Estimation of Contaminant Transport Parameters using Genetic Algorithms

Anirudh Vemula[1], K.S. Hari Prasad[2] and D. N. Ratha[3]

Abstract

Reliable estimates of contaminant transport parameters are of great importance for arriving at effective remedial measures for groundwater decontamination. The transport parameters are usually estimated by conducting column experiments on saturated soil columns and the optimal parameters are obtained by using a suitable optimization technique. In the present study, the applicability of Genetic Algorithm (GA) optimization technique is assessed in estimating the transport parameters. For this purpose, the genetic algorithms technique is coupled with the McCormack scheme based numerical technique for solving Advection – Dispersion equation in Groundwater. The results indicate that GA technique is an efficient and reliable technique for estimating the transport parameters. GA avoids the subjectivity, large computational time and ill-posed ness often associated with conventional optimization techniques.

Introduction

Groundwater is a valuable natural resource of water which is being used extensively for drinking, irrigation and industrial purposes. Due to the increase in population and rapid industrialization, this resource has been increasingly contaminated. The main sources of ground water contamination are waste disposal sites, industrial effluents, agricultural and urban watersheds [Ward and Giger, 1985]. Mathematical models have been effectively used to study the transport of contaminants in groundwater and to take appropriate measures for remediation [Charbeneau, 2000]. With greater model sophistication comes a need for more intensive data requirements and precision in model prediction hinge on our ability to accurately determine the model parameter [Kool et al., 1987]. The transport parameters of a soil medium are commonly determined by imposing restrictive initial and boundary conditions so that the governing equations can be inverted by analytical or semi-analytical methods. However, these direct inversion methods also have a number of limitations which restrict their practicality, in particular when used for calibrating field scale models. The direct

[1]*Undergraduate Student, Department of Civil Engg., Indian Institute of Technology, Roorkee.*
[2]*Associate Professor, Department of Civil Engg., Indian Institute of Technology, Roorkee.*
[3]*Doctoral Student, Department of Civil Engg., Indian Institute of Technology, Roorkee.*

methods are time consuming and hence costly owing to the need to meet conditions requisite for the explicit calculation of model parameters. An alternative and more flexible approach to solving the inverse problem is to employ parameter estimation techniques. In such an approach, the model parameters are estimated by minimizing the deviations between the model predicted and field observed output. Contrary to direct methods, the optimization approach does not put any inherent constraint on the form of complexity of the model, on the stipulation of initial and boundary conditions. Thus a major advantage is that experimental conditions can be selected on the basis of convenience and expeditiousness. Variety of optimization techniques such as Steepest Descent, Newton's method, Quasi Newton's method, Conjugate Gradient methods have been used to estimate the flow and transport parameters [Kool and Parker 1988, Kool et al., 1987]. Most of these techniques involve the computation of first derivative of the objective function which some times may not be accurate or are difficult to evaluate. In addition, these traditional methods have the tendency to converge to local minimum. Recently, non traditional optimization algorithms such as Genetic algorithms have been effectively used as optimization tools in water resources [McKinney and Lin (1994)]. Genetic algorithms are heuristic (non exact), probabilistic (stochastic), Combinatorial (discrete) search based optimization techniques and the continuing price/performance improvements of computational systems have made them attractive for various types of optimization problems [Samuel and Jha, 2003]. In the present study, genetic algorithms have been used to estimate the transport parameters in this present work.

Contaminant Transport in one Dimension

The one dimensional form of advection-dispersion equation for non reactive dissolved constituent in saturated, homogeneous, isotropic material under steady state uniform flow is described by (Charbeneau, 2000)

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} - v\frac{\partial c}{\partial x} \tag{1}$$

where c is concentration of contaminant (ML^{-3}). D is the hydrodynamic dispersion coefficient, v is fluid velocity (LT^{-1}), x is the spatial coordinate (L) and t is the time coordinate (T). Usually D is expressed in terms of two components (Bear, 1972) as

$$D = D_m + a_L v \tag{2}$$

where D_m is the coefficient of molecular diffusion (L^2/T) and a_L is the longitudinal dispersivity (L).

