2014

Efficiency Of Ensemble Square-Root Kalman Filter In 3D Subsurface Contaminant Transport Modeling

Torupallab Ghoshal

North Carolina Agricultural and Technical State University

Efficiency of Ensemble Square-Root Kalman Filter in 3D Subsurface Contaminant Transport Modeling

Torupallab Ghoshal

North Carolina A&T State University

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department: Civil, Architectural and Environmental Engineering

Major: Civil Engineering

Major Professor: Dr. Shoou-Yuh Chang

Greensboro, North Carolina

2014

The Graduate School
North Carolina Agricultural and Technical State University
This is to certify that the Master's Thesis of

Torupallab Ghoshal

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

Greensboro, North Carolina
2014

Approved by:

Dr. Shoou-Yuh Chang
Major Professor

Dr. Manoj K Jha
Committee Member

Dr. Stephanie Luster-Teasley
Committee Member

Dr. Sameer A. Hamoush
Department Chair

Dr. Sanjiv Sarin
Dean, The Graduate School

Biographical Sketch

Mr. Torupallab Ghoshal was born on September 04, 1982, in Dhaka, Bangladesh. He received Bachelor of Science degree in Civil Engineering from Bangladesh University of Engineering Technology (BUET) in 2005. He received his MBA degree majoring in Finance from Institute of Business Administration (IBA), University of Dhaka, Bangladesh. He has received Wadaran Kennedy 4.0 GPA Scholar Award (2013) at North Carolina A&T State University. Mr. Torupallab Ghoshal is a candidate for the Master of Science degree in Civil Engineering.

Dedication

This thesis is dedicated to my father, Sushanta Ghoshal, my mother, Sudha Acharyya, my wife, Bidita Banerjee and to my sister, Meghomallar Mukherjee for their love and encouragement.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Shoou-Yuh Chang, for his guidance, support and encouragement throughout my graduate studies. His profound wisdom and experience helped me to accomplish my research goals.

I would also like to thank Dr. Manoj K. Jha and Dr. Stephanie Luster-Teasley for being in my thesis committee. I also want to thank Sikdar Latif, Dr. Godwin Assumaning, Anup Saha, Md. Sayemuzzaman and Somsubhra Chattopadhyay for their kind support throughout my graduate studies. I also wish to thank all of the faculty and staff members in the Department of Civil, Architectural and Environmental Engineering at North Carolina A&T State University for their support and guidance.

This research was funded by the Department of Energy Samuel Massie Chair of Excellence Program under Grant No. DE-NA0000718. The views and conclusions herein are those of the writer's and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the funding agency.

Table of Contents

List of Figures

List of Tables

Abstract

A three dimensional subsurface contaminant transport model with advection, dispersion and reaction has been developed to predict transport of a reactive continuous source pollutant. Numerical Forward-Time-Central-Space (FTCS) scheme has been used to solve the advection-dispersion-reaction transport model and Kalman Filter (KF), Ensemble Kalman Filter (EnKF) and Ensemble Square Root Kalman Filter (EnSRKF) schemes have been used for data assimilation purpose. EnKF and EnSRKF both use Monte Carlo simulation in Bayesian implementation to propagate state estimation. The key difference between EnKF and EnSRKF is that EnSRKF does not require perturbation of observation during analysis stage. In this study, contaminant concentration is the state that has been propagated by this model. Reference true solution derived from analytical solution with added noise has been used to compare model results. Root Mean Square Error (RSME) profile shows that the EnSRKF concentration estimate can improve prediction accuracy better compared to numerical, KF and EnKF approaches. For a 10x12x4 space domain (480 nodes) with 10,000mg/L initial concentration, numerical scheme shows an average error of 127.01 mg/L, whereas EnSRKF shows an average error of 5.47 mg/L, indicating an improvement of 95.69%. KF and EnKF schemes show average error of 26.16 and 5.74 mg/L. Therefore, EnSRKF approach reduces mean RMSE by 79% and 4.70% compared to KF and EnKF approach respectively. Although EnSRKF shows marginal improvement compared to EnKF, EnSRKF is computationally cheaper compared to EnKF for larger problems with more nodes. For a 50x60x4 space domain (12,000 nodes) EnSRKF produces similar accuracy of EnKF with much less execution time. For 12,000-nodes domain, it can reduce computational time by 68% compared to EnKF. EnSRKF also shows better performance than EnKF with small ensemble sizes.

CHAPTER 1

Introduction

Groundwater is the primary source of drinking water for more than half of US population (Nolan et al., 1998) and in most rural areas it is the only source of drinking water supply. According to a USGS water usage report, in 2005 98% of self-supplied withdrawals were from fresh groundwater (Kenny et al., 2009). According to the same report, groundwater supplied 38% of total water usage in 2005 excluding the water usage in thermoelectric power generation (Kenny et al., 2009). Groundwater is also a very vital source of freshwater. Despite this immense importance of groundwater, however, it is always under the threat of contamination. With rapid industrialization, urbanization and increase of usage, the threat of contamination is increasing. Additionally, according to a USGS report, groundwater in the USA has been depleting at an increasing rate. USGS estimated that from 1900 to 2008 a total of 1000 km^3 of groundwater has been depleted in USA. In the recent years of 2000-2008, the depletion rate is highest (Konikow, 2013). Therefore, with increasing rate of groundwater depletion, it is becoming more and more important to preserve the current groundwater reserve in usable condition. With rapid industrialization, urbanization, and an increase in usage, it is practically impossible to keep groundwater completely free from any sort of contamination.

The most common reason for groundwater contamination is human activity. Densely populated areas are more vulnerable towards groundwater contamination. Groundwater can be contaminated from many sources such as, septic systems, improper disposal of hazardous waste, releases and spills from stored chemicals and petroleum products, landfills, surface impoundments, sewers and other pipelines, use of pesticides and fertilizers, drainage wells, etc (USEPA, 1993) .

Contaminants that are released in the environment can reach groundwater in several ways. The most common reason is percolation of the contaminant from land surface to unsaturated (vadose) zone. Different contaminants have different characteristics and can stay in groundwater for different durations of time and can pose different ranges of threats for human. Once groundwater is contaminated, it is very difficult to remove the contaminant and mitigate the problem. Therefore, preventive measures are usually taken whenever there could be possibility of contamination and also there are regulations to prevent contamination of groundwater. Despite all these measures and regulations, however, there are still plenty of serious cases of groundwater contamination in the USA.

Several factors contribute to make contaminant removal a very challenging problem. Expensive monitoring systems, heterogeneity of subsurface environment, different dispersion behavior of contaminants could be attributed to make the problem very difficult. For a successful mitigation procedure, the first and foremost important step is to locate the source and to know the propagation behavior of the pollutant plume. As many of serious contamination cases are point source pollution, it is relatively easy to locate the source of the contaminant. To know the details of the plume behavior, analytical, numerical and more advanced numerical models such as stochastic filtering techniques can be used.

Numerical models are quite popular in subsurface pollutant transport problems. However, numerical models are plagued with various limitations to predict transport of contaminant in subsurface environment. They cannot properly handle the uncertain heterogeneity of subsurface environment. Moreover, randomness of transport process, incorrect assumptions of parameters may contribute to lack of accuracy for numerical models. To solve advection-dispersion equations, numerical methods can be broadly classified as Eulerian, Lagrangian and mixed

Eulerian-Lagrangian methods (Neuman, 1984; Baptista, 1987). Eulerian methods are fixed grid and easy to implement. Several popular Eulerian approaches are finite difference and finite element schemes. But, they suffer from truncation error and they can produce considerable amount of numerical dispersion errors for advection dominated problems. To overcome the problem of numerical dispersion there are some stability check criteria to limit grid spacing and time step sizes. These constraints call for finer grid spacings and smaller time steps to solve transport problem in these methods. This can make the solution of the transport problem very expensive in terms of computational effort (Zheng & Bennett, 2002). These drawbacks can somewhat be overcome by analytical solution. Nonetheless, analytical solution also suffers problems like inaccurate assumptions of homogenous soil layers, assumption of isothermal condition and isotropic porous media, etc. These assumptions do not represent the true subsurface field behavior since true subsurface field contains prevalence of irregularities and heterogeneities.

The system model of numerical methods is based on some certain parameters like porosity, velocity of pollutants, retardation, etc. All these parameters may not be accurate enough to predict transport of a certain contaminant in a certain subsurface environment. Therefore, to improve prediction accuracy, collection of field data is quite important. Observation data can guide the system model of numerical approaches and can help to find the true state of pollutants. Data assimilation methods thus play a very important role in predicting subsurface transport mechanism of contaminants. Therefore, filtering techniques based on data assimilation became more and more popular in recent years for subsurface pollutant transport problems.

Kalman Filter (KF) and some descendents of Kalman filter, especially Ensemble Kalman Filter (EnKF) are extremely popular in hydrological and hydrogeological models. Kalman filter is a recursive data processing algorithm developed by Rudolf E. Kalman in 1960. Kalman filter is a very efficient tool as it estimates state of a process in a way that minimizes mean of squared error. Kalman filter is very robust in a sense that it can be used to estimate past, present and future states of a system (Welch & Bishop, 2006). Kalman filter is best suited for linear problems with relatively smaller number of state variables. When a system is nonlinear and there are large number of state variables Kalman filter could become prohibitively expensive (Bannister, 2012). To overcome these problems Ensemble Kalman Filter (EnKF) provides some better alternatives to estimate state of a system. EnKF uses Monte Carlo simulation for Bayesian estimation. Ensemble Kalman filter can handle large and nonlinear problems with better accuracy. In this paper another data assimilation technique namely Ensemble Square Root Kalman Filter (EnSRKF) is used to compare its performance with EnKF in subsurface contaminant transport modeling problem. EnSRKF can be defined as a variant of EnKF. The key difference between EnKF and EnSRKF is that EnSRKF scheme does not require observations to be perturbed during analysis stage (Whitaker & Hamill, 2002). Square root scheme was first introduced as an alternative implementation of EnKF to improve filtering performance.

In this paper, a synthetic case of subsurface contamination with a generic reactive (nonconservative) pollutant has been presented. The transport problem is an advection, dispersion and reaction problem with a known decay rate. Analytical solution with added noise has been used to determine the reference true solution and deterministic numerical solution with added noise has been used to determine the system state. In this model, contaminant concentration is the state that has been propagated through the various schemes. Simulated true values are obtained to guide state estimate. The numerical model has been coupled with filtering techniques for state estimation and data assimilation purpose. Two different cases are considered to determine the accuracy and efficiency of numerical method, Kalman filter, Ensemble Kalman filter and Ensemble Square Root Kalman filter schemes. In case 1, a domain with a total of 480 nodes and in case 2, another domain with 12,000 nodes has been used. To check performances of filtering techniques and due to presence of very large problems in real environmental modeling it is very important to check model accuracy and required computational effort when domain size increases. For example, despite being a very robust data assimilation technique for smaller domain problems, Kalman filter can be prohibitively expensive for larger domain problems due to its computational effort. Therefore, performances of Ensemble Kalman filter and Ensemble Square Root Kalman filter have been compared for two different scenarios. Although a 12,000 node domain is still small compared to real scenarios, it will help to demonstrate which technique works best when domain size increases. Root Mean Square Error (RMSE) is calculated for each scheme with respect to reference true solution. Reference true solution is obtained by adding random nose with analytical solution.

One objective of this study is to determine accuracy and effectiveness of Ensemble Square Root Kalman filter, Ensemble Kalman filter, Kalman filter and numerical model in a three dimensional subsurface contamination transport model. Another objective is to compare computational efficiency of Ensemble Square-Root Kalman filter and Ensemble Kalman filter in predicting contaminant transport in subsurface when domain size increases.

CHAPTER 2

Literature Review

Modeling of groundwater contaminant transport has been a challenging problem for civil engineers, hydrologists and hydrogeologists for decades. Obtaining accurate data from groundwater is an immense task as groundwater exhibits significant heterogeneity and very small spatial measuring error can produce completely different scenario than the real case. Analytical models are best suited for highly homogenous environment which is completely impractical for uncertain heterogeneity prevalent in subsurface. Traditional numerical models are plagued with various limitations to predict transport of contaminant in subsurface environment. They cannot properly handle the uncertain heterogeneity of subsurface environment. Numerical solution of advection-dominated subsurface transport equation has remained as an "embarrassingly" difficult problem for engineers (Mitchell, 1984). The primary reason for this difficulty is the presence of spatial first derivative term, advection and spatial second derivative term, dispersion in the single governing partial differential equation. If the transport equation contains reaction and if dispersion is considered in all three dimensions, then the problem becomes more complicated. Due to all of these difficulties involved in numerical solution there has been a burgeoning popularity of stochastic techniques to handle these types of prediction problems.

2.1 Studies on Data Assimilation Techniques

Kalman Filter (KF) was first proposed by Rudolf E. Kalman in 1960. It has wide range of applications in any optimal state estimation problem with dynamic nature. Welch and Bishop (2006) has a good discussion and derivation on KF. Despite being a very robust algorithm it cannot handle nonlinear dynamics which lead to some other descendant filtering techniques which can approach to handle nonlinear dynamics. Extended Kalman Filter (EKF) is one of the earlier developments in handling nonlinear dynamics. EKF can handle nonlinear dynamics using tangent linear function of nonlinear state transition matrix. Essentially EKF is a nonlinear approximation of the linear KF (Welch & Bishop, 2006). EKF needs to calculate and store prior and posterior state error covariance calculations which increase computational cost. On the other hand, Ensemble-based filtering techniques use statistical sampling techniques for forecast and analysis errors and thus these techniques can reduce computational cost significantly. Thus, KF, EKF and all other filtering techniques that do not use ensemble-based technique have a major common drawback. These techniques should only be applied for smaller systems requiring a small number of state variables to describe the whole system.

To overcome the inefficacy of KF and limitations of EKF to handle nonlinear dynamics and large problems Ensemble Kalman Filter (EnKF) is proposed. EnKF can handle purely nonlinear dynamics. The term 'ensemble' actually describes statistical samples. In EnKF a single state estimate is replicated by an ensemble of state estimates and the error covariance is calculated from the ensemble members instead of a separate covariance matrix for state. With any statistically representative ensemble size EnKF shows significant work reduction compared to KF and EKF. Tangent linear operator is not used in EnKF which leads to an easier implementation and it may have better handling capacity for nonlinearity (Burgers et al., 1998).

EnKF was originally introduced by Evensen (1994) for use on oceanic models, where state dimensions are usually very large. Subsequent development of EnKF shows that use of an ensemble of pseudo-random measurement perturbations is important to extract the right statistics from analysis ensemble. Houtekamer and Mitchell (1998) independently studied EnKF and showed that EnKF performs better with increase in ensemble size. In fact, for linear dynamics if ensemble size is infinity then EnKF approximation would yield the same result as of Kalman

Filter (KF). However, in practice only a statistically significant number of ensemble members (samples) can produce very good results. More studies and implementations of EnKF in different model can be found in Evensen and Leeuwen (1996), Evensen (1997), Houtekamer and Mitchell (2001).

EnKF has different implementation techniques. The most common one is the perturbed observation implementation. Many works on EnKF was based on perturbed observation technique. Houtekamer and Mitchell (1998) , Burgers et al. (1998) have implemented EnKF with perturbed observation. The reason of observation perturbation was to avoid divergence of the filter. However, several early works of EnKF did not use perturbed observation technique such as Evensen (1994); Evensen and Leeuwen (1996). Several papers have good discussion on Ensemble Kalman filter without perturbed observation. Lermusiaux and Robinson (1999); Anderson (2001); Bishop et al. (2001) and have discussed different approaches of EnKF without perturbed observation.

There are several approaches those do not require perturbations of observations during analysis stage. One approach that does not need perturbed observation is Ensemble Square-Root Filter (EnSRF) or Ensemble Square-Root Kalman Filter (EnSRKF). Whitaker and Hamill (2002) and Tippett et al. (2003) have demonstrated frameworks for Ensemble Square-Root filtering schemes. The 'square-root' term arises from these particular implementations as these implementations consider forecast or analysis errors as the 'square-root' of forecast or analysis covariance matrices (Bannister, 2012).

2.2 Implementation of Data Assimilation Techniques in Hydrology and Water Resources

Kalman filter has numerous implementations in fields of hydrology and water resources.

Ngan and Russel (1986); Stednick and Roig (1989); Yu et al. (1989) have applied KF in various areas of water resources.

Due to computational advantage and accuracy in performance EnKF is now widely used in areas where large dynamical models are present. Areas of Numerical Weather Prediction (NWP) and oceanic modeling have extensive applications of EnKF. However, in recent days EnKF is widely used in hydrology, water resources and environmental engineering also. Reichle et al. (2002) applied EnKF in soil moisture estimation. Authors concluded that EnKF can produce satisfactory results even with moderate ensemble sizes. Huang et al. (2008) used EnKF data assimilation technique to calibrate hydraulic conductivity field and to improve solute transport prediction with unknown initial contaminant source condition. Authors found that EnKF significantly improves the estimation of hydraulic conductivity and solute transport prediction.

Clark et al. (2008) used EnKF and EnSRKF in hydrological data assimilation in which streamflow observations were used to update states in a distributed hydrological model. Authors found that, EnSRKF performs better compared to EnKF to simulate the model. Chen et al. (2013) used EnSRKF to assimilate streamflow data in a flood forecasting model. As discussed earlier, Ensemble Square Root Kalman filter is a particular flavor of Ensemble Kalman filter. Some papers that used square root schemes in hydrology, water resource and environmental engineering may not used the term ‘square-root’, but used the generic term of Ensemble Kalman filter.

Zou and Parr (1995) used Kalman filter in their state-space model to obtain optimal estimation of contaminant plume in their two-dimensional advection-dispersion subsurface transport model. This paper used two independent estimation of plume concentration. One is process modeling and another one is measurement or observation modeling. To apply data assimilation technique successfully use of two separate models is very important. They used analytical model as a reference solution, a finite difference method (FDM) to generate process or system data and Method of Characteristics (MOC) model (Konikow and Bredehoeft, 1984) to generate measurement data. State-space optimal estimation was performed by KF which considers FDM solution as process model and MOC solution as measurement model. KF estimation reduced mean standard deviation by 30% compared to MOC solution and 20% compared to FDM solution. This approach shows that, despite the ability of numerical solution to predict the transport behavior by itself, the coupling of numerical solution with data assimilation technique produces much better results than using numerical solution alone.

