# Low-Rank Kalman Filtering in Subsurface Contaminant Transport Models

Thesis by

**Mohamad El Gharamti, B.Sc.**

Submitted in Partial Fulfillment of the Requirements for the degree of

**Masters of Science**

King Abdullah University of Science and Technology

Division of Physical Sciences and Engineering

Earth Sciences and Engineering Program

Thuwal, Makkah Province, Kingdom of Saudi Arabia

December, 2010

The undersigned approve the thesis of Mohamad El Gharamti

| | | |
|---|---|---|
| Chairperson, Georgiy Stenchikov | Signature | Date |
| Thesis supervisor, Ibrahim Hoteit | Signature | Date |
| Thesis supervisor, Shuyu Sun | Signature | Date |

King Abdullah University of Science and Technology

2010

# ABSTRACT


## Low-Rank Kalman Filtering in Subsurface Contaminant Transport Models

Mohamad El Gharamti

Understanding the geology and the hydrology of the subsurface is important to model the fluid flow and the behavior of the contaminant. It is essential to have an accurate knowledge of the movement of the contaminants in the porous media in order to track them and later extract them from the aquifer. A two-dimensional flow model is studied and then applied on a linear contaminant transport model in the same porous medium. Because of possible different sources of uncertainties, the deterministic model by itself cannot give exact estimations for the future contaminant state. Incorporating observations in the model can guide it to the true state. This is usually done using the Kalman filter (KF) when the system is linear and the extended Kalman filter (EKF) when the system is nonlinear. To overcome the high computational cost required by the KF, we use the singular evolutive Kalman filter (SEKF) and the singular evolutive extended Kalman filter (SEEKF) approximations of the KF operating with low-rank covariance matrices. The SEKF can be implemented on large dimensional contaminant problems while the usage of the KF is not possible. Experimental results show that with perfect and imperfect models, the low rank filters can provide as much accurate estimates as the full KF but at much less computational cost. Localization can help the filter analysis as long as there are enough neighborhood data to the point being analyzed. Estimating the permeabilities of the aquifer is successfully tackled using both the EKF and the SEEKF.

# Dedication

To my mom *"Jihad Awji"* who really cares about me with all nice meanings of sympathy and kindness.

To my dad *"Ali Gharamti"* who always supports me and never doubted that I could make this successful scientific work.

To my lovely future wife *"Berenice Garcia Tellez"* who was there inside my heart giving me hope, courage, and tenderness.

# Acknowledgements

A number of people have made this paper possible. In particular I wish to thank my supervisor *"Prof. Ibrahim Hoteit"* who taught me the basic concepts of filtering and estimations. I am very grateful for his tremendous cooperation and his assistance in reviewing this thesis work.

I would like to thank my second supervisor *"Prof. Shuyu Sun"* who helped me understand the physics and the dynamics of subsurface contaminant models. I would like to thank him as well for the nice internship he found for me in the University of Texas at Austin where I started this work.

My special thanks to *"Prof. Marry Wheeler"* for offering me the usage of her subsurface laboratory in UT Austin in summer 2010.

The quality of this work was also checked by *"Prof. Georgiy Stenchikov" (Chairperson of ErSE Department at KAUST).* Thanks to his gracious assistance and his interseting discussions.

I want to express my deep thanks to "*Dr. Jisheng Kou" (Postdoc, ErSE Department at KAUST)* who supported me with ideas and solution to some problems in my flow and contaminant models at the beginning of the work.

Finally, I am most indebted to my brother *"Mostafa Gharamti" (PhD student studying Mathematical Physics at the University of Edinburgh)* who assisted me in all aspects during this thesis. Thanks to him for teaching me how to write in LaTeX. Thanks to him for explaining me how to use MATLAB 2 years ago. Thanks to him for giving me a final review of this thesis. And thanks for his warm nice supporting words.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Groundwater contamination problem arises when the water in the subsurface becomes polluted by substances of human origin. Since groundwater is one of the most safe sources for people to drink; drinking contaminated water can become a serious issue because, it is very dangerous to consume. Several sources can cause contamination in the groundwater such as chemicals, road salt, bacteria, viruses, medications, fertilizers, and fuel. Groundwater contamination can also occur when factories dump thousands of toxic materials into surrounding waterways, and when polluted runoff from storm drains reaches the aquifer [16]. At the time the aquifer becomes contaminated, it is very difficult to clean up. In some cases, the water can be cleaned using filtration systems, but in other cases, it may be rendered useless.

Subsurface contaminant transport models are very efficient for groundwater quality assessment and risk evaluation [16]. A subsurface contaminant transport model can provide important information about the evolving of the contaminant inside the subsurface geologic system and it can give some estimations and predictions about the future subsurface situation after the migration of the contaminant [15, 16, 41, 42, 43, 44, 45].

Modeling of subsurface dynamic systems is carried out through two main stages starting with the flow modeling and then the contaminant transport model [15, 43, 44]. Unlike the contaminant which evolves with time, the flow model is stationary [42, 43]. The flow model is indispensable to infer correct knowledge of the water heads and the Darcy velocities in the system. The contaminant transport model can be efficiently handled by using the traditional procedure of the state-space approach on the discrete-time formulation. In most

realistic models, uncertainties in the model parameters and configurations are unavoidable and they can mislead the model from the correct trajectoris. Collecting data and using it as a support for the model can help guide the model estimation to much accurate solutions. Such procedure, where the model and the observations are used together, is well known as filtering.

Recursive filtering, such as Kalman filter (KF), can process the received data sequentially rather than dealing with it as a single batch so that it is not necessary to store the complete data set or to reprocess existing data if a new observation becomes available [9, 7].

Basically, the KF is used to provide estimation for linear problems and EKF is used for moderately nonlinear problems where the error distribution is Guassian. On the other hand, EnKF, based on Monte Carlo sampling and KF processes, is used for highly nonlinear problems. All these filters are considered to be good tools for prediction and estimation in dynamic systems as long as they can handle the dimensions of the problem. For instance, the KF is efficient for small dimensions and as the dimension of the problem increases, carrying out the KF will be computationaly more expensive. In some cases where the dimensions become very large, the KF fails to predict and give estimations for the states of the system because of the huge computational cost it requires. To overcome this problem; a singular evolutive Kalman filter, proposed by Pham et al. (1997) [34], will be considered. This SEKF is a new filtering technique for subsurface contaminant models; it is mainly based on a low rank approximation for the full KF and it is well guaranteed to decrease the high computational cost needed by the KF and to give reliable estimation results for high dimension problems.

Recently, KF and extended Kalman filter (EKF) have been applied in surface and subsurface hydrologic systems and water quality modeling. We can refer, for example, to [4, 14, 23, 17, 31, 36, 50, 53, 56]. Apart from modeling, KF has been used also in several areas for water resources [1, 3, 20, 35, 39, 40, 52]. Cheng (2000) applied discrete KF in a three-dimensional subsurface contaminant transport model for a continuous input [13]. Chang and Jin (2005) used KF with regional noises in a two-dimensional subsurface con-

taminant transport model for a pulse input [10]. Chang and Latif (2007) used KF and particle filter in a one-dimensional leachate transport in subsurface [8]. Chang and Latif (2010) used EKF in a two-dimensional contaminant transport model with a pulse input [9].

Several examples of Kalman filtering and extended Kalman filtering were applied in environmental and ecological modeling, analysis, and prediction studies. Pastres et al. (2003) applied EKF to the analysis of high frequency field measurements of dissolved oxygen, water temperature, and salinity collected by multi-parametric sensors in the lagoon of Venice, Italy [33]. Neal et al. (2007) examined the application of a river-flow forecasting approach based on a one-dimensional hydraulic flow simulation model updated using real-time data within an EnKF framework [32]. Goegebeur and Pauwels (2007) compared the performance of the parameter estimation method within the EKF for the estimation of hydrologic model parameters [21]. Franssen et al. (2008) used ensemble Kalman filtering (EnKF) to assimilate hydraulic head data from 90 locations during two years of groundwater flow modeling [18].

The objectives of this thesis are to prove the efficiency of data assimilation in providing more accurate estimates of the contamination state in the system, to test the functioning of the KF in the dynamic model by comparing the estimated states with the true model states, and to reduce the expensive computational cost needed by KF through the usage of low-rank filtering techniques. A low resolution grid is used to implement the full KF and the low-rank KF, then when the low-rank KF is validated it can be then applied on a more realistic high resolution grid. Further and by using contaminant data, state parameter estimation problem will be studied using EKF and singular evolutive extended Kalman filter (SEEKF). The prediction accuracy of the EKF and the SEEKF will be tested on a permeability estimation system using a nonlinear contaminant model.

This thesis is organized in 6 chapters. In chapter 2, the subsurface model is presented, all mathematical derivations are well explained for both the flow and the contaminant model. All filtring tools and estimation techniques are given in details in chapter 3. The numerical experiments and the results for some interesting problems are presented in chapter 4. In

chpater 5, the joint state-parameter estimation problem is tackled. Further discussions and conclusions are found in the last chapter.

# Chapter 2

# Subsurface Model and Discretization

The dynamic flow model is represented by a 2-D uniform saturated flow field that is composed of two rock types with different permeabilities where one is embedded in the other (Figure 2-1). $K_1$ and $K_2$ are the hydraulic conductivities of the two rocks, $a_0$, $b_0$, $a_1$, and $b_1$ indicate the position of the low permeability rock with respect to the main high permeability rock.

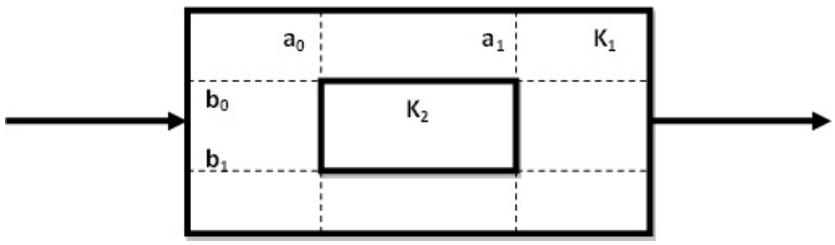


Figure 2-1: 2D saturated flow field with the 2 major rocks having different permeabilities (The small rock with $k_2$ is located at the center of the large rock).

## 2.1 Flow Model

The behavior of the water inside the medium is studied by looking at the water head distribution and the Darcy velocities. To do so, both the Darcy equation and the continuity

equation are considered.

- Darcy Equation

$$U = -\frac{k}{\mu}\left(\nabla P + \rho g \nabla z\right), \tag{2.1}$$

where $U$ is the Darcy velocity, $k$ is the permeability of the porous medium, $\mu$ is the viscosity of water, $P$ is the pressure head, $\rho$ is the density of water, $g$ is the gravitational acceleration, and $z$ is the vertical coordinate.

The flow in the model is taking place in the 2 dimensional space so the vertical coordinate $z$ in (2.1) is ignored. The permeability term $k$ depends only on the porous medium (soil, rock, ...). Another more general term depending on both the fluid and the porous medium, known as hydraulic conductivity (as in Figure 2-1), is introduced in (2.1). Then, the Darcy equation becomes

$$U = -K\nabla h, \tag{2.2}$$

where $K$ is the hydraulic conductivity and $h$ is the water head given by

$$h = \frac{P}{\rho g}. \tag{2.3}$$

- Continuity Equation

$$\frac{\partial\left(\phi\rho\right)}{\partial t} = -\nabla\cdot\left(\rho U\right) + \widetilde{q}, \tag{2.4}$$

where $\phi$ is the porosity of the medium, $t$ is the time, and $\widetilde{q}$ is the source term.

Discretization of the PDEs in (2.2) and (2.1) is done using the Cell Centred Finite Differences (CCFD) approach . CCFD is also known as Block Centered Finite Difference Method, it is based on mass conservation concept; i.e. the net fluid flowing out from a cell is equal to the net injection of fluid into the same cell. The governing equations can be written as

$$U_x = -K_{xx}\frac{\partial h}{\partial x}, \tag{2.5}$$

$$U_y = -K_{yy}\frac{\partial h}{\partial y}, \tag{2.6}$$

$$\frac{\partial U_x}{\partial x} + \frac{\partial U_y}{\partial y} = q, \tag{2.7}$$

$$h = h_B \text{ on } \delta D, \tag{2.8}$$

$$U \cdot n = U_B \text{ on } \delta N, \tag{2.9}$$

where $U_x$ and $U_y$ correspond to the Darcy velocities in $x$ and $y$ directions respectively, $K_{xx}$ and $K_{yy}$ correspond to the hydraulic conductivities in $x$ and $y$ directions respectively, $q$ is the source term, $h_B$ is water head at the boundaries, $U_B$ is the Darcy velocity at the boundaries, and finally $\delta D$ and $\delta N$ represent the Dirichlet and Neumann boundary conditions respectively.

Considering a small rectangular cell inside the porous domain, the net fluid flowing out from the cell will be

$$U^x_{i+1,j+\frac{1}{2}}\left(y_{j+1}-y_j\right)+U^y_{i+\frac{1}{2},j+1}\left(x_{i+1}-x_i\right)-U^x_{i,j+\frac{1}{2}}\left(y_{j+1}-y_j\right)-U^y_{i+\frac{1}{2},j+1}\left(x_{i+1}-x_i\right), \tag{2.10}$$

and the net injection of the fluid into the cell is

$$q\left(x_{i+\frac{1}{2}},y_{j+\frac{1}{2}}\right)\left(x_{i+1}-x_i\right)\left(y_{j+1}-y_j\right)\approx \int_{y_j}^{y_{j+1}}\int_{x_i}^{x_{i+1}} q\left(x,y\right)dxdy, \tag{2.11}$$

where $U^x_{i+1,j+\frac{1}{2}}, U^x_{i,j+\frac{1}{2}},, U^y_{i+\frac{1}{2},j+1}$, and $U^y_{i+\frac{1}{2},j}$ are the Darcy velocities components on the right, left, top, and bottom edges of the cell respectively., and $q\left(x_{i+\frac{1}{2}},y_{j+\frac{1}{2}}\right)$ is the source term at the center of the cell.

The Darcy velocities are written in terms of water heads as follows

$$U^x_{i+1,j+\frac{1}{2}} = -K_{xx}\left(x_{i+1},y_{j+\frac{1}{2}}\right)\frac{h_{i+\frac{3}{2},j+\frac{1}{2}}-h_{i+\frac{1}{2},j+\frac{1}{2}}}{x_{i+\frac{3}{2}}-x_{i+\frac{1}{2}}}, \tag{2.12}$$

$$U^x_{i,j+\frac{1}{2}} = -K_{xx}\left(x_{i},y_{j+\frac{1}{2}}\right)\frac{h_{i+\frac{1}{2},j+\frac{1}{2}}-h_{i-\frac{1}{2},j+\frac{1}{2}}}{x_{i+\frac{1}{2}}-x_{i-\frac{1}{2}}}, \tag{2.13}$$

$$U^y_{i+\frac{1}{2},j+1} = -K_{yy}\left(x_{i+\frac{1}{2}},y_{j+1}\right)\frac{h_{i+\frac{1}{2},j+\frac{3}{2}}-h_{i+\frac{1}{2},j+\frac{1}{2}}}{y_{j+\frac{3}{2}}-y_{j+\frac{1}{2}}}, \tag{2.14}$$

$$U^y_{i+\frac{1}{2},j} = -K_{yy}\left(x_{i+\frac{1}{2}},y_{j}\right)\frac{h_{i+\frac{1}{2},j+\frac{1}{2}}-h_{i+\frac{1}{2},j-\frac{1}{2}}}{y_{j+\frac{1}{2}}-y_{j-\frac{1}{2}}}. \tag{2.15}$$

Equating both (2.10) and (2.11) gives

$$\frac{U^x_{i+1,j+\frac{1}{2}}-U^x_{i,j+\frac{1}{2}}}{x_{i+1}-x_i}+\frac{U^y_{i+\frac{1}{2},j+1}-U^y_{i+\frac{1}{2},j}}{y_{j+1}-y_j}=q\left(x_{i+\frac{1}{2}},y_{j+\frac{1}{2}}\right), \tag{2.16}$$

where $i$ runs from 0 to $m-1$ and $j$ from 0 to $n-1$.