Initial and Boundary Conditions

Eqn. (1) needs the initial and boundary conditions to obtain a unique solution for a particular problem. Initially, before the injection of a contaminant, the concentration in the domain is assumed to be zero.

i.e. $$t = 0, c = 0, 0 \le x \le \infty \tag{3}$$

At time t = 0, a contaminant with concentration c_o is added at the inlet boundary, *i.e.*,

$$t \ge 0,\ c = c_o,\ x = 0 \tag{4}$$

Further, it is assumed that for away from the source, the ground water is not contaminated, i.e.,

$$t \ge 0,\ c = 0,\ x \to \infty \tag{5}$$

Solution of eqns. (1) subject to eqns. (3) to (5) provide the variation of contaminant concentration as a function of space and time. Many analytical and numerical solutions are available in literature for solving eqn. (1) [Ogata and Banks, (1961), Chaudhury (1993)]. Recently Pradeep Verma et al., (2006) developed a MacCormack scheme based numerical model for solving eqn. (1), which provides accurate results for wide range of Peclet numbers and is discussed below.

Numerical Solution

The numerical solution is carried out in two steps. First, the advective part is solved using MacCormack scheme to get the nodal advective concentrations in the solution domain. The nodal advective concentrations so obtained are used in the dispersive part as source/sink terms to obtain the final concentrations [Pradeep Verma et al., 2006]. The procedure is described in detail in the following section.

Solution of the Advective Part

Figure 1 shows the nodal representation to describe the MacCormack scheme.

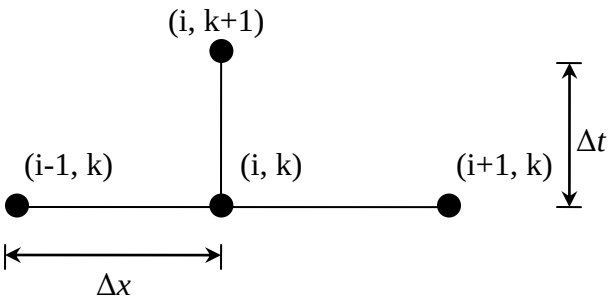

Figure 1 – Nodal representation of MacCormack scheme

For a typical interior node '*i*', the predictor and corrector steps of the MacCormack scheme can be written as: [Chaudhry (1993), Bhallamudi and Singh (1996)]

Predictor: (Using Forward difference)

$$\frac{C_i^* - C_i^k}{\Delta t} = -v\left(\frac{C_i^k - C_{i-1}^k}{\Delta x}\right) \qquad C_i^* \text{ being the predicted concentration.} \quad (6)$$

Corrector: (Using Backward difference)

$$\frac{C_i^{**} - C_i^*}{\Delta t} = -v\left(\frac{C_{i+1}^* - C_i^*}{\Delta x}\right) \qquad C_i^{**} \text{ is the corrector concentration.} \quad (7)$$

Hence, the corrected value of advective concentration at unknown time level ${}_aC_i^{k+1}$ is given by,

$${}_aC_i^{k+1} = \frac{1}{2}\left(C_i^{**} + C_i^k\right) \quad (8)$$

In Eqs. (6) to (8), superscript '*k*' denotes the time level at which the solution is known and '*k+1*' denotes the time level at which solution is sought, Δx is the spatial grid size and Δt is the time increment. For the spatial derivative, backward-differencing is used in the predictor step and forward-differencing is used in the corrector step. MacCormack scheme is an explicit scheme and its stability is governed by the Courant number α which is defined as

$$\alpha = v\frac{\Delta t}{\Delta x} \quad (9)$$

MacCormack scheme is stable for $\alpha \leq 1$.