Chang and Jin (2005) applied KF with regional noise in subsurface contaminant transport model. The authors used a small 220-node two-dimensional synthetic problem with advection-dispersion transport equation to analyze performance of KF. KF reduced contaminant transport prediction error up to 60% compared to finite difference based deterministic model. They used only 4 observation nodes in this whole domain of 220 nodes (1.82% of total nodes are observation nodes) and found that KF can successfully handle this sparse observation with reasonably less error.

Chang and Latif (2009) used KF and Particle filter approach in a one dimensional leachate transport model. The authors found that both schemes can improve prediction accuracy of contaminant transport by around 80% compared to numerical model. Chang and Assumaning (2011) applied KF and Particle filter schemes to model transport of radioactive pollutants in subsurface. Chang and Sayemuzzaman (2014) used Unscented Kalman filter in a two dimensional subsurface contaminant transport model.

Chang and Latif (2010) implemented Extended Kalman Filter (EKF) in 2D subsurface contaminant transport model with advection-dispersion. A finite difference method namely FTCS (Forward-Time and Central-Space) is used to generate process or system variables for EKF. Authors used two cases to determine effectiveness of EKF. In case 1 the number of observation nodes is same as that of state nodes and in case 2 a small number of observation nodes are used to prepare the measurement model. In both cases EKF significantly improves prediction accuracy over numerical scheme. EKF can reduce prediction error by 72% to 82% compared to numerical model.

Assumaning and Chang (2012) applied three different data assimilation (DA) techniques in a three dimensional advection-dispersion-reaction contaminant transport model with instantaneous pollutant source. These DA techniques were Kalman filter, Extended Kalman filter and Particle filter. Authors used a 12x12x3 domain with 432 nodes to describe the state model and 18 nodes to describe the measurement model. The authors concluded that filtering techniques could reduce the error of numerical scheme by about 70%.

CHAPTER 3

Methodology

In this subsurface contaminant transport problem analytical solution and numerical solution are used to prepare the necessary framework for data assimilation techniques. Analytical solution is used to determine the reference true solution and observation. Numerical solution is used to prepare the state model. In this study, traditional advection-dispersion-reaction equation has been used for non-conservative pollutant. This is a three dimensional model with advection in x direction and dispersion in all three directions. This synthetic model has been used to determine the accuracy and efficiency of numerical, KF, EnKF and EnSRKF approaches compared to true solution derived from analytical solution. The subsurface environment is considered as porous and saturated soil. For reaction term a first-order decay rate constant has been used. The three-dimensional form of the advection-dispersion-reaction equation for non-conservative pollutant in a saturated, homogeneous porous media with isotropic materials under uniform flow is given by the following partial differential equation:

$$\frac{\partial C}{\partial t} = \frac{D_x}{R}\frac{\partial^2 C}{\partial x^2} + \frac{D_y}{R}\frac{\partial^2 C}{\partial y^2} + \frac{D_z}{R}\frac{\partial^2 C}{\partial z^2} - \frac{v_x}{R}\frac{\partial C}{\partial x} - kC \tag{1}$$

Where, C = Contaminant concentration, mg/L

t = Time, day

D_x, D_y and D_z = Dispersion coefficients in x, y and z direction respectively, m^2/day

R = Retardation factor, dimensionless;

v_x = Velocity in x-direction, m/day

k = First-order decay rate constant, day^{-1}

x, y, z = Cartesian coordinates, m

The boundary conditions for the three dimensional solute transport model with a continuous contaminant source is expressed as

$$c(x_0, y_0, z_0, t) = C_0; \ \frac{\partial c}{\partial x} = \frac{\partial c}{\partial y} = \frac{\partial c}{\partial z} = 0, \qquad for\, x = y = z = \infty$$

Here (x_0, y_0, z_0) is the contaminant injection point and C_0 is the concentration of continuous source contaminant, mg/L.

3.1 Analytical Solution and Reference True Solution

The analytical solution for the governing partial differential equation is given by the following equation derived by Domenico (1987):

$$C(x, y, z, t) = \frac{C_0}{8} e^{\frac{xv_x}{2D_x}\left[1-\left(1+\frac{4\lambda D_x}{v_x^2}\right)^{\frac{1}{2}}\right]} erfc\left(\frac{x - \dfrac{v_x}{tR_f\left(1 + \dfrac{4\lambda D_x}{v_x^2}\right)^{\frac{1}{2}}}}{2\sqrt{D_x t/R_f}}\right)$$

$$* \left\{\mathrm{erf}\left[\frac{y + \frac{Y}{2}}{2\left(\frac{D_y x}{v_x}\right)^{\frac{1}{2}}}\right] - \mathrm{erf}\left[\frac{y - \frac{Y}{2}}{2\left(\frac{D_y x}{v_x}\right)^{\frac{1}{2}}}\right]\right\} \tag{2}$$

$$* \left\{\mathrm{erf}\left[\frac{z + \frac{Z}{2}}{2\left(\frac{D_z x}{v_x}\right)^{\frac{1}{2}}}\right] - \mathrm{erf}\left[\frac{z - \frac{Z}{2}}{2\left(\frac{D_z x}{v_x}\right)^{\frac{1}{2}}}\right]\right\}$$

Where, C (x, y, z, t) = Contaminant concentration, mg/L

t = time, day

C_0 = Initial concentration of continuous source contaminant

D_x, D_y and D_z = Dispersion coefficients in x, y and z direction respectively, m^2/day

v_x = Velocity in x-direction, m/day

λ = First-order decay rate constant, day^{-1}

R_f = Retardation factor, dimensionless

erf = Error function

erfc = Complementary error function

Y = Width of the contaminant source in saturated zone, m

Z = Depth of the contaminant source in saturated zone, m

x, y, z = Cartesian coordinates, m

Solution of this equation provides approximate solution of the governing partial differential equation. Analytical solution is used by Cheng (2000) for a three dimensional contaminant transport problem for continuous source pollutant. Chang et al. (2012) used analytical solution as reference true solution for their two dimensional contaminant transport model. Chang and Assumaning (2011) also used analytical solution as true solution for their two dimensional contaminant transport model with instantaneous input. In this paper, a random Gaussian error has been added with analytical solution to simulate reference true solution. The error is considered to be 5% of the analytical solution. Therefore, the true solution in this study is analytical solution added with 5% error. This true solution has been used to evaluate performance of numerical method and filtering techniques for a three dimensional model with continuous source pollutant.

3.2 Numerical Solution Approach

Forward-Time Central-Space (FTCS) finite difference method has been used to solve the three dimensional partial differential transport equation numerically. FTCS is an explicit method and therefore it is very efficient in terms of computational effort. Owen (1984) evaluated several mathematical models used in coastal and estuarine regions. One of the results he found was that FTCS scheme can always be used for advective transport with salinity. Chang and Li (2009) used FTCS scheme in their two dimensional transport model. Chang and Latif (2010), Chang et al. (2012) also used FTCS scheme to solve their transport models numerically. After state-space discretization by the FTCS method the following partial derivatives are obtained:

$$\frac{\partial c}{\partial t} \approx \frac{C(i,j,k,t+1) - C(i,j,k,t)}{\Delta t} \tag{3}$$

$$\frac{\partial c}{\partial x} \approx \frac{C(i+1,j,k,t) - C(i-1,j,k,t)}{2\Delta x} \tag{4}$$

$$\frac{\partial^2 c}{\partial x^2} = \frac{C(i+1,j,k,t) - 2C(i,j,k,t) + C(i-1,j,k,t)}{\Delta x^2} \tag{5}$$

$$\frac{\partial^2 c}{\partial y^2} = \frac{C(i,j+1,k,t) - 2C(i,j,k,t) + C(i,j-1,k,t)}{\Delta y^2} \tag{6}$$

$$\frac{\partial^2 c}{\partial z^2} = \frac{C(i,j,k+1,t) - 2C(i,j,k,t) + C(i,j,k-1,t)}{\Delta z^2} \tag{7}$$

Where *i* = spatial coordinate of nodes along *x* direction; *j* = spatial coordinate of nodes along *y* direction; *k* = spatial coordinate of nodes along z direction; *t* = coordinate of time step

And the equation for *kC*:

$$kC = k\frac{C(i,j,k,t+1) + C(i,j,k,t)}{2} \tag{8}$$

Substitution of equations (3) to (8) into the three dimensional partial differential subsurface transport equation (1) yields,

$$\begin{aligned} C(i,j,k,t+1) &= b_1C(i-1,j,k,t) + b_2C(i,j,k,t) + b_3C(i+1,j,k,t) \\ &+b_4C(i,j-1,k,t) + b_5C(i,j+1,k,t) \\ &+b_6C(i,j,k-1,t) + b_7C(i,j,k+1,t) \end{aligned} \tag{9}$$

Where

$$b_1 = (\frac{D_x\Delta t}{R\Delta x^2} + \frac{V\Delta t}{2R\Delta x})\Big/(1 + \frac{k\Delta t}{2}) \tag{10}$$

$$b_2 = (1 - \frac{2D_x\Delta t}{R\Delta x^2} - \frac{2D_y\Delta t}{R\Delta y^2} - \frac{2D_z\Delta t}{R\Delta z^2} - \frac{k\Delta t}{2})\Big/(1 + \frac{k\Delta t}{2}) \tag{11}$$

$$b_3 = (\frac{D_x\Delta t}{R\Delta x^2} - \frac{V\Delta t}{2R\Delta x})\Big/(1 + \frac{k\Delta t}{2}) \tag{12}$$

$$b_4 = b_5 = (\frac{D_y\Delta t}{R\Delta y^2})\Big/(1 + \frac{k\Delta t}{2}) \tag{13}$$

$$b_6 = b_7 = (\frac{D_z\Delta t}{R\Delta z^2})\Big/(1 + \frac{k\Delta t}{2}) \tag{14}$$

For state-space discretization the stability and convergence criteria for FTCS finite difference scheme can be defined in following way;

$$\Delta x < \frac{2D_x}{V} \tag{15}$$

And

$$\Delta t < \frac{1}{(\frac{2\Delta y D_x}{\Delta x^2 R})} \tag{16}$$

Now, based on equations (10) to (14) equation (9) can be rewritten in following matrix form;

$$\mathrm{c}_{(\mathrm{t+1})} = \mathbf{M} * \mathrm{c}_{(\mathrm{t})} \tag{17}$$

Where

$c_{(t+1)}$ is the state variable defined as vector of contaminant concentration at all nodes of the domain at time, t+1;

$c_{(t)}$ is the state variable defined as vector of contaminant concentration at all nodes of the domain at time, t

M is the State Transition Matrix (STM) containing all of the parameters for the model. It is a matrix composed of coefficients b_1 to b_7 of equations (10) to (14) in its main, upper and lower diagonal entries. With a known present time step concentration STM can determine concentration of next time step. Essentially STM propagates through every time step to provide the result of numerical solution approach.

3.3 System Equation Based on Numerical Solution Approach

System or process equation of the contaminant transport model is based on equation (17). To incorporate heterogeneity and stochastic behavior of subsurface environment a random Gaussian error has been introduced in the model derived from numerical solution. This error vector can be termed as process noise. Therefore, stochastic representation of the system equation can be expressed as;

$$\mathbf{x}_{\mathrm{t+1}} = \mathbf{M} * \mathbf{x}_{\mathrm{t}} + \mathbf{p}_{\mathrm{t}} \tag{18}$$

Where

$\mathbf{x}_{t+1}$ is the state variable defined as vector of contaminant concentration at all nodes of the domain at time, t+1

$\mathbf{x}_t$ is the state variable defined as vector of contaminant concentration at all nodes of the domain at time, t

$\mathbf{M}$ is the State Transition Matrix (STM) involving all the model parameters

$\mathbf{p}_\mathrm{t}$ is the process noise i.e. the random Gaussian error introduced in the system equation.

In this study, process noise is introduced as a percentage of the numerical solution for each time step and process noise is considered to be 10% of numerical solution. Therefore, the process equation used in all data assimilation techniques is numerical solution added with 10% random error. $\mathbf{p}_\mathrm{t}$ has zero mean and covariance $\mathbf{Q}_\mathrm{t}$. This stochastic representation of process or system equation is important for later consideration in filtering techniques. To analyze and infer a dynamic system at least two models are necessary. One is based on aforementioned process equation and this model describes the state evolution with time. Another one is to assimilate the noisy measurement to the state (the measurement or observation model).

3.4 Measurement (Observation) Model Based on Analytical Solution

Due to unavailability of field data a set of measurement data is simulated using reference true solution. Gaussian error has been added with true solution to obtain observation data required for measurement model. Measurement data are obtained by the following equation;

$$\mathbf{z}_\mathrm{t} = \mathbf{H} * \mathbf{x}_\mathrm{t}^\mathrm{T} + \mathbf{o}_\mathrm{t} \tag{19}$$

Where

$\mathbf{z}_\mathrm{t}$ is the vector of observed values for all nodes at time step t

$\mathbf{x}_\mathrm{t}^\mathrm{T}$ is the true solution of the state vector for all nodes at time step t

$\mathbf{o}_\mathrm{t}$ is the observation error vector

$\mathbf{H}$ is the observation data pattern matrix.

The observation data pattern matrix has some other popular names such as measurement sensitivity matrix or design matrix. The propagation of the state vector $\mathbf{z}_\mathrm{t}$ is dependent on observation data pattern matrix $\mathbf{H}$ and observation error vector $\mathbf{o}_\mathrm{t}$. $\mathbf{H}$ is an identity matrix to describe the pattern of observed data in the field. Similar to the process noise error, $\mathbf{o}_t$ is considered as the measurement noise. Measurement noise is considered to be 2.5% of the analytical solution. Therefore, 2.5% error is added with analytical solution in each observation node to make the observation matrix. It is an error vector of observation with zero mean and covariance matrix R_t. With the same percentage error approach described in the system model, observation noise is considered to be a percentage of the true solution.

3.5 Data Assimilation with Kalman Filter

Kalman Filter (KF) is a recursive data processing algorithm that can be used for any dynamic state estimation problem. It has two versions: one is discrete time and another one is continuous time. In this paper, the discrete time Kalman filter has been used. Kalman filter can assimilate noisy data in its analysis to recursively improve its estimation of state. It has wide range of application in any navigational state estimate problem such as, guiding rockets and missiles and in case of numerical weather prediction. The stochastic nature of Kalman filter exploits the theorem of Gauss-Markov model. At each discrete time step increment, a linear operator is applied to the current state to generate the new state, adding some noise. Also there is option to add control information on the system if they are known.

As discussed previously, filtering algorithms need at least two sets of models. One is system or process model and another one is measurement or observation model. In Kalman filter the stochastic state estimation equation can be written in following form:

$$\mathbf{x}_{t+1}(+) = \mathbf{x}_{t+1}(-) + \mathbf{K}_{t+1}(\mathbf{z}_{t+1} - \mathbf{H}\mathbf{x}_{t+1}(-)) \tag{20}$$

Where, $\mathbf{x}_{t+1}(+)$is the estimated value of state after the KF adjustment, and $\mathbf{x}_{t+1}(-)$ is the estimated value of state before the KF adjustment, i.e. the predicted value from the model.

Essentially, in the following equations (+) will indicate value after Kalman filter adjustment and (-) will indicate value before Kalman filter adjustment. The matrix K_{t+1} is defined by

$$\mathbf{K}_{t+1} = \mathbf{P}_{t+1}(-)\mathbf{H}^{T}[\mathbf{H}\mathbf{P}_{t+1}(-)\mathbf{H}^{T} + \mathbf{R}_{t+1}]^{-\mathbf{1}} \tag{21}$$

Here P_{t+1} is the optimal estimate of system error covariance matrix and can be estimated by

$$\mathbf{P}_{t+1}(+) = \mathbf{P}_{t+1}(-) - \mathbf{P}_{t+1}(-)\mathbf{H}^{T}[\mathbf{H}\mathbf{P}_{t+1}(-)\mathbf{H}^{T} + \mathbf{R}]^{-\mathbf{1}}\mathbf{H}\mathbf{P}_{t+1}(-) \tag{22}$$

Or rewriting by

$$\mathbf{P}_{t+1}(+) = \mathbf{P}_{t+1}(-)[\mathbf{I} - \mathbf{K}_{t+1}\mathbf{H}] \tag{23}$$

And

$$\mathbf{P}_{t+1}(-) = \mathbf{M}\mathbf{P}_{t+1}(+)\mathbf{M}^{T} + \mathbf{Q}_{t} \tag{24}$$

Here $\mathbf{K}_{t+1}$ introduced in equation (21) is called the Kalman optimal gain or Kalman filter. It determines how much the estimated value using this filtering algorithm can gain from the available observations.

Equations (18) to (22) and equation (24) are the primary six equations of Kalman filter. For prediction of optimal state by using Kalman filter, $\mathbf{x}_{t+1}$ of equation (18) is used and $\mathbf{x}_{t}$ is substituted by $\mathbf{x}_{t+1}$ in equation (19) to get observation vector $\mathbf{z}_{t+1}$. Then using equation (24), (22), (21), and (20) sequentially, optimal estimation of state, $\mathbf{x}_{t+1}(+)$ is estimated. Then this value of $\mathbf{x}_{t+1}$ is used to predict next time step state of $\mathbf{x}$ i.e. $\mathbf{x}_{t+2}$ by using all these aforementioned equations. This recursive algorithm will continue to operate up to the expected time step. Figure 1 describes the general recursive operation of Kalman filter.

