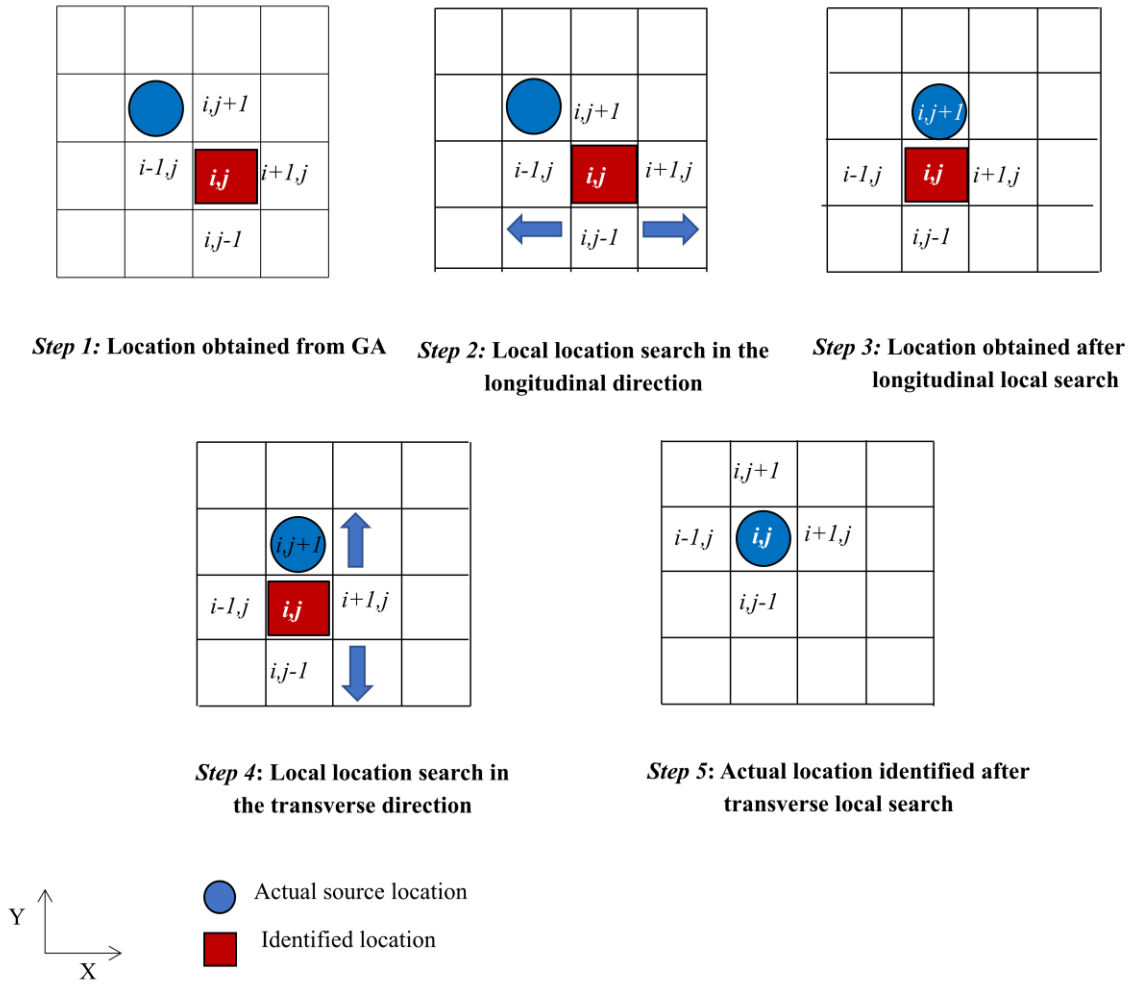


Fig. 5.5: Steps for determining the exact location in LTS algorithm

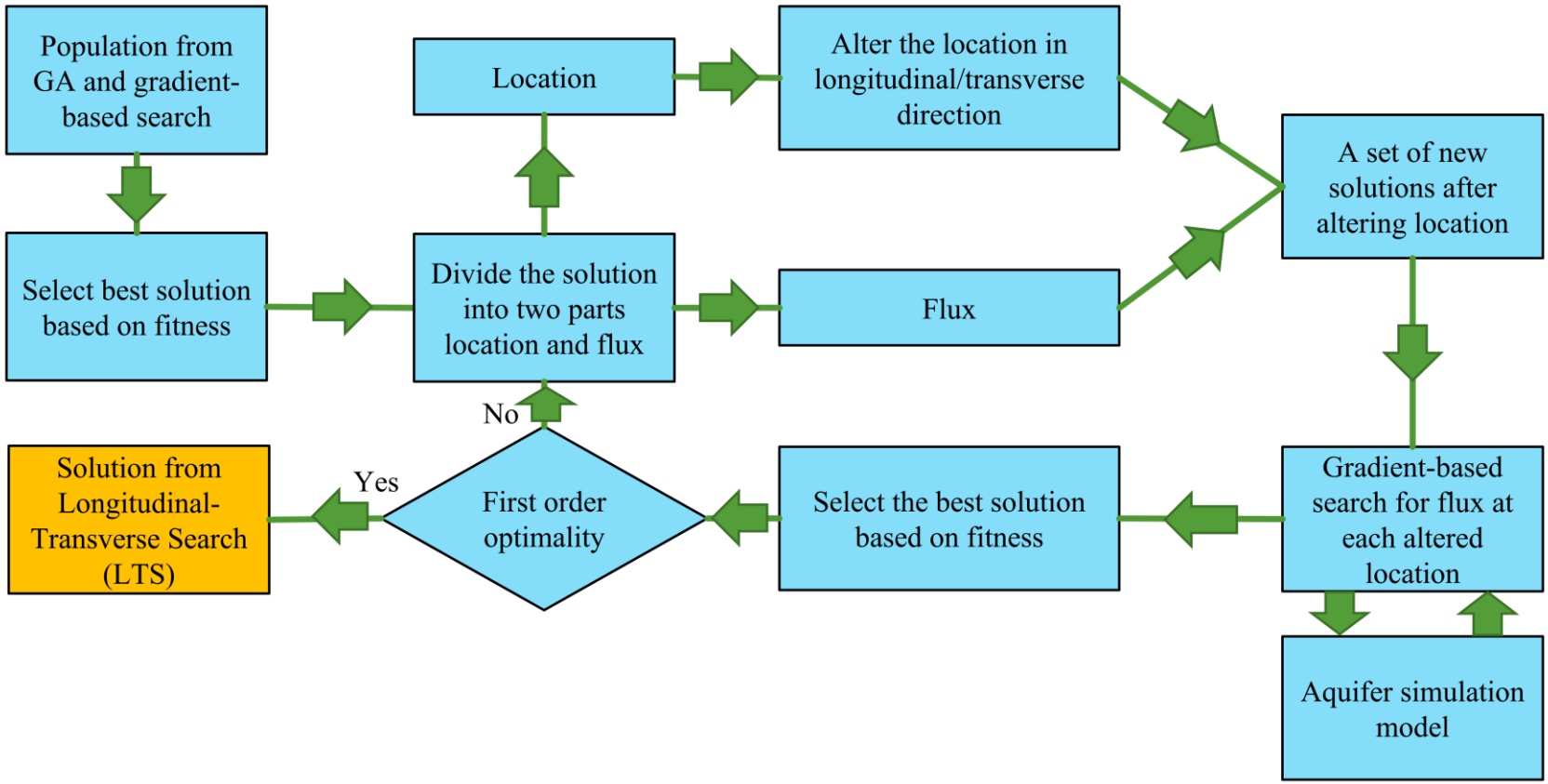


Fig. 5.6: Flowchart describing the LTS algorithm

5.2.2.2.2 Mutation Search (MS)

Mutation can be considered as one of the genetic operators which will maintain diversity and has the capability of even recover the genetic material which might be lost in the selection and crossover phase (Goldberg, 1989). Introduction of mutation will prevent the population from becoming too similar which may increase the function evaluation or even result in stopping the evaluation. But the level of mutation probability applied should be very low (Goldberg, 1989), preferably in the range of 0.001 to 0.01. If it is high it might give unrefined random search. So, there is a possibility that use of mutation may yield better solution.

In the present problem, the best fit population obtained by modified GA if not converge to the best solution will require further alteration. The location obtained by GA and the source flux from classical method are now served as the input. Here, five sets of exact copies of the string comprising of location and source flux will be adopted as shown in Fig. 5.7 (a). The fitness function F_1, $F_{2\ldots}$, F_5 will be evaluated for these set of population strings. Now randomly select three copies from the population (Fig. 5.7 (b)). Mutation will be performed on these locations (Cell ID equal to 30, 24 and 17) with low and high variance on the basis whether the generation is odd or even respectively (Fig. 5.7 (c)).

30	24	17	0	11.69	40.92	11.13	32.50	3.58	2.25	22.50	0	13.19	11.34	47.39
30	24	17	0	11.69	40.92	11.13	32.50	3.58	2.25	22.50	0	13.19	11.34	47.39
30	24	17	0	11.69	40.92	11.13	32.50	3.58	2.25	22.50	0	13.19	11.34	47.39
30	24	17	0	11.69	40.92	11.13	32.50	3.58	2.25	22.50	0	13.19	11.34	47.39
30	24	17	0	11.69	40.92	11.13	32.50	3.58	2.25	22.50	0	13.19	11.34	47.39

Fig. 5.7 (a): Five sets of exact copies of the string

30	24	17	0	11.69	40.92	11.13	32.50	3.58	2.25	22.50	0	13.19	11.34	47.39
30	24	17	0	11.69	40.92	11.13	32.50	3.58	2.25	22.50	0	13.19	11.34	47.39
30	24	17	0	11.69	40.92	11.13	32.50	3.58	2.25	22.50	0	13.19	11.34	47.39

Fig. 5.7 (b): Randomly selected copies from the string

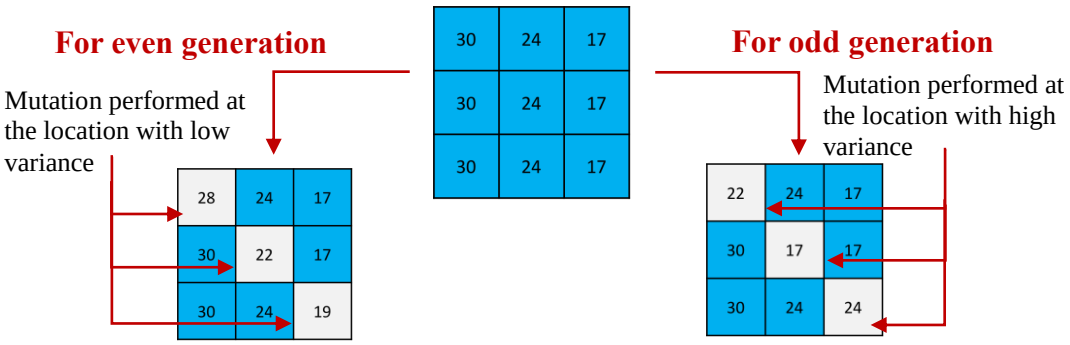


Fig. 5.7 (c): Mutation performed at the selected locations

In an even generation, mutation will be performed on the location of the randomly selected location with a factor 1 or 2 (for the case presented) corresponding to $(i \pm 1, j)$ location. For example, as shown in Fig. 5.7 (c), when the generation is even, mutation will be performed on each location by introducing a factor of 1or 2 on the selected locations. The source locations thus become 28, 22 and 19 which is converging towards the actual locations. If generation is odd similarly mutation will be performed by a factor of 7,8 or 9 (for the case presented) corresponding to $(i, j \pm 1)$ locations. Ultimately the source locations converge to 22,17 and 24. The whole process is also shown in Eq. 5 as

$$\text{Mutation at } x_{ij} = \begin{cases} x_{i \pm 1, j} \text{ with factor 1 or 2, if generation is even} \\ x_{i, j \pm 1} \text{ with a factor of 7, 8 or 9, if generation is odd} \end{cases} \tag{5.3}$$

It may be noted that mutation has been carried out on the location only and the source flux remains the same. With the altered strings, classical optimization is performed and function values are evaluated (*f1*, *f2* and *f3*) with the mutated locations (x_m). A comparison is made between the function values obtained with the mutated locations and the locations using GA-Classical search. If the function values are found to be improved ($f_i \leq F_i$), then it will replace the original locations in the string otherwise the whole steps would be repeated by randomly generating the populations again (Eq. 5.4). The complete procedure involved in MS approach has also been shown in Fig. 5.7.

$$x_{ij} = \begin{cases} x_m, \text{ if } f1, f2, f3 < F1, F2, F3 \\ x_o, \text{ otherwise} \end{cases} \tag{5.4}$$

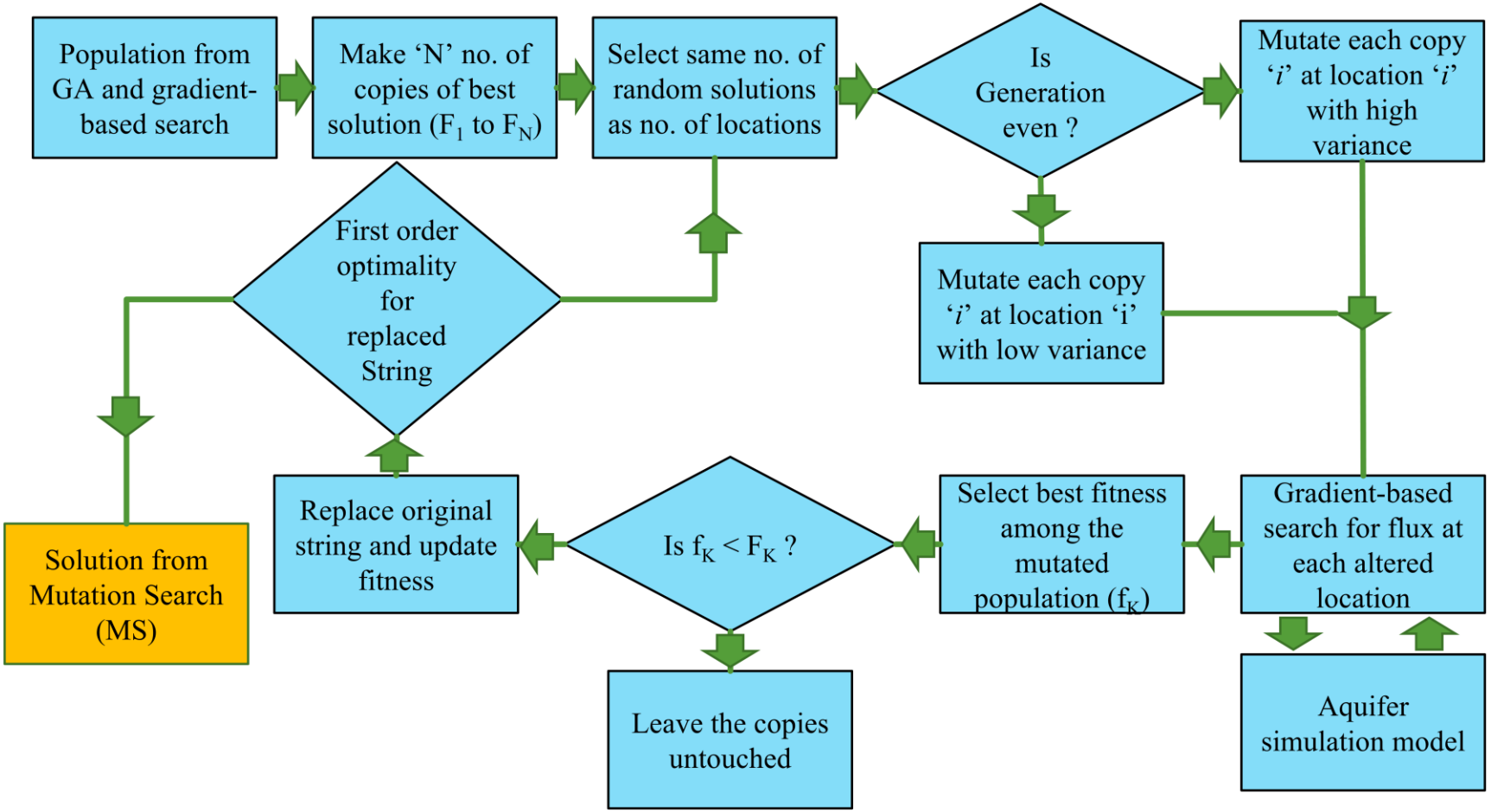


Fig. 5.8: Flowchart describing the MS algorithm

5.2.2.2.3 Ripple-Migration Search (RMS)

In this algorithm, the solutions will migrate towards the best source location which resembles the effect of ripple formation when a stone is dropped in the middle of a pond. Based on this theory, the algorithm is named Ripple-Migration Search. The input parameters are the location from the modified GA and source flux obtained from classical search. In order to modify the population sets, locations can be generated randomly using a random number generation function. These randomly generated locations will be around the actual location.

For example, if a total number of 20 populations are there, 19 of them will be randomly generated except the best location obtained using modified GA. The source flux remains the same as obtained using classical. The fitness function, $F_1, F_2 \ldots F_{20}$ for each of the population is calculated and sorted in ascending order as per the function value. Now from these sorted population, 3 locations are selected randomly (as the number of pollution sources are assumed to be three). The string with the best solution ($i_{b1}, j_{b1}, i_{b2}, j_{b2}$ and i_{b3}, j_{b3}) will be around the exact location as compared with the remaining 3 population strings surrounding as seen in Fig. 5.9.

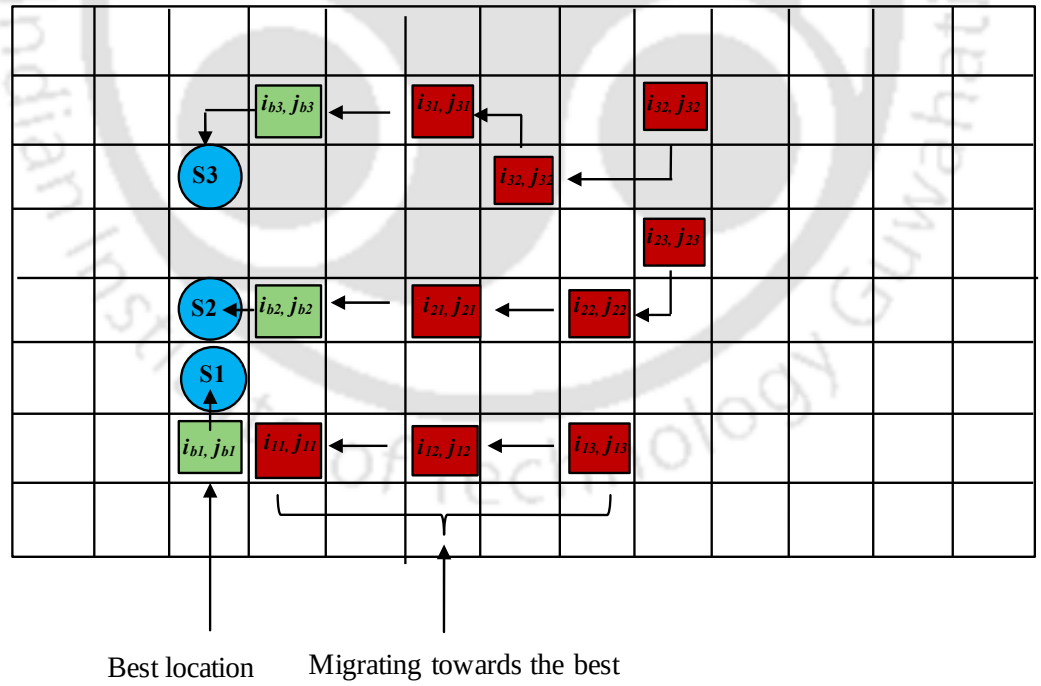


Fig. 5.9: Randomly selected location migrating towards best locations

Now the next successive step is to shift the best population string towards the best one. This can be performed with the condition that if $i_{11} < i_{b1}$ then the location would proceed for $i_{11}+1$ and if $i_{11} > i_{b1}$ then the location would become $i_{11}-1$. Similarly, for $j_{11,}$ the same algorithm is followed. Again, if $(i_{11,} j_{11}) = (i_{b1,} j_{b1})$, the location remain same as it is. Here, i_{b1} and j_{b1} are the best location and i_{11} and j_{11} are the randomly selected location.

