2008

Continuous reservoir model updating by ensemble Kalman filter on Grid computing architectures

Xin Li

Louisiana State University and Agricultural and Mechanical College

Recommended Citation

Li, Xin, "Continuous reservoir model updating by ensemble Kalman filter on Grid computing architectures" (2008). *LSU Doctoral Dissertations*. 585.
https://repository.lsu.edu/gradschool_dissertations/585

CONTINUOUS RESERVOIR MODEL UPDATING BY ENSEMBLE KALMAN FILTER ON GRID COMPUTING ARCHITECTURES

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

Craft & Hawkins Department of Petroleum Engineering

by
Xin Li
B.S. in Petroleum Engineering, Southwest Petroleum University, 1991
M.S., Southwest Petroleum University, 1994
December 2008

Acknowledgments

I would like to express my sincere gratitude to my professors, friends, and family for their selfless support that made my PhD study an incredible achievement through an memorable journey. I feel fortunate to have each single one of them in my life.

First, I appreciate Dr. Christopher D. White, who is the best advisor and mentor one could ever have. He provided me a lot of brilliant ideas and helped me to dig deep insights in research, invaluable guidance and providing me with generous financial support during the past four years at LSU. I also appreciate his patience and encouragement during my difficult moments.

My dissertation committee members have offered valuable instructions, critiques and suggestions in the process of my PhD study, which are deeply appreciated. Dr. Gabrielle Allan introduced me into the world of Grid computing and gave my research a lot of valuable suggestions. Dr. Steven Sears taught me to bear the industrial applicability in mind when developing a work flow. Dr. John R. Smith gave me suggestions and comments to my research. I am grateful to all the PETE faculty members for offering high quality learning experience and making the school a wonderful place to start a career in petroleum engineering.

I thank Dr. Lei Zhou, the research of the Center for Computation and Technology (CCT). His creative work and original middleware code have greatly contributed to the UCoMS project (Ubiquitous Computing and Monitoring System for Discovery and Management of Energy Resources). I would like to extend my thanks to all the other members who have worked for UCoMS: Dayong Huang (Grid data archive), Chongjie Zhang (Grid portal), Promita Chakraborty, John Lewis, Archit Kulshrestha (Grid testbed), and Yaakoub El-Khamra. I also thank the Grid testbeds provided by CCT. I extend thanks to the Department of Energy and Louisiana Board of Regents for providing sponsorship (award No. DE-FG02-04ER46136) that made this research possible.

Licences for the ECLIPSE$^{\text{TM}}$Black Oil Reservoir Simulator were provided by Schlumberger. I thank iReservoir.Inc for providing me internship opportunities from which I gained deep insight to the application of reservoir simulation and automatic history matching methods. During my internship, Mr. James Gilman was a wonderful mentor who taught me a great deal of practical knowledge of reservoir simulation.

My graduate study would not be as enjoyable without all my wonderful colleagues and dear friends: Subhash Kalla, Bobby Kurniawan, Hong Tang, Feng Wang, and many others. They provided encouragement and support throughout my graduate study. I would also like to thank them all for the splendid shared moments.

I am thankful to my husband Shengkai Duan and my daughter Cindy Duan for their unconditional love, and support over the years.

This work is dedicated to my parents, Shilun Li and Yongming Wu, for their guidance and unconditional love.

Table of Contents

List of Tables

List of Figures

Abstract

A reservoir engineering Grid computing toolkit, ResGrid and its extensions, were developed and applied to designed reservoir simulation studies and continuous reservoir model updating. The toolkit provides reservoir engineers with high performance computing capacity to complete their projects without requiring them to delve into Grid resource heterogeneity, security certification, or network protocols.

Continuous and real-time reservoir model updating is an important component of closed-loop model-based reservoir management. The method must rapidly and continuously update reservoir models by assimilating production data, so that the performance predictions and the associated uncertainty are up-to-date for optimization. The ensemble Kalman filter (EnKF), a Bayesian approach for model updating, uses Monte Carlo statistics for fusing observation data with forecasts from simulations to estimate a range of plausible models. The ensemble of updated models can be used for uncertainty forecasting or optimization.

Grid environments aggregate geographically distributed, heterogeneous resources. Their virtual architecture can handle many large parallel simulation runs, and is thus well suited to solving model-based reservoir management problems. In the study, the ResGrid workflow for Grid-based designed reservoir simulation and an adapted workflow provide tools for building prior model ensembles, task farming and execution, extracting simulator output results, implementing the EnKF, and using a web portal for invoking those scripts.

The ResGrid workflow is demonstrated for a geostatistical study of 3-D displacements in heterogeneous reservoirs. A suite of 1920 simulations assesses the effects of geostatistical methods and model parameters. Multiple runs are simultaneously executed using parallel Grid computing. Flow response analyses indicate that efficient, widely-used sequential geostatistical simulation methods may overestimate flow response variability when compared to more rigorous but computationally costly direct methods.

Although the EnKF has attracted great interest in reservoir engineering, some aspects of the EnKF remain poorly understood, and are explored in the dissertation. First, guidelines are offered to select data assimilation intervals. Second, an adaptive covariance inflation method is shown to be effective to stabilize the EnKF. Third, we show that simple truncation can correct negative effects of nonlinearity and non-Gaussianity as effectively as more complex and expensive reparameterization methods.

Chapter 1
Introduction

1.1 Significance of the Research

This research uses the UCoMS (Ubiquitous Computing and Monitoring System for Discovery and Management of Energy Resources) grid for reservoir characterization, design of experiments, and model inversion through continuous reservoir model updating.

1.1.1 Closed Loop Management

Closed-loop model-based reservoir management (Jansen et al. 2005; Figure 1.1) allows real-time decisions to be made that maximize the production potential from a reservoir. These decisions are based on the most current information available from the reservoir models. Essential elements of closed-loop reservoir management are model-based optimisation and data assimilation techniques (automatic history matching). In addition, techniques for model reduction and uncertainty assessment may play a role. The closed-loop model-based reservoir management, also called "smart fields", "e-field" or "digital oil field" technology, it involves the use of reservoir system models in a closed-loop fashion. This dissertation will focus on data assimilation process. Data assimilation is the process that rapidly completes reservoir model inversion and uncertainty estimation conditioned on the measured observations because accurate real-time model calibration may greatly improve predictions and optimization results.

1.1.2 Model Updating Process

Conditioning static geophysical and geological data to sequential production observations to infer more accurate actual estimates of poorly known reservoir model parameters (*e.g.*, k and ϕ for all gridblocks) is termed model updating. Model updating is an important component

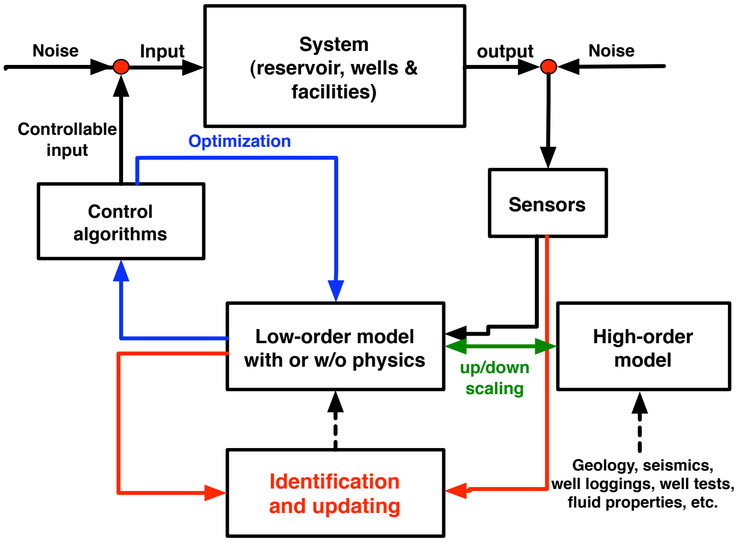


Figure 1.1: Closed-loop work flow for real-time reservoir management (Jansen et al. 2005)

of managing production policies according to the "closed-loop" concept. The term "model updating" emphasizes the sequential nature of the process and implies its Bayesian roots. Related terms for such processes include "history matching," which emphasizes adjustments to reconcile production measurements and model predictions; and "inversion," which is a more general mathematical term (commonly used in geophysics).

In oil and gas fields, increased use of permanent gauges or distributed downhole sensors (such as optical fiber sensors and 4D seismic technology; Brown and Hartog 2002; Manin, Charara, and Delhomme 2002; Haddad, Proano, and Patel 2004; Marschall and Sherlock 2002) for monitoring bottom-hole pressure, temperature, resistivity, flow rate measurement, and flow controls has added impetus to the need for continuous model updating. This move to greater instrumentation is sometimes referred to as using "smart" or "intelligent" well technology (Addiego-Guevara, Jackson, and Giddins 2008). Geoscientists and engineers wish to incorporate these data as soon as they are obtained so that the reservoir model is always up to date, and more useful for decision making. Brouwer et al. (2001) combined the data assimilation and optimization algorithms to decision making of a water flooding case. The increase in recovery obtained varied between 0-20 percent. The delay in breakthrough time achieved by the routine varied from 7-168 percent.

However, the field application of true data assimilation is in its infancy, reservoir engineers have commonly used trial-and-error to find model parameters to reduce the mismatch between actual measurements and model simulation responses. The matching techniques are usually ad-hoc and not focused on the balanced use of responses from various sources. Efficient and reliable techniques for data assimilation are required.

1.1.3 From Current to Emerging Methods: the Ensemble Kalman Filter

Although the history matching process has been more automated, a large programming and computational effort is still required, either in objective function evaluation (nongradient-based minimization methods), or in gradient computation (gradient-based minimization methods). If gradient-based minimization methods are used, the adjoint method is the most efficient way to compute the gradient of a defined objective function (Li, Reynolds, and Oliver 2003). The adjoint system requires modifying the source code of the reservoir simulator, which is time-consuming and sometimes is impossible when a commercial simulator is used. Furthermore, if we choose a different simulator, the adjoint calculations must be recoded as well. Finally, uncertainty estimation with gradient methods requires multiple history matches.

Hence the heavy programming burden, high data-sampling frequency, noise in data, and uncertainty forecasting requirement have motivated reservoir engineers to use the ensemble Kalman filter (EnKF) in closed-loop reservoir management.

EnKF has gained popularity for reservoir monitoring and continuous model updating because of its simple formulation, the ability to account for the possible model noise and error, and the relative ease of implementation for any simulators. It requires no derivation of a tangent linear operator (Tarantola 1997) or adjoint equations (and no integrations backward

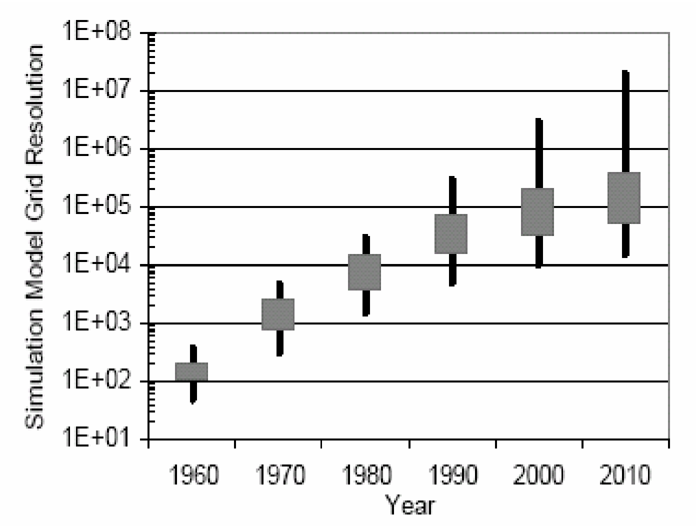


Figure 1.2: Trend of simulation model gridblock resolution for three decades (Fjerstad et al. 2005).

in time, as required by adjoints). This supports efficient uncertainty assessment and diverse data integration. However, the EnKF still requires many runs — an ensemble — to provide stable estimates of covariances.

1.2 Problem Statement

Two kinds of problems associated with EnKF are discussed in the following sections.

1.2.1 Computation Problems

Increases in model gridblock numbers is clearly linked with increasing availability of computation resources over the last 30 years (Figure 1.2). Early coarse models and limited numbers of runs were based on large mainframes. Beginning in the late 80s, engineers could access workstations, and began to use refined grid models, which were more representative of the subsurface geological model. More recently, the evolution of workstations to clusters greatly expanded capabilities. More computation resources and data storage spaces are needed for simulation studies to overcome the limitations in problem size and memory, enabling geologists and reservoir engineers to include the more realistic geologic and engineering details needed for better production optimization. At present, the gridblocks of models used in the oil and gas industry are around $10^5 - 10^6$ and sometimes reach 10^7. Therefore, reservoir engineers need a versatile tool to study reservoir performance and make real-time operational decisions for their reservoir with low computation costs and less run time.

Grid computing technology assembles widely distributed, dynamic, and heterogeneous resources into a virtual organization, and has been viewed as a promising platform for large-scale scientific applications. It decomposes the computation problem, and then distributes its components across a set of computational resources, each to run on the most suitable computing equipment, database server, storage server, or other specialized device. In addition to hardware, the grid encompasses software architectures for parallel computing, communications protocols, scheduling, security and policy mechanisms (Allen et al. 2005). Grid computing is well suited for designed reservoir simulation studies and continuous model updating with EnKF because the independent runs from designs or ensembles are easy to distribute in grid environments which allows parallelism.

However, when multiple distributed computers are used to perform the task, the synchronization of simulations for all ensemble members at data forecast steps affects computation efficiency. Synchronization refers to the requirement that all ensemble member simulations must run to the assimilation point in the same runtime to avoid wasted waiting time on the processors. The diverse reservoir models have different runtimes, which can make synchronization process difficult and increase the overhead time. Grid environments may impose challenges such as multiple queue times and system shut-downs or reboots, further complicating synchronization. In addition, data transfer on Grid imposes an additional burden for network because of large data sets comprising models, Kalman gain and state vector. The more complex reservoir geological models are, the greater the data transferring problem will be.

Many efforts in grid communities have focused on middleware research and development. However, grid application-level tools are needed which can build higher-level functionalities on top of core middleware services. It is useful to provide reservoir engineers with a grid-enabled framework for reservoir simulation studies using experimental design and continuous reservoir model updating.

1.2.2 EnKF Problems

Despite progress on the EnKF, ensemble forecasting is not yet used to its full potential in reservoir management. The EnKF is still a relatively new endeavor and there are many problems which need to be solved. First, "filter divergence" is one of potential problems with the EnKF. It causes an increasing error in model parameter estimate as the EnKF is used to sequentially assimilate observed measurements. Then, the EnKF becomes unstable after a number of assimilation steps and leads to a useless forecasting ensemble. The possible causes, effects and solutions of "divergent filters" phenomena have not been systematically investigated before in continuous reservoir model updating application. Secondly, in the presence of strong nonliearity and non-Gaussianities of state vectors, the EnKF can generate nonphysical model variables which do not honor the future observations. Which method is more efficient, keeps the simplicity of EnKF and is reasonably accurate to solve the problem, EnKF with reparameterization or EnKF with truncation (which is more efficient but not as well proven before?) The performance comparison between the two methods is the best guide for reservoir engineers. Finally, it is important to seek ways to provide answers to one critical question: what is the optimal frequency for the updating of reservoir models?

1.3 Literature Review

Automatic history matching (model inversion/continuous reservoir model updating) in reservoir simulation is an ill-posed problem with non-unique solutions, because of nonlinearity and the large number of model parameters. The problem of reservoir characterization through automatic history matching has been extensively studied since the 1970s. Among existing history matching algorithms, gradient-based methods and ensemble Kalman filters are two important categories. In the following Sections 1.3.1 and 1.3.2, the development and current research status of each algorithm are introduced.

1.3.1 Gradient-Based Methods

Adjoints (Chen et al. 1974) have been used for history match and applied to one-phase flow problems. The method has been modified and improved for multiphase history matching (Wasserman, Emanuel, and Seinfeld 1974; Watson et al. 1979; Lee and Seinfeld 1987; Zhang and Reynolds 2002). These investigators used adjoint models to calculate the gradient of an objective function with the least square error between calculated and observed data, then applied first order gradient-based optimization algorithms to perform the minimization. This is one of the most efficient methods available to solve the history matching problem.

Another gradient-based approach is the sensitivity coefficient method, in which gradients of all observed data with respect to the model parameters (termed sensitivity coefficients) are calculated. This can then be used to calculate the Hessian for use with more efficient second order optimization algorithms such as the Gauss-Newton method. This approach was proposed by Carter et al. (1974) for single-phase problems and later extended in a computationally efficient way for 3D problems by He, Reynolds, and Oliver (1996). Although this approach is efficient, it cannot be used to calculate the sensitivity coefficients for multiphase flow problems. A more general approach is the use of the gradient simulator to calculate the sensitivity coefficients. Procedures of this type have been described by many authors (Anterion, Karcher, and Eymard 1989; Bissel 1994; Landa and Horne 1997; Wu, Reynolds, and Oliver 1999; Wu and Datta-Gupta 2002; Li, Reynolds, and Oliver 2003). Unfortunately, this approach is complex and difficult for large-scale simulation problems as it is expensive to compute sensitivity coefficients when the number of observed data or model parameters exceeds a few hundred.

1.3.2 Ensemble Kalman Filter

The Kalman filter was originally developed to continuously update the states of linear systems to honor the available noisy measurements (Jazwinski 1970). For highly nonlinear models, the ensemble Kalman filter (EnKF) was introduced by Evensen (1994). EnKF uses an ensemble of models from which parameter uncertainties (*i.e.*, correlation between model parameters and responses) can be directly computed. Interest in data assimilation methods using ensembles of prediction models is growing in weather forecasting, oceanography, and hydrology (Houtekamer and Mitchell 1998; Reichle, McLaughlin, and Entekhabi 2002; Margulis et al. 2002). Researchers are interested in characterizing more information about the probability distribution of their dynamic system than can be revealed by a single assimilated state estimate. In these applications, only dynamic state variables are updated. However, in automatic history matching application, both static and dynamic parameters are simultaneously tuned to assimilate new measurements. In petroleum engineering, the method was first applied to well flow modeling (Lorentzen et al. 2001). Navdal, Mannseth, and Vefring (2002) used the EnKF to update the permeability in a near-well model. This approach was later developed to update 2D three-phase reservoir models by continuously adjusting the permeability, saturation and pressure fields at each assimilation step (Navdal et al. 2005). In the last case, the assimilation step was at least once a month and when new wells started production or wells were shut in. Gu and Oliver (2005, 2006) used EnKF to tune porosity, permeability, pressure and saturation in a widely-studied reservoir test case, the PUNQ-S3 model. Furthermore, Brouwer et al. (2004) combine EnKF for continuous model updating with an automated adjoint-based waterflood optimization to optimize waterflooding strategy. Liu and Oliver (2005) applied EnKF to facies estimation in reservoir models. This is a highly nonlinear problem where the distributions of petrophysical properties are multimodal. Results from their previous studies have shown that the EnKF can be efficient and robust. Wen and Chen (2005a, 2005b) have added a conforming step to EnKF to ensure that the updated dynamic and static variables are always consistent. However, the conforming EnKF fails to provide a correct estimates of porosity and permeability because of mismatches between model parameters and state solution used to reinitialize the reservoir flow equations.

To reduce uncertainty in estimation, Skjervheim et al. (2005) and Dong and Oliver (2005) used the EnKF to assimilate seismic 4D data and production data to provide dense information across whole field. A comparison of EnKF with Randomized Maximum Likelihood (RML) within a Bayesian framework show both of them give a reasonable quantification of the uncertainty in performance predictions (Gao, Zafari, and Reynolds 2006). Evensen et al. (2007) used EnKF for model inversion of a North Sea reservoir and successfully adjusted not only porosity and permeability, but also fluid contacts, vertical transmissivity multipliers and fault transmissibility multipliers. Li and Reynolds (2007) presented the iterated ensemble Kalman filter (IEnKF) update used when standard EnKF method fails to give an adequate data match, which is an application of the Gauss-Newton method for approximating a maximum likelihood estimate. First iterative EnKF (IEnKF(1)) requires an adjoint solution back to time zero which is identical to Randomized Maximum Likelihood (RML) using LBFGS (limited memory Broyden-Fletcher-Gikdfarb-Shanno) to minimize the appropriate objective functions (Zhang and Reynolds 2002; Gao, Zafari, and Reynolds 2006). In order to save computational time, the third iterative EnKF (IEnKF(3)) is derived by only requiring an adjoint solution from the current data assimilation time to the prior data assimilation time. The difference of IEnKF(1) and IEnKF(3) are: IEnKF(1) matches all the observation up to the current assimilation time but IEnKF(3) matches the observation measurements at current assimilation step; IEnKF(1) requires far more computational time than the standard EnKF method and IEnKF(3) is highly efficient iterative method; IEnKF(3) gives better future performance predictions than standard EnKF but far less accurate than IEnKF(1).

In recent years, various methods have been proposed to combine the updated reservoir model(s) with optimization control theory to determine optimal operating conditions to maximize hydrocarbon production or net present value (NPV) for the remaining expected life of the reservoir. Brouwer et al. (2004) used an adjoint method for optimization and the EnKF for model updating. Sarma et al. (2005) used an adjoint method for both the model updating and the production optimization. Wang, Li, and Reynolds (2007) accomplished closed-loop reservoir management with EnKF for data assimilation and three methods for production optimization. They concluded that the steepest ascent method with gradients provided by numerical perturbation gave better results than those with gradients provided by either an ensemble or simultaneous perturbation stochastic approximation. Lorentzen et al. (2006) directly adopted the EnKF method to optimize choke settings. They utilized the sensitivity approximated by the ensemble but did not make the process clear and the use of a preset upper limit might need more justification. Chen, Oliver, and Zhang (2008) presented an ensemble-based optimization (EnOpt) method to improve direction searching in the production optimization and maximize the expected net present value mean for an ensemble instead of a single reservoir model.

Computational efficiency of the EnKF depends on ensemble size. Researchers try to improve the efficiency of EnKF by improving initial sampling such that a small ensemble size can provide as large an uncertainty space span as possible. Methods for sampling a given random variable for efficient uncertainty propagation calculation are applied, such as the polynomial chaos expansion or probabilistic collocation method (Xiu and Karniadaskis 2003). Sarma et al. (2005) evaluated such methods for optimal encapsulation of information contained in an input random variables and output random flow variables. These methods require the independent random variables; thus a methodology for representing a spatial correlated random function by a series of independent random variables is required, *e.g.*, a Karhunen-Loveve (KL) expansion, which is a form of eigenvector expansion. A dimension-reduced Kalman filter based on K-L decomposition was proposed by Zhang, Lu, and Chen (2007). This type of operation can be inefficient for large systems because of the eigenvalue and eigenvector calculations. Evensen (2004) has examined resampling of smaller number of ensemble members from the suite of realization sets; he reports a larger uncertainty space than naive selection from the suite of realization sets.

1.3.3 History Matching as a Minimization Problem

In this section, the basic notation for the history matching problem in a Bayesian setting is introduced. We discuss the computation of the maximum a posteriori (MAP) estimate of reservoir variables and randomized maximum likelihood (RML). The uncertainty in the future performance prediction assessed by the RML method is similar to that by EnKF (Gao, Zafari, and Reynolds 2006); both aim to estimate distributions of models rather than single, most-likely values.

The objective function to be minimized for model inversion problem is constructed to integrate all available data. In the early stage of history matching (Jahns 1966; Jacquard and Jain 1965; Chen et al. 1974), only dynamic data were integrated in history matching, and the objective function was a weighted mismatch between observed production data, $\boldsymbol{d}_{obs} \in \boldsymbol{R}^{N_d}$, and predicted production data, $\boldsymbol{g}(\boldsymbol{m}) \in \boldsymbol{R}^{N_d}$.

$$\boldsymbol{O}(\boldsymbol{m}) = [\boldsymbol{g}(\boldsymbol{m}) - \boldsymbol{d}_{obs}]^{\mathrm{T}} \boldsymbol{W}_d [\boldsymbol{g}(\boldsymbol{m}) - \boldsymbol{d}_{obs}], \tag{1.1}$$

In the equation, $\boldsymbol{W}_d$ is an $\boldsymbol{N}_d \times \boldsymbol{N}_d$ weighting matrix and $\boldsymbol{N}_d$ is the number of observations. If $\boldsymbol{W}_d$ is the inverse covariance matrix of the measurement errors of the data integrated $\boldsymbol{C}_D^{-1}$, then $\boldsymbol{C}_D$ quantifies observation errors, and minimization of objective function Eq. 1.1 yields the maximum likelihood estimate of the model. In most model inversion cases, the number of observations is less than the number of model parameters; these are termed underdetermined problems.

The objective function $\boldsymbol{O}(\boldsymbol{m})$ with an added regularization term in Eq. 1.2, leads to estimates for the property fields which should be unique (or at least less rank-deficient), smoothly varying functions of position that have only the amount of variation necessary to satisfy the measured data and the spatial correlation provided by the geostatistical data

(Tarantola 1997).

$$O(\boldsymbol{m}) = [\boldsymbol{g}(\boldsymbol{m}) - \boldsymbol{d}_{obs}]^{\mathrm{T}} \boldsymbol{W}_d [\boldsymbol{g}(\boldsymbol{m}) - \boldsymbol{d}_{obs}] + [\boldsymbol{m} - \boldsymbol{m}_{prior}]^{\mathrm{T}} \boldsymbol{W}_m [\boldsymbol{m} - \boldsymbol{m}_{prior}], \tag{1.2}$$

where $\boldsymbol{W}_m$ is an $\boldsymbol{N}_m \times \boldsymbol{N}_m$ matrix, termed the model weighting matrix, $\boldsymbol{m}_{prior}$ denotes the prior estimate of model parameters, and $\boldsymbol{N}_m$ is the number of model parameters. $\boldsymbol{W}_d$, $\boldsymbol{W}_m$ and $\boldsymbol{m}_0$ are chosen by the information of measurement errors, spatial correlation of model parameters and initial geophysical and geological information. Eq. 1.2 means that the model parameters with the highest probability of being correct (given the current state of information) are those models that are "close" to the prior model and that honor the observation data.

Assuming the prior pdf (probability density function) for $\boldsymbol{m}$ is a multivariate Gaussian random variable with mean $\boldsymbol{m}_{prior}$ and covariance matrix $\boldsymbol{C}_M$, then Eq. 1.3 holds. The conditional pdf (or the *a posteriori* pdf) for model $\boldsymbol{m}$ given observation $\boldsymbol{d}_{obs}$ can be derived using an application of Bayes theorem (Jackson 1979; Tarantola and Valette 1982),

$$\pi_p(\boldsymbol{m}) = a \exp(-\frac{1}{2}(\boldsymbol{m} - \boldsymbol{m}_{prior})^{\mathrm{T}} \boldsymbol{C}_M^{-1} (\boldsymbol{m} - \boldsymbol{m}_{prior})), \tag{1.3}$$

$$p(\boldsymbol{m}|\boldsymbol{d}_{obs}) = \frac{p(\boldsymbol{d}_{obs}|\boldsymbol{m}) p(\boldsymbol{m})}{p(\boldsymbol{d}_{obs})} = c \exp([-\boldsymbol{O}(\boldsymbol{m})]), \tag{1.4}$$

where c is the normalizing constant, and $\boldsymbol{O}(m)$ is the objective function given by,

$$O(\boldsymbol{m}) = \frac{1}{2}[\boldsymbol{g}(\boldsymbol{m}) - \boldsymbol{d}_{obs}]^{\mathrm{T}} \boldsymbol{C}_d^{-1} [\boldsymbol{g}(\boldsymbol{m}) - \boldsymbol{d}_{obs}] + \frac{1}{2}[\boldsymbol{m} - \boldsymbol{m}_{prior}]^{\mathrm{T}} \boldsymbol{C}_M^{-1} [\boldsymbol{m} - \boldsymbol{m}_{prior}], \tag{1.5}$$

For Monte Carlo methods, like EnKF, we need not estimate the value of c. The maximum *a posteriori* (MAP) estimate, $\boldsymbol{m}_\infty$ is obtained by minimizing the objective function given by Eq. 1.5. In practice, minimizing the objective function once merely provides one realization of reservoir field because of the nonlinearity and nonuniqueness of the reservoir simulation equations (discussion of RML, next section). Moreover, the minimization process is very demanding in both gradient and search direction computations.

1.3.4 Uncertainty Assessment

It is difficult to sample the *a posteriori* pdf for nonlinear problems. The rejection algorithm is theoretically correct, but it is impractical for the problems of conditioning a reservoir model to production data (Liu, Betancourt, and Oliver 2001). Markov chain Monte Carlo (MCMC) is another rigorous sampling method. Unfortunately, it appears to be too inefficient computationally for practical applications even with modifications designed to improve its computational efficiency (Oliver, Luciane, and Reynolds 1997). Although it is possible to get some measure of the uncertainty in model parameters by calculating of the *a posteriori* covariance matrix (Oliver 1994; Tarantola 1997), the approach assumes that the *a posteriori* pdf can be approximated by a Gaussian centered at the MAP estimate. However, approximating the pdf by a Gaussian does not appear to provide a good basis for predicting the uncertainty in performance predictions (Liu, Betancourt, and Oliver 2001).Although it is possible to get some measure of the uncertainty in model parameters by calculating the *a posteriori* covariance matrix (Oliver 1994; Tarantola 1997), the approach assumes that the pdf can be approximated by a Gaussian centered at the MAP estimate. However, approximating the *a posteriori* pdf by a Gaussian does not appear to provide a good basis for predicting the uncertainty in performance predictions (Liu, Betancourt, and Oliver 2001). Oliver, He, and Reynolds (1996) proposed using the randomized maximum likelihood (RML) method to generate an approximate sampling of the *a posteriori* pdf. A conditional realization is generated by minimizing the objective function given by

$$O(\boldsymbol{m}) = \frac{1}{2}[g(\boldsymbol{m}) - \boldsymbol{d}_{uc}]^{\mathrm{T}}\boldsymbol{C}_d^{-1}[\boldsymbol{g}(\boldsymbol{m}) - \boldsymbol{d}_{uc}] + \frac{1}{2}[\boldsymbol{m} - \boldsymbol{m}_{uc}]^{\mathrm{T}}\boldsymbol{C}_M^{-1}[\boldsymbol{m} - \boldsymbol{m}_{uc}], \tag{1.6}$$

where $\boldsymbol{m}_{uc}$ is an unconditional realization defined by

$$\boldsymbol{m}_{uc} = \boldsymbol{m}_{prior} + \boldsymbol{C}_M^{1/2}\boldsymbol{z}_M, \tag{1.7}$$

and $\boldsymbol{d}_{uc}$ is obtained by adding noise to the observed data,

$$\boldsymbol{d}_{uc} = \boldsymbol{d}_{obs} + \boldsymbol{C}_D^{1/2}\boldsymbol{z}_D, \tag{1.8}$$

$\boldsymbol{z}_M$ and $\boldsymbol{z}_d$ are N_m and N_d dimensional column vectors of independent traditional random normal deviates and $\boldsymbol{C}_M^{1/2}$ and $\boldsymbol{C}_D^{1/2}$ denote the square root of $\boldsymbol{C}_M$ and $\boldsymbol{C}_D$. If $\boldsymbol{C}_D$ is diagonal, generating the square root simply requires taking the square root of the diagonal elements; for general $\boldsymbol{C}_M$, an LU-decomposition (*e.g.*, Cholesky) is used to compute $\sqrt{\boldsymbol{C}_M}$, which may be expensive (Oliver, He, and Reynolds 1996). For large scale problems, the unconditional realizations $\boldsymbol{m}_{uc}$ may also be generated from the prior model by sequential Gaussian cosimulation. A series of conditional realizations can be then generated by minimizing the objective function Eq.1.6 with different realizations of $\boldsymbol{m}_{uc}$ and $\boldsymbol{d}_{uc}$.