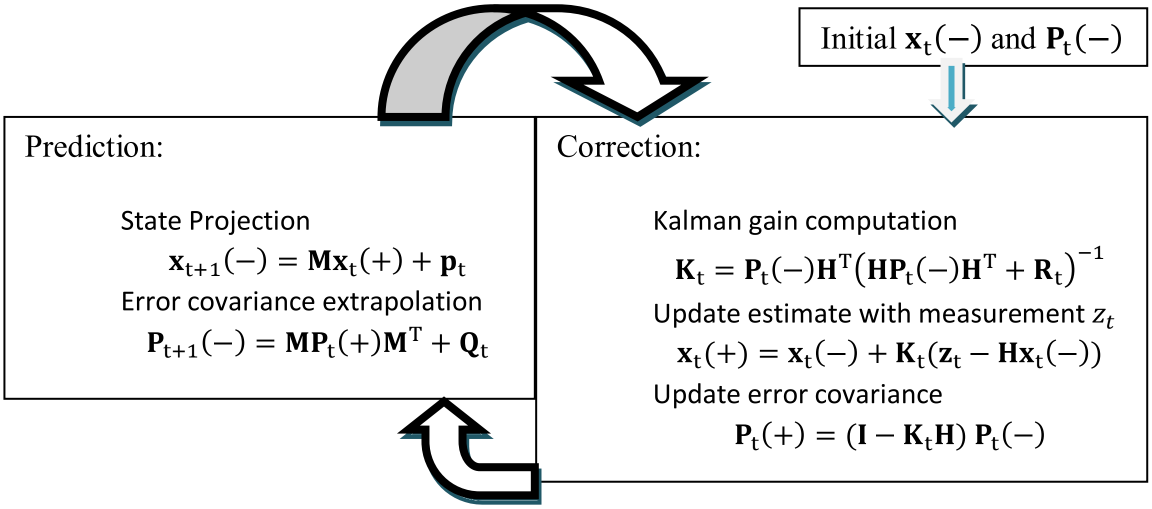
Initial $\mathbf{x}_t(-)$ and $\mathbf{P}_t(-)$
Prediction:
State Projection
$\mathbf{x}_{t+1}(-) = \mathbf{M}\mathbf{x}_t(+) + \mathbf{p}_t$
Error covariance extrapolation
$\mathbf{P}_{t+1}(-) = \mathbf{M}\mathbf{P}_t(+)\mathbf{M}^T + \mathbf{Q}_t$
Correction:
Kalman gain computation
$\mathbf{K}_t = \mathbf{P}_t(-)\mathbf{H}^T(\mathbf{H}\mathbf{P}_t(-)\mathbf{H}^T + \mathbf{R}_t)^{-1}$
Update estimate with measurement z_t
$\mathbf{x}_t(+) = \mathbf{x}_t(-) + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}\mathbf{x}_t(-))$
Update error covariance
$\mathbf{P}_t(+) = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\,\mathbf{P}_t(-)$

Figure 1. Sequential operation of Kalman filter (Welch & Bishop, 2006)

3.6 Data Assimilation with Ensemble Kalman Filter

Despite being a very robust algorithm Kalman filter is not applicable for estimation problems with nonlinear dynamics. Moreover, in environmental engineering, hydrology and hydrogeology research state variables could be very large and system may require a large number of information to describe the state. In this case, application of Kalman filter can become prohibitively expensive as because Kalman filter analysis step will have an inverse operation in each time step involving a very large matrix of error covariance. Ensemble Kalman Filter (EnKF) is an efficient data assimilation technique to deal nonlinear dynamics with lower computational effort compared to classical Kalman filter. In Ensemble Kalman filter the error covariance matrix is represented by stochastic ensemble of model realizations. EnKF uses sequential Monte Carlo simulation to generate required ensemble realizations.

EnKF has many implementations. Here EnKF implementation of Evensen (2003) applied by Chang and Latif (2011) is followed in this paper. Using Monte Carlo sampling the state matrix can be built as an arrangement of the ensemble members $\mathbf{x}_i \in \Re^n, (i = 1, ..., N)$

$$\mathbf{A}^{\mathrm{f}}_{\mathrm{t|t-1}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \Re^{n\times N}, \tag{25}$$

where *n* is the size of model state vector and *N* is the number of ensemble members. The ensemble mean can be calculated by the following operation.

$$\overline{\mathbf{A}}^{\mathrm{f}}{}_{\mathrm{t|t-1}} = \mathbf{A}^{\mathrm{f}}_{\mathrm{t|t-1}}\mathbf{B} \tag{26}$$

where $\overline{\mathbf{A}}^{\mathrm{f}}{}_{\mathrm{t|t-1}}$ is the ensemble mean matrix and $\mathbf{B} \in \Re^{N\times N}$ is the matrix where each element is equal to $1/N$ The ensemble residual matrix can be defined as,

$$\mathbf{E} = \mathbf{A}^{\mathrm{f}}_{\mathrm{t|t-1}} - \overline{\mathbf{A}}^{\mathrm{f}}{}_{\mathrm{t|t-1}} \tag{27}$$

The prior ensemble covariance matrix $\mathbf{P}_{e,t|t-1} \in \Re^{n\times n}$ can be defined as

$$\mathbf{P}_{\mathrm{e,t|t-1}} = \frac{\mathbf{E}\mathbf{E}^{\mathrm{T}}}{N-1} = \mathbf{P}_{\mathrm{e}}. \tag{28}$$

A vector of measurements $\mathbf{z} \in \Re^{m}$, with m being the number of observation nodes, can be considered as the mean vector of observation. In this stochastic model, the observation is explicitly treated as random variable and, therefore, can be replicated according to the Monte Carlo simulation with number of the ensemble members *N*. The perturbed observations are,

$$\mathbf{z}_{\mathrm{j}} = \mathbf{z} + \boldsymbol{\varepsilon}, \mathrm{j} = 1, \ldots, N, \tag{29}$$

and can be stored in the columns of a matrix

$$\mathbf{Z} = [\mathbf{z}_1,\ \mathbf{z}_2, \ldots, \mathbf{z}_N] \in \Re^{m\times N}, \tag{30}$$

while the ensemble of observation errors can be stored in the observation residual matrix,

$$\boldsymbol{\gamma} = [\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_N] \in \Re^{m\times N}, \tag{31}$$

The ensemble measurement error covariance matrix can be represented by the following equation,

$$\mathbf{R}_e = \frac{\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathrm{T}}}{N-1}. \tag{32}$$

The ensemble Kalman gain can be calculated by

$$\mathbf{K}_e = \mathbf{P}_e\mathbf{H}^{\mathrm{T}}\mathbf{W}^{-1}, \tag{33}$$

Where $\mathbf{H}$ is the observation operator; $\mathbf{W}$ is the ensemble innovation covariance matrix and expressed by

$$\mathbf{W} = \mathbf{H}\mathbf{P}_e\mathbf{H}^{\mathrm{T}} + \mathbf{R}_e \tag{34}$$

The residual matrix is defined as

$$\mathbf{Z}'_{\mathrm{e,t|t-1}} = \mathbf{Z}_{\mathrm{e}} - \mathbf{H}\mathbf{X}_{\mathrm{t|t-1}}. \tag{35}$$

The posterior estimate of the state matrix is calculated as

$$\begin{aligned}\mathbf{X}_{\mathrm{t|t}} &= \mathbf{A}^{\mathrm{f}}_{\mathrm{t|t-1}} + \mathbf{K}_{\mathrm{e}}\mathbf{Z}'_{\mathrm{e,t|t-1}} \\ &= \mathbf{A}^{\mathrm{f}}_{\mathrm{t|t-1}} + \mathbf{P}_{\mathrm{e}}\mathbf{H}^{\mathrm{T}}\mathbf{S}_{\mathrm{e}}^{-1}\mathbf{Z}'_{\mathrm{e,t|t-1}} \\ &= \mathbf{A}^{\mathrm{f}}_{\mathrm{t|t-1}} + \mathbf{P}_{\mathrm{e}}\mathbf{H}^{\mathrm{T}}\mathbf{S}_{\mathbf{e}}^{-1}\mathbf{Z}'_{\mathrm{e,t|t-1}} \\ &= \mathbf{A}^{\mathrm{f}}_{\mathrm{t|t-1}} + \mathbf{P}_{\mathrm{e}}\mathbf{H}^{\mathbf{T}}(\mathbf{H}\mathbf{P}_{\mathrm{e}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}_e)^{-1}\mathbf{Z}'_{\mathrm{e,t|t-1}}\end{aligned} \tag{36}$$

The inverse computation entails a potential singularity. (Evensen, 2003) prescribed a pseudo-inverse operation to take care of this potential singularity. Pseudo-inverse uses singular value decomposition approach to handle this inversion of potential singular matrix calculation. There is also another approach of eigenvalue decomposition to handle this inversion of potential singular matrix.

After obtaining the value of $\mathbf{X}_{\mathrm{t|t}} \in \Re^{n \times N}$, the mean analysis or posterior estimation of state is calculated by the following equation:

$$\overline{\mathbf{X}}_{\mathrm{t|t}} = \mathbf{X}_{\mathrm{t|t}}\mathbf{B}, \tag{37}$$

any column of $\overline{\mathbf{X}}_{\mathrm{t|t}}$ gives the analysis or posterior estimate, $\mathbf{x}^{\mathrm{a}}_{\mathrm{t|t}}$.

The posterior ensemble covariance matrix becomes

$$\mathbf{P}_{\mathrm{e,t|t}} = \frac{1}{N}(\mathbf{X}_{\mathrm{t|t}} - \overline{\mathbf{X}}_{\mathrm{t|t}})(\mathbf{X}_{\mathrm{t|t}} - \overline{\mathbf{X}}_{\mathrm{t|t}})^{\mathrm{T}} \tag{38}$$

The posterior state ensemble matrix at time step t, $\mathbf{X}_{\mathrm{t|t}}$ will be used to predict the prior at time step t+1, with a linear state transition operator $\mathbf{M}_{\mathrm{t}}$,

$$\mathbf{X}_{\mathrm{t+1|t}} = \mathbf{M}_{\mathrm{t}}\mathbf{X}_{\mathrm{t|t}}. \tag{39}$$

3.7 Data Assimilation with Ensemble Square-Root Kalman Filter

Ensemble Kalman filter has many different formulations with all of these differences contributed by the different approach in solving the analysis stage. The standard EnKF has an approach of perturbed observation in analysis step. Square root schemes don't need to perturb observation during assimilation procedure. Ensemble Square Root Kalman Filter (EnSRKF) can be termed as an efficient variant of Ensemble Kalman Filter (Evensen, 2003). With Comparison to EnKF, EnSRKF improves efficiency of analysis by avoiding perturbations of observation during assimilation period (Whitaker & Hamill, 2002). EnSRKF also does not have large inversion computation during analysis step which makes it a very efficient tool for data assimilation.

The implementation of EnSRKF is initiated following the implementation of EnKF. Similar to EnKF, EnSRKF also uses Monte Carlo sampling to perform the generation of the ensemble members. Implementation of EnSRKF starts with the equations (25) to (27) and equation (30) described in formulation of EnKF. Bannister (2012) showed a three step analysis procedure for EnSRKF. Author's procedure is mostly followed in following implementation.

The first step is to find a mean analysis state matrix, $\overline{\mathbf{X}}_{\mathrm{t}}^{\mathrm{a}}$:

$$\overline{\mathbf{X}}_{\mathrm{t}}^{\mathrm{a}} = \overline{\mathbf{A}}^{\mathrm{f}}{}_{\mathrm{t|t-1}} + \mathbf{E}\mathbf{S}^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{Z} - \mathbf{H}\overline{\mathbf{X}}_{\mathrm{t|t-1}}), \tag{40}$$

Where $\overline{\mathbf{A}}^{\mathrm{f}}{}_{\mathrm{t|t-1}}$ and $\mathbf{E}$ are ensemble mean matrix and ensemble residual matrix defined in equations (26) and (27) respectively. $\mathbf{H}$ is the observation data pattern matrix, $\mathbf{Z}$ is the observation data or measurement matrix (equation 30) and **S** and **C** matrices are defined below:

$$\mathbf{S} = \mathbf{HE} \in \Re^{m \times N}$$

And

$$\mathbf{C} = \mathbf{S}\mathbf{S}^{\mathrm{T}} + (N-1)\mathbf{R}_{\mathrm{e}} \in \Re^{m \times m},$$

where $\mathbf{R}_{\mathrm{e}}$ is the $m \times m$ observation error covariance matrix defined in equation (32). Then a matrix **G** is defined as: $\mathbf{G} = \mathbf{S}^{\mathrm{T}}\mathbf{C}\mathbf{S}$. Then eigenvectors **V** and eigenvalues **D** of the matrix **G** is calculated.

The second step is to calculate analysis perturbations, **A**:

$$\mathbf{A} = \mathbf{E}\mathbf{V}(\mathbf{I} - \mathbf{D})^{1/2}\mathbf{V}^{\mathrm{T}}. \tag{41}$$

Inversion of $(\mathbf{I} - \mathbf{D})^{1/2}$ involves potential singularity. Therefore, pseudo-inverse approach discussed in EnKF formulation has been used to avoid this potential singularity.

The final step is to assemble the full ensemble using analysis perturbation and eventually this model propagates this ensemble to next time step.

$$\mathbf{A}_{\mathrm{t}}^{\mathrm{a}} = \overline{\mathbf{X}}_{\mathrm{t}}^{\mathrm{a}} + \mathbf{A}, \tag{42}$$

$$\mathbf{A}_{\mathrm{t+1|t}}^{\mathrm{f}} = \mathbf{M}_{\mathrm{t}}(\mathbf{A}_{\mathrm{t}}^{\mathrm{a}}). \tag{43}$$

Where posterior state matrix at time t, $\mathbf{A}_{\mathrm{t}}^{\mathrm{a}}$ is used to update prior state matrix, $\mathbf{A}_{\mathrm{t+1|t}}^{\mathrm{f}}$ at time t+1 with a linear state transition operator matrix $\mathbf{M}_{\mathrm{t}}$.

3.8 The 3D Space Grid and Model Inputs

Three dimensional volumetric domain grids are used to illustrate the performance of contaminant transport models in this study. To evaluate performance of different data assimilation algorithms two domains are used. First domain is a smaller domain with less number of nodes and another domain has more number of nodes. First domain has been created with 10 nodes in X axis, 12 nodes in Y axis and 4 nodes i.e. layers in Z axis. In total it has 480 nodes. Grid spacing Δx, Δy and Δz in between each node is 5, 5 and 3 meters respectively. 30 time steps have been used to simulate the transport of concentration and duration of each time step, Δt is 0.75 days. A continuous pollutant source with an injection rate of 10,000mg/L is inserted in grid point (1, 6, 1). 9 observation nodes has been used in each layer, that means in four layers a total of 36 observation nodes has been used in a domain of 480 nodes. The other parameters are chosen according to suggestion by Zou and Parr (1995). These parameters are: velocity 0.5 m/d, retardation factor 1.5, Dispersion coefficients, $D_x = 3.0$ m^2/d, $D_y = 0.6$ m^2/d and $D_z = 0.7$ m^2/d and first order decay rate is 0.3/d. Second domain has been created with 50 nodes in X axis, 60 nodes in Y axis and 4 nodes i.e. layers in Z axis. In total, it has 12,000 nodes. As the second domain is very large compared to the first one, values of some parameters have been changed to get a larger shape of the contaminant plume. Otherwise, if same parameter values of small domain are used in case of the larger domain, the plume would look too small in a large domain and performance of the different data assimilation techniques will not be easy to distinguish visually. Table 1 is a summary of all these model parameters.

Table 1

Value of parameters for two different sized models

Parameters	Parameter values for Case 1	Parameter values for Case 2
Grids in X direction	10	50
Grids in Y direction	12	60
Grids in Z direction	4	4
Total nodes (n)	480	12,000
Node Spacing $\Delta x = \Delta y =$	5m	5m
Node Spacing $\Delta z =$	3m	3m
Total volume	50m X 60m X 12m	250m X 300m X 12m
Time step size, $\Delta \mathrm{t}$	0.75d	0.75d
Total time steps	30	40
Total simulation time	22.5 days	30 days
Initial concentration	10,000 mg/L	20,000 mg/L
Concentration input node	(1,6,1)	(1,30,1)
Decay rate	0.3/d	0.1/d
Velocity	0.5 m/d	1.2 m/d
Retardation coefficient	1.5	1.0
Dispersion coefficient at X direction, D_x	3.0 m^2/day	4.0 m^2/day

Table 1

Cont.

Dispersion coefficient at Y direction, D_y	0.6 m^2/day	2.0 m^2/day
Dispersion coefficient at Z direction, D_z	0.7 m^2/day	1.3 m^2/day
Ensemble size, N	100	100
Process noise	10.0%	10.0%
Measurement noise	2.5%	2.5%

Other than these parameters, 9 observation data points are considered to be located in each layer of these 3D six layer models. Therefore, in total there are 36 observation nodes in these example domains. For all of the simulations in this thesis, a system with 64-bit operating system with 2.5 GHz processor and 6.00 GB ram has been used.

The plan view of the small domain is shown in Figure 4. The initial point source of continuous source contaminant is located at node (1,6,1) for the small domain problem and shown by a large dot in the figure. The direction of the flow is shown by an arrow.