As there are 3 pollution source locations, the migration towards the best location is repeated for all the combinations. With the migrated locations, the classical optimization with a first order optimality tolerance (10^{-1}) would be carried out and the function value f_1, f_2, f_3 for each of these populations are being evaluated. Now if the function values $(f_i \leq F_i)$ then the original location will be replaced by the migrated locations, if not the original locations will be retained. The schematic representation of RMS is shown in Fig. 5.10.

5.2.2.3 Classical (gradient-based) optimization

The solution obtained using modified GA or from the local location search will serve as the starting points for the gradient search. The search approach adopted to find the exact source fluxes is the gradient descent method. The algorithm is performed using *fmincon* function available in MATLAB. In this case the tolerance criterion is the step length, which is equal to 10^{-6}. When the step length is lesser than this, the optimization terminates. The tolerance for optimality is customized based on two levels.

If the algorithm is called from the local location search, then the first order optimality is increased to be coarser at 10^{-1} tolerance. This is because, there is no need to spend much time on the locations which are not optimal locations. This can be easily known at a first order optimality tolerance. If the first order optimality is satisfied and at the same time the function value is comparatively much lesser than that of the other locations, it can be taken as the best location for that particular iteration. If the algorithm is called after the local location search, then the final locations have already been obtained. This means that only the flux value needs to be modified. Thus, a finer tolerance is used for first order optimality and function tolerance. Both of them can be left as default at 10^{-6}, or can be further customised much finer to get the accurate flux values.

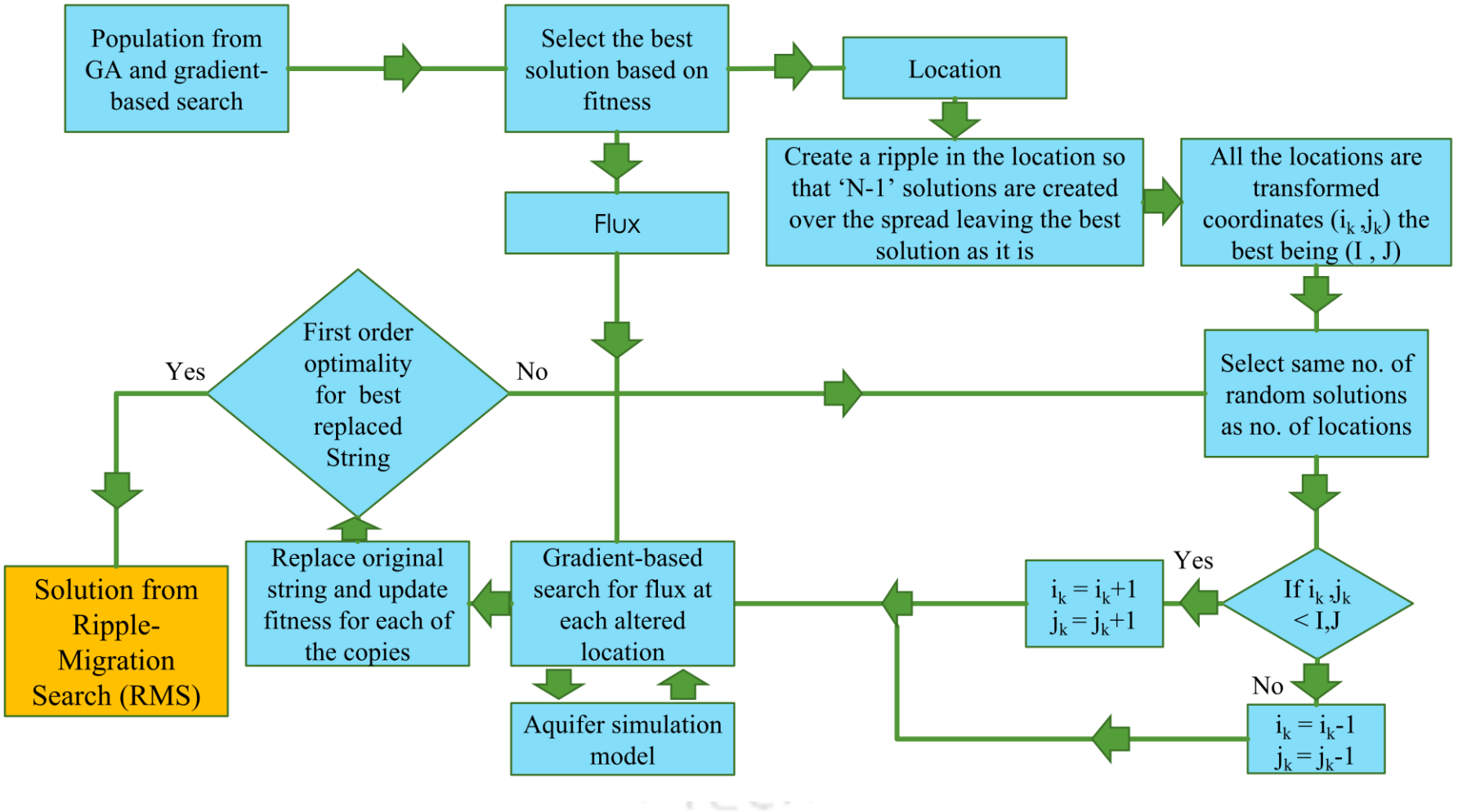


Fig. 5.10: Flowchart describing the RMS algorithm

5.3 Simulation model

The groundwater flow process has been simulated using MODFLOW which solves the groundwater flow equation using the finite difference technique. The governing equation to simulate the groundwater flow process in MODFLOW can be written as

$$\frac{\partial}{\partial x_i}\left(K_{ij}\frac{\partial \phi}{\partial x_j}\right) = S_s\frac{\partial \phi}{\partial t} \tag{5.5}$$

Where, K_{ij} is the hydraulic conductivity tensor (LT^{-1}); ϕ is the hydraulic head (L); S_s is the specific storage coefficient; t is the time (T). Using the equation (5.5) flow field for the aquifer is evaluated. The hydraulic head (ϕ) will be further implemented for calculating the velocity in the groundwater flow (v) in Darcy's law as

$$v_i = -\frac{K_{ij}}{\theta}\frac{\partial \phi}{\partial x_j} \tag{5.6}$$

Here, v is the velocity vector (LT^{-1}; and θ is the porosity for the porous media. After evaluating the average velocity for the entire aquifer, it will be used in simulating the groundwater transport process.

Groundwater transport equation is a very complex process. It is mainly governed by advection, dispersion and diffusion. A MATLAB code is written to simulate the groundwater transport equation following the standard finite difference scheme. The three-dimensional contaminant transports in groundwater (Bear, 1979) can be expressed as

$$\frac{\partial(\theta C)}{\partial t} = \frac{\partial}{\partial x_i}\left(\theta D_{ij}\frac{\partial C}{\partial x_j} - \theta v_i C\right) + q_s C_s - q'_s C \tag{5.7}$$

Where, C is the dissolved concentration in the groundwater (ML^{-3}); θ is the porosity of the subsurface medium; t is the time(T); x_i is the distance along the respective Cartesian co-ordinate axis (L); q_s is the volumetric flow rate per unit volume of aquifer representing fluid sources (positive) and sinks (negative) (T^{-1}); $q_{s'}$ is the rate of change in transient groundwater storage (T^{-1}); C_s is the concentration of the source or sinks flux (ML^{-3}). D_{ij} is the hydrodynamic dispersion coefficient tensor (L^2T^{-1}) which can be evaluated as (Burnett and Frind, 1987).

$$D_{ij} = \begin{pmatrix} \alpha_L \frac{v_x^2}{|v|} + \alpha_{TH} \frac{v_y^2}{|v|} + \alpha_{TV} \frac{v_z^2}{|v|} & (\alpha_L - \alpha_{TH}) \frac{v_x v_y}{|v|} & (\alpha_L - \alpha_{TV}) \frac{v_x v_z}{|v|} \\ (\alpha_L - \alpha_{TH}) \frac{v_x v_y}{|v|} & \alpha_{TH} \frac{v_x^2}{|v|} + \alpha_L \frac{v_y^2}{|v|} + \alpha_{TV} \frac{v_z^2}{|v|} & (\alpha_L - \alpha_{TV}) \frac{v_y v_z}{|v|} \\ (\alpha_L - \alpha_{TV}) \frac{v_x v_z}{|v|} & (\alpha_L - \alpha_{TV}) \frac{v_y v_z}{|v|} & \alpha_{TV} \frac{v_x^2}{|v|} + \alpha_{TV} \frac{v_y^2}{|v|} + \alpha_L \frac{v_z^2}{|v|} \end{pmatrix} \quad (5.8)$$

v_i is the seepage or linear pore water velocity (LT^{-1}); Also, the magnitude of velocity vector can be expressed as $|v| = \sqrt{v_i^2 + v_j^2}$; α_L, α_{TH} and α_{TV} are the longitudinal, transverse dispersivity horizontal and transverse dispersivity vertical respectively.

5.4 Study Area

For evaluating the proposed methodology, the well-known hypothetical study area (Fig.5.11) adopted by Mahar and Datta (2001) is taken. The aquifer has an area of 1.04 km^2. It is a homogeneous and isotropic aquifer with constant head boundary on the left and right-hand boundaries and no flow on the upper and lower boundaries of the aquifer. There are 13 x 8 columns and rows in the study area and is discretized into 100m x 100m. A total number of 8 observation wells are placed sparsely. The hydrological parameters adopted in the study area are shown in Table 5.1.

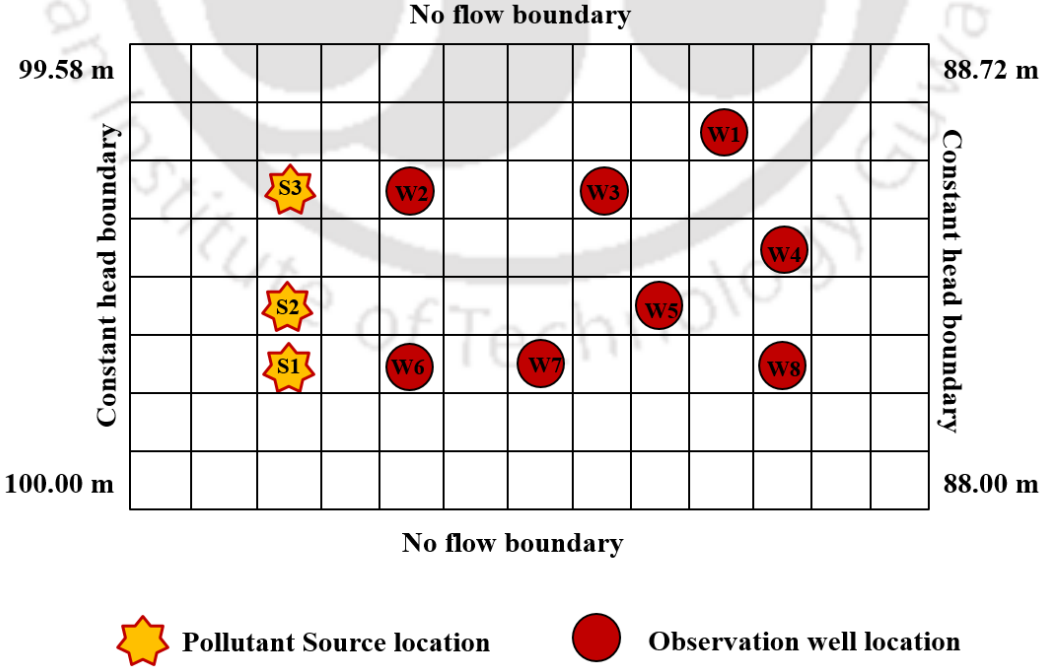


Fig. 5.11: Illustrative study area showing pollutant source locations, observation wells.

Table 5.1: Hydrological parameters used in the study area

Parameters	Values
Hydraulic conductivity in x direction, K_{xx}(m/s)	0.0002
Hydraulic conductivity in y direction, K_{yy} (m/s)	0.0002
Porosity, ε	0.25
Thickness of the aquifer, b (m)	30.5
Longitudinal dispersivity, α_L (m)	40
Transverse dispersivity, α_T (m)	9.6
Time steps, Δt (months)	3

The observation wells are designated as W1, W2…W8. Three pollutant sources are present in the aquifer. However, the location and pollutant flux are completely unknown to the problem and have to be identified using the present methodology. The groundwater flow and transport processes are simulated for 5 years at an interval of three months. It is assumed that the pollution sources are active for 4 time steps i.e. the source releases pollutant concentration for a year. The magnitude of the pollutant sources is shown in Table 5.2.

Table 5.2: Source fluxes at different time steps (g/s)

Sources	Time Step 1	Time Step 2	Time Step 3	Time Step 4
S1	47	15	37	0
S2	0	0	0	0
S3	30	58.8	0	35

5.5 Results and Discussion

The performance of the developed methodology in identifying the groundwater pollution sources is explained subsequently below.

5.5.1 Contour and surface plot of single pollution source

To understand the complexity of the problem and for easy visualization of the objective function with the location space, the same study area is taken initially with only one source (Fig. 5.12). The objective function is calculated initially at all the locations for a same random flux. Then the local optimization is carried out at each location for 10 iterations. As the number of grid size is 13x8, there are 104 locations and for each location 100 function evaluations are carried out approximately for local optimality. Therefore, a total of 10400 function evaluations are required solely to visualize the data.

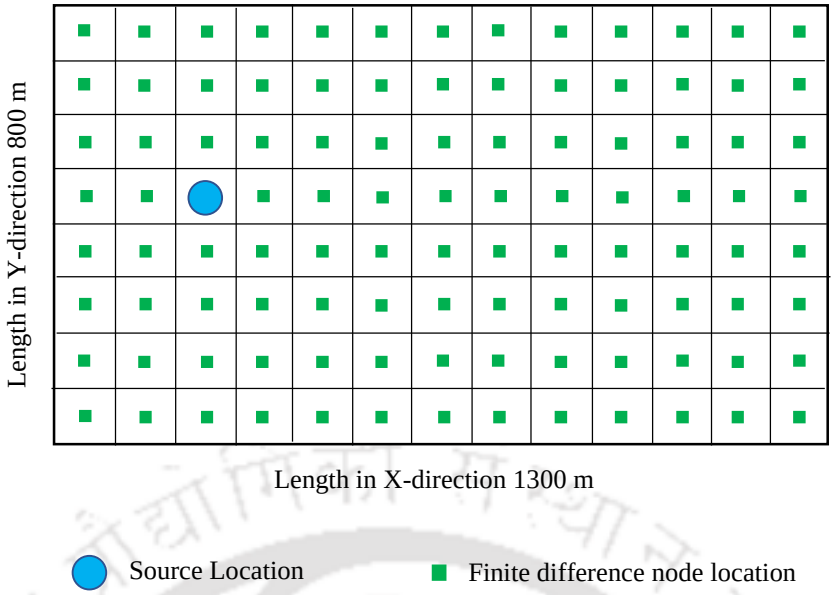


Fig. 5.12: Hypothetical study area to visualize the data

Fig. 5.13 (Fig. 5.13 (a) to Fig. 5.13 (d)) shows the behaviour of the function from the 1st iteration to the 10th iteration at all locations in contour and surface plots. In Fig. 5.13 (a) it can be observed that there are many local optimal solutions, not revealing the behaviour of the function on the first go. However, as the local optimality at each location is reached in 5th iteration as seen in Fig. 5.13 (b), the function value is decreasing showing a depression at the actual location as well as near the actual locations. As the local optimization continues, it can be clearly seen that the local optimality itself gives a precisely perfect location as the source and the near optimality conditions too disappear, revealing a steep gradient for the function value at actual location by the end of 10th iteration (Fig. 5. 13 (c) and Fig. 5.13 (d)). Thus, the designed algorithm works perfectly well for such source identification problems with location unknown.