Plugging equations (2.12), (2.13), (2.14), and (2.15) in (2.16) will give the general discretized flow equation

$$\frac{\left[-K_{xx}\left(x_{i+1},y_{j+\frac{1}{2}}\right)\frac{h_{i+\frac{3}{2},j+\frac{1}{2}}-h_{i+\frac{1}{2},j+\frac{1}{2}}}{x_{i+\frac{3}{2}}-x_{i+\frac{1}{2}}}+K_{xx}\left(x_i,y_{j+\frac{1}{2}}\right)\frac{h_{i+\frac{1}{2},j+\frac{1}{2}}-h_{i-\frac{1}{2},j+\frac{1}{2}}}{x_{i+\frac{1}{2}}-x_{i-\frac{1}{2}}}\right]}{(x_{i+1}-x_i)}$$

$$+$$

$$\frac{\left[-K_{yy}\left(x_{i+\frac{1}{2}},y_{j+1}\right)\frac{h_{i+\frac{1}{2},j+\frac{3}{2}}-h_{i+\frac{1}{2},j+\frac{1}{2}}}{y_{j+\frac{3}{2}}-y_{j+\frac{1}{2}}}+K_{yy}\left(x_{i+\frac{1}{2}},y_j\right)\frac{h_{i+\frac{1}{2},j+\frac{1}{2}}-h_{i+\frac{1}{2},j-\frac{1}{2}}}{y_{j+\frac{1}{2}}-y_{j-\frac{1}{2}}}\right]}{(y_{j+1}-y_j)}$$

$$=q\left(x_{i+\frac{1}{2}},y_{j+\frac{1}{2}}\right), \tag{2.17}$$

here $i$ runs from 1 to $m-1$ and $j$ from 1 to $n-1$.

Closing the system of equations, Dirichlet boundary conditions (constant water heads) are imposed at the eastern and western boundaries. Top and bottom boundaries are impermeable (i.e. the Darcy velocities across these boundaries vanish).

## 2.2 Contaminant Transport Model

The spatial distribution of the water heads and the Darcy velocities obtained from the flow model are used to solve for the contaminant's concentration in the following transport equation

$$\frac{\partial(\phi C)}{\partial t}+\nabla\cdot(UC-D(U)\nabla C)=r(C)+qC^*, \tag{2.18}$$

where $C$ is the concentration of the contaminant commonly referred as the amount of species in a unit volume of water, $D$ is the dispersion/diffusion term, $r$ is the reaction/adsorption term, and $C^*$ is the upwind concentration.

Solving for the contaminant's concentration in (2.18), the upwind scheme of the CCFD is used in order to get a stable solution with no oscillations. The upwind scheme emphasizes the idea that the concentration at the center of the cell is affected by the concentration of the upwind cells around it. If the velocity of water is pointing from left to right, the concentration has to get information from the cells left, up, and bottom of it and vise versa.

Since this transport equation is time dependent, an additional initial condition has to be imposed together with the boundary conditions. This initial condition will represent the spatial contaminant spread inside the aquifer at time zero.

Just like the flow model, the mass conservation idea is applied here but this time the fluid flow is replaced by a mass flow. In other words, the net mass inflow rate has to be equal to the mass accumulation. Considering a small cubic volume

- The mass inflow of the species of interest
  - across the surface $x - \frac{\Delta x}{2}$ is

$$(U_x C)_{x-\frac{\Delta x}{2},y,z}\, \Delta y \Delta z, \tag{2.19}$$

  - across the surface $y - \frac{\Delta y}{2}$ is

$$(U_y C)_{x,y-\frac{\Delta y}{2},z}\, \Delta x \Delta z, \tag{2.20}$$

  - across the surface $z - \frac{\Delta z}{2}$ is

$$(U_z C)_{x,y,z-\frac{\Delta z}{2}}\, \Delta x \Delta y. \tag{2.21}$$

- The mass outflow of the species of interest
  - across the surface $x + \frac{\Delta x}{2}$ is

$$(U_x C)_{x+\frac{\Delta x}{2},y,z}\, \Delta y \Delta z, \tag{2.22}$$

  - across the surface $y + \frac{\Delta y}{2}$ is

$$(U_y C)_{x,y+\frac{\Delta y}{2},z}\, \Delta x \Delta z. \tag{2.23}$$

  - across the surface $z + \frac{\Delta z}{2}$ is

$$(U_z C)_{x,y,z+\frac{\Delta z}{2}}\, \Delta x \Delta y, \tag{2.24}$$

Mass accumulation term in $\Delta x \Delta y \Delta z$ is

$$\frac{\partial}{\partial t}\left(\Delta x \Delta y \Delta z \phi C\right). \tag{2.25}$$

where $\Delta x$, $\Delta y$, and $\Delta z$ represent the length of the cell in $x$, $y$, and $z$ directions respectively.

Equating the mass accumulation in (2.25) to the net mass inflow rate, we will get

$$\frac{\partial}{\partial t}\left(\phi C\right) = -\frac{\partial}{\partial x}\left(U_x C\right) - \frac{\partial}{\partial y}\left(U_y C\right) - \frac{\partial}{\partial z}\left(U_z C\right). \tag{2.26}$$

Discretization is then applied on this equation (2.26) considering the upwind values for C. So, the final equation will take the following form

$$\begin{aligned}
&\frac{d}{dt}\left[\phi\left(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}\right) C_{i+\frac{1}{2},j+\frac{1}{2}}\right]\left(x_{i+1}-x_i\right)\left(y_{j+1}-y_j\right) \\
&= \left[\max\left(U^x_{i,j+\frac{1}{2}}, 0\right) C_{i-\frac{1}{2},j+\frac{1}{2}} + \min\left(U^x_{i,j+\frac{1}{2}}, 0\right) C_{i+\frac{1}{2},j+\frac{1}{2}}\right]\left(y_{j+1}-y_j\right) \\
&- \left[\max\left(U^x_{i+1,j+\frac{1}{2}}, 0\right) C_{i+\frac{1}{2},j+\frac{1}{2}} + \min\left(U^x_{i+1,j+\frac{1}{2}}, 0\right) C_{i+\frac{3}{2},j+\frac{1}{2}}\right]\left(y_{j+1}-y_j\right) \\
&+ \left[\max\left(U^y_{i+\frac{1}{2},j}, 0\right) C_{i+\frac{1}{2},j-\frac{1}{2}} + \min\left(U^y_{i+\frac{1}{2},j}, 0\right) C_{i+\frac{1}{2},j+\frac{1}{2}}\right]\left(x_{i+1}-x_i\right) \\
&- \left[\max\left(U^y_{i+\frac{1}{2},j+1}, 0\right) C_{i+\frac{1}{2},j+\frac{1}{2}} + \min\left(U^y_{i+\frac{1}{2},j+1}, 0\right) C_{i+\frac{1}{2},j+\frac{3}{2}}\right]\left(x_{i+1}-x_i\right).
\end{aligned} \tag{2.27}$$

Attention has to be given for the time step taken in each iteration because it plays an essential role in the stability of the system according to the CFL condition

$$\left(\frac{U_x \Delta t}{\phi \Delta x} + \frac{U_y \Delta t}{\phi \Delta y}\right) < c, \tag{2.28}$$

where $c$ is constant for the *CFL* condition (Courant-Friedrichs-Levy condition). We note that small time steps will insure stability.

The transport equation (2.27) can be simplified more in the form

$$N\frac{dC}{dt} + BC = b, \tag{2.29}$$

where $N$ and $B$ are defined as follows

$$\begin{aligned}
N &= diag\left(\phi S\right), \\
B &= \left(B^W - B^E\right)\left(diag\left(U_x^+\right) B^W + diag\left(U_x^-\right) B^E\right) \\
&+ \left(B^S - B^N\right)\left(diag\left(U_y^+\right) B^S + diag\left(U_y^-\right) B^N\right),
\end{aligned}$$

and $b$ and $S$ are the source term vector and the area of each cell respectively.

The new terms and matrices in $B$ are

$$U_x^+ = \max\left(U_x.h_y, 0\right), \qquad U_x^- = \min\left(U_x.h_y, 0\right),$$
$$U_y^+ = \max\left(U_y.h_x, 0\right), \qquad U_y^- = \min\left(U_y.h_x, 0\right),$$

and

$$B^W = \begin{bmatrix} \begin{pmatrix} O_{1\times m} \\ I_{m\times m} \end{pmatrix}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \begin{pmatrix} O_{1\times m} \\ I_{m\times m} \end{pmatrix}_n \end{bmatrix},$$

$$B^E = \begin{bmatrix} \begin{pmatrix} I_{n\times m} \\ O_{1\times m} \end{pmatrix}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \begin{pmatrix} I_{n\times m} \\ O_{1\times m} \end{pmatrix}_n \end{bmatrix},$$

$$B^S = \begin{bmatrix} O_{m\times(mn)} \\ I_{(mn)\times(mn)} \end{bmatrix},$$

$$B^N = \begin{bmatrix} I_{(mn)\times(mn)} \\ O_{m\times(mn)} \end{bmatrix},$$

where $h_x$ and $h_y$ are the horizontal and the vertical length of each cell respectively, and $m$ and $n$ are the total number of cells in x and y directions respectively.

Reorganizing the terms of (2.29) and applying forward Euler's methods will give

$$C^{k+1} = N^{-1}\left(t^{k+1} - t^k\right)\left(b - BC^k\right) + C^k. \tag{2.30}$$

Note that $N$ is a diagonal matrix, and thus its inverse $N^{-1}$ is readily available.

## 2.3 Model Physics and Uncertainties

The model grid is defined on a 2-D plane domain. There are 100 cells in both x and y directions making a total of $10,000$ grid points. The horizontal and vertical lengths of the grid are $H = 1000$ m and $L = 500$ m respectively. The permeabilities of the large and the small (embedded) rocks are 100 millidarcy and 10 millidarcy respectively. The embedded rock is positioned exactly in the center of the domain. The density of water is 1000 Kg/m$^3$ and the viscosity is 1 cP. The gravitational acceleration is taken as 9.81 m/s$^2$. The water head is 100 m-water along the western boundary and 10 m-water along the eastern boundary. The water head distribution and the Darcy velocity streamlines inside the domain are shown in figure 2-2. The transport of the contaminant is modeled for 50 years and the time step is 2 months.

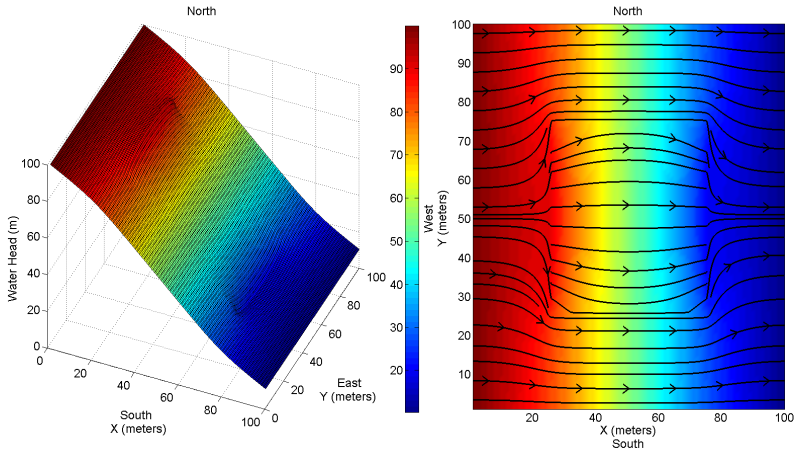


Figure 2-2: Spatial distribution of the water heads (left) and the Darcy velocity streamlines (right) in the porous medium.

The porous medium is considered as totally homogenous within each subdomain, and it contains just solid matrix with no fractures. Dispersion and reaction terms in (2.18) are ignored. Precipitation and dissolution can play an important role as a good source of watering for the aquifer, but in this model all water sources that might take place are

suppressed, thus the $b$ term in (2.30) vanishes. So, the general time dependent transport model will take this form

$$C^{k+1} = AC^k, \tag{2.31}$$

where

$$A = -N^{-1}\left(t^{k+1} - t^k\right)B + I. \tag{2.32}$$

# Chapter 3

# Data Assimilation into Contaminant Models

Data assimilation is the process of combining information from a numerical model and observations to determine the best possible description of the state of a dynamic system. As will be discussed later, "best possible" refers to the fact that the "best estimate" is often difficult to compute because of the large dimension of the system under study, and because of our poor knowledge of the system uncertainties. Roughly speaking, the observations guide the model towards a realistic trajectory, while the model provides a spatiotemporal dynamics interpolation for the observations [25]. Assimilation methods generally fall down into two categories: sequential methods based on statistical estimation theory where the state estimation is carried out sequentially in time with observations, and variational methods based on the deterministic inverse problems theory where the optimization is done for the whole system at once [19]. The work carried out in this thesis is based on the first category coming from the well-known Kalman Filter (KF).

The KF is a well-known statistical data assimilation scheme that provides the best estimate, in the sense of minimum variance, of the state of a linear system with Gaussian errors using all observations up to the estimation time (Kalman 1960). The application of the KF to realistic underground water problems often encounters two major difficulties, nonlinearity of the governing equations and computational cost. The transport contaminant model in this study (2.32) is linear but the model state can be of huge dimension depending on the resolution and the size of the area of interest. The KF relies on the model to

integrate the state estimate in time when no-observations are available, in what is called the KF forecast step. Because models outputs often strongly depend on the input model parameters, determining accurate state estimates would therefore requires good knowledge of the system parameters. In this study, we use the joint state-parameter approach to simultaneously determine estimates of the state and the model parameters using the KF. It is important to note here that even if the model is linear function of the state, the parameter estimation problem is very often nonlinear. Here we resort to the extended Kalman filter (EKF) approach and its low-rank variant to tackle this problem.

## 3.1 Kalman Filter (KF)

The KF can be described by a set of mathematical equations that provides an efficient computational recursive algorithm means to estimate the state of a dynamic system. The optimality criteria of the KF relies on the minimization of the mean squared estimation error. The filter is very powerful in several aspects;

- It can handle estimations of past, present, and even future states,
- It can do so even when the precise nature or real parameters and inputs of the simulated system is poorly unknown.
- It provides estimates not only of the state, but also of the underlying uncertainties, and

The KF uses a form of feedback control to estimate the quantities of interest; the filter gives a prediction for the process state at some time and then obtains feedback in the form of (noisy) measurements. As such, the KF operates in two steps

- Time update equations known as “Forecast Step”
- Measurement update equations known as “Correction Step”

The time update equations project the current state and its error covariance estimates forward in time to provide the a priori estimates for the following time step. The correction step updates the a priori estimate with new observation before the next forecast step takes place.

### 3.1.1 KF Algorithm

We follow the usual notation of the data assimilation community which was proposed by Ide et al. (1997) to describe the algorithm of the KF.