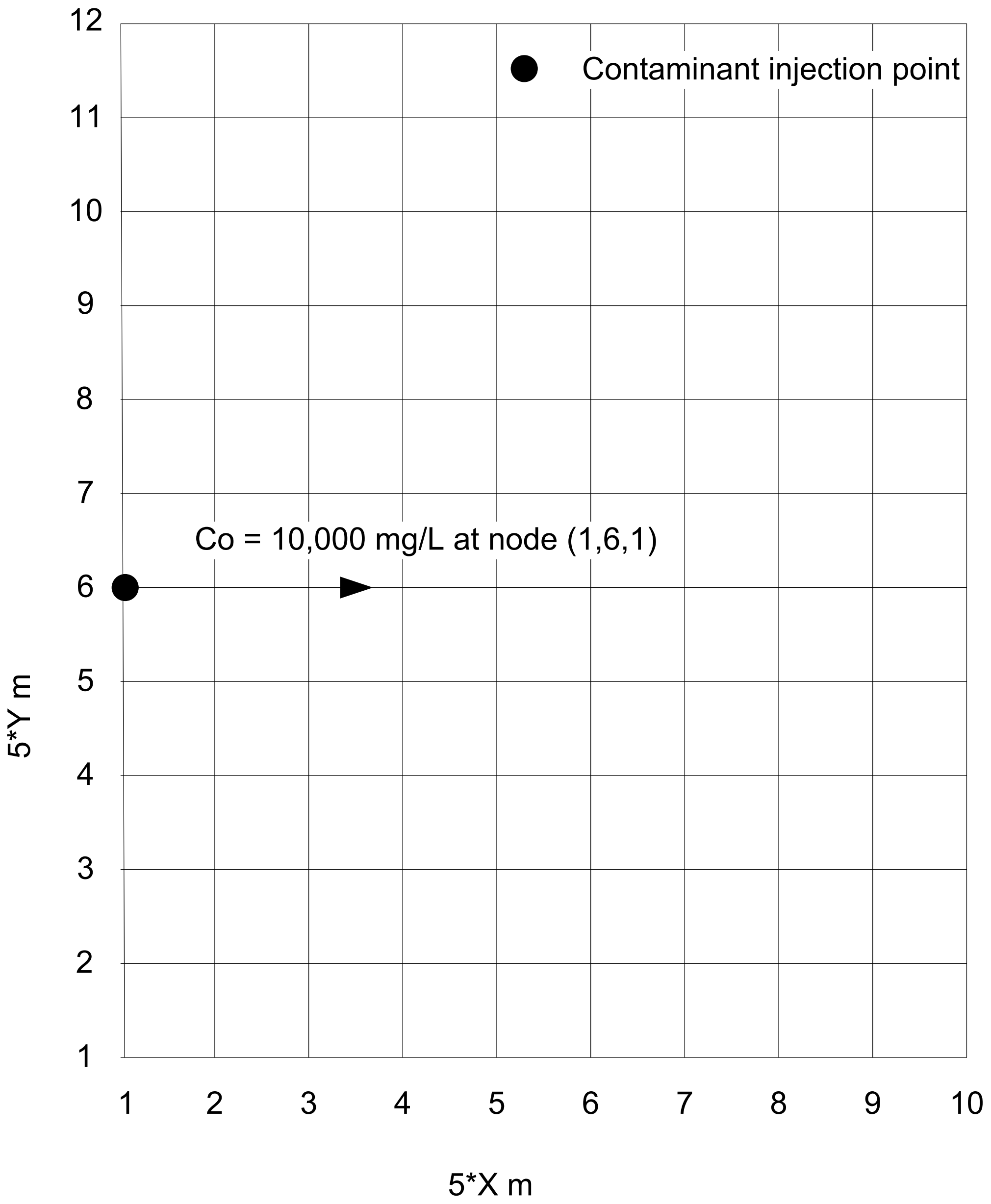

Figure 2. Top layer of the 3D experimental domain grid

3.9 Prediction Effectiveness and Accuracy Test

The effectiveness and accuracy of numerical, KF, EnKF and EnSRKF schemes are measured by comparing the respective model predicted results with the true value. The root mean square error (RMSE) is used as an effectiveness estimator. RMSE of each approaches are calculated using the following equation,

$$RMSE\ (t) = \sqrt{\frac{1}{n-1}\sum[C_P(i,j,k,t) - C_T(i,j,k,t)]^2} \tag{44}$$

Here $RMSE\ (t)$ is the error (mg/L) at time step t. n is number of nodes. $C_P(i,j,k,t)$ is the model predicted value of concentration in 3D coordinate system at time, t. $C_T(i,j,k,t)$ is the true solution of concentration in 3D coordinate system at time, t. To calculate an average of RMSE in all of the instances where RMSE profile is plotted an average of RMSE is calculated by summing up root mean square error of all time steps and dividing that sum by number of time steps.

CHAPTER 4

Results and Discussion

4.1 Case 1 with a Domain of 480 Nodes

A computer code is developed to run the mathematical models of FTCS numerical solution, KF, EnKF and EnSRKF data assimilation techniques. A three dimensional model with 10x12x4 grids (480 nodes) has been constructed to simulate contaminant transport prediction of each techniques. The simulations are run for 30 time steps. To evaluate spreading behavior of plumes, a contaminant concentration profile is drawn for each layer. In Figure 3, the concentration distribution of each approach is presented after time step 5 for layer 1.

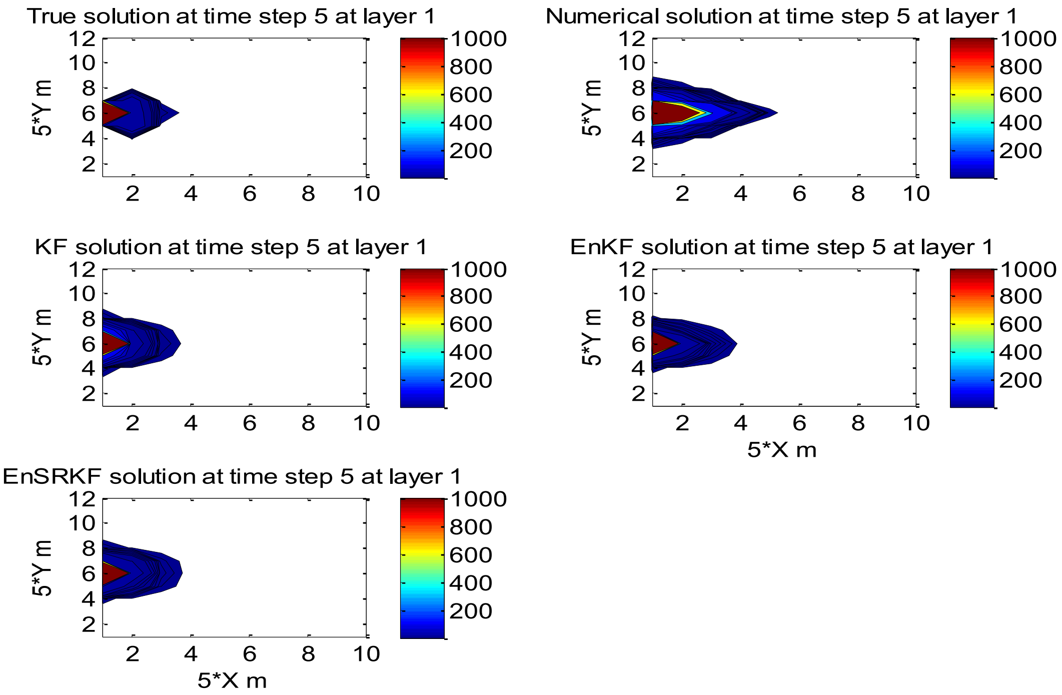

Figure 3. Contaminant concentration contour profile after time step 5 at layer 1 for True, Numerical, KF, EnKF and EnSRKF solution.

From Figure 3 it is clear that, after time step 5 numerical solution is moving faster compared to true solution. At this early stage it is difficult to distinguish KF, EnKF and EnSRKF plumes; however, these have better shape compared to the numerical solution. MATLAB generated colorbar adjacent to each plot indicates the range of contour line concentrations. As contaminant is injected through layer 1, concentration of contaminant would be highest in layer 1 and it is reflected in the colorbars as all of the colorbars have a highest value of 1000 mg/L. Next, in figure 4 and 5, layer 2 and layer 3 contour profiles are plotted after time step 5.

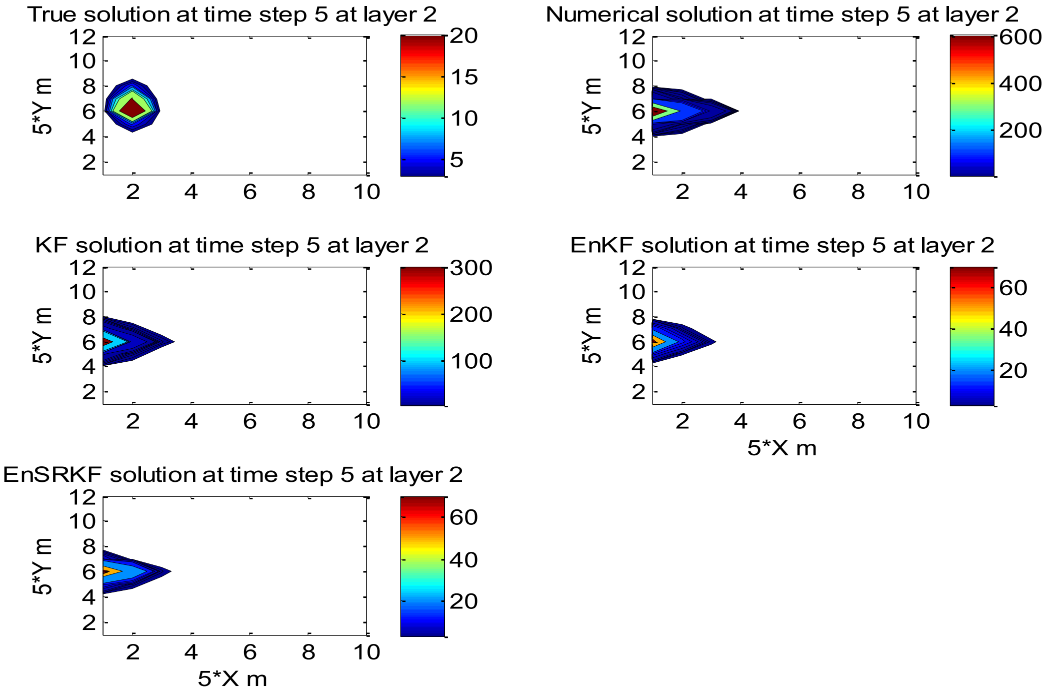

Figure 4. Contaminant concentration contour profile after time step 5 at layer 2 for True, Numerical, KF, EnKF and EnSRKF solution.

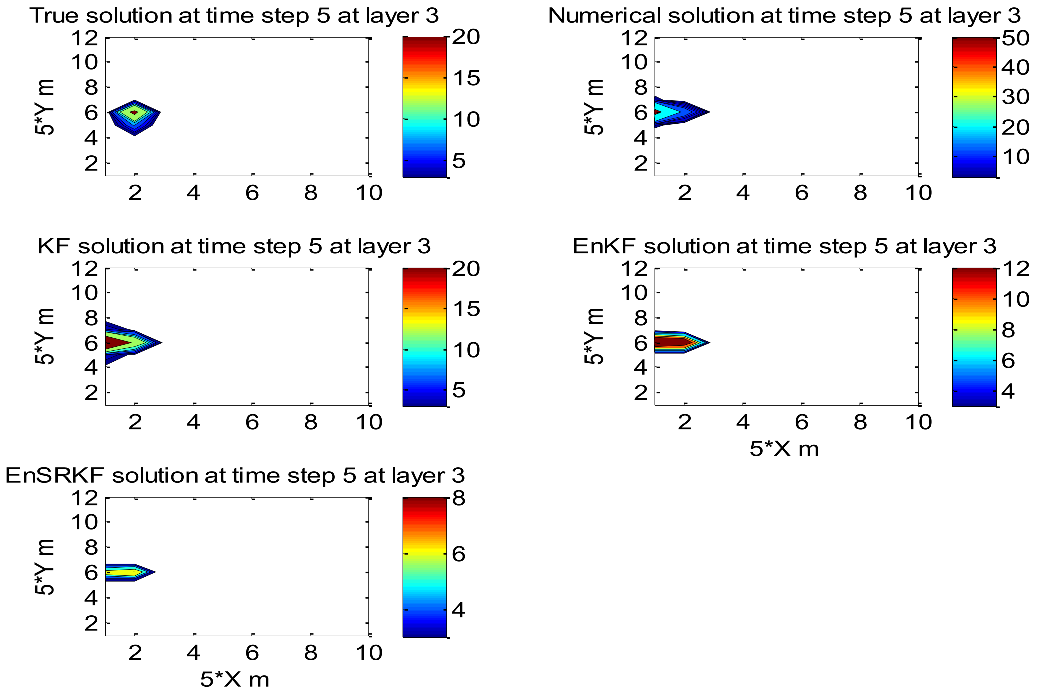

Figure 5. Contaminant concentration contour profile after time step 5 at layer 3 for True, Numerical, KF, EnKF and EnSRKF solution.

From Figure 4 and 5 it is clearly visible that, concentrations of contaminant plumes are becoming reduced with layer level. To clearly identify the change the MATLAB colorbars with auto-generated scaling associated with contour profiles can be noticed. For example, in Figure 3 the range of colorbar for contours was from 1 to 1000 mg/L for all approaches, however, for Figure 5 the colorbar shows a significant change in concentration range. True, EnKF, EnSRKF solutions have ranges with smaller values in the colorbars. This is quite reasonable and expected because layer 1 should have highest level of contour concentration profile as contaminant is injected at layer 1. As contaminant is injected in a point at the top layer (node (1,6,1) in 3D coordinate system) the contaminant concentrations farther from this point will remain lower compared to the contaminant concentrations closer to this point. Next, in Figure 6 concentration profiles of contaminants are shown for time step 10 at layer 1. Figure 6 shows that at time step 10 at layer 1, the contaminant is more spread out compared to the time step 5 depictions shown in Figure 3. Here it can be noticed again that, numerical solution is moving faster compared to True solution, KF and ensemble techniques.

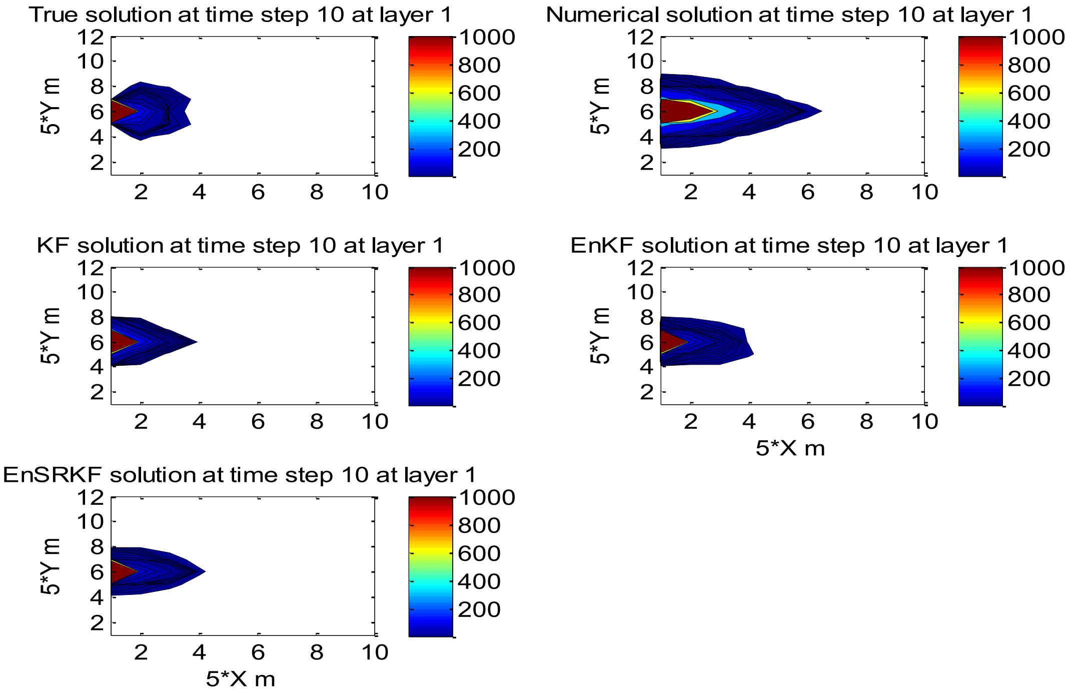

Figure 6. Contaminant concentration contour profile after time step 10 at layer 1 for True, Numerical, KF, EnKF and EnSRKF solution

To evaluate performances of different approaches at time step 10, another plot of concentration profiles is presented at Figure 7 to compare performance of the different approaches. Figure 7 provides concentration profiles after time step 10 at layer 3. Here, ranges of colorbars of true, KF, EnKF, EnSRKF solutions are very similar and the plume distribution of true solution is more similar to that of EnKF and EnSRKF solution.

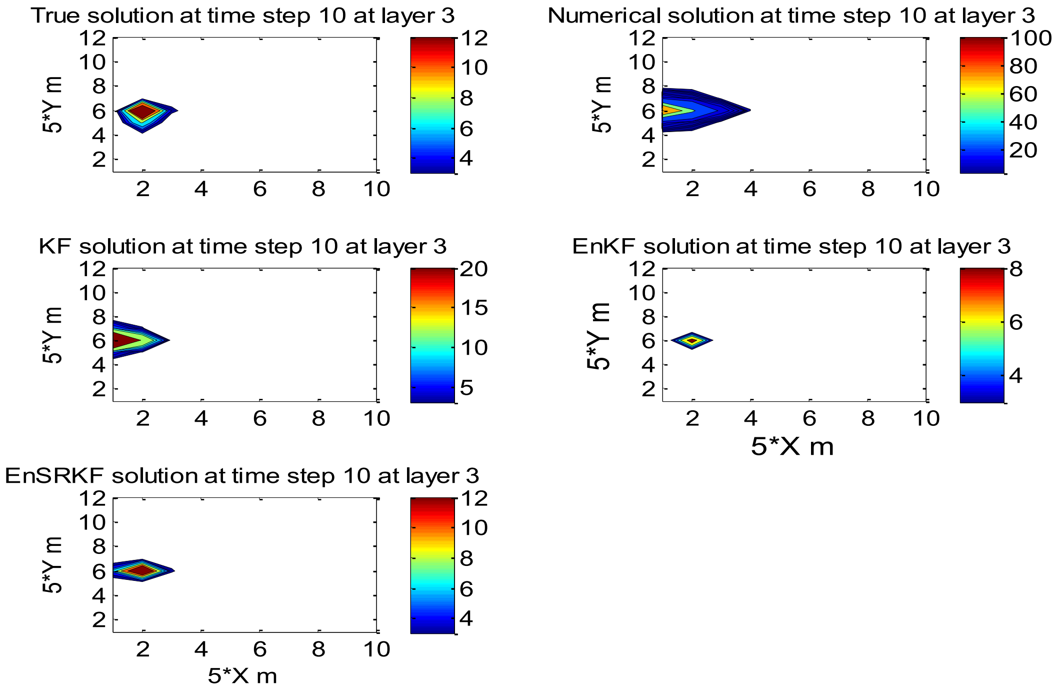

Figure 7. Contaminant concentration contour profile after time step 10 at layer 3 for True, Numerical, KF, EnKF and EnSRKF solution

Range of colorbar for true solution is 1 to 12 which is equal to the range of colorbar of EnSRKF. EnKF has a range of 1 to 8 and KF has a range of 1 to 20. However, range of colorbar of numerical solution is quite different from that of true solution. For example, numerical solution has a range of 1 to 100. This indicates that filtering techniques work better compared to the numerical method in predicting the contaminant transport. The concentration profile after time step 20 at layer 1 is shown in Figure 8. In figure 8, concentration profile for each scheme is more spread out as 20 time steps out of total 30 time steps of simulation have already passed.