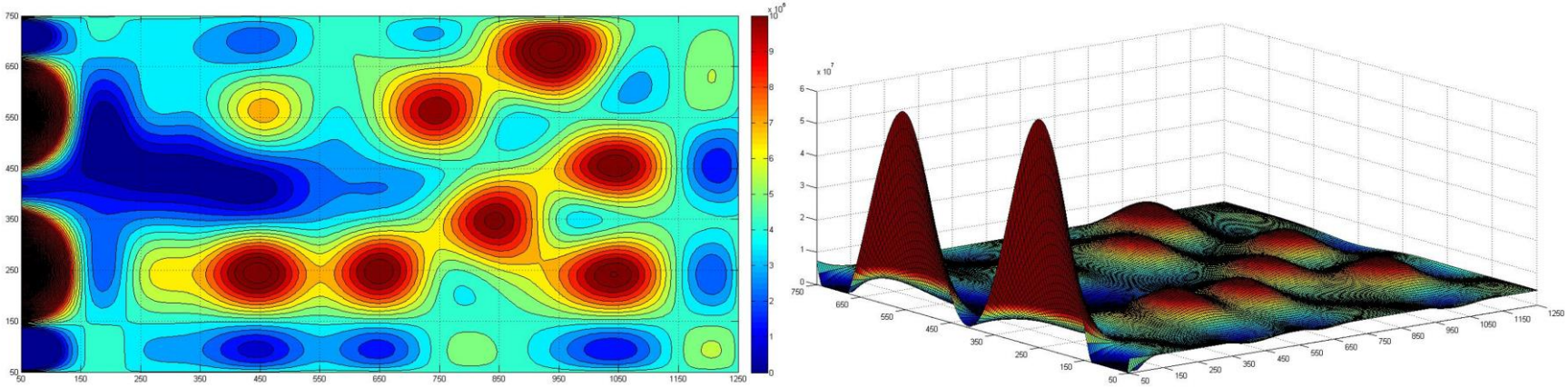


Fig. 5.13 (a): Contour and Surface plots of the function values at 1st iteration

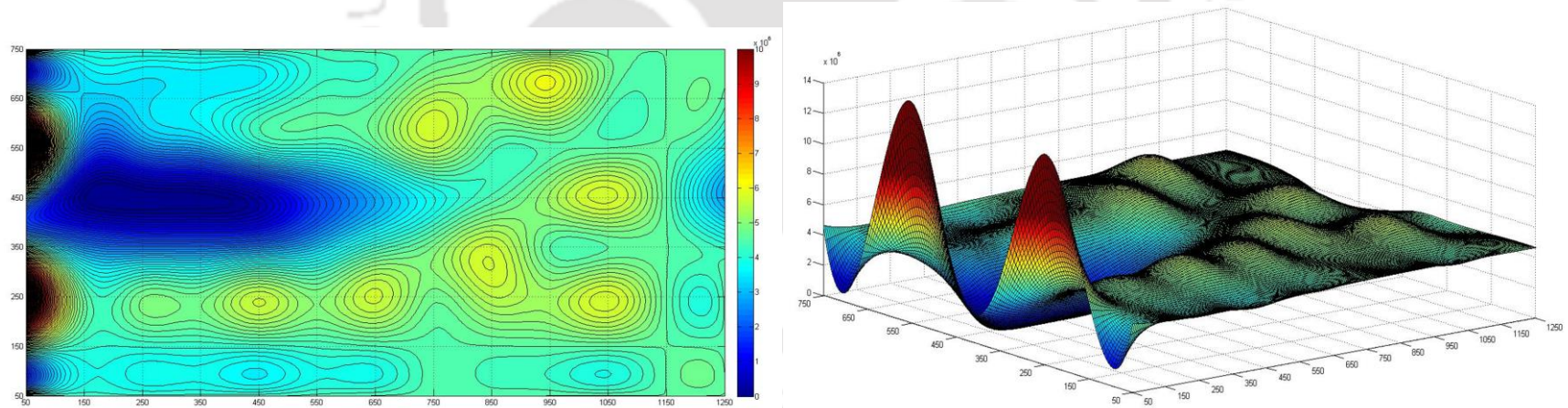


Fig. 5.13 (b) Contour and Surface plots of the function values at 5th iteration

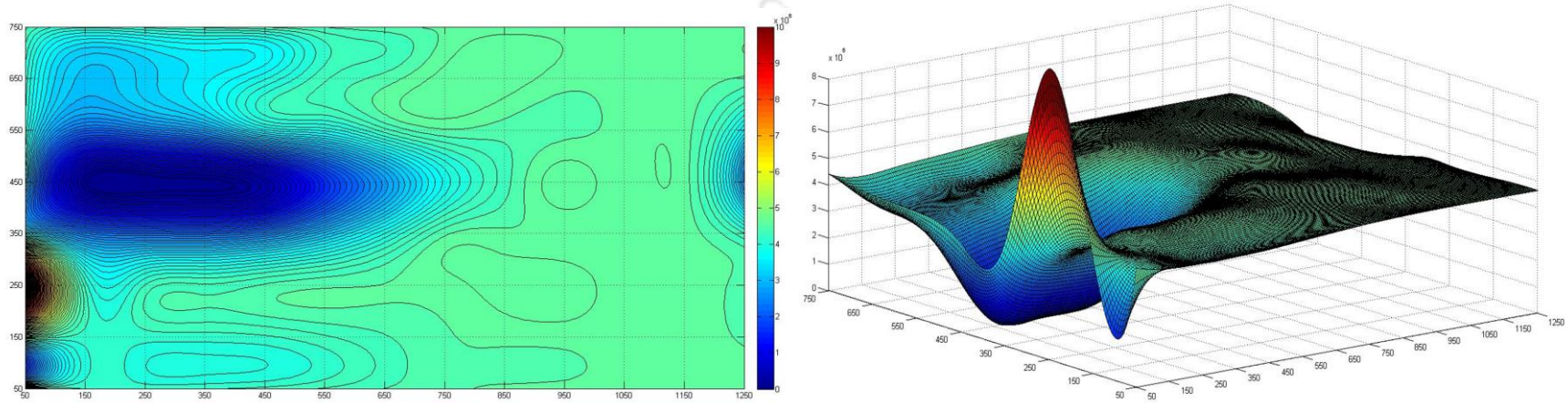


Fig. 5.13 (c) Contour and Surface plots of the function values at 8th iteration

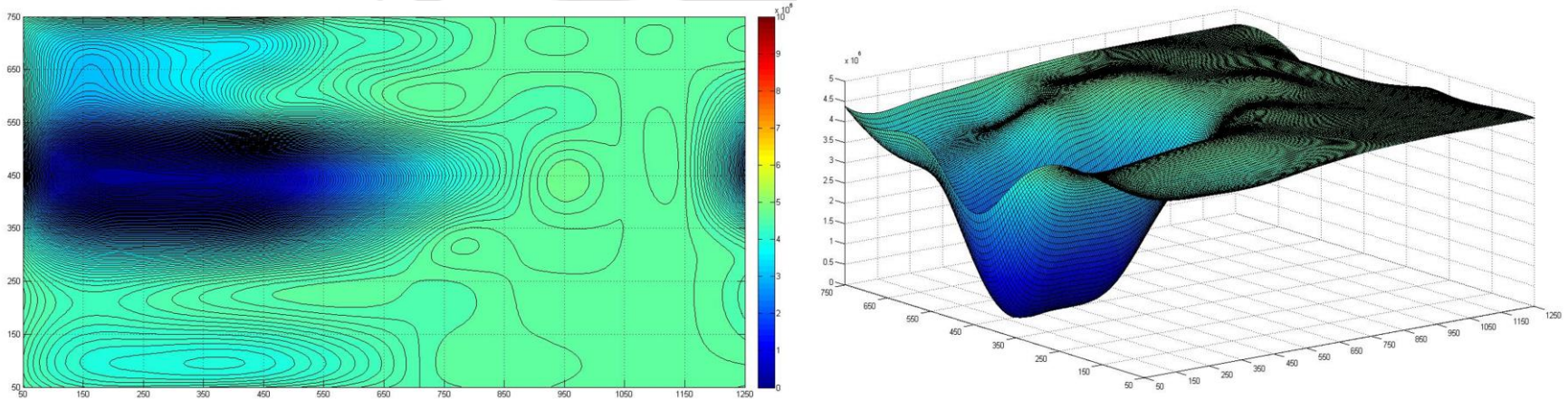
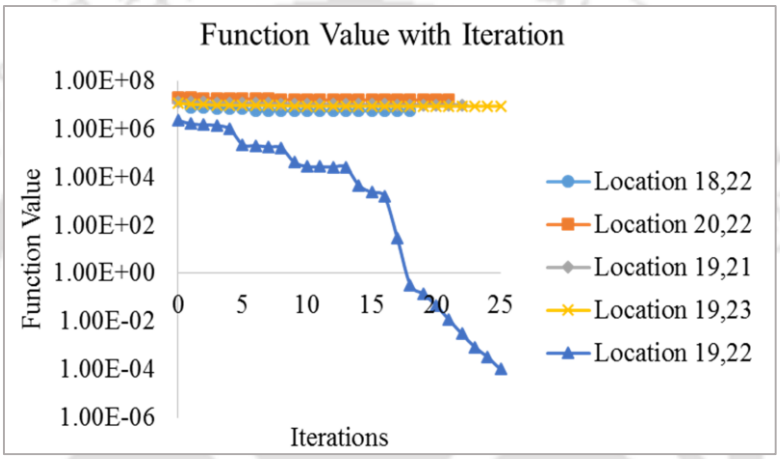


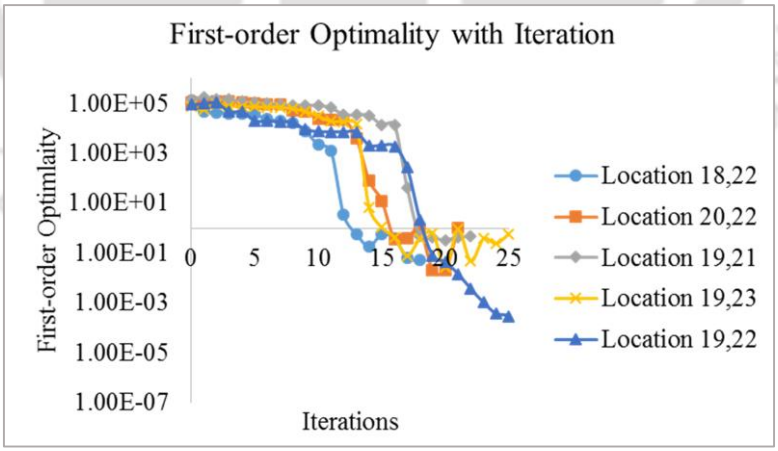
Fig. 5.13 (d) Contour and Surface plots of the function values at 10th iteration

5.5.2 First order optimality and function evaluation

Initially when the Modified GA is performed, the level of search is in the 1st iteration. There is a possibility of many local optimal solutions and just subtle changes can lead to a completely different solution. Thus, the technique of introducing stall location criteria, and diverging the location as it reaches optimality works perfectly well. But the flux values are not clear by the end of first step. Then the local location search is performed which improves the location as well as the source fluxes. At this stage, the source locations are near the optimal locations, the local location search is performed to find the exact location of the sources.



(a)



(b)

Fig. 5.14: Steepness of the (a) Function values compared to the (b) First order optimality plotted with iterations of gradient based optimization

At each location, the steepness is observed based on the first order optimality and the corresponding change in the function value (this steepness of function values is described

in Fig. 5.14 for a different example, where the actual locations is (19, 22). This search increases the possibility to get to the solution much quicker. Once the location is exactly identified, the local optimization based on gradient search gives the best results.

As discussed in earlier sections, the groundwater pollution sources are identified using three different algorithms *viz.*GA-LTS-GR, GA-MS-GR and GA-RMS-GR. Here, as it can be observed in Fig. 5.13 (a, b, c and d) there is a set of ridges (contours) which gets to the optimal location in longitudinal and transverse direction if the depressions are followed. Thus, the search space decreases to a minute level, which gives the location exactly, if these ridges were followed. Thus, out of all the three local location search algorithms, GA-LTS-GR shows the best results.

5.5.3 Performance of GA-LTS-GR, GA-MS-GR and GA-RMS-GR

The actual problem proposed by Mahar and Datta (2001) is tested and the results are presented for that case in the following section. Table 5.3, 5.4 and 5.5 show the comparison between actual and estimated pollution sources using GA-LTS-GR, GA-MS-GR and GA-RMS-GR respectively. It may be observed that the actual locations and the source fluxes are exactly identified by all the three types of approaches. The source locations and the fluxes are identified accurately but the presented models must be computationally efficient also.

5.5.3.1 Comparison between actual sources and estimated sources using GA-LTS-GR

In GA-LTS-GR, the number of function evaluation is found to be 7910 (Table 5.3). It indicates that the convergence towards the exact solution is quite good. Initially, the search for source location and the flux is performed by modified GA. At beginning, modified GA could only give the near optimal location globally. However, this non-gradient search does not converge towards the actual solution in terms of flux. The initial guess for the flux values are reliable when it comes to the solution output from modified GA. With no further improvement in the stall locations, the algorithm stops and proceeds towards gradient search.

With the entry of the classical algorithm and the local location search, much improvement could be observed. But with single cycle of the gradient search, the exact solution could not be recovered. Subsequently the GA-LTS-GR algorithm could effectively converge towards the exact solution with the application of transient and longitudinal location

search as discussed earlier. Whereas the other algorithms did not converge as fast as compared to LTS algorithm (Fig. 5.15).

Table 5.3: Comparison between actual and estimated pollution sources using GA-LTS-GR

Actual Location (Cell ID)	Estimated Location (Cell ID)	Actual source Flux				Estimated source Flux				No. of Func. Eva.
		Time Step 1	Time Step 2	Time Step 3	Time Step 4	Time Step 1	Time Step 2	Time Step 3	Time Step 4	
19	19	30	58.8	0	37	30	58.8	0	37	8910
21	17	0	0	0	0	0	0	0	0	
22	22	47	15	37	0	47	15	37	0	

5.5.3.2 Comparison between actual sources and estimated sources using GA-MS-GR

In case of GA-MS-GR, the total number of fitness function evaluation is found to be 24208 (Table 5.4) which is very large compared with the GA-LTS-GR algorithm. Although the exact source locations and flux are identified, the computational efficiency is a major concern.

Table 5.4: Comparison between actual and estimated pollution sources using GA-RMS-GR

Actual Location (Cell ID)	Estimated Location (Cell ID)	Actual source Flux				Estimated source Flux				No. of Func. Eva.
		Time Step 1	Time Step 2	Time Step 3	Time Step 4	Time Step 1	Time Step 2	Time Step 3	Time Step 4	
19	19	30	58.8	0	37	30	58	0	37	
21	17	0	0	0	0	0	0	0	0	24208
22	22	47	15	37	0	47	15	37	0	

5.5.3.3 Comparison between actual sources and estimated sources using GA-RMS-GR

The third algorithm, GA-MS-GR is found to be computationally expensive. The number of function evaluation is 32377 (Table 5.5) which is more than three times than that of the GA-LTS-GR algorithm. This search technique also converges towards the actual locations and the fluxes but has found to be computationally inefficient. So, when the three algorithms are compared, the first Local-Location search (LTS) technique can successfully identify the source locations and fluxes at lesser computational evaluation.

Table 5.5: Comparison between actual and estimated pollution sources using GA-RMS-GR

Actual Location (Cell ID)	Estimated Location (Cell ID)	Actual source Flux				Estimated source Flux				No. of Func. Eva.
		Time Step 1	Time Step 2	Time Step 3	Time Step 4	Time Step 1	Time Step 2	Time Step 3	Time Step 4	
19	19	30	58.8	0	37	30	58	0	37	32377
21	17	0	0	0	0	0	0	0	0	
22	22	47	15	37	0	47	15	37	0	

Fig. 5.15 shows the comparison of the three algorithms *viz.* GA-LTS-GR, GA-MS-GR and GA-RMS-GR on function value with number of iterations. Initially, it is observed that the behaviour of the three algorithms are almost same because the modified GAs could converge to a near best solution. But the abrupt drop of the GA-LTS-GR curve, it can be said that the fitness function of the GA-LTS-GR algorithm converges to minimum value at 64 iterations. In case of GA-MS-GR and GA-RMS-GR, the fitness function value remains constant for some iterations showing no improvement in the value before converging to minimum values at 120 and 113 respectively. It signifies that the convergence of the GA-LTS-GR towards optimal solution is achieved and has been found to be an effective algorithm when compared with GA-MS-GR and GA-RMS-GR algorithms.

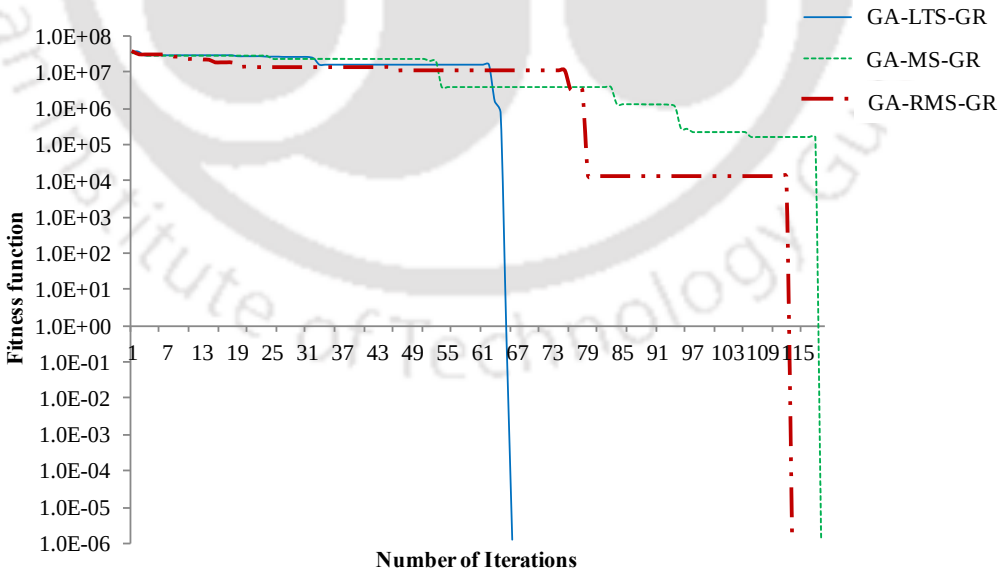


Fig. 5.15: Comparison plot of the three algorithms GA-LTS-GR, GA-MS-GR and GA-RMS-GR on function value with number of iterations.

5.5.3.4 Performance of the GA-LTS-GR algorithm

Table 5.6: Relative error for different source fluxes using GA-LTS-GR and function evaluation for 10 numbers of runs

No. of Run	Actual Location	Estimated Location	Relative Error (%)				No. of function evaluation
			Time Step 1	Time Step 2	Time Step 3	Time Step 4	
1	19	19	0.003	0.017	0.000	0.057	5453
	21	59	-	-	-	-	
	22	22	0.042	0.067	0.054	-	
2	19	19	0.007	0.032	-	0.314	4461
	21	104	-	-	-	-	
	22	22	0.042	0.067	0.270	-	
3	19	19	0.001	1.547	-	0.342	6123
	21	100	-	-	-	-	
	22	22	0.021	0.067	0.054	-	
4	19	19	0.067	0.002	0.000	0.286	7174
	21	79	-	-	-	-	
	22	22	0.213	0.067	0.027	-	
5	19	19	0.034	1.564	-	0.286	5134
	21	50	-	-	-	-	
	22	22	0.213	0.009	-	-	
6	19	19	0.037	0.017	-	0.343	5180
	21	104	-	-	-	-	
	22	59	0.149	0.133	0.270	-	
7	19	19	0.004	0.001	-	0.314	6083
	21	94	-	-	-	-	
	22	22	0.064	0.013	0.297	-	
8	19	19	0.023	0.170	-	0.286	4845
	21	26	-	-	-	-	
	22	22	0.191	0.133	0.270	0.000	
9	19	19	0.007	0.290	-	0.067	6407
	21	29	-	-	-	-	
	22	22	0.149	0.149	0.270	-	
10	19	19	0.037	0.153	-	0.040	5803
	21	104	-	-	-	-	
	22	22	0.149	0.133	0.076	-	

The effectiveness of the GA-LTS-GR cannot be guaranteed with only one observation result. Thus, for further affirmation a total number of 10 trials were carried out with the

same hydrological parameters as given in the earlier section. It was found that the pollution source locations were exactly identified in all these 10 runs.

Furthermore, for verifying the effectiveness in identifying the source fluxes, the relative errors with respect to the actual source fluxes were evaluated (Table 5.6). It is seen that the calculated relative error values are found to be very negligible for all the 10 runs. Table 5.6 also displays the number of function evaluations for all the 10 runs. Some variations could be observed in the function values but the differences can be neglected as majority of the function values falls under same range.

It can be concluded that the GA-LTS-GR is an efficient algorithm for identifying the unknown groundwater pollution sources. The algorithm is also superior in terms of number of function evaluation. Fig. 5.16 shows the function evaluation in different runs. The maximum number of function evaluation is found to be 7174 and the minimum is 4461 with a mean of 5666.

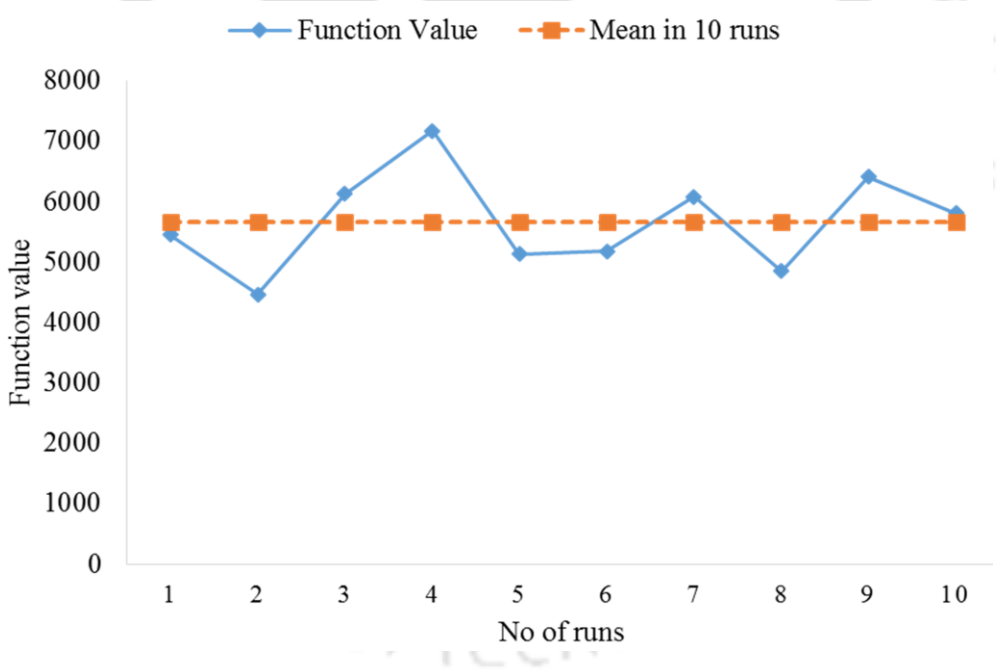


Fig. 5.16: Number of function evaluations for 10 runs

5.5.4 Comparison between NLP model (Mahar and Datta, 2001) and GA-LTS-GR algorithm

A comparative analysis has also been carried between the source flux obtained using the GA-LTS-GR algorithm and the NLP optimization model (Mahar and Datta, 2001) for the same hypothetical study area. Table 5.7 shows the average sets of the source fluxes for five number of realization. The comparison shows that the flux recovered using the GA-

LTS-GR algorithm closely resembles the actual one. Relative error is also calculated for all the source fluxes with respect the actual flux. It can be observed that the relative error of GA-LTS-GR algorithm is found to be much lesser than that of NLP model (Mahar and Datta, 2001), which signifies that the present GA-LTS-GR is found to be more efficient than the NLP model when applied to the same hypothetical model.