Liu, Betancourt, and Oliver (2001) compared the results obtained by different sampling methods with a 1-D single phase test example. They generated sets of 5,000 realizations using different sampling methods. The results show that the RML method produces distributions of reservoir properties similar to MCMC (which is known to be correct in the limit of very large samples).

1.4 Objectives and Research Scope

The main goal of this research is to develop grid-based software for designed reservoir simulation studies and continuous reservoir updating with EnKF (in cooperation with researchers from the Center for Computation Technology). Then use the software to explore the characteristics and performance of EnKF. The focus is on the following parts:

1. To collaborate with computer science researchers to implement a designed reservoir simulation studies work flow in high-performance Grid computing environments and evaluate geostatistics algorithms with the software – ResGrid;
2. To develop the EnKF "plug-in" for the continuous reservoir model updating work flow and solution extraction codes from simulation output. To implement the EnKF on Grid by extending the ResGrid;

3. To investigate the behavior and performance of the EnKF, especially focus on studies on two potential problems of the EnKF: the reasons causing its failure to converge to its solutions, the efficient way to solve the strongly nonlinear, non-Gaussian problem.
4. Explore the efficiency of grid computing for this application area.

There are nine chapters in this dissertation. Chapter 1 (this chapter) gives a brief review of research objectives, a statement of problems and the relevant literature. Chapter 2 illustrates the grid computing module (ResGrid) development, its components and the flow chart. An application of ResGrid to a geostatistical algorithm comparison project is included in the chapter. Chapter 3 includes an introduction of terminologies used in continuous reservoir model updating, background of Kalman filter, extended Kalman filter, and ensemble Kalman filter. The differences between Kalman filter and its variant are explained. Chapter 4 introduces the EnKF on Grid workflow adapted from ResGrid and a 2-D waterflooding case study by using the workflow. The computational efficiency and applicability of grid computing are also discussed in this part. In Chapter 5, we generate a prototype workflow for the well known 3-D three-phase test case, PUNQ-S3, to better understand the EnKF behaviors and limitations. The influence of assimilation frequency on the EnKF results and the causes and solutions for the instability of EnKF are investigated in Chapter 6. Chapter 7 discusses how to use EnKF with a reparameterization method to solve a nonlinear and non-Gaussian state vector problem. The complete comparisons between EnKF with reparameterization and EnKF with simple truncation are demonstrated in this part. In Chapter 8, some topics of particular interest about the EnKF on Grid are discussed. Finally, the summary and conclusions of the study are summarized in Chapter 9.

Chapter 2
Grid Computing for Designed Simulation

In this chapter, we begin by discussing the features of Grid computing environment in Section 2.1, emphasizing those features that enhance reliability and usability, and that contribute to decreased computing costs via a future market in secure, distributed Grid computing. A workflow for Grid-based reservoir simulation is then outlined. This workflow includes middleware for model building and execution, and a web portal for invoking those scripts. This project uses softwares such as the Condor and Globus Toolkit to build and manage workflow, Grid Security Infrastructure, Gridsphere for portal creation and management.

The workflow is demonstrated for a geostatistical study of three-dimensional displacements in heterogeneous reservoirs. A regularly gridded, 3-D and multiphase reservoir simulator is used. A suite of 1,920 simulations assesses the effects of geostatistical methods and parameters. Much of the pre- and post-processing is automated in this workflow, which is based on experimental design. Multiple runs are simultaneously executed using parallel Grid computing. Grid services manage security, data acquisition, resource brokering and allocation, flow response analysis, and visualization; the reservoir engineer is freed from micromanaging these workflow components.

2.1 Grid Environment

Grid computing emerged in the mid 1990s with the goal of making computer power as easy to access as an electric power grid (Foster and Kesselman 1999). Grid computing is a subset of distributed computing, distinguished by its focus on large-scale resource sharing, innovative applications, and high-performance orientation. Grid computing is sometimes confused with cluster computing. The key difference is that a cluster is a single set of nodes sitting in one location, while a Grid may be composed of many machines or clusters with diverse architectures, in addition to other resources (*e.g.*, networks and storage facilities):

> Therefore Grid computing enables the virtualization of distributed computing and data resources such as processing, network bandwidth and storage capacity to create a single system image, granting users and applications seamless access to improved IT capabilities. Just as an Internet user views a unified instance of content via the Web, a Grid user essentially sees a single, large virtual computer.
>
> —Foster and Kesselman (1999)

Moreover,

> At its core, Grid computing is based on an open set of standards and protocols *e.g.*, Open Grid Services Architecture (OGSA) that enable communication across heterogeneous, geographically dispersed environments. With Grid computing, organizations can optimize computing and data resources, pool them for large capacity workloads, share them across networks and enable collaboration.
>
> —(IBM Redbooks 2008)

Grid computing technology is likely to play a significant role in future high-performance computing environments. It may enable new classes of computation-intensive applications, as the Internet fostered the development of new classes of information-oriented applications. In computational Grids, today's large-scale computing challenges, such as reservoir simulation, could become routine, and reservoir engineers would be able to explore a new generation of tools that use teraflop computers and petabyte storage systems interconnected by gigabit networks. Grid computing not only allows reservoir engineers to share files, but also resources. That is it not only enhances communication, but also fosters full collaboration toward common goals. Depending on the Grid that is used, these aggregated resources might comprise the majority of the supercomputers in the state or region.

Technologies such as Grid Security Infrastructure (GSI), Globus Toolkit, Condor-G, GridSphere, and Simple API for Grid Application (SAGA), were applied in this research.

GSI is a specification for secure and authenticatable communication in the Grid computing environment (Foster and Kesselman 1998).

Globus Toolkit (Globus 2008) is an open-source toolkit for building Grids, which integrates or implements GSI, remote resource allocation, data location service, information infrastructure, *etc* (Foster and Kesselman 1999).

Condor-G provides a job submission queue system for Grid computing. GridSphere is an open-source portal framework, which offers web-based user management, access control, and data and execution integration *via* its portlet-based architecture (Frey et al. 2002).

SAGA is the standard means for Grid application programming abstraction (Goodale1 et al. 2006), which is now on the threshold of becoming an Open Grid Forum (OGF) technical recommendation (early 2008).

SAGA has the following properties suitable for constructing a designed simulation study work flow (SAGA 2008):

1. Simple: easy to use, install, administer and maintain.
2. Uniform: provides support for different application programming languages as well as consistent semantics and style for different Grid functionality.
3. Scalable: Contains mechanisms for the same application (source) code to run on a variety of systems ranging from laptops to high performance resources.
4. Generic: adds support for different Grid middleware, even concurrent ones.
5. Modular: provides a easily extendable framework.

2.1.1 ResGrid Implementation

The ResGrid development includes Grid portal and middleware.

2.1.1.1 Grid Portal Design

The Grid portal acts as the entry point for reservoir engineers to access Grid resources. A reservoir engineer can submit and manage simulation jobs and keep track of a large number of data files through a user-friendly web page with no need to type in any command lines.

First, the portal manages security. A Grid Security Infrastructure certificate is retrieved from a proxy to provide authentication to access Grid resources. Secondly, the portal provides web pages to specify all the geological and engineering data for building flow models. Then the user can submit an ensemble of simulation jobs and track the progress of computation and view the simulation results via this ResGrid portal. GridSphere and GridPortlets are chosen for the ResGrid portal because they can speed the process of developing and deploying an application portal. GridSphere is a free, open-source portal framework developed by the GridLab project, which focused on developing Grid application tools and middleware.

2.1.1.2 Grid Middleware Design

ResGrid (Lei et al. 2006) is implemented in four modules:

1. **Resource Broker module** manages Grid resources to share loads across a Grid. It captures resource information and uses load balancing strategies to dispatch the simulation runs. Two resource attributes are considered: computational capability and architecture. A matrix describes a computing resource, including CPU number, CPU speed, CPU load averages, network bandwidth, memory size and local resource management system load. These features summarize the computational capability of a resource. The architecture factor can be employed to decide which type of geostatistics algorithms and reservoir simulators should be staged in to which resource. We can see there are three resources available in Table 2.1. One is a 256-node Linux cluster with PBS as LRMS (Local Resource Management System). The second one is a 16-node Linux cluster with PBS as LRMS. The third one is a 14-node AIX machine with loadleveler. Additionally, there is a work directory (WORK_DIR) for each facility. It

Table 2.1: Resource list example

CPU#	**OS**	**HOSTNAME**	**WORK_DIR**	**LRMS**
256	LINUX	eric.loni.org	/home1/xli	pbs
16	LINUX	cangrid.louisiana.edu	/home1/xli	pbs
14	AIX	pelican.cct.lsu.edu	/home/xli	loadleveler

is the home directory of the local account of a Grid user. The Stage In/Out module uses the work directory to update the executable and reservoir models, and download the results. In the current step of research, CPU number (N_Π) and CPU speed (S_Π) of a resource are critical because both geological modeling and reservoir flow simulation are sequential processes. The computational capability of a resource χ_i is measured as follows:

$$\chi_i = N_\Pi \times S_\Pi \quad \forall i \in \{1, 2, \ldots, N_{hosts}\} \tag{2.1}$$

where N_{hosts} is the number of possible sites to distribute to, and

$$\chi_T = \sum_{i=1}^{N_{hosts}} \chi_i \tag{2.2}$$

where χ_T is the total computational capability of all possible sites. The load balancing strategy dispatches certain number of simulations to a resource according to its computational capability. The following equation computes the number of simulations Ns_j submitted to a resource j:

$$Ns_j = Ns_t \times \frac{\chi_j}{\sum_{k=1}^{N_{hosts}} \chi_k} \tag{2.3}$$

where χ_i and χ_k are the computational capability of the resource i and k; Ns_t is the total number of simulation runs.

2. **Stage In/Out module** deploys the model data and executables to remote resources, retrieves the simulation output from the remote resources and GridFTP the results back to users' master machine. Before uploading the executables, the Stage In checks the Resource Broker module to obtain the type of operating system on a remote resource. Then it chooses the correct executable codes for geostatistical algorithm and simulator. After retrieving the load balancing results from the Resource Broker module, the Stage In/Out module calculates how many and which simulation models should be run on a particular remote resource. This module also accesses the work directory of a user on a particular remote resource. After obtaining the required information, this module transfers the datasets back to users' master machine.
3. **Invoking module** handles remote executions. The module communicates with the LRMS on a remote resource and submits simulations to the corresponding LRMS.
4. **Status Monitoring module** communicates with LRMS. There are two levels of queues for status monitoring: the resource queue on submission master machines and the LRMS job queue on each particular resource. Each resource which is running simulations has an entry in resource queue. On a particular resource, the job queue of LRMS is checked periodically. When all the simulations dispatched to the resource have been done; the corresponding resource entry in resource queue is removed. A user can get the simulation status from the resource queue.

2.1.2 ResGrid Description

ResGrid provides an environment for reservoir simulation studies. Its interfaces specifies the geostatistical or engineering parameter, invoke stochastic and flow model simulations across the Grid, monitor the simulation processing, and analyze and visualize simulation results. The architecture of ResGrid is illustrated in Figure 2.1. Typically, the steps in ResGrid can be summarized as following:

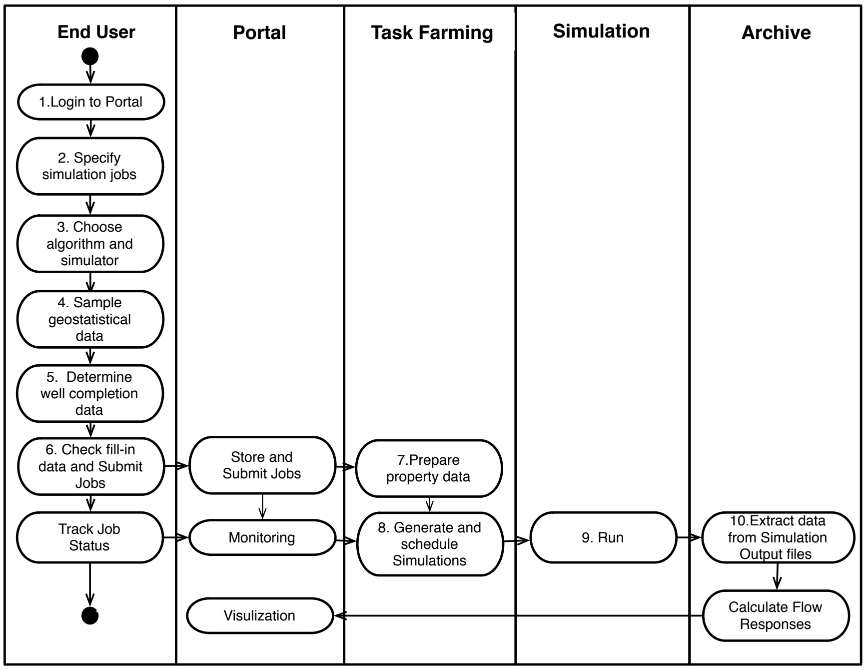


Figure 2.1: ResGrid usage scenario. All the operations on large-scale reservoir uncertainty analysis are completed via the Grid portal, interacting with various Grid services and resources.

1. Users log in the ResGrid portal and retrieve a GSI certificate from a proxy server. The certificate authorizes the user to access Grid resources and implement secure data transfer (Figures 2.2(a), 2.2(b)).

2. On the job description page, users specify the general information of simulation jobs, the name and description of the job and check the template if users have saved data in database before. The functionality of the template is to avoid tedious repetition by using a previously saved template by users.

3. On the model geometry page, users fill in the model geometry information and choose the stochastic simulation algorithm and numerical simulator.

4. On the spatial variability page, the geostatistical data used for the reservoir property fields are specified by the users according to the core, log data and geological knowledge (Figure 2.3(a)).

5. On the well completion page, the numbers of wells, well types and well locations are specified by users. The portal will automatically check if the well locations are reasonable. If not, an error will highlight in red on the top of the page.

6. Before submission, all the information could be displayed again on one page when users can double check the fill-in data (Figure 2.3(b)).

   After the job submission, the Grid services can be summarized as following:

7. The first service is reservoir modeling. This triggers a data-archiving tool and generates reservoir permeability and/or porosity fields by using the geostatistical parameters and algorithms specified by the user in Steps 3 and 4.

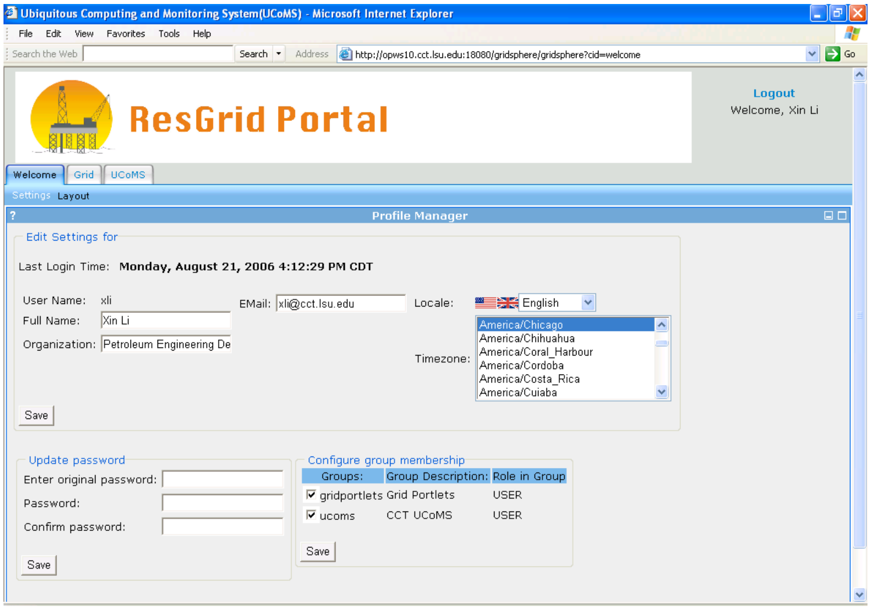
8. The resource-brokering service captures the dynamic information from the information service provided by the Grid, makes a decision on the appropriate resource for each simulation run with the help of load balancing strategies.

9. The simulation jobs are distributed to available Grid resources and then the simulation execution service invokes the simulation runs.

10. Once all the simulation runs have been completed, the useful information is automatically extracted to calculate the simulation responses, and the statistical analysis service is activated to analyze the simulation results.

11. The user views the results on the ResGrid portal which are generated by the visualization service.

The workload of a reservoir engineer can be reduced by using ResGrid, because they only interact with the Web-based Grid portal designed for reservoir study. The ResGrid services take care of security, data acquisition, resource management, result analysis, and visualization. It is not necessary for end users to manually manage these activities.



http://opws10.cct.lsu.edu:18080/gridsphere/gridsphere?cid=welcome

Last Login Time: **Monday, August 21, 2006 4:12:29 PM CDT**

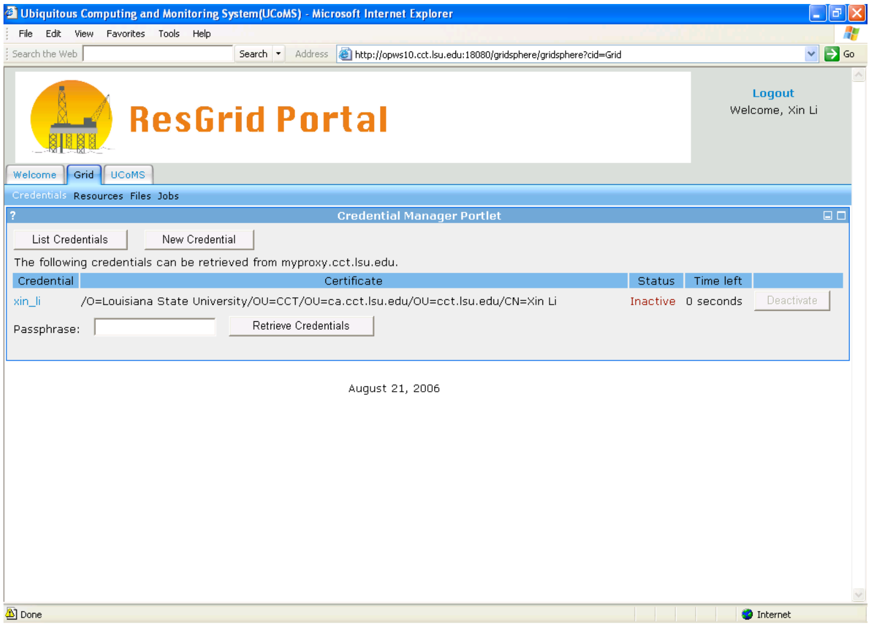
User Name: xli
Full Name: Xin Li
Organization: Petroleum Engineering De
EMail: xli@cct.lsu.edu
Locale: English
Timezone: America/Chicago

America/Chicago
America/Chihuahua
America/Coral_Harbour
America/Cordoba
America/Costa_Rica
America/Cuiaba

Enter original password:
Password:
Confirm password:

Groups:	Group Description:	Role in Group
☑ gridportlets	Grid Portlets	USER
☑ ucoms	CCT UCoMS	USER

(a) Welcome page for ResGrid Portal. After login, user information will be shown in this page. At the configure group membership part, check the Gridportlets and UCoMS



http://opws10.cct.lsu.edu:18080/gridsphere/gridsphere?cid=Grid

The following credentials can be retrieved from myproxy.cct.lsu.edu.

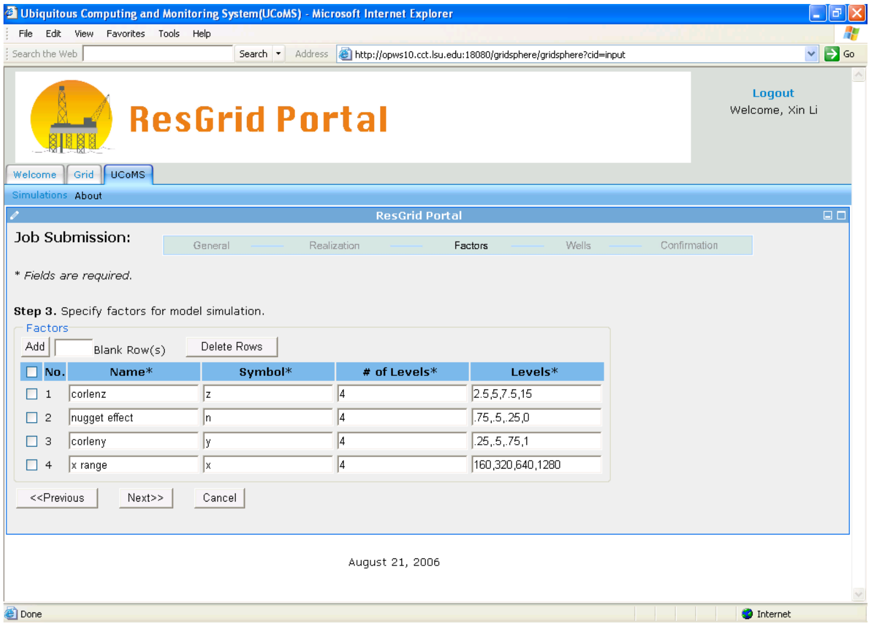
Credential	Certificate	Status	Time left
xin_li	/O=Louisiana State University/OU=CCT/OU=ca.cct.lsu.edu/OU=cct.lsu.edu/CN=Xin Li	Inactive	0 seconds

Passphrase:

August 21, 2006

(b) Grid security. A Grid Security Infrastructure certificate is retrieved from a proxy to provide user authentication to access Grid resources.

Figure 2.2: UCoMS portal login



Job Submission:

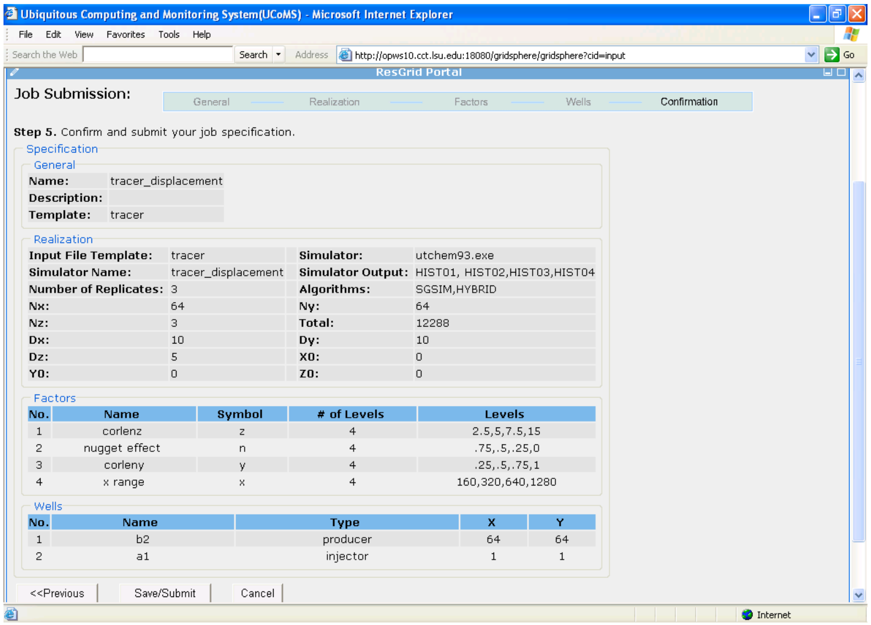
** Fields are required.*

Step 3. Specify factors for model simulation.

Factors

☐	No.	Name*	Symbol*	# of Levels*	Levels*
☐	1	corlenz	z	4	2.5,5,7.5,15
☐	2	nugget effect	n	4	.75,.5,.25,0
☐	3	corleny	y	4	.25,.5,.75,1
☐	4	x range	x	4	160,320,640,1280

(a) Variogram parameters levels and values



Job Submission:

Step 5. Confirm and submit your job specification.

Specification

General

Name:	tracer_displacement
Description:	
Template:	tracer

Realization

Input File Template:	tracer	**Simulator:**	utchem93.exe
Simulator Name:	tracer_displacement	**Simulator Output:**	HIST01, HIST02,HIST03,HIST04
Number of Replicates:	3	**Algorithms:**	SGSIM,HYBRID
Nx:	64	**Ny:**	64
Nz:	3	**Total:**	12288
Dx:	10	**Dy:**	10
Dz:	5	**X0:**	0
Y0:	0	**Z0:**	0

Factors

No.	Name	Symbol	# of Levels	Levels
1	corlenz	z	4	2.5,5,7.5,15
2	nugget effect	n	4	.75,.5,.25,0
3	corleny	y	4	.25,.5,.75,1
4	x range	x	4	160,320,640,1280

Wells

No.	Name	Type	X	Y
1	b2	producer	64	64
2	a1	injector	1	1

(b) Job submission. Before the users submit jobs or save the templates, they can check for typos or other mistakes. Users click the submit/save button, the job will be submitted to available resources or saved as template according to user's requirement.

Figure 2.3: UCoMS portal usage

2.1.2.1 Open Source Geostatistical Library and Reservoir Simulator

- The reservoir modeling software GSLIB, is an acronym for **G**eostatistical **S**oftware **LIB**rary. This name was originally used for a collection of geostatistical programs developed at Stanford University over the last 15 years. The code used in ResGrid is Fortran 77 and can be downloaded freely from GSLIB: Geostatistical Software Library 2008.
- Reservoir Simulator UTCHEM (University of Texas Chemical Compositional Simulator) is a three-dimensional, multiphase, multicomponent, compositional, variable temperature, finite-difference numerical simulator. The Fortran 77 source code can be compiled and run at a variety of Unix workstations, which can be downloaded from UTCHEM 2008.
- SPECSIM and HYBRID - Both algorithms are in-house C and Fortran codes (developed as part of this dissertation by author).

2.2 Geostatistical Studies

This study compares and evaluates four different stochastic simulation algorithms. We identify and quantify how the geological factors influence the determination of effective properties and production behaviors using different simulation methods. A range of variogram factors are studied, using Design of Experimental (DOE), F-test,t-test and Response Surface Models (RSM).

2.2.1 Motivations

Stochastic simulation creates reservoir property fields that match the available information at the sparse wells and reproduce the pattern of spatial variability between wells described by the variogram. Stochastic simulations can be categorized into direct (LU Decomposition Gaussian Simulation) and sequential approaches (Sequential Gaussian Simulation). LU Decomposition Simulation (LUSIM) is rigorous but slow. Sequential Gaussian Simulation

(SGSIM) is quicker but potentially inaccurate. We create a hybrid simulation (HYBRID) to take the advantage of the best of the two approaches. Spectral simulation is fast because it applies fast Fourier transforms instead of directly solving a kriging system at each simulated node. Three flow responses are computed from the simulation results to find differences among these simulation algorithms.

2.2.2 Design of Experiments (DOE) and Response Surface Model (RSM)

Flow simulations are employed to assess the effects of four variogram parameters describing permeability distributions on three flow responses in three directions. The parameters are varied using a 32-run four-level four-factor orthogonal array (OA) design (Kalla and White 2007) for all permeability model sets. Five realizations are run at each design point to model the stochastic permeability fields.

2.2.2.1 Factors Examined

Factors are input parameters that are varied during experimental design (White and Royer 2003). Four uncertain variogram factors are examined in this study (Willis and White 2000; White et al. 2001). The range of four factors are listed in Table 2.2).

1. N, the variogram nugget effect is related to sources of variation that operate over distances smaller than the shortest sampling interval. We assume the worst case of nugget effect is three quarters, the best scenario is no nugget effect. Half and one quarter are the medium cases.
2. R, the variogram range of 90 degree azimuth which is the major direction of continuity. In this study, the experimental variogram is modeled by exponential expression. The inferred variogram ranges for structure use quarter of structure length as the low case and twice the structure length as the high case. Half the reservoir length and equal to the length are the two middle cases. This is the x-direction for the flow-simulation study.

Table 2.2: Designed factors and scaling

Index	Factor Name	Symbol	Units	0	1	2	3
1	Nugget effect	N	fraction	0.75	0.5	0.25	0
2	X range	R	m	160	320	640	1280
3	Anisotropy ratio	A	fraction	0.25	0.5	0.75	1
4	Z range	Z	m	2.5	5	10	20

3. A, the variogram geometric anisotropy ratio is the directional variogram which has the same shape and sill but smaller range values compared with the range of maximum continuity direction R. The low case considers the quarter of R, the high case is the same as R.
4. Z, the vertical variogram range. The quarter of the thickness of structure is taken as the low case. The thickness of structure is the high case in the study.

2.2.2.2 Reservoir Permeability Fields

The heterogeneity of permeability fields are modeled by unconditional simulations. The logarithm permeability ($\ln k$) has a Gaussian histogram with mean and variance of 5.5 and 1, respectively. The variogram is an exponential model with principal directions of continuity at 90 degrees and 180 degrees.