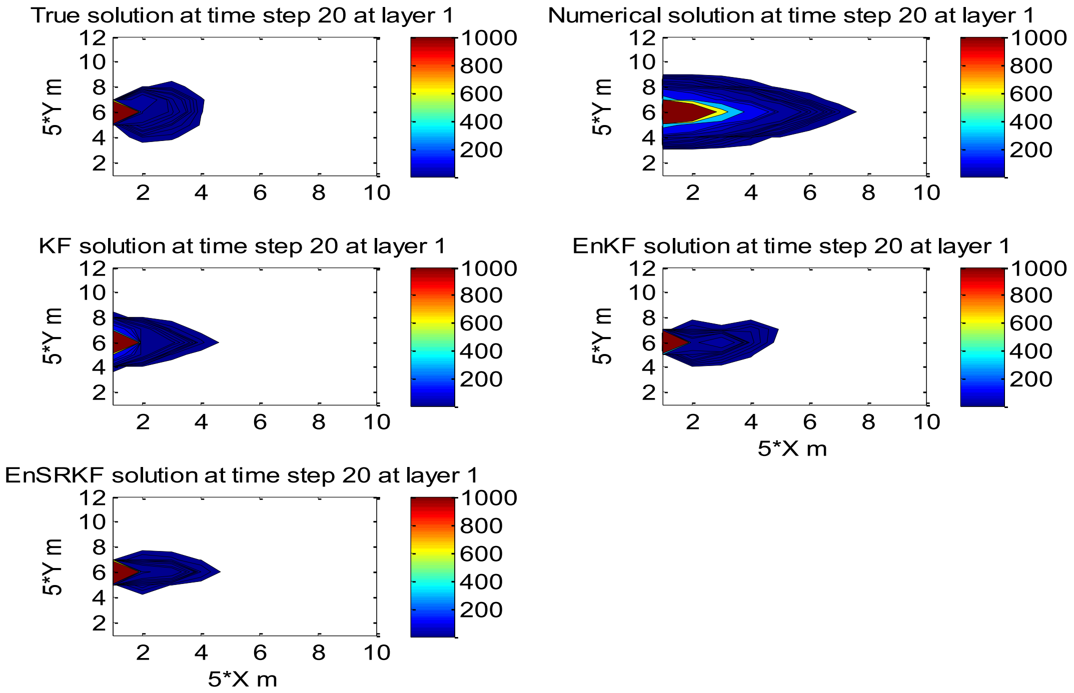

Figure 8. Contaminant concentration contour profile after time step 20 at layer 1 for True, Numerical, KF, EnKF and EnSRKF solution

Figure 8 displays that, when none of the other schemes have reached the halfway of the domain, numerical solution has already beyond the halfway. Numerical solution adds up errors of each time step and therefore, it is always showing more errors in predicting the concentration plume. Numerical solution, without having any data assimilation technique, therefore, does not perform well to predict contaminant transport in subsurface. Concentration plumes of true, EnKF and EnSRKF solutions are nearly similar to each other. However, from layer 1 it is not possible to distinguish each other as layer 1 has a continuous source of 10,000 mg/L.

Figure 9 provides contour profile of concentration after time step 20 at layer 3.

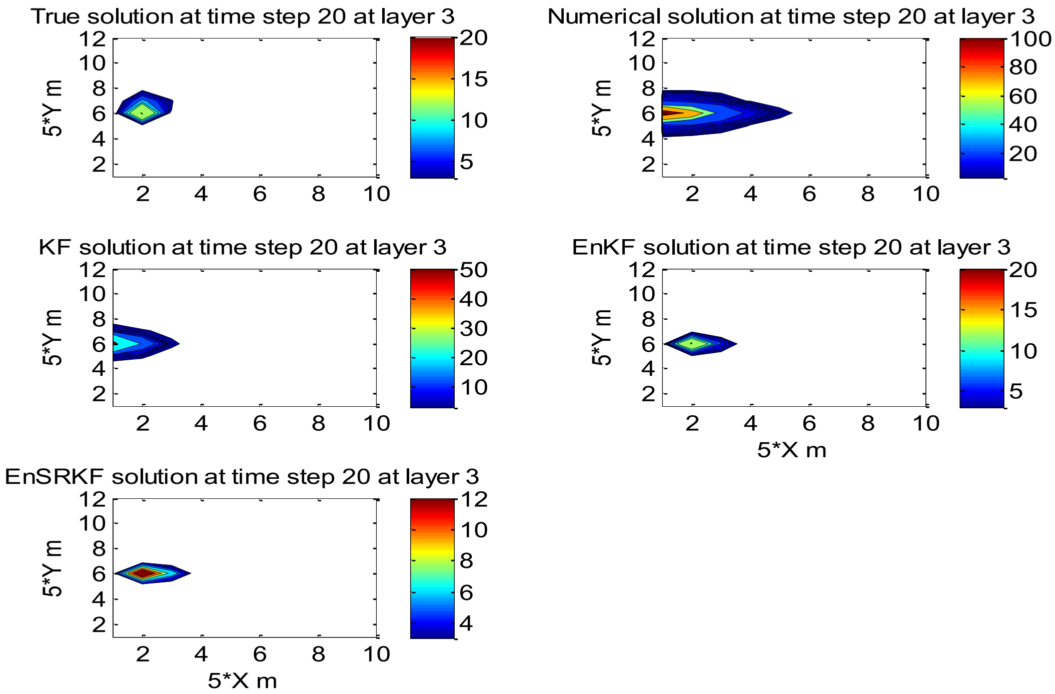

Figure 9. Contaminant concentration contour profile after time step 20 at layer 3 for True, Numerical, KF, EnKF and EnSRKF solution

Figure 9 shows that, shapes of concentration plume generated by true solution is better predicted by EnSRKF and EnKF schemes. Kalman filter has more errors compared to ensemble-based techniques but better prediction ability compared to numerical scheme. According to the colorbar, range of contours for true solution is from 1 to 20 mg/L. EnKF has the same range of 1 to 20 mg/L, EnSRKF has a range of 1 to 12 mg/L and numerical scheme has a range of 1 to 100 mg/L. Although, EnSRKF has a small range of concentration, a closer inspection indicates a larger area of EnSRKF has the peak concentration of its color range. EnSRKF has highest contour value of 12 mg/L shown in red in colorbar and in the center of the plume it has reasonably larger area with red color. On the other hand, despite EnKF has a peak contour value of 20 mg/L indicated by red color in its colorbar, the center of the plume shows a green area which indicates much of it highest concentration is between 10 to 15 mg/L. Therefore, despite having two marginally different contour ranges, EnKF and EnSRKF actually have reasonably similar range of contour plumes. Figure 10 shows contour profile of concentration after the last time step of 30 at layer 1.

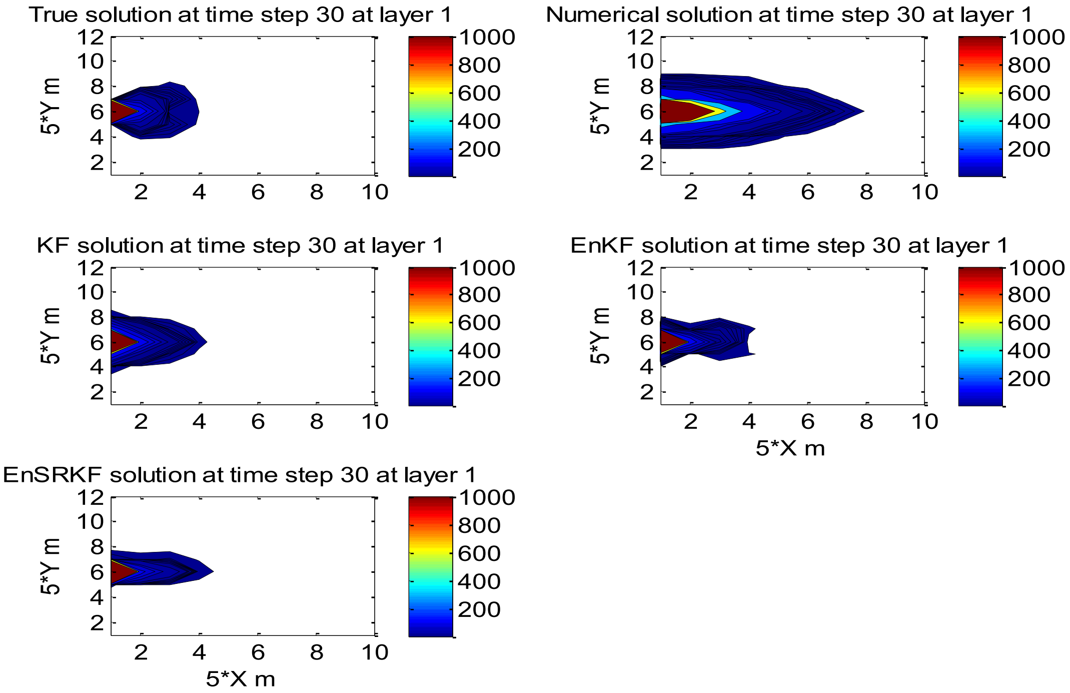

Figure 10. Contaminant concentration contour profile after time step 30 at layer 1 for True, Numerical, KF, EnKF and EnSRKF solution

At layer 1 of the domain all of the contours are well spread and numerical scheme is overpredicting the concentration plume as seen in earlier plots. Filtering techniques are working better to assimilate the plume generated by the true solution. To evaluate the performance of the data assimilation techniques another profile is plotted after time step 30 at layer 3.

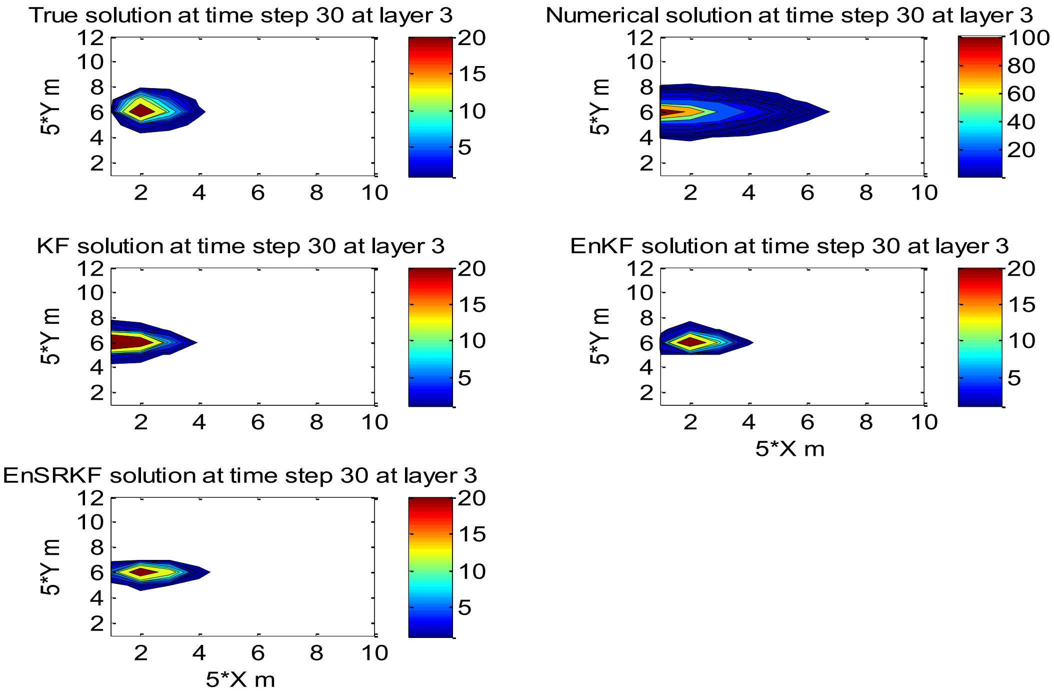

Figure 11. Contaminant concentration contour profile after time step 30 at layer 3 for True, Numerical, KF, EnKF and EnSRKF solution

Figure 11 shows that, shape of contour for true solution, EnKF and EnSRKF are reasonably similar. The colorbar shows that, the true, KF, EnKF and EnSRKF solution has a same scale of 1 to 20 mg/L. Numerical solution has a scale of 1 to 100 mg/L. In Figure 12, 13, 14 and 15 a side-by-side comparison between true solution and numerical, KF, EnKF and EnSRKF has been shown after time step 30 for all layers. Contour plumes of only time step 30 is chosen as after 30 time steps simulation ends and the plume is well developed at this stage.

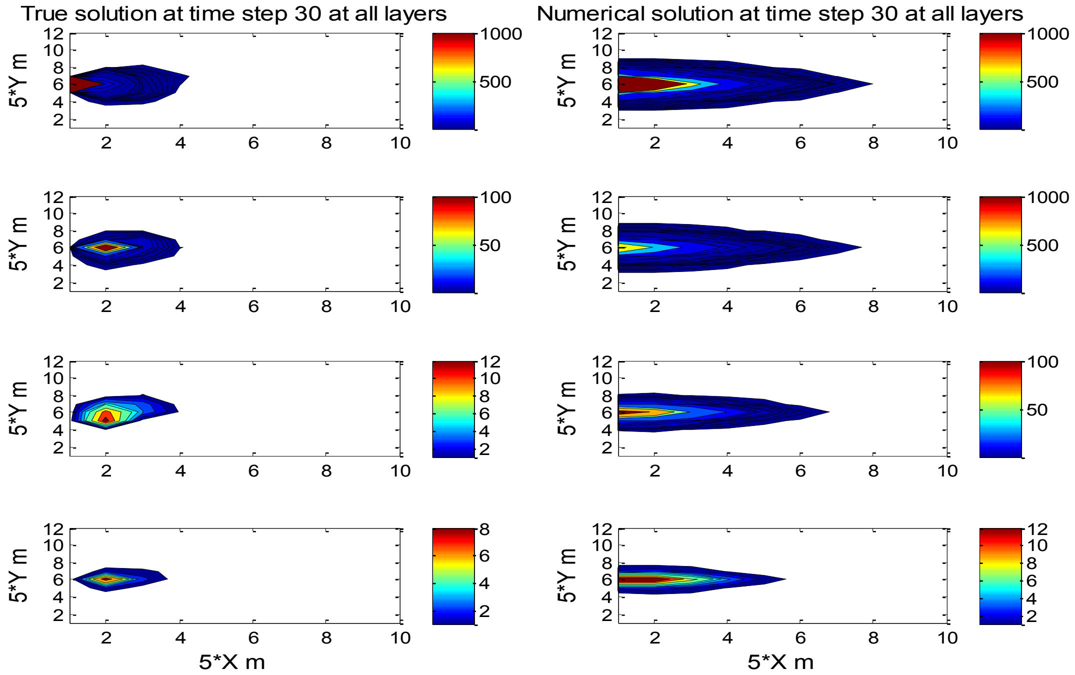

Figure 12. Comparison of true and numerical solutions after time step 30 at all layers

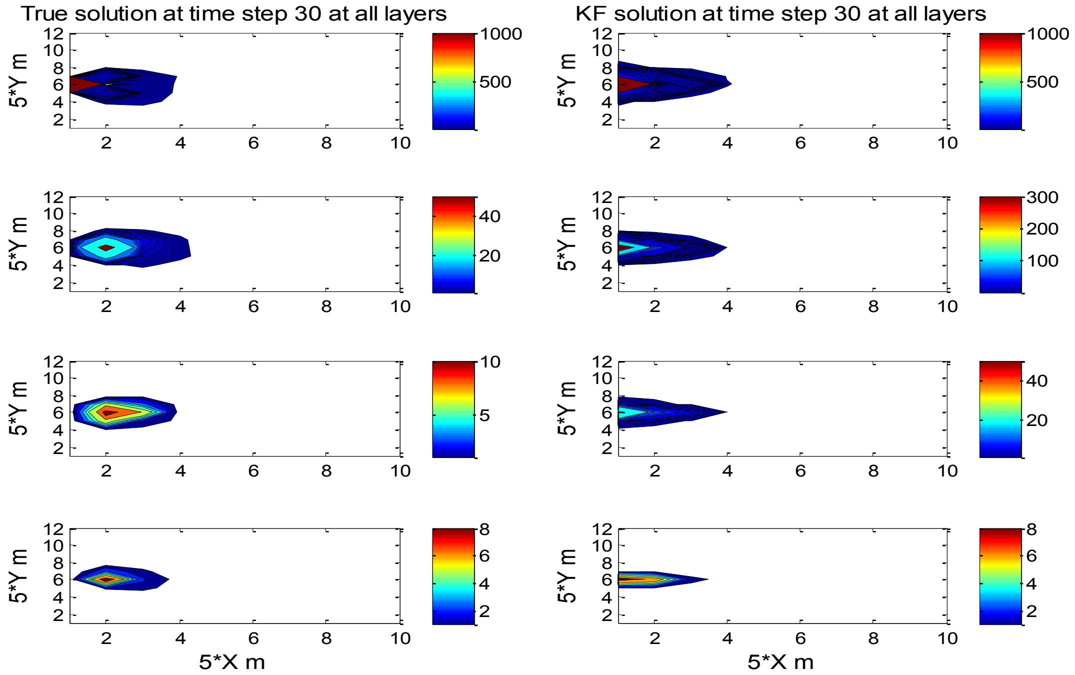

Figure 13. Comparison of true and KF solutions after time step 30 at all layers

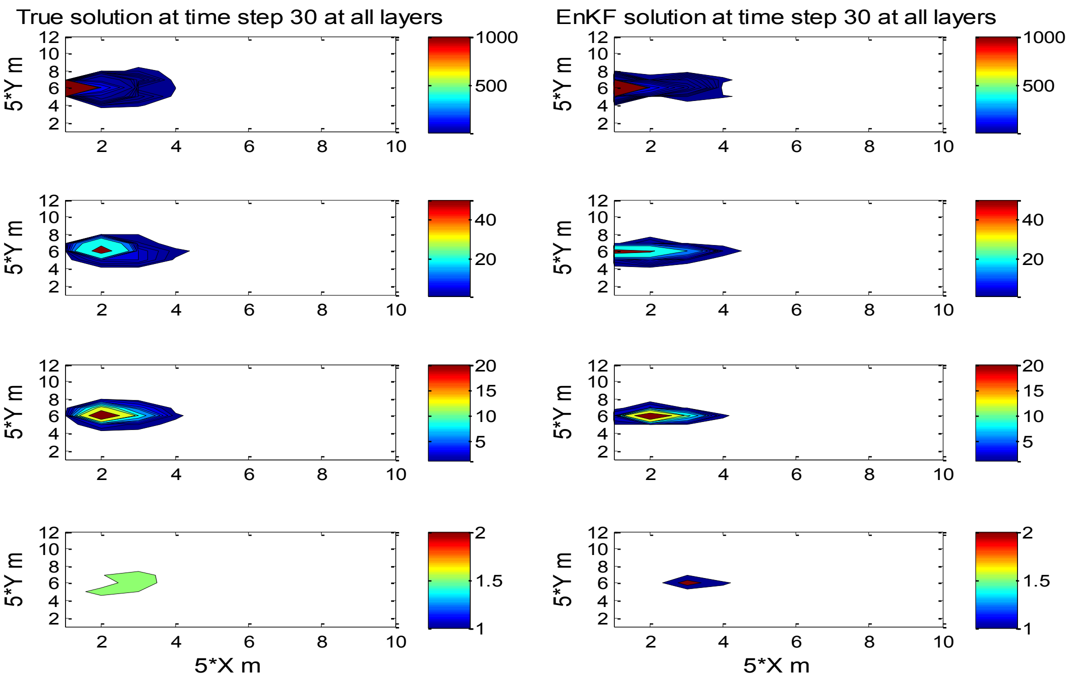

Figure 14. Comparison of true and EnKF solutions after time step 30 at all layers

