2011

State And Parameter Estimation With A Sequential Monte Carlo Method In A Three Dimensional Transport Model

Tushar Chowhan
North Carolina Agricultural and Technical State University

STATE AND PARAMETER ESTIMATION WITH A SEQUENTIAL MONTE CARLO METHOD IN A THREE DIMENSIONAL TRANSPORT MODEL

by

Tushar Chowhan

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE

Department: Civil, Architectural & Environmental Engineering
Major: Civil Engineering
Major Professor: Dr. Shoou-Yuh Chang

North Carolina Agricultural and Technical State University
Greensboro, North Carolina
2011

School of Graduate Studies
North Carolina Agricultural and Technical State University

This is to certify that the Master's Thesis of

Tushar Chowhan

has met the thesis requirements of
North Carolina Agricultural and Technical State University

Greensboro, North Carolina
2011

Approved by:

_____	_____
Dr. Shoou-Yuh Chang Major Professor	Dr. Stephanie Luster-Teasley Committee Member
_____	_____
Dr. Manoj K. Jha Committee Member	Dr. Sameer Hamoush Department Chairperson

_____
Dr. Sanjiv Sarin
Dean of Graduate Studies

BIOGRAPHICAL SKETCH

Tushar Chowhan was born on January 7, 1984, in Mymensingh, Bangladesh. He received the Bachelor of Science degree in Civil Engineering from Bangladesh University of Engineering and Technology in Bangladesh in 2007. He is a candidate for the M.S. degree in Civil Engineering.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my advisor, Dr. Shoou-Yuh Chang, for his support, guidance and encouragement throughout the study.

I wish to give special thanks to Dr. Manoj. K. Jha and Dr. Stephanie Luster-Teasley, for being part of my thesis committee and for reviewing this thesis.

I also want to express gratitude to Mr. Sikdar Latif for sharing his thought and experience with me during this research study.

This work was funded by the Department of Energy Samuel Massie Chair of Excellence Program under Grant No. DF-FG01-94EW11425.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES **PAGE**

LIST OF NOMENCLATURE

SMC	Sequential Monte Carlo
PF	Particle filter
SIS	Sequential Importance Sampling
SIR	Sequential Importance Resampling
KF	Kalman filter
EKF	Extended Kalman filter
FTCS	Forward Time Centered Space
A	State transition matrix (STM)
C	Concentration of contaminant in solute phase
C_0	Initial concentration
M_0	Initial mass input
D_x, D_y,D_z	Dispersion coefficients in x, y, and z directions respectively
Δt	Time step
V	Linear pore liquid velocity
x,y,z	Cartesian coordinates
Δx, Δy, Δz	Space interval along x, y, and z directions respectively
δ	Dirac delta function
η	Porosity of porous medium
R	Retardation factor
N_s	Number of samples

x_k	Vector of concentration at time step k
z_k	Observations at time step k
P	Probability
w	Weight
d	First-order decay rate
N	Normal distribution
ε	Error
ϒ	Norm
RRMSE (t)	Relative Root-Mean-Squared-Error at time step t

ABSTRACT

Chowhan, Tushar. STATE AND PARAMETER ESTIMATION WITH A SEQUENTIAL MONTE CARLO METHOD IN A THREE DIMENSIONAL TRANSPORT MODEL. **(Major Advisor: Dr. Shoou-Yuh Chang),** North Carolina Agricultural and Technical State University.

Due to the inherent randomness and heterogeneity of the transport process, macrodispersion, non-fickian motion, and ergodicity, general assumptions of linearity and Gaussian distribution do not hold for the real field. Therefore, a state-space transport model for the non-linear and non-Gaussian system is proposed in this study. In this study, the state variable (concentration vector) and parameter (first-order decay) are updated with the available measurements. The probabilistic state-space formulation and updating of information on receipt of new measurements is formulated in the Bayesian framework. particle filter, a sequential Monte Carlo method, provides a rigorous general framework for dynamic state estimation problems in the Bayesian scheme. Here the reactive contaminant transport in subsurface is treated as a dynamic state and parameter estimation problem. A type of particle filter, commonly called Sequential Importance Resampling (SIR) is used for this subsurface transport problem. The model estimation is compared with a reference true random field. A promising improvement of the estimation accuracy is attained with the SIR particle filter while compared with a traditional deterministic approach. The standard deviations of the residuals were calculated for the comparison purpose. The particle filter data assimilation scheme reduces the prediction error by 48% in estimation accuracy. In case of having fixed parameters in the model, a

standard technique to perform parameter estimation consists of extending the state with the parameter to transform the problem into optimal filtering problem. This approach requires the use of special particle filtering techniques which suffer from several drawbacks. An alternative statistical approach was adopted here to combine parameter estimation with the particle filter scheme. The concept of Euclidian norm was introduced in order to address the sequential weight assignment to the parameter estimation. The SIR particle filter scheme successfully estimated the parameter (first-order decay). With the use of the updated parameter in the state prediction, prediction error of the SIR particle filter data assimilation scheme became 78% smaller than the error from the deterministic model.

CHAPTER 1

INTRODUCTION

Groundwater accounts for approximately 20% of the total water usage: 53% of the population drink groundwater, 80 billion gallons of groundwater is withdrawn daily, and 90% of the freshwater supply is groundwater (MDEQ 2003). Contamination of the subsurface environment is pervasive, with pollutants ranging in source from manufacturing, mining, agriculture, municipalities, energy, and defense industries (Yeh et al. 2010). The transport of different types of contaminants has long been one of the greatest concerns to environmental engineers. The contaminant usually enters the groundwater system from the land surface, percolating down through the aerated soil and unsaturated or vadose zone (Pye and Jocelyn 1984). Prevention and control of groundwater contamination can better be understood if the sources of contamination, type of contaminant, and movement of contaminant through porous media are taken into consideration.

Mathematical modeling of the contaminants in the subsurface is important to predict the spread of the plume as well as for risk assessment. This prediction is also sometimes largely dependent on the parameters used in the model. Deterministic model is traditionally used to study this complex subsurface environment. Numerical modeling provides a viable means of analyzing contamination problems before a remediation option is chosen and implemented. Many techniques that are widely used for forecasting contaminant movement and their resulting risks to the linked ecosystems are composed of

mathematically based subsurface models. Finite element methods (Ren and Zheng 1999, Kim and Parizek 1999) are the most popular for one-dimensional and two-dimensional problems. They often make use of Galerkin`s method of weight residuals, and their complex geometries are easily handled by creating polygons from the node points (Schnoor 1996). The finite element techniques are useful in keeping the numerical dispersion at a minimum, which is important because the reaction terms are concentration dependent. Large concentration gradients arise in subsurface remediation problems due to the sharp boundaries of contamination. Also the techniques are complicated by non-linearities and stiffness. However, the errors arising from the numerical model can bring unavoidable prediction deviations from the real world; which is associated with increasing uncertainty. The numerical model may include numerical errors from model mechanisms, time and space limits of numerical schemes, and boundary conditions.

Methods of probabilistic prediction and data assimilation (DA) for quantification and reduction of state uncertainty have been extensively explored in the atmospheric and oceanic sciences. Their application in the hydrological sciences is relatively new, although deterministic hydrological prediction and parameter estimation have become reasonably mature. Most of the current interests in simulation-based methods of sequential Bayesian analysis of dynamic models have been focused on improved methods of filtering for time-varying state vectors. Researchers have been using discrete numerical approximations to sequentially updated posterior distribution in various “mixture modeling” frameworks. Simulation-based methods were developed in the late 1980 (Pole

and West 1990, Pole et al. 1988). Parallel developments in the early 1990`s, further led to the publication of many different but related approaches (West 1993; Gordon et al. 1993).

During the past decade, particle filters have developed rapidly and have been successfully applied in a number of different areas (Arnaud et al. 2001). There have been limited applications of particle filters in process engineering. Examples include the state estimation of a non-linear dynamic process (Chen et al. 2004a, Han and Li 2008), and the state estimation with initial condition rectification, which was implemented using a Markov chain Monte Carlo approach (Chen et al. 2004b).

Parameter estimation has been conducted mainly by using deterministic (manual or automatic) calibration techniques that tend to ignore model structural errors and measurement errors (Duan et al. 1992). Recently, stochastic data assimilation methods have been developed and applied to parameter estimation problems (Thiemann et al. 2001).

In order to predict the real field scenario in a subsurface contaminant transport, the objectives of this study are as following:

- Construct a Sequential Importance Resampling (SIR) particle filter scheme to interpret the contaminant transport with a instantaneous input in a three-dimensional subsurface model.
- Estimate the unknown parameter using the SIR particle filter algorithm.
- Examine the effectiveness of the SIR particle filter scheme with and without the parameter estimation process.

CHAPTER 2

LITERATURE REVIEW

Typically, the source of the hydraulic parameters and data initialization in environmental transport models are field observations, such as hydraulic conductivities from tracer tests and data network systems, such as the geographic information system. However, laboratory and field observations indicate that a high degree of heterogeneity may exist for hydraulic properties in natural subsurface flow system. This variability is unavoidable (Heuvelink and Webster 2001). In order to address uncertainty in hydrologic modeling, there are three distinct yet related aspects to be considered: understanding, quantification, and reduction of uncertainty. Arguably, understanding uncertainty is an integral part of any application of uncertainty quantification and/or reduction.