2.2.2.3 Flow Responses

Responses are the results of the flow model. The responses used in the study are (Li and White 2003):

1. Upscaled permeability $\bar{k}$ is defined as the ratio of flow rate to pressure draw-down computed from simulation results.
2. Breakthrough time τ_{BT} is a dimensionless time in pore volumes. It is the total tracer injection volume when the outlet tracer concentration exceeds 1 percent. The dimensionless time is:

   $$\tau(t) = \frac{\int_0^t q_{rc}(t')dt'}{V_p} \tag{2.4}$$

   where t is the time, q is the volumetric flow rate at reservoir conditions, V_p is the total pore volume. If $t = t_{BT}$ is the time when the outlet tracer concentration first reaches 1 percent of the injected concentration, $\tau = \tau_{BT}$.
3. Sweep efficiency N_{pD1} is the fraction of the initial tracer free water recovered after one pore volume of injection:

   $$N_{pD1} = \int_0^1 (1 - c(\tau))d\tau \tag{2.5}$$

where $c(\tau)$ is outlet tracer concentration. The post-processing part of ResGrid can automatically extract the responses from simulation output and calculate the responses required by users.

2.2.2.4 Flow Model Description

The simulated displacement process is ideal tracer flow. There is no buoyancy, capillary pressure, relative permeability, or viscosity contrast effects (Calhoun and Tittle 1968). A tracer displacement is used as the model process because it is quick to simulate. In addition, tracer flow isolates the heterogeneity of permeability, fewer factors influence responses, and truncation errors can be reduced for fully miscible systems.

2.2.2.5 Response Surface Model.

Based on 1,920 simulation runs, which are calculated by $3 \times 4 \times 32 \times 5$ (where 3 is the number of well patterns, 4 is number of geostatistical algorithms, 32 is the simulation number by OA design and 5 is the number of realizations for each factor combination), the least squares

Table 2.3: Simulation model parameters summary

Model parameter	Value	Units
xyz grid size	$64 \times 64 \times 3$	
Total blocks	12,288	
Active blocks	12,288	
Block size	$10 \times 10 \times 5$	ft
Porosity	19%	
Pore volume	1,167,360	ft^3
Initial pressure	14.7	psi
Well patterns	horizontal injector and producer	quarter nine-spot

Table 2.4: Box-Cox transformation result analysis for LUSIM

response	upscaled permeability		breakthrough time		sweep efficiency	
	before	after	before	after	before	after
R^2	0.7679	0.7771	0.6725	0.6821	0.7451	0.7735
R^2_{adj}	0.7523	0.7621	0.6506	0.6608	0.7280	0.7583

method builds a first order polynomial response surface model. The R^2 and R^2_{adj} are low for linear regressions. The regression models do not fit the responses well. The Box-Cox transformation and weighted least squares (WLS) have been used to improve the regression (Tables 2.4 and 2.5). The results show that the transformations and WLS did not produced a significant improvement to the regression results. The regression models fit responses from 2D flow much better than that of 3D. The reason could be that the injector and producer are horizontal wells penetrated in the second layer in the 3D model, which may cause nonuniform and erratic flow. Therefore, direct comparisons between each combination are made in the following section.

Table 2.5: Weighted Least Square result analysis for LUSIM

response	upscaled permeability		breakthrough time		sweep efficiency	
	before	after	before	after	before	after
R^2	0.7679	0.9132	0.6725	0.6684	0.7451	0.7735
R^2_{adj}	0.7523	0.9074	0.6506	0.6460	0.7280	0.7583

2.2.3 Model Discrimination

The reproduced model statistics are used to compare LUSIM and SGSIM models. The variance curve comparison of the two models for five realizations with same variogram (R= 320 m; A= 1; Z= 10 m; N= 0.25) is shown in Figure 2.4. Ergodic fluctuations exist for both models. But variances fluctuate become more seriously if sequential simulation is used. The reproduced variances of LUSIM are closer to model variance. The variance of SGSIM is higher than that of LUSIM. The 5 realizations of SGSIM model have quicker breakthrough and slower tracer concentration build-up in fractional flow than the LUSIM model (Figure 2.5). More flow sweeps the high permeable parts and by-pass the low permeability area. These features appear to exert a strong influence on flow behavior.

2.2.4 Geostatistical Analysis

Flow simulations of the geostatistical models are analyzed using analysis of variance. Standard t- and F-statistics assess whether flow responses of LUSIM models are different from the other simulation methods at each combination. All responses are computed for mean flow along the directions of maximum (y) and minimum (x) continuity.

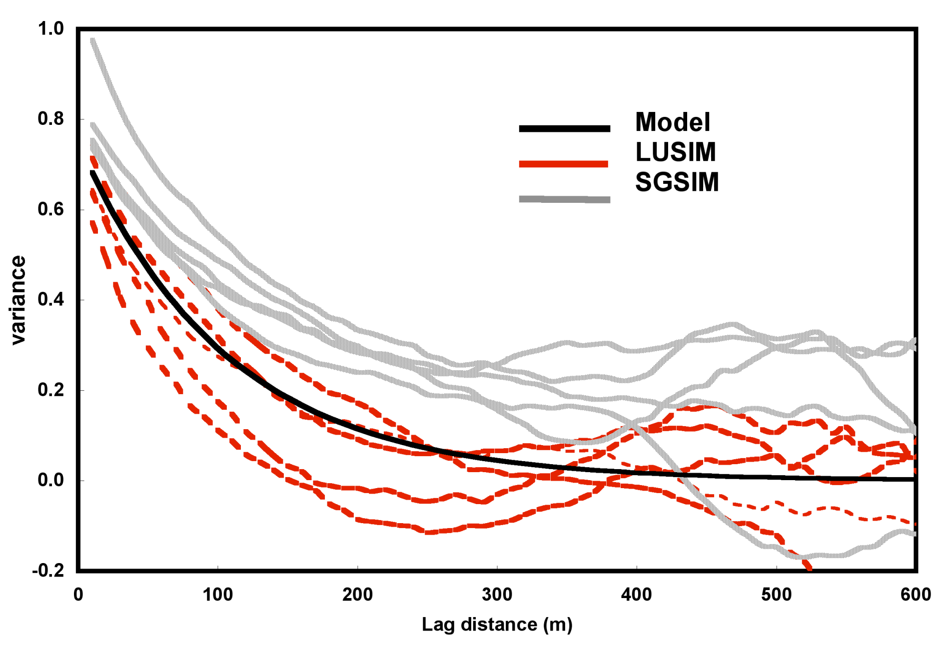


Figure 2.4: Reproduced variance models of SGSIM,LUSIM compared with given model

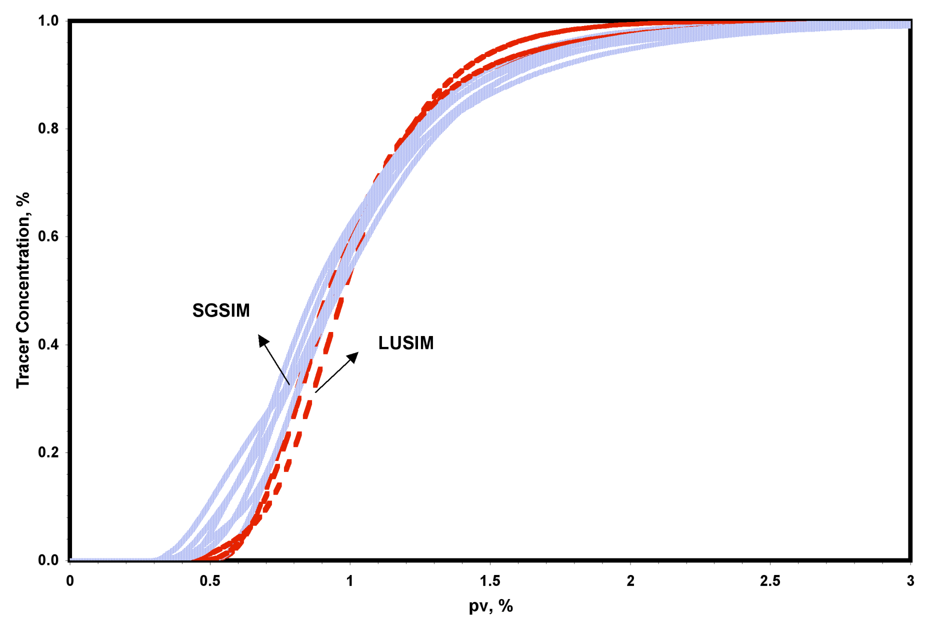


Figure 2.5: Comparison of fractional flow between LUSIM and SGSIM

Upscaled permeability ($\bar{k}$). The SGSIM mean differs from the LUSIM (reference) mean, especially when the z-range is short (Table A-1). The differences between LUSIM and SGSIM means are 5 to 20 percent when the z-range is small. F-tests indicate that SGSIM significantly overstates response variability, especially for high nugget, short range and anisotropic scenarios; the variance ratio is as high as 200. Thus, there are significant errors for SGSIM models.

Breakthrough time (τ_{BT}). The breakthrough time response t-tests show significant differences (Table A-3). In the maximum continuity direction, SGSIM predicts (mean) breakthrough 5 to 26 percent earlier than LUSIM (Table A-3). In the minimum continuity direction, the error is from 16 to 36 percent. The F-test shows that the difference in the maximum continuity direction is significant, with the variances differing from 6 to 41 percent.

Sweep efficiency (N_{pD1}). The t-test results between LUSIM and SGSIM are significantly different, but the mean difference is modest compared with the other responses (Table A-2 and A-5). The F-tests indicate no difference in variance estimates (Table A-2).

HYBRID method is different from the LUSIM method at 95 percent confidence level in Table A-7 and A-8. It seems that the conditional data from LUSIM does not have enough constraints on the sequential Gaussian method.

Table A-9 and A-10 shows the spectral method could not be distinguished from the LUSIM method at 95% confidence level.

2.3 Discussion

Our work focuses on creating an integrated, secure, and easy-to-use problem-solving environment for reservoir simulation study across a Grid. This chapter described ResGrid application: a Grid portal, data management and execution management.

Table 2.6: Grand average of mean difference and variance ratio of responses between LUSIM and SGSIM

Response	direction	upscaled permeability	sweep efficiency	breakthrough time
$(R_{SG} - R_{LU})/R_{LU}$	x	11.3	3.8	-14.1
	y	13.3	4.8	-22.3
σ_{SG}/σ_{LU}	x	29.3	2.2	7.8
	y	17.9	9.5	5.4

1. A GridSphere-based Grid portal provides an ubiquitous, friendly, and secure interface for reservoir engineers. Reservoir engineers can easily submit and manage their simulation jobs without addressing the Grid complexity underneath.
2. The essential part of ResGrid is a data management tool, which has been implemented on top of the GAT. With this tool, a reservoir engineer can archive and search simulation datasets conveniently.
3. In order to conduct execution management, a task farming framework has been developed. The resource brokering module captures Grid resource information and uses load balancing strategies to dispatch reservoir simulations on Grid resources. The invocation module is used to invoke reservoir simulation runs combined with geostatistics algorithms across a Grid.

ResGrid can be applied to designed simulation studies to enumerate influential factors and discriminate models, and yield response and sensitivity estimates over the range of all factors. Factor lists and automate permeability construction, simulation data deck assembly, execution, and summary tabulation. The automatic workflow was applied to simulate, analyze and discriminate some frequently used geostatistical algorithms. Standard t- and F-test results show that when the reservoir is heterogeneous, the upscaled permeability, sweep efficiency and breakthrough time of the LUSIM model are significantly different from the SGSIM. The overestimate of permeability by SGSIM has more influence on breakthrough time and upscaled permeability. The response differences between LUSIM and SPECSIM are small. Improvement with the HYBRID method is not significant. The positions and number of samples to condition the sequential Gaussian simulation may affect the performance of HYBRID. Future work should focus on monitoring and steering capabilities at runtime during the execution of a given simulation run, checking job status and terminate the job if an error occurs and applying ResGrid to other simulation areas, such as automatic history matching with ensemble Kalman filter.

Chapter 3
Background and Terminology

Using EnKF, an ensemble of reservoir models that is tuned to the most current observations of production data is always available. Thus, the estimations of reservoir model parameters, and their associated uncertainty, as well as the forecasts are always up-to-date.

The chapter explains the basics of the EnKF method for continuous model updating and introduces the equations for the time evolution of the error covariance matrix. Section 3.1 introduces the governing equation system used to describe multiphase flow in porous media. In Section 3.2, the terminology commonly used in model inversion is reviewed. Section 3.3 briefly outlines the Kalman filter (KF) originally proposed by Kalman (1960). In Section 3.3.3 and 3.3.4 the basic introduction is given to the ensemble Kalman filter proposed by Evensen (1994) and its implementation. The last section will introduce the frequently used statistical measurements for performance of EnKF.

3.1 Flow Equations in Porous Media

In reservoir simulation, the flow equations governing the multiphase flow in porous media are obtained by combining forms of Darcy's Law and the equation of mass conservation. The general equations for multiphase flow with "black oil" assumptions (the gas component dissolves in the oileic phase, but the aqueous phase is pure brine and the vapor phase is pure gas):

$$C_1 \nabla \cdot [\lambda_o (\nabla p_o - \gamma_o \nabla z)] = \frac{1}{C_2} \frac{\partial}{\partial t} [\frac{\phi S_o}{B_o}] + \tilde{q}_o \tag{3.1}$$

$$C_1 \nabla \cdot [\lambda_w (\nabla p_w - \gamma_w \nabla z)] = \frac{1}{C_2} \frac{\partial}{\partial t} [\frac{\phi S_w}{B_w}] + \tilde{q}_w \tag{3.2}$$

$$C_1 \nabla \cdot [R_s \lambda_o (\nabla p_o - \gamma_o \nabla z) + \lambda_g (\nabla p_g - \gamma_g \nabla z)] = \frac{1}{C_2} \frac{\partial}{\partial t} [\phi (\frac{R_s}{B_o} S_o + \frac{S_g}{B_g})] + R_s q_o + \tilde{q}_g \tag{3.3}$$

where transmissibilities λ_l are defined as

$$\lambda_l = \frac{k_{rl}k}{\mu_l B_l} \tag{3.4}$$

Eqs. 3.1 to 3.3 contain six dependent variables, Three additional relationships are needed to complete the multiphase flow description:

$$S_o + S_w + S_g = 1 \tag{3.5}$$

$$P_{cow} = p_o - p_w = f(S_w, S_g) \tag{3.6}$$

$$P_{cgo} = p_g - p_o = f(S_w, S_g) \tag{3.7}$$

where the subscripts o, w, and g stand for oleic, water, and gas phases, respectively; $C_1 = 1.127 \times 10^{-3}$ and $C_2 = 5.615$ when oil field units are used; the unit of oil and water formation factor (B_o and B_w) is RB/STB; the gas formation factor is RB/scf; R_{so} is the solution gas-oil ratio with units of scf/STB; the viscosity μ_m is in unit of cp; $\gamma_m = \frac{\rho_m g}{144 g_c}$ is the specific density; D is the vertical distance from a datum level in ft; q_m is the source/sink term, $\tilde{q}_o$ and $\tilde{q}_w$ are in units of $\frac{\text{stb}}{\text{rb}\cdot\text{day}}$; $\tilde{q}_g$ is in units of $\frac{\text{scf}}{\text{cf}\cdot\text{day}}$; for production, $\tilde{q}_m < 0$, for injection, $\tilde{q}_m > 0$; the pressure is in units of psi; P_{cow} is the capillary pressure assuming the water is the wetting phase in the presence of oil and water; P_{cgo} the capillary pressure assuming the oil is the wetting phase in the presence of oil and gas; the saturation S_ℓ is dimensionless and varies between 0 and 1; porosity $\phi \in [0,1]$ is the fraction of pore space in the reservoir rock; $\vec{K}$ is the diagonal permeability tensor $\vec{K} = \begin{bmatrix} k_x & 0 & 0 \\ 0 & k_y & 0 \\ 0 & 0 & k_z \end{bmatrix}$, the nonzero entries are the absolute permeability of the reservoir rock along x, y, and z directions; the permeability has the dimension of [L^2] and for oil field units is md; k_{rl} is the relative permeability and a function of saturation; and $\nabla\cdot$ is a gradient operator, for Cartesian coordinate system,

$$\nabla = \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{bmatrix}$$

.

The above flow conservation equations are subject to the following assumptions (White 2007).

1. In a black oil model, the water and oil components may occur in their respective phases (aqueous and oleic) only, whereas the gas component may occur in the vapor or the oleic phases. Some extended black-oil models include gas soluble in the aqueous phase; volatile oil models allow the oil component to dissolve in the vapor phase; some extensions allow other possibilities such as water in the vapor phase.
2. Mass conservation only, no energy or momentum conservation (although momentum conservation is implicit in Darcy's law(Hubbert (1969, Bear (1988)).
3. The diffusive flux is neglected. Only advection is modeled because mainly the immiscible "field" is simulated in the study.

In Chapter 7, the 1-D two-phase water flood equation is derived by assuming incompressible flow, and the velocity and densities are constant. In that case, the oil equation 3.1 can be written:

$$u_T \frac{\partial f_o}{\partial x} + \phi \frac{\partial S_o}{\partial t} = 0 \tag{3.8}$$

where u_T is the total velocity, $u_w + u_o$ in two-phase flow. f_o can be calculated by rock-fluid properties. This is the classic Buckley-Leverett problem using a hyperbolic equation(Lake 1989). With initial and boundary conditions, the PDEs become a well-defined Initial Boundary Value problem. Usually, the reservoir is discretized into many gridblocks. In each gridblock, the three governing equations are still valid, but much simplified because porosity, permeability, viscosity, and B_ℓ are assumed to be constant within the grid at each iterative step. Along with methods to linearize the equations (*e.g.*, Newton's method; Aziz and Settari 1979), this allows algebraic solution of the system of coupled equations for all components and blocks. The solution provides insight into porous fluid distribution in the subsurface. The program conducting this computation process is the reservoir simulator. In this study, the simulator is based on finite difference methods. More detailed literature on reservoir simulation can be found in Aziz and Settari (1979).

3.2 Terminology in Continuous Reservoir Model Updating by EnKF

The most commonly used terms in history matching are summarized below.

1. **Model parameters**: these variables are uncertain and do not change with time. Therefore, they are referred to as **static model variables**. They include rock properties, such as porosity and permeability; transmissivities for faults and vertical permeability; and rock-fluid properties, *e.g.* endpoint of relative permeability. For sequential data assimilation methods like EnKF, estimates of these properties change as data are integrated, but porosity and permeability are understood to be constant in fact if not in the algorithm. The continuous model updating is always begun by defining a set of first-guess parameters $m^f \in \Re^n$, which is spatial mean, and their assumed stochastic error $m' \in \Re^n$, with mean equal to zero and known covariance $C_{MM} \in \Re^{n\times n}$. This can be shown in the following equation

$$m = m^f + m' \tag{3.9}$$

2. **Model solutions**: these variables are solutions from the simulator, and vary with time; they define the dynamics of the system. In contrast to static model variables, they are called **dynamic model variables**. The uncertainty in these variables comes from the uncertainty of model parameters and from mathematical modeling errors (which are not addressed here). However, we assume the relationships between model parameters and model solutions are deterministic — the solution is stochastic only because the model parameters are. Model solutions could be phase pressures, saturations of all fluid phases or solution gas-oil ratio R_s. These variables are solutions of systems of differential (or finite difference) equations in Section 3.1. If the reservoir model is

valid, and the model parameters are given, then the model solutions can be computed with initial conditions that are considered certain in this dissertation. However, other investigators have considered initialization uncertainty in inverse models. For example, Evensen (2007) identified initial uncertainties in the water saturation distribution, *e.g.*, fluid contacts such as water oil contact (z_{WO}) and gas oil contact (z_{GO});

3. **Observation responses**: these are observable quantities directly related to the model solutions and indirectly to the model parameters. For oil and gas reservoirs, these data could include surface flow rates, well bottom-hole pressure, well tubing head pressure, water cut (or water oil ratio), gas oil ratio, amplitude of seismic reflection, and other production, petrophysical and geophysical measurements over the reservoir life. Crucially, observed responses always have some errors or noises associated with them, and many data assimilation methods, including EnKF, require that the errors be specified.

The reservoir model parameter estimation problem with EnKF can be formulated as

> ...how to find the joint pdf of the model parameters and model solutions, given a set of measurement and a dynamical model with known uncertainty.
>
> — Evensen (2007),

which is vastly different from the traditional model inversion method (*op cit.*)

> ...how to find the parameters resulting in a model solution which is “closest” to a set of measurements.
>
> —Evensen (2007),

In reservoir history matching application, $\boldsymbol{m}$ is used to denote the model parameters, $\boldsymbol{f}(\boldsymbol{m})$ is denoted as the model solutions, and model solutions are the function of model parameters, and $\boldsymbol{g}(\boldsymbol{m})$ is used to denote the forecast of observations.

The observed responses are given by $\boldsymbol{d}_{obs}$,

$$\boldsymbol{d}_{obs} = \boldsymbol{g}(\boldsymbol{m}^{true}) + \boldsymbol{\epsilon} \tag{3.10}$$

where $\boldsymbol{\epsilon}$ is the measurement noise. It is usually assumed that $\boldsymbol{\epsilon}$ is unbiased and Gaussian, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{C}_D)$, *i.e.*, $E[\boldsymbol{\epsilon}] = \mathbf{0}$ and $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}] = \boldsymbol{C}_D$. $\boldsymbol{C}_D$ is the measurement error covariance matrix. It is a diagonal matrix if the measurement errors are uncorrelated with each other, which is assumed in this study.

In the Kalman filter literature, the

> **state vector** for a dynamic system is composed of any set of quantities sufficient to completely describe the unforced motion of the system. Given the state vector at a particular point in time and a governing equation from that point in time forward, the state at any other time can be computed.
>
> —Gelb (1979),

A joint state vector of a system can be defined as followed:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{f}(\boldsymbol{m}) \\ \boldsymbol{g}(\boldsymbol{m}) \end{bmatrix} \tag{3.11}$$

where $\boldsymbol{y}$ denotes the (augmented) state vector. It consists of model parameters, model solutions and observation responses. The poorly known parameters are concatenated with the model solutions because they are updated simultaneously in the EnKF. Now we have a combined parameter and state estimation problem. Using the true state vector, the relationship between the observed response and the true, error-free observations is

$$\boldsymbol{d}_{obs} = \boldsymbol{H}\boldsymbol{y}^{true} + \boldsymbol{\epsilon}, \tag{3.12}$$

where $\boldsymbol{H}$ is an operator matrix or row vector (depending on the number of observations). The expected value of the observation can be calculated by applying $\boldsymbol{H}$ to the joint state vector $\boldsymbol{y}$ (Eq. 3.11), which is equivalent to applying the possibly nonlinear function $\boldsymbol{h}$ to independent variable x. $\boldsymbol{H}$ allows applying ensemble filters in the joint state space. $\boldsymbol{H}$ is a trivial matrix whose elements are only ones and zeros. It is

$$\boldsymbol{H} = [\mathbf{0}|\boldsymbol{I}], \tag{3.13}$$

In practice, the construction of $\boldsymbol{H}$ is not necessary, while it is a way of writing column and row selections algebraically; this can be done more efficiently with simple operations in actual computer implementations.

3.3 Background on the Kalman Filter

The Kalman filter (KF), which introduces an equation for the time evolution of the error covariance matrix, was originally proposed by Kalman (1960), The Kalman filter has been used for state estimation of linear systems that evolve with time from noisy measurement, which is a widely used sequential data assimilation method for the purpose of obtaining a least squares estimation of the state of the system (Maybeck 1979). At each assimilation, there are actually two steps: a **forecast step** and an **update step**. In the following, t_k is the current time step. The **forecast step** evolves the state vector to the next measurement time.

If a discrete linear system can be modeled as Eq. 3.14

$$\boldsymbol{y}_k^{true} = \boldsymbol{\Psi}_{k-1}\boldsymbol{y}_{k-1}^{true} + \boldsymbol{\omega}_{k-1}, \tag{3.14}$$

The evolution of the state vector is

$$\boldsymbol{y}_k^{p} = \boldsymbol{\Psi}_{k-1}\boldsymbol{y}_{k-1}^{u}, \tag{3.15}$$

where k and $k-1$ are time step indices for measurement time t_k and t_{k-1}, respectively, when measured data are available; $\boldsymbol{y}_k$ is the state vector at time step k; the superscript p represents *prior*, meaning that the values are direct output of the dynamic system before updating (Eq. 3.15); u represents *updated*, meaning that the values are after observed response assimilation; $\boldsymbol{\Psi}_{k-1}$ is the state transition matrix that transits the state vector from time t_{k-1} to t_k with dimensions equal to $n \times n$; $\boldsymbol{\omega}_{k-1}$ is the unbiased Gaussian model error with covariance matrix $\boldsymbol{Q}_{k-1}$, *i.e.* $E[\boldsymbol{\omega}_{k-1}] = 0$ and $E[\boldsymbol{\omega}_{k-1}\boldsymbol{\omega}_{k-1}^{\mathrm{T}}] = \boldsymbol{Q}_{k-1}$.

The estimate of the state vector $\boldsymbol{y}_{k-1}^{u}$ is conditioned to measurements up to time t_{k-1}. Thus, $\boldsymbol{y}_k^{p}$ is also regarded as conditional to observed responses up to time t_{k-1}. The collection of measurements up to time t_{k-1} is denoted by $\boldsymbol{D}_{obs,k-1}$. The measurement errors are independent in time.

$$\boldsymbol{D}_{obs,k-1} = \{d_{obs,i} | 1 \leq i \leq k-1\}, \tag{3.16}$$

Assume that the prior of $\boldsymbol{y}_k^p$ is Gaussian distribution. Then

$$p(\boldsymbol{y}_k^p | \boldsymbol{D}_{obs,k-1}) \sim N(\boldsymbol{y}_k^p, \boldsymbol{C}_{Y,k}^p), \tag{3.17}$$

where $\boldsymbol{y}_k^p$ is computed by Eq. 3.15; and $\boldsymbol{C}_{Y,k}^p$ is prior covariance matrix associated with the prior estimate. It is explicitly computed by propagating an assumed initial covariance matrix of the state vector at time 0, $\boldsymbol{C}_{Y,0}^u$, through time,

$$\boldsymbol{C}_{Y,k}^p = \boldsymbol{\Psi}_{k-1} \boldsymbol{C}_{Y,k-1}^u \boldsymbol{\Psi}_{k-1}^{\mathrm{T}} + \boldsymbol{Q}_{k-1}, \tag{3.18}$$

where $\boldsymbol{C}_{Y,k-1}^u$ is the posterior covariance matrix after data are assimilated at time t_{k-1}. $\boldsymbol{C}_{Y,k}^p$ is associated with $\boldsymbol{y}_k^p$, which is,

$$\boldsymbol{C}_{Y,k}^p = E[\boldsymbol{e}_k^p (\boldsymbol{e}_k^p)^{\mathrm{T}}] = E[(\boldsymbol{y}_k^{true} - \boldsymbol{y}_k^p)(\boldsymbol{y}_k^{true} - \boldsymbol{y}_k^p)^{\mathrm{T}}], \tag{3.19}$$

Now $\boldsymbol{e}_k^p$ is the mismatch between true state and the state propagated from t_{k-1}. At the **update step**, with the new observation responses assimilated, $\boldsymbol{d}_{obs,k}$, the best estimate of $\boldsymbol{y}$ at each step of data assimilation is

$$\boldsymbol{y}_k^u = \boldsymbol{y}_k^p + \boldsymbol{K}_k (\boldsymbol{d}_{obs,k} - \boldsymbol{H}_k \boldsymbol{y}_k^p), \tag{3.20}$$

$\boldsymbol{d}_{obs,k} - \boldsymbol{H}_k \boldsymbol{y}_k^p$ is termed the measurement *innovation* or the *residual.* $\boldsymbol{K}_k$ is the Kalman gain matrix; it can be obtained by minimizing the error covariance matrix associated with $\boldsymbol{y}_k^u$.

$$\begin{aligned} \boldsymbol{K}_k &= \boldsymbol{C}_{Y,k}^p \boldsymbol{H}_k^T (\boldsymbol{H}_k \boldsymbol{C}_{Y,k}^p \boldsymbol{H}_k^T + \boldsymbol{C}_{D,k})^{-1} \\ &= \frac{\boldsymbol{C}_{Y,k}^p \boldsymbol{H}_k^T}{\boldsymbol{H}_k \boldsymbol{C}_{Y,k}^p \boldsymbol{H}_k^T + \boldsymbol{C}_{D,k}} \end{aligned} \tag{3.21}$$

where $\boldsymbol{C}_{D,k}$ is the measurement noise covariance at time t_k, $\boldsymbol{C}_{D,k} = E[\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^T]$, and $\boldsymbol{\epsilon}_k$ is the noise of measurement $\boldsymbol{d}_{obs,k}$. $\boldsymbol{Q}_{k-1}$ is the process noise covariance associated with model at

t_{k-1}, $\boldsymbol{Q}_{k-1} = E[\boldsymbol{\omega}_{k-1}\boldsymbol{\omega}_{k-1}^T]$. $\boldsymbol{\omega}_{k-1}$ is model errors, because the model is assumed to contain errors, *e.g.*, due to neglected physics and numerical approximations (*i.e.* truncation error and numerical dispersion). $\boldsymbol{C}_{Y,k}^p$ is the error covariance matrix associated with $\boldsymbol{y}_k^p$ and $\boldsymbol{y}_k^{true}$, $\boldsymbol{C}_{Y,k}^p = E[\boldsymbol{e}_k^p(\boldsymbol{e}_k^p)^{\mathrm{T}}]$. $\boldsymbol{Q}_{k-1}$ is included in $\boldsymbol{C}_{Y,k}^p$.

Looking at Eq. 3.21, we see that as the measurement error covariance $\boldsymbol{C}_{D,k}$ approaches zero, the denominator of the gain $\boldsymbol{K}_k$ becomes small, then the $\boldsymbol{K}_k$ increases, it weights the *residual* more heavily.

On the other hand, as the *a priori* estimate error covariance $\boldsymbol{C}_{Y,k}^p$ approaches zero, the gain $\boldsymbol{K}_k$ weights the *residual* less heavily. Specifically,

$$\lim_{\boldsymbol{C}_{Y,k}\to\boldsymbol{0}} \boldsymbol{K}_k = \boldsymbol{0} \tag{3.22}$$

Another way of thinking about the weighting by $\boldsymbol{K}_k$ is that as the measurement error covariance $\boldsymbol{C}_{D,k}$ approaches zero, the actual measurement $\boldsymbol{d}_{obs,k}$ is trusted more and more, while the predicted measurement $\boldsymbol{H}_k\boldsymbol{y}_k^p$ is trusted less and less. On the other hand, as the a priori estimate error covariance $\boldsymbol{C}_{Y,k}^p$ approaches zero the actual measurement $\boldsymbol{d}_{obs,k}$ is trusted less and less, while the predicted measurement $\boldsymbol{H}_k\boldsymbol{y}_k^p$ is trusted more and more. The state vector remains unchanged.