Table 5.7: Comparison between source fluxes estimated using NLP (Mahar and Datta, 2001) and GA-LTS-GR algorithm

Time Steps	Source Location	Actual Source Flux	NLP (Mahar and Datta, 2001)	GA-LTS-GR	Relative error (%)	
			Avg. values (5 runs)	Avg. values (5 runs)	NLP	GA-LTS-GR
1	S1	47	46.23	46.98	1.63	0.04
	S2	0	0	0	-	-
	S3	30	32.11	30	7.03	0.01
2	S1	15	17.23	15.01	14.87	0.61
	S2	0	0	0	-	-
	S3	58.8	54.39	57.78	7.5	0.03
3	S1	37	32.83	36.98	11.27	0.27
	S2	0	0	0	-	-
	S3	0	1.22	0	-	-
4	S1	0	1.67	0	-	-
	S2	0	0	0	-	-
	S3	35	34.70	34.91	0.86	0.29

5.6 Summary and Conclusions

Numerous studies on identification of groundwater pollution source were performed in the past. Many have concluded that a single algorithm could not effectively determine the source locations along with their fluxes. One of the famous heuristic approaches, GA has the ability to determine discrete variables which makes the algorithm a competent one in identifying the location of the pollution sources. On addition to this, the gradient approach is capable of determining the continuous variables of the problem. Hence, this can be utilized in identifying the pollution source flux. So, considering the advantages of these two approaches, these two techniques were combined to form an improved one. However, this approach may be a robust one theoretically, but it cannot be always assured that the exact pollution source location and the flux will be identified as there are many more complexities in the function behaviour at different levels of optimality. Henceforth,

initially GA is modified for obtaining near optimal locations. The algorithm is called as Modified GA. The solution obtained from this modified GA approach has been further refined. The best solution obtained using modified GA will be the input for the three different algorithms *viz*. LTS, MS and RMS.

As the source locations obtained using modified GAs are the near best ones, the exact locations are around it and hence with an effective algorithm, the model can certainly converge to optimal solution. The first technique LTS algorithm searches for the exact location source in the transient or longitudinal direction alternatively. The second logic is RMS based on the principle that the ripple formed on a water body will migrate towards the centre portion having the maximum strength. Similarly, the location obtained will migrate towards the best one based on the fitness function value and will be replaced by the best. In MS algorithm, some of the best population members are randomly selected and mutation is performed on these selected strings. The location of the mutation point is selected on the basis whether odd or even generation takes place from the initial GA-classical cycle. For even generation, mutation is performed by adding or subtracting 1 or 2 factors around the exact location column wise. If odd, mutation will be performed by adding or subtracting 7 or 8 factors in row wise. By following this pattern, the ultimate location and flux could be obtained.

Even though all the three algorithms could identify the exact source locations and fluxes, the computational efficiency of the three algorithms are different. When compared, the GA-LTS-GR algorithm is found to be the most efficient one as it required minimum function evaluation out of the three algorithms. A comparison between the source flux obtained using NLP model (Mahar and Datta, 2001) and the GA-LTS-GR algorithm also reveals that the GA-LTS-GR algorithm performs much better than the NLP model for the same hypothetical study area. Also, the search space increases and as the number of pollution sources increases. This method of source identification gives the best technique to handle source identification problem with unknown locations most efficiently.

In the present modified GA, the initial candidate solutions are randomly generated. As the initial solutions are randomly generated, the chance of producing exact solution reduces and it may take more iterations to reach the optimal solution. Therefore, in the next chapter a methodology is presented which will improve the initial population and assists in converging to optimal solution efficiently.

Chapter 6

Identification of Groundwater Pollution Sources using Pool Population

The present chapter proposes a technique for obtaining a solution comprising of the locations and strength of groundwater pollution sources which are closed to the true value of the problem. The first section describes how the previous model generated the initial population randomly which may not yield the global optimal solution of the problem in every run of the model. The methodology explained in the second section involves the selection of the initial solution from a generated pool which comprises of the probable locations. The third section describes the results obtained by using the present methodology.

6.1 Introduction

The previous chapter presented a Modified GA-Local Location Search-Gradient based approach for identifying the groundwater pollution sources locations and the source fluxes. The GA is modified to effectively handle the discrete and continuous variables i.e. the source locations and the source fluxes of the source identification problems. The first step in GA is initiated by randomly generating the initial population. Subsequently, the better individuals are propagating generation to generation after suitably modified by the genetic operators. However, the solution achieved and the generation required to obtain the optimal solution depend on the initial randomly generated population. Also, the required number of function evaluations for convergence towards optima varies for each run as the solution starts with randomly generated initial population. Also, the feasibility of producing an initial population with same traits does not happen every time. Hence, the solution obtained by modified GA may not always be the near global optimal solution. It has also been seen that the solution obtained using the modified GA is further utilized as the initial points for the gradient search. Therefore, the overall performance of the algorithm in terms of the number of function evaluations varies largely with the initial solution of the modified GA. Although this algorithm gives the

best solution using a very small number of function evaluations, it might not be very effective in handling a large number of sources or larger study areas.

To overcome this problem, a better methodology is proposed that reduces the likelihood of selecting the redundant locations as sources. This methodology (Model 4), to be discussed in this chapter further modified the GA to give a much better initial solution. Then the LTS algorithm which is quite efficient in improvising the location is used in the next step to get to the optimal solution. The performance of the present methodology is checked using illustrative study areas of different grid sizes and a different number of pollution sources.

6.2 Methodology

The methodology proposed in this chapter focuses on the initial solution of the problem for the modified genetic algorithm. It is well known that genetic algorithm has the capability to reach the global optimal solution or near the global optimal solution. However, an important information being ignored by the modified genetic algorithm in obtaining the global optimal is the information about the observation wells. The observation well data is a set of concentrations. But when these concentrations are plotted with respect to the time at which the data is collected, it gives a set of breakthrough curves. These curves give a much valuable information about the location and source flux values at the sources. This information can be used to narrow down the search space so that the solution obtained by the modified genetic algorithm is always a near global optimal solution. The algorithm proposed in this chapter is explained in Fig. 6.1. The calculation of the probability of each of the location to be a source location in the entire area is carried out initially, as discussed in the next section. Then a pool of probable locations is generated based on these probabilities. This means that when a random location is selected from the pool, the probability of that location to be an actual location is predetermined. The population of modified GA is always called from this generated pool of the population ensuring that the possible location has the probability to be picked as an initial solution. Then the aquifer simulation model is called and the simulated concentrations at all the observation well locations are calculated. The corresponding fitness is calculated and then the best fitness is checked with the stall location tolerance. If the tolerance is not reached then the population goes through the modified genetic operators to improve the population. If the tolerance is reached then the algorithm stops and gives the location as the solution of the modified

GA. The solution obtained by the modified GA is then passed through the gradient-based search and Local location search (LTS) algorithms to get to the final solution which is similar to the process described in Chapter 5.

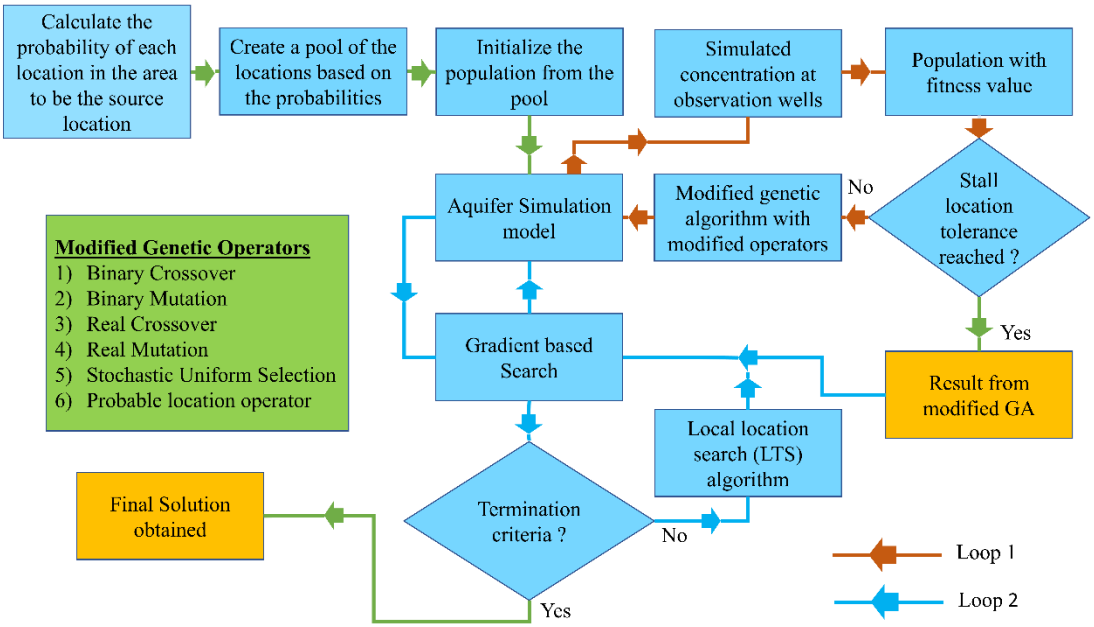


Fig. 6.1: Description of the present methodology

6.2.1 Probability of each location to be a source

The process of calculation of the probability of each of the location to a potential candidate for a source location is explained in Fig. 6.2.

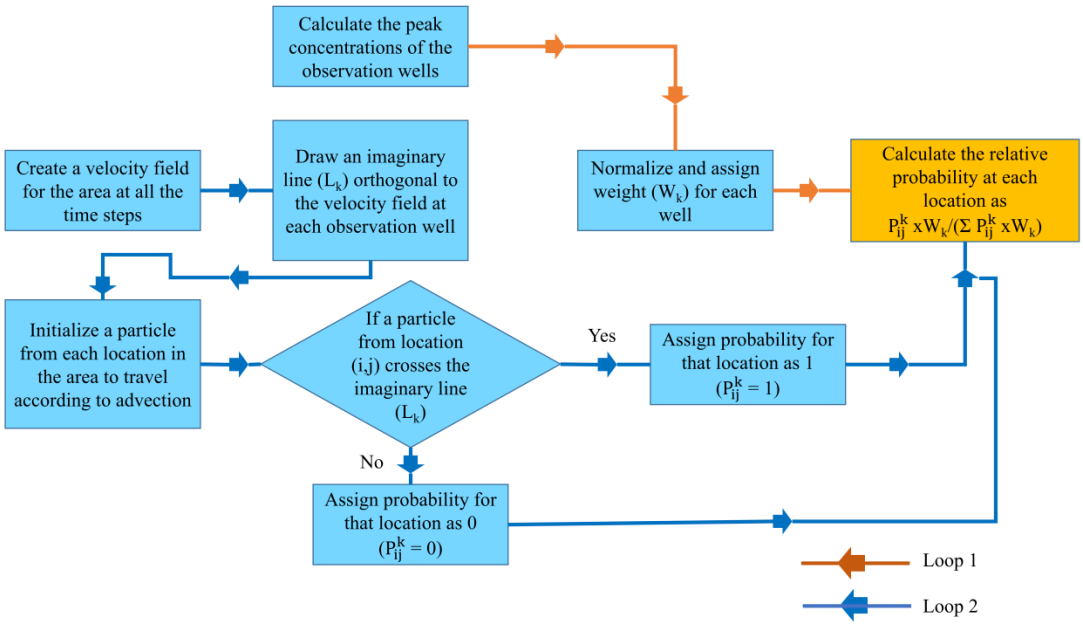


Fig. 6.2: Steps involved in calculation of probability at each location point

The probability calculated at each location point of the study area will help in forming a pool based on the probability values. The pool comprises of good locations which will help in converging to the optimal solution at much faster rate. In groundwater contaminant transport, the process of advection is considered to be more dominant than the dispersion as the process of advection goes along with the groundwater flow. Considering the dominance of advection process in groundwater processes, the effect of velocity is also used in the present study. As such, the velocity field is created at each location based on the information from the flow equation and the description of velocities at each time step (Fig. 6.3). The velocity field is further utilized in the subsequent steps while calculating the probability (explained in succeeding sections). The probability of each location to be the source is calculated and normalized so that the sum of all the probabilities over the area is 1.

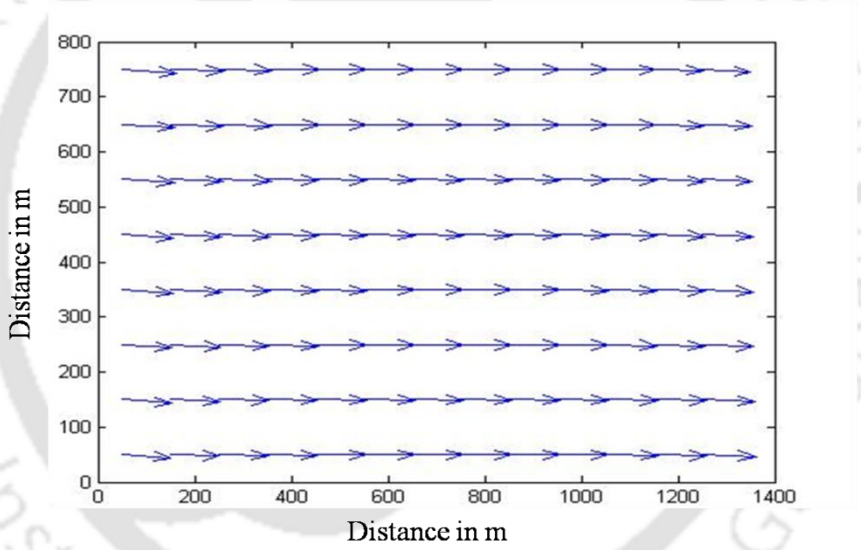


Fig. 6.3: Velocity field created for the study area at all-time steps

Another factor that is considered in the present methodology is the peak concentration of the observation well. The reason being that breakthrough curves can give the information about the location of the pollution sources. By calculating the peak concentration at the observation wells present an idea about the position of the sources from the observation wells. The variation of concentration peaks is shown in the Fig 6.4 for an example problem defined in the section 6.2.1.1 By studying the trend of the curve from Fig. 6.4, a brief idea about the source location can be made. The peak value of observation well W6 and W2 are quite high which signifies that they are close to

pollution sources. Whereas for W8 and W1, the peak values are quite low, suggesting that they are located at a far distance from the pollution sources.

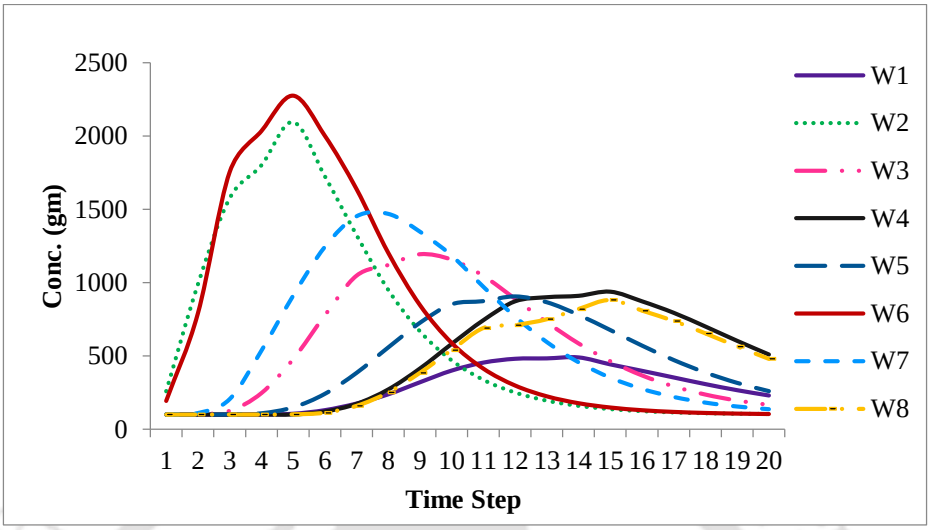


Fig. 6.4 Concentration curve showing peak values for all the eight observation wells

So, in order to acquire the contribution of each of the observation wells, a normalized weight based on the peak values of the concentrations is assigned to each observation well as described in equation (6.1)

$$W_k = \frac{C_p^k}{\sum_{k=1}^{k=N} C_p^k} \tag{6.1}$$

Here, C_p^k is the concentration at the peak of well k, W_k is the weightage assigned at well *'k'*, *N* is the total number of observation wells. The weightages at each of the observation wells for the example are shown in the Table 6.1.

Table 6.1: Weightage values for all the eight observation wells

Wells	W1	W2	W3	W4	W5	W6	W7	W8
Weightage	0.0425	0.1788	0.1043	0.0949	0.1251	0.2082	0.1485	0.0976

6.2.1.1 Orthogonal position of the observation well

In this advection dominated process, eight observation wells are placed randomly throughout the study area (13 columns x 8 rows) of the aquifer as shown in Fig. 6.6. The location of the eight observation wells are (7,10), (6,5), (6,8), (5,11), (4,9), (3,5), (3,7) and (3,11). For each of the observation wells, orthogonal direction with respect to

the natural advection of the present aquifer is identified. The orthogonal lines (L_1,L_2, L_3, L_4, L_5, L_6, L_7 and L_8) represent the set of lines which are responsible for the contamination concentration to be recorded in the observation locations. The particles that are tracked so that they cross theese lines are given in the section 6.2.1.2.