The hydrologic literature has seen various applications of data assimilation and/or uncertainty analysis in hydrology ranging from characterization of soil moisture and/or surface energy balance. One critical issue for hydrologic modeling is how the DA methods used in atmospheric and related sciences can best be adapted and combined with hydrologic methods to cope with the uncertainties arising from hydrologic modeling in a cohesive, systematic way to maximally reduce and adequately quantify the predictive hydrologic uncertainty (Liu and Gupta 2007).

There are three main areas where actions can be taken toward reducing uncertainty in hydrologic predictions: (1) acquisition of more informative and higher quality hydrological data (including data of new types) by developing improved

measurement techniques and observation networks; (2) development of improved hydrologic models by incorporating better representations of physical processes and using better mathematical techniques; and (3) development of efficient and effective techniques that can better extract and assimilate information from the available data via the model identification and prediction processes.

While hydrologic science has witnessed astonishing advances in the availability of hydrologic data (area 1) and the complexity/reliability of hydrological models (area 2), there is an urgent need for techniques that effectively and efficiently assimilate important information from the data into the models to produce improved hydrological predictions (area 3). Such techniques are generally referred to as data assimilation (DA) methods, which is defined as procedures that aim to produce physically consistent representations or estimates of the dynamical behavior of a system by merging the information present in imperfect models and uncertain data in an optimal way to achieve uncertainty quantification and reduction (Liu and Gupta 2007).

It is worth mentioning that this description of the DA problem is broadly encompassing, not being limited only to problems of ‘‘state estimation’’ as the term is often applied to in the literature. Instead, it describes the more comprehensive problem of ‘‘merging models with data’’ and therefore includes the three related problems of system (structure) identification, parameter estimation, and state estimation, which are all critical to the reduction of uncertainty in model predictions.

Many uncertainty analysis frameworks have been introduced in the hydrologic literature, including the generalized likelihood uncertainty estimation (GLUE)

methodology, the Bayesian recursive estimation technique (BaRE), the Shuffled Complex Metropolis algorithm (SCEM) , the multi-objective extension of SCEM, the dynamic identifiability analysis framework (DYNIA), the maximum likelihood Bayesian averaging method (MLBMA), the dual state-parameter estimation methods and simultaneous optimization and data assimilation algorithm (SODA) (Liu and Gupta 2007). However, few of these methods completely address all the above three critical aspects of uncertainty analysis in an explicit and cohesive way.

One of the most successful and popular approximation techniques is Sequential Monte Carlo (SMC), which is referred to as particle filtering (PF) in the Bayesian filtering domain. State estimation can be considered as an optimal filtering problem within a Bayesian framework. If the state equations are linear and the posterior density (at every time step) is Gaussian, the Kalman filter (KF) is an optimal solution to the state estimation problem. However, when these assumptions do not hold, there exists no analytical solution and therefore approximations need to be made. For example, the extended Kalman filter (EKF) has been widely applied to estimate non-linear state space models (Kiparissides et al. 2002, Kozub and MacGregor 1992). The EKF assumes a Gaussian posterior density and adopts a first-order Taylor series expansion to provide a local approximation to the current state. However, when state equations are highly non-linear and the posterior density is non-Gaussian, the EKF may give a high estimation error. To avoid the Gaussian assumption, one approach was to approximate the posterior density by discretizing the continuous state variables into grids (Terwiesch and Agarwal 1995, Bucy and Senne 1971). This methodology was termed point-mass filters

or probability-grid filters. However, the computational cost of point-mass filters was found to increase exponentially with the state dimension, thus limiting its widespread application in process engineering. All such approaches involved methods of evolving and updating discrete sets of sampled state vectors, and the associated weights on such sampled values as "particles."

Particle filters are an extension of point-mass filters. The basic idea is that a large number of samples (particles) are generated using Monte Carlo methods to approximate the posterior probability of the states. Thus, the particles are adaptively concentrated in regions of high probability. This is in contrast to point-mass filters which adopt a pre-defined discretization approach to the state space problem, resulting in the particles being assumed to be uniformly distributed over all the space. Chen et al. (2004a) estimated the state of a non-linear dynamic process with initial condition rectification using a Markov Chain Monte Carlo approach. They used a particle filter to the highly non-linear batch process by developing a benchmark batch polymerization process.

Yu and Cheng (2006) developed the particle filter for mobility tracking. The model was used to describe the maneuvering target tracking problem. Li et al. (2004) proposed the use of a Rao-Blackwellised particle filter to estimate parameters in a linear state-space model. A particle filter based on the sequential Monte Carlo method was used to estimate both the state and parameter (Chen et al. 2004a). A novel sequential hydrologic data assimilation approach was explored to estimate model parameters and state variables by using a sequential importance resampling (SIR) particle filter. The particle filter approach was used to model the behavior of chlorobenzene leaching from a

landfill into a soil environment at discrete time intervals in a one-dimensional space (Chang and Li 2006). A two-dimensional subsurface contaminant transport modeling was used to generate numerical and particle filter results spatially and temporally (Li 2006). She estimated BOD and decay using the boot- strap particle filtering approach. A three-dimensional subsurface transport model was used by Cheng (2000) to generate the analytical, numerical, and Kalman filter results spatially and temporally under continuous contaminant input conditions.

Parameter estimation has been conducted mainly by using deterministic (manual or automatic) calibration techniques that tend to ignore model structural errors and measurement errors (Duan et al. 1992). Recently, stochastic data assimilation methods have been developed and applied to parameter estimation problems (Thiemann et al. 2001). The particle filters approach was used for data assimilation in a high-dimensional non-linear ocean model (Kivman 2003). Kivman estimated three state variables and two parameters in the Lorennz model by using the particle filter data assimilation method. In situation where the model has fixed parameters, a standard technique was developed to perform parameter estimation. This technique consists of extending the state with the parameters to transform the problem into optimal filtering problem (Doucet and Tadić 2003). This approach requires the use of special particle filtering techniques which suffer from several drawbacks. In this research, newly emerged stochastic data assimilation method has been used for parameter estimation due to the limitation of the traditional deterministic model calibration methods. Such method operates within Bayesian updating framework for estimation of predictive uncertainty.

CHAPTER 3

METHODOLOGY

3.1 Three-Dimensional Contaminant Transport Model

The conceptual model or governing equation most widely used to represent solute transport in hydrologic systems is the advection–dispersion reaction equation. The three-dimensional solute transport equation for a conservative solute in a uniform, saturated groundwater flow field with the direction of flow parallel to the x-axis is:

$$\frac{\partial C}{\partial t} = \frac{D_x}{R}\frac{\partial^2 C}{\partial x^2} + \frac{D_y}{R}\frac{\partial^2 C}{\partial y^2} + \frac{D_z}{R}\frac{\partial^2 C}{\partial z^2} - \frac{V}{R}\frac{\partial C}{\partial x} - kC \qquad (1)$$

where C =solute concentration, ML^{-3}

t=time, T

x, y, z=cartesian coordinates, L

D_x, D_y, D_z=dispersion coefficient in x, y and z directions respectively, $L^2 T^{-1}$

V =linear velocity of flow field in the x direction, LT^{-1}

k = first-order degradation rate constant, T^{-1}

R= dimensionless retardation factor.

The retardation factor is defined as:

$$R = 1 + K_d\left(\frac{\rho_b}{\eta}\right) \qquad (2)$$

where ρ_b = bulk density of the porous medium, ML^{-3},

η = effective porosity, dimensionless, and

K_d = distribution coefficient, L^3M^{-1}.

The initial condition is assumed as:

$$C(x,y,z)\,\big|_{t=0} = C_0(x_0,y_0,z_0) \tag{3}$$

3.2 Analytical Solution for Subsurface Model

For the instantaneous input subsurface transport model, the analytical solution is obtained based on the literature in the subsurface area (Cheng 2000). The analytical solution for a pollutant with an initial mass, M_o, that is injected (Figure 3.1) instantaneously at t=0 is:

$$C(x,y,z,t) = \frac{M_o R^{3/2}}{8\eta(\pi t)^{3/2}\left(D_x D_y D_z\right)^{1/2}} \exp\left[-\frac{(x-Vt/R)^2 R}{4D_x t} - \frac{y^2 R}{4D_y t} - \frac{z^2 R}{4D_z t} - kt\right] \tag{4}$$

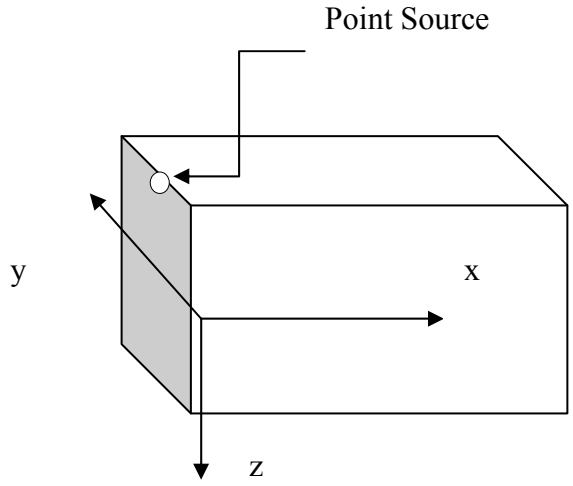

Figure 3.1. Three-dimensional contaminant transport with an instantaneous input

3.3 Subsurface Transport Scheme

In order to incorporate the particle filter scheme, we are going to use the state-space form to represent a mathematical model that simulates the dynamic process of the transport phenomenon. Owen (1984) compared several mathematical modeling methods used in coastal and estuarine regions. Owen found that the Forward-time and Central-Space (FTCS) method is always applicable for the advective transport of salinity.