Solution of Dispersive Part:

The dispersive part is written as

$$\frac{\partial C}{\partial t} - D_x \frac{\partial^2 C}{\partial x^2} = 0 \tag{10}$$

For a typical interior node, a fully implicit finite difference scheme for the solution of dispersive part can be written as

$$\frac{{}_d C_i^{k+1} - {}_a C_i^{k+1}}{\Delta t} - D_x \frac{{}_d C_{i+1}^{k+1} - 2{}_d C_i^{k+1} + {}_d C_{i-1}^{k+1}}{(\Delta x)^2} = 0 \tag{11}$$

It should be noted that in Eqn.(11), the advective concentration obtained from MacCormack scheme is used in the time derivative term which acts as sink term. Rearranging Eqn.(11) to bring all the unknown terms on the left hand side and all the known terms on the right hand side, one gets

$$\left(-\frac{D_x}{(\Delta x)^2}\right) C_{i+1}^{k+1} + \left(\frac{1}{\Delta t} - \frac{2D_x}{(\Delta x)^2}\right) C_i^{k+1} + \left(-\frac{D_x}{(\Delta x)^2}\right) C_{i-1}^{k+1} = \left(-\frac{1}{\Delta t}\right)_a C_i^{k+1} \tag{12}$$

The parameter that characterizes the effect of dispersion is the Peclet number which is defined as

$$P_c = \frac{u\Delta x}{D_s} \tag{13}$$

Computationally, a flow is considered to be dispersion dominated if $P_c \leq 1$ and advection dominated if $P_c \geq 5$ (Anderson, 1979).

Inverse Problem

Knowing the parameters a_L, D_m and v eqn. (1) can be solved either analytically or numerically to obtain the solute concentrations as a function of space and time. This constitutes the forward problem. In contrast, the inverse problem involves the estimation of the parameters a_L, D_m and v from the observed concentration data in an experiment. Inverse procedure consists of estimating the model parameters iteratively such that the deviation between the model predicted concentrations and those observed for a stipulated initial and boundary conditions, is minimized.

Formulation of objective Function

The objective function (fitness function in Genetic algorithm terminology) is defined as the sum of the squares of difference between the models' predicted and experimentally observed contaminant concentrations. If, c_{ij}^{obs} is the experimentally observed concentration at time t_i and at

location x_j, and c_{pre}^{ij} is the corresponding model predicted concentrations, then the fitness function SSD can be defined as

$$SSD = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(c_{ij}^{obs} - c_{ij}^{pre} \right)^2 \tag{14}$$

where I, J is the total number of time and space observations respectively. In eqn. (14), c_{pre}^{ij} depend on the model parameters v and D for a stipulated initial v and D boundary conditions.

Estimation of Transport Parameters

The parameter a_L, D_m and v are estimated by minimizing the fitness function (14). In the present study, genetic algorithm optimization technique is used to estimate these parameters. Genetic algorithms are computerized search and optimization algorithms based on the mechanics of natural selection and natural genetics [Goldberg, 1989]. Genetic algorithms are robust search methods that seek to reproduce mathematically the mechanisms of natural selection and population genetics according to the biological processes of survival and adaptation.

Genetic Representation

Initially, a population of parameters proportional to the total string length is generated using a random generator. Generally, genetic algorithms have been developed by using binary coding in which a string (or chromosome) is represented by a string of binary bits that can encode integers, real numbers. In this study, the parameters are encoded as substrings of binary digits having a specific length. These substrings are joined together to form longer strings representing a solution. The length of the substring (*i*) is determined according to the desired solution accuracy and is dependent on the range of the parameter and the precision requirement of the parameter. McKinney and Lin (1994), gave the relationship between the substring length and the obtainable precision as

$$\left(D_k - C_k\right) 10^{\alpha_k} \leq 2^l - 1 \tag{15}$$

Where, i_k = desired precision for the i^{th} parameter, i is the substring length, C_k and D_k = lower and upper bounds of the k^{th} parameter, k = index for the parameter:

The value of k^{th} parameter is decoded using the following equation.

$$Z_k = C_k + [(D_k - C_k)/(2^l - 1)] \sum_{j=0}^{l} I_j \, 2^l \tag{16}$$

Where Z_k = value of k^{th} parameter and l_j = binary digit at the j^{th} position in the string starting from right. Once the initial population is randomly generated, the fitness of each string is determined using eqn. (16).