The posterior of the composite vector $\boldsymbol{y}_k$ is

$$p(\boldsymbol{y}_k^u|\boldsymbol{D}_{obs,k}) \sim N(\boldsymbol{y}_k^u, \boldsymbol{C}_{Y,k}^u), \tag{3.23}$$

After assimilating the observed responses, the error covariance matrix $\boldsymbol{C}_{Y,k}^u$ associated with $\boldsymbol{y}_k^u$ is

$$\begin{aligned}
\boldsymbol{C}_{Y,k}^u &= (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k)\boldsymbol{C}_{Y,k}^p(\boldsymbol{I} - \boldsymbol{H}_k^{\mathrm{T}}\boldsymbol{K}_k^{\mathrm{T}}) + \boldsymbol{K}_k\boldsymbol{R}\boldsymbol{K}_k^{\mathrm{T}} \\
&= \boldsymbol{C}_{Y,k}^p - \boldsymbol{K}_k\boldsymbol{H}_k\boldsymbol{C}_{Y,k}^p - \boldsymbol{C}_{Y,k}^p\boldsymbol{H}_k^{\mathrm{T}}\boldsymbol{K}_k^{\mathrm{T}} + \boldsymbol{K}_k(\boldsymbol{H}\boldsymbol{C}_{Y,k}^p\boldsymbol{H}_k^{\mathrm{T}} + \boldsymbol{R})\boldsymbol{K}_k^{\mathrm{T}} \\
&= (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k)\boldsymbol{C}_{Y,k}^p,
\end{aligned} \tag{3.24}$$

The term $(\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H}_k)$ ensures and implies that the covariance should (for stable and nonzero $\boldsymbol{K}_k$) decrease as more observed responses are integrated.

The two steps illustrated above at one measurement time are summarized in Table 3.1. With the generation of the state vector at the initial time 0 (the initial state vector is generated by sampling from a multivariate Gaussian distribution with the prior information about its mean and covariance matrix, $\boldsymbol{C}_{Y,0}^{u}$), the recursive process of the Kalman filter can be initiated. The two-step procedure is repeated until the last data are assimilated. The solution that the Kalman filter offers is based on maximizing the posterior PDF of the state vector within the context of Bayesian statistics. It is equivalent to minimizing the variances of the posterior covariance matrix in Eq. 3.24 with the assumption that the following variables are Gaussian (Maybeck 1979):

Table 3.1: Summary of two-step procedure of the Kalman filter at measurement t_k.

1.Evolve the state vector and covariance matrix forward in time	Eq. 3.15 $\boldsymbol{y}_k^p = \boldsymbol{\Psi}_{k-1}\boldsymbol{y}_{k-1}^u$ Eq. 3.18 $\boldsymbol{C}_{Y,k}^p = \boldsymbol{\Psi}_{k-1}\boldsymbol{C}_{Y,k-1}^u\boldsymbol{\Psi}_{k-1}^{\mathrm{T}} + \boldsymbol{Q}_{k-1}$
2.Update the state vector and its covariance matrix using the new data	Eq. 3.20 $\boldsymbol{y}_k^u = \boldsymbol{y}_k^p + \boldsymbol{K}_k(d_{obs,k} - \boldsymbol{H}_k\boldsymbol{y}_k^p)$ Eq. 3.24 $\boldsymbol{C}_{Y,k}^u = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k)\boldsymbol{C}_{Y,k}^p$

- model errors, ω_{k-1}
- measurement errors, ϵ_k
- estimate of state vector at the initial time 0

In addition to the Gaussian assumptions, the model and measurement errors are assumed to be unbiased and *white*. Whiteness implies that the noises are not correlated with time. The textbook definition of white noise is

> A **white random sequence** $x_n, n = 1, 2, ...$ is a Markov sequence for which
>
> $$p(x_k \mid x_l) = p(x_k) \quad (k > l) \tag{3.25}$$
>
> That is, all the x_k's are mutually independent. As a result, knowing the realization of x_l in no way helps in predicting what x_k will be. A white sequence is totally

> random or completely unpredictable. If the x_k's are all normally distributed, the x_k sequence is called a **white Gaussian random sequence**. The fact is that noise due to the superposition of a large number of small, independent, random effects is always Gaussian distribution because of central limit theorem. The importance and usefulness of white Gaussian sequences stems from the fact.
>
> — Jazwinski (1970),

If the dynamic model Eq. 3.14 is linear and the estimation of the initial state vector is Gaussian distribution, both the *a prior* and *a posterior* (Eqs. 3.17 and 3.23) pdfs would be Gaussian. But non-Gaussian variables violate assumptions of the Kalman filter and cause filter failure. For Gaussian variables, the mean and covariance (second order moments) are sufficient to describe a distribution. Higher moments are required for non-Gaussian distributions. In that case, estimates from the Kalman filter may not be optimal. Furthermore, when dealing with large and nonlinear systems, the propagation of the error covariance matrix by system dynamic is the main bottleneck for the Kalman filter method, imposing an unacceptable computation burden. The extended Kalman filter (EKF) proposed by Evensen (2003) uses the ideas of the Kalman filter to nonlinear systems based on first order linearizations and a closure assumption. The closure assumption is that only the first and second moments have been preserved. However, when nonlinearities are severe or the model size becomes large, the EKF is not an optimal estimator. In addition, if the initial estimate of the state is wrong, or if the process is modeled incorrectly, the filter may quickly diverge, owing to its linearization. Another problem with the extended Kalman filter is that the estimated covariance matrix tends to underestimate the true covariance matrix and therefore risks "filter divergence" without the addition of "stabilizing noise".

For large scale non-linear systems, a more promising approach is the ensemble Kalman filter. It is essentially a Monte Carlo approach, using an ensemble of model realizations to evaluate necessary statistics.

3.3.1 Ensemble Kalman Filter

Most problems in reservoir engineering are nonlinear and involve many variables, often two or more variables per simulator gridblock (in a black oil model with uncertain k and ϕ, 2 model parameters and 3 dynamic variables for a total of 5 unknowns per gridblock). Thus, the ensemble Kalman Filter for nonlinear systems (Evensen 1994) has received a lot of attentions. EnKF is a Monte Carlo method in which an ensemble of reservoir models is used for continuous model updating and reservoir monitoring. The correlation between reservoir responses (*e.g.*, pressure and phase saturation) and model parameters (*e.g.*, permeability and porosity) can be estimated from the ensemble.

The basic method of the EnKF is similar to the linear Kalman filter. It consists of the *forecast step* and the *assimilation step.* The *forecast step* is to advance the state vectors from the previous time step to the next time step, using any available solution method (for reservoir engineering, typically a finite difference simulator). At the time step where observations are available, the forecast pauses and the data are assimilated to update the state vectors. The following paragraphs introduce the EnKF using reservoir characterization terminology.

If forward modeling is achieved using a reservoir simulator, the state vector typically includes model parameters such as porosity ϕ, log permeability $\ln k$; and state variables such as pressure p_ℓ, and saturations S_ℓ at each reservoir simulation gridblock. Besides the model parameters and solutions, the state vector also includes the reservoir response output from the reservoir simulator, such as well bottom-hole pressure, water cut or water oil ratio and gas oil ratio. Thus, the state vector of a typical two-phase reservoir can be written as the following joint state vector:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{\phi} \\ \ln \boldsymbol{k} \\ \boldsymbol{p} \\ \boldsymbol{S}_w \\ \boldsymbol{p}_{wf} \\ \boldsymbol{q}_o \\ \vdots \end{bmatrix} (j, k) \tag{3.26}$$

where j is the ensemble member index and k is the time step index. If the number of gridblocks is N_m and the number of computed responses is N_d, the dimension of $\boldsymbol{y}_k$ is $4 \times N_m + N_d$ (for this particular choice of model parameters and state variables). When $k = 0$, the initial pressure vector p_0 and saturation S_{w0}, are initial conditions of the reservoir. In current EnKF applications, the initialization is not considered as uncertain, though they might be uncertain in actual reservoir (Evensen 2007). Similarly, the uncertainties of initial pressure and phase saturation can be taken into consideration the same as model parameters.

3.3.2 Forecast Step for Reservoir Models

The forecast step is used to evolve the state vector forward in time between two consecutive measurement times to get the state solutions. The dynamic system model in a mathematical form should be

$$\boldsymbol{y}^p_{j,k} = \boldsymbol{\Psi}(\boldsymbol{y}^u_{j,k-1}) \ \ (j = 1, 2, \ldots, N_e), \tag{3.27}$$

where j is the ensemble member index and N_e is the number of ensemble members, using u to denote *updated* and p to denote *predicted*. $\boldsymbol{\Psi}$ is the porous media partial differential equations introduced in Section 3.1. Note that only model solutions, *i.e.*, pressure and saturations, and the computed responses change between $k-1$ and k. The static variables, *i.e.*, porosity and permeability, or other reservoir properties remain unchanged during the same time interval, that is $\boldsymbol{m}^p_{j,k} = \boldsymbol{m}^u_{j,k-1}$. They are however adjusted as well as the dynamic variables during the assimilation step.

Simulators are used to model flow in porous media. In this research, the following simulators are used:

1. Academic simulator, UTCHEM (UTCHEM 2008)
2. Commercial simulator, Eclipse (ECLIPSE 100 2008)

In EnKF, the evolution of all ensemble members may take considerable time, especially reservoir models with complex permeability heterogeneity. However, because of the independence of the ensemble models at the forecast step, the evolution of multiple models can proceed simultaneously using parallel computing, such as Grid computing.

3.3.3 Update Step for Reservoir Models

At time step k, some measurement data are obtained and the state vectors are updated using their prior values from the forecast step and Kalman gain matrix, $\boldsymbol{K}_{e,k}$.

$$\boldsymbol{y}^u_{j,k} = \boldsymbol{y}^p_{j,k} + \boldsymbol{K}_{Y,e,k}(\boldsymbol{d}_{obs,k} - \boldsymbol{H}_k \boldsymbol{y}^p_{j,k}) \quad (j = 1, 2, \ldots, N_e), \tag{3.28}$$

The subscript e indicates that the items are calculated from the ensemble. In contrast to the Kalman filter, $\boldsymbol{K}_{Y,e,k}$ is computed from the ensemble members using the following equation,

$$\boldsymbol{K}_{Y,e,k} = \boldsymbol{C}^p_{Y,e,k} \boldsymbol{H}^T_k (\boldsymbol{H}_k \boldsymbol{C}^p_{Y,e,k} \boldsymbol{H}^T_k + \boldsymbol{C}_{D,k})^{-1}, \tag{3.29}$$

In the Kalman filter, the *prior* error covariance matrix $\boldsymbol{C}^p_{Y,k}$ is always calculated explicitly by time evolution equation Eq. 3.18, posing a significant computation burden for large or nonlinear problems. But in the ensemble Kalman filter, the *prior* covariance matrix $\boldsymbol{C}^p_{Y,e,k}$ is estimated from ensemble members by the statistics in Eq. 3.30.

$$\boldsymbol{C}^p_{Y,e,k} = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (\boldsymbol{y}^p_{k,j} - \bar{\boldsymbol{y}}^p_k)(\boldsymbol{y}^p_{k,j} - \bar{\boldsymbol{y}}^p_k)^{\mathrm{T}}, \tag{3.30}$$

where $\bar{\boldsymbol{y}}_k^p$ is the averaged variables in the state vector and is computed by

$$\bar{\boldsymbol{y}}_k^p = \frac{1}{N_e}\sum_{j=1}^{N_e}\boldsymbol{y}_{j,k}, \tag{3.31}$$

Note that $N_e - 1$ instead of N_e is used in Eq. 3.30 to ensure the estimate is unbiased. Although the preliminary loop is a good starting point, it is not a practical algorithm because of the matrix computation expense. In Eq. 3.30 (for our choice of two model parameter and two state solution), the dimension of $\boldsymbol{C}_{Y,e,k}^p$ is $(4 \times N_m + N_d) \times (4 \times N_m + N_d)$. To explicitly form and store this matrix requires too much work, which makes the loop impractical even for a medium size problem. Therefore, to avoid the explicit computation of $\boldsymbol{C}_{Y,e,k}^p$, Eq. 3.30 can be rewritten as

$$\boldsymbol{C}_{Y,e,k}^p = \frac{1}{N_e - 1}\Delta\boldsymbol{Y}_k^p(\Delta\boldsymbol{Y}_k^p)^{\mathrm{T}}, \tag{3.32}$$

where $\Delta\boldsymbol{Y}_k^p$ consists of N_e column vectors, each of which is the difference between an ensemble state vector and the averaged state vector. Applying Eq. 3.32 to Eq. 3.29, the Kalman gain matrix has the new form:

$$\begin{aligned}\boldsymbol{K}_{Y,e,k} &= \frac{1}{N_e - 1}\Delta\boldsymbol{Y}_k^p(\Delta\boldsymbol{Y}_k^p)^{\mathrm{T}}\boldsymbol{H}_k^{\mathrm{T}}\left(\boldsymbol{H}_k\frac{1}{N_e - 1}\Delta\boldsymbol{Y}_k^p(\Delta\boldsymbol{Y}_k^p)^{\mathrm{T}}\boldsymbol{H}_k^{\mathrm{T}} + \boldsymbol{C}_{D,k}\right)^{-1} \\ &= \frac{1}{N_e - 1}\Delta\boldsymbol{Y}_k^p(\boldsymbol{H}_k\Delta\boldsymbol{Y}_k^p)^{\mathrm{T}}\left(\frac{1}{N_e - 1}\boldsymbol{H}_k\Delta\boldsymbol{Y}_k^p(\boldsymbol{H}_k\Delta\boldsymbol{Y}_k^p)^{\mathrm{T}} + \boldsymbol{C}_{D,k}\right)^{-1},\end{aligned} \tag{3.33}$$

If we define

$$\boldsymbol{A} = \boldsymbol{H}_k\Delta\boldsymbol{Y}_k^p, \tag{3.34}$$

Then Eq. 3.33 becomes

$$\boldsymbol{K}_{Y,e,k} = \frac{1}{N_e - 1}\Delta\boldsymbol{Y}_k^p\boldsymbol{A}^{\mathrm{T}}\left(\frac{1}{N_e - 1}\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{C}_{D,k}\right)^{-1}, \tag{3.35}$$

In practice, it is not necessary to compute an approximation of the covariance matrix, because only the product of $\boldsymbol{C}_{Y,e,k}^p\boldsymbol{H}_k^T$ is required to compute the weight matrix (Eq. 3.33). The covariance matrix can have fairly large dimensions to $N_{y,k} \times N_{y,k}$, whereas $\boldsymbol{C}_{Y,e,k}^p\boldsymbol{H}_k^T$ has reduced dimensions, $N_{y,k} \times N_{d,k}$. $N_{y,k}$ is the length of state vector and $N_{d,k}$ is the number of observed responses.

3.3.4 Implementation of EnKF

$\boldsymbol{H}_k$ is the joint covariance forward observation operator. The expected value of an observation can be calculated by applying $\boldsymbol{H}_k$ to the state vector $\boldsymbol{y}$, which is equivalent to applying the possible nonlinear operator $\boldsymbol{g}$ to independent variables $\boldsymbol{x}$. Since $\boldsymbol{H}_k$ is the operator matrix used to extract entries corresponding to data, $\boldsymbol{H}_k \Delta \boldsymbol{Y}_k^p$ does not involve matrix computation. In addition, only $\Delta \boldsymbol{Y}_k^p$ needs to be formed and stored. Since the dimension of $\Delta \boldsymbol{Y}_k^p$ is $(4 \times N_m + N_d) \times N_e$, this approach is more efficient than computing $\boldsymbol{C}^p_{Y,e,k}$. Thus, a feasible EnKF algorithm is formed,

a) Input the ensemble state vectors into the reservoir simulator and advance them in time. If it is the first time step, all the vectors are filled with initial values.

b) At the time step k when the observations are available, stop advancing and fill the state vectors with the $k-1$ step model parameters and the k time step model solutions.

c) Compute the averaged state vector using,

$$\bar{\boldsymbol{y}}_k^p = \frac{1}{N_e} \sum_{j=1}^{N_e} \boldsymbol{y}_{k,j}^p, \tag{3.36}$$

d) Form the difference matrix $\Delta \boldsymbol{Y}_k^p$ and take entries from $\Delta \boldsymbol{Y}_k^p$ using $\boldsymbol{H}_k$, which is $\boldsymbol{H}_k \Delta \boldsymbol{Y}_k^p$.

e) Compute the Kalman gain matrix using Eq. 3.33.

f) Update the ensemble state vectors using Eq. 3.28.

g) If the current time step is the final step, then STOP. Otherwise, go back to Step a).

There are three numbers related to the dimension of the matrices: N_y, N_d, and N_e. For reservoir models, N_y could easily exceed 10^6 for field-scale problems. N_d is the number of observed responses available at one measurement time, usually the number of production rate measurements of wells (order 10^1–10^2). N_e is the number of ensemble members which is determined according to estimation reliability, the complexity of the flow model, and computational resources. Usually for serial work flow, the number is $O(100)$. In the Grid computing environment, the number could be $O(1000)$. The evolution of the state vectors dominates the computational cost in the EnKF. The total computation time of serial work flow equals to the simulation run time for all ensemble members plus overhead time involved in matrix calculation at the update steps. In Grid environment, the simulation run time can be greatly shortened by distributed computing. But the queue time and time cost for data transferring through the network also need to be considered.

3.4 Statistical Measures of the EnKF Performance

The spread about the truth case ($\boldsymbol{\sigma}_t^2$) is a standard criterion used to measure the difference between the estimate mean and the true state vector for synthetic case in Eq. 3.37, or more directly using model parameters in Eq. 3.38:

$$\boldsymbol{\sigma}_t^2 = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (\boldsymbol{y}_{i,j} - \boldsymbol{y}_i^{true}) \boldsymbol{C}_{Y,e,k}^{-1} (\boldsymbol{y}_{i,j} - \boldsymbol{y}_i^{true})^{\mathrm{T}} \quad (i = 1, 2, \ldots, N_m) \tag{3.37}$$

$$\boldsymbol{\sigma}_t^2 = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (\boldsymbol{m}_{i,j} - \boldsymbol{m}_i^{true}) \boldsymbol{C}_{M,e,k}^{-1} (\boldsymbol{m}_{i,j} - \boldsymbol{m}_i^{true})^{\mathrm{T}} \quad (i = 1, 2, \ldots, N_m) \tag{3.38}$$

where N_m is the number of the gridblocks of the model; $\boldsymbol{y}_{i,j}^{true}$ and $\boldsymbol{m}_{i,j}^{true}$is the truth of state vector and reservoir properties, respectively. And $\boldsymbol{y}_{i,j}$ and $\boldsymbol{m}_{i,j}$ is the state vector and model parameters of ensemble members, respectively. As pointed by others (Zafari 2007), it is difficult to define metrics to provide a reliable characterization of uncertainty in the model parameters. $\boldsymbol{\sigma}_t^2$ is a measure of the accuracy of the ensemble mean.

But with ensemble models, we can calculate the spread of the samples at each gridblock:

$$\boldsymbol{\sigma}_e^2 = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (\boldsymbol{m}_{i,j} - \bar{\boldsymbol{m}}_i) \boldsymbol{C}_{M,e,k}^{-1} (\boldsymbol{m}_{i,j} - \bar{\boldsymbol{m}}_i) \quad (i = 1, 2, \ldots, N_m) \tag{3.39}$$

which is representative of the estimated uncertainty of the ensemble. If the EnKF estimate of uncertainty is correct, the $\boldsymbol{\sigma}_t^2$ and ensemble spread $\boldsymbol{\sigma}_e^2$ should be almost identical. The observation error represents the mismatch between computed observation from simulation and measurement,

$$\sigma_d^2 = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (\boldsymbol{d}_{obs} - \boldsymbol{g}(\boldsymbol{m})) \boldsymbol{C}_d^{-1} (\boldsymbol{d}_{obs} - \boldsymbol{g}(\boldsymbol{m})) \tag{3.40}$$

3.5 Discussion

From the introduction of the Kalman filter, it is obvious the there are two problems when KF is used for data assimilation in high dimensional and nonlinear system. The first is related to storage and computational issues. If the reservoir has N_m unknowns in the state vector, then the error covariance matrix $C^p_{Y,k}$ has N^2_m unknowns. The evolution of the error covariance matrix in time requires cost of $2N_m$ model integrations. Thus, KF and EKF can only be applied to fairly low-dimensional models. The second issue is related to the use of the EKF with nonlinear models, which requires a linearization when evolving the error covariance to new time step. This linearization leads to a poor error covariance evolution and unstable error covariance grows for some strongly nonlinear models. This may be resolved using higher order closure schemes. But the method is not practical for high dimensional model, since the fourth order moment requires storage of N^4_m elements.

In the EnKF approach, the difficulties that prevented the adjoint and sensitivity methods from being widely applied are minimized. First, the EnKF is applicable to problems with frequent data acquisition based on its Bayesian formulation. Second, because it is derivative-free, it does not depend on the specific reservoir simulator because adjoint or sensitivity does not need to be computed explicitly. It only requires output from the simulator, such as pressure, phase saturation and production data. Coding for the EnKF algorithm can be adapted to any reservoir simulator on a “plug-in” basis (Gu and Oliver 2005). Third, EnKF reduces a nonlinear minimization problem in a huge parameter space involving the minimization of an objective function with multiple local minima to a statistical minimization problem in the ensemble space. Thus, by searching for the mean rather than the (many) mode(s) of the posterior pdf, the method avoids getting trapped in local minima (Evensen et al. 2007). Finally, the ensemble Kalman filter (EnKF) method takes one simulation run per reservoir model realization and each one is independent with each others. Therefore, the simulations of the reservoir models in the ensemble are ideal for distributing to supercomputing environments or Grid computing. There are far fewer runs than other methods, such as randomized maximum likelihood, and it samples more efficiently than most MCMC methods do (Gao, Zafari, and Reynolds 2006).

Chapter 4

Application of the Ensemble Kalman Filter to Continuous Model Updating on Grid Environment

Real-time Model updating by EnKF is processing-intensive because it requires simulation of many reservoir models. In addition, large datasets (comprising model, state vector and Kalman gain) must be transferred from member-specific simulation processors to the Kalman gain processor, and then back to the member processor at each assimilation step. Between assimilations, the simulation runs of the ensemble members are independent, and therefore they can be distributed to Grid resources.

These features motivate a distributed computing solution across multiple machines to increase throughput. However, because the Kalman gain computation requires synchronization of all ensemble members at each assimilation step (Evensen 2003), the EnKF is constrained by the slowest simulation (caused by slower processors or more difficult flow modeling). Moreover, some ensemble member results may be lost because of algorithmic or hardware failure on distributed clusters. This requires a work flow which can reduce time spent waiting on synchronization or rerunning failed simulations *i.e.*, overhead (Lei and Allen 2007). Section 4.1 introduced the EnKF on Grid work flow adapted from ResGrid. Section 4.2 is a 2-D water flood example to validate the distributed EnKF.

4.1 Description of Grid Computing for EnKF

The simulation management for EnKF is based on previous work on a Grid computing based task-farming toolkit, ResGrid (Lei et al. 2006), introduced in Chapter 2. ResGrid combines elements of experimental design, response surface models, uncertainty and sensitivity analysis. The workflow has been adapted to the EnKF. In addition, the ResGrid portal provides a web-based entry point for reservoir engineers to access the Grid, concealing many complexities and technical details of job scheduling and resource management from end users.

Dynamic Assignment and Task Container (DA-TC) concepts are used to shorten queue time when synchronizing ensemble members at the assimilation points; the goal is to avoid reëntering at the back of the queue at each forecast step. Task containers are queued on remote clusters as normal jobs. Any cluster may host multiple containers, depending on resources and the load balancing strategy (Lei and Allen 2007). DA-TC dynamically assigns tasks to task containers as containers obtain cluster resources. Once a container has been allocated resources, it persists as a job and therefore retains the resources until all simulations for all members assigned to the container are completed. Thus, all simulation runs assigned to a container have only one queue wait, and dynamic task assignment allows containers with high-performance resources to execute more tasks. With many members and simulations per container, this reduces total queue time. The workflow of EnKF on Grid is illusrated in the following Figure 4.1. The essential characteristics of the workflow are:

1. Use model parameter sampler to sample the initial ensemble members.
2. Distribute the ensemble metadata over Grid resources (so called "task farming").
3. Build flow models with ensemble metadata and flow model parameters in parallel.
4. At time k when the new sensor data are recorded, distribute the flow model to remote resources. In this step, the simulation jobs are submitted to DA-TC containers which are pre-submitted to Grid resources according to resources status and loading balance strategies.
5. Advance the simulations from t_{k-1} to t_k.
6. After all machines finish the tasks, the Stage Out module transmits the new state vectors from each machine (or ensemble member) to the Kalman gain processor and the Kalman gain is computed by integrating observation data from sensors and observation error models.
7. Task farming the ensemble Kalman gain to member processors.

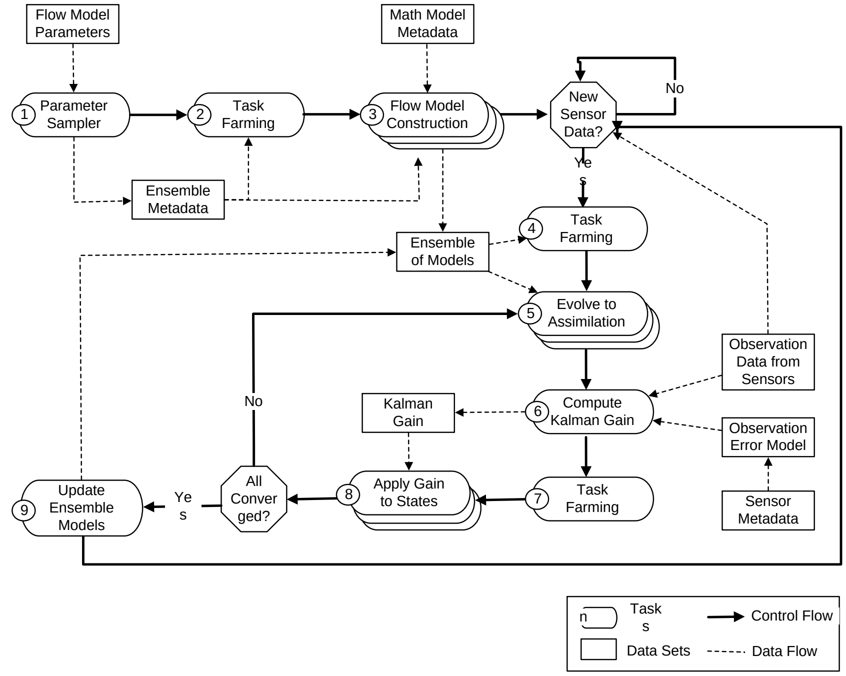


Figure 4.1: Workflow for Automatic history matching by EnKF on Grid

8. Update the state vectors of ensemble, then EnKF convergence check to ensure the updated state vectors are within the physical boundaries.
    (a) If the convergence check is true, go on to the next step.
    (b) If the convergence check is false, then truncate the nonphysical value or use reparameterization method (Chen, Oliver, and Zhang 2007).
9. Update the ensemble models.
10. Wait for new observation data from sensor.

The data transfer load of GridFTP (main function of Stage Out Module) and resources are balanced in the work flow design to improve the computational efficiency. According to the complexity of simulation models, two designs are implemented for Step 8. One is suitable for small- to medium-scale reservoir simulation cases (gridblocks in the magnitude of $O(10^5)$). Taking the following case as an example, the state vector plus the simulation data deck for one ensemble member is about 300 KB (30 MB for 100 realizations) and UTCHEM executive file size is about 2.6 MB. It takes trivial seconds of transfer time for high-bandwidth networks. We can neglect the cost for data transfer and persue the flexible job submission to containers based on load strategy to save more time. Therefore, after GridFTP brings the state vectors back from distributed machines in Step 6, all the history files there are deleted. Both the model update and convergence check are complete at Kalman gain processor. In the next forecast step, the updated models (including updated state vectors) are distributed to remote resources.

For larger, more complex reservoir models, the total transfer data may reach 10 GB and the costs of data transportation at network must be taken into consideration. In Step 7, GridFTP only brings the model solutions at ensemble processors to the Kalman gain processor. The Kalman gain is then calculated and transferred back to Grid resources, both the update step and convergence check will take place at member processors based on locally-stored state vectors.

The total computation time includes the simulation time for all ensemble models, plus time for Kalman gain calculation at the update steps, plus queue waiting time (from job submission to execution) at Grid resources and transfer time. The following example study has 400 ensemble members. Assimilation occurs every 10 days up to 160 days in the case study. If there are nonphysical values generated for updated water saturation (*e.g.*, $S_w \notin [0, 1]$) or other state variables in the assimilation step, the non-physical values are truncated after the assimilation. There are 400 members $\times$ 32 assimilations, or 12,800 member updates.

Average processor time per simulation by UTCHEM is 10 minutes; the total time for simulation of the ensemble is about $2,133$ hrs (88 days) using 1 processor. In our case, all the simulations are submitted to 256, 15, or 14 processor clusters (running a mix of Linux and AIX). We use 5 to 10 containers depending on cluster size; each container uses 1 processor. Using 10 containers for 256 processor machines, 5 containers for 15 and 14 processor clusters, the execution time is $\sim$ 106 hours. Assuming the queue waiting time is 5 hours for each forecast step, the total time for this EnKF processing is $106 + 5 \times 3$ (number of machines) = 121 hours using the DA-TC mechanism because there is only one wait for each cluster. The total time will increase to $100 + 5 \times 32$ (assimilation times) = 260 hours if DA-TC were not used and the queue reëntered for each forecast step. The queue waiting time depends on the Grid cluster status, and may range from minutes to days. If the production history is quite long, the cumulated queue time will increase and reduce or even eliminate the advantages of Grid computing.