Jin (1996) used the basic FTCS differences to develop the state-space form of the system equation for a two-dimensional transport model. Zou and Parr (1995) also used this finite-difference method (FDM) in their research to predict the pollutant transport in a two-dimensional aquifer. For this three- dimensional scheme, the term in vertical direction (z-axis) is introduced. Let $C(i,j,k,t) = C(x_i, y_j, z_k, t)$, the form of equation based on the FTCS method is:

$$C(i,j,k,t+1) = b_1 C(i-1,j,k,t) + b_2 C(i,j,k,t) + b_3 C(i+1,j,k,t)$$

$$+b_4 C(i,j-1,kt) + b_5 C(i,j+1,k,t) + b_6 C(i,j,k-1,t) + b_7 C(i,j,k+1,t) \quad (5)$$

The matrix form based on these equations is,

$$C(t+1) = \mathbf{A}\, C(t) \quad (6)$$

where $C(t) =$ the vector of contaminant concentration at all nodes at time $(t)\Delta t$,

$C(t+1) =$ the vector of contaminant concentration at all nodes at time $(t+1)\Delta t$,

$\mathbf{A} =$ State Transition Matrix.

For this three dimensional scheme, $\mathbf{A}$ is constructed with the seven coefficients $b_1, b_2, b_3, b_4, b_5, b_6$ and b_7. The seven coefficients represents that the concentration

effects of one node at time $(t+1)\Delta t$ come from the concentrations at time $(t)\Delta t$ in six directions and itself (seventh terms). The concept of effect (as mentioned above) represents the concentration flow between two nodes.

The boundary condition adopted here is used in the FTCS model to control the operation of the State Transition Matrix. For each time period Δt, the concentration distribution vector is improved at one step by multiplying the matrix. The concentration vector is built using the concentrations from the whole plume. Thus, the boundary condition is applied before each multiplication to eliminate the effects between nodes which are not adjacent to each other, such as two boundary nodes. However, for the nodes located on the boundary, there are no six-direction effects available since some of the directions are the boundary of the sample aquifer. For example, in the top layer of the plume, only five-direction effects exist because there is no higher node on this one. In this case, the State Transition Matrix has to be modified to re-count the effects eliminated during the operation of the boundary condition such as the nodes in the top layer; the concentration effect with coefficient b_7 for higher node is disappeared after the multiplication. Therefore, we have to change b_2 to $b_2 + b_7$ to recount the lost concentration.

3.4 Bayesian Estimation of State Space Model

At least two models are required to analyze and make inference about a dynamic system. The first model is needed to describe the evolution of the state with time (the system model). The second model is needed to relate the noisy measurements to the state

(the measurement model). Here it is assumed that these models are available in a probabilistic form. The probabilistic state-space formulation and the requirement for updating of information upon receipt of new measurements are ideally suited for the Bayesian approach. In the Bayesian approach to dynamic state estimation, the posterior probability density function (pdf) of the state is constructed based on all available information, including the set of received measurements. A pdf embodies all available statistical information and then represents the complete solution to the estimation problem. In principle, an optimal (with respect to any criteria) estimate of the state may be obtained from pdf (Arulampalam et al. 2002). Also the measure of the accuracy of the estimate may be obtained from the pdf. A recursive filter is a convenient solution in this case. This filter processes data sequentially rather than as a batch so that it is not necessary to store the complete data set nor to reprocess existing data if a new measurement becomes available. This kind of filter consists of essentially two stages: prediction and update. In the prediction stage system model is used to predict the state pdf forward from one measurement time to the next. As the state is usually subject to unknown disturbances (modeled as random noise), the prediction generally translates, reforms, and spread the state pdf. In the update operation the measurement is used to modify the prediction pdf. All these are achieved by the Bayesian theorem, which is the mechanism for updating the knowledge about the target state in light of extra information obtained from the new data.

Consider the following state space model with non-linear state and measurement functions, f_k and h_k, respectively:

$$x_k = f_k\left(x_{k-1}, v_{k-1}\right) \tag{7}$$

$$z_k = h_k\left(x_k, n_k\right) \tag{8}$$

where k is the time index, x is a state vector, and z is the measurement vector. v and n are independent and identically distributed noise for the process and measurements, respectively.

The objective of state estimation is to sequentially calculate the state vector, x_k using the given measurements z_k. In real processes, some states are very difficult to measure on-line, such as the molecular weight of polymers and the concentration of reactant, while others are unmeasurable. Therefore, one of the challenges in state estimation is to infer all the states from limited measurements.

From a Bayesian perspective, the aim of state estimation is to infer the probability function of the state x_k given the measurement sequence $z_{1:k} = \{z_i, i = 1, \ldots, k\}$ i.e., $p\left(x_k \middle| z_{1:k}\right)$. Assuming the initial conditions (expressed in the form of a probability distribution function $p\left(x_0 \middle| z_0\right) \equiv p\left(x_0\right)$) are available, $p\left(x_k \middle| z_{1:k}\right)$ can be obtained sequentially through prediction.

Suppose that the required pdf $p\left(x_{k-1} \middle| z_{1:k-1}\right)$ at time $k-1$ is available. The prediction stage will then involve using the system model Equation (7) to obtain the prior pdf of the state at time k via the Chapman-Kolmogorov equation:

$$p\left(x_k \middle| z_{1:k-1}\right) = \int p\left(x_k \middle| x_{k-1}\right) p\left(x_{k-1} \middle| z_{1:k-1}\right) dx_{k-1} \tag{9}$$

and then update it as follows:

$$p(x_k | z_{1:k}) = \frac{p(z_k | x_k) P(x_k | z_{1:k-1})}{p(z_k | z_{1:k-1})} \tag{10}$$

where $p(z_k | z_{1:k-1})$ is a normalizing factor independent of the state x_k.

Equations (9) and (10) are the optimal solutions from a Bayesian perspective to the non-linear state estimation problem. In general, the posterior probability, $p(x_k | z_{1:k})$, cannot be determined analytically. Thus approximate filters are used to provide suboptimal solutions. The widely used EKF may work poorly for highly non-linear systems because of the Taylor approximation. In addition, even if $p(x_{k-1} | z_{k-1})$ is Gaussian, $p(x_k | z_k)$ is no longer Gaussian due to the non-linear state function, which invalidates the underlying assumption of the EKF. An alternative approach is through particle filters, when the posterior pdf is non-Gaussian.

3.5 Sequential Importance Sampling (SIS)

The sequential importance sampling (SIS) algorithm is a Monte Carlo (MC) method that forms the basis for most sequential MC filters developed over the past decades (Arnaud et al. 2001, Doucet et al. 2000). This sequential MC (SMC) approach is also known variously as bootstrap filtering (Gordon et al. 2002), and particle filtering (Carpenter et al. 1999). It is a technique for implementing a recursive Bayesian filter by MC simulations. The key idea is to represent the required posterior density function through a set of random samples with associated weights and then to compute estimates based on these samples and weights. As the number of samples become very large the

MC characterization becomes an equivalent representation to the usual functional description of the posterior pdf, and the SIS filter approaches the optimal Bayesian estimate.

The basic idea of SIS filters is to approximate $p(x_k|z_{1:k})$ through using a set of random samples (also called particles) $\{x_k^i, i=1,.....,N\}$ with associated weights $\{w_k^i, i=1,.....,N\}$, where $\sum_{i=1}^{N} w_k^i = 1$

$$p(x_k|z_{1:k}) \approx \sum_{i=1}^{N} w_k^i \delta(x_k - x_k^i) \tag{11}$$

where, $\delta(x)$ is an indicator function which is equal to unity if $x = 0$; otherwise it is equal to zero.

The key step is to generate random samples from $p(x_k|z_{1:k})$. However, as $p(x_k|z_{1:k})$ is not of the conventional form of a probability density function, such as Gaussian or Cauchy, direct sampling is not possible. Therefore importance sampling (Bergman 1999, Doucet et al. 2000) is then used to obtain the particles and their associated weights. The first step in importance sampling is to define an importance density $q(x_k|z_{1:k})$ from which samples x_k^i can be drawn (e.g. a standard Gaussian distribution function). Thus the weights are defined as:

$$w_k^i \propto \frac{p(x_k^i|z_{1:k})}{q(x_k^i|z_{1:k})} \tag{12}$$

For the sequential estimation problem, at time point k, the particles which approximate $p(x_{k-1}|z_{1:k-1})$ will be passed through the state function and updated with a new measurement, z_k to approximate $p(x_k|z_{1:k})$. It was shown (Arulampalam et al. 2002) that if the importance density is only dependent on the current measurement, z_k, and the past state, x_{k-1}, the weights can be updated as:

$$w_k^i \propto w_{k-1}^i \frac{p(z_k|x_k^i)\, p(x_k^i|x_{k-1}^i)}{q(x_k^i|x_{k-1}^i, z_k)} \tag{13}$$

Using these particles and associated weights, the estimated state vector, $\hat{x}_k$, is the mean of $p(x_k|z_{1:k})$ and is calculated as:

$$\hat{x}_k = \sum_{i=1}^{N} w_k^i x_k^i \tag{14}$$

3.6 Sequential Importance Resampling (SIR)

A common problem with the SIS particle filter is the degeneracy problem phenomenon, as after a few iterations, all but one particle will have negligible weight. It has been shown (Doucet et al. 2000) that the variance of the importance weight can only increase over time, and thus, it is impossible to avoid the degeneracy phenomenon. This degeneracy implies that a large computational effort is devoted to updating particles whose contribution to the approximation to $p(x_k|z_{1:k})$ is almost zero. Alternative solution to this problem can be achieved by any of the two methods: 1) a good choice of importance density and 2) the use of resampling. Here we will limit our discussion to the resampling method only.