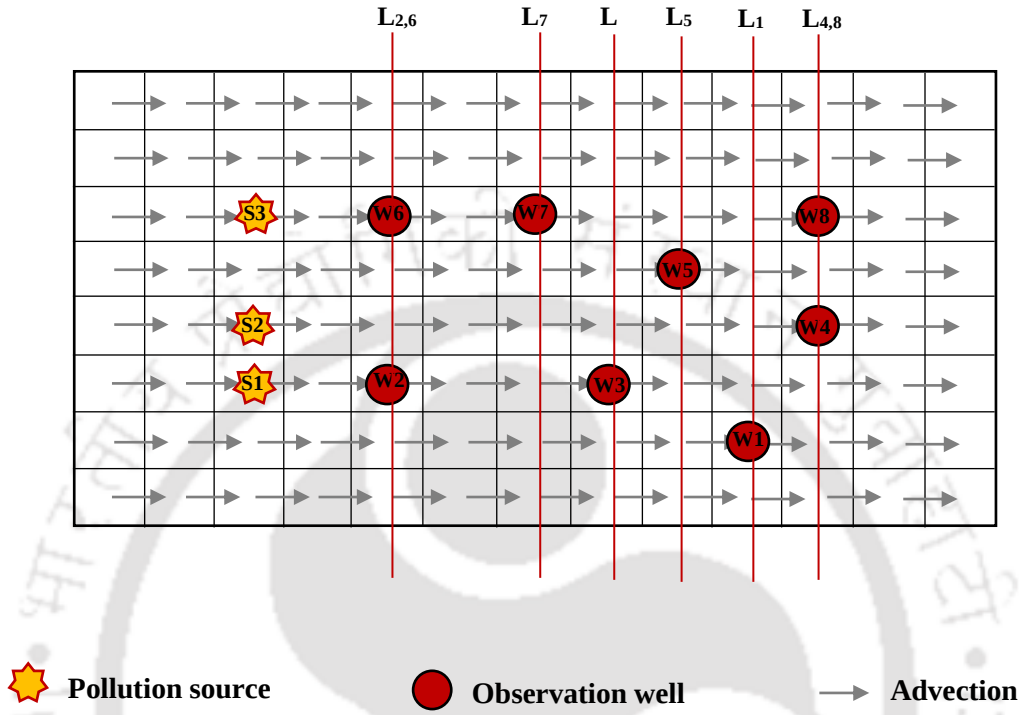


Fig. 6.5: Eight observation wells are placed in the advection dominated process of transport process

6.2.1.2 Particle tracking

By using the particle tracking theory, the future locations of the particles are calculated based on the velocity field. The particle location is represented (x, y). As the velocity field (V_x, V_y) are known at each point, the position after dt time can be updated as (x + $V_x$dt , y + $V_y$dt). Following this procedure, the position of the particles by the end of time steps can be evaluated.

6.2.1.3 Assigning probability

As the orthogonal positions for each of the observation wells locations are known, then the crossing probability of the particle can be evaluated. If the particle crosses each of the orthogonal position (L_k), then the probability (P_{ij}^k) will be equal to 1 otherwise 0. This can be visualized in the Fig. 6.6. Particle $p_1$is crossing the orthogonal imaginary line but particle p_2 is not, therefore the probabilities are assigned as 1 and 0 respectively.

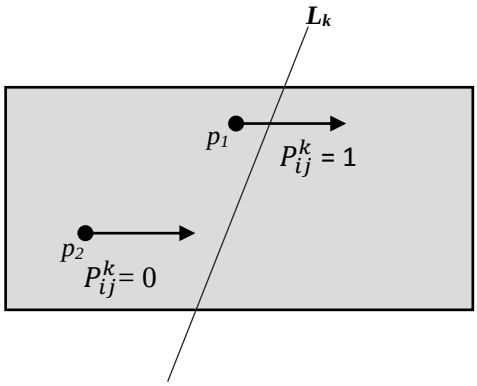


Fig. 6.6: Probabilities of the particles based on orthogonal lines

The probabilities of each location *(i, j)* associated with each observation well *'k'* is given by P_{ij}^k. Then each well has the weightage *'W_k'* representing the strength of that particular well to result a concentration corresponding to the source. The probability of each location to be source (P_{ij}) is thus linked with the probability P_{ij}^k and weightage W_k as shown in equation 6.2.

$$P_{ij} = \frac{\sum_{k=1}^{N} P_{ij}^k w_k}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{N} P_{ij}^k w_k} \tag{6.2}$$

Table 6.2 shows the probabilities for each of the location (grid points) for the example taken up to explain the present methodology.

Table 6.2: Probabilities calculated at each location of the study area

0	0.0073	0.0169	0.0197	0.0157	0.0134	0.0136	0.0129	0.0126	0.0092	0.0036	0	0
0	0.0097	0.0145	0.0206	0.0154	0.0135	0.0131	0.0138	0.0114	0.0092	0.0036	0	0
0	0.0097	0.0169	0.0182	0.0164	0.0133	0.0132	0.0133	0.0124	0.0080	0.0036	0	0
0	0.0097	0.0169	0.0182	0.0164	0.0133	0.0132	0.0133	0.0124	0.0080	0.0036	0	0
0	0.0097	0.0169	0.0182	0.0164	0.0133	0.0132	0.0133	0.0124	0.0080	0.0036	0	0
0	0.0121	0.0145	0.0192	0.0161	0.0134	0.0127	0.0143	0.0112	0.0080	0.0036	0	0
0	0.0121	0.0145	0.0192	0.0161	0.0134	0.0127	0.0143	0.0112	0.0080	0.0036	0	0
0	0.0121	0.0145	0.0192	0.0161	0.0134	0.0127	0.0143	0.0112	0.0080	0.0036	0	0

As there are 13 columns and 8 rows in the study area, a total number of 104 probabilities are calculated. As the probabilities are calculated for each of the location

points, a pool is generated based on the probabilities. From the pool, the modified GA picks up the location ensuring that the most possible location has the probability to be picked in the initial population.

6.2.1.4 Generating a population from the pool

The pool is a box containing the samples of locations based on their probabilities to be a source location. After calculating the probabilities of each of the location to be an actual source, the total number of samples to be present in the pool is decided (say 500). The probabilities at each location (*i, j*) is then multiplied by the total number of samples (500 samples), which gives the number of copies of that location (*i, j*) in the pool. By following the same procedure, the number of copies of each location are calculated.

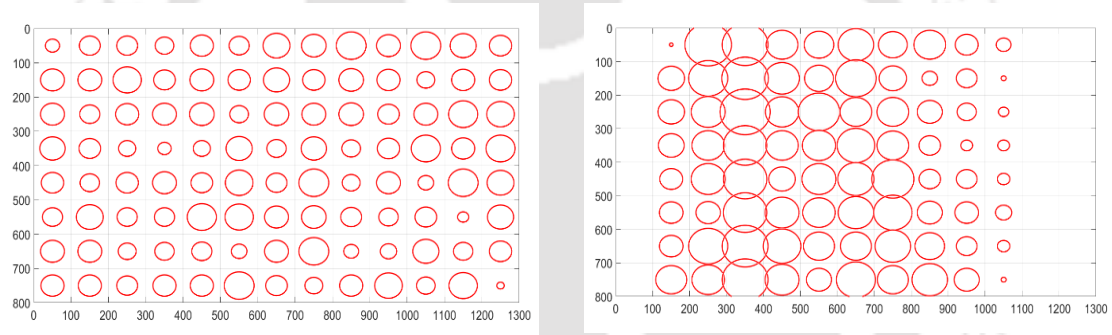
Then the population is taken randomly from the pool. This assures that the probability of each location to be selected is decided by the predetermined probabilities. For example, for a total sample of 500, the number of copies of location (3, 3) is 0.0169 x 500 (the value of 0.0169 is from table 6.2) which is equal to 8 copies. If a random sample is selected from the pool, the probability of location (3, 3) to be selected is 8/500 which is equal to 0.016. Higher precision can be obtained by selecting the total number of samples in the pool to be high. Table 6.3 shows the number of copies of each location to be in the pool taking the total samples to be 500.

Table 6.3: Number of copies of each location to be in pool

i,j	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	0	4	8	10	8	7	7	6	6	5	2	0	0
2	0	5	7	10	8	7	7	7	6	5	2	0	0
3	0	5	8	9	8	7	7	7	6	4	2	0	0
4	0	5	8	9	8	7	7	7	6	4	2	0	0
5	0	5	8	9	8	7	7	7	6	4	2	0	0
6	0	6	7	10	8	7	6	7	6	4	2	0	0
7	0	6	7	10	8	7	6	7	6	4	2	0	0
8	0	6	7	10	8	7	6	7	6	4	2	0	0

Generating the population randomly and generating the population based on the pool gives a significant improvisation in the initial locations. The random population generated randomly and from the pool are shown for 1500 samples in the figure 6.7. Here the number of times each location selected is represented as a circle with corresponding radius. Figure 6.7 (a) shows the samples selected based on random numbers between lower bound and upper bound. Figure 6.7 (b) shows the samples selected randomly from the pool.

The information of observation wells for Fig. 6.7 are for 4 sources at locations (250m,250m); (450m,250m); (550m,250m); (350m,450m). The radii of the circles at these locations are bigger in the pool generated population (Fig. 6.7(b)), whereas it is small at randomly generated population (Fig. 6.7(a)). Therefore, it is better to use the probabilities of locations as initial information to generate the population, and also as an operator in the modified GA which is explained in the next section 6.2.2.



a) Population generated randomly b) Population generated from pool

Fig. 6.7: Frequency of each location to be selected based on random and pool selections

6.2.2 Modified Genetic operators and Algorithm

The modified GA operators are called when the stall location tolerance is not reached from the main algorithm. The input of the population string along with the fitness is initially selected according to fitness using the stochastic uniform selection. Then the strings are divided into two portions, i.e. locations and the fluxes. For the flux portion of the chromosome, the real coded crossover and mutation have been used to obtain the new flux. The stall location ratio expressed as a percentage (x %) and the percentage of the number of sources to the population (y %) are calculated. Out of all the populations for the locations, x % are mutated using binary mutation and (100-x-y) % goes through crossover using the binary crossover. The rest y % of the population are modified every

generation by randomly selecting the population according to the probability at each location.

These modified locations and the fluxes are then combined to form new population and new fitness is calculated using the aquifer simulation model. The stall location tolerance is then checked for the best chromosome. If it is satisfied, then the solution is given as the final solution, otherwise the algorithm continues with the selection operator again and the cycle continues until termination. The modified genetic operators used in the algorithm are described in Fig 6.8.

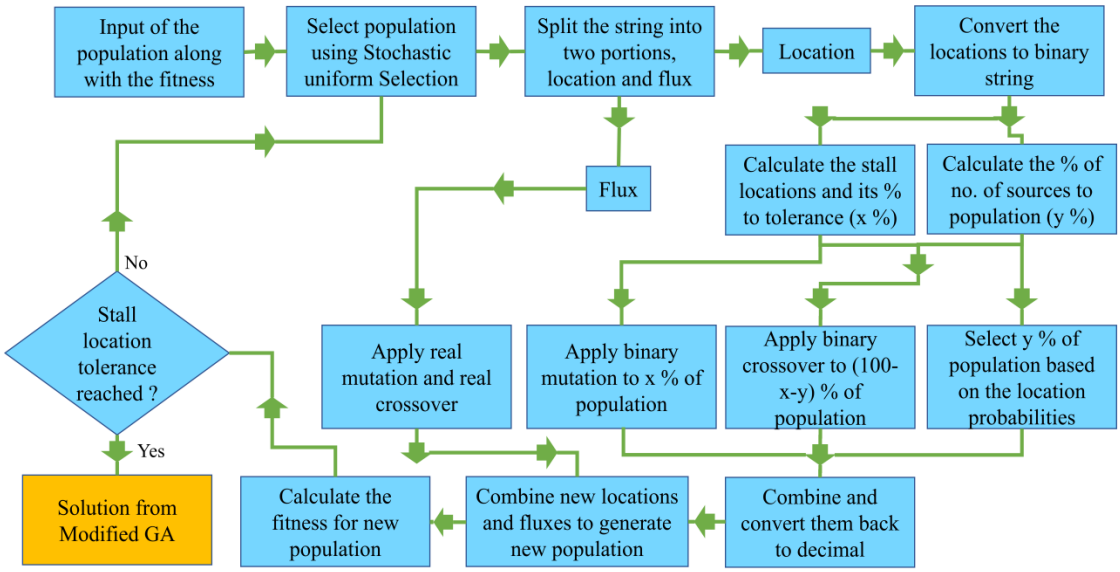


Fig. 6.8: Steps involved in Modified GA operators

6.3 Results and Discussions

The performance of the present methodology after utilizing the generated pool is discussed below. Here, different scenarios of the source identification problem are studied. It has been mentioned in the earlier sections that the complexity of identification of groundwater pollution sources intensifies with the increase in the size of the aquifer. In the linked simulation-optimization model, executing a large domain requires a large number of function evaluations and hence can be computationally very expensive. Therefore, the performance of the present methodology is checked by evaluating different number of grid cells for the aquifer. As such different simulation models are developed in the Groundwater Modeling System (GMS) environment of varying number of grids.

Another factor which may escalate the function evaluation of the objective function is the number of pollution sources available in the affected aquifer. Henceforth, more of pollution sources will have more decision variables which enhance the degree of complexity. Considering the above circumstances, two cases have been considered.

6.3.1 Case 1 – Different number of pollution sources

This section presents the result for different number of pollution sources. Fig. 6.9 shows different study areas adopted in the present case. Initially, it is assumed that only one number of source (3,3) is present in the aquifer. There is a possibility of having more than one number of polluted sites in the affected aquifer. Henceforth, the number of pollution sources are subsequently increased as two, three and four for the same study area (1.04 km^2).

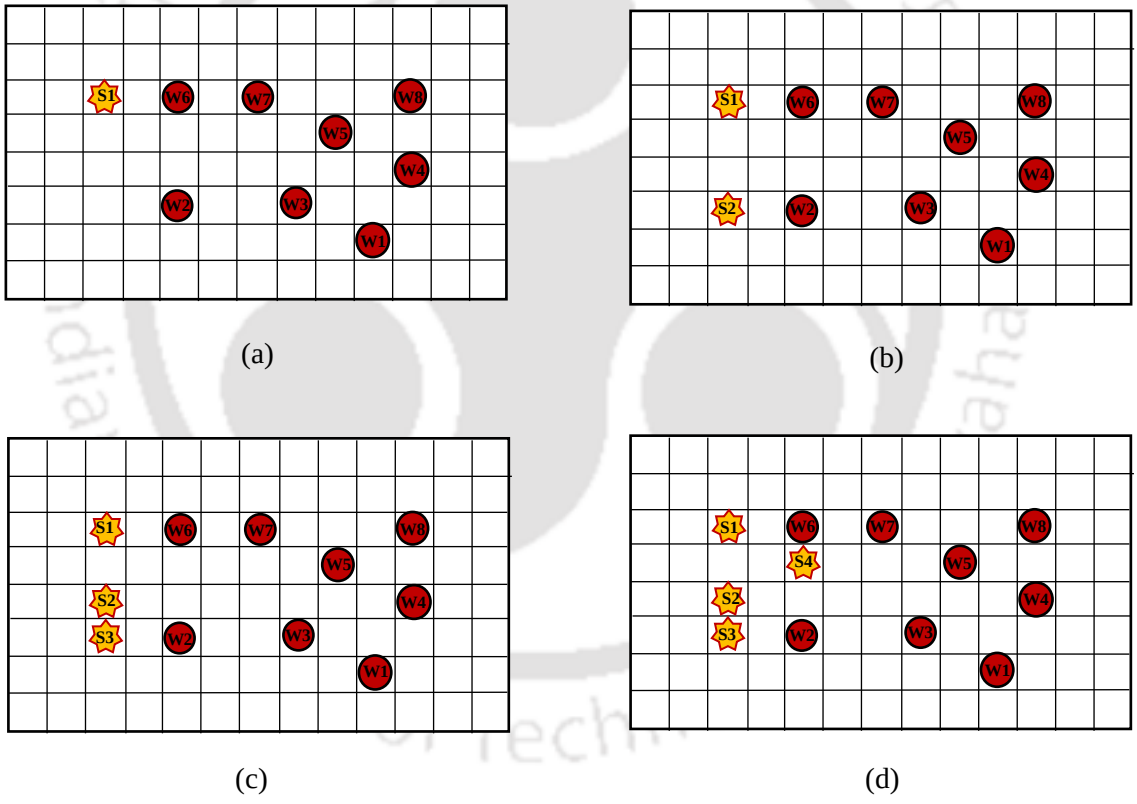


Fig.6.9: Study areas with (a) One source (b) Two sources (c) Three sources and (d) Four sources

Table 6.4 represents the hydrological parameters used in all the study areas. The first task in source identification problem is to check whether the source locations and the source fluxes are exactly identified or not. Table 6.5 showed the identified source locations and mean value of function evaluations recovered from ten simulations. It is

observed that the model has identified the exact source locations for all the ten runs. The mean value of function evaluation also rises gradually and the trend is discussed subsequently.

Table 6.4: Hydrological parameters used in the study areas

Parameters	Values
Hydraulic conductivity in x direction,K_{xx}(m/s)	0.0002
Hydraulic conductivity in y direction,K_{yy} (m/s)	0.0002
Porosity, ε	0.25
Thickness of the aquifer, *b* (m)	30.5
Longitudinal dispersivity, α_L (m)	40
Transverse dispersivity, α_T(m)	9.6
Time steps, Δt (months)	3

Table 6.5: Estimated source locations and mean value of function evaluations

No. of source	Actual Location (*i,j*)	Estimated Location (*i,j*)	Mean value of function evaluations
One	(3,3)	(3,3)	593
Two	(3,3)	(3,3)	1740
	(6,3)	(6,3)	
Three	(3,3)	(3,3)	9064
	(5,3)	(5,3)	
	(6,3)	(6,3)	
Four	(3,3)	(3,3)	34282
	(5,3)	(5,3)	
	(6,3)	(6,3)	
	(4,5)	(4,5)	

Table 6.6: Comparison between estimated source locations using Modified GA-LTS and Modified GA with modified operator- Pool location

Model	No. of sources	Actual Location	Estimated Location	Mean value of function evaluations
Modified GA-LTS	2 + (1 dummy)	(3,3)	(3,3)	5666
		(6,3)	(6,3)	
		(5,3)	(1,3)	
Modified GA operator- Pool location	2	(3,3)	(3,3)	1740
		(6,3)	(6,3)	

A comparison is made between the source locations (for two sources) identified using modified GA-LTS chapter and the modified GA operator-pool location (Table 6.6) for the same study area. It could be inferred from Table 6.5 that the number of function evaluations required by the present Modified GA operator- pool location model is three times lesser than the Modified GA-LTS model for converging towards the optimal solution with the same number of active sources.

Fig. 6.10 shows the number of function evaluation for a different number of sources. For all the 10 simulation runs, the range of function evaluations for one, two and three number of sources show a similar trend. However, in case of four sources, the number of function evaluation fluctuates at every run. Moreover, it is also seen that there is an abrupt rise in the number of function evaluations for four sources. This suggests that the model is very effective for one, two and three sources in terms of function evaluations but some inconsistency is seen with four pollution sources. Even though this model shows some discrepancies in range of function evaluation for all the runs of the four sources, butare capable of converging towards the optimal solutions.

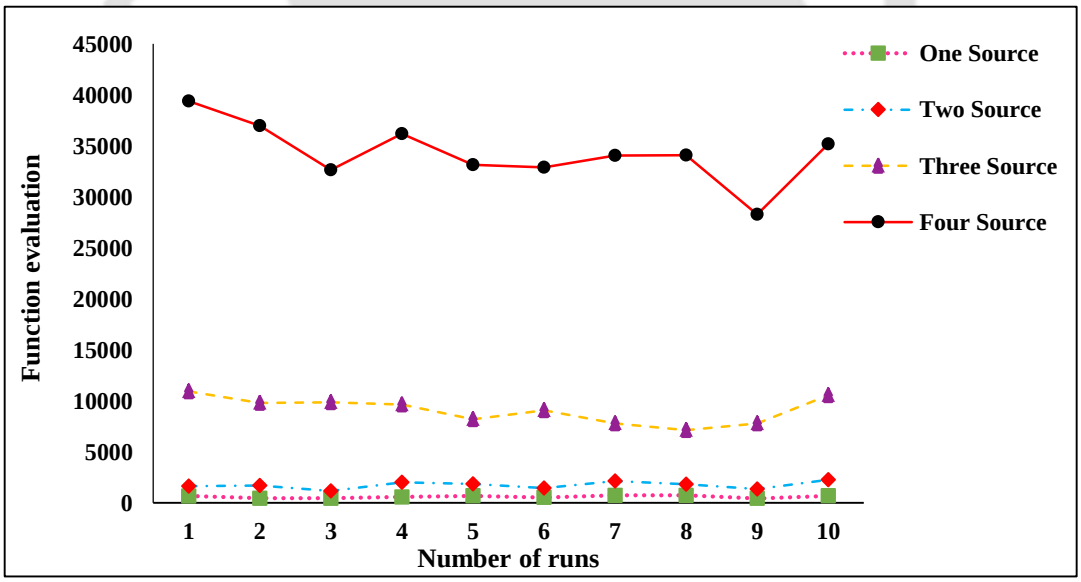


Fig. 6.10: Comparison of function evaluations for one, two, three and four number of pollution sources

The further identification capability of the present model is evidently visible in the average values of estimated source flux estimated (Table 6.6). It is found that the average value of source fluxes closely matches with the actual source fluxes for different number of pollution sources.