4.2 2-D Waterflood Analysis Using the EnKF

The EnKF workflow is applied to a 2D waterflooding reservoir model updating using injection and production rates. The goal is to explore the characteristics of EnKF and sensitivity of EnKF to various factors. A 2D geostatistical permeability field is used. The model size is $16 \times 16 \times 1$ blocks with gridblock dimensions of $60 \times 60 \times 10$ ft. The simulation area is $960 \times 960 \times 10$ ft. The initial realizations are generated by unconditional LU decomposition simulation (Goovaerts 1997). The log permeability is assumed to follow Gaussian distribution. The mean and variance of $\ln k$ are 5.5 and 1 respectively, and it is assumed to be second-order stationary with a exponential covariance function, which is defined as

$$C_{\ln k}(\boldsymbol{h}) = \sigma^2_{\ln k}(\exp[-\frac{|h_x|}{a_x} - \frac{|h_y|}{a_y}]), \tag{4.1}$$

The variogram is exponential where $\boldsymbol{h} = (h_x, h_y)^{\mathrm{T}}$ is the vector and a_x and a_y are ranges of 10 and 5 gridblocks (600 and 300 ft) in the directions of 45 and 135 degrees (relative to x). Four hundred and one realizations are generated and one realization is randomly chosen as the “truth” for comparison with EnKF results.

The simulation is a five-spot well pattern which is initially at uniform, irreducible 20 percent water saturation. The injector (in the center) has constant bottom-hole pressure of 4,500 psi. Four producers (at each of the corners) have bottom hole pressure constraints of 1,500 psi. Quadratic relative permeability curves are used with 0.2 residual saturation for oil. Compressibility and capillary pressure are neglected. Measurement error is assumed to be 5 percent for oil and water production rates before water breakthrough, 8 percent for water production rates after water breakthrough at producers. The noises are mutually uncorrelated and therefore the error covariance of matrix $\boldsymbol{C}_{D,k}$ of the observations is diagonal (but not constant). The measurement errors are simplified here but in the real oil and gas field, the measurement errors of oil production rate and water cut may be correlated with each other because they are measured in the same separator. The EnKF system can handle more complicated cases and is not limited to the simplification like this.

If the covariance matrix is not diagonal, a singular value decomposition (SVD) can be performed on $\boldsymbol{C}_{D,k}$ (Zafari 2007). The prior joint state ensembles can be projected onto the singular vectors and the assimilation can proceed using this new basis, in which $\boldsymbol{C}'_{D,k}$, the observation covariance matrix, is diagonal. Upon completion of the assimilation computation, the updated state vectors can be projected back to the original state space. Sequential assimilation observations can have general (correlated) Gaussian distributions by the application of SVD. The method is useful in the case, for example, the oil production rate is correlated to the water cut. Here, we assume the observation errors are uncorrelated spatially and temporally.

In the example, the state vector has 3×256 elements plus 9 observations: The adjusted model parameter is $\ln k$; the model solutions are pressure and saturation of gridblocks; the observation part includes the simulated oil and water cuts of producers (4 producers $\times$ 2 observations = 8 measured data) and injection rate of the injector (the 9^{th} datum). The "true" observations are read from the simulation using the reference $\ln k$ (Figure 4.3(a)).

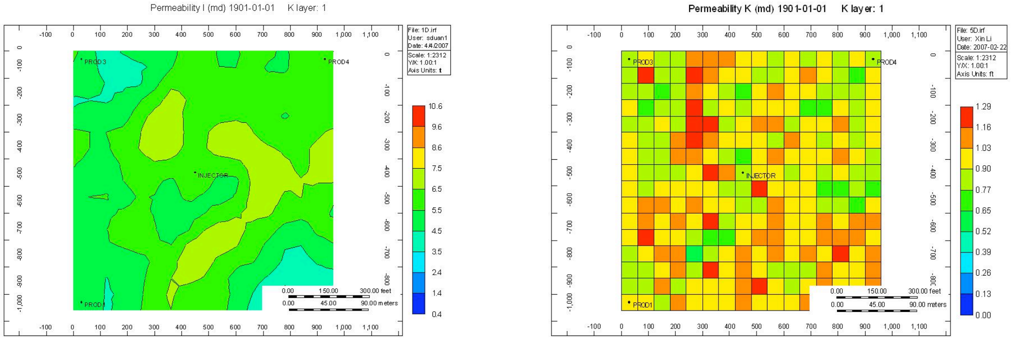
In this model, the total simulation time is 160 days (1.24 pv, 1 pv = 1 pore volume, $V_p = \sum_{i=1}^{N_m} (\Delta x \Delta y \Delta z \phi)_i$). The assimilation step is ten days ($\sim$ 0.08 pv). The porosity is 20 percent throughout the model. Because the same mathematical model is used for reference production forecast and ensemble forward forecast step, the model errors, such as truncation error, are not considered in this EnKF study.

4.2.1 Behavior of the EnKF

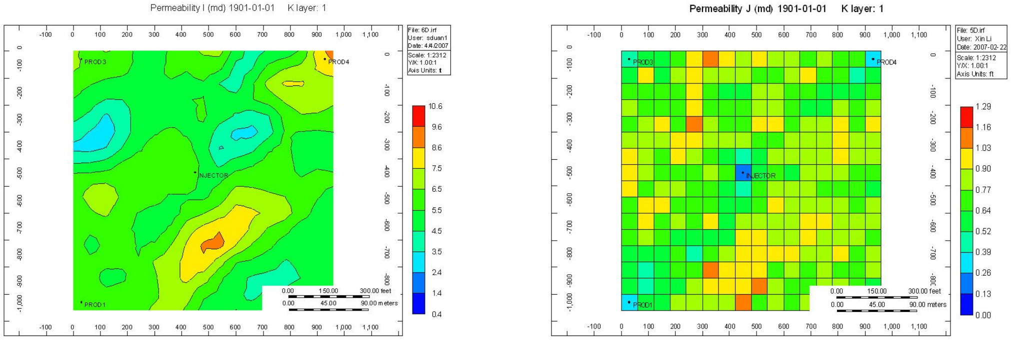
The ensemble mean and variance of the initial 400 realizations, and the estimated mean of $\ln k$ fields at 0.18 pv and 0.70 pv for assimilation steps are plotted in Figure 4.2. Figure 4.2(a) shows a constant mean for every gridblock because of second-order stationarity, but the individual ensemble member, including the prior geostatistical data of the reservoir characterization, has the different reservoir property. Figure 4.2(b) shows the initial fluctuation of $\ln k$ at each gridblock. The 0.18 pv assimilations (Figs. 4.2(c)) indicate that the ensemble mean $\ln k$ gradually includes the major features of reference distribution, and reduces the ensemble variability among ensemble members (Figs. 4.2(d)).

Prior to assimilation the ensemble variance is as high as 1.29 at the initial time (Figure 4.2(b)). The variance decreases dramatically (Fig. 4.2(d)), especially along the main flow direction between injector (I1) and producers (P1 and P4). This is because the flow rate is higher in the main flow direction, which leads in large changes in pressure and saturation, therefore, more information is involved in the observations around the main flow direction and hence a quick reduction of the estimation variance. Later in the simulations (Fig. 4.2(f)), the water has a breakthrough at all wells except P2, the ensemble variance is low throughout the reservoir, but still the flow portions between the injector and producers have less variance than the far-removed part with less information from wells. After the 0.70 pv assimilation, the ensemble mean shows the filter divergence characteristics: the $\ln k$ values are out of range and different maximum continuity direction from the truth.

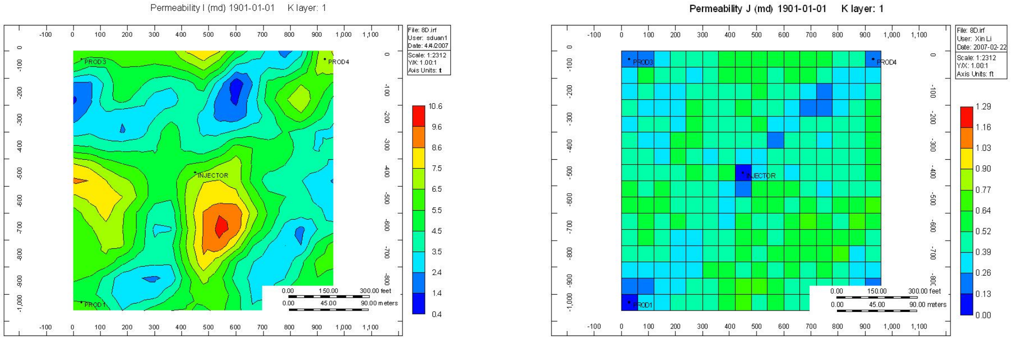
The $\ln k$ fields of two randomly selected individual members after assimilation at $\tau = 0.70$ are compared with truth and ensemble mean (Fig. 4.3). Although they obey the geostatistical data, the properties are significantly different from truth initially. After the production data assimilations, these two fields become fairly similar with each other and ensemble mean, while the updated permeability distributions are not “close” to the reference permeability, whether one considers the mean or particular ensemble members. The inversion is most



(a) Ensemble mean of the initial realizations (b) Ensemble variance of the initial realizations

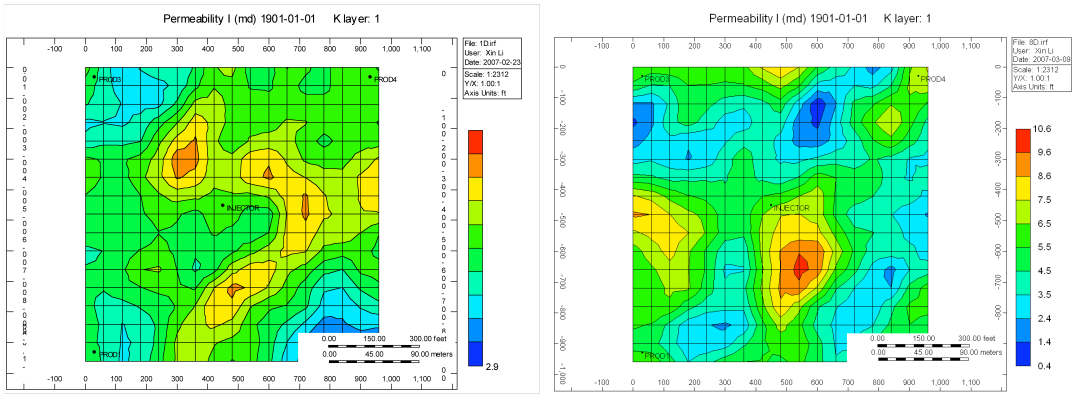


(c) Ensemble mean, τ=0.18 assimilation (d) Ensemble variance, τ=0.18 assimilation



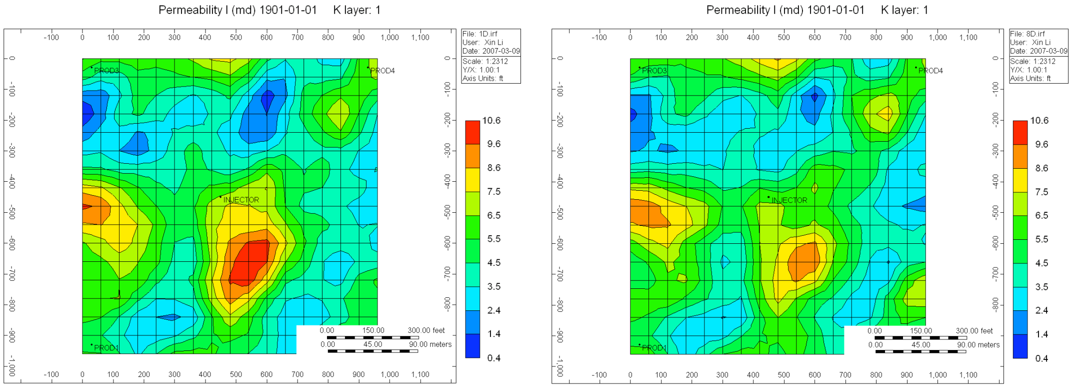
(e) Ensemble mean, τ=0.70 assimilation (f) Ensemble variance, τ=0.70 assimilation

Figure 4.2: The evolution of mean and variance contours of 400 realizations, τ=0.18 and 0.7 assimilations.



(a) $\ln k$ reference field

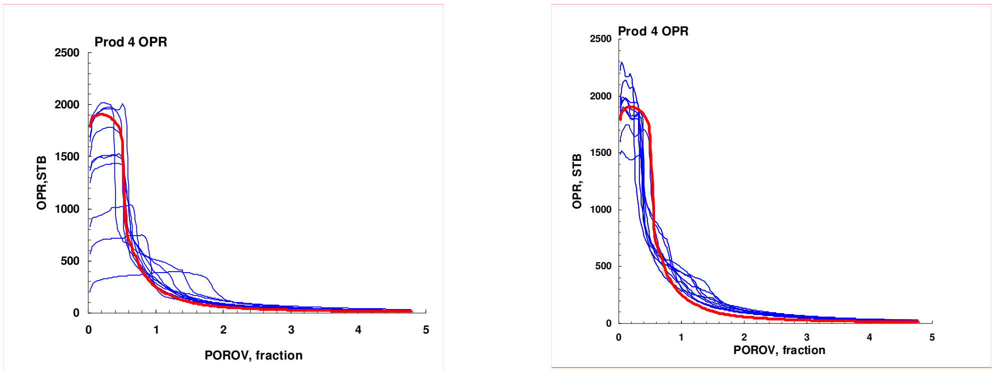
(b) $\ln k$, ensemble mean, τ=0.7 assimilation



(c) $\ln k$, member No.150, τ= 0.7 assimilation

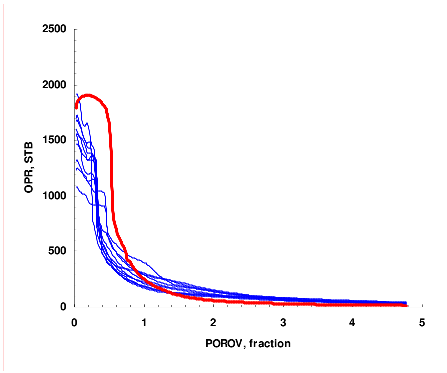
(d) $\ln k$, member No. 350, τ=0.7 assimilation

Figure 4.3: $\ln k$ contours comparison among ensemble members, ensemble mean and reference.



(a) Initial ensemble members

(b) τ=0.18 assimilation



(c) τ=0.7 pv assimilation

Figure 4.4: Oil production rate forecast of P4 in 200 days (red is reference).

accurate where fluxes are high (*e.g.*, along the streamlines connecting the injector to producers). The uncertainties are reflected among realizations, especially at locations where the heterogeneity features are less informed by flow between wells.

Fig. 4.4 indicates the prediction of well performance improved after more update steps. However, after the 0.7 pv assimilation step, the forecast ensemble shows the divergent trend.

4.3 Discussion

In this chapter, EnKF is applied to continuously update reservoir models by assimilating injection and production observations from wells. EnKF provides updated estimates of model parameters and their uncertainty. We used the EnKF to estimate and predict transient 2-phase flow in heterogeneous reservoir and analyze the predictability of assimilated models with a synthetic 2D examples by EnKF on Grid workflow.

1. Reservoir model updating and performance forecasting can be obtained relatively quickly by efficiently using a distributed EnKF on Grid environments.

2. The estimation of model parameters improved with integrating more observation data. The inverted permeability distributions are close to the reference permeability, whether one consider the mean or particular ensemble members. The inversion is most accurate where fluxes are high (*e.g.*, along the streamlines connecting the injector to producers). The uncertainties are reflected among realizations, especially at the part where flow does not sweep through.

3. The forecast mismatch indicates the predictability of model improved after several update steps. However, after some time, the forecast ensemble diverges, *e.g.*, the 0.7 pv assimilation in the example. The topic will be discussed in Chapter 6.

Chapter 5
Model Inversion of a 3-D Synthetic Case

In Chapter 4, we have shown a 2-D synthetic water flood example. In this chapter, a serial work flow was built on a single processor linux machine as a prototype to better understand the behavior of EnKF. A much bigger and more realistic 3-D reservoir model, PUNQS3, shows that the work flow is effective and robust.

5.1 Introduction to the PUNQ-S3 Model

The model history, geological setting, model properties and production are introduced in this section.

5.1.1 Model History

PUNQ-S3 is a synthetic reservoir engineering model based on a field operated by Elf Exploration and Production. The PUNQ project is a joint effort of 10 European companies, universities, and research centers supported by the European Union to compare methods for quantifying uncertainty assessment in history matching. PUNQ is an acronym for Production forecasting with UNcertainty Quantification.

A detailed description of the PUNQ-S3 reservoir simulation model can be found in Floris et al. (2001) or Barker, Cuypers, and Holden (2001). All the data are also available on the website of Department of Earth Science and Engineering of Imperial College (PUNQ-S3 Model for Quantifying Uncertainty in Production Forecast 2008).

5.1.2 Model Properties

The top depth of PUNQ-S3 reservoir is 2340 m. The dip angle is about 1.5 degree. It is bounded by a fault to the east and south and with a fairly strong aquifer on the north and

Table 5.1: PUNQ-S3 model summary

Parameter	Value	Units
xyz grid size	$19 \times 28 \times 5$	
Total blocks	2,660	
Active blocks	1,761	
Block size	$180 \times 180 \times 3.7$	meter
Gas-oil contact	2395	meter

west. Because of the strength of the aquifer, no injection wells are drilled. There is a small gas cap in the reservoir top. Model dimensions are summarized in Table 5.1.

Six producers are denote as black dots in Figure 5.1. They are located near the initial gas-oil contact, their position and perforated layers are listed in Table 5.2. Positions for five extra infill wells (X1-X5) are denoted as white dots in the figure, but we shall not discuss any results with the infill wells. The detailed information of porosity and permeability fields generation with geostatistical techniques is addressed in Appendix B.

5.2 Observation Data

The revealed true reservoir simulation model at the PUNQ web page was taken and run on the simulator to provide the true production data of 16.5 years; this will serve as our "truth" case. Only production data from the first 8 years are used to calibrate an ensemble of geological models. The final corrected models are used to predict recovery for the next 8.5 years of production for a given specified scheme. The prediction of the total oil production at the end of 16.5 years is compared with the results obtained by Barker, Cuypers, and Holden (2001). The production history of six production wells are summarized as follows,

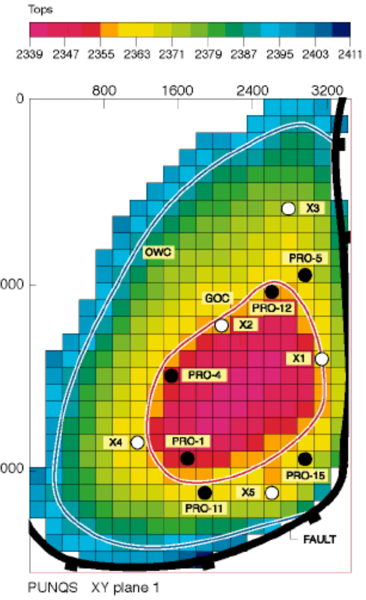


Figure 5.1: Top of structure map for the PUNQ-S3 reservoir. The units on the x, y, and "Tops" scale are in meters

Table 5.2: PUNQ-S3 well locations and perforated layers

Well name	Location: (x, y)	Perforated layers
PRO-1	(10, 22)	4, 5
PRO-4	(9, 17)	4, 5
PRO-5	(17, 11)	3, 4
PRO-11	(11, 24)	3, 4
PRO-12	(15, 12)	4, 5
PRO-15	(17, 22)	4

(1) an extended well testing period during the first year (four different flow rates, each lasting three months)

(2) a shut-in period for the following three years

(3) a 12-year production period with fixed oil production rate, 150 sm^3/day

(4) each well is shut in two weeks for testing each year in production period

PRO-4 has water breakthrough during the 7th year because it is close to the strong aquifer in the west. PRO-1 and PRO-4 start to produce free gas during the 4th and 5th years, respectively. Within the 16.5 year production period, each well has a target oil rate of 150 sm^3/day and a minimum bottom-hole flowing pressure 120 bar (1bar = 14.5 psi); if the maximum gas-oil ratio is greater than 200 sm^3/sm^3, the oil production rate is cut back by a factor of 0.75.

Bottom-hole pressure, gas-oil ratio, water cut and oil production rate are used as the assimilation data. Although the target oil production rate is identical for all reservoir models, the actual oil production rates vary because wells in some models are unable to attain the target rate 150 sm^3/sm^3, which may change the well to the bottom-hole pressure constraint 120 bar. The "predicted" bottom-hole pressure is set equal to the minimum bottom-hole pressure and thus no longer represents a prediction based on the corrected model by EnKF. This diminishes the reliability of the EnKF (and other inversion methods), and can cause errors ("filter divergence"), which is discussed in Chapter 6. The problem can be solved by applying an additional assimilation with the oil production rate (with a small measurement error). If a model changes to bottom-hole pressure constraint, the assimilation with a small error allows the filter to bring the rate data back to the historic production rate.

The state vector is in the following form,

$$
\begin{aligned}
\boldsymbol{y} =& [\phi_1, \ldots, \phi_N, \ln k_1, \ldots, \ln k_N, \ln k_{z,1}, \ldots, \ln k_{z,N}, \\
& p_1, \ldots, p_N, S_{w,1}, \ldots, S_{w,N}, S_{g,1}, \ldots, S_{g,N}, \\
& d_1, \ldots, d_{N_d}]^{\mathrm{T}}
\end{aligned}
\tag{5.1}
$$

where ϕ is porosity; $\ln k$ and $\ln k_z$ are horizontal and vertical permeability, respectively; p is pressure; S_w and S_g are water and gas saturation; d are observed data; N is the number of active cells, $N = 1,761$; N_d is the number of observation data, maximum $N_d = 24$ in the study.

In regions in which oil is under-saturated, gas saturation (S_g) is not a valid state solution and solution gas-oil ratio (R_s) should be used as the state vector. However, the difference between bubble point and reservoir pressure of the PUNQ-S3 model is small and the use of S_g as a state solution did not result in a significant problem. In this application, the initial pressure and phase saturation distributions are not treated as random variables. All the ensemble models use the same initial pressures and saturations.

The kinds and amount of observed data available at different times vary in Table C-1 at Appendix C. During the history matching period of the first 8 years (0–2,936 days), the Gaussian standard noises used for observation data perturbation are presented in Table 5.3:

Table 5.3: Standard deviations of the noises added to data computed from the true reservoir simulation model. (b.t. stands for breakthrough)

Data	STD of noise
Shut-in pressure	1 bar
Flowing pressure	3 bar
Gas-oil ratio before gas b.t.	10%
Gas-oil ratio after gas b.t.	25%
Water cut	1%
Oil production rate	10^{-4} sm^3/day

5.3 Production History Match Results

The performance forecasts of six producers by corrected models during the history matching and prediction phases have been improved compared with the initial models. In summary,

(1) Well oil production rate: After EnKF correction, most wells can produce at the specified well production target rate. Even in the prediction phase, most of the corrected models can produce at the desired rate while some models change to 120 bar bottom-hole pressure constraint (*e.g.*, PRO-4). The results indicate that the assimilation of oil production rate with small measurement error effectively addresses changes in well constraints.

(2) Well bottom-hole pressure: The matched results in the history match phase are “close” to the truth, and the spread in the models in the prediction phase (which reflects uncertainty) is reduced significantly compared with the initial models. The standard deviations of flowing pressure for initial models range from 20 to 50 bar. In some initial models, PRO-11 and PRO-15 change to bottom-hole constraint within history match phase because of the poor rock property estimate. After the EnKF correction, the standard deviations of corrected models range from 1 to 2 bar, which is comparable to the measurement error 3 bar for flow pressure. The only exception is PRO-12. The updated models underestimate the pressure draw-downs in the forecast phase.

(3) Well gas-oil ratio: The comparison shows substantial improvement in the gas-oil ratio match. Some of the initial models produce much more free gas than the true model does in most of wells, whereas the corrected models have their gas-oil ratio distributed close to the truth. In history match phase, the largest error of gas-oil ratio is from PRO-4, that is, 16 percent of 120 sm^3/sm^3 (“truth”), which is less than the assumed measurement error (25 percent) . Note during the first year of extended well testing, some of the initial models have gas-oil ratios as high as 400 sm^3/sm^3, even without any gas-oil ratio data assimilated during that time. However, the bottom-hole pressure

assimilation brings the high values down to 150 $\mathrm{sm}^3/\mathrm{sm}^3$. For well PRO-11, the free gas production is corrected during the history matching phase, however, shortly after the history matching phase is ended, the gas-oil ratio goes up again in the prediction phase.

(4) Well water cut: In history match phase, there are no water breakthough occur in any wells. Therefore the prior error is less than assumed measurement error 1 percent. There is no water production for PRO-1 during the 16.5 years, whereas some of the initial and corrected models predict water breakthroughs within the prediction period. However, the water breakthrough time predicted by the corrected models is delayed compared to the initial models. The initial models do not give correct timing for the water breakthrough but after data assimilation, the timing is captured better for PRO-11. The updated model does not provide a good prediction for water cut at PRO-12.

For this case, we are interested in comparing the variability of the cumulative oil production after 16.5 years from the ensemble of final corrected models using EnKF to the results from the initial model. The variance decreases after the EnKF corrections. The mean of corrected models is (practically) equal to the true production value (3.872×10^6 sm^3) and with standard deviation of 71,365 sm^3; this gives a coefficient of variation ($= \sigma/\mu$) of 0.018 compared to 0.3 before assimilation. This is a vast improvement — the P_{90} to P_{10} range has gone down by a factor of 96 percent which corresponds to 3.872×10^6 sm^3. The ensemble mean and standard deviation of cumulative oil recovery from the corrected model using EnKF are compared with other history matching methods in Figure 5.2. The accuracy of the EnKF is similar to a particular implementation of Markov chain Monte Carlo (MCMC) and better than the pilot point (PP) method, important sampling (IS1 and IS2) and genetic algorithm (GA1 and GA2). summarized by Barker, Cuypers, and Holden 2001).

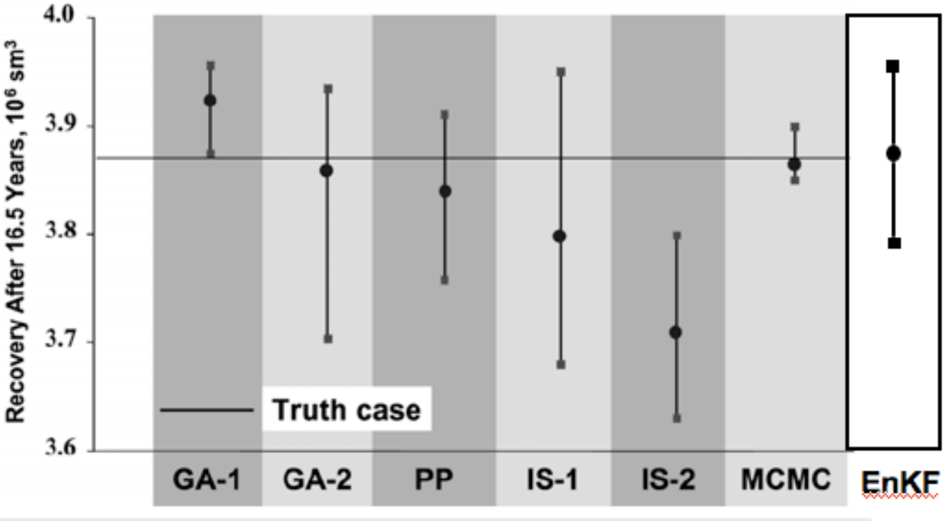


Figure 5.2: Comparison with uncertainty estimate from PUNQS3 study. Mean and STD of the cumulative oil recovery prediction from corrected models by EnKF, and results summarized by Barker et. al (2001). The horizontal line crossing the entire plot denotes the computed total oil production value from the true reservoir simulation model.

5.4 Porosity Estimates

Figures 5.3 to 5.7 plot the porosity estimates of the five layers at the 0, 4th (274 days), 11th (2008 days), and 20th (2936 days) assimilation steps. From these figures, we have some general observations:

Layer 1. The gas cap is situated at layer 1. No well is perforated in this layer because free gas production might influence the ultimate recovery. In layer 1, there are two channels in the truth. One is in the middle and another is on the upper right corner. Actually, there are no obvious channel sands in the initial realization. After the 274 days assimilation, the estimates resolved the middle channel sand. Because no well is perforated in layer 1, and only PRO-5 at layer 3 is perforated at the closest region (Figure 5.3(c)), it is difficult to change the porosity at the upper right corner by EnKF. Therefore the uncertainty of porosity is not reduced much at those far-removed parts. The middle channel sand "disappears" after the 2936 days estimate, as the property estimate diverges from the "truth" (Figure 5.3(e)).

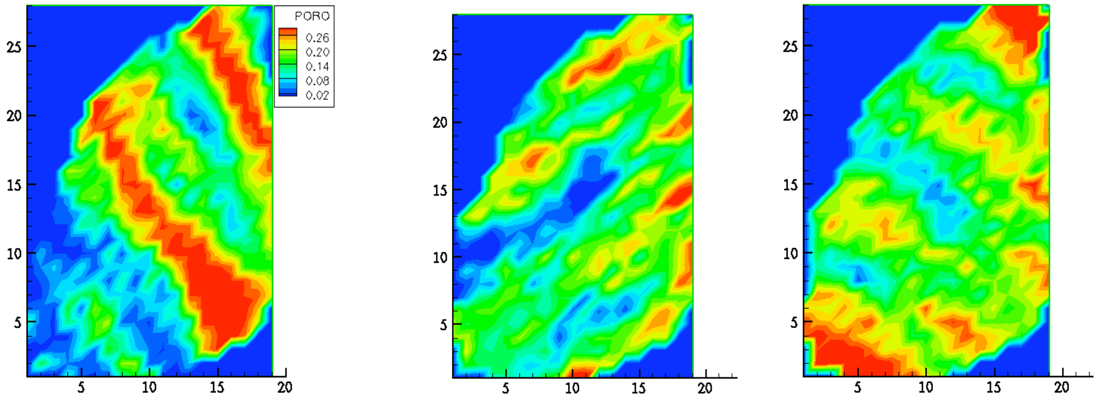
Layer 2. Because the facies of this layer is lagoonal shale, no well produces the layer. The initial realization is different from the true model. After the 274 days, the high porosity area emerges at the low bottom (Figure 5.4(c)). The porosity field has few changes from 274 to 2008 days.

Layer 3. There are three channel sands in this layer. PRO-5 and PRO-11, in the north and south, respectively, are perforated this layer. The EnKF captures the middle channel sand after assimilation at 274 days (Figure 5.5(c)). The lower bottom channel sand is difficult to recover because it is close to the fault in the south, the flow is slight compared with the strong aquifer in the west.