A suitable measure of the degeneracy of the algorithm is the effective sample size N_{eff} introduced in (Bergman 1999) and defined as:

$$N_{eff} = \frac{N_s}{1 + \mathrm{Var}(w_k^{*i})} \tag{15}$$

where, w_k^{*i} is referred as the "true weight" and N_s is the number of samples.

As this cannot be evaluated exactly, an estimate $\hat{N}_{eff}$ of N_{eff} can be obtained by:

$$\hat{N}_{eff} = \frac{1}{\sum_{i=1}^{N_s} (w_k^i)^2} \tag{16}$$

where w_k^i is the normalized weight obtained using Equation (13).

Notice that when $\hat{N}_{eff} \le N_s$, a small value of N_{eff} indicates severe degeneracy. Therefore, when N_{eff} falls below some threshold N_T, the SIR is used (Arulampalam et al. 2002). The basic idea of resampling is to eliminate the particles that have small weights and to concentrate on the particles with large weights. The resampling step involves generating a new set of $\left\{x_k^{i*}\right\}_{i=1}^{N_s}$ by resampling (with replacement) N_s times from an approximate discrete representation of $p\left(x_k \middle| z_{1:k}\right)$ given by:

$$p\left(x_k \middle| z_{1:k}\right) \approx \sum_{i=1}^{N_s} w_k^i \delta\left(x_k - x_k^i\right) \tag{17}$$

where $\Pr\left(x_k^{i*} = x_k^j\right) = w_k^j$.

The resulting sample is in fact an i.i.d. sample from the discrete density. Therefore, the weights are now reset to $w_k^i = 1/N_s$. The operation of SIR particle filter is represented in Figure 3.2.

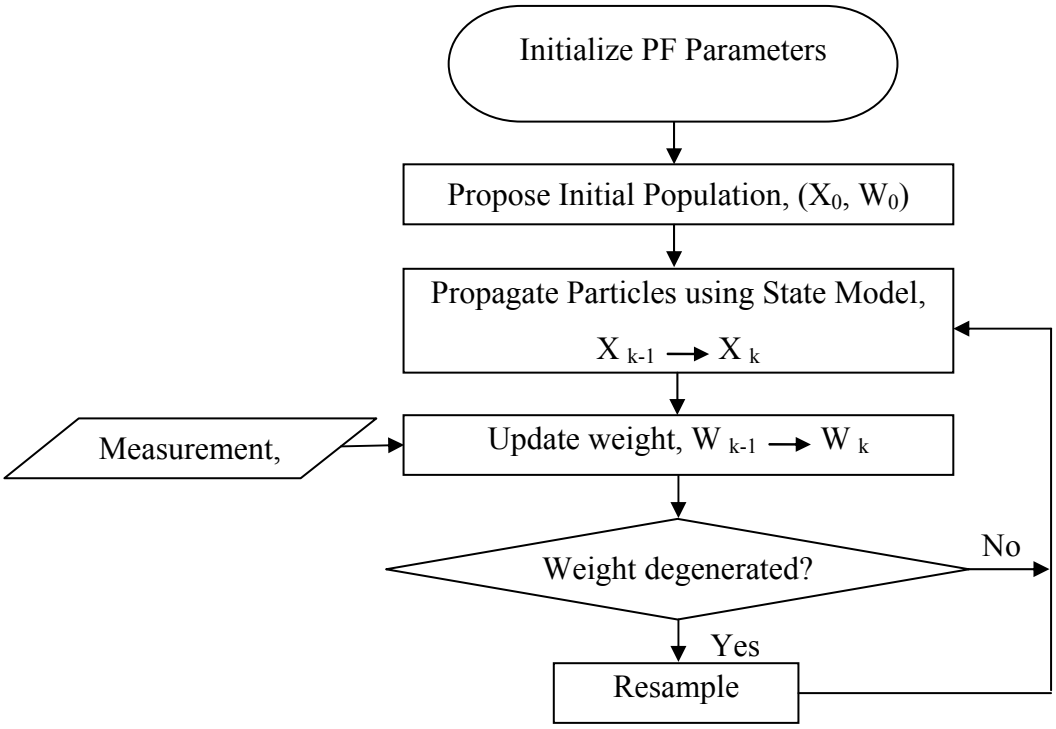

Figure 3.2. Operation of SIR particle filter

3.7 Coupling Parameter Estimation with Sequential Monte Carlo Method

Parameter estimation has been conducted mainly by using deterministic approach. Recently, stochastic data assimilation methods have been developed and applied to parameter estimation problems. One of our main objectives of the research is to estimate the parameter (decay) along with the state (concentration). For this research, particle filter

state (concentration) estimate, $\hat{\mathbf{c}}_{d_i\text{-}1,t\text{-}1}$ at time $\mathbf{t-1}$, and observation (concentration) $\mathbf{z_t}$ at time, $\mathbf{t}$, and the particle filter estimate of the parameter $\hat{d}_{\mathbf{t-1}}$, at time $\mathbf{t-1}$ are available. The primary objective is to find the particle filter estimate of parameter at time $\mathbf{t}$. Then this estimated parameter is used to find the particle filter estimate of the state, $\hat{\mathbf{c}}_{d_i,\mathbf{t}}$ at time $\mathbf{t}$.

3.7.1. Derivation of Weight for Parameter Estimation

In state estimation, the traditional way of assigning weight to the samples at each time step is based on the boot-strap particle filter method. Due to the limitation of the traditional approach in parameter estimation process, a new statistical approach was proposed in our study. The basic assumption for this approach is: $\text{probability} \propto \frac{1}{\text{norm}}$, where, norm is the distance from the origin to the point of interest. For a sample size n, the parameter $\hat{d}_{\mathbf{t-1}}$ can be sampled as a normally distributed sample. The form of the distribution can be written as: $\boldsymbol{N}(\hat{d}_{t-1}, \sigma^2) = \left[d_1, d_2, d_3, \ldots\ldots d_n\right]_t$

Using Equation (6) the state equation for concentration can be written as:

$$\left[A_{k_i,t}\right]\left[\hat{C}_{t-1}\right] = \left[\hat{C}_{k_i,t}\right] \tag{18}$$

From the observation, $\mathbf{z_t}$ at time step $\mathbf{t}$, the error matrix can be formulated as:

$$\left[\boldsymbol{\varepsilon}_{d_i,t}\right] = \left[z_t - \hat{\boldsymbol{c}}_{d_i,t}\right] \tag{19}$$

For the n number of samples the error matrix is a column vector of size n.

$$\left[\varepsilon_{d_i,t}\right] = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix} \tag{20}$$

Using the concept of Euclidean norm, the norm for d_i can be written as:

$$\left[\Upsilon_{d_i,t}\right] = \sqrt{\sum_{j=1}^{n} \varepsilon_j^2} \tag{21}$$

Using the assumption of, weight $\propto \frac{1}{\text{norm}}$, the weight can be formed as :

$$w'_{d_i} = \frac{\sum_{i=1}^{n} \Upsilon_{d_i} - \Upsilon_{d_i}}{\sum_{i=1}^{n} \Upsilon_{d_i}} \tag{22}$$

After normalizing, the final weight for d_i can be written as:

$$\omega_{d_i} = \frac{\omega'_{d_i}}{\sum_{i=1}^{n} \omega'_{d_i}} \tag{23}$$

The weights for all the samples are calculated using Equation (23). With these weights, the parameter estimation process enters the update stage of the traditional SIR particle filter method (Figure 3.2) and moves to the next time step.

3.8 Filter Effectiveness Measurement

The effectiveness of the SIR particle filter can be is demonstrated by comparing the results from the numerical (FTCS) model and the SIR particle filter model. Although

different indices can be compared, we chose relative-root-mean-squared error (RRMSE). The expression of RRMSE is as following:

$$\mathbf{RRMSE(t)} = \frac{\sqrt{\frac{\mathbf{1}}{\mathbf{N-1}}\sum_{\mathbf{m=1}}^{\mathbf{N}}\left[\mathbf{x_m(t) - z_m(t)}\right]^2}}{\frac{\sum_{\mathbf{m=1}}^{\mathbf{N}}\left[\mathbf{z_m(t)}\right]}{\mathbf{N}}} \tag{24}$$

where, $\mathbf{RRMSE(t)}$ = the residuals at time step t;

$\mathbf{x_m(t)}$ = the simulated observation of node m at time step t;

$\mathbf{z_{(m)}(t)}$ = the estimation of node m at time step t;

$\mathbf{N}$ = the total number of nodes.

The numerator of Equation (24) is also known as RMSE. The RMSE is normalized by the mean of the estimated concentrations of all the nodes at a time step to generate the RRMSE.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Model and Parameter Description

With the deterministic transport model and particle filter algorithm described in the previous section, a three-dimensional contaminant model is constructed to simulate the contaminant transport processes and predict the contaminant plumes` evolution. The system parameters are assumed on the basis of the research of Cheng (2000). He assumed the horizontal dispersion, $D_x = 1.00\,m^2/day$, $D_y = 0.50\,m^2/day$, the vertical dispersion $D_z = 0.70\,m^2/day$, porosity=0.30, velocity= $0.8\,m/day$, retardation $R = 1.5$ and degradation rate $k = 0.3/day$. We set the model grid size, $dx = dy = dz = 2.00m$. Each time step is 0.75 day and the number of total simulation time steps is 30. The number of grid points in x direction =10, number of grid points in y direction=9, and number of grid points in z direction =6. The number of all nodes in the transport scheme is 10*9*6=540. The initial condition is a instantaneous contaminant source of 10,000 ppm seeping into a location with the central coordinates C (1, 5, 1). In this study, the conception of "layers" was introduced to indicate the horizontal sections in the different vertical depth. That is to say, the "first layer" represents the top aquifer plane (z =1), the "second layer" represent the next aquifer plane (z = 2), and so on.