Reproduction

The string generated in the initial population is chosen for participation in the reproduction process based on their fitness values. Many selection schemes such as deterministic sampling, stochastic sampling with or without replacement, stochastic tournament selection and fitness proportionate selection, are used for reproduction (Goldberg, 1989). In this work, the fitness proportionate selection is chosen, where a string is selected for the reproduction process with a probability proportional to its fitness. Thus the probability p_i of an individual member string i being selected is given by

$$p_i = \frac{S_i}{\sum_{i=1}^{p} S_i} \tag{17}$$

where s_i = fitness of individual string *i*, and p is the population size. The scheme is implemented with the simulation of a roulette wheel with its circumference marked for each string proportionate to the string's fitness. The roulette wheel is rotated p times, each time selecting a copy of the string chosen by the roulette wheel printer. As the circumference of the wheel is marked according to a string's fitness, the roulette wheel mechanism makes S_i / S_{mean} copies of the i^{th} string in the re production where S_{mean} is the average fitness of the population given as

$$S_{mean} = \frac{1}{p} \sum_{i=1}^{p} S_i \tag{18}$$

The string with a higher fitness value represents a larger range in the cumulative probability values, and hence has a higher probability of being copies into the mating pool.

Mating (Cross over)

The general theory behind the cross over operation is that by exchanging important building blocks between two strings that perform well, the GA attempts to create new strings that preserve the best material from two parent strings. The cross over is a recombinant operator that selects two strings from the mating pool at random and cuts their bits at a randomly chosen position. This produces two "head" segments and two "tail" segments. The tail segments are then swapped over to produce two new full length strings. The number of strings participating in mating depends on cross over probability. If a cross over probability of p_c is used, only $100 \times p_c$ percent strings in the population are used in the cross over operation. Cross over has a wide range of possible types, i.e., one point, multipoint, uniform, intermediate arithmetical and entered arithmetical (Goldberg 1989). The effect of cross over depends on the site at which cross over takes place.

Mutation

Mutation is an important process that permits new genetic material to be introduced to a population. A mutation probability is specified that permits random mutations to be made to individual genes (e.g. changing *1* to *0* and vice versa for binary GAs). Mutation operator facilities the convergence towards an optimal solution even if initial population is far from the optimal solution. The binary GAs used mutation probability at very low rates ranging from 0.001 to 0.01. Fig. 3 shows the flow chart of the estimation of transport parameters usually Genetic algorithms. A computer code is written in C language to implement the algorithm.