Consider a dynamic system

$$X^t(t_k) = M(t_k, t_{k-1})\, X^t(t_{k-1}) + \beta(t_k), \tag{3.1}$$

where $X^t(t_k)$ denotes the vector representing the true state at time $t_k$, $M(t_k, t_{k-1})$ is the transition operator that integrates the system states from time $t_{k-1}$ to time $t_k$, and $\beta(t_k)$ is the system noise vector representing uncertainties in the model. Here we assume that the model $M$ is linear. As will be discussed later, $M$ represents the contaminant model described in (2.32).

At each time $t_k$, we assume that the observations of the state are obtained from the following observation system

$$Y_k^o = H_k X^t(t_k) + \varepsilon_k, \tag{3.2}$$

where $H_k$ is the observational operator that relates the state to the observation, and $\varepsilon_k$ is the observational noise.

We assume that the process and the observational noises have normal probability distributions with zero means as follows

$$p(\beta) = N(0, Q), \tag{3.3}$$

$$p(\varepsilon) = N(0, R), \tag{3.4}$$

where $Q$ is the process noise covariance and $R$ is the measurement noise covariance.

The KF algorithm is a succession of a forecast step and a correction step. The KF has to be initialized prior to these two steps. The initialization the filter will be discussed in section 4.4.

1. Forecast Step:

   At time $t_{k-1}$ an estimate $X^a(t_{k-1})$ of the system state and its corresponding error covariance matrix $P^a(t_{k-1})$ are available. The forecast step $X^f$ and the associated error covariance matrix $P^f$ are computed by integrating the model forward in time

$$X^f(t_k) = M(t_k, t_{k-1}) X^a(t_{k-1}), \tag{3.5}$$

$$P^f(t_k) = M(t_k, t_{k-1}) P^a(t_{k-1}) M(t_k, t_{k-1})^T + Q_k, \tag{3.6}$$

2. Correction Step:

   Every time a new observation $Y_k^o$ is available, the KF corrects the forecast with the analysis equations

$$X^a(t_k) = X^f(t_k) + G_k \left[ Y_k^o - H_k X^f(t_k) \right], \tag{3.7}$$

$$P_k^a = (I - G_k H_k) P_k^f, \tag{3.8}$$

   where $Y_k^o$ is the new observation at time $t_k$, and $G_k$ is the Kalman gain matrix

$$G_k = P_k^a H_k^T \left( H_k P_k^f H_k^T + R_k \right)^{-1}, \tag{3.9}$$

## 3.2 Covariance Inflation

In some cases where $Q$ is very hard to estimate, an inflation factor is introduced in the covariance equation (3.6) as follows

$$P^f(t_k) = \alpha M P^a(t_{k-1}) M^T, \tag{3.10}$$

where $\alpha$ is the inflation factor, commonly referred to as $1/\rho$, and $\rho$ is a forgetting factor [2].

As shown in equation (3.3), the error covariance matrix represents the uncertainties in the model, but when we cannot estimate it we resort to inflations. The main role of inflations

is to account as much as possible to the missing error covariance matrix. Inflations can help in guiding the filter to the best possible estimate by trusting the observations. In most of the real cases, the observations are noisy but still we can rely on them because they do not include such large uncertainties like the model.

## 3.3 Singular Evolutive Kalman Filter (SEKF)

If the system state has a dimension $N$, the error covariance matrix $P$ will have a dimension $N \times N$. Manipulating the error covariance matrix for large dimensional systems becomes computationaly very expensive and even impossible because of the huge memory storage it requires. The SEKF has been introduced as a way to reduce the computational cost of the KF arising in large dimensional systems. The main idea is to approximate the error covariance matrix of the KF by a singular matrix with low rank $r << N$ which allows the decomposition

$$P = LUL^T, \tag{3.11}$$

where $L$ is of size $N \times r$ and $U$ is simply $r \times r$. Using this decomposition in the KF algorithm, we obtain the algorithm of the SEKF in which only $L$ and $U$ are used. $P$ can still be computed but not needed, therefore drastically reducing computational burden of the KF. This resulting SEKF applies the KF correction only along certain directions called correction directions of the filter, parallel to a linear subspace of dimension $r$. It was shown that these directions are those for which the error is not sufficiently attenuated by the system dynamics [25, 26]. Just as the KF, the SEKF proceeds in two stages apart from an initialization stage based on Empirical Orthogonal Functions "EOFs" (See Chapter 4, section 4.4.1).

1. Forecast Step:

   At time $t_{k-1}$, an estimate $X^a(t_{k-1})$ of the state and its corresponding error covariance matrix $P^a(t_{k-1})$, in the factorized form $L_{k-1}U_{k-1}L_{k-1}^T$, are available. The SEKF

updates the analysis and the correction directions with the model.

$$X^f(t_k) = M(t_k, t_{k-1}) X^a(t_{k-1}), \tag{3.12}$$

$$L_k = M(t_k, t_{k-1}) L_{k-1}. \tag{3.13}$$

The forecast error covariance matrix is then

$$P^f(t_k) = L_{k-1} U_{k-1} L_{k-1}^T + Q_k. \tag{3.14}$$

2. Correction Step:

The KF correction step is then applied as

$$X^a(t_k) = X^f(t_k) + G_k \left[ Y_k^o - H_k X^f(t_k) \right], \tag{3.15}$$

with the Kalman gain now given as

$$G_k = L_k U_k L_k^T H_k^T R_k^{-1}, \tag{3.16}$$

with

$$U_k^{-1} = \left( U_{k-1} + \left( L_k^T L_k \right)^{-1} L_k^T Q_k L_k \left( L_k^T L_k \right)^{-1} \right)^{-1} + L_k^T H_k^T R_k^{-1} H_k L_k. \tag{3.17}$$

The corresponding filter analysis error covariance matrix is then equal to

$$P^a(t_k) = L_k U_k L_k^T. \tag{3.18}$$

Again, equations (3.14) and (3.18) are only included for better interpretation of the filter's algorithm the results, but are not needed in the algorithm. It is also important to note that equation (3.17) was obtained after projection of the model error in the filter correction directions $L$ as described in [34]. This is needed to avoid an unbounded increase in the rank of $P$ [34].

When inflation is used to represent the model error in the SEKF, equation (3.17) becomes

$$U_k^{-1} = \alpha U_{k-1}^{-1} + L_k^T H_k^T R_k^{-1} H_k L_k. \tag{3.19}$$

Even if the statistics of the model are known, the use of inflation in the SEKF is also needed to mitigate for the underestimation of $P$ by low-rank matrices and because of the projection $Q$ on $L$.

## 3.4 Singular Fixed Kalman Filter (SFKF)

The SFKF carries the same low-rank idea just as the SEKF; it only aims to decrease the computational cost more by fixing $L$ with time. This means that the correction directions are obtained just once without further updates by the model. This can introduce more uncertainties and noise in the system but still it can be manipulated to estimate future state estimates. This filter can be used for high resolution grids where the size of the dynamic system $N$ is extremely large and estimating the correction directions at each step (3.13) can slow down the speed of the algorithm and take a huge memory storage.

The main change is the error covariance matrix which will be written as

$$P^a(t_k) = LU_kL^T. \tag{3.20}$$

The algorithm of the SFKF is exactly the same as the SEKF and the only difference is that the evolving equation for $L$ (3.13) does not exist anymore.

## 3.5 Extended Kalman Filter (EKF)

The EKF was introduced to allow the application of the KF to (moderately) nonlinear systems. The main idea is to linearize the system about the most recent state estimate before applying the KF [19]. Linearization can be done by several ways including Taylor expansions, finite differences, ... [46]. It is now more customary to use the Ensemble Kalman filter (EnKF) for nonlinear data assimilation problems [27]. In our study, the EKF was found efficient enough to compute accurate estimates of the model parameters. Future work will consider EnKF-based methods for more accuracy.

The algorithm for the EKF is quite similar to the one of the KF, only the linearized operators of the model and the observational operators are now in the KF algorithm as follows

1. Forecast Step:

$$X^f(t_k) = M(t_k, t_{k-1}) X^a(t_{k-1}), \tag{3.21}$$

$$P^f(t_k) = \mathbf{M}(t_k, t_{k-1}) P^a(t_{k-1}) \mathbf{M}(t_k, t_{k-1})^T + Q_k, \tag{3.22}$$

where $\mathbf{M}(t_k, t_{k-1})$ is the gradient of $M(t_k, t_{k-1})$ evaluated at $X^a(t_{k-1})$.

2. Correction Step:

$$X^a(t_k) = X^f(t_k) + G_k \left( Y_k^o - H_k X^f(t_k) \right), \tag{3.23}$$

$$G_k = P_k^a \mathbf{H}_k^T \left( \mathbf{H}_k P_k^f \mathbf{H}_k^T + R_k \right)^{-1}, \tag{3.24}$$

$$P_k^a = (I - G_k \mathbf{H}_k) P_k^f, \tag{3.25}$$

where $\mathbf{H}_k$ is the gradient of $H_k$ evaluated at $X^f(t_k)$.

The computation of the forecast error covariance matrix $P^f$ requires the manipulation of matrix of order $N$, and at least $N$ model integrations. Therefore, approximations are unavoidable. The SEEK filter is a good approach to reduce the cost of the EKF [26].

## 3.6 Singular Evolutive Extended Kalman Filter (SEEKF)

The SEEKF is the extended version of the SEKF, where linearization is incorporated in the algorithm as in the EKF. After initialization, the forecast and the correction steps are as follows

1. Forecast Step:

$$X^f(t_k) = M(t_k, t_{k-1}) X^a(t_{k-1}), \tag{3.26}$$

$$L_k = \mathbf{M}(t_k, t_{k-1}) L_{k-1}. \tag{3.27}$$

2. Correction Step:

$$X^a(t_k) = X^f(t_k) + G_k\left(Y_k^o - H_k X^f(t_k)\right), \tag{3.28}$$

$$G_k = L_k U_k L_k^T \mathbf{H}_k^T R_k^{-1}, \tag{3.29}$$

$$U_k^{-1} = \left(U_{k-1} + \left(L_k^T L_k\right)^{-1} L_k^T Q_k L_k \left(L_k^T L_k\right)^{-1}\right)^{-1} + L_k^T \mathbf{H}_k^T R_k^{-1} \mathbf{H}_k L_k. \tag{3.30}$$

## 3.7 Localization of the Filter Analysis

As can be seen in equations (3.15) and (3.16), the filter correction is only applied in the directions of $L$. The low-rank approximation used in SEKF and SFKF, therefore results in very few degrees of freedom for the filter analysis to fit available observations. Another problem, but closely related, can be due to the bad representation of long-term correlations of a covariance matrix using low-rank approximation. Based on Houtekamer and Mitchel (2001), the simplest strategy to deal with this problem is to exclude observations greatly distant from the grid point being analyzed. By doing so, short-range correlations in the filter's error covariance matrices will be preserved, and long-range correlations will be filtered out. In other words, localization can fit the data and filter out spurious unrealistic long correlations of the covariance matrices dominated by large scale signals.

Localization is now considered as a necessary tool for a successful implementation for a low-rank KF, including Ensemble KFs [49, 28].

Applying localization would require changes only in the analysis step, where only a specific number of observations falling within a distance from the point being estimated are used. This localization idea is defined by means of a radius of influence around the analyzed point. All data located outside this area of influence are discarded. The analysis equations of the SEKF will take the form

$$X_j^a(t_k) = X_j^f(t_k) + L_{k,j} U_{k,j} x_k, \tag{3.31}$$

$$x_k = L_k^T H_{k,j}^T R_{k,j}^{-1} \left( Y_{k,j} - H_{k,j} X^f \left( t_k \right) \right), \tag{3.32}$$

$$U_{k,j}^{-1} = U_{k-1,j}^{-1} + L_k^T H_{k,j}^T R_{k,j}^{-1} H_{k,j} L_k, \tag{3.33}$$

where $H_{k,j}$ and $R_{k,j}$ are the observation and the measurement noise covariance matrices for every single point in the state vector respectively, $Y_{k,j}$ corresponds to the observations located within the radius of influence of the analyzed point, $L_{k,j}$ is the $j^{th}$ row of $L$ matrix, and $U_{k-1,j}$ is the initial $U$ at point $j$.

## 3.8 Computational Requirements of the Filters

One of the basic criteria followed to compare the filters is by looking at their computational requirement. Obviously, the low-rank filters require the least effort that is $r+1$ time the cost of the numerical integration of the model (to compute the evolution of $L$). The full Kalman; however, requires $N$ times the cost of the model integration which is much larger than $r+1$. In the extended Kalman filters, there is another computational effort given for updating the model around the most recent parameters. If localization is done, the computational effort will increase in the analysis step. It depends on how much observations are found in the neighborhood of the point to be analyzed.

## 3.9 Implementation of KFs on Contaminant Transport Models

Once the filters algorithms are well coded, we can plug the contaminant state vector and its covariance matrix in these filters. The contaminant state correspond to the concentration value of the contaminant in each cell of the domain. Contaminant data at specific locations in the domain have to be collected before starting filtering. The output from the filter will give an idea about the position of the contaminant plume after some time. Other model parameters such as the permeability, the spatial distribution of the rocks in the

porous medium, ... can be estimated using the KFs. All these estimation problems will be discussed in the following chapter.

# Chapter 4

# Numerical Applications

Several experiments were performed, each with specific configurations, using different parameters and inputs and focusing on different objectives. Based on the information given in the previous chapter, we considered the transport contaminant model on two grids, low and high resolution, and the reason for that is to study the impact of the dimension on the problem. The KF can be applied on the coarse mesh grid (CMG) whereas, the SEKF can be used for both the coarse and the fine mesh grid (FMG). The meshing properties of the two grids are detailed in Table 4.1.

Table 4.1: Meshing properties for the low and the high resolution model grids.

| Meshing Properties | CMG | FMG |
|---|---|---|
| The total number of cells | 2500 | 10000 |
| The total number of nodes | 2601 | 10201 |
| The number of edges | 5100 | 20200 |
| The number of boundary edges | 200 | 400 |
| The length of each cell in x-direction ($m$) | 20 | 10 |
| The length of each cell in y-direction ($m$) | 10 | 5 |
| The area of each cell ($m^2$) | 200 | 50 |

As it can be seen from the table above, the FMG has double more cells than the CMG. The CMG has a lower resolution than the FMG and this is clearly shown in the areas of

each cell in both grids. Smaller cell areas means better representation of the contaminant distribution inside the domain.

Obviously, the FMG would require more computational time because our objective is to know the contaminant concentration at each cell location. The computational effort in the CMG is faster but less accurate because the contaminant is averaged on larger areas. That's why it is important to be able to manipulate contaminant model on the FMG to get accurate results and good contaminant image.

## 4.1 Twin Experiments

Twin experiments are used as a tool to assess the performances and the capabilities of our filters. These experiments work in the following manner:

1. A reference experiment is performed and the reference contaminant states are saved to be compared later with the estimations of the filters.

2. Pseudo-measurements are then extracted from these reference states based on our choice.

3. Later, we run the assimilation experiments using the contaminant model and the collected pesudo-observations.

The twin experiments can validate or dis-validate the filter's performance based on the resulting estimations. We expect the estimation of the contaminant to improve and get closer to the reference contaminant states throughout filtering if we are using correct observations and true model. If the results get worse, then we know that the filter is not functioning properly and the reason could be a problem in the observations or the model.