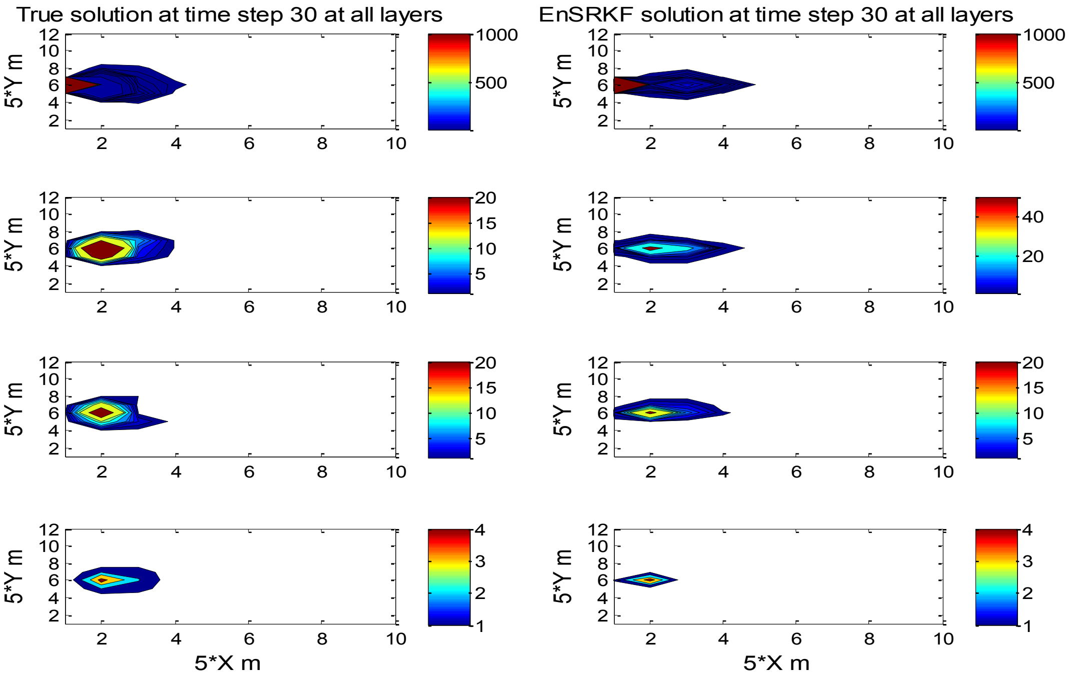

Figure 15. Comparison of true and EnSRKF solutions after time step 30 at all layers

Figures 12, 13 14 and 15 provide a depiction how numerical. KF, EnKF and EnSRKF schemes performs compared with true solution. From shapes of contours and from scales of colorbars it is quite evident that. For each layer EnSRKF, EnKF and KF have better prediction accuracy compared to numerical and KF solutions.

To get a clear idea how all these approaches are working a RMSE profile is drawn to estimate root mean square error of each scheme compared to the true solution. Sometimes plotting only the contour plumes may not give a definitive indication of performance of each scheme as contour profiles may show very similar plume distributions visually. Therefore, a RMSE profile is plotted in Figure 16 to present the performances of each scheme in a single plot.

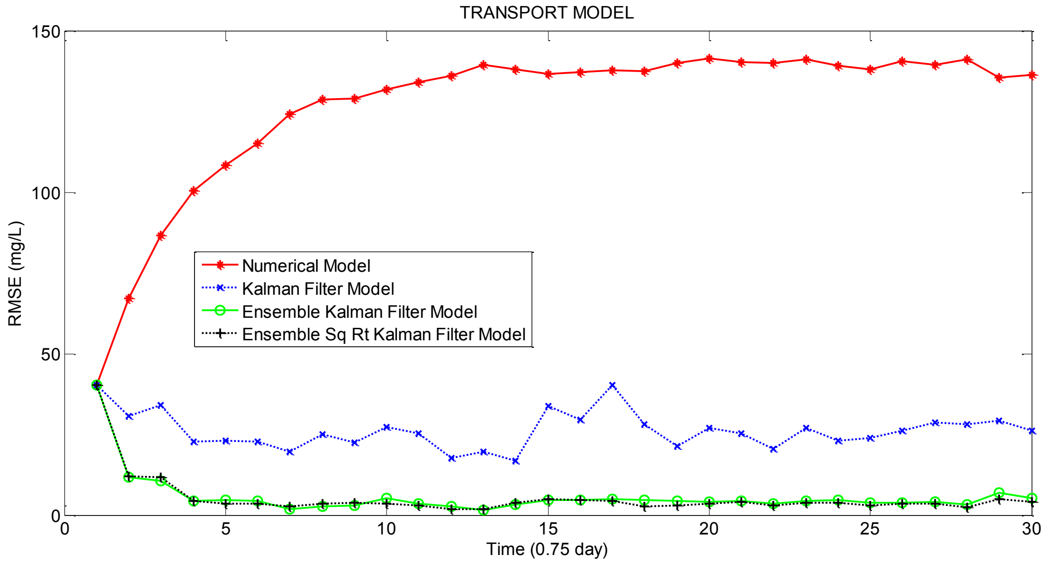

Figure 16. Root Mean Square Error (RMSE) profiles of Numerical, KF, EnKF and EnSRKF solutions

Figure 16 plots RMSE profile of each scheme for 30 time steps. Error of Numerical solution continues to increase until around time step 15, after that it stabilizes. Kalman filter RMSE shows that it converges quickly and remains relatively flat throughout the entire simulation period. EnKF and EnSRKF have least errors compared to numerical and KF schemes. However, it is difficult to distinguish which scheme has less error between EnKF and EnSRKF solution. In fact, EnKF and EnSRKF both are types of ensemble Kalman filters and as EnSRKF can be termed as a special flavor of EnKF, it may be expected that both schemes may have similar accuracy. To determine average error of each scheme, the errors of each scheme for each time steps can be summed up and divided by 30 to get an average of RMSE. Before calculating the average RMSE for each approaches the entire code was run five times to check if all of the schemes are converging properly. A decision based on a single run can have a scope of more errors compared to a decision based on multiple runs and taking their average. In Figure 17, results of RMSE profile are shown for five different runs in a single plot.

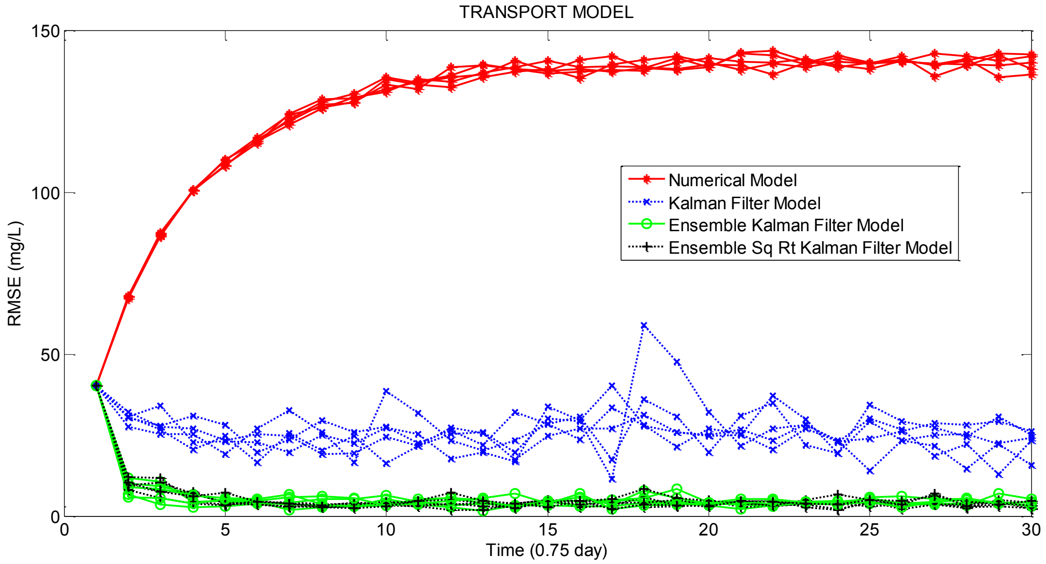

Figure 17. Root Mean Square Error (RMSE) profiles of Numerical, KF, EnKF and EnSRKF solutions for five different runs.

Figure 17 shows that each of the schemes maintaining their trend more or less in each run and therefore, it can be concluded that all the schemes are converging properly. RMSE of 5 different runs and their average is shown in Table 2.

Table 2

Results of five different RMSE profile runs and the average RMSE of each scheme

Scheme	Run 1	Run 2	Run 3	Run 4	Run 5	**Average**
Numerical RMSE(mg/L)	126.6	127.44	127.04	127.40	126.57	**127.01**
KF RMSE(mg/L)	26.19	25.86	26.39	25.77	26.60	**26.16**
EnKF RMSE(mg/L)	5.70	5.58	5.68	6.00	5.76	**5.74**
EnSRKF RMSE(mg/L)	5.29	5.42	5.16	5.80	5.70	**5.47**

Taking average of RMSE for five different runs yield that Numerical solution has the highest value in RMSE as it was seen in the Figure 17. Numerical solution has an average RMSE of 127.01 mg/L, Kalman filter solution has an average RMSE of 26.16 mg/L, Ensemble Kalman filter solution has an average RMSE of 5.74 mg/L and Ensemble Square Root Kalman filter solution has an average RMSE of 5.47 mg/L. EnSRKF solution can improve prediction accuracy by 95.6%, 79% and 4.7% compared to numerical, KF and EnKF solutions. Therefore, it can be concluded that, EnSRKF solution provides much better results compared to numerical and KF solutions and marginally better results compared to EnKF solution.

Results from case 1 shows that, EnKF and EnSRKF schemes perform very well to predict contaminant transport. In terms of RMSE calculation these two schemes have very similar results; although EnSRKF has slightly better RMSE compared to EnKF.

4.2 Case 2 with a Domain of 12,000 Nodes

To explore the computational efficiency of filtering techniques a larger domain problem is used. In case of large domain problems application of Kalman filter is infeasible as it calculates and stores large system and measurement covariance matrices explicitly. To demonstrate this infeasibility of Kalman filter, execution times of major functions have been recorded for different domain sizes. MATLAB profiler can estimate required time for each subroutine of a code. Using MATLAB profiler execution times are recorded. For a domain with total 8640 nodes with 36 nodes in x direction, 40 nodes in y direction and 6 nodes (layers) in z direction the execution time indicates that the entire simulation time is 2 hour and 21 minutes. Kalman filter itself occupied 2 hour and 8 minutes out of this 2 hour 21 minutes. Following Table 3 provides a brief summary of major functions which take significant time to execute. The table provides execution time of 4 different domain-size problems.

Table 3

Major operations that take much time to execute in four different domains

Time in seconds	Domain 1 10x12x4 Grids (480 nodes)	Domain 2 10x12x6 Grids (720 nodes)	Domain 3 20x24x6 Grids (2880 nodes)	Domain 4 36x40x6 Grids (8640 nodes)
Total execution time	6.95	11.86	208.21	8440.70
State Transition Matrix	1.3	3	46.24	415.07
Kalman filter	0.82	2.65	142.46	7932.44
EnKF	0.86	1.43	8.7	56.77
EnSRKF	1.02	1.59	6.6	23.20

Table 3 shows that, for smaller domain problems Kalman filter can run relatively quickly, however, as number of nodes increases computation time for Kalman filter increases enormously. Therefore, to concentrate properly on performances of EnSRKF and EnKF the Kalman filter algorithm is dropped for the case 2 problem. The case 2 has a new domain with 50 nodes in x axis, 60 nodes in y axis and 4 nodes (layers) in z axis (total 12,000 nodes). Parameter sets described for 2nd case of Table 1 is used to run this larger domain problem. The reason of separate parameters is that, as case 2 domain is 25 times larger than that of case 1 domain, keeping unchanged parameters for this larger domain produces contours with too small plumes. Therefore, few parameter values are modified. This model has been run for 40 time steps with duration of each time step being equal to 0.75 day. Therefore, total simulation time for this run is 30 days. In the previous model, total 30 time steps were used with same duration of each time step. The motivation behind increasing simulation time is to provide ample time to concentration plumes to develop properly. As this domain is relatively large, plume would look very small in earlier time steps as it would not be properly developed at earlier time steps. Therefore, contour profiles of time steps below 10 is not shown here. Figure 18 describes concentration profile of true, numerical, EnKF and EnSRKF after time step 10 at layer 1. Figure 19 provides concentration profile of true, numerical, EnKF and EnSRKF after time step 10 at layer 3. From Figure 18 and 19 it is not clearly visible which schemes have better prediction ability. Therefore, in Figure 20, 21, 22 and 23 concentration profiles are plotted after time step 30 at layer 1, time step 30 at layer 3, time step 40 at layer 1 and time step 40 at layer 3.

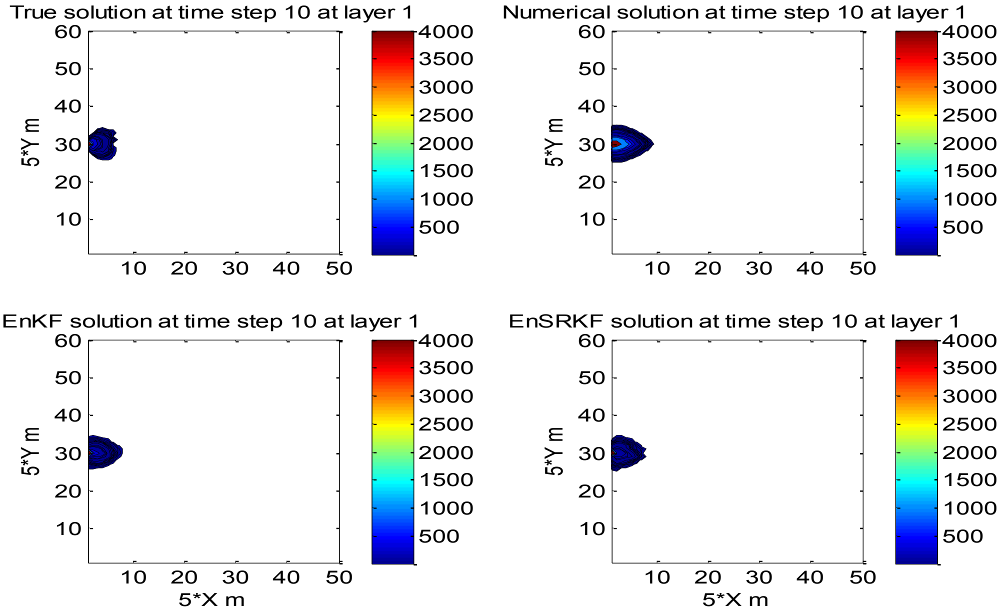

Figure 18. Contaminant concentration contour profile after time step 10 at layer 1 for True, Numerical, EnKF and EnSRKF solution

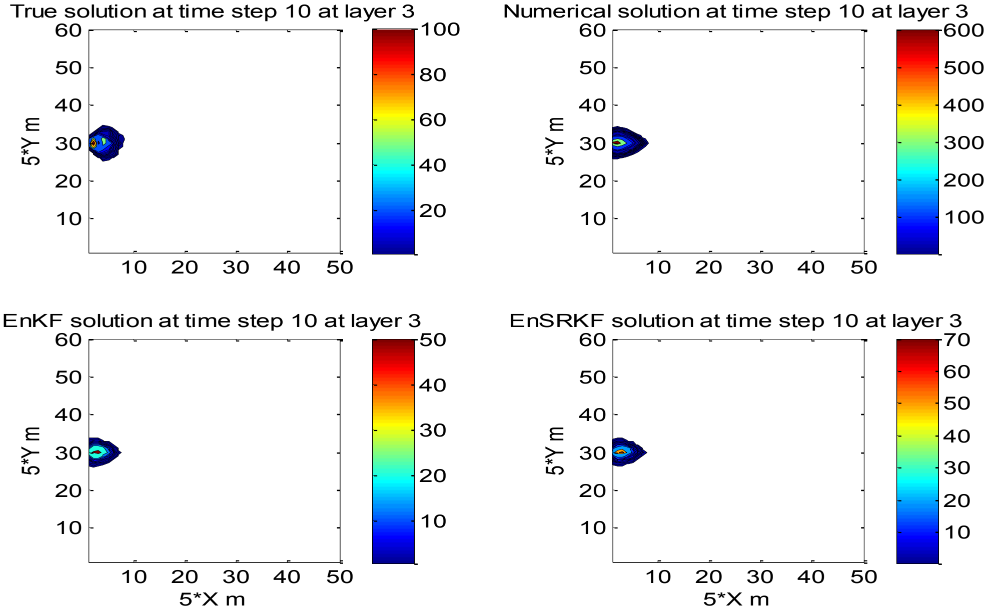

Figure 19. Contaminant concentration contour profile after time step 10 at layer 3 for True, Numerical, EnKF and EnSRKF solution