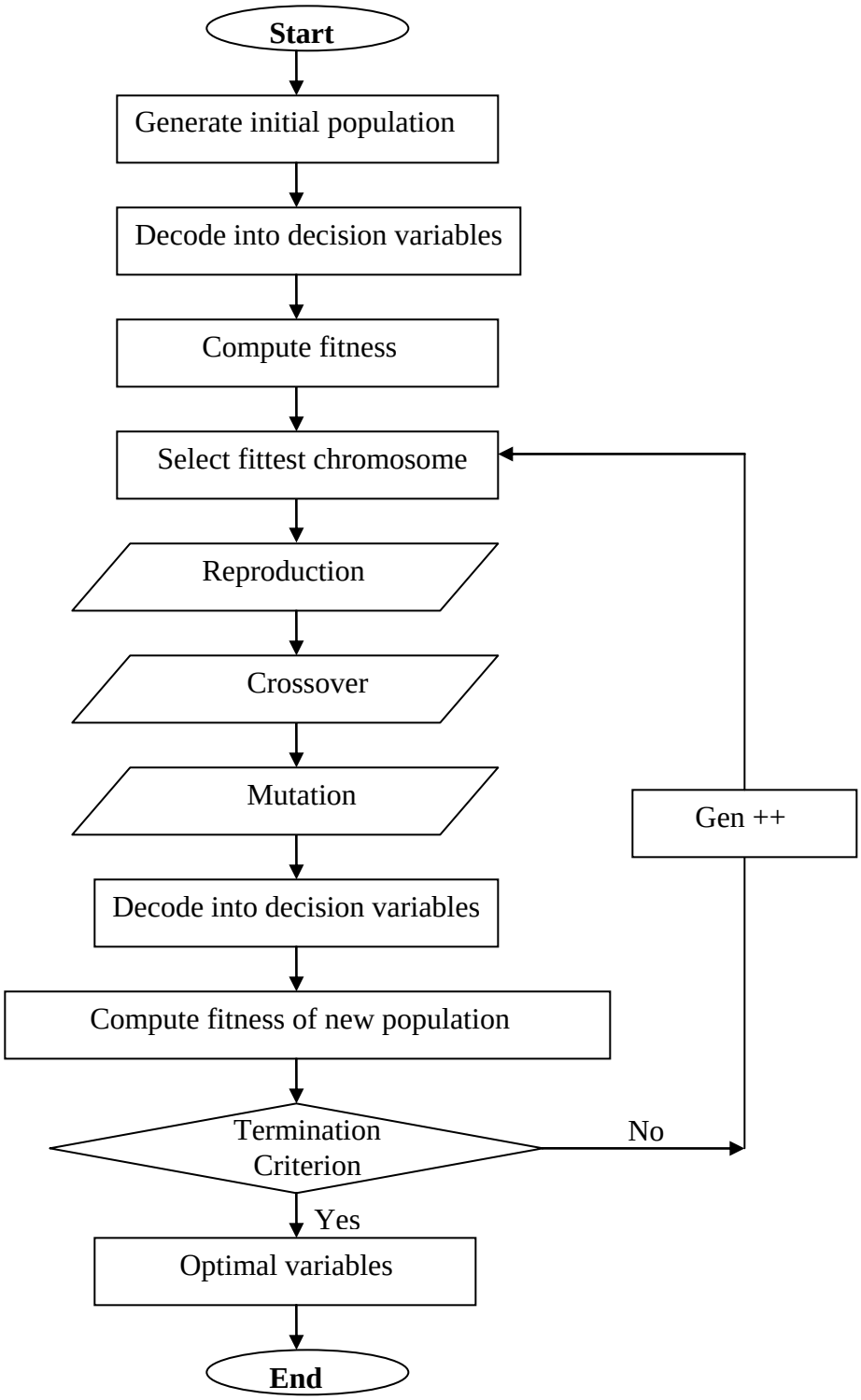

Figure 2 – Flowchart of Genetic Algorithm

Results and Discussion

In order to evaluate the performance of the genetic algorithm assisted parameter estimation technique, hypothetical transient concentration data is generated for different values of the parameters a_L, D_m and v. For the generation of concentrations data, initially the concentration in groundwater is assumed to be zero. A continuous source concentration (C_0) of 100 mg/l is applied at one end of the domain and transient concentration data is generated. The data generated involve both advective and dispersive dominated concentration profiles. [Hari Prasad et al. 2004]. Table 1 presents the different cases considered in the study along with the corresponding true values of the parameters $a_{L1} v$ and D_m used for generating data.

Table 1: True values of parameters used for Concentration Data Generation

	a_L (cm)	v (cm/day)	D_m (cm^2/day)
Case 1	0.05	4	2
Case 2	0.05	60	2
Case 3	15	4	2
Case 4	15	60	2

Transient concentration data is generated for all the four cases and these data is used in estimating the parameters using Genetic Algorithms. The genetic algorithm parameters used in optimization are: population size = 100; String length = 60; Number of Generations = 100; Mutation probability = 0.01.

Table 2 presents the parameter estimation details for the four cases considered. It is clear from Table 2 that the parameter estimates obtained from the Genetic algorithms are almost identical to the true values of the parameters. Figs. 3 to 6 show the comparison of model predicted concentrations using the optimum parameters with the observed concentrations for all the four cases. It can be seen from Figs. 3 to 6 that model predicted concentrations match very well with the observed concentrations resulting in minimum residual mean square error (RMS error). Thus, it can be concluded that genetic algorithms are robust in estimating solute transport parameters and can be used as an effective tool for parameter estimation.