## 4.2 Reference States

The reference state, or what we what we refer to as "the truth", includes all the states that are used to evaluate the performance of the filters and to study their behaviors. These

states result from the simulation of the dynamic model, using the correct parameters and initial conditions. In our study, we collect the true states every 2 months by running the transport model in (2.31) for 50 years using the model parameters as mentioned section (2.3). Chang and Latif (2010), considered a pulse mass input of 1604 g, producing an initial concentration ($C_0$) of $10,000$ mg/l; this contaminant is injected at a single point in the grid [9]. In our model, we consider a contaminated area (plume) in a fully pure aquifer. The contaminant is then transported into the medium with the water flow. For the low resolution grid, the contaminant plume is located close to the western boundary in the subdomain $[4, 6] \times [5, 45]$ , and it moves towards the eastern side. The same initial condition is considered for the high resolution grid, but this time the contamination plume is located in the subdomain $[8, 12] \times [10, 90]$. The initial concentration of the contaminant is 100 ppm. The flowing water entering the aquifer goes around the low permeability layer because the values of the Darcy velocities in that subdomain are very small. Figures 4-1 and 4-2 show the evolution of the contaminant every 5 years inside the aquifer for the CMG and the FMG respectively.

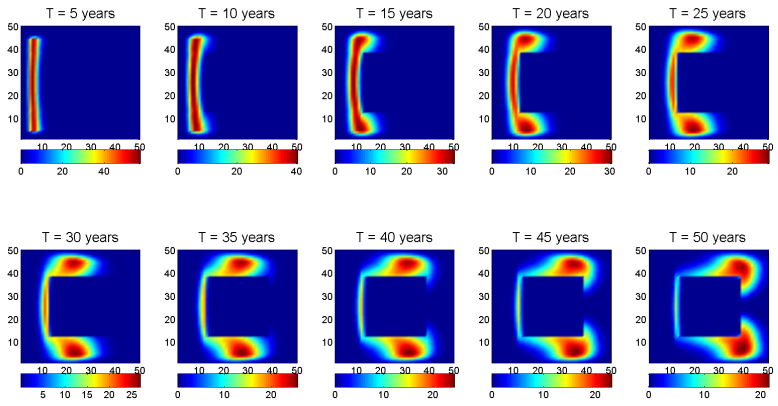


Figure 4-1: The reference "true" states of the contaminat transport model (CMG). The initial contaminant concentration is 100 ppm.

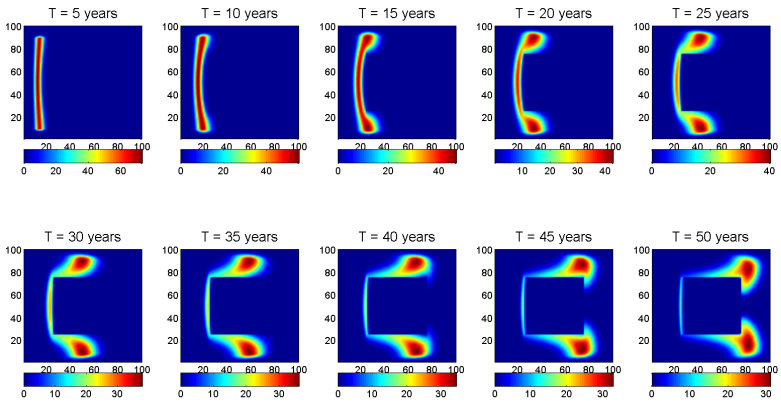


Figure 4-2: The reference "true" states of the contaminat transport model (FMG). The initial contaminant concentration is 100 ppm.

One can see that the spread of the contaminant is thinner in the FMG than the CMG. This is because the larger number of cells inside the FMG, meaning that the concentration values are being assigned to small areas rather than approximating them for larger areas as in the CMG case.

## 4.3 Pseudo-Observations

We collect pseudo-observations for both the CMG and the FMG. Since there are more cells in the FMG, we choose to extract more data from the reference states of this grid. In the real life case, it's extremely hard and expensive to get observations from the whole domain area. We collect observations from the grid points located along the path of the moving contaminant.

In total, there are 305 reference states and observations have to be collected from each of these states, i.e. every 2 months. From the CMG, we collect observations - in a vertical manner every 200 m - from 160 cells out of 2500. In the FMG we choose to collect - in the same vertical manner but every 100 m - 720 observations out of $10,000$ cells (Figure 4-3).

Observations represent the true state of the contaminant only if there is absolutely

no measurement error. In practical measurement, this ideal situation does not often take place. The measurement is incorporated with unavoidable error termed as measurement noise. This may take place due to an inaccurate reading of the measuring instrument, lack of proper instrument calibration, and insensitivity of the measuring instrument ... [9].

Perturbation is imposed on the observations to include measurement noise so that the experiments are constructed in a more realistic setting. The observation error is simulated by adding randomly generated Gaussian noise with zero mean. We assume that the observational errors are not correlated, and thus we write the error covariance matrix $R$ as a diagonal matrix having the variances of the observational errors as its entries. In most of the cases, we assume 10% and 30% observational errors of the total variance.

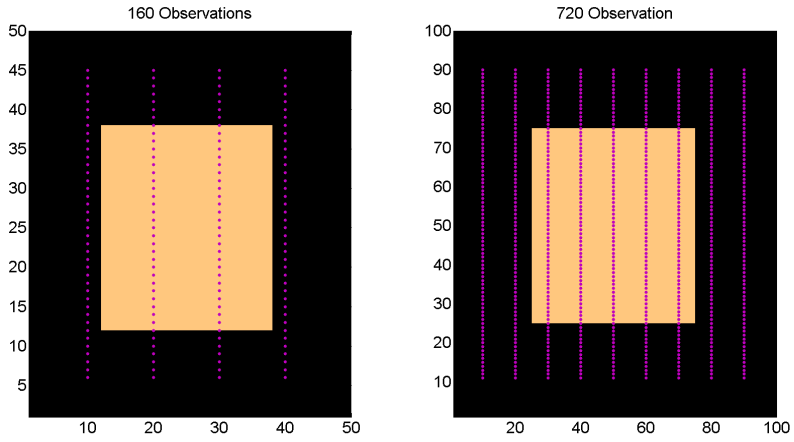


Figure 4-3: "Pesudo" observations taken from each state of the CMG (left) and the FMG (right).

## 4.4 Initialization of the Filters

As discussed in section (3.1.1), to initialize the filters one needs some initial estimate of the state vector $X^a(t_0)$, and its initial error covariance matrix $P^a(t_0)$. The choice of $X^a(t_0)$ and $P^a(t_0)$ is usually not very important on the long term behavior of the filter. To

initialize the KF, we take $X^a(t_0)$ as the average of the simulated state vectors from the reference states (Figure 4-4), and $P^a(t_0)$ as the sample covariance matrix of these vectors. To initialize the SEKF, we take $X^a(t_0)$ as the initial state (similar to KF) and $P^a(t_0)$ as the low-rank approximation of the sample covariance matrix which we compute using empirical orthogonal function analysis. Such an analysis can provide the initial $L_0$ and $U_0$ needed to start the SEKF.

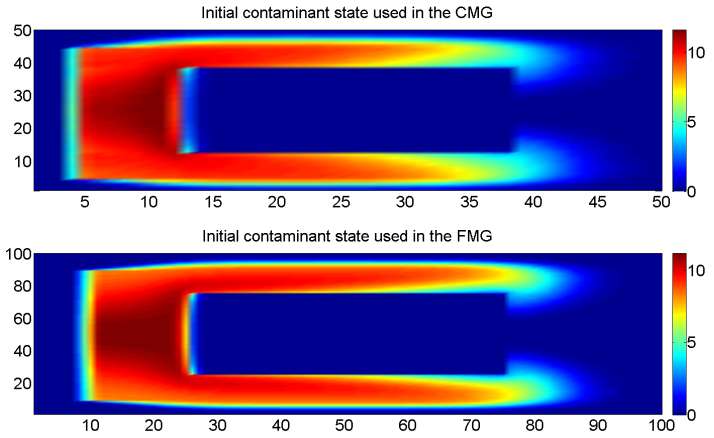


Figure 4-4: Initial contaminant states (mean of all 305 states) used in the filters for both, the low and the high resolution grids.

### 4.4.1 EOF Analysis

The Empirical Orthogonal Function (EOF) analysis, is a method to split the temporal variance of spatially distributed data into orthogonal spatial patterns called EOFs. It can be viewed as a method of compressing data contained in a set of subsurface states, by summarizing the correlation of their variables in a few vectors, called EOFs. The EOFs are the eigenvectors of the sample covariance matrix of the set of states.

The relative importance of any individual EOF to the total variance in the field is measured by its associated eigenvalues. The theorem of Taylor-Young also demonstrates

that the EOF analysis provides the best low-rank $r$ approximation of sample covariance matrix $P$ (in the sense of least squares) decomposed in the form

$$P \approx LUL^T, \tag{4.1}$$

where $U$ is a diagonal matrix containing the eigenvalues of $P$; $\lambda_1, ..., \lambda_r$ ranked in decreasing order on its diagonal [24]. In many earth sciences applications, only few eigenvalues are found significant, whereas the rest are very small suggesting that a drastic rank reduction is possible.

### 4.4.2 Calculation of the EOFs

The sample covariance matrix $P$ can be obtained by

$$P = \frac{1}{N} \sum_{i=1}^{Ns} \left(X_i - \overline{X}\right) \left(X_i - \overline{X}\right)^T, \tag{4.2}$$

where $X_i$ is the $i^{th}$ contaminant state, $\overline{X}$ is the mean of all states, and $Ns$ is the total number of states (i.e. 305 in our case).

After getting $P$, we can compute the eigenvalues and the eigenvectors of this matrix. We sort the eigenvalues in decreasing order on the main diagonal of matrix $U$ and the eigenvectors in $L$. Since the eigenvectors are orthogonal so we can write the following equality

$$X_i - \overline{X} = LL^T \left(X_i - \overline{X}\right), \tag{4.3}$$

where $LL^T$ is identity. If we take the first $r$ eigenvectors $L_r$ associated with the largest $r$ eigenvalues we can approximate our centered states in (4.3) as

$$X_i - \overline{X} \approx L_r L_r^T \left(X_i - \overline{X}\right). \tag{4.4}$$

This approximation means that the centered states can be projected on a smaller subspace of dimension $r$ if multiplied by the transpose of $L_r$, then we can reconstruct the states in the original space by multiplying the projected states by $L_r$.

The two bar plots in figure 4-5 show the first 50 eigenvalues from the CMG and the FMG. It is clear that only the first few eigenvalues are the significant ones. The FMG has larger eigenvalues than the CMG and this is expected because of the large dimension of the FMG. This means that more information is needed to represent the FMG. We can also understand this better by plotting the inertia of each eigenvalue (Figure 4-6). We see that for both grids the first 10 EOFs account for more than 90% of the inertia of the sample.

Based on these two figures (4-5 and 4-6) we can choose the number of retained EOFs in all assimilation experiments.

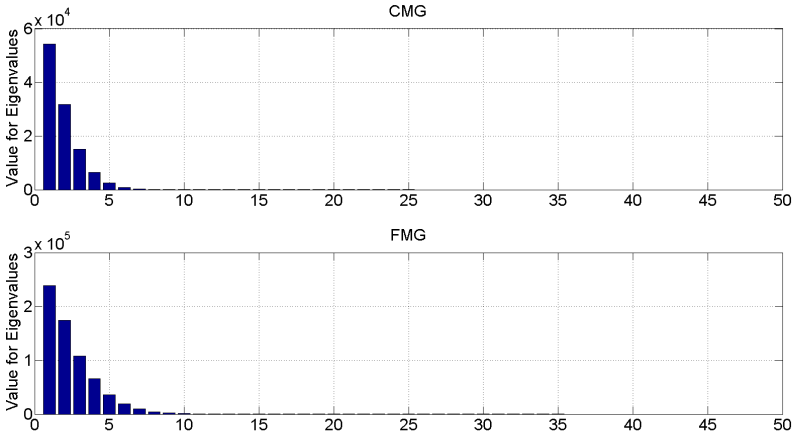


Figure 4-5: Bar plot showing the values for the first 50 eigenvalues from the CMG (in total there are 2,500 eigenvalues) and the FMG (in total there are 10,000 eigenvalues).

## 4.5 Forecast Model

The forecast model is the model used to integrate the state estimates forward in time. The forecast model is said to be perfect if we use the same model that has been used to generate the observations. In this case, no model errors are considered. In the other case, the forecast model is said to be imperfect.

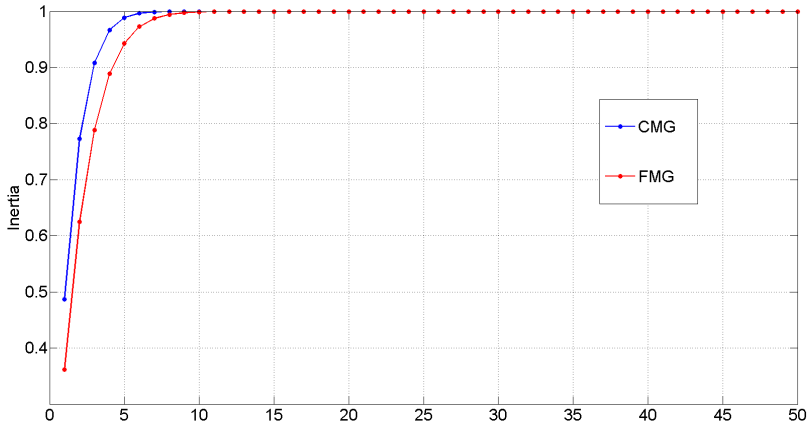


Figure 4-6: Percentage of inertia versus the number of EOFs from both grids.

Apart from the assimilation experiments, we apply free model runs on the CMG and the FMG with no assimilation. The purpose of these experiments is to evaluate the performances of the filters and to show that assimilation improves the behavior of the model. We compare the results of the filters to that obtained from a model run starting from the filters initial conditions and running without assimilation using the forecast model.

**Case 1** *Estimation with perfect model forecast*

In this first case, the forecast of the contaminant state is computed using the "true" model, i.e. using the model parameters discussed in section (2.3) and the initial conditions shown in figure 4-4. These experiments allow us to evaluate the filters performances without the influence of the model errors. The only difference in the filters performances are due to the formulations of the filters. We carry out several experiments using 3 different filtering techniques namely;

**KF** Applied only on the CMG using observations from the data points in the domain (Figure 4-3).

**SEKF** Carried out for both, the CMG and the FMG. A number of filter ranks are considered and analyzed.

**SFKF** Same application like the SEKF.

Note that no model error covariance matrix $Q$ was used in the filters runs as the model was perfect.

**Case 2** *Estimation with imperfect forecast model*

In real applications, the model is subject to model errors. To test the filters in a more realistic settings, we used a perturbed forecast model. More precisely, we decreased the rock permeabilities by 20% to make $k1 = 80$ millidarcy and $k2 = 0.8$ millidarcy. We apply the same filtering techniques as in the first case and we compare the efficiency of the filters by including the process noise covariance matrix $Q$ and using inflation factors.

$Q$ is estimated as the sample covariance matrix between the correct solution states and the perturbed model states. Then in the filter's algorithm, at each time step when a new state is analyzed, $Q$ can be updated.

We also apply localization for the low rank filters using 3 influence areas. The searching criteria for observations is implemented by considering a rectangular area and the grid point being analyzed is placed at the center of the rectangle. The areas were chosen as follows

1. $R1$; starting from the grid point we search in horizontal direction 40 m to east and 40 m to west. In vertical direction, we search 20 m to north and 20 m to south.

2. $R2$; in the same manner but using a larger influence area. We use 100 m in east and west directions and 50 m in north and south directions.

3. $R3$; is the largest searching area. We look for data located within a distance of 200 m to the east and west and 100 m in north and south directions.

Since in most of the real cases we face imperfect models, we apply localization just in this case. For the localization idea, we check from the first case whether the SEKF is working well with the perfect model then we go to the imperfect model and apply localization.