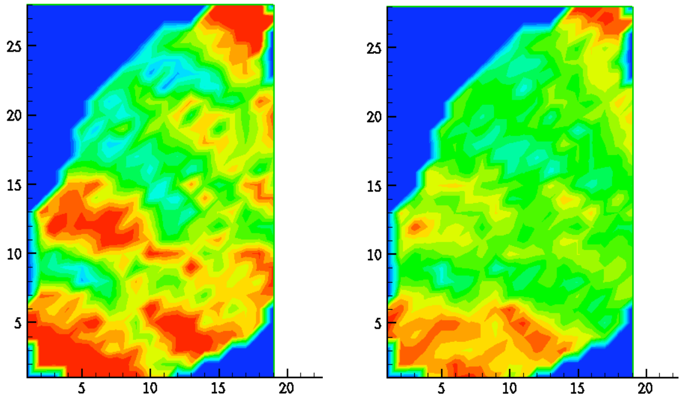
Layer 4. All six producers are drilled and perforated the layer. The porosity of sand channel is greater than 20 percent. The initial realization conditioned to the well data has more than 50 percent of channel sands. However, the geological description shows only 30 percent of the volume of the whole layer is channel sand. Therefore, the initial realization is not reasonable and cannot be used to forecast the future performance. Compared with layer 2, the porosity of layer 4 captures more reservoir characteristics. The uncertainty reduction of porosity in the layer is significant whereas the filter divergence is the most obvious among other layers after the last assimilation step.

Layer 5 There are two channel sands that are very close to each other in the truth and three producers produce this layer. With data assimilation, the features of the true porosity field are able to be recovered gradually, the alternating low and high porosity streaks in this layer; but the changes become smaller after 274 days assimilation (Figure 5.7(c)). The final result shows that the channel sands are reasonably near the "truth" case position after 2936 days (Figure 5.7(e)).

The production data are less sensitive to the rock properties in layer 3 than in layer 4 and 5 because more production data are available in those layers. The correction of model



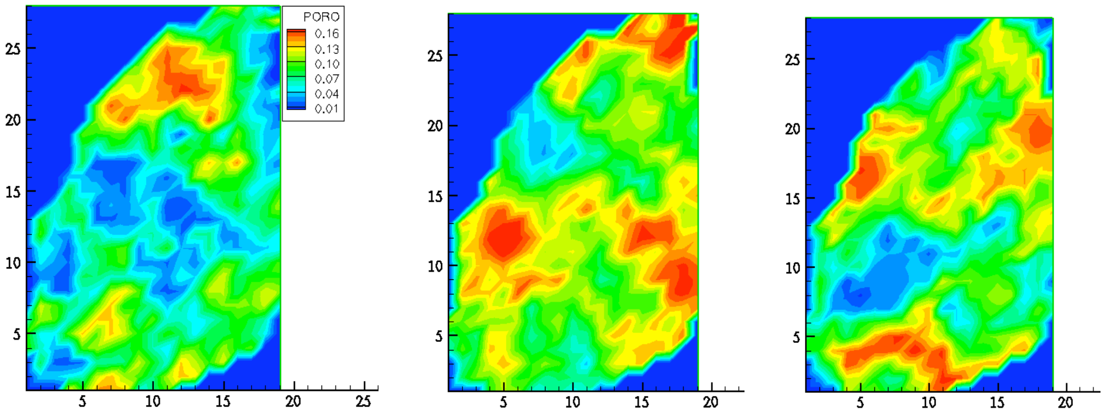
(a) True model (b) Initial (c) 274 days



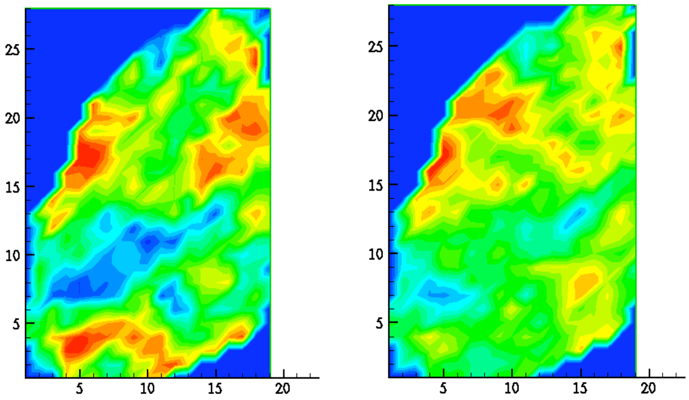
(d) 2008 days (e) 2936 days

Figure 5.3: Porosity of a realization for layer 1 at different assimilation steps

parameters in layers 1, 3 and 5 is greater than in layers 2 and 4 because the correlation lengths (variogram range) in layers 1, 3 and 5 are nearly twice as long as those in layers 2 and 4. Longer correlation length means that if a location is sensitive to a model parameter in one grid block, then it is also sensitive to model parameters in grid blocks in a larger area around that grid block. In this study, we found that the completely wrong information on the statistical anisotropy (*i.e.*, azimuth) may generate a nonphysical state vector, especially porosity in the initial assimilation steps and has a long-lasting effect on the spatial pattern of the inferred rock properties and may be corrected with more observations in space. Figure 5.8 plots the σ_t and σ_e of the porosity estimates. The σ_t of porosity decreases initially, then increases at 2938 days. The spread of ensemble estimates decreases monotonically.

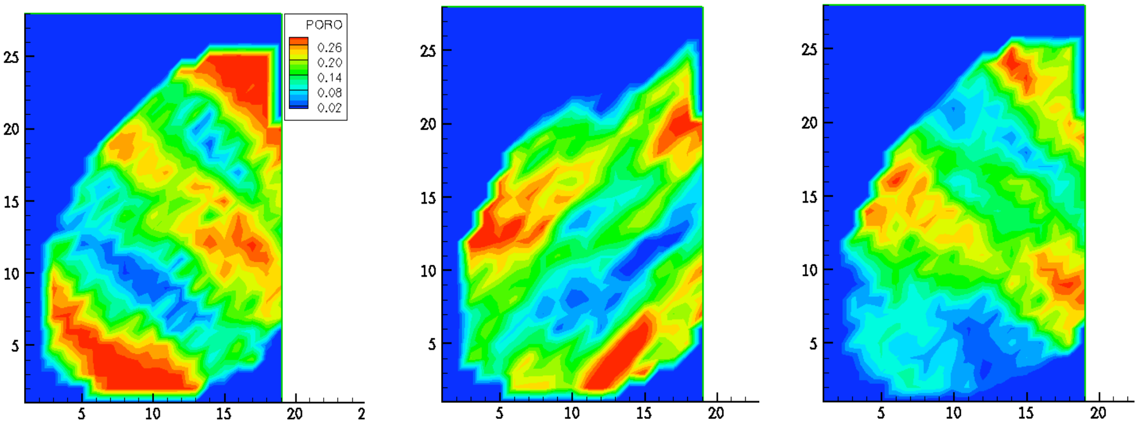


(a) True model (b) Initial (c) 274 days

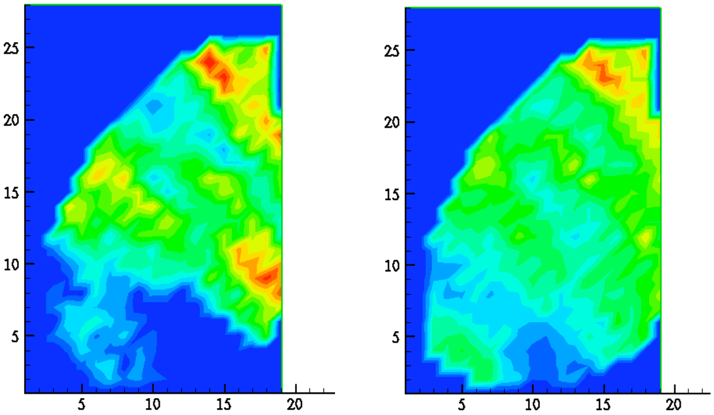


(d) 2008 days (e) 2936 days

Figure 5.4: Porosity of a realization for layer 2 at different phases

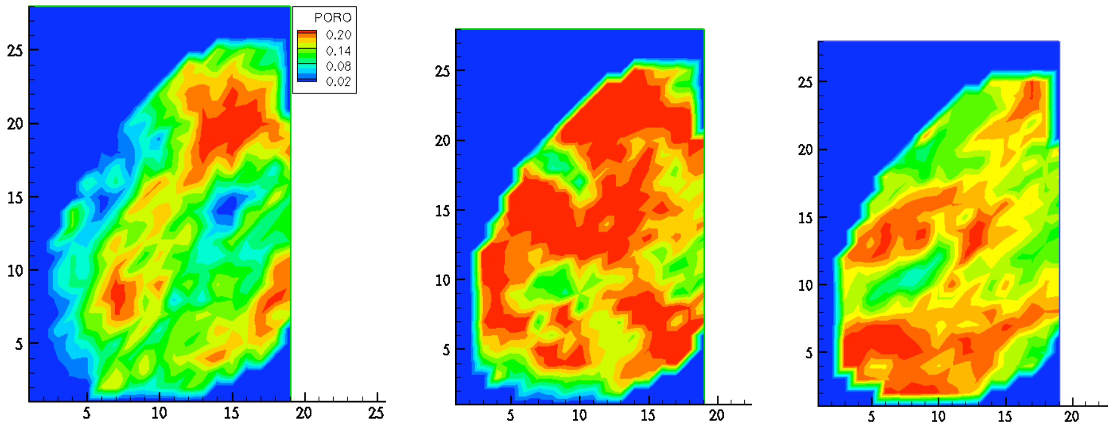


(a) True model (b) Initial (c) 274 days



(d) 2008 days (e) 2936 days

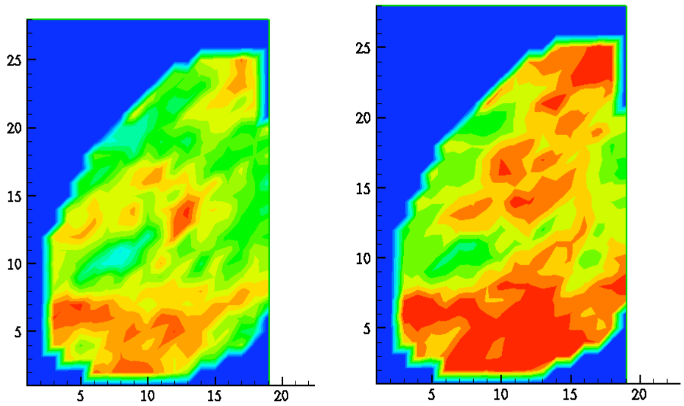
Figure 5.5: Porosity of a realization for layer 3 at different phases



(a) True model

(b) Initial

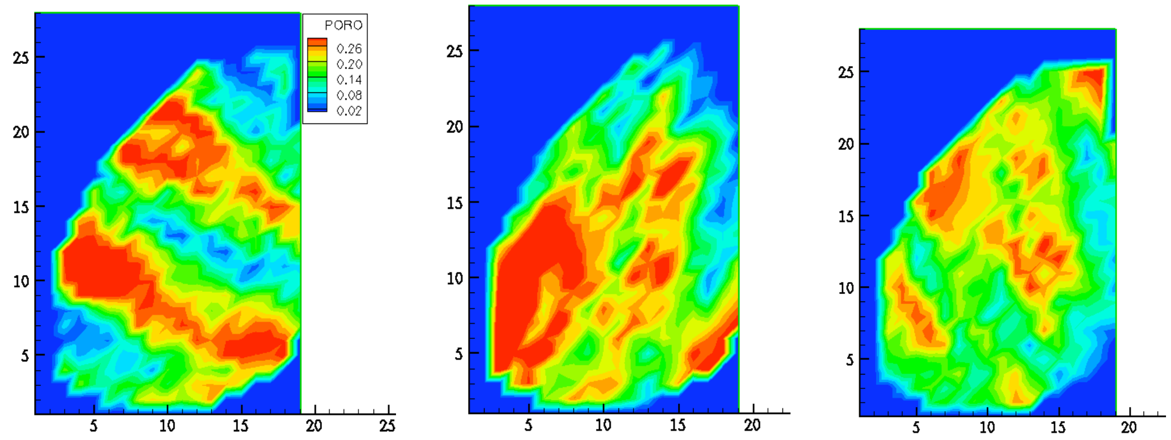
(c) 274 days



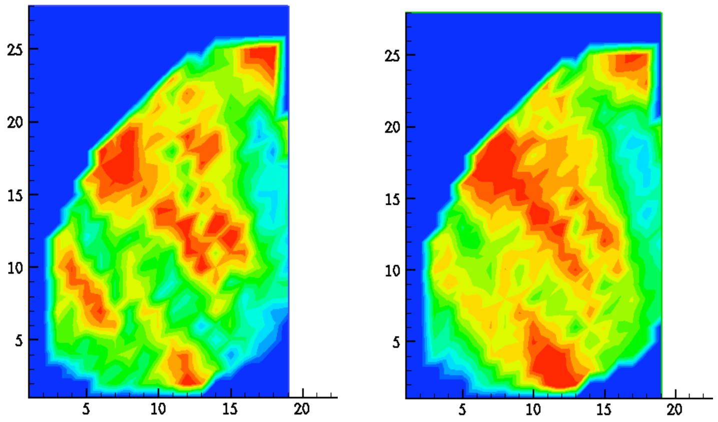
(d) 2008 days

(e) 2936 days

Figure 5.6: Porosity of a realization for layer 4 at different phases



(a) True model (b) Initial (c) 274 days



(d) 2008 days (e) 2936 days

Figure 5.7: Porosity of a realization for layer 5 at different phases

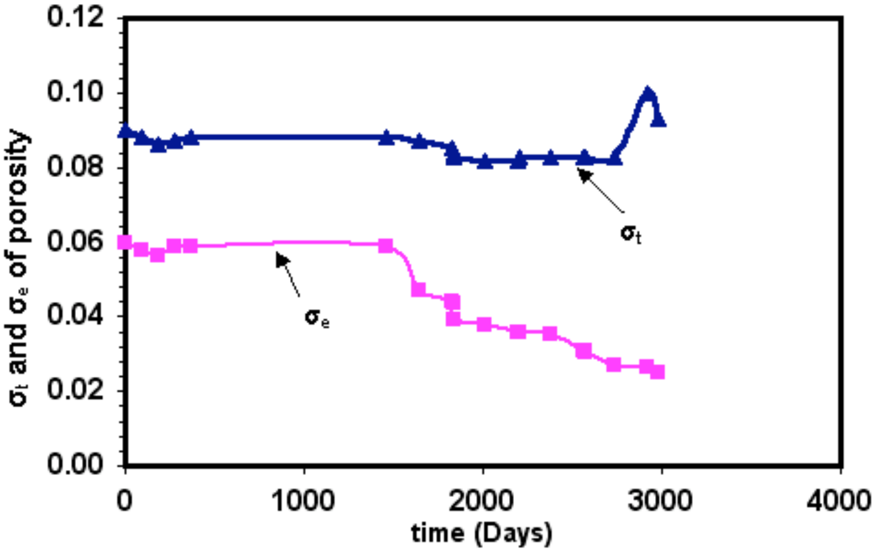


Figure 5.8: σ_t and σ_e for porosity estimates.

5.5 Discussion

The prototype work flow demonstrated the successful application of EnKF to a 3-D, 3-phase model with complex geology — a more realistic model inversion — problem on a single processor Linux machine. There are 20 analysis steps, assimilating 265 observed production data. It is efficient and robust, taking two days to finish the history match and forecast phases. The prototype work flow will be applied to a geological models screening method to choose essentially diverse yet manageably small prior ensembles to improve the efficiency of EnKF in the future work.

The estimate of the model parameters improves after assimilating with dynamic observations. In this case, the estimate captures the main characteristics of the reservoir after the first 4 or 5 assimilation steps. More generally, this will depend on the flow characteristics, the well locations, and the measurement types (*e.g.*, rate data used in the case).

The prior geologic knowledge of the reservoir, mainly the statistical data, plays an important role in data assimilation. The spatial continuity has a large influence on the structure of error covariance matrix in Kalman gain calculation. The large correlation between the greatly separated grid points produce spurious covariances (which will be addressed with localization in Section 6.3).

Chapter 6

Exploration of Ensemble Kalman Filter Divergence

As discussed in Chapter 4, the grid-based distributed EnKF has a number of desirable properties:

(i) The work flow renders the EnKF computationally feasible by using parallel and Grid computing. It is an efficient tool useful for reservoir engineers in assimilating large numbers of observations and large ensembles.

(ii) The work flow is flexible, and can be adapted easily to an in-house simulator Cactus BlackOil, which is designed to scale thousands of processors.

(iii) The work flow can be extended to the "closed-loop" reservoir management process, perhaps using ensemble optimization method (Chen, Oliver, and Zhang 2008).

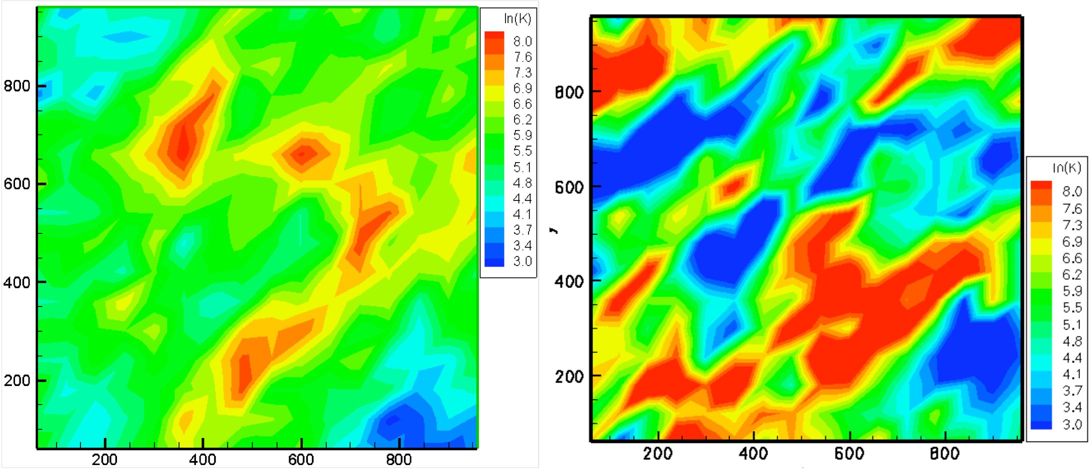
Despite the appeal of the distributed EnKF approach as a history matching technique, and its efficiency in the grid environment, there is much to be learned before it will gain wide use in the industry. In this chapter, the impact of "filter divergence" on the EnKF is investigated. "Filter divergence" causes an increasing error in reservoir property estimates as the EnKF is used to sequentially assimilate observed responses. This indicates a potential problem with the EnKF becoming unstable after a number of assimilation steps. In Section 6.1, we investigate the relationship between assimilation frequency and filter divergence. Then, in Section 6.2, the continuously reduced error covariance (or "ensemble collapse") that causes filter divergence is discussed. Next, inflation of the error covariance is used to stabilize the EnKF. Several topics, such as how much inflation is needed, its dependence on model and ensemble sizes are discussed. Finally, in Section 6.3, the performance of localized EnKF is compared with the traditional EnKF correction.

6.1 Analysis of Errors as Function of Assimilation Frequency

In situ permanent sensors provide continuous observations from near the completion intervals of wells. Frequent assimilation of these large datasets poses a significant computational burden. How should engineers choose the most appropriate assimilation frequency? Here we examine the effect of refining some of the measurements in time as a simple scheme in dealing with high assimilation frequency, or skipping some of the measurements in time as a consequence of low assimilation frequency which some flow information may be missed.

In principle, the updates should resolve the time scales of the model dynamics; otherwise, the estimate may diverge from the true state between each model update. This is analogous to how the simulator "chops" the time step when the iterations can not converge to a solution, especially when dealing with strongly nonlinear systems. Unfortunately, the required assimilation frequency is case-dependent, as shown in the following. It appears that reservoir engineers have to depend on experience and trial-and-error to determine the most appropriate assimilation timing. Nevertheless, it is in general true that assimilation is needed whenever significant flow behavior changes occur, such as water breakthrough, sharp changes of water cut, adding new wells in the reservoir system, well shut-in or well converting from producer to injector. Fortunately, the significant changes that drive the assimilation interval are the changes in the observations (which are shared by all ensemble members) rather than individual member behavior. Thus, an experienced engineer can formulate an assimilation schedule based on these observations; automation and detailed guidelines are desirable but beyond the scope of this work.

The 2-D water flood example in Chapter 4 is chosen for the assimilation frequency study. First, the assimilation step is chosen as every 0.1 pv (about 10 days), which is termed base case. After 0.7 pv measurements are assimilated, the $\ln k$ contour of the ensemble mean (Figure 6.1(b)) demonstrates geological characteristics differing from the reference



(a) $\ln k$ reference contour

(b) Divergent $\ln k$ for ensemble mean after large assimilation step (base case)

Figure 6.1: Contour of $\ln k$ comparisons between truth and base case

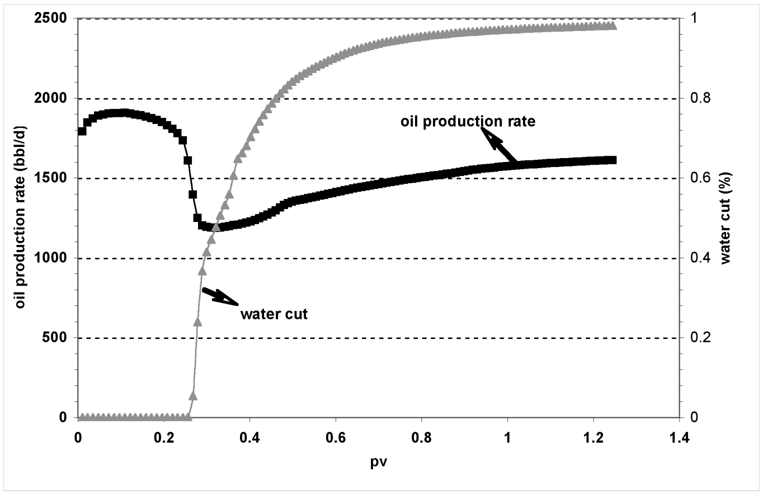


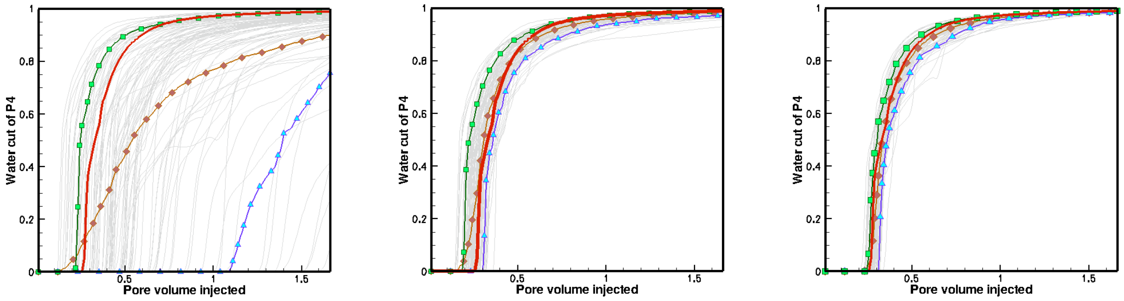
Figure 6.2: Production performance of P4 from “truth” model

(Figure 6.1(a)) and the values of $\ln k$ are out of range. The EnKF functions as a least squares linear regression and when the measured data between each model update change significantly, traditional EnKF update can cause problems because of its linear assumption. The true production performance of P4, which is located at the right upper corner in a high permeability zone, is shown at Figure 6.2. Water breakthrough occurs in P4 after 0.25 pv injection and the water cut reaches 55 percent after 0.32 pv injection. The water cut changes almost 50 percent within time period of 0.08 pv, which is less than the assimilation step 0.1 pv. After one or two assimilation steps, the overcorrected estimate of reservoir properties begins to diverge from the true state. The increasing errors from the forecast steps accelerate the overcorrections. Finally, the oil production rate forecast of P4 (black line) deviates from the reference (red line) as shown in Figure 4.4(c). The EnKF process fails.

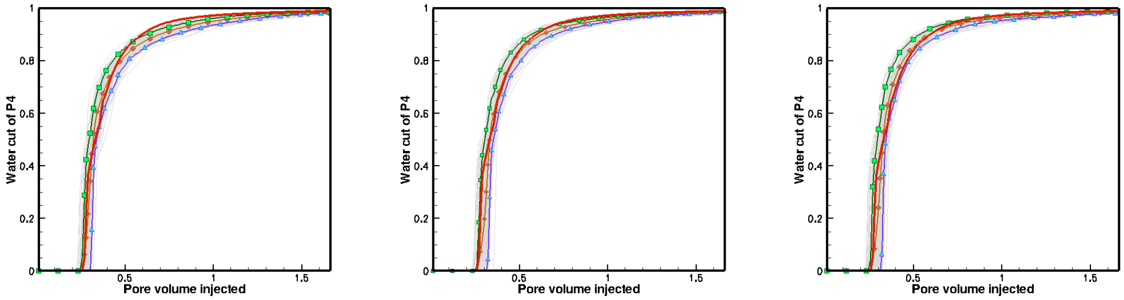
Now use a higher assimilation frequency, each 0.05 pv, which is termed Case I: the predicted performance of the updated models by EnKF improves significantly. As shown in Figure 6.3, the water cut predictions of P4 (the updated models) at different measurement times (black line) closely matches the corresponding observed value (red line) as more observed responses are assimilated. The P_{90} to P_{10} range has gone down by a factor of 80 percent which corresponds to 90 % difference from Figures 6.3(a) to 6.3(b). We obtain more accurate and less uncertain results as more features of the reservoir have been attained from the well data. The match of P_{90}, P_{10} and mean of ensemble forecast match the truth from the 0.35 pv assimilation step. The permeability at the region between injector I1 and producer P4 is important for matching production performance of P4. The similarity between the ensemble mean and "truth" is shown in Figure 6.4. The ensemble forecasts indicate that the permeability features were captured by the early assimilation steps (less than 0.12 pv). Therefore, the 0.05 pv better resolves reservoir dynamics than 0.1 pv interval (in this case).

Alternative assimilation frequencies were considered to further examine this behavior. Case II assumes using every 0.05 pv as an assimilation step similar to Case I, but the assimilation frequency is increased to every 0.01 pv (1 day) once water breakthrough occurs in any producers until production from the well stabilizes. Then assimilations repeat every 0.05 pv until water breakthrough on another well occurs. There are 38 assimilations in total, which is denoted as 0.01 and 0.05 pv combination. The Case III assumes the assimilation step is taken every 0.01 pv (1 day). There are 121 assimilation steps in total.

The σ_t of $\ln k$ for the three cases are drawn as a function of injected pore volume in Figure 6.5. The final results for all the cases are shown in Table 6.1. The σ_t of Case III, with every

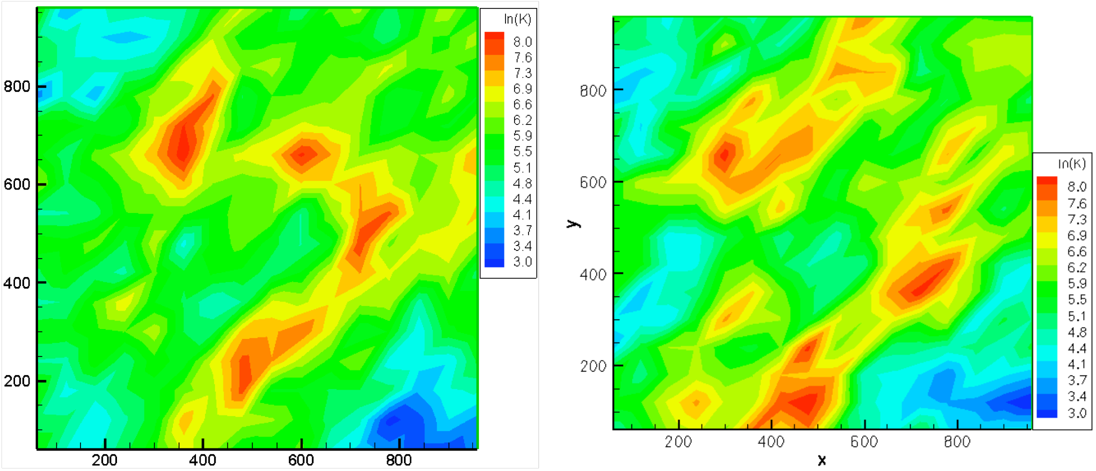


(a) Initial model forecast for P4 (b) 0.12 pv assimilation (c) 0.35 pv assimilation



(d) 0.64 pv assimilation (e) 0.94 pv assimilation (f) 1.24 pv assimilation

Figure 6.3: The P4 water cut forecast by updated model (Case I). The red line denotes the truth. P_{90} is denoted by green square line. Ensemble mean is brown in diamond line and P_{10} is blue delta line.



(a) $\ln k$ reference contour (b) $\ln k$ contour for ensemble mean, every 0.05 pv assimilation (Case I)

Figure 6.4: Contour of $\ln k$ comparisons between reference and ensemble mean from Case II

Table 6.1: The final results for all the cases

Case	Assimilation interval	Assimilation times	σ_t
Base	0.1 pv	7	1.232
I	0.05 pv	21	0.830
II	0.05 and 0.01 pv combination	39	0.824
III	0.01 pv	121	0.963

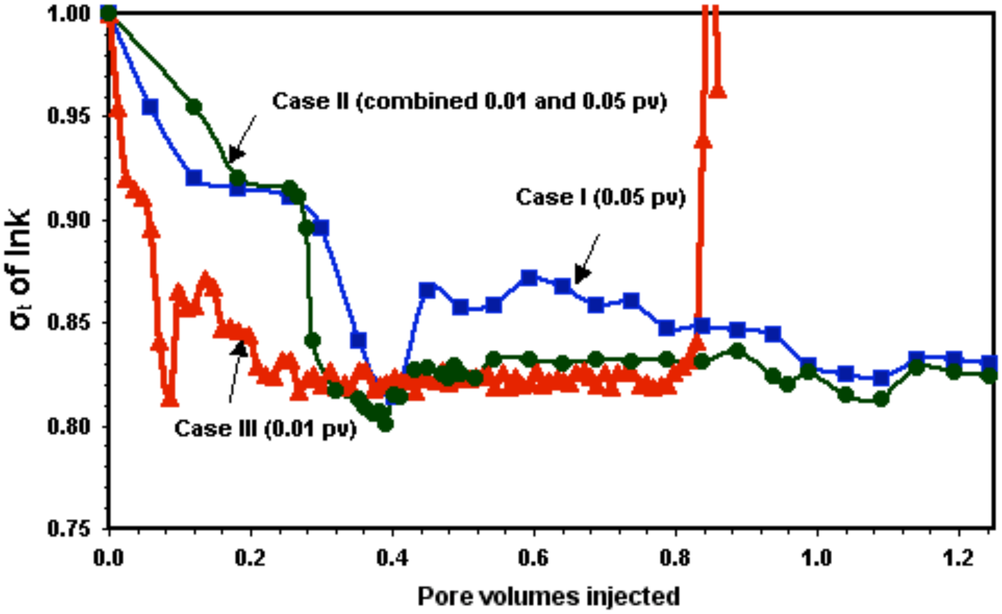


Figure 6.5: Influence of assimilation frequency

0.01 pv assimilation frequency, is divergent at 0.8 pv. This may be caused by continuously reduced error covariance and stochastic errors from too many assimilation steps. The reason will be discussed in Section 6.2. Case II with combined assimilation frequency has some improvement from the Case I but the improvement is not significant. This may be because such closely spaced observations are redundant. It appears that early observations give the largest contribution to the reduction in the σ_t. In this case the transient flow at the early time observations provide more information than the observations from a steady state flow. Under actual condition, production data might be observed at different frequency and permanent sensors could acquire data at very high frequencies. Too frequent updates leads to many restarts and slows down the system. In addition, the updates may introduce numerical or stochastic errors into the system, which may be larger than and therefore mask the benefits of the (possibly redundant) measurements. Because production data may be correlated, and no useful information can be absorbed when the selected points are too closely spaced in time.