Table 2: Parameter Estimates by Genetic Algorithms

Parameter	True value	Estimated value
Case 1		
a_L (cm)	0.05	0.0498
v (cm/day)	4	4.0001
D_m (cm^2/day)	2	2.0005
RMS Error = 0.0032 $(mg/l)^2$		
Case 2		
a_L (cm)	0.05	0.05001
v (cm/day)	60	60.0007
D_m (cm^2/day)	2	1.9998
RMS Error = 0.0026 $(mg/l)^2$		
Case 3		
a_L (cm)	15	14.9997
v (cm/day)	4	4.0003
D_m (cm^2/day)	2	1.9997
RMS Error = 0.0028 $(mg/l)^2$		
Case 4		
a_L	15	15.0003
v	60	60.0002
D_m	2	2.0002
RMS Error = 0.0022 $(mg/l)^2$		

Conclusions

In the present study, the applicability of Genetic Algorithm (GA) optimization technique is assessed in estimating the groundwater contaminant transport parameters. For this purpose, the genetic algorithms technique is coupled with the McCormack scheme based numerical technique for solving Advection – Dispersion equation in Groundwater. The performance of the genetic algorithm based optimization technique is studies by estimating the transport parameters a_L, D_m and v from synthetically generated transient concentration data. The data includes both advective and dispersive

dominated data. It is observed that in all the cases, parameters estimated by genetic algorithms are almost identical to the true values. The results indicate that GA technique is an efficient and reliable technique for estimating the transport parameters. GA avoids the subjectivity, large computational time and ill-posed ness often associated with conventional optimization technique.

References

1. Anderson, M. P. (1979) Using models to simulate the movement of contaminants through groundwater flow systems, Crit. Rev., Control 9(2): 97-156.
2. Bear, J., (1972) Hydraulics of Groundwater, McGraw-Hill, Englewood Cliffs, NJ.
3. Bhallamudi, S.M. and Singh. V. (1996) Complete Hydrodynamic Border-strip Irrigation, Model. J. Irrigation and Drainage Eng. ASCE, Vol. 122(4), pp, 189-197.
4. Charbeneau, R. J., "Groundwater Hydraulics and Pollutant Transport", Prentice Hall, (2000).
5. Chaudhry, M. H. (1993). Open Channel Flow. Prentice-Hall, Englewood Cliffs, NJ.
6. Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley.
7. Hari Prasad K. S., Mohan Reddy M. M. and Ojha, C. S. P. (2004) Stochastic Analysis of Solute Transport in Soils, Hydrology Journal, 27(3-4):29-40.
8. Kool, J.B. and Parker, J.C. (1988) Analysis of the inverse problem for transient unsaturated flow. Water Resources Research, 24(6), 817-830.
9. Kool, J.B., Parker, J.C. and Van Genuchten, M.Th. (1987). Parameter estimation for unsaturated flow and transport models- A review. Journal of Hydrology, 91, 255-293.
10. McKinney, D. C., and Lin, M. D. (1994). ''Genetic algorithm solution of groundwater management models.'' Water Resour. Res., 30(6), 1897– 1906.
11. Ogata, A., and Banks, R.B.(1961). A solution of the differential equation of longitudinal dispersion in porous media, Prof. paper No. 411-a, U.S. Geological Survey, Washington, D.C.
12. Pradeep Verma, Hari Prasad, K. S. and Ojha, C. S. P., (2006), MacCormack Scheme Based Numerical Solution of Advection – Dispersion Equation, ISH Journal of Hydraulic Engineering, 12(1):27-38.

13. Samuel, M. P., and Jha, M. K. (2003). “Estimation of Aquifer Parameters from Pumping Test Data by Genetic Algorithm Optimization Technique”. Jl. Irrig. and Drainage Engg. ASCE,129, 348- 359.