4.2 Prediction from Numerical Scheme

At the first stage of experiment, the deterministic model with the specified initial condition described in Equation (3) was formed. A program coded in MATLAB was developed to solve the model and to estimate the concentration. Figure 4.1 shows the model prediction at t=15 days. The pollutant contour lines from the numerical model simulated the theoretical advection–dispersion–reaction transport process. As shown, the pollutant distribution from the model is symmetrical due to the numerical dynamics and the assumed velocity in the x direction.

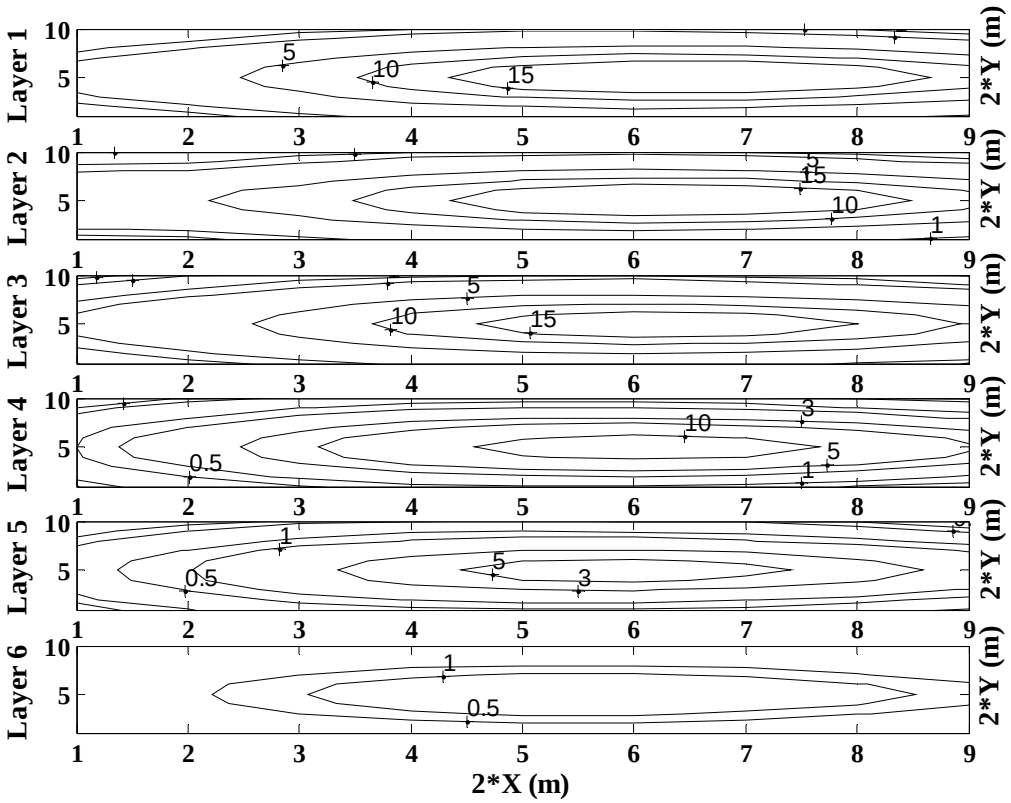

Figure 4.1. Numerical concentrations (mg/L) at different layers after 15 days

The relatively smooth shape of the contaminant plume is a result of the approximation made to the numerical model used. The numerical scheme is characterized with error coming from the assumptions made on the parameters and the model used in estimation. The parameters used in this approach were assumed to be constant.

4.3 Simulated True Field Prediction

Figure 4.2 depicts the analytical field scenario for time step 20, i.e. after 15 days of the contaminant transport. The prediction of the analytical scheme was made using the Equation (4). Afterwards, a randomly distributed noise of was chosen and added to the analytical solution to simulate the true states.

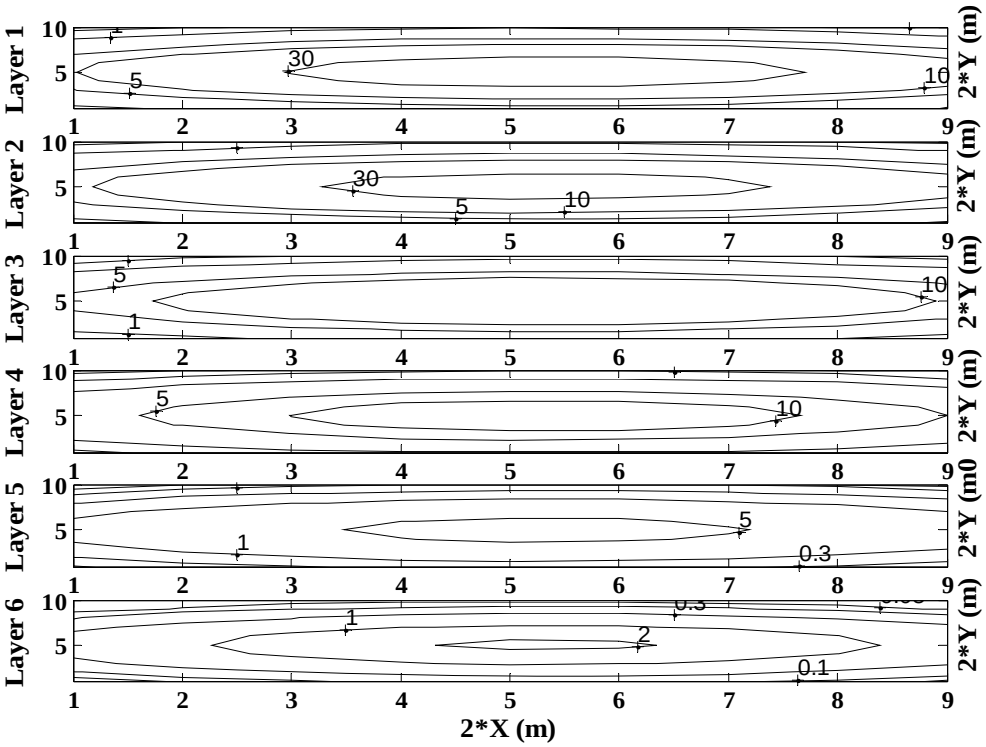

Figure 4.2. Analytical concentrations (mg/L) at different layers after 15 days

4.4 Observation Data Generation

A random Gaussian error was added to the true field to obtain simulated observation data or measurement (Figure 4.3) for all time steps. The observation error introduced reflects the randomized nature of real-life field data of contaminant concentrations owing to human and instrument errors. An observation error of 5% was chosen and added to the true value to simulate the dynamic observation states.

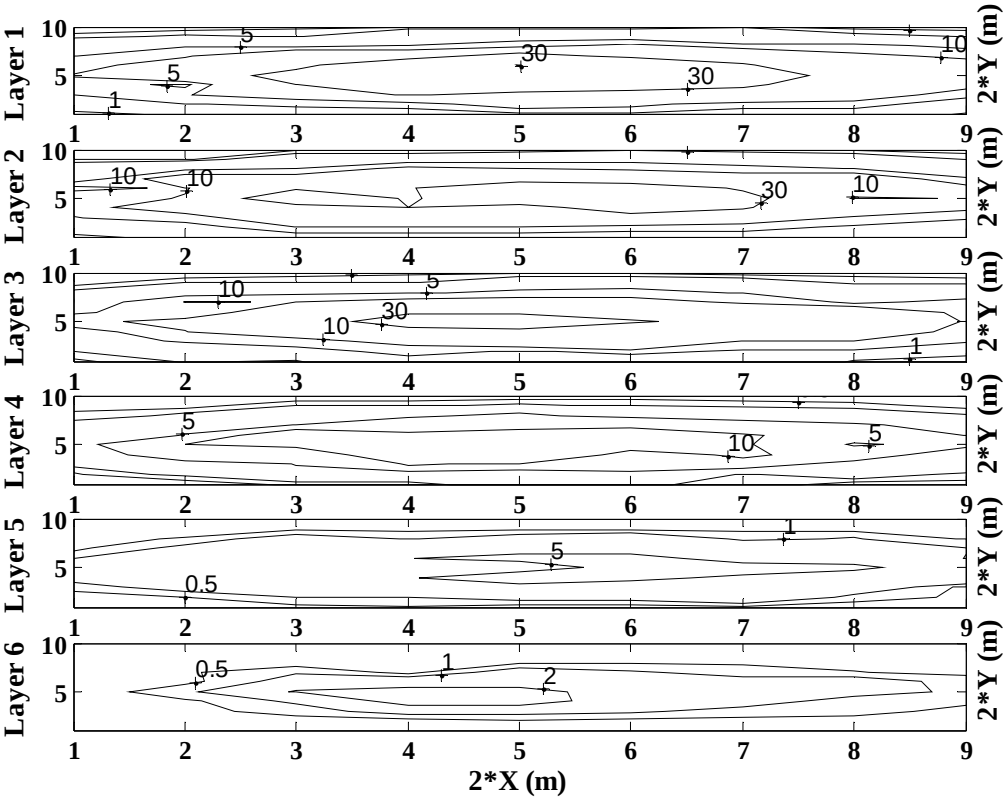

Figure 4.3. Simulated true field concentrations (mg/L) at different layers after 15 days

4.5 SIR Particle Filter Estimate

By using both the numerical and the SIR particle filter scheme, the model dynamics were assimilated with observation data at each time step to give the estimated value for the contaminant concentration. The numerical model serves as a guide in estimating the state of the model. The contours of the particle filter results are relatively closer to the true value than the numerical solution shown previously. The particle filter results are directed by the observation data hence the closeness in results. Figure 4.4 shows the contaminant plume evolution by using the particle filter at time step 20.