Table 6.7: Comparison between the actual and the estimated source fluxes (Average values)

No. of source	Source Flux (g/s)							
	Actual				Estimated (Average)			
	Time Step 1	Time Step 2	Time Step 3	Time Step 4	Time Step 1	Time Step 2	Time Step 3	Time Step 4
One	47	15	37	0	46.93	15.00	36.99	0.00
Two	47	15	37	0	46.99	15.00	36.98	0.00
	30	58.8	0	35	30.00	58.79	0.00	34.99
Three	47	15	37	0	47.00	14.99	36.99	0.00
	30	58.8	0	35	30.00	58.78	0.00	34.99
	36.50	0	25.45	12.23	36.51	0.00	25.44	12.23
Four	47	15	37	0	47.00	14.99	36.99	0.00
	30	58.8	0	35	30.00	58.79	0.00	34.98
	35.99	0.00	11.95	12.67	35.99	0.00	11.94	12.67
	17.83	39.46	24.35	0.00	17.82	39.46	24.35	0.00

The above discussions have shown that the model is capable of converging towards the optimal solutions and computationally efficient in terms of the number of function evaluation. The next description describes the trend of convergence (Fig. 6.11).

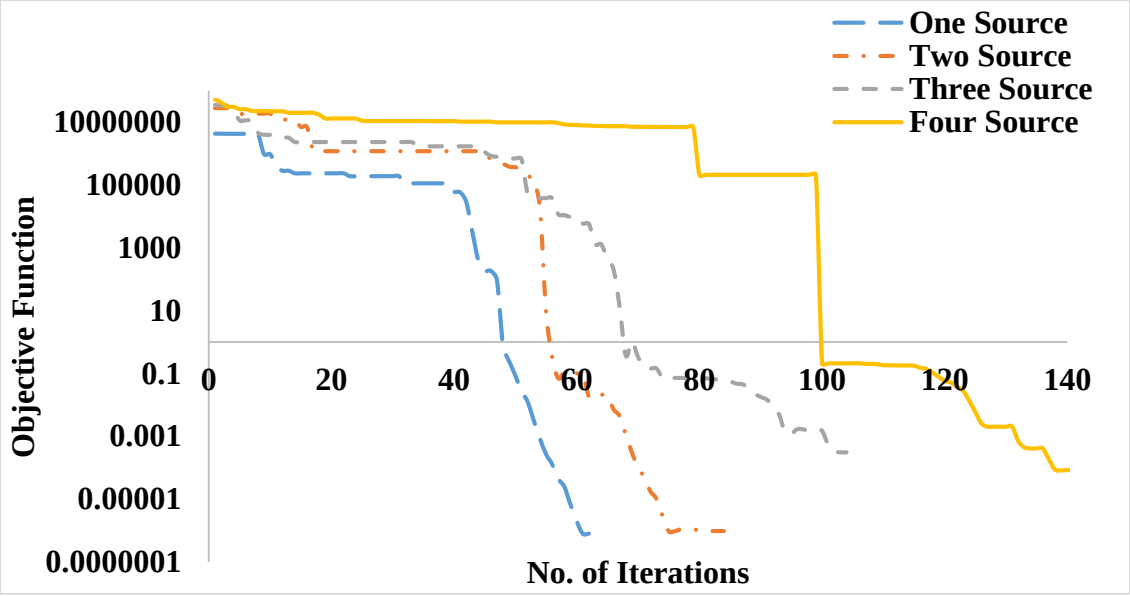


Fig. 6.11 comparison of objective function for different number of pollution of sources

In Fig. 6.11, it can be seen that initially, the objective function value for one number pollution source is 4.17 x 10^6. This function value depicts the performance of the genetic algorithm at the initial phase of the algorithm. Later it slowly converges to 0.0268 at 51th iteration. At this point, it may be mentioned that the model has converged towards the optimal solution and with further fine-tuning of the objective function value, it reaches 7.89 x 10^{-7} and achieved the optimal solution.

In the next subsequent steps a total number of two, three and four number of pollution source are considered in the same hypothetical area. With two number of pollution source, the objective function starts with 2.78 x 10^7. The grid location of the two pollution sources are (6,3) and (3,3). It took 84 hefty iterations to converge towards the optimal solution at an objective function of 9.62 x 10^{-7}. For three number of pollution source the locations are (6,3), (5,3) and (3,3). The starting objective function value is found to be 3.52 x 10^7 and it gradually minimizes to 3.10 x 10^{-4} at 104th iteration. Some fluctuation in the objective function value could be seen in the middle of the iterations which is the reason for extended increase in iteration. With four number of pollution sources ((6,3), (5,3), (3,3) and (4,5)), the initial objective function value is found to be 5.29 x 10^7 and it the gently converges to the optimal solution at 8.37 x 10^{-5} at 141th iteration. When the number of pollution sources increases, the computational cost increases as the number of variables increases gradually making the problem more complex. For this reason, the objective function value for three and four number of pollution sources initiates almost with the same objective function value and converges towards the solution at the same number of iterations.

6.3.2 Case 2 – Different number of grid cells

In this section, the performance of the model with varying number of grids is studied. A total number of four different study areas with different number of grid cells are adopted. Table 6.8 shows the list of the adopted grid numbers and the details of the pollution sources and the observation wells present in each of the particular area. All the four study areas are simulated for ten years and the source is active for one year at an interval of three months. The different grid numbers imply that the affected aquifer is of varying sizes and each of the grid sizes signifies a distance of 100 m. Thus, the area of the four study areas are 1.04 km^2, 2.09 km^2, 4.06 km^2 and 6.08 km^2.

The increase in the area of the aquifer will enhance the complexity of the optimization model as the search space will be increased. Fig. 6.12 shows the linear relation between the number of grid cells and the number of observation wells. The number of observation wells to be placed on the different study areas are calculated proportional to the number of grids. The hydrological parameters adopted by the four study areas are given in Table 6.9.

Table 6.8: Detail of the different study areas

Study Area	No. of Column x Row	Total No. of Grid cells	Area (km^2)	Location of pollution source (*i,j*)	Location of observation wells (*i,j*)
1	13 x 8	104	1.04	(3,3) (5,3) (6,3)	(3,11) (3,5) (3,7) (4,9) (5,11) (6,5) (6,8) (7,10)
2	11 x 19	209	2.09	(3,4) (6,3) (8,4)	(2,6) (5,6) (7,7) (8,9) (3,9) (6,10) (8,11) (5,12) (3,13) (4,16)
3	14 x 29	406	4.06	(4,5) (7,6) (10,5)	(4,8) (7,11) (10,9) (5,12) (8,14) (10,12) (7,15) (10,17) (10,22) (7,22) (11,14) (7,19) (4, 21) (5,16)
4	16 x 36	608	6.08	(5,6) (8,7) (12,6)	(5,9) (8,11) (13,9) (12,11) (7,13) (11,13) (4,12) (5,15) (8,15) (11,19) (7,18) (11,19) (5,20) (8,22) (11,23) (4,24) (7,25) (13,26) (8,29)

Table 6.9: Hydrological parameters adopted by the four study areas

Parameters	Values
Hydraulic conductivity in x direction, $\boldsymbol{K_{xx}}$(m/s)	0.0002
Hydraulic conductivity in y direction, $\boldsymbol{K_{yy}}$ (m/s)	0.0002
Porosity, $\boldsymbol{\varepsilon}$	0.25
Thickness of the aquifer, *b* (m)	30.5
Longitudinal dispersivity, $\boldsymbol{\alpha_L}$ (m)	40
Transverse dispersivity, $\boldsymbol{\alpha_T}$ (m)	9.6
Time steps, $\Delta \boldsymbol{t}$(months)	3

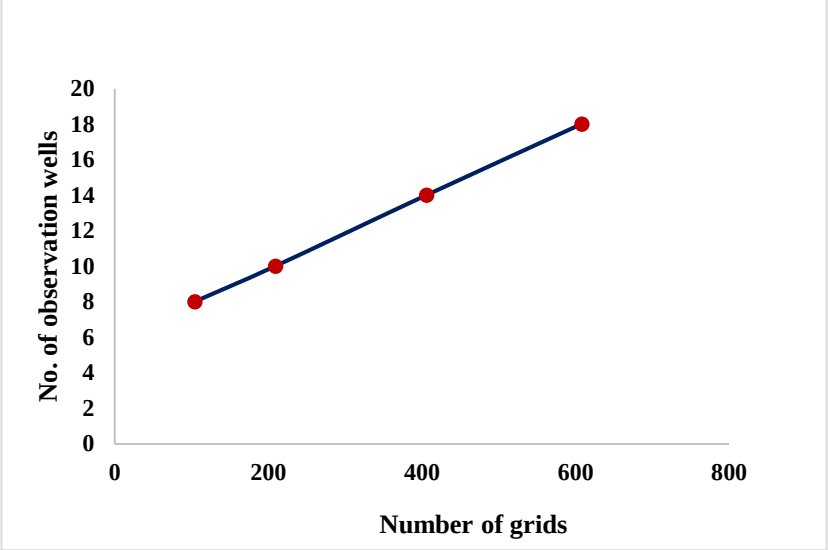


Fig. 6.12: Linear relation plotted between the number of grid and the number of observation wells to be adopted.

The observation wells are spread at different distances of the aquifer the locations of pollution sources the contaminant concentration in different stress periods. Fig. 6.13 (a, b, c and d) shows the locations of pollution sources and the observation wells placed at different grid positions of the aquifer. The pollution source locations are located at different grid points for all the four study areas.

Table 6.10 shows the estimated source location and the fluxes using the present methodology. When compared with the actual ones, it is observed that the source locations are exactly identified and the source fluxes for all the four areas are found to be very close to the actual source flux. As the model identifies the same flux for smaller to the large size study area, the proposed methodology is capable of efficiently identifying the groundwater pollution source locations and source fluxes.

With the increase in the number of grids, the number of function evaluation also increases as depicted in Fig. 6.14. The number of function evaluation taken up by the model to converge towards the optimal solution for 104, 209 and 406 grids aquifer are 5196, 6136 and 7060 respectively. However, in case of 608 number of grid cells, the number of function evaluation accelerates to 11345 but still produces the exact solution. This number of function evaluation of 11345 can still be considered as an acceptable number for a large aquifer equal to 6 km^2.

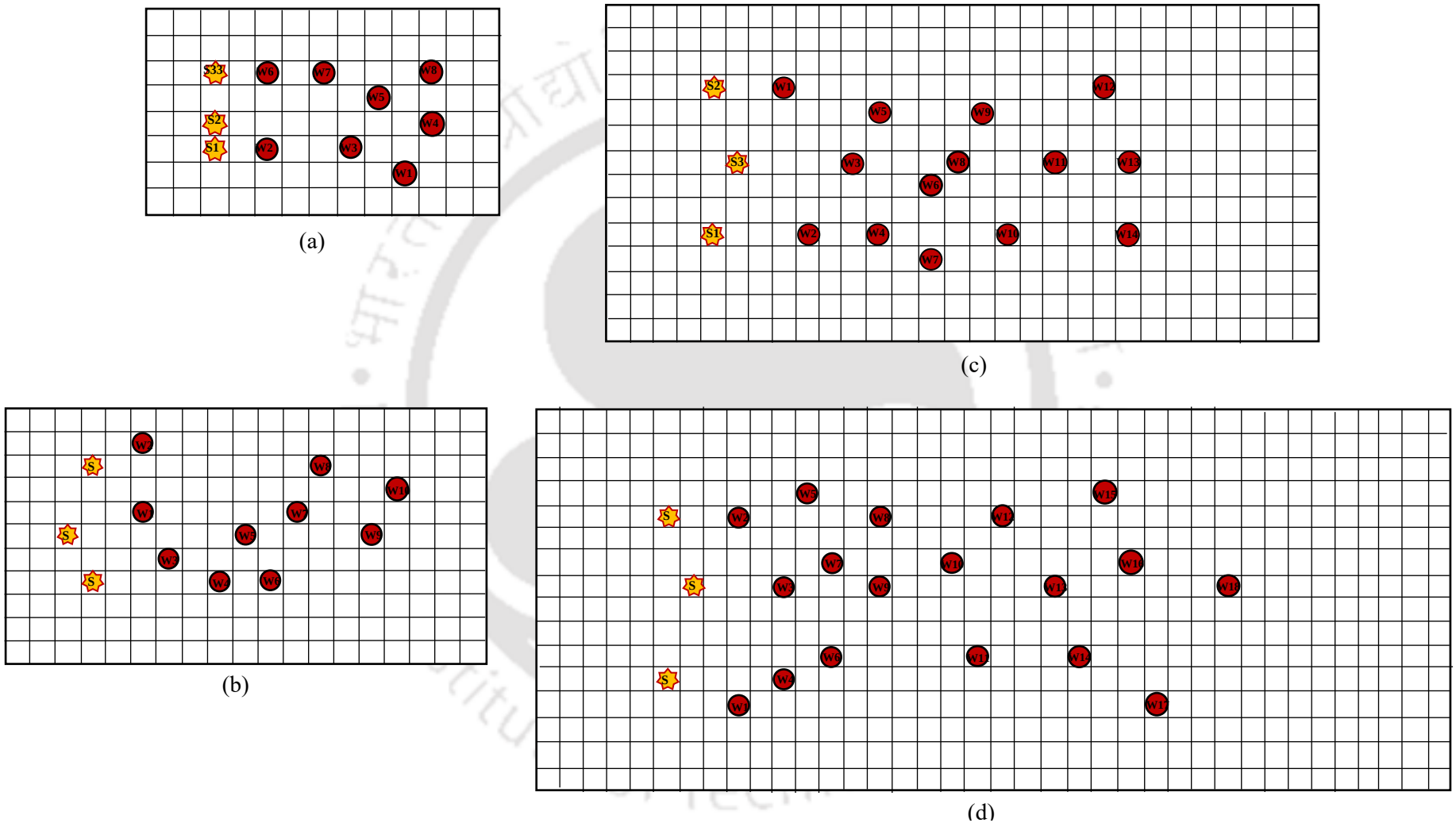


Fig. 6.13: Locations of pollution source and observation well location for (a) 104 grids (b) 209 grids (c) 406 grids and (d) 608 grids

Table 6.10: Source location and the source flux estimated using the present model

No. of Grids	Source Location		Actual source flux				Estimated source flux			
	Actual	Estima-ted	Time Step 1	Time Step 2	Time Step 3	Time Step 4	Time Step 1	Time Step 2	Time Step 3	Time Step 4
104	(3,3)	(3,3)	30	58.8	0	35	30.00	58.78	0.00	34.99
	(5,3)	(5,3)	36.50	0	25.45	12.23	36.51	0.00	25.44	12.23
	(6,3)	(6,3)	47	15	37	0	47.00	14.99	36.99	0.00
209	(3,4)	(3,4)	47	15	37	0	46.62	15.01	36.86	0
	(6,3)	(6,3)	60.30	0	29	37	60.26	0	29.05	37.50
	(8,4)	(8,4)	30	58.8	0	35	30.01	58.68	0	34.89
406	(4,5)	(4,5)	47	15	37	0	46.35	15.35	36.95	0
	(7,6)	(7,6)	60.30	0	29	37	60.26	0	29.35	37.09
	(10,5)	(10,5)	30	58.8	0	35	30.04	58.79	0	34.91
608	(5,6)	(5,6)	47	15	37	0	46.98	15.05	36.94	0.02
	(8,7)	(8,7)	60.30	0	29	37	60.74	0	28.93	37.18
	(12,6)	(12,6)	30	58.8	0	35	30.08	58.54	0.05	34.91

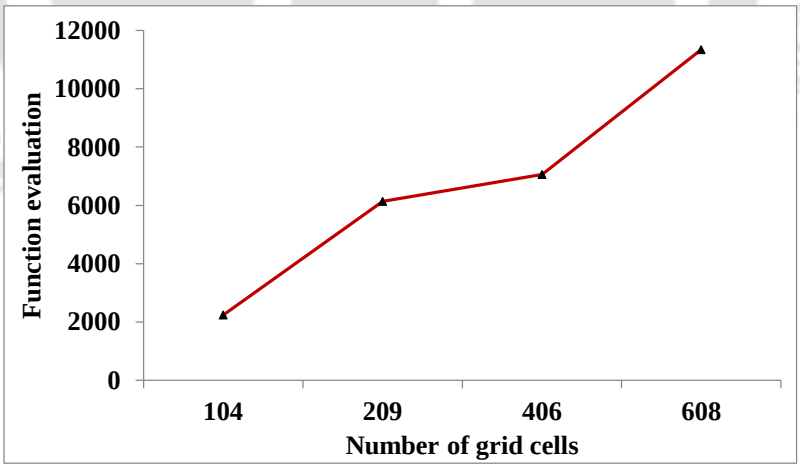


Fig. 6.14: Comparison of function evaluations for different number of grid cells

Fig. 6.15 shows the variation of the objective function for different grid numbers. The initial function value for 104 grids is found to be 4.28 x 10^6. Not much improvement is seen in the function value until 62nd iteration. There is an abrupt drop in the function value in the 51st iteration and converges to the optimal solution in 62nd iteration. For

209 grids, the initial function value is found to be 2.08 x 10^7 whereas for 406 grid size it is 1.93 x 10^7.

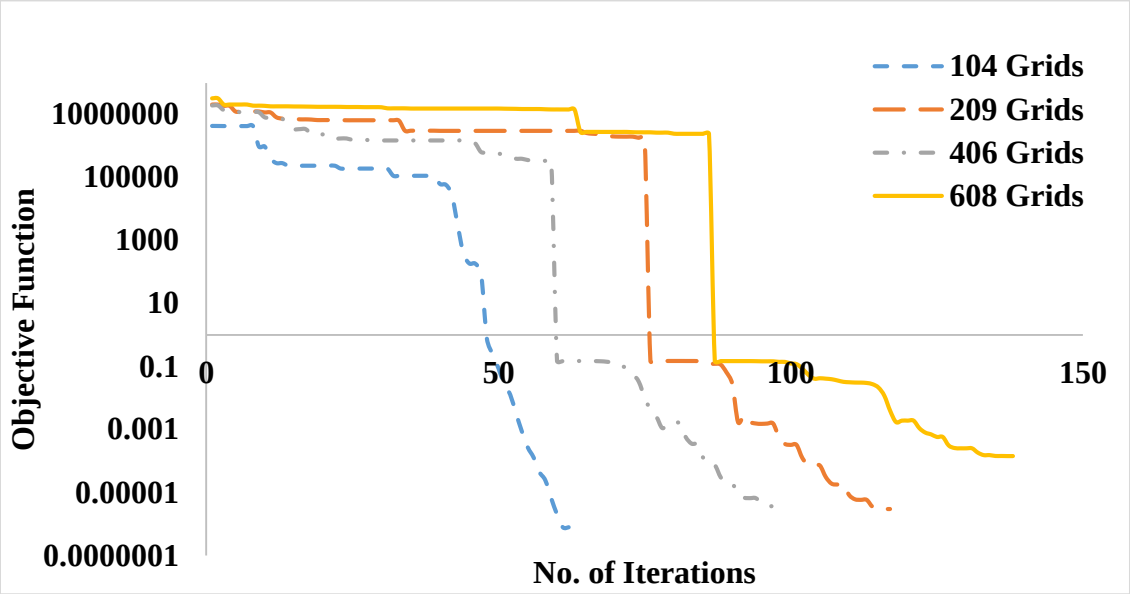


Fig. 6.15: Variation of objective function values for different number of grid cells

However, with few simulations runs, the objective function for 406 grid sizes drop down to 3.67 x 10^{-6} at 99th iteration and for 209 grids the function value drops to 2.98 x10^{-6} at 117th iterations. In case of the 608 grids, the function value has increased by more than double fold and is found to be 3.27 x 10^7 but converges towards the optimal solution at 1.45 x 10^{-4}. The objective function values for 608 grids may not be an improved one but the number of simulation runs is found to be almost in the same trend as 209 and 406 grids as it reaches the exact solution at 139th iteration. This indicates that the convergence power of the present model is quite effective even for the large study area. The overall results signified that introducing pool location could effectively enhance the performance of the present source identification model.