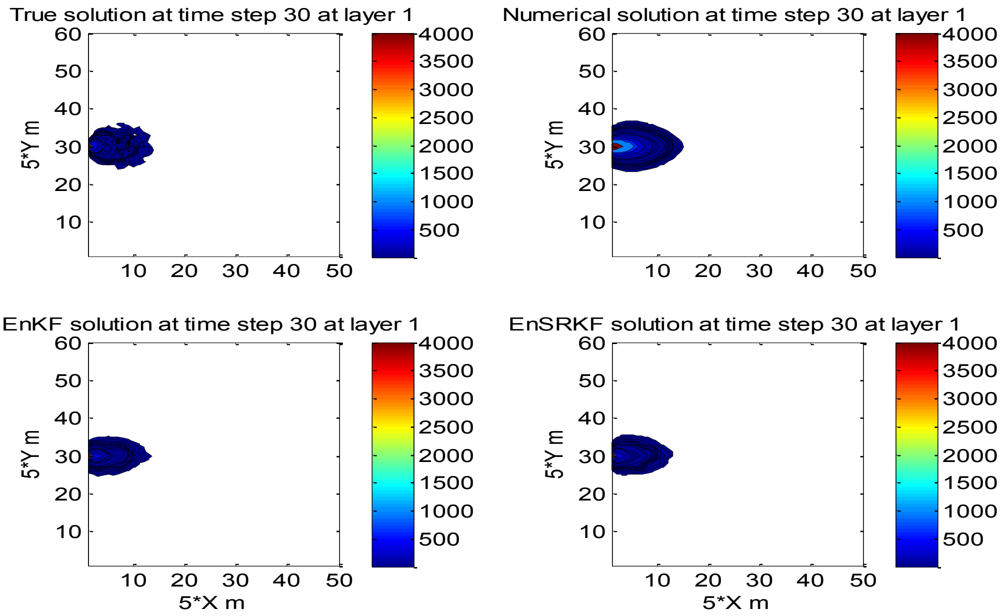

Figure 20. Contaminant concentration contour profile after time step 30 at layer 1 for True, Numerical, EnKF and EnSRKF solution

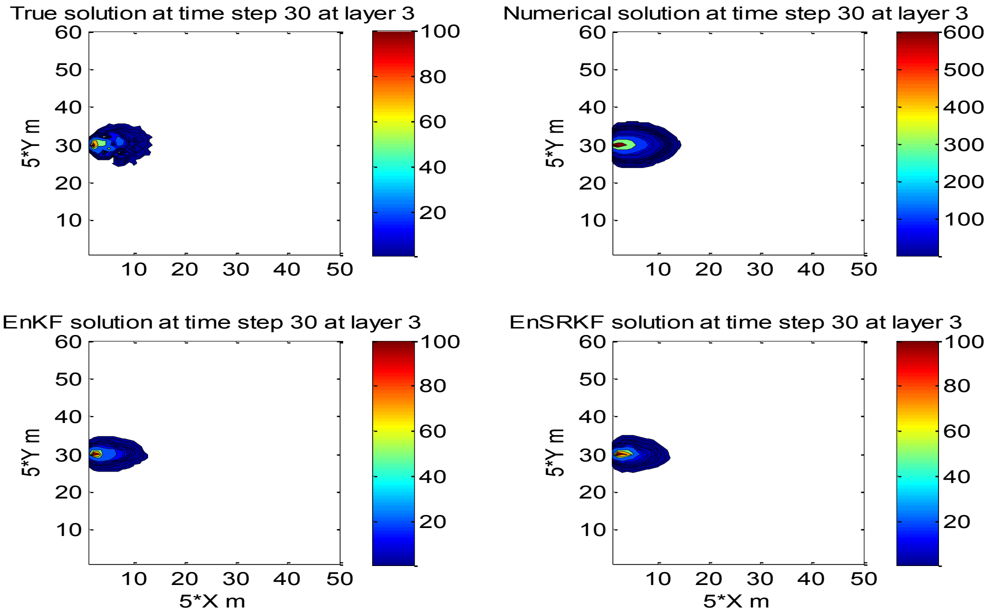

Figure 21. Contaminant concentration contour profile after time step 30 at layer 3 for True, Numerical, EnKF and EnSRKF solution

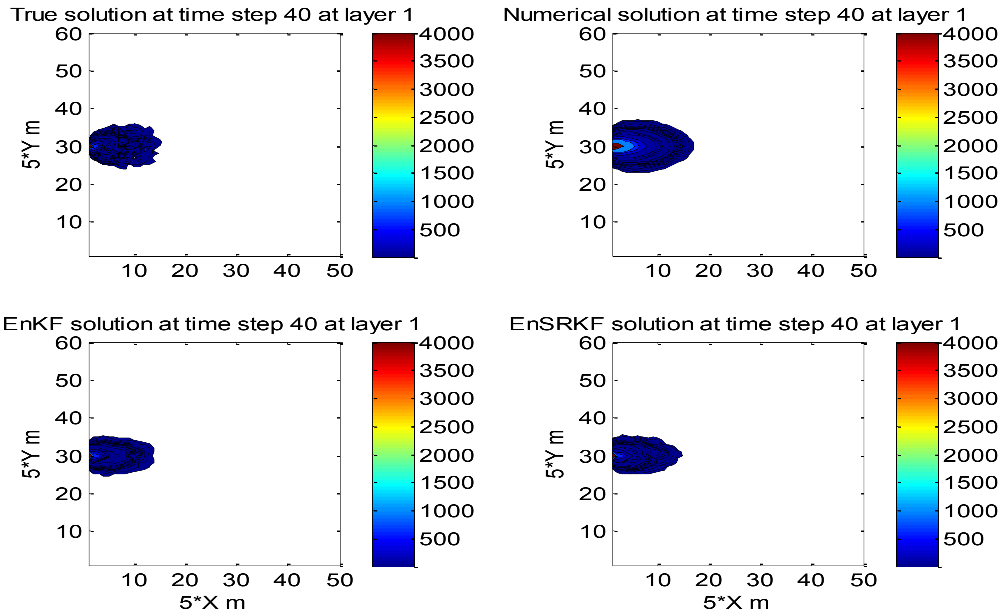

Figure 22. Contaminant concentration contour profile after time step 40 at layer 1 for True, Numerical, EnKF and EnSRKF solution

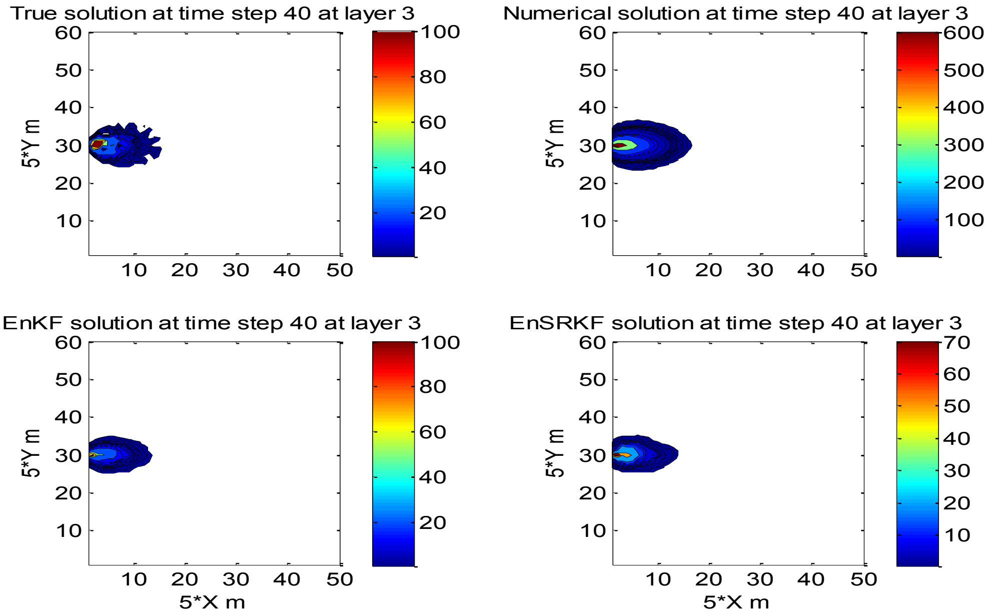

Figure 23. Contaminant concentration contour profile after time step 40 at layer 2 for True, Numerical, EnKF and EnSRKF solution

From Figure 20 and 21 shows that contaminant plume of true solution is better replicated by EnKF and EnSRKF compared to numerical solution. Solutions after time step 30 at layer 3 presented in Figure 21 shows that range of colorbar scales of true solution, EnKF and EnSRKF are same. Each of these has a range of 1 to 100 mg/L whereas numerical solution has a range of 1 to 600 mg/L. Figure 20 and Figure 22 shows contour profiles of layer 1 after time step 30 and 40 respectively. As seen in case 1, layer 1 shows highest range of concentration plume distribution due to the presence of a continuous source pollutant at layer 1. Figure 23 presents that, after the last time step of 40, at layer 3 plume of true solution has better similarity with that to EnKF and EnSRKF plumes. Also ranges of scales shown in colorbars substantiate that EnKF and EnSRKF can better predict concentration plume compared to deterministic numerical solution. To compare prediction accuracy of numerical, EnKF and EnSRKF solutions Figures 24, 25 and 26 are plotted for all layers after the end of the simulation period i.e. 40 time steps. Figure 24 shows comparison between true solution and numerical solution. Layer 2 and 3 shows significant difference in scales of colorbars between true and numerical solution. Figure 25 presents that, shape and scale range of true solution and those of EnKF solutions are more similar than these found in Figure 24 between true solution and numerical solution. Figure 26 presents contour profiles of true solution and EnSRKF solution after time step 40 at all layers. From Figure 25 and 26 it is clear that, EnKF and EnSRKF have completely same range of colorbar scales which indicates these two have similar prediction accuracy. These results are congruent with the results of case 1.

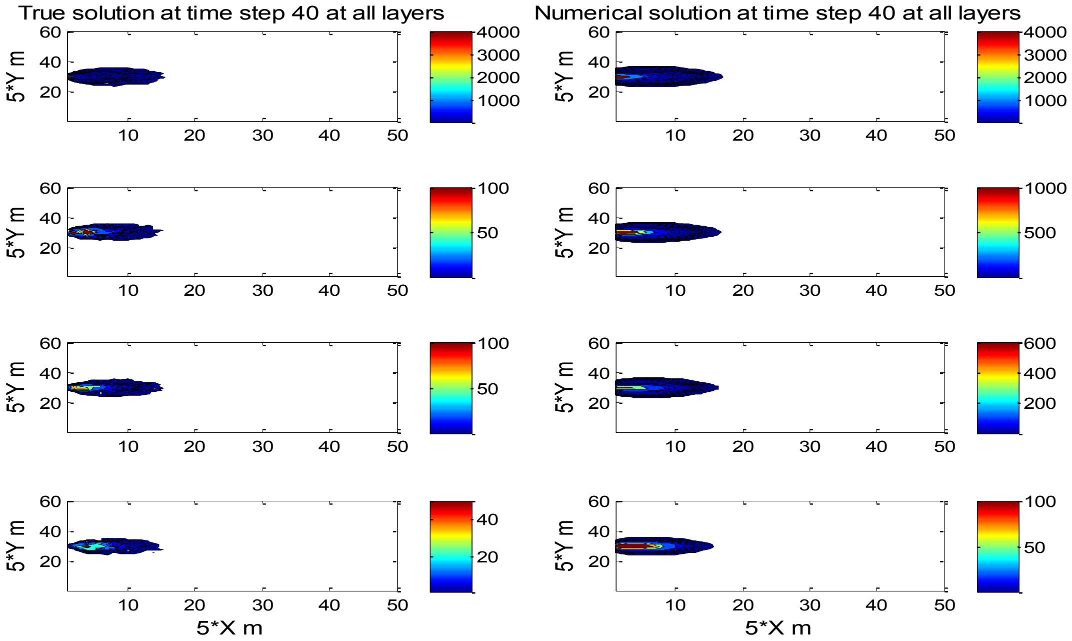

Figure 24. Comparison of true and numerical solutions after time step 40 at all layers

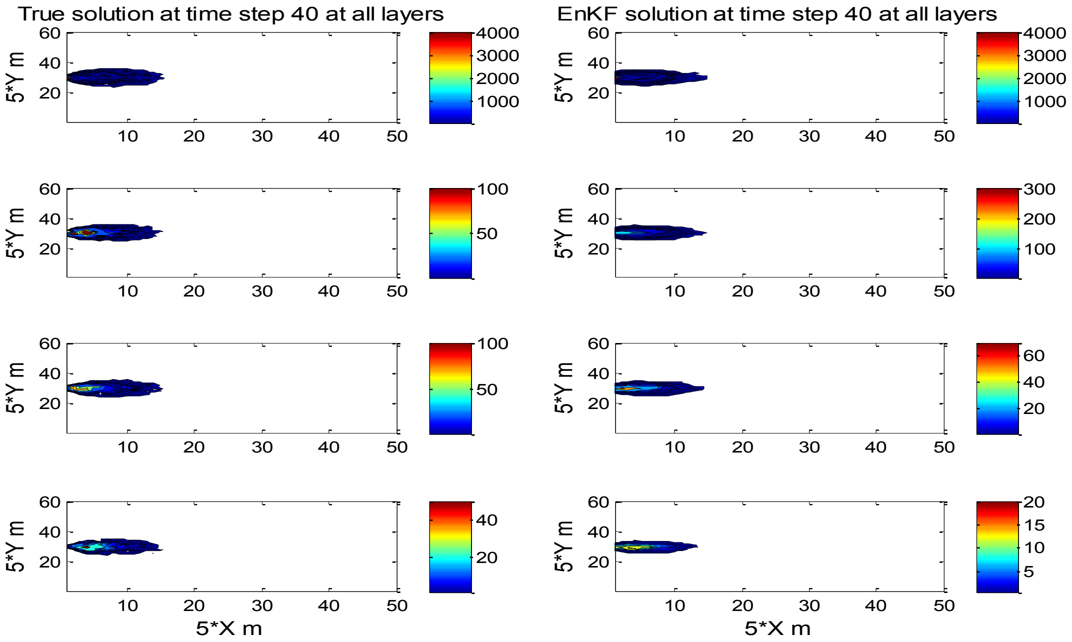

Figure 25. Comparison of true and EnKF solutions after time step 40 at all layers

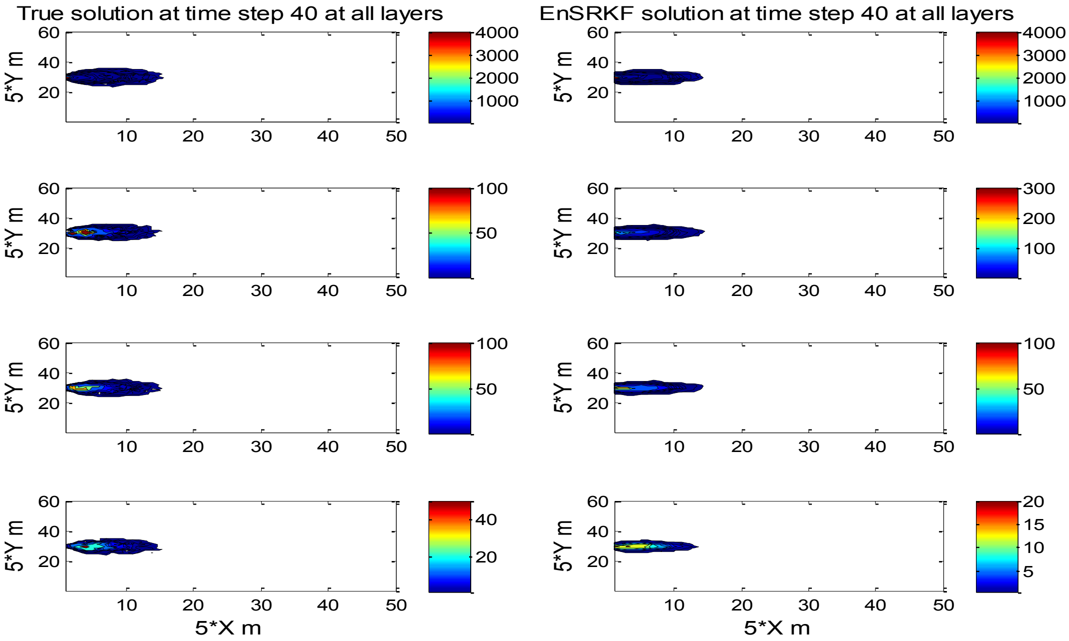

Figure 26. Comparison of true and EnSRKF solutions after time step 40 at all layers

Contour profiles provide a visual depiction of each scheme's effectiveness to predict distribution of contaminant plume. To check the accuracy of each scheme numerically the RMSE profile is plotted for numerical, EnKF and EnSRKF solutions. RMSE is calculated using true solution as a reference solution. Figure 27 provides RMSE profile of each scheme for entire simulation period of 40 days. RMSE profile Numerical solution continues to increase until around time step 15 and after which it stabilizes. EnKF and EnSRKF RMSE profile show early convergence and they provide a nearly similar profile as seen in smaller domain problem. Calculated average RMSE for numerical scheme is 109.05 mg/L, for EnKF scheme is 7.73 mg/L and that for EnSRKF is 7.69 mg/L. Therefore, both EnKF and EnSRKF shows significant improvement over numerical solution in prediction of contaminant plume in subsurface contaminant transport model.

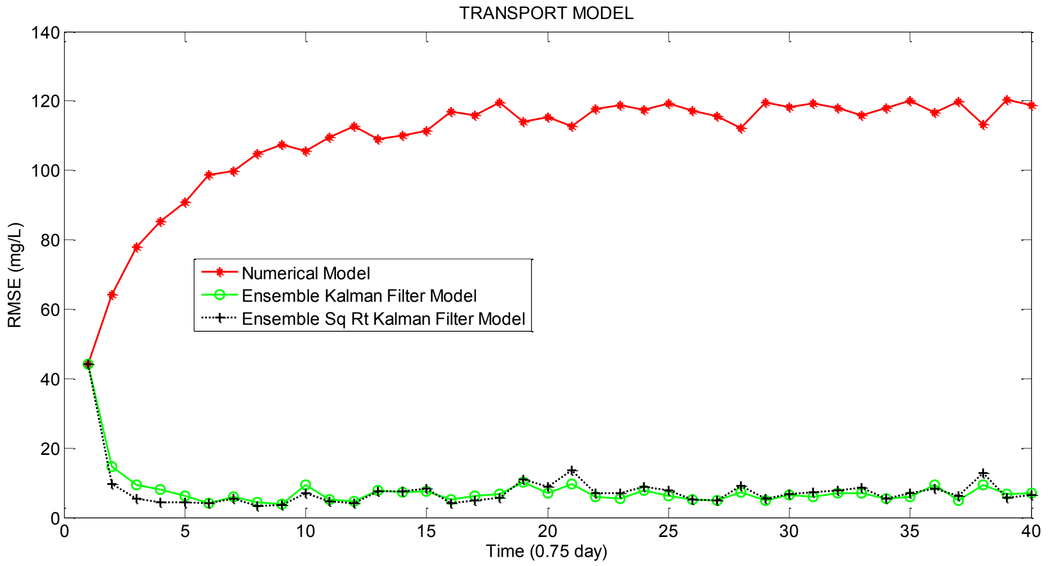

Figure 27. Root Mean Square Error (RMSE) profile of Numerical, EnKF and EnSRKF solutions.