## 4.6 Evaluation of the filters solution

To account for the analysis errors in both cases, we look at the Root Mean Square Error (RMSE). The RMSE measures the difference between the state predicted or estimated by the filter and the reference state. It can be calculated as follows

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (X_{Model,i} - X_{True,i})^2}{N}}. \tag{4.5}$$

We also look at the estimated spatial distribution of the contaminant inside the aquifer and we compare it with the reference states in figure 4-1.

## 4.7 Numerical Results

### 4.7.1 Estimations using the Perfect Model

The primary point to prove is the importance of data assimilation and how can the observations guide the model to the true trajectories. We start our experiments with the first case from the forecast model.

#### Free Run and KF

We used the perfect model and we compared the free run results with the KF estimations. Figures 4-7 and 4-8 show the evolution of the contaminant in time from the free run and the KF estimations respectively. One can see that the spatial distribution of the contaminant predicted by the KF is more accurate and closer to the reference states (Figure 4-1) than the free run estimation.

The free run underestimates the concentration of the contaminant after 50 years by almost 10 ppm compared to the reference states. We looked at the analysis errors from both runs (Figure 4-9) and we noticed that the RMSE values were also consistent with the results given by the spatial distribution of the contaminant. The RMSE of KF is large at the early assimilation steps, but quickly decreases after assimilating the data into the

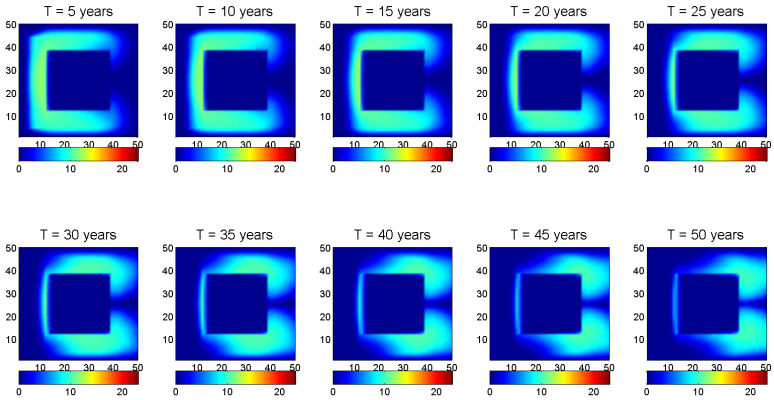


Figure 4-7: Concentration of the contaminant every 5 years as obtained from the free run. (Model run without assimilation starting from the filter's initial condition).

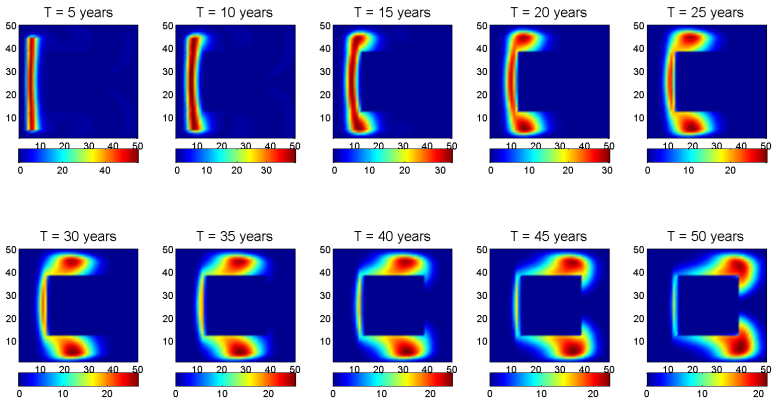


Figure 4-8: Concentration of the contaminant every 5 years as estimated by the KF with 30% observational errors.

model. The RMSE of the free run also decreases in time but with much slower pace than the KF, reflecting in a way the stable dynamics of the contaminant model.

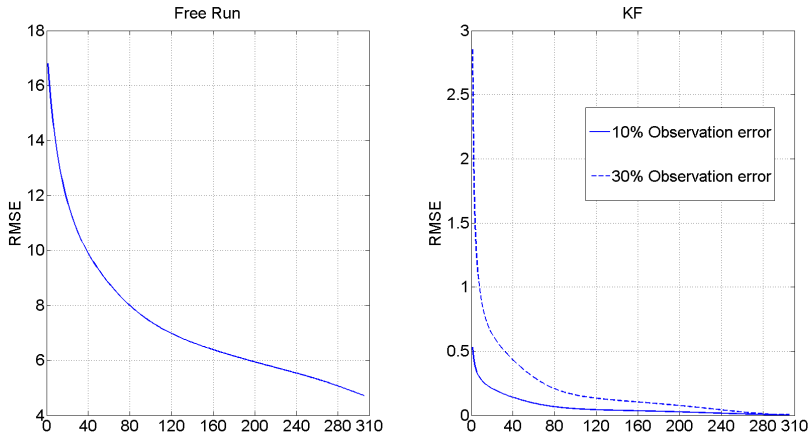


Figure 4-9: Variation of the analysis errors in time from the free run (without data assimilation) and the KF (with 10% and 30% observational errors).

**KF versus SEKF**

In this set of experiments, we study the behavior of the SEKF and compare its performance to the KF. What we are trying to find is an accurate and fast estimation on the same time. Here we use the SEKF that is guaranteed to support us with the least time and memory storage that we seek.

The first question we try to answer is how to choose a good rank for the SEKF. Based on figures 4-5 and 4-6, the CMG has very few eigenvalues that are significant ($\sim 10$) and they account to more than 90% of the total inertia. Note that the larger the rank, the more model integrations are needed to evolve the filter correction directions implying increase in the computational burden. Based on that, we conducted 2 sets of experiments and we implemented the SEKF with different rank values 6, 8, and 10. The results are then compared them to the ones of the KF.

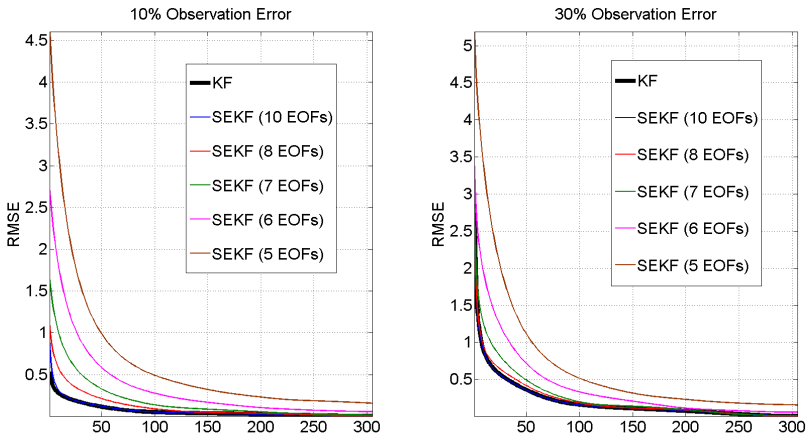


Figure 4-10: Comparison between the analysis errors given by the KF and the ones of the SEKF based on the number of EOFs (with 10% and 30% observational errors).

Looking at figure 4-10, the first thing that can be noticed is the blue curve representing the analysis errors of the SEKF using 10 EOFs. It almost coincides with the bold black curve representing the KF meaning that the low-rank approximation has only a marginal impact on the accuracy of the estimation. We did not lose any essential information while going to a smaller subspace because we are still able to get almost the same estimations with the SEKF as in the full Kalman. This accurate estimation was obtained with a drastic decrease in computational time and memory storage. Storing 2 matrices $L$ of size $n \times r$ and $U$ of size $r \times r$ instead of a covariance matrix of size $n \times n$ can significantly decrease the memory storage.

As we decrease the rank of the filter, we start losing information and the estimation errors start to increase. This is expected because the first 10 eigenvalues are the largest ones and ignoring some of them will lead to a less accurate estimation.

### Testing SFKF

In the following experiment, we use the SFKF for the sake of increasing the speed of the algorithm more. We compare the RMSE values with those of the SEKF in figure 4-11.

Obviously, the RMSE values for SEKF estimations are better than those of the SFKF but the SFKF is significantly faster. To certain extent, the RMS errors given by the SFKF can be tolerated given the high speed of the algorithm when compared to the KF.

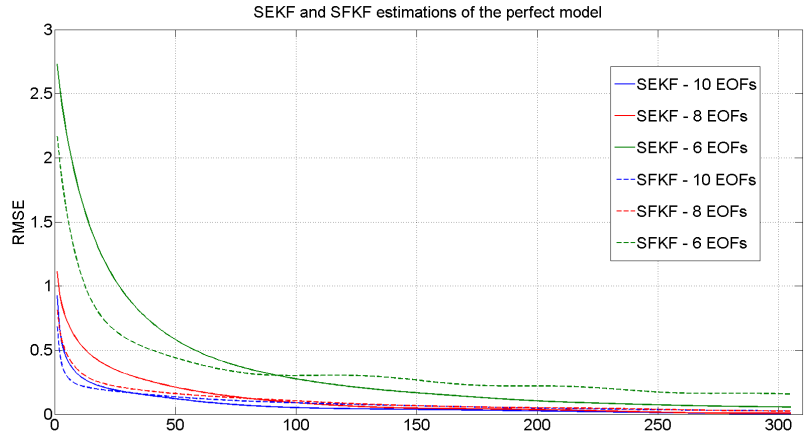


Figure 4-11: The SEKF (solid curves) and the SFKF (dashed curves) analysis errors for the perfect model with different EOFs at 10% observational errors.

This interesting result can tell us that we can increase the rank of the SFKF and obtain better results than the SEKF (running with smaller rank) without increasing the computational cost as this does not require any new model integrations. Only storage and the analysis step would be more demanding but not like integrating the model.

### High Resolution Grid

The last experiment with the true forward model was carried out with the model solved on the FMG instead of the CMG. In this setup, it was not possible to implement the KF because we ran out of memory. MATLAB could not afford handling a covariance matrix of size $10,000 \times 10,000$. It is important to mention here that the CMG runs could validate for us the usage of the SEKF by comparing its estimations with those of the full Kalman. Now, we run this FMG using the SEKF and this is another advantage for the SEKF that

it can handle large dimension problems at times where the KF can not. We got the lowest analysis errors using the first 15 EOFs as in Figure 4-12. We run the same model in a free run mode (Figure 4-13) and again we see the important effect of data assimilation in guiding the filter toward the true solution. The RMSE values after 50 years by the SEKF using just 5 EOFs is less than 1 and almost 0 for 15 EOFs; however, it is still greater than 6 in the free run simulation.

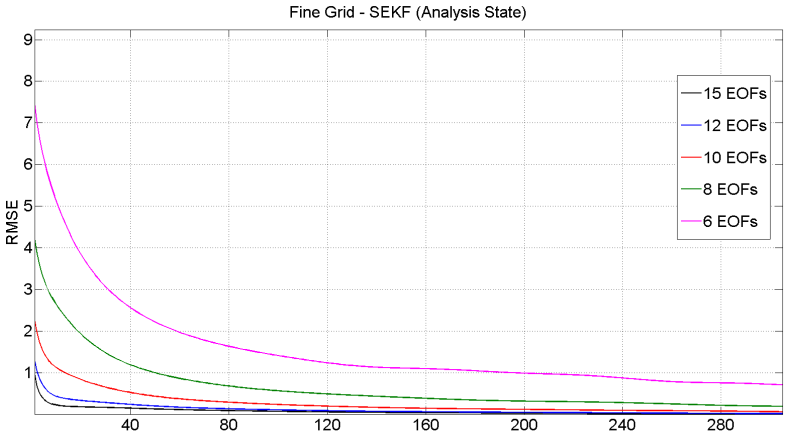


Figure 4-12: RMSE values resulted from the SEKF when implemented on the FMG with different filter ranks at 10% observational errors.

### 4.7.2 Estimations using the Imperfect Model

We consider now the case of the imperfect forecast model which is the true model with perturbed initial conditions and permeabilities. Clearly, the filter estimates should be less accurate than the ones we obtained with the true model. This is a more realistic case when we do not know exactly the reservoir parameters. We start from some estimates, collect some (noisy) data, and then apply filtering. In all the experiments in this section, we impose 10% errors on the observations.

The use of an imperfect model to forecast the state of the KF significantly degrades

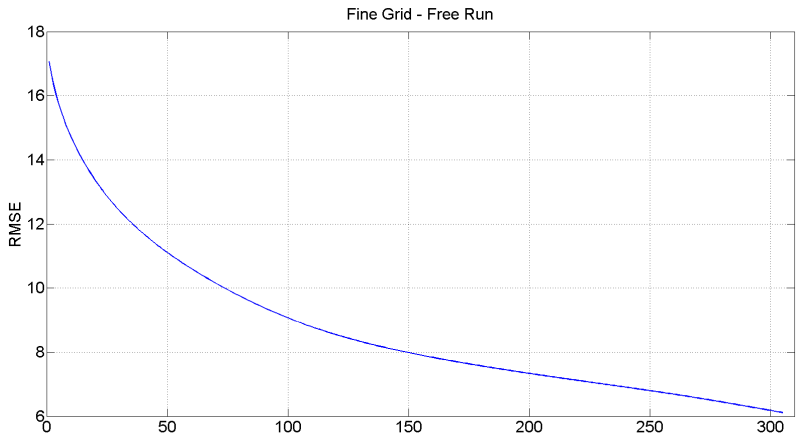


Figure 4-13: RMSE values for the free run when applied on the FMG.

the KF estimates. The RMSE increases in time because the filter does not fully use the information in the observations as more weight is given to the enormous forecast (Figure 4-14). Adding inflations increases the uncertainties on the model forecast and push the filter to trust more the observations, this results in a better filter performance. As expected, inflation stabilizes the filter RMSE at the early assimilation window before allowing the filter to converge towards the true state at the end of the assimilation period. However, increasing inflation beyond 1.08 increases the error and at some point it caused the filter to diverge. A more preferable strategy is take into account the uncertainties in the model and to compute the error covariance matrix. It is important to note that this is however based on the assumption of additive noise which is likely to be not true in this configuration. As discussed in section (4.5), $Q$ was estimated as the sample covariance matrix between the true reference run and the perturbed model. The bold dashed curve in figure 4-14 shows how including $Q$ in the filter's algorithm greatly improves the stability of the filter; unlike when using inflations where we see the estimation error varying in a more irregular pattern. The filter with $Q$ cannot however decrease the RMSE values while for certain inflations, the filter RMSE continuously decreases in time.

Further we tested the behavior of the SEKF and the SFKF with the same imperfect

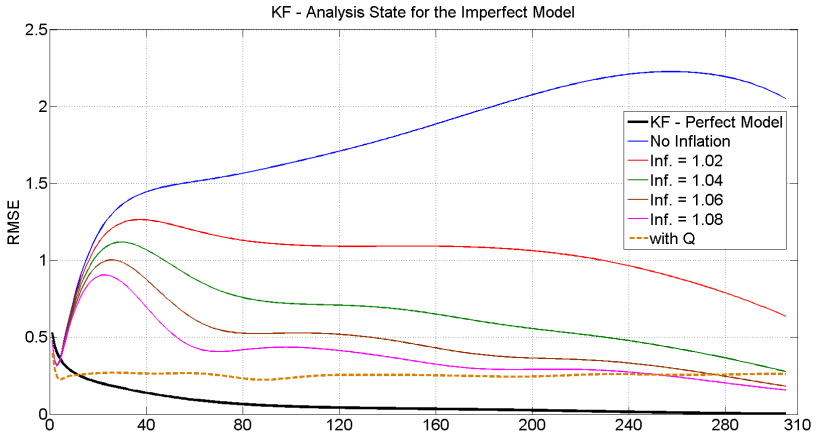


Figure 4-14: KF analysis errors for the imperfect low resolution model with $Q$ and inflations (observational errors are 10% of the total variance).

model. Figure 4-15 shows the RMSE values from the SEKF (with 10 EOFs) using both inflations and $Q$. There are 2 main interesting features that one can get from this plot:

1. Firstly, the filter could now handle more inflation than the KF, up to 1.14. This can be expected because the low-rank approximation of the SEKF underestimates the filter covariance matrices allowing for more inflation than in the KF. More inflation again continuously decreases the RMSE to about 0.12 with an inflation $\alpha = 1.14$, but after certain level, the filter diverges as for the KF.