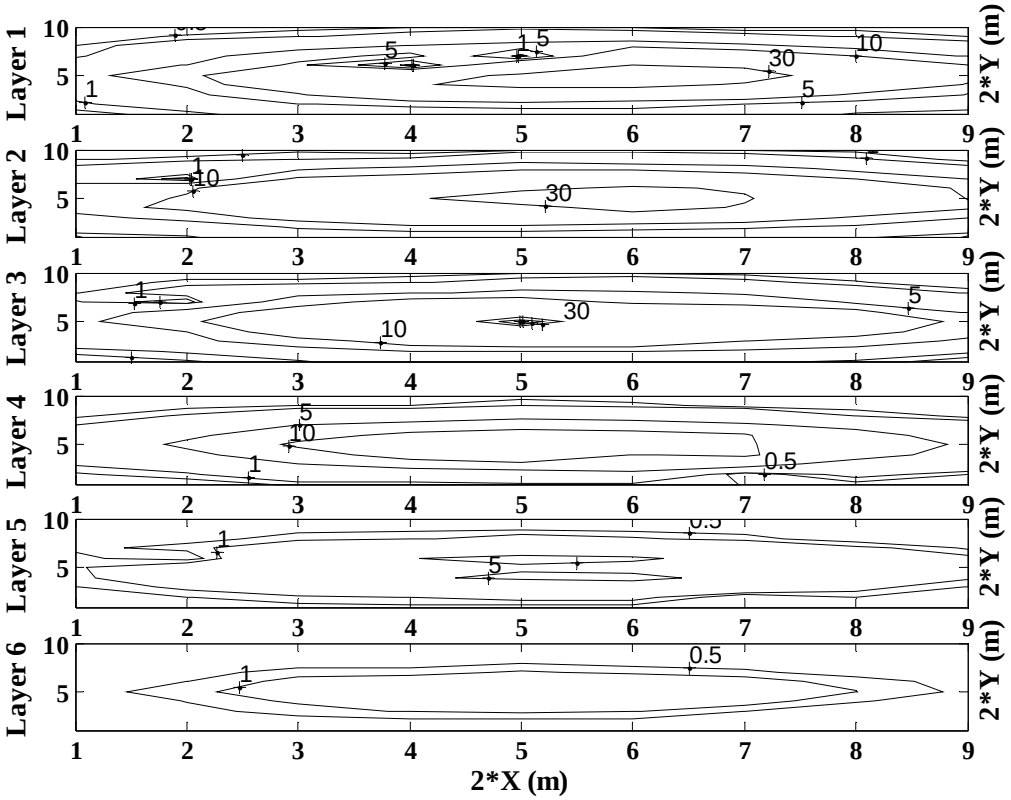

Figure 4.4. SIR particle filter concentrations (mg/L) at different layers after 15 days

4.6 Effectiveness of Numerical and SIR Particle Filter Scheme

The effectiveness of the numerical and SIR particle filters scheme is determined by comparing both the results with the simulated true value for each time step. The changes in the RRMSE (Figure 4.5) indicate that as the assimilation progressed, the estimated value for the concentration is getting closer to the reference true value, which results in the smaller RRMSE over time. The bigger error is largely attributable to the linearity of the model used, initial averaging of samples and the random noise introduced into the filtering scheme.

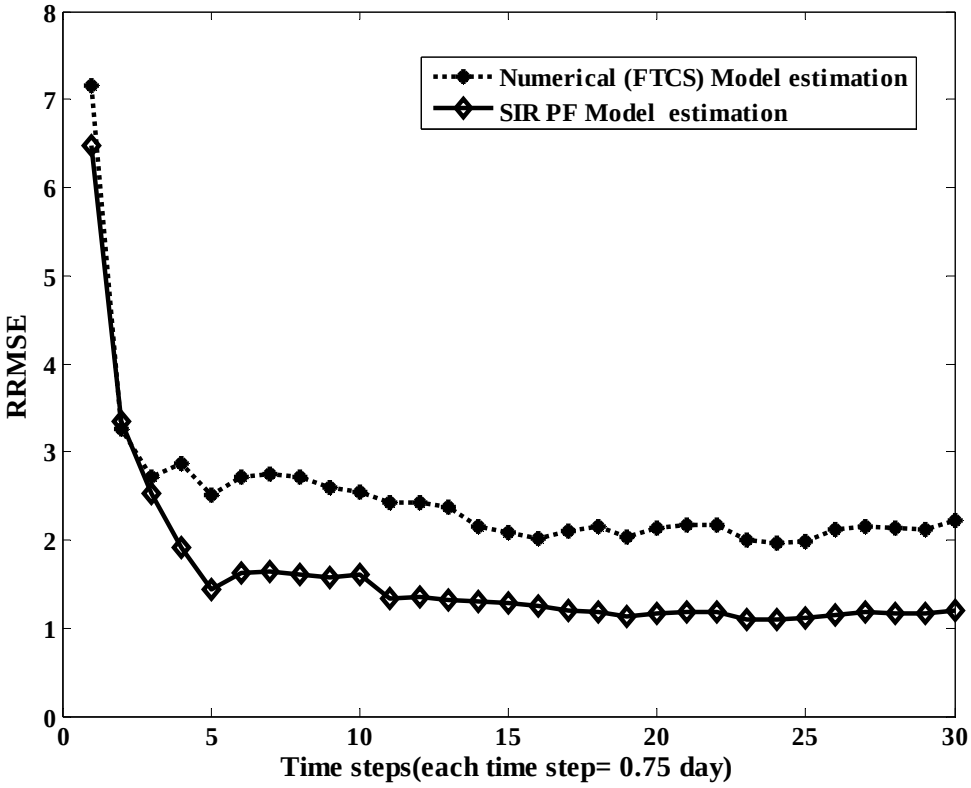

Figure 4.5. RRMSE for the numerical model and the SIR particle filter model

From the RRMSE profile, the numerical scheme shows more errors at all time steps. The approximation and assumptions made to the model introduced a certain amount of error. The SIR particle filter scheme reduces the RRMSE to 1.2 from 2.3. This is about 48% improvement of the particle filter over the deterministic FTCS model prediction results.

4.7 Parameter Estimation

In our experiment one parameter (first-order decay) was estimated and used to update the state (concentration) predictions at every time step. The main challenge was to develop weights for the parameter to couple with the particle filter at every time step. The problem was resolved using the statistical concept of Euclidean norm to generate weights for the particles. Initial sampling of decay was done based on an assumed mean of 0.3/day and a variance of 10% of the mean, which is randomly distributed with 300 samples. At every time step, norm was generated using the error from observation and particle filter estimate. Assuming that norm is proportional to weight, weights of all the particles were calculated. With the updated decay the state estimation was done to predict the concentration plume`s evolution. The assimilation result of a single run is shown in Figure 4.6. The results show the adaptation of the process with the reference true value. As the parameter estimation was a random process, the curve started from the vicinity of 0.3/day and finally converges towards the reference true value of 0.05/day.

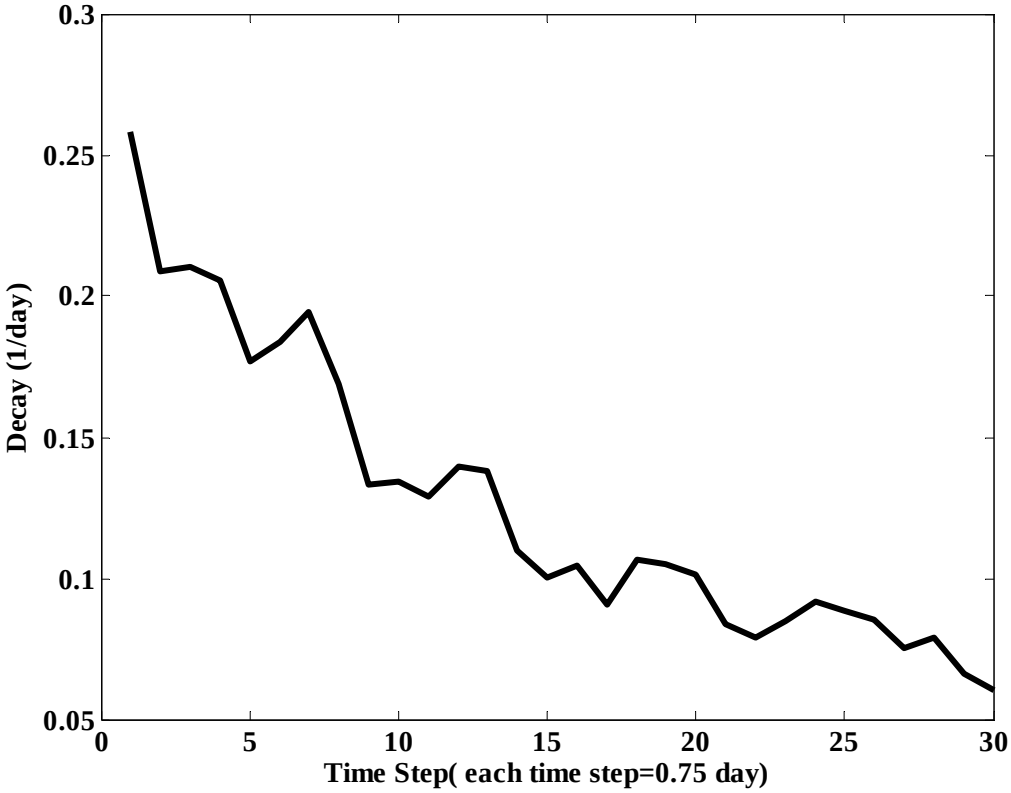

Figure 4.6. First-order decay vs. number of time steps with random noises (single run)