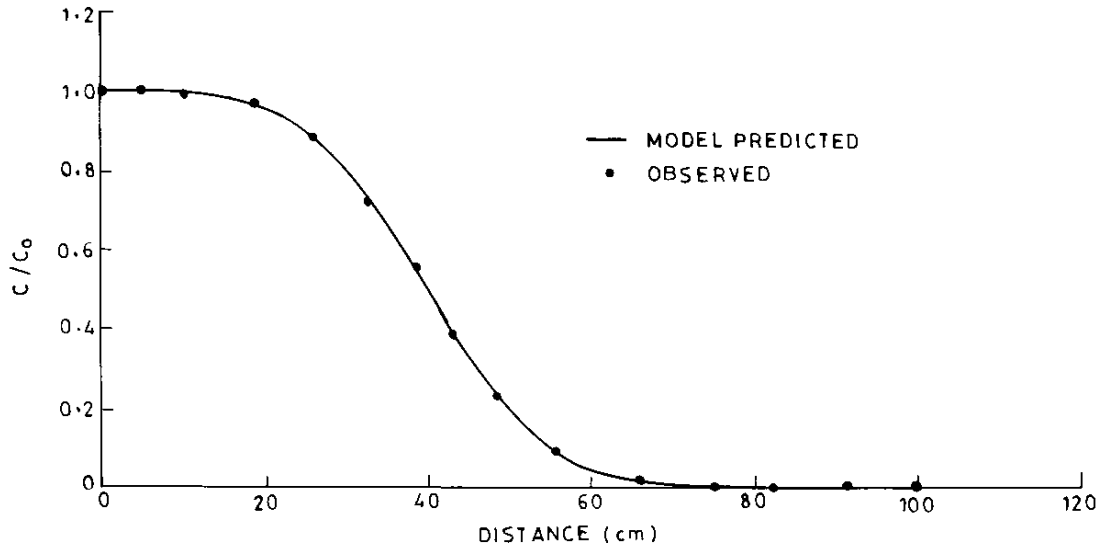

Fig 3 Comparison of observed and model predicted concentrations with optimum parameter estimates- Case I

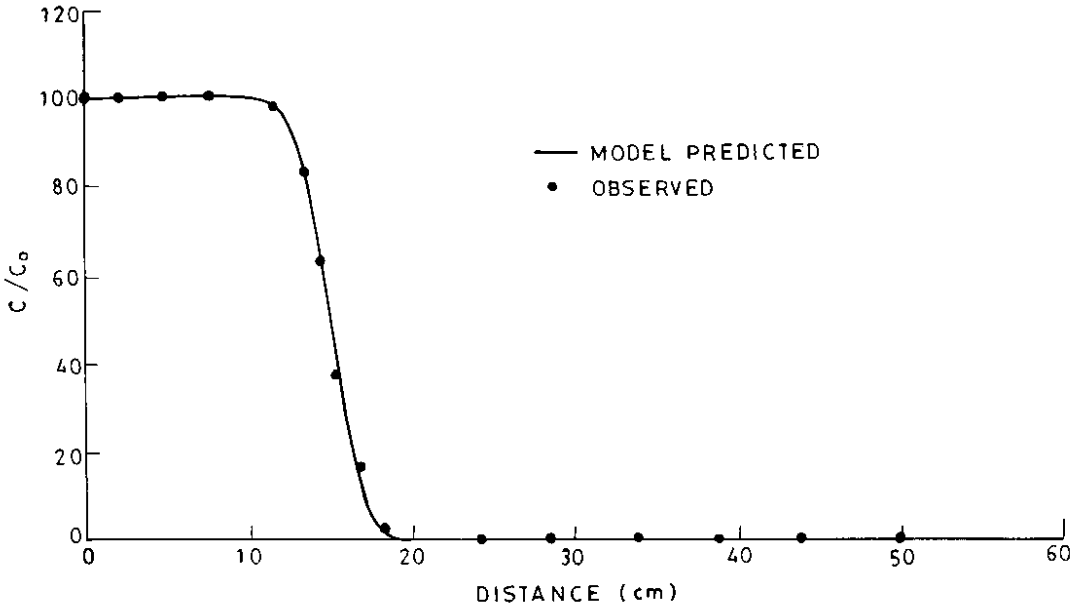

Fig 4 Comparison of observed and model predicted concentrations with optimum parameter estimates- Case II