2. Secondly, using $Q$ in this SEKF interestingly decreases the error to 0.08 and this contributes to a more accurate and much stable estimation than all inflation cases.

For the SFKF (with 10 EOFs), again using $Q$ gave slightly better performance than using inflation, but the filter did not handle as much inflation as the SEKF (Figure 4-16). This can be explained by the invariant correction directions that are used to parametrize the filter covariance matrices. The filer can be then sometimes overestimated and adding inflation might degrade the results. The error with SFKF decreased after 50 years to 0.58 which is larger than both the KF and the SEKF.

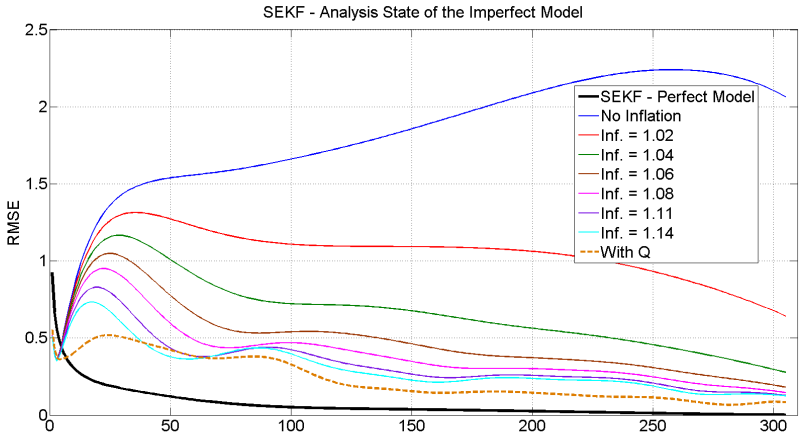


Figure 4-15: SEKF analysis errors for the imperfect low resolution model with $Q$ and inflations (observational errors are 10% of the total variance).

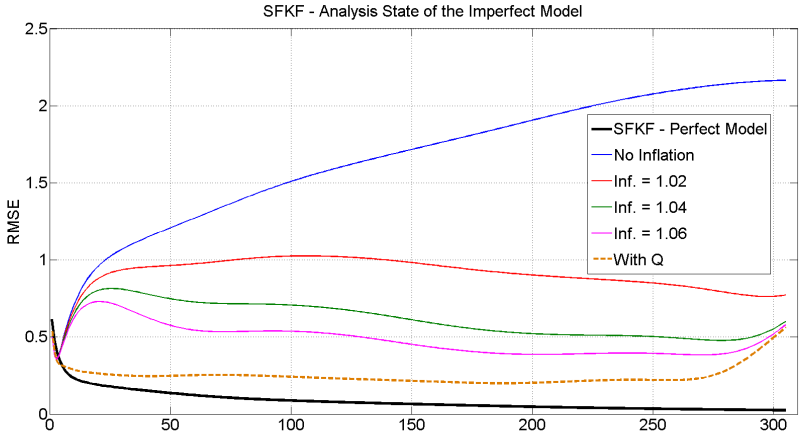


Figure 4-16: SFKF analysis errors for the imperfect low resolution model with $Q$ and inflations (observational errors are 10% of the total variance).

To summarize the results of this section, we compare the 3 performances of the filters (Figure 4-17) in terms of using inflation and $Q$. The least accurate one in both cases is the SFKF, the error is seen to decrease and later increases towards the end of the assimilation window. This shows that omitting the evolution of the correction directions in the filter's algorithm degrades the performances and lead to less accurate estimates. Concerning the KF and the SEKF, surprisingly the low-rank approximation is found to provide better estimates than the full Kalman. The RMSE values of the SEKF were smaller than those of the KF in both cases. Also, the SEKF with $Q$ lead to less errors and was more stable than the KF. As a tentative exploration, we hypothesize that the low-rank approximation filters out some of the model noise to better behave in the presence of model uncertainties.

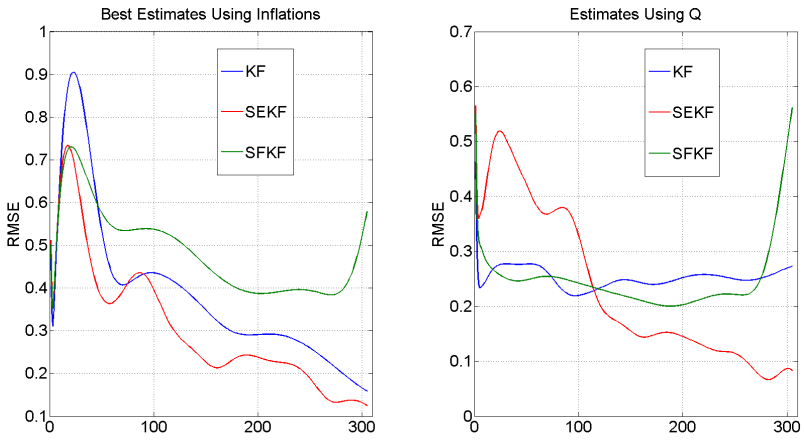


Figure 4-17: Comparison between the 3 filters (using the imperfect CMG model) based on their RMSE values and the usage of $Q$ and inflations.

### 4.7.3 Effect of Localization on the Estimated Contaminant States

As discussed in section (3.7), the localization is applied for the analysis step in the SEKF and the SFKF filters. We tested localization only with the CMG because applying it on the FMG would be very slow. The domain grid in the FMG is very large and for each

point we have to look for data around it, so computationaly it can be implemented but not so feasible. The results shown in Figures 4-19 and 4-20 were obtained from assimilation runs with three influence areas as described in the forecast model, section (4.5). In all assimilation experiments, we impose 10% noise as perturbation on the observations.

In both filter runs, we see that the larger the influence area the better the estimation becomes. Based on the local support idea, if the point gets its analysis from the data close to it we should get a better estimation. Well, this is correct as long as the neighborhood around the point include observation points; if not we will end up having a case where no data will be assigned to the point and its value would remain unchanged. This is the case mainly for $R1$ and $R2$, most of the grid points had no observations around them so their values ended up being uncorrected. Especially for $R1$, the RMSE curve resembles to the free run to certain extent because what is really going on is just forecasting with very small corrections. The fact that the RMSE values at the beginning are large is due to the same problem. What is being estimated at the beginning is so much affected by the initial condition (mean of all the states) because there are no enough observations for correction (Figure 4-18).

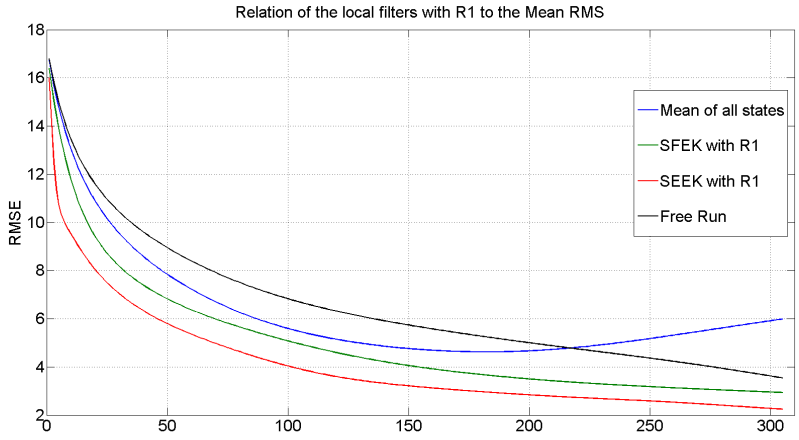


Figure 4-18: RMSE values using localization with $R1$ searching area and the free run (CMG).

In the SEKF, we see that when using localization with $R2$ we were able to obtain more accurate estimates than the SEKF case with no inflations and it even becomes much better with $R3$. For the SFKF case, the error with $R2$ and $R3$ was also not that as accurate as the filter run without inflation. From the computational side, as the area of influence around the point increases, estimations become more expensive requiring more computational time.

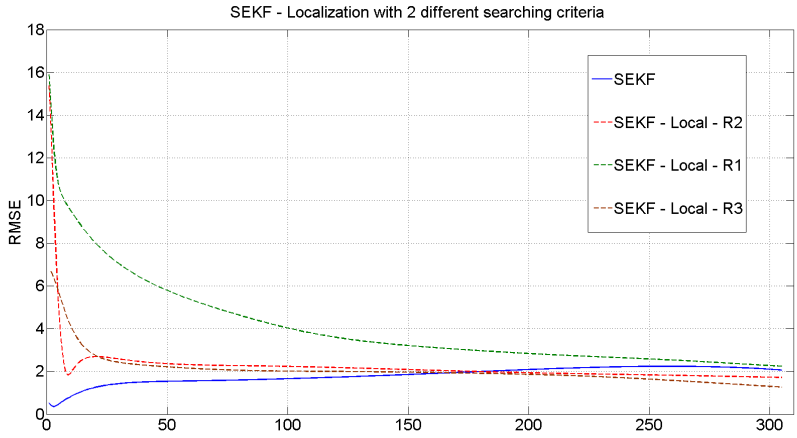


Figure 4-19: RMSE values when applying localization to the SEKF analysis with different influence areas.

The reason for the weak impact of localization on the filters performances is mainly due to the assimilation of "pesudo" observations in the present work. These observations are fully consistent with the model dynamics and the long-distance correlations summarized in the correction directions are likely to be correct. Another reason is that, as for the model states, the total variance of the observations can be also represented at very high accuracy with a very few modes. The direction of the filter correction subspace is therefore not an issue and the filter should be able to extract most of the information in the observations even with few correction directions. In these conditions, localization might not necessarily enhance the filters performances and the results of our experiments support our analysis.

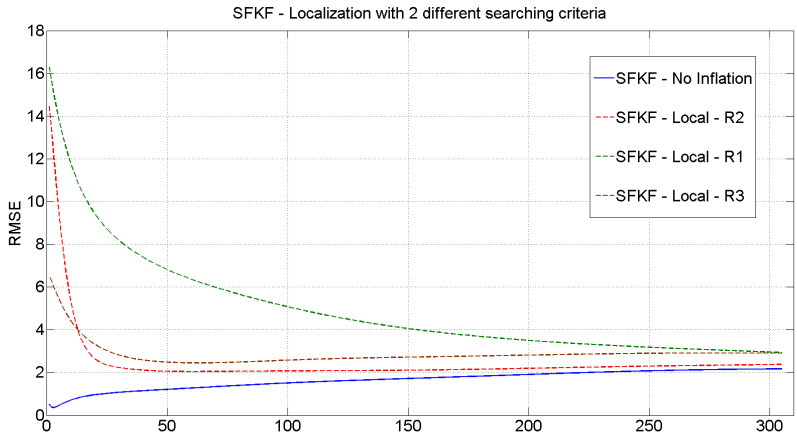


Figure 4-20: RMSE values when applying localization to the SFKF analysis with different influence areas.

# Chapter 5

# Joint State-Parameter Estimation with KF

In all the previous data assimilation discussion, we stated the problem where we use filtering procedures to estimate the model state variables. In most models, we face a problem where we need to estimate some model parameters together with the state variables. This is often encountered in very complicated models with inaccurate configurations and parameters. To solve this, we consider in this chapter the combined state-parameter estimation problem, in which both the model state and parameters are estimated simultaneously.

One approach for the combined estimation problem is given by the joint estimation where the states and the parameter vectors are added together in a single joint state vector; commonly referred to as state augmentation approach [5, 6, 37, 47, 48, 55]. In other words, the parameters are treated in a similar way just like the state variables. What differs these parameters from the normal state variables is the fact that they are not observed and they can be nonlinear function of the model even if the standard system is linear.

The joint state-parameter estimation problem is generally nonlinear and for this we use the EKF and the SEEKF assigned for nonlinear dynamic systems. It is essential to mention that for some systems where the non-linearity is strong, the system with EKF can become unstable. To deal with this drawback, generally the Ensemble Kalman filter (EnKF), based on Monte Carlo method, is used.

## 5.1 General State-Parameter Estimation with the KF

Based on the KF algorithm discussed in section 3.1.1, the model state in the forecast step is updated in time by the model operator as in equation (3.5). Using the augmented state-parameter estimation approach, the update of the model state will be represented in a different way because the state vector now includes some model parameters. An important point to remember is that, the parameters are time invariants and so the model operator will not project them forward in time as the state variables.

Assume that the parameter vector is denoted as $Z$, then the forecast equation (3.5) of the model state in the KF algorithm will be splitted into 2 equations as follows

$$X^f(t_k) = M(t_k, t_{k-1})\, X^a(t_{k-1}), \tag{5.1}$$

$$Z^f(t_k) = Z^a(t_{k-1}). \tag{5.2}$$

This configuration is then joined in a single equation

$$\widetilde{X}^f(t_k) = \widetilde{M}(t_k, t_{k-1})\, \widetilde{X}^a(t_{k-1}), \tag{5.3}$$

where $\widetilde{X}$ and $\widetilde{M}$ correspond to the state and the model operator of the joint system approach and can be written as

$$\widetilde{X} = \begin{bmatrix} X \\ Z \end{bmatrix}, \tag{5.4}$$

$$\widetilde{M} = \begin{bmatrix} M & 0 \\ 0 & I \end{bmatrix}, \tag{5.5}$$

where $I$ is an identity matrix having the same size as the parameter vector $Z$.

## 5.2 Joint State Estimation of the Contaminant System

In the case when imperfect forecast model is used, the model was perturbed using inaccurate values of the permeabilities. In this section, we try to estimate the 2 rock permeabilities alongside of the contaminant state using the state parameter approach. This is a challenging

problem as the information about the permeabilities come from observing the contaminant concentration not the flow. To do so, we add the two permeabilities to the state vector we are estimating making its dimension $N+2$. Since the contaminant model operator $A$ in (2.32) is a nonlinear function of the parameters, the estimation problem becomes nonlinear. In this case, we use the EKF and the SEEKF in which the model is linearized around the previous state estimate.

In the EKF and the SEEKF setups, based on the previous section, the model state $X^a(t_{k-1})$ and the model operator $M(t_k, t_{k-1})$ in equation (3.21) become

$$X^a(t_{k-1}) = \begin{bmatrix} C_{1(t_{k-1})} \\ C_{2(t_{k-1})} \\ \vdots \\ C_{n(t_{k-1})} \\ k_{1(t_{k-1})} \\ k_{2(t_{k-1})} \end{bmatrix}_{(N+2)\times 1}, \tag{5.6}$$

$$M(t_k, t_{k-1}) = \begin{bmatrix} & & 0 & 0 \\ \left[A\left(k_{1(t_{k-1})}, k_{2(t_{k-1})}\right)\right]_{N\times N} & & \vdots & \vdots \\ & & 0 & 0 \\ 0 & \dots & 1 & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix}_{(N+2)\times(N+2)}, \tag{5.7}$$

where $C_{1(t_{k-1})}$ represents the concentration of the contaminant of the first grid point at $t_{k-1}$. $k_{1(t_{k-1})}$ and $k_{2(t_{k-1})}$ correspond to the estimated rock permeabilities at $t_{k-1}$.