To evaluate computational expense of EnKF and EnSRKF execution time is recorded for both schemes. On an average, for a 50x60x4 grid domain with 12,000 nodes EnKF operation took 167.43 seconds and EnSRKF operation took only 53.17 seconds to run. With similar prediction accuracy EnSRKF shows much greater computational efficiency compared to EnKF algorithm. For this larger domain problem EnSRKF takes 68% less computational time compared to EnKF scheme. Table 4 provides a summary of computational time in five different runs.

Table 4

Computational time of EnKF and EnSRKF schemes for 12,000 nodes problem

Scheme	Run 1	Run 2	Run 3	Run 4	Run 5	**Average**
EnKF run time (Seconds)	164.5	171.72	163.44	167.99	169.5	**167.43**
EnSRKF run time (Seconds)	52.97	53.76	53.01	50.29	55.8	**53.17**

4.3 Sensitivity Analysis with Change in Ensemble Sizes

To compare performances of EnKF and EnSRKF a sensitivity analysis is performed changing ensemble sizes. EnKF and EnSRKF both are Monte Carlo simulation techniques and use statistical ensembles or samples to calculate mean and covariances. Therefore, it is important to know how many ensembles are necessary to produce acceptable results. To compare prediction accuracy of EnKF and EnSRKF six different ensemble sizes of 10, 30, 50, 100, 200 and 400 have been used. In Figure 28, RMSE profiles of EnKF and EnSRKF schemes are plotted for these 6 different ensemble sizes.

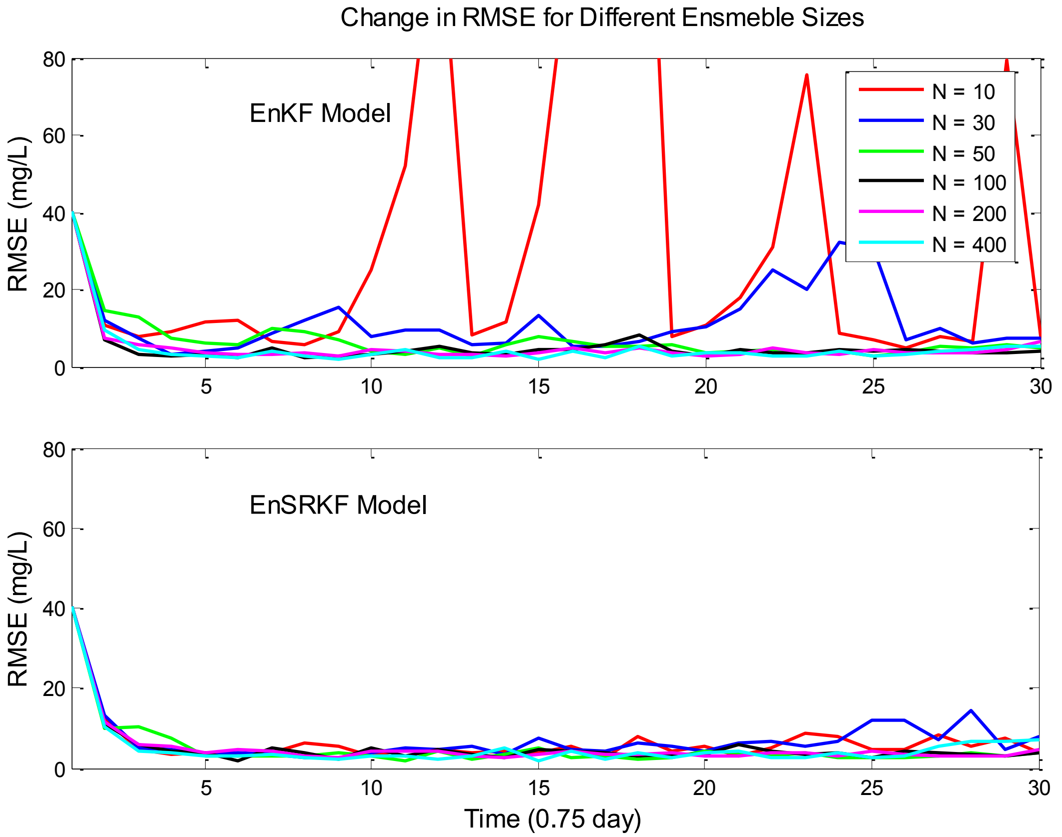

Figure 28. Change in RMSE profile with change in ensemble size. Vertical scale is customized to 0 to 80 mg/L to have a closer view of the RMSE profiles.

Figure 28 shows that for ensemble sizes of 10 and 30 EnKF has abrupt RMSE profile. As ensemble size increases EnKF profiles becomes smoother. On the other hand, for EnSRKF they exhibit much better profile even when ensemble size is 10 and 30. However, it also shows better trend as ensemble sizes increase. Table 5 shows comparison of EnKF and EnSRKF mean RMSE and execution time for different ensemble sizes.

Table 5

Mean RMSE and execution time of EnKF and EnSRKF for different ensemble sizes

Ensemble size, *N*	10	30	50	100	200	400
Mean EnKF RMSE (mg/L)	37.11	11.89	6.95	5.24	5.17	4.83
Mean EnSRKF RMSE (mg/L)	6.68	7.23	5.44	5.26	5.23	5.11
EnKF execution time (s)	0.21	0.33	0.46	0.82	1.55	3.26
EnSRKF execution time (s)	0.16	0.37	0.53	1.05	2.3	6.72

Table 5 shows that, for small ensemble sizes of 10 and 30, EnKF has relatively large mean RMSE values of 37.11 and 11.89 respectively. As ensemble size increases mean RMSE value for EnKF becomes lower gradually. On the other hand, for smaller ensemble sizes of 10 and 30 EnSRKF shows much better results compared to EnKF with mean RMSE values of 6.68 and 7.23 mg/L respectively. Table 5 also shows that as ensemble sizes increase both scheme shows gradual decrease in mean RMSE. This sensitivity analysis is performed for case 1 with 10x12x4 grids (480 nodes). Execution time for this problem shows that although for larger domains EnSRKF is computationally cheaper compared to EnKF, in case of increase of ensemble size for a given node size EnSRKF shows larger increase in execution time compared to EnKF. For example, when ensemble size is 400 EnSRKF schemes requires 6.72 seconds to execute whereas EnKF scheme takes 3.26 seconds to execute. However, as in most hydrological modeling cases very reasonable results are found using only statistically significant amount of ensemble sizes, the use of large ensemble sizes like 200 or 400 is not required in most real cases.

CHAPTER 5

Conclusion

In this study a three dimensional subsurface advection-dispersion-reaction contaminant transport model is developed to examine performances of data assimilation techniques. FTCS numerical solution with added noise provided the required system or process model and analytical solution with added noise provided the required measurement or observation model for the filtering approaches. Numerical solution cannot predict transport of contaminant properly due to some factors. Few of these are limited parameters to describe a complex process of contaminant transport in subsurface, heterogeneity of subsurface environment, truncation error to avoid higher order terms during Taylor's series expansion, etc. Data assimilation techniques do not have these limitations and with two different models (system and measurement) they can predict transport quite effectively. However, in this paper two separate cases are analyzed and demonstrated that for larger domains KF operation becomes too much expensive in terms of computational cost. EnKF and EnSRKF approaches use Monte Carlo simulation based ensemble data assimilation techniques and performs quite well to predict transport of contaminant in subsurface. With similar accuracy of EnKF, EnSRKF takes much less time compared to EnKF and it is found to be more suitable for subsurface contaminant transport prediction problem.

For the case 1problem a 10x12x4 grid (480 nodes) was used and found that mean error of EnSRKF scheme is 5.47 mg/L, that of EnKF scheme is 5.74 mg/L, that of KF scheme is 26.16 mg/L and that of numerical scheme is 127 mg/L. Therefore, on an average EnSRKF can improve prediction accuracy compared to numerical scheme by 95%, compared to KF scheme by 79% and compared to EnKF scheme by 4.70%. For case 1, execution time of EnKF and EnSRKF both are very small. EnKF took 0.86 seconds and EnSRKF took 1.02 seconds to execute. However, in real fields, state dimensions are very high and it is very important to analyze performance and efficiency of the model for problems with more nodes.

To compare computational performance of EnSRKF and EnKF case 2 was examined with 50x60x4 grid size with 12,000 nodes. Case 2 has 25 times more nodes than the case 1. KF approach was not included for case 2 as it takes astronomically high time to execute due to explicit calculation and storage of large covariance matrices. For example, for a domain with 36x40x6 grid size with 8640 nodes total execution time for the entire model is 8441 seconds and KF itself takes 7932 seconds i.e. around 94% of total execution time. For case 2 with 12,000 nodes EnKF and EnSRKF approaches exhibit similar accuracy. Mean RMSE for EnKF scheme is 7.73 mg/L and that for EnSRKF is 7.69 mg/L. However, EnSRKF shows significant improvement in average execution time i.e. for EnSRKF approach average execution time is 53.17 seconds and for EnKF average execution time is 167.43 seconds. Therefore, execution time of EnSRKF scheme is 68.2% less compared to EnKF scheme.

It can be concluded that, for both case 1 and case 2, EnSRKF shows marginal improvement in prediction accuracy compared to EnKF. However, as node numbers increase EnSRKF becomes more and more computationally efficient compared to EnKF.

To examine performances of EnKF and EnSRKF more closely, a sensitivity analysis is performed to examine change in RMSE with respect to change in ensemble size. In case of ensemble-based data assimilation techniques it is important to know performances of techniques with smaller ensemble size. It was found that for smaller ensemble sizes EnSRKF shows less error compared to EnKF. As ensemble size increases differences of errors between two schemes diminishes gradually.

References

Anderson, J. L. (2001). An ensemble adjustment filter for data assimilation. *Monthly Weather Review, 129*, 2884-2903.

Assumaning, G., & Chang, S. (2012). Use of Simulation Filters in Three-Dimensional Groundwater Contaminant Transport Modeling. *Journal of Environmental Engineering, 138*(11), 1122-1129. doi: 10.1061/(ASCE)EE.1943-7870.0000578

Bannister, R. N. (2012). A Square-Root Ensemble Kalman Filter Demonstration with the Lorenz model. Retrieved from http://www.met.reading.ac.uk/~darc/training/lorenz_ensrkf/.

Baptista, A. E. de M. (1987). Solution of Advection-Dominated Transport by Eulerian-Lagrangian Methods Using the Backward Method of Characterstics (Doctoral Dissertation). *MIT, Cambridge, MA*.

Bishop, Craig H., Etherton, Brian J., & Majumdar, Sharanya, J. (2001). Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Monthly Weather Review, 129*, 420-436.

Burgers, Gerrit, Leeuwen, P. J. V., & Evensen, Geir. (1998). Analysis Scheme in the Ensemble Kalman FIlter. *Monthly Weather Review, 126*, 1719-1724.

Chang, Shoou-Yuh, Chowhan, Tushar, & Latif, Sikdar. (2012). State and Parameter Estimation with an SIR Particle Filter in a Three-Dimensional Groundwater Pollutant Transport Model. *Journal of Environmental Engineering, 138*(11), 1114-1121. doi: 10.1061/(asce)ee.1943-7870.0000584

Chang, Shoou-Yuh, & Jin, An. (2005). Kalman Filtering with Regional Noise to Improve Accuracy of Contaminant Transport Models. *Journal of Environmental Engineering, 131*(6), 971-982. doi: 10.1061//asce/0733-9372/2005/131:6/971

Chang, Shoou-Yuh, & Latif, Sikdar M. I. (2009). Use of Kalman Filter and Particle Filter in a One Dimensional Leachate Transport Model. In E. Nzewi, G. Reddy, S. Luster-Teasley, V. Kabadi, S.-Y. Chang, K. Schimmel & G. Uzochukwu (Eds.), *Proceedings of the 2007 National Conference on Environmental Science and Technology* (pp. 157-163): Springer New York.

Chang, Shoou-Yuh, & Latif, Sikdar M. I. (2011). Ensemble Kalman Filter to Improve the Accuracy of a Three Dimensional Flow and Transport Model with a Continuous Pollutant Source *World Environmental and Water Resources Congress 2011* (pp. 1109-1117): American Society of Civil Engineers.

Chang, Shoou-Yuh, & Latif, Sikdar Muhammad Istiuq. (2010). Extended Kalman Filtering to Improve the Accuracy of a Subsurface Contaminant Transport Model. *Journal of Environmental Engineering, 136*(5), 466-474. doi: 10.1061//asce/ee.1943-7870.0000179

Chang, Shoou-Yuh, & Li, Xiaopeng. (2009). Organic Pollutant Transport in Groundwater Using Particle Filter. In E. Nzewi, G. Reddy, S. Luster-Teasley, V. Kabadi, S.-Y. Chang, K. Schimmel & G. Uzochukwu (Eds.), *Proceedings of the 2007 National Conference on Environmental Science and Technology* (pp. 165-171): Springer New York.

Chang, Shoou-Yuh., & Assumaning, G. (2011). Subsurface Radioactive Contaminant Transport Modeling Using Particle and Kalman Filter Schemes. *Journal of Environmental Engineering, 137*(4), 221-229. doi: doi:10.1061/(ASCE)EE.1943-7870.0000317

Chang, Shoou-Yuh., & Sayemuzzaman, M. (2014). Using Unscented Kalman Filter in Subsurface Contaminant Transport Models. *Journal of Environmental Informatics, 23*(1), 14-22. doi: 10.3808/jei.201400253

Chen, He, Yang, Dawen, Hong, Yang, Gourley, Jonathan J., & Zhang, Yu. (2013). Hydrological data assimilation with the Ensemble Square-Root-Filter: Use of streamflow observations to update model states for real-time flash flood forecasting. *Advances in Water Resources, 59*(0), 209-220. doi: http://dx.doi.org/10.1016/j.advwatres.2013.06.010

Cheng, X. (2000). Kalman Filter scheme for three-dimensional subsurface transport simulation with a continuous input. MS thesis, North Carolina A&T State University, Greensboro, NC.

Clark, Martyn P., Rupp, David E., Woods, Ross A., Zheng, Xiaogu, Ibbitt, Richard P., Slater, Andrew G., . . . Uddstrom, Michael J. (2008). Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources, 31*(10), 1309-1324. doi: 10.1016/j.advwatres.2008.06.005

Domenico, P. A. (1987). An analytical model for multidimensional transport of a decaying contaminant species. . *Journal of Hydrology, 91*(1-2), 49-58.

Evensen, Geir. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research, 99*(C5), 10,143-110,162.

Evensen, Geir. (1997). Advanced data assimilation for strongly nonlinear dynamics. *Monthly Weather Review, 125*(6), 1342.

Evensen, Geir. (2003). The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics, 53*(4), 343-367. doi: 10.1007/s10236-003-0036-9

Evensen, Geir, & Leeuwen, P. J. V. (1996). Assimilation of Geostat Altimeter Data for the Agulhas Current Using the Ensemble Kalman Filter with a Quasigeostrophic Model. *Monthly Weather Review, 124*(1), 85-96.

Houtekamer, P. L., & Mitchell, Herschel L. (1998). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review, 126*(3), 796.

Houtekamer, P. L., & Mitchell, Herschel L. (2001). A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation. *American Meteorological Society, 129*, 123-137.

Huang, Chunlin, Hu, Bill X., Li, Xin, & Ye, Ming. (2008). Using data assimilation method to calibrate a heterogeneous conductivity field and improve solute transport prediction with an unknown contamination source. *Stochastic Environmental Research and Risk Assessment, 23*(8), 1155-1167. doi: 10.1007/s00477-008-0289-4

Kenny, J. F., Barber, N. L., Hutson, S. S., Linsey, K. S., Lovelace, J. K., & Maupin, M. A. (2009). Estimated use of water in the United States in 2005: U. S. Geological Survey Circular 1344. 52 p.

Konikow, Leonard F. (2013). Groundwater Depletion in the United States (1900-2008). *U. S. Geological Survey Scientific Investigations Report 2013-5079*, 63 p. Retrieved from http://pubs.usgs.gov/sir/2013/5079.

Lermusiaux, P. F. J., & Robinson, A. R. (1999). Data assimilation via error subspace statistical estimation. Part I: Theory and schemes. *Monthly Weather Review, 127*, 1385-1407.

Mitchell, A. R. (1984). Recent developments in the finite element method. *Computational Techniques and Applications, CTAC-83*, 2-14.

Neuman, S. P. (1984). Adaptive Eulerian-Lagrangian finite element method for advection dispersion. *International Journal for Numerical Methods in Engineering, 20*, 321-337.

Ngan, P., & Russel, S. O. (1986). Example of flow forecasting with Kalman filter. *Journal of Hydrologic Engineering, 112*(9), 818-832.

Nolan, B. T., Hitt, K. J., & Ruddy, B. C. (1998). Probability of Nitrate Contaminant of Recently Recharged Ground Waters in the Conterminous United States. *Environmental Science and Technology, 36*(10), 2138-2145.

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

Reichle, Rolf H., McLaughlin, Dennis B., & Entekhabi, Dara. (2002). Hydrologic Data Assimilation with the Ensemble Kalman Filter. *Monthly Weather Review, 130*(1), 103-114.

Stednick, J. D., & Roig, L. C. (1989). Kalman filter calculation of sampling frequency when determine annual mean solution concentrations. *Water Resource Bulletin, 25*(3), 672-682.

Tippett, Michael, K., Anderson, J. L., Bishop, Craig H., Hamill, Thomas M., & Whitaker, Jeffrey S. (2003). Ensemble Square Root Filters. *Monthly Weather Review, 131*, 1485-1490.

USEPA. (1993). Wellhead Protection: A Guide for Small Communities. (EPA Number: 625R93002), 156p. Retrieved from http://www.epa.gov/region151/students/pdfs/gwc151.pdf.

Welch, Greg, & Bishop, Gary. (2006). An Introduction to the Kalman Filter. Retrieved from http://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf.

Whitaker, Jeffrey S., & Hamill, Thomas M. (2002). Ensemble Data Assimilation without Perturbed Observations. *Monthly Weather Review, 130*(7), 1913-1924.

Yu, Y. S., Heidari, M., & Guang-Te, W. (1989). Optimal estimation of contaminant transport in ground water. *Water Resource Bulletin, 25*(2), 295-300.

Zheng, C., & Bennett, G. D. (2002). *Applied Contaminant Transport Modeling, Second Edition.*, New York, NY: John Wiley & Sons.

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