6.2 Analysis of Errors as Function of Inflation Factor

The error covariance statistics control EnKF behavior, because the EnKF uses covariance of ensemble to approximate the pdf of model parameters, solutions and observed response (instead of evolving it in time, as it was for the KF (Jazwinski 1970) and EKF (Evensen 2003).

6.2.1 Revisiting the Update Scheme of EnKF

To better understand the reason for filter divergence, it is useful to first revisit the analysis scheme of the EnKF. The modified analysis step of the EnKF consists of update applied to each of ensemble members:

$$\boldsymbol{y}^{u}_{j,k} = \boldsymbol{y}^{p}_{j,k} + \boldsymbol{K}_{Y,e,k}(\boldsymbol{d}_{obs,k} - \boldsymbol{H}_k \boldsymbol{y}^{p}_{j,k}) \quad j \in \{1 \ldots, N_e\} \tag{6.1}$$

Recall that $\boldsymbol{H}$ is a projection of the model state onto the measurement space, that is, $\boldsymbol{H}_k \boldsymbol{y}^{p}_{j,k}$ simply extracts the predictions $\boldsymbol{d}^{p}_{j}$ from the state vector, so the term in parentheses, $\boldsymbol{d}^{p}_{j,k} \equiv (\boldsymbol{d}_{obs} - \boldsymbol{H}_k \boldsymbol{y}^{p}_{j,k})$, is the vector of mismatches of production data of model j to the observations. EnKF updates each ensemble members with the same measurement. That is, the model parameter $\boldsymbol{m}_{j,k}$ for the ensemble $j \in \{1 \ldots N_e\}$ is updated proportional to its mismatch $\Delta \boldsymbol{m}_{j,k} = \boldsymbol{K}_{Y,e,k}(\boldsymbol{d}_{obs,k} - \boldsymbol{d}^{p}_{j,k})$ using the same value from the Kalman gain matrix. This is the linearization approximation inherent in the EnKF: the gain estimates the *average, linear* proportionality between mismatches and parameter values. The mean and standard deviation of ensembles are corrected by the same Kalman gain in the following equations,

$$\bar{\boldsymbol{y}}^{u}_{k} = \bar{\boldsymbol{y}}^{p}_{k} + \boldsymbol{K}_{Y,e,k}(\boldsymbol{d}_{obs,k} - \boldsymbol{H}_{k}\bar{\boldsymbol{y}}^{p}_{k}) \tag{6.2}$$

$$(\boldsymbol{\sigma}^{2}_{k})^{u} = (\boldsymbol{\sigma}^{2}_{k})^{p} + \boldsymbol{K}_{Y,e,k}\boldsymbol{H}_{k}(\boldsymbol{\sigma}^{2}_{k})^{p} \tag{6.3}$$

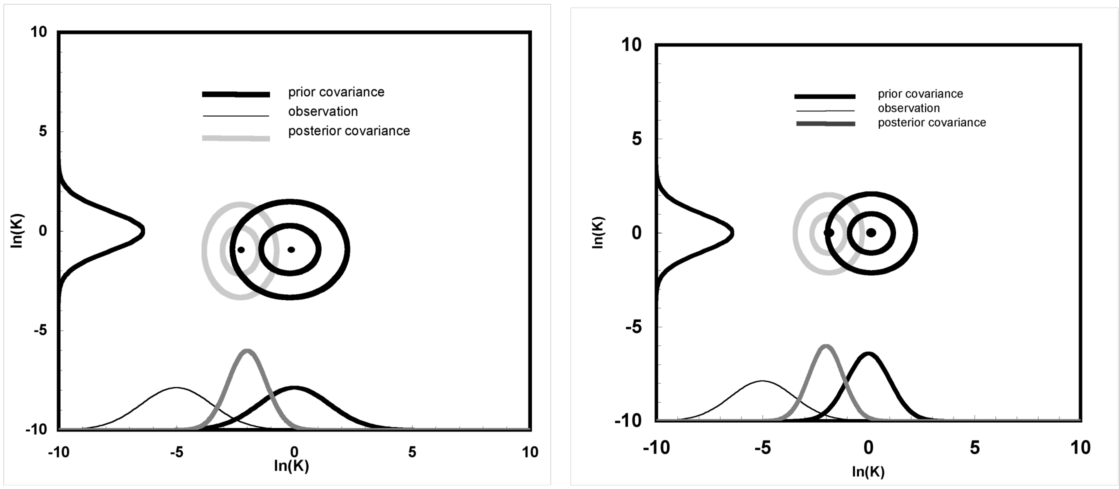
The EnKF uses an ensemble of model states to represent the posterior pdf. The formulations show that the EnKF is searching for the mean in the ensemble space. With wisely chosen initial realizations, the initial ensemble should represent most of variance of the true parameter solution. From Eq. 6.2, it is clear that the EnKF update neglects any non-Gaunssian contribution in the predicted pdf, when the update increments are computed, since these only take into account the covariances. The error covariance of the analyzed ensemble can be reduced to:

$$\begin{aligned}
\boldsymbol{C}^{u}_{Y,e,k} &= E[(\boldsymbol{y}^{u}_{j,k} - \bar{\boldsymbol{y}}^{u}_{k})^{2}] + \boldsymbol{O}(N_{e}^{-1/2}) \\
&= E[(\boldsymbol{y}_{j,k} - \bar{\boldsymbol{y}}_{k}) + \boldsymbol{K}_{Y,e,k}(\boldsymbol{d}_{obs,k} - \bar{\boldsymbol{d}}_{obs,k} - \boldsymbol{H}y_{j,k} + \boldsymbol{H}\bar{\boldsymbol{y}}_{k})^{2}] + \boldsymbol{O}(N_{e}^{-1/2}) \\
&= E[(\boldsymbol{y}_{k} - \bar{\boldsymbol{y}}_{j,k})^{2}] + E[\boldsymbol{K}_{Y,e,k}(\boldsymbol{d}_{obs,k} - \bar{\boldsymbol{d}}_{obs,k})^{2}] + E[\boldsymbol{K}_{Y,e,k}(\boldsymbol{H}y_{j,k} - \boldsymbol{H}\bar{\boldsymbol{y}}_{k})^{2}] \\
&\quad + \boldsymbol{O}(N_{e}^{-1/2}) \\
&= (\boldsymbol{I} - \boldsymbol{K}_{Y,e,k}\boldsymbol{H})\boldsymbol{C}^{p}_{Y,e,k}(\boldsymbol{I} - \boldsymbol{K}_{Y,e,k}\boldsymbol{H})^{\mathrm{T}} + \boldsymbol{K}_{Y,e,k}\boldsymbol{\epsilon}_{k} + \boldsymbol{O}(N_{e}^{-1/2})
\end{aligned} \tag{6.4}$$

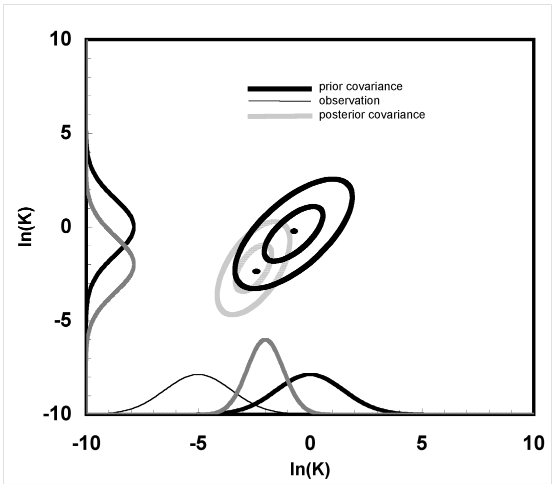
Compared to the analyzed error covariance of the Kalman filter, the error covariance of the forecasted ensemble mean consists of the ensemble covariance, the measurement error $\boldsymbol{\epsilon}_{k}$ and sampling error of Monte Carlo method which is $\boldsymbol{O}(N_{e}^{-1/2})$. The error covariance of the updated ensemble mean will tend to reduce too much because all ensemble members are updated with the same measurements. Therefore, the EnKF's model forecast covariance is easily underestimated by its implementation. Eq. 6.4 also implies that the inappropriate measurement errors $\boldsymbol{\epsilon}_{k}$ and sampling errors can also introduce stochastic errors in error covariance. For example, a spurious relation caused by sampling errors for filter divergence has been reported by Gu (2006). The porosity and permeability estimates move away from the "truth" in a 1-D example when water first reached the producer. Filter divergence problems also occur in her 2-D and 3-D examples. One reason the filter divergence occurs earlier in her EnKF implementation is the sampling errors caused by a rather small ensemble size of 40. Large ensemble sizes and improved sampling methods improve convergence of the EnKF. In the following section, we focus on methods to stabilize the error covariances of the EnKF.

6.2.2 The Impact of Error Covariance on Filter Divergence

Much of the problem is a consequence of using the ensemble to estimate the ensemble covariance matrix in the EnKF. The EnKF may diverge due to errors in covariance estimates. These errors are examined using a simple two-variable cases, in which we make an observation (with error) of a $\ln k$ at x, and examine the effects on the posterior at x and y as the dependence on the covariance estimate (Fig. 6.6). The x-axis is the prior distribution of $\ln k$ (black thick line) at one gridblock close to a well where the observation responses are obtained. The y-axis is the prior pdf (black thick line) of $\ln k$ at a far-removed gridblock from the well. Because the two locations are widely separated, they are *a priori* uncorrelated, and this is reflected in the widely dispersed prior covariance (Fig. 6.6(a)). After the EnKF update, the prior pdf of $\ln k$ at x is tuned toward the observation only to an extent consistent with the Kalman gain to get the posterior Gaussian distribution (gray thick line). If the pdf of $\ln k$ at x is estimated correctly, the variance of $\ln k$ at x decreases compared with that of the prior distribution. Because the covariance of $\ln k$ at x and y is zero, the posterior distribution of $\ln k$ at y is unchanged. But if the prior variance of the $\ln k$ distribution is underestimated, the measurement is ignored in the Kalman gain and the posterior unavoidably looks like the prior (Figure 6.6(b)). During subsequent assimilation steps, the variance-deficient ensemble thus further underestimates the model errors, disregarding even more the influence of the



(a) True variance of $\ln k$ (b) Underestimated variance of $\ln k$



(c) Underestimated variance of $\ln k$

Figure 6.6: Filter divergence analysis (Hamill and Whitabker 2001)

new observations. This problem can progressively worsen, resulting in a useless ensemble of forecasts.

Filter divergence can occur as a result of a wrongly estimated covariance relationship (Fig. 6.6(c)). Similar to Figure 6.6(b), it is assumed that the "flow information" (*e.g.*, pressure transients after rate changes) does not propagate rapidly across large distances in the reservoir, and that observations tend to be most related to model parameters in their vicinity. However, if the magnitude of covariance between a well and this location is overestimated because of the poorly known geological information, the pdf of the far-removed location will be artificially correlated, and therefore "corrected," even if there is no flow information there. This can make the a posterior pdf at those areas biased and/or reduce variance; that is, the posterior distribution has insufficient probability at the location near to true state. This may lead to filter divergence.

Many authors have suggested approaches to lessen or prevent the trend toward this kind of filter divergence. Houtekamer and Mitchell (1998) proposed a "double EnKF". They used parallel ensemble data assimilation cycles in which the error covariance estimated by one ensemble is used to calculate the Kalman gain for the other. This procedure compensates for the biases associated with nonlinearity in the error covariance and helps prevent the assimilation process toward smaller and smaller prior error covariances. Anderson (2007) proposed the "Hierarchical ensemble Kalman filter", in which an ensemble of ensemble filters is used to lessen the effects of spurious correlations among an observation and model parameters. Anderson shows that even small numbers of groups appear to lead to good estimates of sampling error in ensembles. But computation costs are an important considerations in the application of these methods; hierarchical models require more simulation runs to ensure enough members in each of the subensembles.

One simple approach to filter divergence is to add some (white) noise to the prior distribution to "broaden" distribution and enhance the impact of observations in the Kalman gain calculations, which is termed "inflation" (Jazwinski 1970). Hence, in the following discussion, we examine the behavior of EnKF with an inflation of the covariance, and investigate the sensitivity of inflation factor to ensemble size. From the discussion in Section 6.1, to mitigate the influence of assimilation frequency and ensure the piecewise linearity of each assimilation step, Case II (with a combination of 0.01 and 0.05 pv assimilation frequencies) is used in the following studies.

Before the first observation is assimilated, the deviations from the ensemble mean are inflated by $\gamma \geq 1.0$ (typically, $\gamma \leq 1.10$):

$$\boldsymbol{y}^p_{j,k} = \gamma(\boldsymbol{y}^p_{j,k} - \bar{\boldsymbol{y}}^p_k) + \bar{\boldsymbol{y}}^p_k \quad (j = 1, 2, \ldots, N_e) \tag{6.5}$$

where k is the time step, $\boldsymbol{y}_{j,k}$ is the state vector and γ is an inflation factor.

As noted in Section 3.4, the goal of assimilation is to produce a random sample of the conditional probability distribution that is consistent with the "truth" while minimizing σ_t of the ensemble from the truth case. The effects of inflation are examined for an uninflated case (Case II, earlier) and an EnKF with an inflation of variance $\gamma = 1.01$ (Fig. 6.7). During the first several steps, σ_t of the traditional EnKF and EnKF with an inflation factor are similar. However, the traditional EnKF diverges from the "truth" at 1.24 pv injection. In contrast, the EnKF with inflation keeps converging after more assimilation steps. The standard error of ensemble σ_e of the two methods drop steadily with data assimilation. The differences between traditional EnKF and EnKF with inflation keeps increasing with time. While the empirical constant inflation factor 1.01 broadens the prior distribution artificially, the divergence problem appears to be avoided and the implied prior distribution tends to be better preserved, as appropriate. However, larger γ may result in a filtering where the observations are given too much weight and do not yield the best unbiased estimate of state variables; therefore, γ must be chosen with care.

It is likely that 1 percent is not an optimal factor for different ensemble sizes. In general, the only viable method for choosing the best γ is trial and error. A search of covariance inflation values is made until a minimum value of σ_t for the ensemble mean is found for

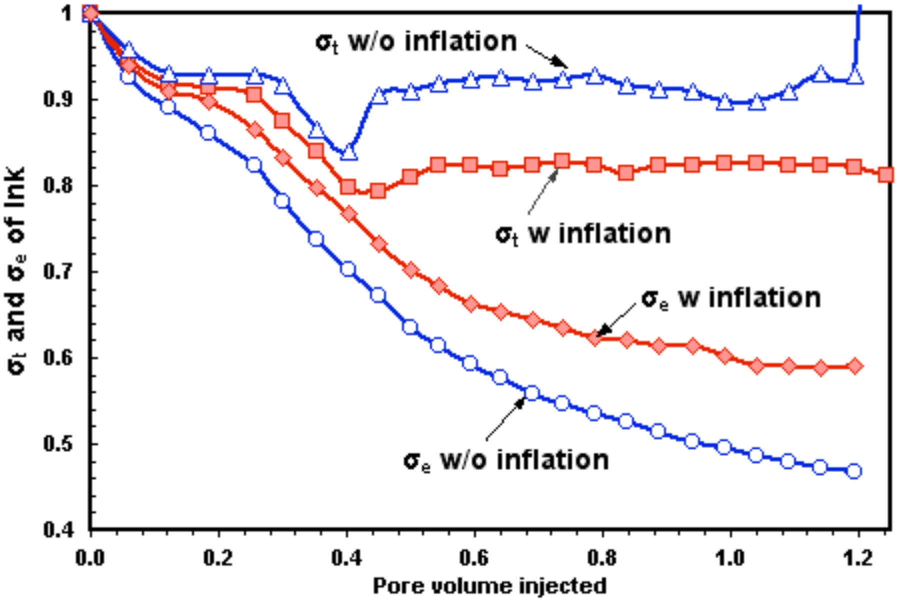


Figure 6.7: The σ_t of the ensemble without inflation (triangle), and the σ_t of the ensemble with inflation (square), the σ_e with inflation (diamond), the σ_e without inflation (circle).

synthetic case which the “truth” is known. It is computationally expensive to do trial-and-error for a range of inflation factors to find the relationship between inflation factor and ensemble size. Different inflation factors *versus* ensemble sizes are tried for Case II using high performance grid computing (Figure 6.8). The optimal inflation factor is a function of ensemble size and errors can be decreased by choosing suitable inflation factor. For the 100 ensemble member, 1 or 2 percent seems optimal, and for the 400 ensemble members , a 0.25 -1.0 percent inflation factor produces the minimum σ_t. Note that 1 percent inflation factor is nearly optimal for all ensembles. Results are only reported for this case. Tuning a filter for a real reservoir is complicated by the limited number of observations, the lack of geological knowledge of the reservoir, and the presence of systematic model errors. In order to get the best estimation by filtering techniques, trial-and-error for a suitable inflation factors is needed before EnKF process.

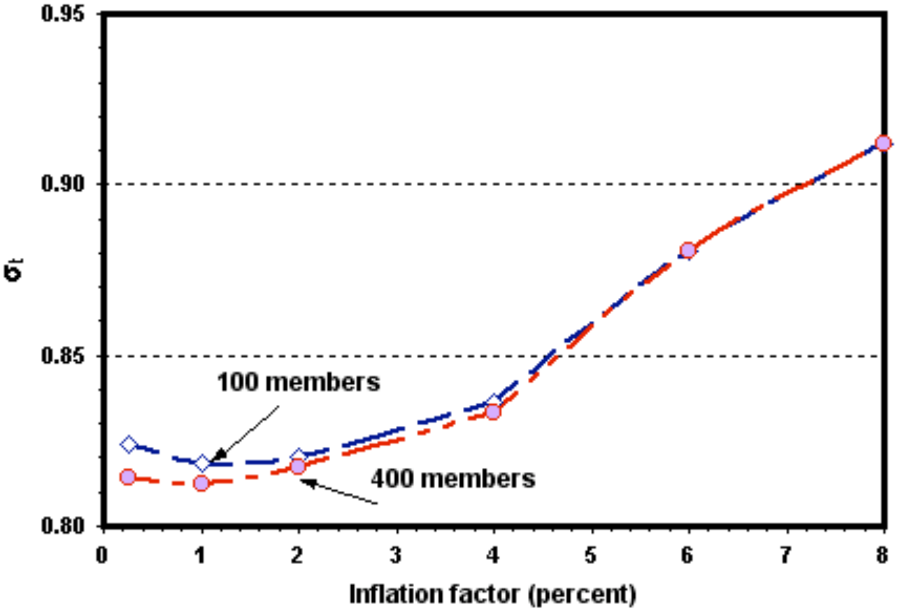
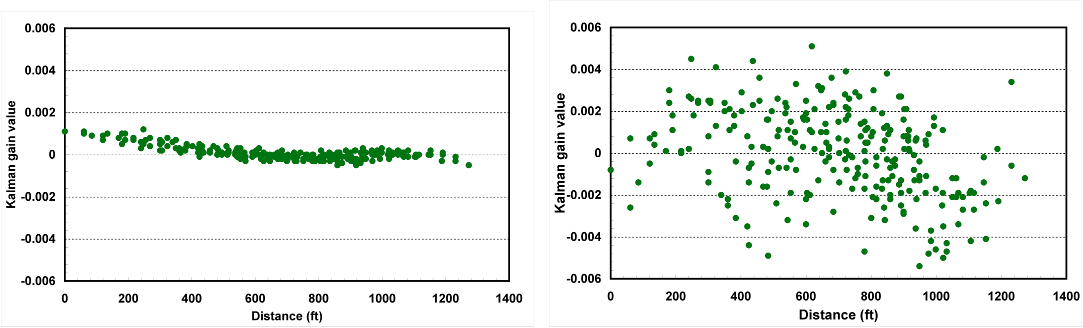


Figure 6.8: Sensitivity of inflation factor to ensemble size. The σ_t of the 100 ensemble members and 400 ensemble members are shown as square dash line and diamond dash dot line, respectively

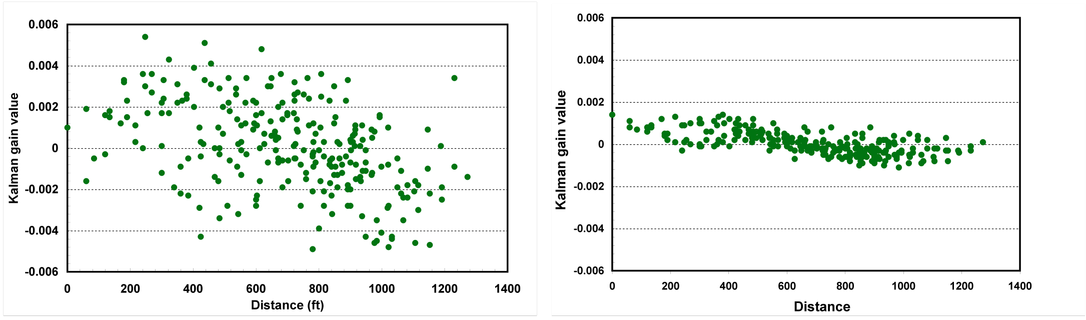
6.3 Analysis of Errors as a Function of Localization

In order to investigate the relationship between the error covariance and distance during the EnKF update, the 2^{nd} columns of Kalman gain matrices (777×9) during the first to seventh assimilation steps are drawn versus distance (Figures 6.9(a) to 6.9(d)). y-axix reflects the weight computed from oil production rates of P1, x-axix is the distance between the gridblocks and P1. At the first assimilation step (Figure 6.9(a)), the distance that measurements appear to affect model parameters is about 600 ft, which matches the direction of maximum continuity in variogram. But in the second and third assimilation steps (Figures 6.9(b) and 6.9(c)), the scattered spots show that almost all the gridblocks have either positive or negative values. Intuitively, observations should not so strongly affect reservoir models parameters (*e.g.*, grid block permeabilities) that are so far away. This can cause divergence because of misestimation of the covariance (or Kalman gain), as noted previously. Many of the observa-



(a) After first assimilation

(b) After second assimilation



(c) After third assimilation

(d) After seventh assimilation

Figure 6.9: Kalman gain computed from well P1 *vs.* distance at different assimilation time

tions are expected to be physically unrelated to a particular state variable because they are observations of physically remote quantities. However, some of these observations may be highly correlated with the state variable purely by chance and these spurious correlations will have affect the updated ensemble. The effects of spuriously correlated remote observations can overwhelm more relevant observations.

The finite ensemble size causes the estimated correlations to be noisy. To filter out the error covariance correlation associated with remote observation. The covariance between state variables and observation in the joint state space are multiplied by a correlation function. The correlation function, called Schur product, is a fifth-order piecewise rational function derived by Gaspari and Cohn (1999). In meteorology, it is widely accepted that geopotential-height forecast error correlations should be set to zero beyond distances of a few thousand kilometers in the troposphere. The Schur product method is implemented by multiplication of the sample covariances between the observations and state variable by the distance-dependent factor. It is similar to the inflation factor used in Section 6.2, but the multiplier is a Schur product, ρ^o, which is a compactly supported correlation matrix in Eq. 6.6 which parameterize smoothness for the random field and reduce the computational burden.

$$\boldsymbol{C}^p_{Y,e,k} = \rho^o[\frac{1}{N_e - 1}\sum_{j=1}^{N_e}(\boldsymbol{y}^p_{k,j} - \bar{y}^p_k)(\boldsymbol{y}^p_{k,j} - \bar{y}^p_k)^{\mathrm{T}}], \tag{6.6}$$

We substitute Eq. 6.6 into the Kalman gain equation

$$\boldsymbol{K}_{Y,e,k} = \frac{1}{N_e - 1}\rho^o(\Delta\boldsymbol{y}^p_k\boldsymbol{A}^{\mathrm{T}})\left(\frac{1}{N_e - 1}\rho_o\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{C}_{D,k}\right)^{-1}, \tag{6.7}$$

If we define

$$\boldsymbol{A} = \boldsymbol{H}_k\Delta\boldsymbol{y}^p_k, \tag{6.8}$$

The function ρ^o depends upon the observation location; it is a maximum of 1.0 at the observation location and typically decreases monotonically to zero at some finite distance from the observation. It is a relatively broad function. The fifth-order function proposed by Gaspari and Cohn (1999) is defined as the following Eq. 6.9. It is important to note that the ρ is nonzero only for separation distances less than twice the value of a, which is the critical distance.:

$$\boldsymbol{\rho}^o(a,b) = \begin{cases} -\frac{1}{4}(\frac{b}{a})^5 + \frac{1}{2}(\frac{b}{a})^4 + \frac{5}{8}(\frac{b}{a})^3 - \frac{5}{3}(\frac{b}{a})^2 + 1 & 0 \leqslant b \leqslant a \\ \frac{1}{12}(\frac{b}{a})^5 - \frac{1}{2}(\frac{b}{a})^4 + \frac{5}{8}(\frac{b}{a})^3 + \frac{5}{3}(\frac{b}{a})^2 - 5(\frac{b}{a}) + 4 - \frac{2}{3}(\frac{b}{a})^{-1} & a < b \leqslant 2a \\ 0 & b > 2a \end{cases} \tag{6.9}$$

Where a is defined as a critical distance, which is assumed to be twice the range of variogram. b is the distance between the gridblock and the observation locations.

The effect of Schur product is evaluated with the difficult case of a small ensemble size. For the experiment, the Case II is used: observation is combined every 0.01 and 0.05 pv; there are 40 ensemble members. The critical distance is equal to the range 600 ft, which is the critical distance a. An examination of the σ_t of EnKF with and without Schur product (Figure 6.10) clearly shows the benefits of using correlation function. If the small ensemble size is chosen,

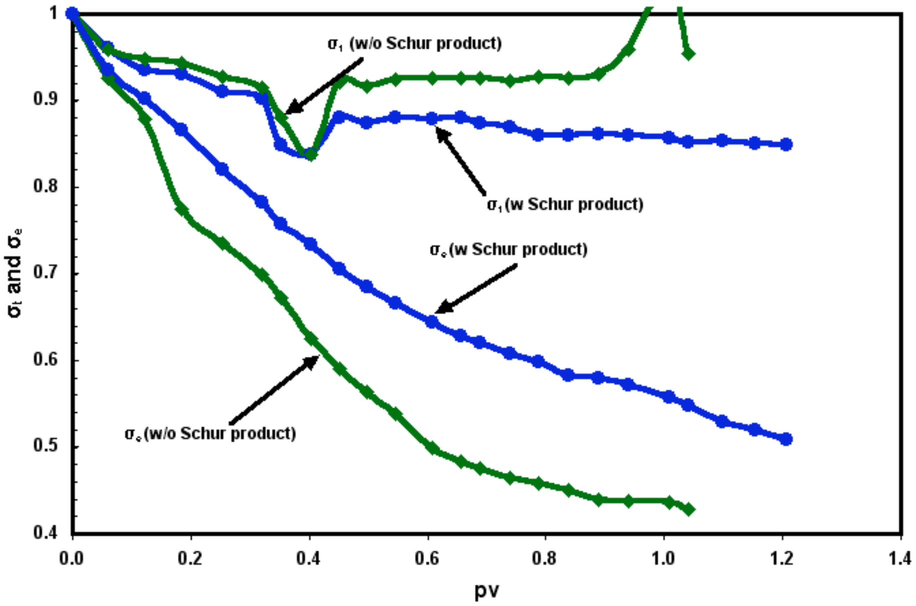


Figure 6.10: Effects of localization on the performance of the EnKF. The σ_t of the ensemble without localization (triangle), and the σ_t of the ensemble with localization (square), the standard error of ensemble σ_e with localization (diamond), the standard error of ensemble σ_e without inflation (circle).

the filter divergence problem can not be avoided. The effect of localization is to increase the effective number of ensemble members. It filters out the small correlations associated with remote observations and the correlation is smooth and monotonically decreasing, produces smooth updating steps.

6.4 Discussion

In this chapter, a 2-D, two phases water flood reservoir model is continuously updated by using EnKF. We examine the case to understand how errors covariances vary with assimilation frequency, ensemble sizes and observation locations. The two main reasons for the instability of filter are the error covariance is systematically underestimated and overestimated. “Underestimate” means that the magnitude of variances were reduced as more data were assimilated. “Overestimate” means the observations are in a distance-dependent manner that the noise (the error) of that observations is larger than the signal (magnitude of the covariance).

We investigate the covariance inflation method by multiplying the ensemble covariance by a uniform covariance inflation factor to compensate for the usual bias of the EnKF to underestimate the analysis uncertainty. For the problem discussed above, the filter remains stable. However, covariance inflation assumes that the model error grows proportional to the error covariance, otherwise the method may cause data assimilation problem. Therefore, a more accurate result could have been obtained if a more complicated inflation approaches is used, *e.g.*, a location-dependent inflation factor could solve some of the problem. We leave more exploration of tuning the EnKF dynamically for future work.

The effect of localization is to increase the effective number of ensemble members. With localization, corrections dependent on the observation location, introducing extra degree of freedom. As the number of ensemble member increases and as the noise in the estimate of weak distance correlation diminishes. The major obstacle must be surmounted in order to apply the localization method widely is the uncertain geological information, *e.g.*, critical distance and the major continuity direction, which is needed to tuning the elements in Schur product. Given a certain ensemble size, an appropriate critical distance can be specified.