The observation operator $H$ will be exactly the same as before augmented by two additional zero columns after the $N^{th}$ column. The reason for this, is that we are only observing the concentration of the contaminant, and the permeability by itself is something intangible, meaning that it can not be observed directly. $H$ is therefore a linear operator and no linearization is required so the gradient of $H$ is the same matrix as $H$.

The gradient of $M\left(t_k, t_{k-1}\right)$ in (3.22) includes some derivative as follows

$$\mathbf{M}\left(t_k, t_{k-1}\right) = \begin{bmatrix} A\left(k_{1(t_{k-1})}, k_{2(t_{k-1})}\right) & \left(\frac{\partial A}{\partial k_{1(t_{k-1})}}\right) C\left(t_{k-1}\right) & \left(\frac{\partial A}{\partial k_{2(t_{k-1})}}\right) C\left(t_{k-1}\right) \\ 0 \quad \ldots \quad 0 & 1 & 0 \\ 0 \quad \ldots \quad 0 & 0 & 1 \end{bmatrix}, \tag{5.8}$$

where $C\left(t_{k-1}\right)$ denotes the contaminant vector as

$$C\left(t_{k-1}\right) = \begin{bmatrix} C_{1(t_{k-1})} \\ C_{2(t_{k-1})} \\ \vdots \\ C_{n(t_{k-1})} \end{bmatrix}. \tag{5.9}$$

Here the derivatives are approximated using a second order centered finite difference method as

$$\left(\frac{\partial A}{\partial k_{1(t_{k-1})}}\right)_{k_{2(t_{k-1})}} \approx \frac{A\left(k_{1(t_{k-1})} + \varepsilon\right) - A\left(k_{1(t_{k-1})} - \varepsilon\right)}{2\varepsilon} + \vartheta\left(\varepsilon^2\right), \tag{5.10}$$

$$\left(\frac{\partial A}{\partial k_{2(t_{k-1})}}\right)_{k_{1(t_{k-1})}} \approx \frac{A\left(k_{2(t_{k-1})} + \xi\right) - A\left(k_{2(t_{k-1})} - \xi\right)}{2\xi} + \vartheta\left(\xi^2\right), \tag{5.11}$$

for some small values $\varepsilon$ and $\xi$.

The covariance matrix $P^a\left(t_{k-1}\right)$ in (3.22) must include the variances that account for the uncertainties in these permeabilities together with the covariance matrix of the state vectors as follows

$$\widetilde{P}^a\left(t_{k-1}\right) = \begin{bmatrix} & & & 0 & 0 \\ & P^a\left(t_{k-1}\right) & & \vdots & \vdots \\ & & & 0 & 0 \\ 0 & \ldots & 0 & var_1 & 0 \\ 0 & \ldots & 0 & 0 & var_2 \end{bmatrix}_{(N+2)\times(N+2)}, \tag{5.12}$$

where $P^a\left(t_{k-1}\right)$ stands for the error covariance of the state vectors having a size of $N \times N$, $var_1$ and $var_2$ are initial estimates of the variances of $k_1$ and $k_2$ respectively. These are assumed to be 100 for $var_1$ and 0.01 for $var_2$. These values are just rough estimations,

so we can adjust them depending on the change of the poor permeability estimates in the filter.

In the SEEKF, the error covariance matrix $P^a(t_{k-1})$ is decomposed into $L_{k-1}U_{k-1}L_{k-1}^T$ then $\widetilde{P}^a(t_{k-1})$ can be decomposed as

$$\widetilde{L}_k = \begin{bmatrix} & & & 0 & 0 \\ & L_{k-1} & & \vdots & \vdots \\ & & & 0 & 0 \\ 0 & \dots & 0 & 1 & 0 \\ 0 & \dots & 0 & 0 & 1 \end{bmatrix}_{(n+2)\times(r+2)}, \tag{5.13}$$

$$\widetilde{U}_k = \begin{bmatrix} & & & 0 & 0 \\ & U_{k-1} & & \vdots & \vdots \\ & & & 0 & 0 \\ 0 & \dots & 0 & var_1 & 0 \\ 0 & \dots & 0 & 0 & var_2 \end{bmatrix}_{(r+2)\times(r+2)}, \tag{5.14}$$

where $L_{k-1}$ and $U_{k-1}$ are given by the EOF analysis and the index $r$ denotes the rank of the SEEK filter. One can then apply the SEEKF without any changes using $\widetilde{L}_k$ and $\widetilde{U}_k$ in the filter's algorithm.

## 5.3 Estimation of the Aquifer Permeabilities

Just as the imperfect forecast model case used in the previous experiments, we impose 20% perturbation on the permeabilities, and we collect the same observations from the CMG as in figure 4-3. The initial contaminant state is the same as the one shown for the low resolution grid in figure 4-4. As mentioned in the previous section, since the model is nonlinear we use the EKF and after linearization of the system. Achieving this is not an easy job because computationaly it is very expensive and the reason is that we need to update our model operator after each iteration around the estimates of $k1$ and $k2$. This is in addition to the high computational cost required by the EKF itself. Using the SEEKF

solves the problem associated with the model integrations but it cannot avoid the important computational time needed to update the model operator at each iteration.

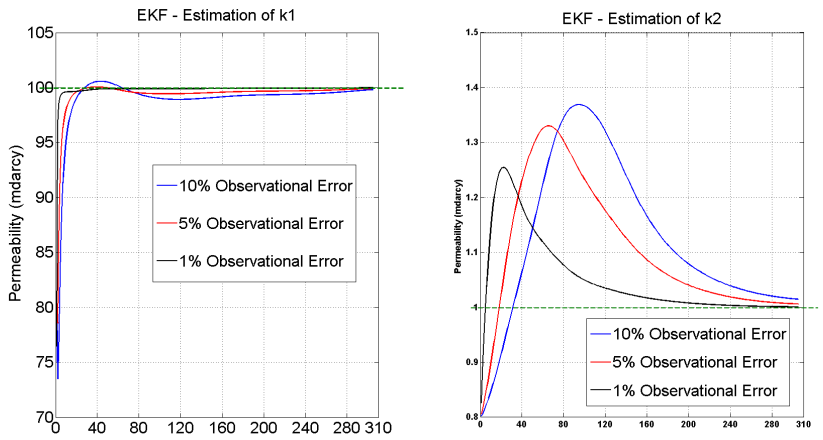


Figure 5-1: Permeability estimation using EKF with different observational errors.

We conducted with the EKF 3 experiments using 10%, 5%, and 1% errors on the observations (Figure 5-1). In the 3 experiments we were able to estimate back the true permeabilities with a very small error. The 2 plots also show how the estimates become more accurate as we decrease the error on the observations. The only question that these 2 plots may bring to us is the different estimation behavior for each permeability. Estimating $k1$ looks faster and more stable than the estimation of $k2$. We can only understand this estimation behavior if we go back and look at the size of the rocks for each permeability. From figure 2-1, we see that $k1$ is assigned for the large rock occupying 75% of the whole domain whereas, $k2$ comes from the small embedded rock occupying the remaining 25% of the aquifer. This fact makes it harder for the EKF to estimate $k2$ in a fast way as $k1$ because most of the grid points are located in the large rock where the flow is taking place.

Then, we applied the SEEKF and we compared it with the EKF. The same idea like before, we see that the more EOFs we use the more accurate solutions we get (Figure 5-2). Based on the large variances we assigned for each permeability, we were not able to decrease

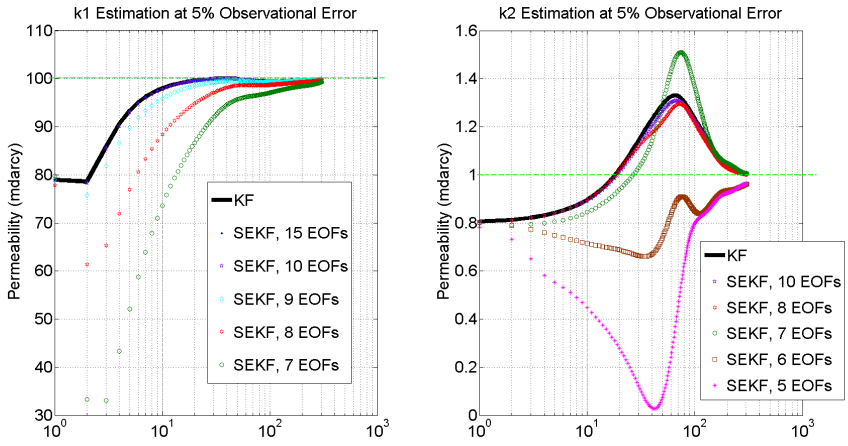


Figure 5-2: Permeability estimation using the SEEK filter using different EOFs.

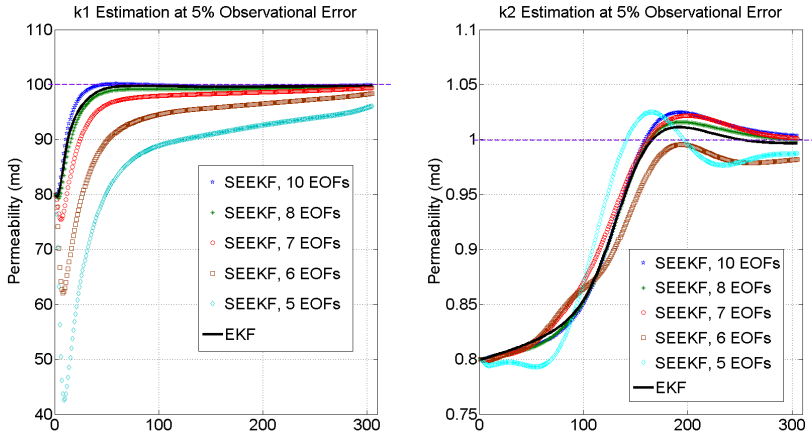


Figure 5-3: Permeability estimation using the SEEK filter with different EOFs and smaller variances.

the number of EOFs more than 7 when estimating $k1$. We noticed that when the variance is large, estimating the permeability becomes harder and at some point the permeabilities can take negative values which is wrong. If we decrease the variance for both permeabilities by taking $\sigma_1^2 = 1$ and $\sigma_2^2 = 10^{-4}$, the estimation gets better and the large changes decrease (Figure 5-3).

# Chapter 6

# Conclusions and Discussions

In this thesis, we studied some applications of a low-rank Kalman filter on subsurface models. This low rank KF is a new filtering technique that have never been used before in subsurface contaminant and flow models. It has the same mathematical equations like the KF including some approximations depending on the rank $r$ of the filter. The main purpose for using this low rank idea is to get fast and trustful estimations by reducing the expensive computational cost of the KF.

We used a coupled model incorporating both flow and contaminant information. We ran assimilation experiments and our objective was to locate the contaminant plume in the 2D domain correctly after some period of time. We used mainly a perfect and imperfect models for assimilation. In the perfect model case, the free run results were not good when compared to the filters estimates, and this shows the importance of data assimilation in improving the overall estimation. The SEKF and the SFKF estimations require less time and memory than the KF; moreover, we found that the SEKF gives less prediction errors when considering only 10 EOFs. If we look at Figure 6-2, we see that towards the end of the simulation the best estimate comes from the SEKF not the KF. So, this tells us that we did not only reduce the computational cost but also estimated our contaminant state more accurately. The physical interpretation for this result arises from the eigenvalues of the large covariance matrix $P$ in the KF. By definition, this matrix is symmetric positive definite; it comes from the product of the contaminant state and the transpose of it as shown in equation (4.2). So, we expect the eigenvalues for this matrix to be all positive, but due to some numerical errors we still have very small negative eigenvalues. In figure

6-1, we plot all the $2,500$ eigenvalues in 2 semi-log plots and it appears that almost half of the eigenvalues are greater than zero while the others are less than zero. These negative eigenvalues can introduce noise to the system and mislead the KF. Since we decrease the rank in the SEKF, we ignore these small eigenvalues and thus the SEKF can filter out all this noise resulting in slightly better estimates.

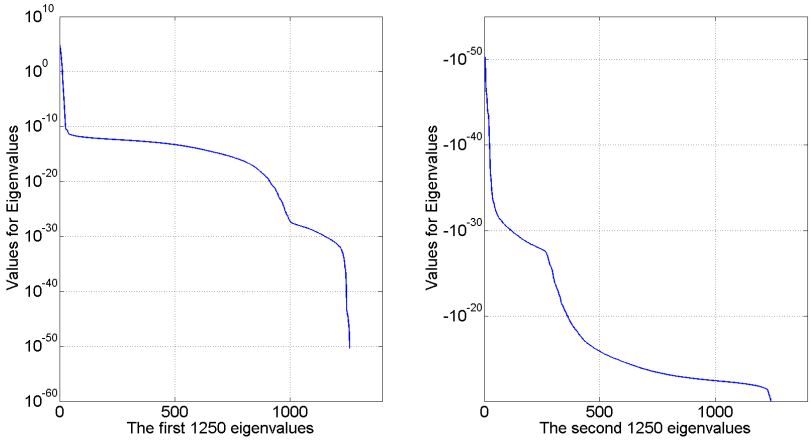


Figure 6-1: Plot of all eigenvalues of the covariance matrix in KF.

The SEKF was also a better choice than the KF for large models. When we increased the resolution of the model, the KF failed to give estimations; nevertheless, the SEKF did. In SEKF, we use some correction directions and the dimension of the problem decreases drastically; however, in KF the dimension is still large and implementing it was not possible.

In the imperfect model case, we used the process noise covariance $Q$ and inflation to get better estimates. SEKF and SFKF with $Q$ provided better results than with inflation. For the KF, estimates with inflation were slightly better than with $Q$. In Figure 6-3, we plot the best estimates as in the perfect model. In here as well, the low rank filter was better than the full Kalman because of the noise coming from the negative eigenvalues and the model uncertainties.

In both models, the SFKF was less accurate than the SEKF and the KF. The RMS

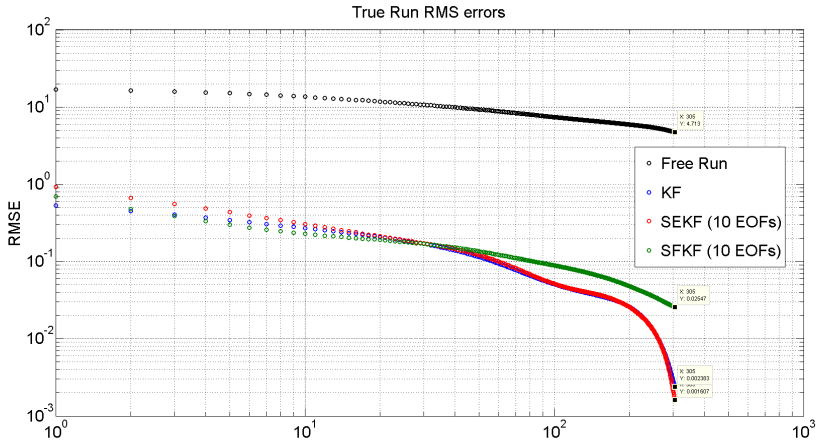


Figure 6-2: All estimates for the true model.

error for this filter was not too bad, still acceptable especially if we are dealing with very large dimensions such as the atmosphere and the ocean.