4.8 Effectiveness of Numerical and SIR Particle Filter Scheme with Parameter Estimation

Figure 4.7 shows the RRMSE for the numerical model (FTCS) and the SIR particle filter model with and without the parameter estimation. The SIR particle filter with the parameter estimation reduced the RRMSE to 0.50 from 2.3. The improvement of the new method is about 78% compared to the deterministic FTCS method while the earlier PF method without the parameter estimation has a 48% improvement.

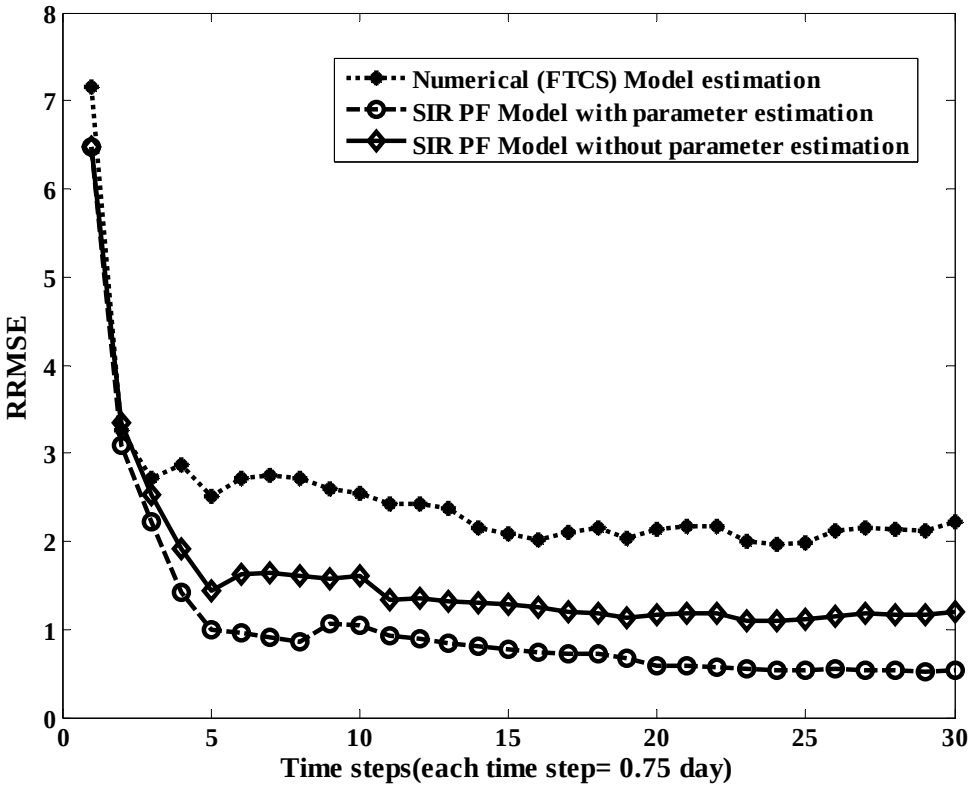

Figure 4.7. RRMSE for the Numerical model and the SIR particle filter model with and without parameter estimation

4.9 Sensitivity Analysis of the Parameter Estimation

To test the sensitivity of the parameter estimation, 10 run of the parameter estimate was made. The result from the runs is shown in Figure 4.8. The trend of the figure clearly shows improvement of the parameter estimation accuracy with time. Here the initial sampling of decay was done based on an assumed mean of 0.3/day and with a variance of 10% of the mean. Due to this initial sampling the estimation started from the assumed mean of 0.3/day and eventually merges towards the true value of 0.05/day after 30 time steps. The result indicates the new method of weight assignment to the parameter`s samples work efficiently in the particle filter scheme.

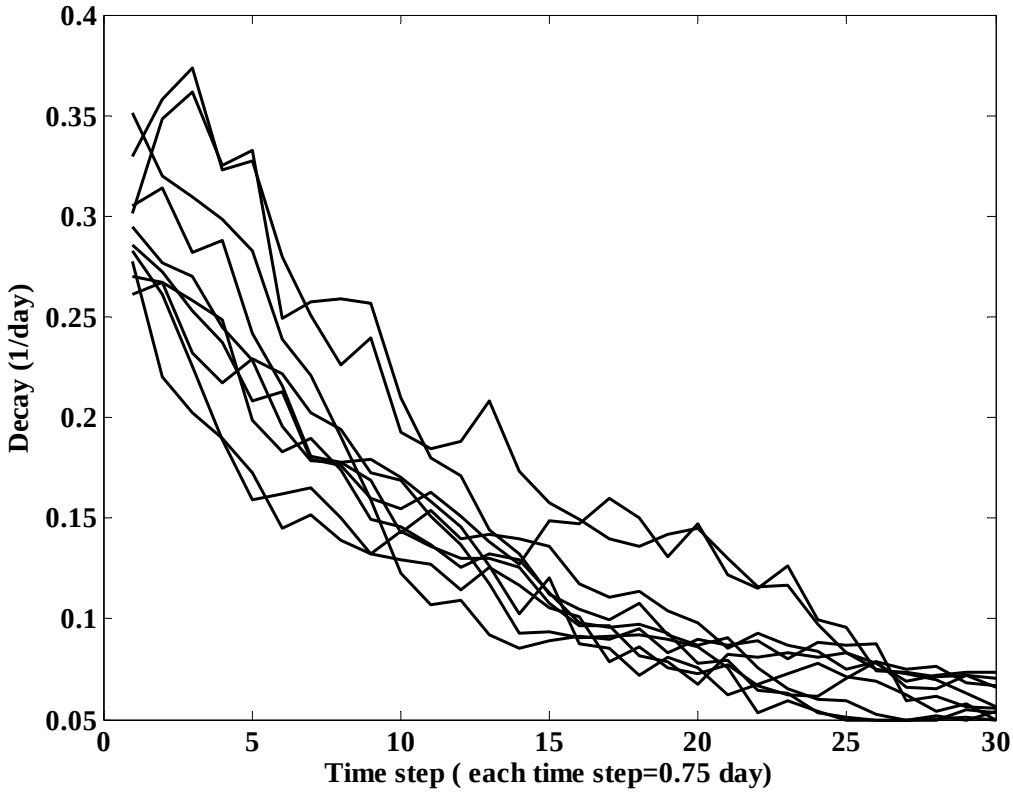

Figure 4.8. First-order decay vs. number of time steps with random noises (10 run)

As the observation value of the parameter was not available, the state observation and particle filter state estimate were used in the parameter estimation process. Weights of samples were formulated by taking inference from these two states. To investigate the effect of the simulated observation on the parameter estimation process, two different kinds of noises were used in the reference true solution. The first set of noises was created by using fixed random noises in the reference true solution. The idea was to use the same random noises for every time step. Without using different random noises at every time step, we generated these noises only once and used it for all the following time steps.

The second set of noise used in the sensitivity analysis was fixed noise. Rather than using random noise, a fixed noise was added to the simulated true field. The main theme of this experiment was to add a fixed noise at every time step which is a percentage of the true solution obtained from the previous time step. In this study, the concentration for each of the 540 nodes was increased by 10% to generate the simulated true field. Figure 4.9 shows the sensitivity analysis of the parameter estimation process.