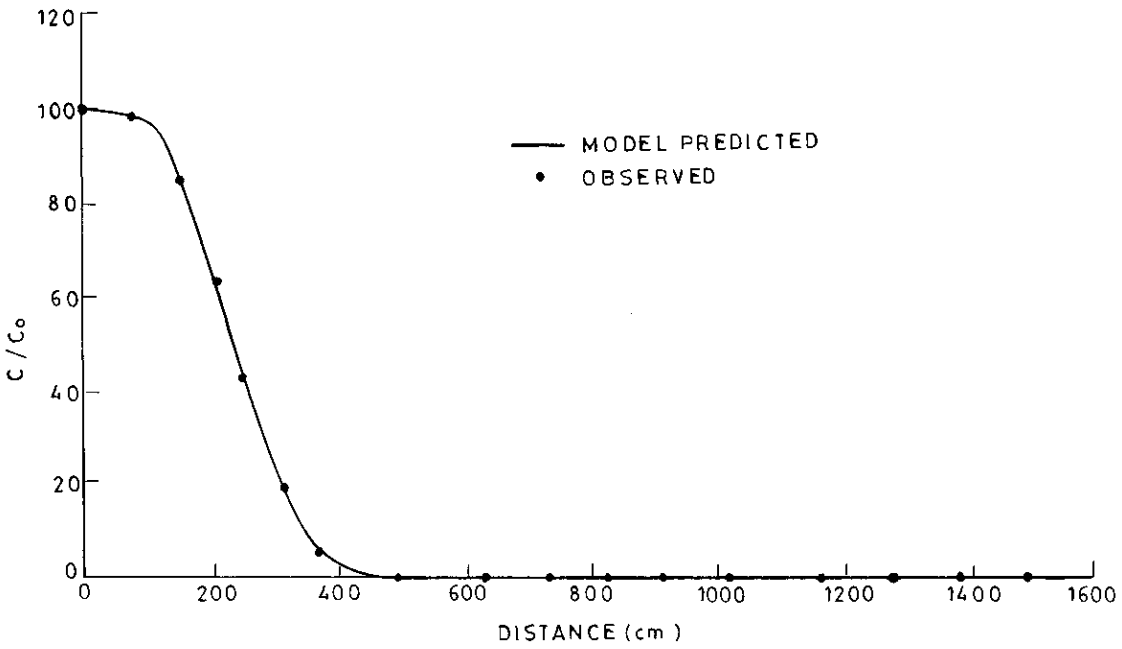

Fig 5 Comparison of observed and model predicted concentrations with optimum parameter estimates- Case III

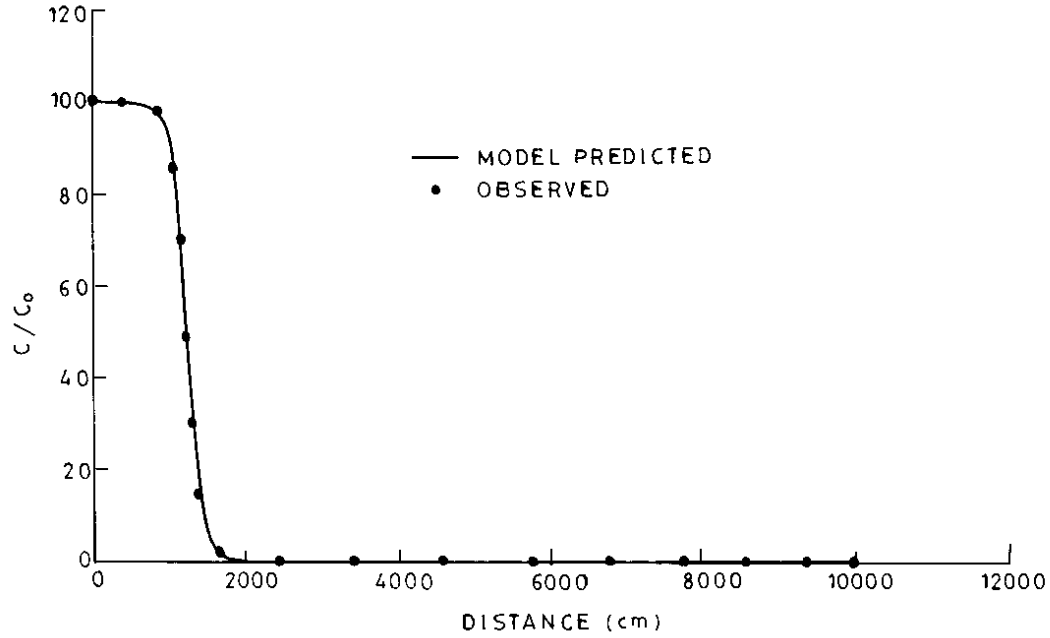

Fig 6 Comparison of observed and model predicted concentrations with optimum parameter estimates- Case IV