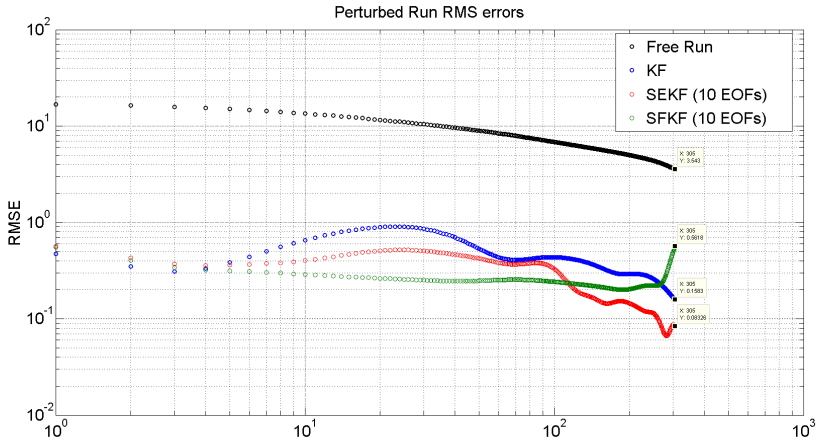


Figure 6-3: All estimates for the perturbed mode.

Next, we continued with the perturbed model and applied localization for the low rank filter analysis aiming to reduce the RMS errors more. The results were very sensitive to the chosen searching area; small influence areas did not help the overall estimation because the

majority of the grid points did not have observations close to them. For large areas, the RMSE improved a little bit for the SEKF, but not for the SFKF. Generally, localization is better applied on very large domains where there are more observations, and the data varies extremely between different parts of the domain. In such cases, it is really important to correlate the grid point just with the data around it, rather than including very far points.

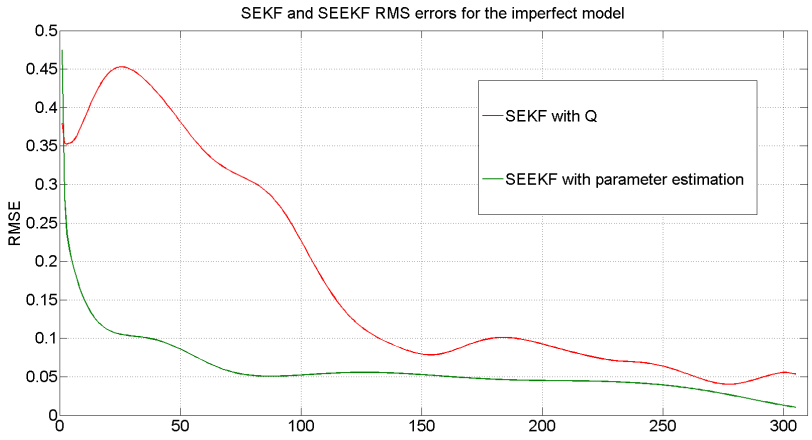


Figure 6-4: Comparison between the the RMSE of the SEKF with $Q$ and the SEEKF from state-parameter estimation at 5% observational errors.

Finally, we apply joint state-parameter estimation using the imperfect model to estimate the true permeabilities of the aquifer. The recovered permeabilities were very close to the true ones using both the EKF and the SEEKF with an error not exceeding 1%. In terms of the computational effort, the SEEKF is faster and as we include more EOFs it becomes more accurate.

We saw that the process noise covariance can account to the model uncertainties and in the state-parameter estimation problem we removed the uncertainties from the system. It is good to compare the RMSE with $Q$ from the SEKF and the RMSE from the SEEKF. Figure 6-4 shows the RMSE values from the filters with 5% observational errors. We notice that the SEEKF is more accurate than the SEKF meaning that we were able to beat the

SEKF with $Q$ by recovering the true permeabilities. What we did is that we removed the uncertainties from the model and in time the estimates were better than the other case where we leave the uncertainties and account for them with $Q$.

## 6.1 Future Work

This work is a first step towards a full comprehensive study of developing efficient data assimilation tools for improving the accuracy of the underground contaminant models. Throughout the progress of this work, interesting ideas for future work came up. Other approaches to treat the model errors exist, but were not all explored in the present study. These need to be implemented and evaluated against the approaches that are considered in this study. Another ideas could be to extend the complexities of the contaminant model by including some dispersion and adsorption terms with external water sources. In this case the system will be nonlinear function of the states and EnKF-like methods might be performed. It would be also interesting to look at compositional flow models that are essential and widely used in reservoir simulations. Working on multiple phase flows in porous media might be considered as well, because in such models the permeability becomes a function of phase saturation. For this case, new filtering schemes need to be developed to be able to estimate functions of more than one variable.

*"There is still too much to learn but as long as we are seeking for knowledge, we will be able to reach the heart of Science."* *Mohamad El Gharamti*

# Bibliography

[1] Ahn, H. (2000). "Ground water drought management by a feedforward control method." J. Am. Water Resour. Assoc., 36, 501-510.

[2] Astrom, K. J. and Wittenmark, B. (1989). "Adaptive Control. AddisonWiesley, Reading, MA."

[3] Bidwell, V. J. (1998). "State-space moxing cell model of unsteady solute transport in unsaturated soil." Environ. Modell. Software, 14, 161-169.

[4] Bowles, D. S., and Grenner, W. J. (1978). "Steady state river quality modeling by sequential extended Kalman filter." Water Resour. Res., 14, 84-96.

[5] Bras, R.L. and Restrepo-Posada, P. (1980). "Real time automatic parameter calibration in conceptual runoff forecasting models." In: Proceedings of the Third International Symposium on Stochastic Hydraulics, p. 61–7.

[6] Bras, R.L. and Rodriguez-Iturbe, I. (1985). "Random functions and hydrology." Reading, MA, USA: Addison Wesley; 559.

[7] Budhiraja, A., Chen, L., and Lee, C. (2007). "A survey of numerical methods for nonlinear filtering problems." Physica D, 230, 27-36.

[8] Chang, S. Y., and Latif, S. M. I. (2007). "Use of Kalman filtering and particle filtering in a one dimensional leachate transport model." Proc., 3rd National Conf. on Environmental Science and Technology, Springer, Greensboro, N.C.

[9] Chang, S. Y., and Latif, S. M. I. (2010). "Extended Kalman filtering to improve the accuracy of a subsurface contaminant transport model." J. Enviro. Eng., 466-474.

[10] Chang, S. Y., and Jin, A. (2005). "Kalman filtering with regional noise to improve accuracy of contaminant transport models." J. Environ. Res., 14, 84-96.

[11] Chen, Y., Oliver, D. S., and Zhang, D. (2008). "Efficient ensemble-based closed-loop production optimization." Proc., SPE Improved Oil Recovery Symp., Society of Petroleum Engineering, Tulsa, Okla.

[12] Chen, Y., Oliver, D. S., and Zhang, D. (2009). "Data assimilation for nonlinear problems by ensemble Kalman filter with reparametrization." J. Pet. Sci. Eng., 66, 1-14.

[13] 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, N.C.

[14] Cosby, B. J., Hornberger, G. M., and Kelly, M. G. (1984). "Identification of photosynthesis-lightmodels for aquatic systems II: Application to a macrophyte dominated stream." Ecol. Modell., 23, 25-51.

[15] Dawson, C., Sun, S., and Wheeler, M. F., (2004). “Compatible algorithms for coupled flow and transport”, Computer Methods in Applied Mechanics and Engineering, 193, 2565-2580.

[16] Dong, C., Sun, S., and Taylor, G. A., (2010). “Numerical modeling of contaminant transport in fractured porous media using mixed finite element and finite volume methods”, Journal of Porous Media.

[17] Ferraresi, M., and Marinelli, A. (1996). "An extended formulation of the integrated finite difference method for ground water flow and transport." J. Hydrol., 175, 453-471.

[18] Franssen, H. H., Kuhlman, U., Kaiser, H., Stauffer, F., and Kinzelbach, W. (2008). "The ensemble Kalman filter for real-time groundwater flow modeling of the upper Liammat aquifer in Zurich (Switzerland)." Geophysical Research Abstract, Copernicus Publications, Vienna, Vol. 10.

[19] Ghil, M. and P. Malanotte0Rizzoli, (1991). "Data assimilation in meteorology and oceanography." Adv. Geophys., 33, 141-266.

[20] Godfrey, J. T., and Foster, G. D. (1996). "Kalman filter method for estimating organic contaminant concentration in Major Chesapeake Bay Tributaries." Environ. Sci. Technol., 30, 2312-2317.

[21] Goegebeur, M., and Pauwels, V. R. N. (2007). "Improvement of the PEST parameter estimation algorithm through extended Kalman filtering." J. Hydrol., 337, 436-451.

[22] Graham, W., and McLaughlin, D. (1989). "Stochastic analysis of nonstationary subsurface solute transport 2. Conditional moments." Water Resour. Res., 25(11), 2331-2355.

[23] El Harrouni, K., Ouazar, D., Wrobel, L. C., and Cheng, A. H. D. (1997). "Aquifer parameter estimation by extended Kalman filtering and boundary elements." Eng. Anal. Boundary Elem., 23, 25-51.

[24] Hoteit, I., Pham, D., and Blum, J. (2001). "A semi-evolutive partially local filter for data assimilation." J. Elsevier Science Ltd.

[25] Hoteit, I., Pham, D., and Blum, J. (2002) "A simplified reduced order Kalman filtering and application to altimetric data assimilation in Tropical Pacific." J. Marine Sys. 36, 101-127.

[26] Hoteit, I., and Pham, D. (2003). "Evolution of the reduced state space and data assimilation schemes based on the Kalman filter." J. Metero. Sci. Japan, vol.81, No.1, 21-39.

[27] Hoteit, I., Triantafyllou, G., and Korres, G. (2007). "Using low-rank ensemble kalman filters for data assimilation with high dimensional imperfect models." J. of Numerical Analysis, vol. 2, no. 1-2.

[28] Houtekamer, P. L., and Mitchel, H. (2001). "A sequential ensemble Kalman filter for atmospheric data assimilation." Mon. Wea. Rev., 23, 1–15.

[29] Ide, K., A.F. Bennett, P. Courtier, M. Ghil and A.C. Lorenc, (1995). "Unified notation for data assimilation: operational, sequential and variational." J. Meteor. Soc. Japan, 75 (1B), 181-189.

[30] Liu, N., and Oliver, D. S. (2005). "Ensemble Kalman filter for automatic history matching of geologic facies." J. Pet. Sci. eng., 47, 147-161.

[31] McLaughlin, D. (2002). "An integrated approach to hrologic data assimilation: interpolation, smoothing and forecasting." Adv. Water Resour. 25, 1275-1286.

[32] Neal, J. C., Atkinson, P. M., and Hutton, C. W. (2007). "Flood inundation model updating using an ensemble Kalman filter and spatially distributed measurements." J. Hydrol., 336, 401-415.

[33] Pastres, R., Ciavatta, S., and Solidoro, C. (2003). "The extended Kalman filter (EKF) as a tool for the assimilation of high frequency water quality data." Ecol. Modell., 170, 227-235.

[34] Pham, D. T., Verron, J., and Rouband, M. C. (1998). "Singular evolutive Kalman filter with EOF initialization for data assimilation in oceanography." J. Marine Sys., 16, 323-340.

[35] Piatyszek, E., Voignier, P., and Graillot, D. (2000). "Fault detection on a sewer network by a combination of a Kalman filter and a binary sequential probability ratio test." J. Hydrol., 230, 258-268.

[36] Porter, D., Bruce, G., Jones, W., Huyakorn, P., Hamm, L., and Flach, G. (2000). "Data fusion modeling for groundwater systems." J. Contam. Hydrol., 42, 303-335.

[37] Reichle, R.H., McLaughlin, D., and Entekhabi, D. (2002). "Hydrologic data assimilation with the ensemble Kalman filter." Monthly Weather Rev;130:103–1.

[38] Saad, G. A. (2007). "Stochastic data assimilation with application to multi-phase flow and health monitoring problems." Doctoral dissertation, University of Southern California, Los Angeles.

[39] Schreider, S. Y., Yong, P. C., and Jakeman, A. J. (2001). "An application of the Kalman filtering technique for streamflow forecasting in the upper Murray basin." Math. Comput. Model. Dyn. Syst., 33, 733-743.

[40] Skaggs, T. H., and Mohanty, B. P. (1998). "Water table dynamics in tile-drained fields." Soil Sci. Soc. Am. J., 62, 1191-1196.

[41] Sun, S., and Wheeler, M. F., (2006). "Anisotropic and dynamic mesh adaptation for discontinuous Galerkin methods applied to reactive transport", Computer Methods in Applied Mechanics and Engineering, 195(25-28), 3382-3405.

[42] Sun, S., and Wheeler, M. F., (2006). "A posteriori error estimation and dynamic adaptivity for symmetric discontinuous Galerkin approximations of reactive transport problems", Computer Methods in Applied Mechanics and Engineering, 195, 632-652.

[43] Sun, S., and Wheeler, M. F., (2005). "Discontinuous Galerkin methods for coupled flow and reactive transport problems", Applied Numerical Mathematics, 52(2-3), 273-298.

[44] Sun, S., and Wheeler M. F., (2006). "Projections of velocity data for the compatibility with transport", Computer Methods in Applied Mechanics and Engineering, 195, 653-673.

[45] Sun, S., and Wheeler M. F., (2005). "Symmetric and nonsymmetric discontinuous Galerkin methods for reactive transport in porous media", SIAM Journal on Numerical Analysis, 43(1), 195-219.

[46] Sun, X., Jin, L., and Xiong, M. (2008). "Extended Kalman Filter for Estimation of Parameters in Nonlinear State-Space Models of Biochemical Networks." PLoS ONE 3(11): e3758. doi:10.1371/journal.pone.0003758.

[47] Todini, E., Szollosi-Nagy, A., and Wood, E.F. (1976). "Adaptive state–parameter estimation algorithm for real time hydrologic forecasting; a case study." In: IISA/WMO Workshop on the Recent Developments in Real Time Forecasting/Control of Water Resources Systems, Laxemburg (Austria).

[48] Todini, E. (1978) "Mutually interactive state–parameter (MISP) estimation, Applications of Kalman filters to hydrology, hydraulics and water resources." In: Chiu C-L, editor. Proceedings of AGU Chapman Conference.

[49] Triantafyllou, G., Korres, G., Hoteit, I., Petihakis, G., and Banks, A. C. (2006). "Assimilation of ocean colour data into a biochemical flux model of the eastern mediterranean sea." Ocean Sci. Discuss., 3, 1569-1608.

[50] Van Geer, F. C. (1982). "An equation based theoretical approach to network design for ground water levels using Kalman filters." Int. Assoc. Sci. Hydrol., 136, 241-250.

[51] Welch, G., and Bishop, G. (2006). "An introduction to the Kalman filter." UNC-Chapel Hill, TR 95-041.

[52] Wendroth, O., Rogasik, H., Koszinski, S., Ritsema, C. J., Dekker, L. W., and Nielsen, D. R. (1999). "State-space prediction of field-scale soil water content time series in a sandy loam." Soil Tillage Res., 50, 85-93.

[53] Whitehead, P. G., and Hornberger, G. M. (1984). "Modeling algal behavior in the river Thames." Water Res., 18(8), 945-953.

[54] Yangxiao, Z., Te Stroet, C. B. M., and Van Geer, F. C. (1991). "Using Kalman filtering to improve and quantify the uncertainty of numerical ground water simulation: Application to monitoring network design." Water Resour. Res., 27, 1995-2006.

[55] Young PC. Advances in real-time flood forecasting. Philos Trans Roy Soc Lond 2002;360:1433–50.

[56] Yu, Y.-S., Heidari, M., and Guang-Te, W. (1989). "Optimal estimation of contaminant transport in ground water." Water Resour. Bull., 25, 295-300.