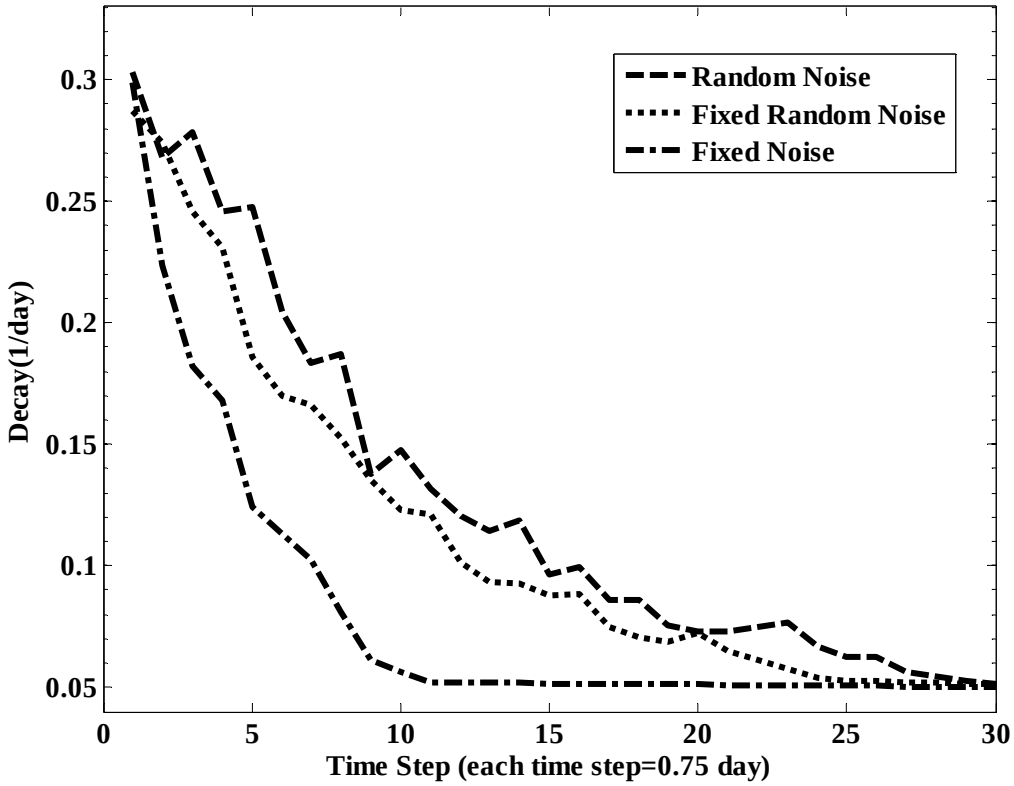

Figure 4.9. First-order decay vs. number of time steps with variable noises (single run)

CHAPTER 5

CONCLUSIONS

In Bayesian state-space theory, the system model, which might start with a very weak knowledge about the initial state, can achieve more and more accurate information about the state through assimilation with the observation data. In the three-dimensional prediction model the particle filter reduces the deviation in each time step by combining observation data within model dynamics. In this study, the effectiveness of the proposed Monte Carlo scheme was demonstrated based on a three-dimensional numerical platform. An advection–dispersion–adsorption subsurface transport model was constructed in MATLAB to predict contaminant plume. A randomly generated noise scheme was designed to represent the real world groundwater contaminant transport. A Sequential Importance Resampling (SIR) particle filter with 300 samples was constructed and operated as a data assimilation scheme with the stochastic system. The relative root mean square error (RRMSE) results indicate that the prediction error of the SIR particle filter data assimilation scheme is 48% smaller than the error from the deterministic model. By comparison of the plume contour figures, the SIR particle filter scheme also has the ability to give predictions that are much closer to any irregular contour shapes of true realities than the deterministic model does.

Parameter estimation was a significant part of the research. We adopted a different statistical approach towards coupling parameter estimation with the sequential Monte Carlo method. The main challenge was to develop a fitness function for weights

generation. The problem was resolved using the statistical concept of Euclidean norm to generate weights for the particles. Using the SIR particle filter unknown parameter (decay) value was predicted successfully. With the use of the updated parameter in the state prediction, prediction error of the SIR particle filter data assimilation scheme became 78% smaller than the error from the deterministic model. Future works include the use of the developed fitness function in Genetic Algorithm and Neural Network frameworks.

REFERENCES

Arnaud, D., Nondo, F., and Neil, G. (2001). "Sequential Monte Carlo methods in Practice." Springer-Verlag, New York.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking." *IEEE Transactions on Signal Processing*, 50(2), 174.

Bergman, N. (1999). "Recursive Bayesian Estimation Navigation and Tracking Applications." Linköping University.

Bucy, R. S., and Senne, K. D. (1971). "Digital synthesis of non-linear filters." *Automatica*, 7(3), 287-298.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999). " Improved particle filter for nonlinear problems. " *Institution of Electrical Engineers*, Stevenage, Royaume-Uni.

Chang, S. Y., and Li, X. (2006). "Modeling of chlorobenzene leaching from a landfill into a soil environment using particle filter approach." Proc., 2006 Int. Conf. on Environmental Informatics, Int. Society for Environmental Information Sciences, Bowling Green, KY.

Chen, T., Morris, J., and Martin, E. (2004a). "Particle filters for the estimation of a state space model." *Computer Aided Chemical Engineering*, A. Barbosa-Póvoa, and H. Matos, eds., Elsevier, 613-618.

Chen, W. -S., Bakshi, B. R., Goel, P. K., and Ungarala, S. (2004b). "Bayesian Estimation via Sequential Monte Carlo Sampling: Unconstrained Nonlinear Dynamic Systems." *Industrial & Engineering Chemistry Research*, 43(14), 4012-4025.

Cheng, X. (2000). "Kalman filter scheme for three-dimensional subsurface transport simulation with a continuous input." M.S. thesis, North Carolina A&T State Univ., Greensboro, NC.

Doucet, A., Godsill, S., and Andrieu, C. (2000). "On sequential Monte Carlo sampling methods for Bayesian filtering." *Statistics and Computing*, 10(3), 197-208.

Doucet, A., and Tadić, V. (2003). "Parameter estimation in general state-space models using particle methods." *Annals of the Institute of Statistical Mathematics*, 55(2), 409-422.

Duan, Q., Sorooshian, S., and Gupta, V. (1992). "Effective and efficient global optimization for conceptual rainfall-runoff models." *Water Resour. Res.*, 28(4), 1015-1031.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (2002). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." *Radar and Signal Processing, IEE Proceedings F*, 140(2), 107-113.

Han, X., and Li, X. (2008). "An evaluation of the nonlinear/non-Gaussian filters for the sequential data assimilation." *Remote Sensing of Environment*, 112(4), 1434-1449.

Heuvelink, M. G. B., and Webster, R. (2001). "Modelling soil variation : past, present, and future. " *Geoderma,* 100(3-4), 269-301.

Jin, A. (1996). " An optimal estimation scheme for subsurface contaminant transport model using Kalman-Bucy filter." Graduate student thesis, North Carolina A&T State Univ., Greensboro, NC.

Kim, J.-M., and Parizek, R. R. (1999). "Three-dimensional finite element modelling for consolidation due to groundwater withdrawal in a desaturating anisotropic aquifer system." *International Journal for Numerical and Analytical Methods in Geomechanics*, 23(6), 549-571.

Kiparissides, C., Seferlis, P., Mourikas, G., and Morris, A. J. (2002). "Online Optimizing Control of Molecular Weight Properties in Batch Free-Radical Polymerization Reactors." *Industrial & Engineering Chemistry Research*, 41(24), 6120-6131.

Kivman, G. A. (2003). "Sequential parameter estimation for stochastic systems." *Nonlinear Proc Geoph*, 10(3), 253-259.

Kozub, D. J., and MacGregor, J. F. (1992). "State estimation for semi-batch polymerization reactors." *Chemical Engineering Science*, 47(5), 1047-1062.

Li, P., Goodall, R., and Kadirkamanathan, V. (2004). "Estimation of parameters in a linear state space model using a Rao-Blackwellised particle filter." *IEE Proc.: Control Theory Appl*., 151(6), 727–738.

Li, X. (2006). "State and parameter estimation using the particle filter approach: Application on organic pollutant transport in ground water." M.S. thesis, North Carolina A&T State Univ., Greensboro, NC.

Liu, Y., and Gupta, H. V. (2007). "Uncertainty in hydrologic modeling : Toward an integrated data assimilation framework." *Water Resour. Res.*, 43, W07401.

Michigan Department of Environmental Quality (MDEQ). (2003). "Ground water statistics." MDEQ Water Division, Groundwater Section. (http://www.michigan.gov/documents/deq/deq-wd-gws-wcu-groundwaterstatistics_270606_7.pdf).

Owen, A. (1984). "Artificial diffusion in the numerical modelling of the advective transport of salinity." *Applied Mathematical Modelling*, 8(2), 116-120.

Pole, A., and West, M. (1990). "Efficient bayesian learning in non-linear dynamic models." *Journal of Forecasting*, 9(2), 119-136.

Pole, A., West, M., and Harrison, P.J. (1988). " Non-normal and non-linear dynamic Bayesian modelling." *Bayesian Analysis of Time Series and Dynamic Models*.

Pye, V. I., and Kelley, J. (1984). “The extent of groundwater contamination in the United States”. Chapter 1, Groundwater contamination, National Academy Press, Washington, DC, 23–34.

Ren, L., and Zhang, R. (1999). "Hybrid Laplace transform finite element method for solving the convection-dispersion problem." *Advances in Water Resources*, 23(3), 229-237.

Schnoor, J. L. (1996). " Environmental Modeling: Fate of Chemicals in Water, Air and Soil." John Wiley & Sons, New York.

Terwiesch, P., and Agarwal, M. (1995). "A discretized nonlinear state estimator for batch processes." *Computers & Chemical Engineering*, 19(2), 155-169.

Thiemann, M., Trosset, M., Gupta, H., and Sorooshian, S. (2001). "Bayesian recursive parameter estimation for hydrologic models." *Water Resour. Res.*, 37(10), 2521-2535.

Yeh, G.-T., Fang, Y., Zhang, F., Sun, J., Li, Y., Li, M.-H., and Siegel, M. (2010). "Numerical modeling of coupled fluid flow and thermal and reactive biogeochemical transport in porous and fractured media." *Computational Geosciences,* 14(1), 149-170.

Yu, Y., and Cheng, Q. (2006). "Particle filters for maneuvering target tracking problem." *Signal Process.*, 86(1), 195-203.

Zou, S., and Parr, A. (1995). "Optimal Estimation of Two-Dimensional Contaminant Transport." *Ground Water,* 33(2), 319-325.