6.4 Summary and Conclusions

The current chapter presented a new source identification methodology where initial populations are selected from a set of the pool comprising of the most probable locations. These locations are further called by the modified GA operator for producing an enhanced optimal solution using the earlier developed methodologies. The robustness of the present methodology is checked by studying different scenarios of source identification problems. In the first case, it is assumed that a different number of pollution sources exists in the aquifer. For one, two and three number of pollution

sources, the optimal solutions are obtained with lesser number of function evaluations but with the increase in the number of sources, a hike in function evaluation is seen but could converge to the optimal solution. It signifies that the function evaluation of the present model is effected with a large number of pollution sources. The second scenario was to see the effect of the model with varying area ranging from 104 grids to 608 grids considering same grid size for all the four-study areas. For this scenario also, the proposed model could converge towards optimal solution successfully and the number of function evaluation ranges from 5196 to 11345 with the increase in number of grids. However, the function evaluations can be considered under acceptable ranges because the area ranges from 1.04 km^2 to 6.08 km^2 which signifies very large aquifer. This shows the robustness of the present model for large study areas.

The next chapter presents a brief summary of the present research work. Based on the proposed methodologies discussed in the earlier and the current chapters, conclusions have been made. The future scope of the work is subsequently discussed.

Chapter 7

Conclusions and Future Scope of the Present Study

In this chapter, a brief description about the summary of the present study has been given. The second section describes the conclusions of the developed methodologies of the present research work. The conclusion is based on the developed methodologies and the results evaluated from each of the illustrative study areas. The third and the final section describe the future scope of the present study that can be carried out.

7.1 Summary of the present study

The identification of groundwater pollution sources becomes an essential task for the researchers working in the field of groundwater management studies. The initial step for identifying the groundwater pollution sources is to identify the source locations and the source fluxes accurately. The groundwater pollution sources can be identified using inverse optimization approach. In this approach, there involves incorporation of simulation model with the optimization model. Although numerous techniques have been proposed by researchers, the linked simulation-optimization approach has proven to be the most desirable technique as it can handle the limitations of other approaches effectively. For a productive linked simulation-optimization model an efficient optimization algorithm is required for solving the model. The gradient-based classical optimization approach and the non-classical approach are the two techniques for solving the inverse optimization model. However, some limitations persist with the gradient based, henceforth non-classical techniques were found to be much superior and have been adopted in various groundwater related studies. The non-classical approaches are a global search technique, e.g. Genetic Algorithm (GA), Simulated Annealing (SA), Particle Swarm Optimization (PSO) etc. which may be adopted to solve the source identification problems. Among all these, GA has been proven to be one of the robust algorithms for finding the global optimal solution of the non-linear non-convex problems.

Considering the above factors, the present research work has presented four different methodologies for identifying groundwater pollution sources. For identifying the groundwater pollution sources, the linked simulation-optimization approach is adopted. Here, the groundwater numerical simulator can be used for simulating the groundwater flow and transport processes. However, linking of such type of numerical groundwater simulator with the optimization model will diminish the computational efficiency of the model. To overcome this computational burden, an ANN-GA based model has been proposed where the ANN model has been used as a surrogate model in place of the numerical groundwater simulator. Though, ANN-GA could effectively identify the source fluxes, but the source locations have to be known to the problem. In real case scenario of groundwater source identification, the number of pollution sources, source locations and the source fluxes are completely unknown. Considering these realistic scenarios, the second methodology is proposed which identifies the number, source locations and source fluxes in an iterative manner. The iterative based model could identify the exact number and the locations of the pollution sources. However, some discrepancy has been seen in the estimated source fluxes. This is so because, groundwater source identification problem being a mixed integer problem, a methodology is required which can handle the discrete variable (source location) and continuous variables (source fluxes) efficiency. As such, a GA-gradient based methodology is developed for identifying the source location as well as the source flux effectively. For accurately identifying the pollution sources, three local location search algorithms i.e. LTS, MS and RMS are also developed. The evaluation of the results shows that an optimal solution can be achieved by using this hybrid optimization model. However, the use of GA in the model is always associated with some randomness, as GA is initiated by randomly generating the initial population. As a result, the number of function evaluations may be large sometimes. Therefore, to reduce the number of function evaluation and also to overcome the randomness encountered in the initial population, the fourth methodology is developed for proposing the initial solutions to the GA model. In this technique, a pool of initial solution is made comprising of most probable locations from which GA selects the best probable locations to start the algorithm. This has reduced the effect of randomness in the initial solution and increases the chances of converging towards the exact optimal solution of the problem.

7.2 Conclusions

Based on the studies carried on the developed methodologies, the following conclusions are drawn as:

- ✓ ANN-GA based model is capable of identifying the pollution sources very effectively even for a large study area. When the relative error was calculated for the estimated source flux with respect to the actual source flux, it could be observed that the error values are all under acceptable range.
- ✓ The present optimization model is capable of selecting the optimal number of the wells for different management period. With the knowledge of plume movement for different years, remedial measures may be adopted.
- ✓ Considering the real case scenario, the linked simulation-optimization model could successfully identify the pollution source locations and the source fluxes.
- ✓ The effect of measurement error in the measured concentration has minimal effect on identification of source location. The optimization model could exactly identify the pollution source locations under the different level of errors. However, in case of source flux, some fluctuations have been observed at the higher level of noise.
- ✓ The source locations and the source fluxes could be successfully identified by the GA-classical based hybrid approach. The source location is a discrete variable and can be effectively identified by GA. The continuous variables i.e. the source fluxescan be estimated by the classical optimization approach. This may be regarded a global-local search technique.
- ✓ The solution obtained by GA has been taken as the inputs for the gradient-based classical approach. As GA is a global technique but does not use the gradient information, it tends to give solution around the actual locations. Thus the efficiency of the proposed hybrid (GA-Classical) technique solely depends on the locations identified by the GA.
- ✓ The local location search algorithms, i.e. Longitudinal-Transverse Search (LTS), Mutation Search (MS) and Ripple Migration Search (RMS) could effectively locate the actual locations of the sources. Out of these three methods, LTS has been emerged to be a computationally efficient one as it requires less function evaluation to converge towards the optimal solution.

- ✓ The comparative study performed between NLP optimization model (Mahar and Datta, 2001) and the LTS algorithm shows that the source flux recovered using LTS were very close to the actual values.
- ✓ The population pool generated using the information obtained from the break-through curve and velocity field of the aquifer enhanced the efficiency of the hybrid approach proposed in the study. Application of the algorithm on source identification problem with different grid sizes and different numbers of pollution sources shows that the model is efficient in identifying the pollution sources available on the aquifer.
- ✓ When the computational efficiency is considered in terms of function evaluation, it is found that the function evaluation of the present model increases abruptly with more number of pollution sources. However, it is quite effective for a large study area.

7.3 Future scope of the present research work

- ✓ The methodologies developed in the present research work proved to be quite effective in identifying the unknown pollution sources of an aquifer. Therefore, the methodology can further be extended for solving groundwater management problem.
- ✓ The starting time of the pollution source activity is assumed to be known to the groundwater source identification problem. However, in most of the real case the activity time is completely unknown. Considering this, the time frame can be included as one of the decision variable for an accurate solution.
- ✓ In the present source identification study, GA is used to solve the inverse optimization model. However, there are numerous non-classical optimization algorithms such Simulated Annealing (SA), Particle Swarm Algorithm (PSA), Firefly algorithm etc. which can be tried to solve the problem.
- ✓ A methodology is developed considering a realistic scenario where limited information about the pollution sources is available. Based on this idea, an experimental study may be carried out.
- ✓ In actual cases, the pollution sources are generally found to be reactive. So, the effect of reactive pollution sources can be included in the developed methodologies.

- ✓ The present methodologies are developed for two-dimensional. Further extension of this study can be performed by considering three-dimensional problem.

References

1. Aguado, E., and Remson, I. (1974). "Groundwater hydraulics in aquifer management."
   J. Hydraulic Division, ASCE, 100(1), 103-118.
2. Ahfeld, D.P., and Heidari, M. (1994). "Application of optimal hydraulic control to groundwater systems." *J. Water Resour. Plann. and Manage*., 120(3), 350-365.
3. Ahlfeld, D.P. (1990). "Two-stage groundwater remediation design." *J. Water Resour. Plann. and Manage*., 116(4), 517-529.
4. Ahlfeld, D.P., Mulvey, J.M., and Pinder, G.F. (1988). "Contaminated groundwater remediation design using simulation, optimization, and sensitivity theory 1. Analysis of a field site." *Water Resour. Res.*, 24(3), 431-441.
5. Alley, W.M., Aguado, E, and Remson, I. (1976). "Aquifer management under transient and steady-state conditions." *JAWRA, J. American Water Resourc. Assoc*., 12(5), 963-973.
6. Aly, A.H., and Peralta, R.C. (1999). "Optimal design of aquifer clean-up systems under uncertainty using a neural network and a genetic algorithm." *Water Resour. Res*., 35(8), 2523-2532.
7. Amirabdollahian, M., and Datta, B. (2013). "Identification of contaminant source characteristics and monitoring network design in groundwater aquifers: an overview." *J. of Environ. Pro*., 4, 26-41.
8. Aral, M. M., Guan, J., and Maslia, M.L. (2001). "Identification of contaminant source location and release history in aquifers." *J. of Hydrol. Eng*., 6(3), 225–234.
9. Aral, M. M., & Guan, J. (1997). "Optimal groundwater remediation system design with well locations selected as decision variables." *Multimedia Environmental Simulations Laboratory*, Georgia Tech, Atlanta, Georgia.
10. Aral, M.M., and Guan, J. (1996). "Genetic algorithms in search of groundwater pollution sources. In advances in groundwater pollution control and remediation." *Springer, Dordrecht*, 347-369.
11. ASCE Task Committee on Application of Artificial Neural Networks in Hydrology. (2000). Artificial neural networks in hydrology. I: Preliminary concepts." *J. Hydrol. Eng.*, 5(2), 115-123.
12. Atmadja, J., and Bagtzoglou, A.C. (2001). "State of the art report on mathematical methods for groundwater pollution source identification." *Environ. Forensics* 2, 205–

214.

13. Ayvaz, M. T. (2016). "A hybrid simulation–optimization approach for solving the areal groundwater pollution source identification problems." *J. Hydrol*. 538, 161-176.
14. Ayvaz, M.T., (2015). "A new simulation-optimization approach for simultaneously identifying the spatial distribution and source fluxes of the areal groundwater pollution sources." 36th IAHR *World Congress*, The Hague, The Netherlands, 1-7.
15. Ayvaz, M. T., and Elçi, A. (2013). "A groundwater management tool for solving the pumping cost minimization problem for the Tahtali watershed (Izmir-Turkey) using hybrid HS-Solver optimization algorithm." *J. of hydro*., 478, 63-76.
16. Ayvaz, M.T. (2010). "A linked simulation-optimization model for solving the unknown groundwater pollution source identification problems" *J. Contam. Hydrol.*, 117(1-4), 46-59.
17. Ayvaz, M.T. (2009). "Application of harmony search algorithm to the solution of groundwater management models." *Adv. in Water Resour*., 32(6), 916-924.
18. Bagtzoglou, A.C., Atmadja, J. (2005). "Mathematical methods for hydrologic inversion: the case of pollution source identification. In: Kassim, T. (Ed.), *Water Pollution*, Handbook of Environmental Chemistry, 3, 65–96.
19. Bagtzoglou, A.C., Dougherty, D.E., and Tompson, A.F. (1992). "Application of particle methods to reliable identification of groundwater pollution sources." *Water Resour. Manage*., 6(1), 15-23.
20. Bashi-Azghadi, S.N., Kerachian, R., Bazargan-Lari, M.R., and Solouki, K. (2010). "Characterizing an unknown pollution source in groundwater resources system using PSVM and PNN." *Expert System with Application*, 37(10), 7154-7161.
21. Bear, J., and Cheng, A.H.D. (2010). "Modeling groundwater flow and contaminant transport." (Vol. 23). *Springer Science & Business Media*.
22. Bhattacharjya, R., and Datta, B. (2009). "ANN-GA-based model for multiple objective management of coastal aquifers." *J. Water Resour. Plann. And Manage.*, 135 (5), 314-322.
23. Bhattacharjya, R.K., Datta, B., and Satish, M.G. (2007). "Artificial neural networks approximation of density dependent saltwater intrusion processes in coastal aquifers." *J. Hydrol. Eng.*, 12(3), 273-282.
24. Bhattacharjya, R.K., and Datta, B. (2005). "Optimal management of coastal aquifer using linked simulation optimization approach.' *Water Resour. Manage*., 19(3), 295-320.

25. Bodardi, J.J., Gupta, A.D., and Jiang, H.Z. (1991). "Search beam method: A promising way to define non-dominated solutions in multi-objective groundwater development." *Inter. J. Water Resour. Dev*., 7(4), 247-258.
26. Borah, T. and Bhattacharjya, R.K. (2014). "Solution of Source Identification problem by using GMS and MATLAB." *J Hydrol. Eng.*, 19(3), 297-304.
27. Burnett, R. D., and Frind, E. O. (1987). "Simulation of contaminant transport in three dimensions: 1. The alternating direction Galerkin technique." *Water Resour. Res*., 23(4), 683-694.
28. Chadalavada, S., Datta, B., and Naidu, R. (2011). "Uncertainty based optimal monitoring network design for a chlorinated hydrocarbon contaminated site." *Environ. monitor. and assess.*, 173(1-4), 929-940.
29. Chandalavada, S., and Datta, B. (2008). "Dynamic optimal monitoring network design for transient transport of pollutants in groundwater aquifers." *Water Resour. Manage*., 22(6), 651-670.
30. Cieniawski, S.E., Eheart, J. W., and Ranjithan, S. (1995). "Using genetic algorithms to solve a multiobjective groundwater monitoring problem." *Water Resour. Res*., 31(2), 399-409.
31. Coppola Jr, E., Szidarovszky, F., Poulton, M., and Charles, E. (2003). "Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping, and climate conditions." *J. Hydrol. Eng*., 8(6), 348-360.
32. Das, A., and Datta, B. (2000). "Optimization based solution of density dependent seawater intrusion in coastal aquifer." *J Hydrol Eng., 5(1), 82-89.*
33. Datta, B., Prakash, O., Cassou, P., and Valetaud, M. (2014). "Optimal unknown pollution source characterization in a contaminated groundwater aquifer: evaluation of a developed dedicated software tool." *J. Geo. Environ. Protect*., 2, 41-51.
34. Datta, B., Prakash, O., and Campbell, S., and 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(14), 4959-4976.
35. Datta, B., Chakrabarty, D., and Dhar, A. (2011). "Identification of unknown groundwater pollution sources using classical optimization with linked simulation." *J. Hydro-Environ. Res.*, 5(1), 25-36.

36. Datta, B., Chakarbarty, D., and Dhar, A. (2009). "Simultaneous identification of unknown groundwater pollution sources and estimation of aquifer parameters." *J. Hydrol.,* 376(1-2), 48-57.
37. Datta, B., Chakrabarty, D., and Dhar, A. (2009). "Optimal dynamic monitoring network design and identification of unknown groundwater pollution sources." *Water Resour. Manage.*,23(10), 2031-2049.
38. Datta, B., and Chakrabarty, D. (2003). "Optimal identification of unknown pollution sources using linked optimization simulation methodology." *Symposium on Advances in Geotechnical Engineering (SAGE 2003). 368-379.*
39. Datta, B., and Dhiman, S.D. (1996). "Chance constrained optimal monitoring network design for pollutants in groundwater." *J. Water Resour. Plann. and Manage.,* 122(3), 180-188.
40. Datta, B., Beegle, J.E., Kavvas, M.L., and Orlob, G.T. (1989). "Development of an expert system embedding pattern recognition techniques for pollution source identification." *National Technical Information Service,* Springfield.
41. Datta, B., and Peralta, R.C. (1986). "Optimal modification of regional potentiometric surface design for groundwater contaminant containment." *Transactions of the ASAE,* 29(6), 1611-1623.
42. Deb, K. (2001). "Multi-objective optimization using evolutionary algorithms." 16, *John Wiley and Sons.*
43. Deb, K., and Goyal, M. (1996). "A combined genetic adaptive search (GeneAS) for engineering design." *Computer Science and informatics,* 26, 30-45.
44. Deb, K. and Agarwal, R.B. (1995). "Simulated binary crossover for continuous search space." Complex Systems, 9, 115-148.
45. Dhar, A., and Datta, B. (2009). "Logic based design of groundwater monitoring network for redundancy reduction." *J. Water Resour. Plann. And Manage.,* 136(1), 88-94.
46. Dhar, A., and Datta, B. (2007). "Multi-objective design of dynamic monitoring networks for detection of groundwater pollution." *J. Water Resour. Plann. And Manage.,* 133(4), 329-338.
47. Diersch., H.J.G. (2002). "FEFLOW finite element subsurface flow and transport simulation system-User's manual/Reference manual/White papers. Release 5.1." WASY Ltd., Berlin.

48. Dorigo, M. (1992). “Optimization, Learning and Natural Algorithms.” PhD thesis, Politechnico di Milano, Italy.
49. Eberhart, R., and Kennedy, J. (1995, October). “A new optimizer using particle swarm theory. In Micro Machine and Human Science, 1995. MHS'95., Proceedings, *6th International Symposium*, 39-43, IEEE.
50. Elango, K., and Rouve, G. (1980). “Aquifer: Finite element linear programming model.” *J. Hydraulic Division*, ASCE, 106(10), 1641-1658.
51. Emch, P.G., and Yeh, W.W. (1998). “Management model for conjunctive use of coastal surface water and ground water.” *J. Water Resour. Plann. Manage.*, 124(3), 129-139.
52. Espinoza, F.P., Minsker, B.S., and Goldberg, D.E. (2005). “Adaptive hybrid genetic algorithm for groundwater remediation design.” *J. Water Resour. Plann. Manage*., 131(1), 14-24.
53. Frind, E.O., and Pinder, G.F. (1973). “Galerkin solution of the inverse problem for aquifer transmissivity.” *Water Resour. Res*., 9(5), 1397-1410.
54. Futagami, T., Tamai, N., and Yatsuzuka, M. (1976). “FEM coupled with LP for water pollution control.” *J. Hydraul. Div.*, ASCE, 102 (7), 881-897.
55. Galeati, G., and Gambolati, G. (1988). “Optimal dewatering schemes in the foundation design of an electronuclear plant.” *Water Resour. Res*., 24(4), 541-552.
56. Gaur, S., Ch, S., Graillot, D., Chahar, B. R., and Kumar, D. N. (2013). “Application of artificial neural networks and particle swarm optimization for the management of groundwater resources.” *Water Resour. Manage.*, 27(3), 927-941.
57. Goldberg, D.E. (1989). “Genetic algorithms in search, optimization and in machine learning.” *Addison Willey*, Bangalore, India.
58. Gorelick, S.M. (1982). “A model for managing sources of groundwater pollution.” 18(4), *Water Resour. Res.*, 773-781.
59. Gorelick, S. M., Evans, B., and Remson, I. (1983). “Identifying Sources of Groundwater Pollution: An Optimization Approach.” *Water Resour. Res.*, 19(3), 779-790.
60. Gorelick, S.M., and Remson, I. (1982). “Optimal dynamic management of groundwater pollutant sources.” *Water Resour. Res*., 18(1), 71-76.
61. Guan, J., Kentel, E., and Aral, M. M. (2008). “Genetic algorithm for constrained optimization models and its application in groundwater resources management.” *J. Water Resour. Plann. Manage*., 134(1), 64-72.

62. Guan, J., and Aral, M.M. (1999). "Progressive genetic algorithm for solution of optimization problems with nonlinear equality and inequality constraints." *Appl. Math. Modeling*, 23, 329-343.
63. Gurarslan, G., and Karahan, H. (2015). "Solving inverse problems of groundwater-pollution-source identification using a differential evolution algorithm." *Hydrolgeo. J.*, 23(6), 1109-1119.
64. He, L., Huang, G.H., and Lu, H.U. (2009). "A coupled simulation-optimization approach for groundwater remediation design under uncertainty: An application to a petroleum-contaminated site." *Environ. poll.*, 157(8-9), 2485-2492.
65. Heidari, M., and Ranjithan, S.R. (1998). "A hybrid optimization approach to the estimation of distributed parameters in two-dimensional confined aquifers." JAWRA, *J. Amer. Water Resour. Assoc.*, 34(4), 909-990.
66. Heidari, M. (1982). "Application of linear system theory and linear programming of groundwater management in Kansas." *JAWRA, J. American Water Resour. Assoc.*, 18(6), 1003-1012.
67. Holland, J.H. (1975). "Adaptation in natural and artificial systems." *University of Michigan*, Michigan.
68. Huyakorn, P.S., and Pinder, G.F. (1983). "Computational methods in subsurface flow." *Academic Press*, Orlando, FL.
69. Huo, Z.L., Feng, S.Y., Kang, S.Z., Cen, S.J., and Ma, Y. (2007). "Simulation of effects of agricultural activities on groundwater level by combining FEFLOW and GIS." *New Zealand J. Agri. Res.*, 50(5), 839-846.
70. Jha, M., and Datta, B. (2015). "Application of dedicated monitoring–network design for unknown pollutant-source identification based on dynamic time warping." *J. Water Resour. Plann. And Manage.*, 141(11), 04015022.
71. Jha, M., and Datta, B. (2012). "Three-dimensional groundwater contamination source identification using adaptive simulated annealing." *J. Hydrol. Eng.*, 18(3), 307-317.
72. Jha, M.K., and Datta, B. (2011). "Simulated Annealing based simulation-optimization approach for identification of unknown contaminant sources in groundwater aquifer." *Desa. Water Treamt.*, 32(1-3), 79-85.
73. Jiang, S., Zhang, Y., Wang, P., Zheng, M. (2013). "An almost parameter-free harmony search algorithm for groundwater pollution source identification." Water Sci. Tech., 68(11), 2359-2366.

74. Kirkpatrick, S., Gelatt, C.D., and Vecchi, M. P. (1983). "Optimization by simulated annealing." *Science*, 220(4598), 671-680.
75. Kleinecke, D. (1971). "Use of linear programming for estimating geo-hydrologic parameters of groundwater basins." *Water Resour. Res.*, 7(2), 367-374.
76. Kollat, J.B., Reed, P.M., and Maxwell, R.M. (2011). "Many-objective groundwater monitoring network design using bias-aware ensemble Kalman filtering, evolutionary optimization and visual analytics." *Water Resour. Res.*, 47(2), W02529.
77. Kollat, J.B., and Reed, P.M. (2007). "A computational scaling analysis of multi-objective evolutionary algorithms in long-term groundwater monitoring applications." *Adv. Water Resour.*, 30(3), 408-419.
78. Kourakos, G., and Mantoglou, A. (2011). "Simulation and multi-objective management of coastal aquifers in semi-arid regions." *Water Resour. Manage.*, 25(4), 1063-1074.
79. Leichombam, L., and Bhattacharjya, R.K. (2016). "Identification of unknown groundwater pollution sources and determination of optimal well locations using ANN-GA based simulation-optimization model." *J Water Resour Prot.*, 8(3), 411-424.
80. Li, Y., and Hilton, A.B C. (2007). "Optimal groundwater monitoring design using an ant colony optimization paradigm." *Environ. Model. and Soft.*, 22(1), 110-116.
81. Li, G.S., Tan, Y.J., Cheng, J., and Wang, X.Q. (2006). "Determining magnitude of groundwater pollution sources by data compatibility analysis." *Inverse Problm. Sci. Eng.*, 14(3), 287–300.
82. Liu, C., and Ball, W.P. (1999). "Application of inverse methods to contaminant source identification from aquitard diffusion profiles at Dover AFB, Delaware." *Water Resour. Res.*, 35(7), 1975-1985.
83. Lin, H.J., Richards, D.R., Talbot, C.A., Yeh, G.T., Cheng, J.R., Cheng, H.P., and Jones, N.L. (1997). "A three-dimensional finite-element computer model for simulating density dependent flow and transport in variable saturated media: Version 3.1." *U.S. Army Eng. Res., Dev., Center*, Visksburg, Miss.
84. Loaiciga, H.A. (1989). "An optimization approach for groundwater quality monitoring network design." *Water Resour. Res.*, 25(8), 1771-1782.

85. Mategaonkar, M., and Eldho, T.I. (2012). "Groundwater remediation optimization using a point collocation method and particle swarm optimization." *Environ. Model. Soft.*, 32, 37-48.
86. Mahar, P.S., and Datta, B. (2001). "Optimal identification of ground-water pollution sources and parameter estimation." *J. of Water Resour. Plann. and Manage.*, 127(1), 20-29.
87. Mahar, P.S., and Datta, B. (2000). "Identification of pollution sources in transient groundwater system." 14(3), 209-227.
88. Mahar, P.S., and Datta, B. (1997). "Optimal monitoring network and groundwater pollution source identification." *J. Water Resour. Plann. and Manage.*, 123(4), 199-207.
89. Mahinthakumar, G., and Sayeed, M, (2006). "Reconstructing groundwater source release histories using hybrid optimization approaches." *Environ. Foren.,* 7(1), 45-54.
90. Mahinthakumar, G., and Sayeed, M. (2005). "Hybrid genetic algorithm-local search methods for solving groundwater source identification inverse problems." *J. Water Resour. Plann. And Manage.,* 13(1), 45-57.
91. Maddock, T. (1972). "Algebraic technological function from simulation model." *Water Resour. Res.*, 8(1), 129-134.
92. Meyer, P.D., Valocchi, A.J., and Eheart, J. W. (1994). "Monitoring network design to provide initial detection of groundwater contamination." *Water Resour. Res., 30*(9), 2647-2659.
93. Meyer, P.D., and Brill, E.D. (1988). "A method for locating wells in a groundwater monitoring network under conditions of uncertainty." *Water Resour. Res., 24(8),* 1277-1282.
94. McCulloch, W.S., and Pitts, W. (1943). "A logical calculus of the ideas immanent in nervous activity." *The bulletin of mathematical biophysics*, 5(4), 115-133.
95. McDonald, M.G. and Harbaugh, A.W. (1984). "A modular three-dimensional finite difference groundwater flow model." USGS Report.
96. McKinney, D. C., and Lin, M. D. (1994). "Genetic algorithm solution of groundwater management models." *Water Resour. Res.*, 30(6), 1897-1906.
97. McPhee, J., and Yeh, W.W.G. (2008). "Groundwater management using model reduction via empirical orthogonal functions." *J. Water Resour. Plann. And Manage., 134(2), 161-170.*

98. Michalak, A. M., and Kitanidis, P. K. (2004). “Estimation of historical groundwater contaminant distribution using the adjoint state method applied to geostatistical inverse modeling.” *Water Resour. Res.*, 40(8).
99. Navarro, A. (1977). “A modified optimization method of estimating aquifer parameters.” *Water Resour. Res.,* 13(6), 935-939.
100. Newman, M., Hatfield, K., Hayworth, J., Rao, P. S. C., and Stauffer, T. (2005). “A hybrid method for inverse characterization of subsurface contaminant flux.” *J. Cont. Hydrol*., 81(1-4), 34-62.
101. Newman, S.P. (1973). “Calibration of distributed parameter groundwater flow models view as multiple-objective decision process under uncertainty.” *Water Resour. Res*., 9(4), 1006-1021.
102. Noori, R., Khakpour, A., Omidvar, B., and Farokhnia, A. (2010). “Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic.” *Expert Systems with Applications*, 37(8), 5856-5862.
103. Nutbrown, D.A. (1975). “Identification of parameters in linear equation of groundwater flow.” *Water Resour. Res*., 11(4), 581-588.
104. Pan, L., and Wu, L. (1998). “A hybrid global optimization method for inverse estimation of hydraulic parameters: Annealing-Simplex Method.” *Water Resour. R*es., 34(9), 2261-2269.
105. Peralta, R.C., and Datta, B. (1990). “Reconnaissance-level alternative optimal ground-water use strategies.” J. *Water Resour. Plann. And Manage*. 116(5), 676-692.
106. Peralta, R.C., Azarmnia, H., and Takahashi, S. (1991). “Embedding and response matrix techniques for maximizing steady-state groundwater extraction: computational comparison.” *Groundwater,* 29(3), 357-364.
107. Peralta, R.C., and Kowalski, K.G. (1986). “Optimizing the rapid evolution of target groundwater potentiometric surfaces.” *Transactions of the ASAE,* 29(4), 940-947.
108. Prakash, O., and Datta, B. (2015). “Optimal characterization of pollutant sources in contaminated aquifers by integrating sequential-monitoring-network design and source identification: methodology and an application in Australia.” *Hydro. J.*, 23(6), 1089-1107.

109. Prakash, O., and Datta, B. (2014). "Characterization of groundwater pollution sources with unknown release time history." *J. Water Resour.and Protec, 6, 337-350.*
110. Prakash, O., and Datta, B. (2013). "Multiobjective monitoring network design for efficient identification of unknown groundwater pollution sources incorporating genetic programming–based monitoring." *J. Hydrol. Eng.*, 19(11), 04014025.
111. Rajeev Gandhi, B.G., Bhattacharjya, R.K., and Satish, M.G. (2016). "Simulation–Optimization-Based Virus Source Identification Model for 3D Unconfined Aquifer Considering Source Locations and Number as Variable." *J. of Hazard., Toxic, and Radio. Waste*, 21(2), 04016019.
112. Rao, S.V.N., Sreenivasulu, V., Bhallamundi, S.M., Thandaveswara, B.S., and Sudheer, K.P. (2004). "Planning groundwater in coastal aquifer." *Hydrol. Sci. J.*, 49(1), 155-170.
113. Reilly, T.E., and Goodman, A.S. (1987). "Analysis of saltwater upcoming beneath a pumping well." *J. of Hydrol.*, 89(3-4, 169-204.
114. Remson, I., Hornberger, G.M., and Molz, F.J. (1971). "Numerical methods in subsurface Hydrology." *Wiley-Interscience*, New York.
115. Reed, P.M., and Minsker, B.S. (2004). "Striking the balance: long-term groundwater monitoring design for conflicting objectives." *J. Water Resour. Plann. Manage.*, 130(2), 140-149.
116. Reed, P., Minsker, B., and Valocchi, A.J. (2000). "Cost-effective long-term groundwater monitoring design using genetic algorithm and global mass interpolation." Water Res. Res., 36(12), 3731-3741.
117. Ritzel, B.J., Eheart, J.W., and Ranjithan, S. (1994). "Using genetic algorithms to solve a multiple objective groundwater pollution containment problem." *Water Resour. Res.*, 30(5), 1589-1603.
118. Rosenwald, G.W., and Green, D.W. (1974). "A method for determining the optimum location of wells in a reservoir using mixed-integer programming." *Soc. of Petrol. Enginr. J.*, 14(1), 44-54.
119. Sharma, Y.C., Mukherjee, A.K., Srivastava, J., Mahato, M., and Singh, T. N. (2008). "Prediction of various parameters of a river for assessment of water quality by an intelligent technique." *Chemical Product and Process Modeling*, 3(1).

120. Shieh, H.J., and Peralta, R.C. (2005). "Optimal in-situ bioremediation design by hybrid genetic algorithm-simulated annealing." *J. Water Resour Plann. Manage*., 131(1) 67-78.
121. Singh, R.M., and Datta, B. (2007). "Artificial neural network modeling for identification of unknown pollution sources in groundwater with partially missing concentration observation data." *Water Resour. Manage*., 21(3), 557-572.
122. Singh, R.M., and Datta, B. (2006). "Identification of groundwater pollution sources using GA-based linked simulation optimization model." *J. Hydrol. Eng.*, 11(2), 101-109.
123. Singh, R.M., Datta, B., and Jain, A. (2004). "Identification of unknown groundwater pollution sources using artificial neural networks." *J. Water Resour. Plann. Manage.*, *130*(6), 506-514.
124. Singh, R.M., Datta, B., and Jain, A. (2002). "Identification of unknown groundwater pollution sources using artificial neural network." Proceedings of the International Conference on *Advances in Civil Engineering* (ACE-2002), Kharagpur, India: Indian Institute of Technology, 83–93.
125. Skaggs, T.H., and Z.J. Kabala. (1995). "Recovering the release history of a groundwater contaminant." *Water Resour. Res*., 30(1), 71-79.
126. Snodgrass, M.F., and Kitanidis, P.K. (1997). "A geostatistical approach to contaminant source identification." *Water Resour. Res*., 33(4), 537-546.
127. Sun, A.Y., Painter, S.L, and Wittmeyer, G.W. (2006b). "A robust approach for contaminant source location and release history recovery." J. *Contam. Hydrol.*, 88(3-4), 29-44.
128. Tung, Y.K., and Kolterman, C.E. (1985). "Some computational experiences embedding technique for groundwater management." *Groundwater*, 23(4), 455-464.
129. Tyson, H.N., and Weber, E.M. (1964). "Ground-water management for the nation's future: computer simulation of ground-water basins." *J. of the Hydraul. Div*., 90(4), 59-77.
130. Voss, C.I. (1984). "A finite-element simulation model for saturated-unsaturated, fluid-density-dependent ground-water flow with energy transport or chemically-reactive single-species solute transport." *US Geological Survey*, 409.
131. Wagner, B.J. (1992). "Simultaneously parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modelling." *J. Hydrol*. 135, 275-303.

132. Wagner, B.J., and Gorelick, S.M. (1986). "A statistical methodology for estimating transport parameters: theory and applications to one-dimensional advective-dispersive systems." *Water Resour. Res.*, 22(8), 1303-1315.
133. Wang, M., and Zheng, C. (1998). "Ground water management optimization using genetic algorithms and simulated annealing: Formulation and comparison." *JAWRA J. Amer. Water Resour. Assoc.*, 34(3), 519-530.
134. Wilby, R.L., Wigley, T.M.L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., and Wilks, D. S. (1998). "Statistical downscaling of general circulation model output: a comparison of methods." *Water Resour. Res.*, 34(11), 2995-3008.
135. Willis, R., and Finney, B.A. (1985). "Optical control of nonlinear groundwater hydraulics: Theoretical development and numerical experiments." *Water Resour. Res.*, 21(10), 1476-1482.
136. Willis, R., and Liu, P. (1984). "Optimization model for groundwater planning." *J. Water Resour. Plann. And Manage.*, 110(3), 333-347.
137. Willis, R., and Newman, B.A. (1977). "Management model for groundwater development." *J. Water Resour. Plann. And Manage.*, 103(1), 159-171.
138. Willis, R. (1976). "A management model for determining the effluent standards for the artificial recharge of municipal and industrial wastewaters." *American Water Res. Assoc.* (AWRA), Bethesda, Md., 227-308.
139. WWAP (World Water Assessment Programme). (2012). "The United Nations World Water Development Report4: Managing water under uncertainty and risk." Paris, UNESCO.
140. Yazicigil, H., Al-Layla, R.I., and Jong, R.L. (1987). "Optimal management of the regional quifer in eastern Saudi Arabia." *JAWRA, J. American Water Resour. Assoc.*, 23(3), 423-434.
141. Yeh, H. D., Lin, C. C., and Yang, B. J. (2014). "Applying hybrid heuristic approach to identify contaminant source information in transient groundwater flow systems." *Math. Prob. Eng.*, 2014.
142. Yeh, H. Der., Chang, T.H., and Lin, Y.C. (2007). "Groundwater contaminant source identification by a hybrid heuristic approach." *Water Resour*. Res. 43, 1–16.
143. Yeh, M.S., Lin, Y.P., and Chang, L.C. (2006). "Designing an optimal multivariate geostatistical groundwater quality monitoring network using factorial kriging and genetic algorithm." *J. Environ. Geo.*, 50(1), 101-121.

144. Yeh, W.W.G. (1986). “Review of parameter identification procedures in groundwater hydrology: The inverse problem.” *Water Resour. Res.*, 22(2), 95-108.

145. Yenigul, N.B., Elfeki, A.M., Gehrels, J.C., van den Akker, C., Hensbergen, A.T., and Dekking, F. M. (2005). “Reliability assessment of groundwater monitoring networks at landfill sites.” *J. of Hydrol.*, 308(1-4), 1-17.

146. Zhao, C., Poulet, T., and Regenauer-Lieb, K. (2015). “Numerical modeling of toxic non-aqueous phase liquid removal from contaminated groundwater systems: mesh effect and discretization error estimation.” *Intern. J for numer and Analy. Metho. In Geomecha*., 39(6), 571-593.

147. Zhao, C., Wang, Y., Chen, X.I., and Li, B. (2005). “Simulation of the effects of groundwater level on vegetation change by combining FEFLOW software.” Ecol. Model., 187(2-3), 341-351.

List of Publications

1) Leichombam, S., and Bhattacharjya, Rajib Kumar (2018),“A Hybrid Optimization Methodology for Optimal Identification of Pollution Sources Considering the Source Locations and Fluxes as Unknown“. Journal of Hazardous, Toxic and Radioactive Waste, ASCE, (DOI: 10.1061/(ASCE)HZ.2153-5515.0000431.)
2) Leichombam, S., and Bhattacharjya, Rajib Kumar (2016), "Identification of Unknown Groundwater Pollution Sources and Determination of Optimal Well Locations Using ANN-GA Based Simulation-Optimization Model", Journal of Water Resource and Protection, 8(3), 411-424.
3) Leichombam Sophia, and Bhattacharjya, Rajib Kumar (2015), "Identification of Unknown Groundwater Pollution Sources and Simultaneous Design of Monitoring Network Using ANN-GA based Simulation Optimization Model.", 9th World Congress of European Water Resources Association, Istanbul, Turkey, 9-13, June 2015.
4) Leichombam Sophia, and Bhattacharjya, Rajib Kumar (2015), "Identification of Unknown Groundwater Pollution Sources Using GMS-GA based Linked Simulation Optimization Model.", Emerging Trends in Science and Engineering Research (ETSER-2015), NIT Manipur, 2-4, December 2015.