Chapter 7

Data Assimilation in Strongly Nonlinear and Non-Gaussian Problems

In this chapter, a synthetic problem is selected and analyzed to investigate one primary concern with the application of the traditional EnKF: is it possible to update state variables whose distribution is non-Gaussian? The synthetic problem is a one-dimensional, two-phase Buckley-Leverett problem (a hyperbolic partial differential equation). In order to alleviate the influence from nonphysical updated state variables, Chen, Oliver, and Zhang (2007) presented an EnKF with reparameterization to attain a more realistic water saturation distribution. However, this method increases computation and data storage requirements. In the following, a simple truncation of nonphysicasl state variables is presented as an alternative to EnKF with reparameterization. The comparison between the two methods is discussed in Section 7.4.

7.1 Example 1: One-Dimensional Buckley-Leverett Problem

For nonlinear or non-Gaussian problems, the EnKF can generate posterior ensembles with nonphysical state variables. For example, in water displacing oil simulations, water saturations may take large values behind the water flooding front ($S_{wc} < S_w < 1 - S_{or} - S_{wc}$), and small values ahead of the front ($S_w \approx S_{wc}$). The distribution of water saturation in gridblocks of the ensemble near the water front is usually bimodal and is not well represented by the mean and variance of a nearly Gaussian distribution.

7.1.1 Generation of the Initial Reservoir Model

The synthetic case is a one-dimensional 32-gridblock reservoir model. The model contains one injector **I1** at grid 1 to inject water and one producer **P**1 at grid 32 to produce oil and water.

One hundred reservoir models with varying porosity and permeability are generated using LU Decomposition simulation (Goovaerts 1997). The porosity is normally distributed with a mean of 0.2 and a standard deviation of 0.04. The permeability is log-normally distributed with a mean of 5.5 (k is in md; 1 md $= 0.9896 * 10^{-12}$ m^2) and the standard deviation of $\ln k$ is 0.5. The two variables have a cross-correlation coefficient of 0.5. The range of the porosity is 18 gridblocks and the variogram model is exponential.

The initial covariance function for porosity in the one-dimensional grid system is

$$C_\phi(i_1, i_2) = \sigma_\phi^2 \exp(-\frac{3(i_1 - i_2)}{a_1}), \tag{7.1}$$

where a_1 is the range of variogram, i_1 and i_2 denote two locations in space. The initial covariance function for $\ln k$ is

$$C_{\ln k}(i_1, i_2) = \sigma_{\ln k}^2 \exp(-\frac{3(i_1 - i_2)}{a_1}), \tag{7.2}$$

and the initial cross-covariance function for porosity and $\ln k$ is

$$C_{\phi,\ln k}(i_1, i_2) = \rho \sigma_\phi \sigma_{lnk} \exp(-\frac{3(i_1 - i_2)}{a_1}), \tag{7.3}$$

A "true" porosity field is created by unconditional LU Decomposition (Goovaerts 1997) the same as the ensemble members. The procedure for the property model construction is

(i) Construct the covariance matrix C of 32 × 32,

$$C(i_1, i_2) = exp(-\frac{3(i_1 - i_2)}{a_1}), \tag{7.4}$$

where we assume the $\ln k$ and ϕ covariances are equal, then $\boldsymbol{C}_{\ln k} = \boldsymbol{C}_\phi$ and $\boldsymbol{L}_{\ln k} = \boldsymbol{L}_\phi$. Therefore in the following equations, both are simplified to $\boldsymbol{C}$ and $\boldsymbol{L}$.

(ii) Decompose it using the Cholesky decomposition

$$\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}, \tag{7.5}$$

where $\boldsymbol{L}$ is the "square root" of the covariance matrix $\boldsymbol{C}$ with Cholesky decomposition.

(iii) Generate the porosity fields

$$\boldsymbol{\phi} = \mu_\phi + \sigma_\phi \boldsymbol{L}\boldsymbol{Z}_1, \tag{7.6}$$

where μ_ϕ is the prior mean of porosity; and Z_1 is a vector of uncorrelated random normal deviates, $Z_1 \in N(0,1)$.

(iv) Generate the permeability field.

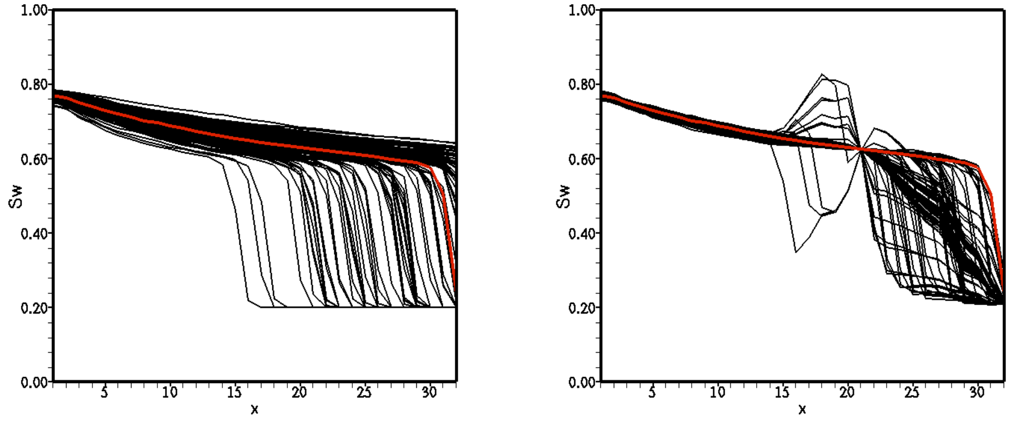
$$\ln \boldsymbol{k} = \mu_{\ln k} + \sigma_{\ln k} L(\rho \boldsymbol{Z}_1 + \sqrt{1-\rho^2}\boldsymbol{Z}_2), \tag{7.7}$$

where $\mu_{\ln k}$ is the prior mean of $\ln k$; $\boldsymbol{Z}_2$ is a vector of uncorrelated random normal deviate, $\boldsymbol{Z}_2 \in N(0,1)$; $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are uncorrelated.

The well constraints used are: bottom-hole pressure for the injector is 4500 psi and 1500 psig for producer. A synthetic, true production record is generated by using the "true" porosity and permeability with initial pressure of 4500 psi, water saturation of 20 percent and perturbed Gaussian noise. The data to be assimilated are water injection rate, oil production rate and the water saturation at grid 21 (an "observation" well). Measurement errors for rate observation are assumed to be Gaussian with mean 0 and standard deviation of 3 percent magnitude of production or injection rate. Usually the error distributions of field observations are poorly known. According to a literature survey, Oglesby (2006) concluded that the current conventional well testing accuracy for determining the flow rates for high water cut wells can range from ±3 percent to ±50 percent. In order to simplify the study, a 3 percent standard error is used when the water cut of producer is less than 20 percent, a 8 percent standard error is used after the water cut of producer is greater than 20 percent.

7.1.2 Solving the Problem with Updated Saturations

The initial assimilation steps should capture most spatial variation features with reduced uncertainty around and between well areas. The following example indicates the importance of assimilating early transient flow data to gain fast recognition of reservoir heterogeneity.



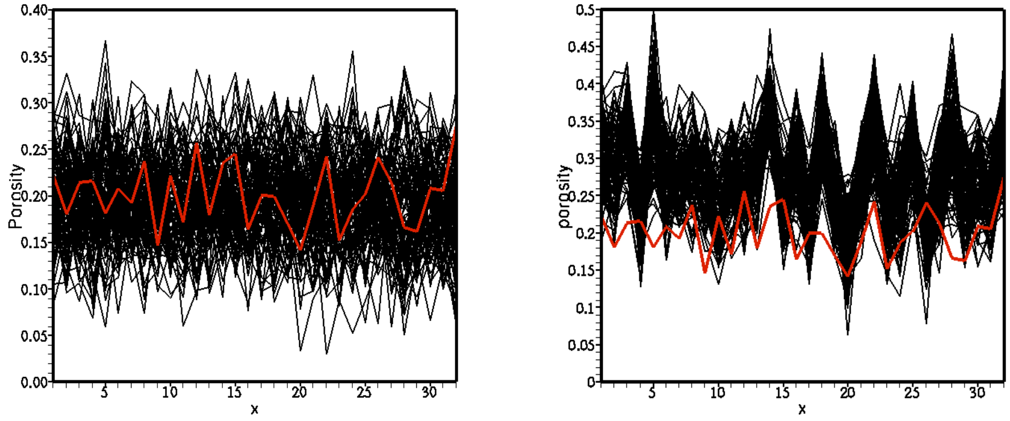
(a) Water saturations before EnKF and water breakthrough in truth.

(b) Updated water saturations after EnKF.

Figure 7.1: Mismatched water saturation profiles because of late EnKF assimilation. Late EnKF assimilation means that the first assimilation is taken when water breakthrough occurs in measurement. Truth denoted by the red curve.

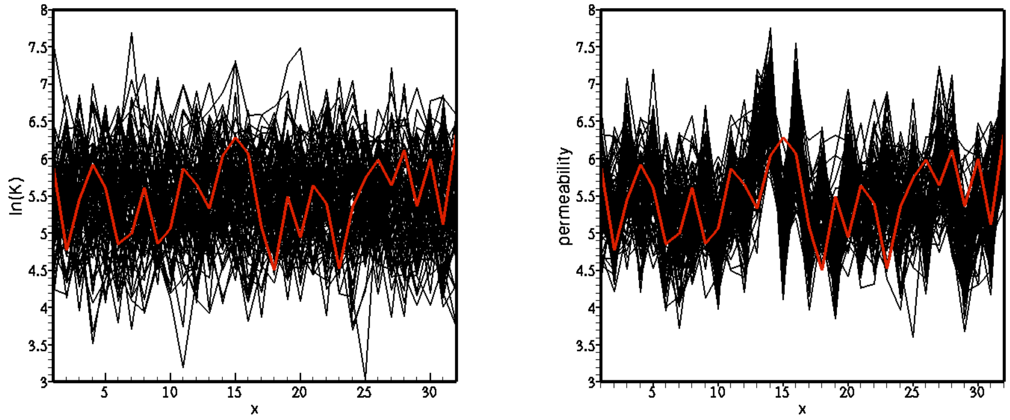
In this case, the first assimilation step (Figure 7.1(a)) begins after water has broken through the producer in the "true" reservoir. Because the initial guess of ensemble properties are generated randomly, some ensemble members with high porosity and permeability have water productions before the first EnKF correction (Figure 7.1(a)), while others have no water breakthrough because of inaccurate property estimates. However (Fig. 7.1(b)), the updated water saturations of all ensemble members (black lines) fail to match the "true" saturation profile (red line). The filter is divergent after several assimilation steps, the updated porosity and permeability differ greatly from truth in Figure 7.2.

Another EnKF failure is observed when insufficient observation data constraints are provided. In this case, assimilation begins at 0.38 pv when the water front does not reach the observation well at the 21st gridblock. The updated porosities of the ensemble are divergent from the "truth" (Figure 7.2(b)). The reason for EnKF failure is the water saturation at gridblock 21 remains at 0.2 for five consecutive assimilation steps. Only the inlet and outlet injection and production rates provide useful flow information to infer average properties of the reservoir and no information between the injector and producer. The properties (es-



(a) Initial porosity of ensemble member.

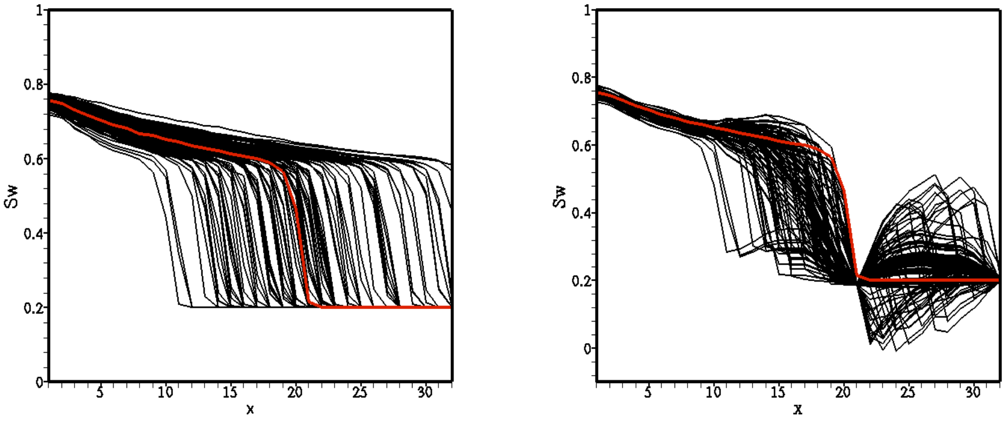
(b) EnKF corrections fails to match the true porosity property.



(c) Initial logarithm permeability of ensemble member.

(d) EnKF correction fails to match the true logarithm permeability property

Figure 7.2: Divergent reservoir properties after late EnKF assimilation timing. Truth denoted by the red curve.



(a) Water saturation before EnKF (b) Water saturation after EnKF

Figure 7.3: First update by EnKF. Water saturation profiles at 65 days with truth denoted by the red curve.

pecially porosity) of the flooded area affects the position of the water front. More detailed information of gridblock property at water flooding area comes from the 21st gridblock water saturation. Therefore, the first assimilation time should be when the water front reaches the 21st gridblock as observed in the "true" case — this observation would be available in an analogous field case. Prior to the first ensemble Kalman filter correction at 65 days, the variation of water front is considerable, the range is between the 12th and 32nd gridblocks (Figure 7.3). The "true" water front has reached observation well at 21st gridblock.

Based on this analysis, 65 days is the first assimilation time when the water front arrives at the observation well. We can see that prior to the first application of correction, the fastest models have water breakthrough, while the slowest model only reaches gridblock 11. The water saturations of ensembles are homogeneously spread around the "truth". After the first EnKF correction, Figure 7.3(b) shows that the saturation difference between ensemble members becomes smaller. However, water saturation profiles do not always decrease monotonically from the injector to the producer and nonphsical values occur $S_w \notin [S_{wc}, 1 - S_{or}]$. The reason is that the distribution of water saturations is non-Gaussian and bimodal (Figure 7.4), whereas the EnKF assumes Gaussian distributions for state vectors.

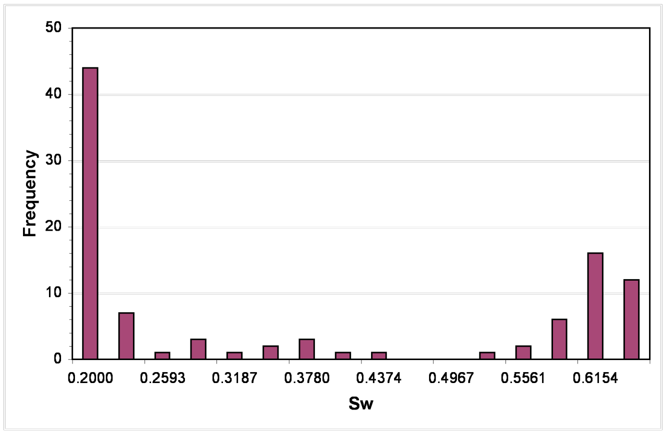
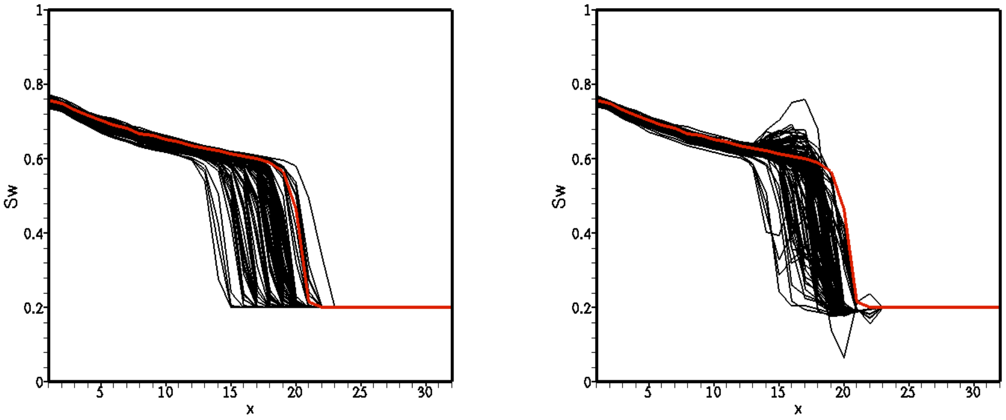


Figure 7.4: Water saturation histogram at gridblock 20

Gu and Oliver (2006) have applied three methods to resolve the problem: (1) use of normal score transform to the saturation front location as a state vector; (2) use of location of saturation front, instead of saturation as the state vector; and (3) iterating the update. Unfortunately, after the normal score transform, the water saturation still oscillates between its high and low value. Besides, this method is not straightforward for two- or three-dimensional problems. The second method is not applicable to two- and three-dimensions either. Other alternative approaches, such as logarithms and power transformations, are not be able to fully transform the bimodal distribution of water saturation to normal distribution. Therefore, these approaches are not suitable for solve our problem. The third method is used whenever the saturation is detected outside the range ([0.2,0.8] in this case), or does not monotonically decrease $[S_w(x + \ell\Delta x) \geq S_w(x), \ell \geq 1]$. The simulator will rerun the previous assimilation interval to attain the current time step dynamic variables (pressure and saturation) again but using the updated model parameters (porosity and permeability) obtained at the current assimilation. After the forecast step, the state vectors and the observations are used to apply EnKF correction again. The iteration continues until the corrected saturation profiles satisfy the physical boundary or the iteration exceeds a preset maximum (3 times in the study because of computation costs). Figure 7.5(a) displays the simulated water saturation profiles at first iteration with updated model parameters during the first mea-



(a) Water saturation before EnKF

(b) Water saturation after EnKF

Figure 7.5: Water saturation contours after iteration at 65 days with truth denoted by the red curve.

surement is available. Comparing with Figure 7.3(a), the saturation profiles are improved with smaller differences among ensemble members because of EnKF correction. After the second correction by the EnKF, the saturation profiles becomes as shown in Figure 7.5(b). Unlike Figure 7.3, the results have improved but some of the saturation values are out of physical boundary or do not monotonically decrease from the injector to producer. The non-physical saturation values can occur from an EnKF update step if there is a wide range of model parameters and bimodal parameter vectors in the ensemble. One problem of iteration is the computational burden. Another problem with iteration is the observation data are assimilated more than once, the error covariances of state vector become too small during the iteration procedure which tends to incorrectly reduce predicted uncertainties of reservoir properties and may result in filter divergence. In the following section, we show that Zafari (2007) has proved mathematically that the iteration is not correct to be applied to resolve non-Gaussian and nonlinear problems by using a set of linear relationship. The iteration may lead to inconsistency for the EnKF.

7.1.3 Error Covariance Problem with the Iteration Method

Zafari (2007) indicated the inconsistency of the confirming EnKF (Wen and Chen 2005a) by using a set of linear relationships. The same method is used here to prove the error covariance problem caused by the iteration. In the EnKF, to avoid the explicit computation of $\boldsymbol{C}^p_{Y,e,k}$, substitute $\boldsymbol{C}^p_{Y,e,k} = \frac{1}{N_e-1}\Delta\boldsymbol{Y}^p_k(\Delta\boldsymbol{Y}^p_k)^{\mathrm{T}}$ in Kalman gain, then the update state vector by EnKF can be rewritten in terms of matrix of ensembles as,

$$\begin{aligned}
\boldsymbol{Y}^u &= \boldsymbol{Y}^p + \frac{\Delta\boldsymbol{Y}^p\Delta\boldsymbol{D}^{\mathrm{T}}}{N_e-1}\left(\frac{\Delta\boldsymbol{D}\Delta\boldsymbol{D}^{\mathrm{T}}}{N_e-1} + \boldsymbol{C}_D\right)^{-1}(\boldsymbol{D}_{obs} - \boldsymbol{H}\boldsymbol{Y}) \\
&= \boldsymbol{Y}^p\big(\boldsymbol{I} + \frac{\Delta\boldsymbol{D}^{\mathrm{T}}}{N_e-1}[\frac{\Delta\boldsymbol{D}\Delta\boldsymbol{D}^{\mathrm{T}}}{N_e-1} + \boldsymbol{C}_D])^{-1}(\boldsymbol{D}_{obs} - \boldsymbol{D})\big) - \bar{\boldsymbol{Y}}^p\delta\boldsymbol{D} \\
&= \boldsymbol{Y}^p(\boldsymbol{I} + \boldsymbol{\delta}(\boldsymbol{D})) - \bar{\boldsymbol{Y}}^p\delta\boldsymbol{D}
\end{aligned} \tag{7.8}$$

where $\boldsymbol{D} = \boldsymbol{H}_k\Delta\boldsymbol{Y}^p$ and

$$\boldsymbol{\delta}D = \frac{\Delta\boldsymbol{D}^{\mathrm{T}}}{N_e-1}[\frac{\Delta\boldsymbol{D}\Delta\boldsymbol{D}^{\mathrm{T}}}{N_e-1} + \boldsymbol{C}_D]^{-1}(\boldsymbol{D}_{obs} - \boldsymbol{D}) \tag{7.9}$$

Note that model parameters, model solutions and computed responses are all updated using the same coefficient matrix, $(\boldsymbol{I} + \boldsymbol{\delta}(\boldsymbol{D}))$ and the constant vector, $\bar{\boldsymbol{Y}}^p\delta\boldsymbol{D}$. Eq. 7.8 indicates the updating coefficient is only a function of the matrix $\boldsymbol{D}$. Suppose we have the following linear relationships between model solution $\boldsymbol{P}$, data $\boldsymbol{D}$ and model parameter $\boldsymbol{M}$,

$$\boldsymbol{P}^{k+1} = \boldsymbol{F}_k\boldsymbol{M}^k + \boldsymbol{A}_k\boldsymbol{P}^k + \boldsymbol{\alpha}_k\bar{\boldsymbol{I}} \tag{7.10}$$

$$\boldsymbol{D}^{k+1} = \boldsymbol{G}_k\boldsymbol{M}^k + \boldsymbol{B}_k\boldsymbol{P}^k + \boldsymbol{\beta}_k\bar{\boldsymbol{I}}, \tag{7.11}$$

where $\boldsymbol{F}_k$, $\boldsymbol{A}_k$, $\boldsymbol{G}_k$ and $\boldsymbol{B}_k$ are matrices and $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ are known vectors. k is the time step, $\boldsymbol{P}^{k+1}$ and $\boldsymbol{D}^{k+1}$ are state solutions and data at time step $k+1$, which is the assimilation time. $\boldsymbol{D}^{k+1}$ is given as a function of $\boldsymbol{M}^k$ and $\boldsymbol{P}^{k+1}$, or a function of $\boldsymbol{M}^k$ and $\boldsymbol{P}^k$ but the coefficient matrices would be different. The following equations are the same equations but using form of matrix of ensembles, where $\bar{\boldsymbol{I}}$ is an N_e-dimensional row vector with each entry equal to 1. When there are data at $(k+1)^{th}$, from Eq. 7.10, if the standard EnKF is used to update the model parameters,

$$\begin{aligned}\boldsymbol{P}^{k+1,u} &= \boldsymbol{P}^{k+1}(\boldsymbol{I}+\delta(\boldsymbol{D}^{k+1}))\\ &= \boldsymbol{F}_k\boldsymbol{M}^k(\boldsymbol{I}+\delta(\boldsymbol{D}^{k+1}))+\boldsymbol{A}_k\boldsymbol{P}^k(\boldsymbol{I}+\delta(\boldsymbol{D}^{k+1}))+\alpha_k\bar{I}\end{aligned} \tag{7.12}$$

But if the iterative step is taken, the equation becomes,

$$\boldsymbol{P}^{k+1,r} = \boldsymbol{F}_k\boldsymbol{M}^k(\boldsymbol{I}+\delta(\boldsymbol{D}^{k+1}))+\boldsymbol{A}_k\boldsymbol{P}^k+\alpha_k\bar{I} \tag{7.13}$$

Now we continue assimilation of data for one more time. For the case that we rerun the time step in iteration would become,

$$\begin{aligned}\boldsymbol{P}^{k+1,r,u} = \boldsymbol{F}_k\boldsymbol{M}^k(\boldsymbol{I}+\delta(\boldsymbol{D}^{k+1}))(\boldsymbol{I}+\delta(\boldsymbol{D}\prime^{k+1}))+\boldsymbol{A}_k(\boldsymbol{F}_k\boldsymbol{M}^k(\boldsymbol{I}+\delta(\boldsymbol{D}^{k+1}))\\ +\boldsymbol{A}_k\boldsymbol{P}^k+\alpha_k\bar{I})(\boldsymbol{I}+\delta(\boldsymbol{D}'^{(k+1)}))+\alpha_k\bar{I}\end{aligned} \tag{7.14}$$

Eq. 7.12 and 7.14 are not the same, as can be seen in the second term on right hand side of Eq. 7.14, the current updated model parameters has been used for last time step. Moreover, the $\delta(\boldsymbol{D}'^{(k+1)})$ is different from $\delta(\boldsymbol{D}^{k+1})$. The inconsistency of the current updated parameters used in the last forecast step might cause problems. Thus, the iteration method does not appear to be feasible to solve the nonlinear and/or non-Gaussian problem.

7.2 Ensemble Kalman Filter with Reparameterization

In the forecasting step of EnKF, simulation is able to propagate the full probability distribution of the model, while at the updating step, the EnKF relies on the ensemble mean and variance. The Gaussian assumption makes the updating step easy to implement but may result in errors when this assumption is violated. The non-Gaussian probability distribution of the model state vector most likely comes from the nonlinearity of the dynamic model. A normally distributed initial state could become strongly non-Gaussian after evolving it with the nonlinear model.

There are some methods available to solve the problem caused by nonlinear models without the Gaussian assumption. Miller, Carter, and Blue (1999) constructed the pdf function for the model evolution and calculated the conditional density in terms of the analytical function according to the maximum-likelihood analysis. Their method is less limiting than the EnKF since the assumption of Gaussianity is not required. But it is computationally demanding when applied to large scale problems.

7.2.1 State Vector Construction

In this section, EnKF with reparameterization is applied for a one-dimensional Buckley-Levertt waterflooding problem. The objective is to obtain better performance of EnKF through improving the Gaussian distribution of the model solutions, mainly water saturation. The EnKF with reparameterization method evolves from the fact that the water saturation distribution away from the water front is approximately Gaussian, whereas the distribution in the water front region is bimodal. Thus, the reparameterization needs to be done only for the front area where the gridblock water saturation distribution is non-Gaussian. The water front arrival time is correlated to the reservoir properties, especially porosity. And it is well modeled by a Gaussian distribution. Since the water front arrival time is defined for gridlocks, there are some additional steps compared with traditional EnKF updating with more computational burdens and storage requirements.

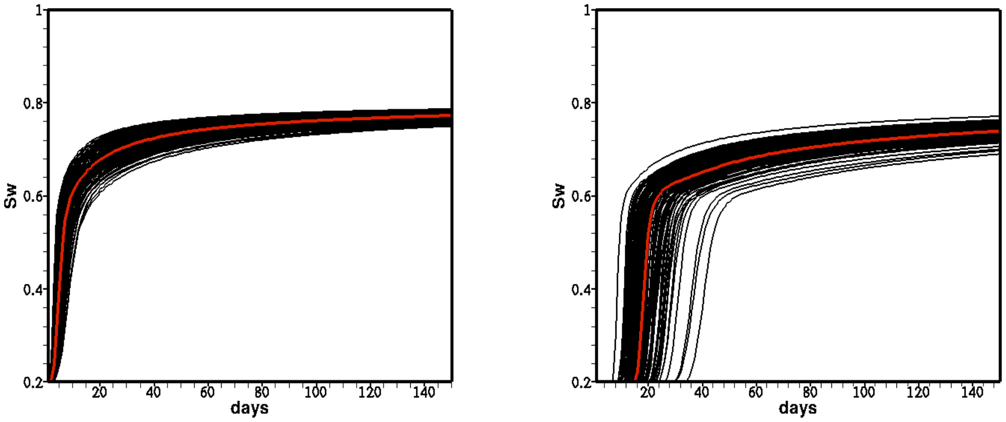
As mentioned previously, the reparameterization approach uses water front arrival time but saturation as the state vector in the water front domain. Thus, the new state vector consists of porosity (ϕ), logarithm permeability ($\ln k$), and pressure (p) at every gridblock, water front arrival time (T) at the gridblocks within the water front area, and water saturations (S_w) in the remaining gridblocks:

$$\begin{aligned}\boldsymbol{y} =& \{\phi_1, \ldots, \phi_N, \ln k_1, \ldots, \ln k_N, p_1, \ldots, p_N, S_{w,1}, \ldots, S_{w,N_1-1}, \\ & T_{N_1}, \ldots, T_{N_2}, S_{w,N_2+1}, \ldots, S_{w,N}\}\end{aligned} \tag{7.15}$$

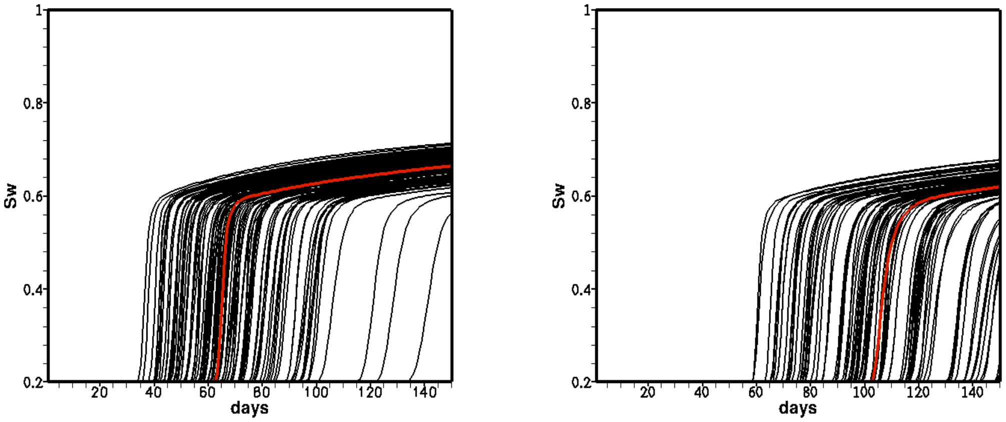
where N_1 and N_2 are the starting and the ending gridblocks of the front area, respectively; N is number of gridblocks. The workflow of EnKF with reparameterization is different from traditional EnKF. Some steps of the distributed EnKF work flow calls for some changes. The modified steps are summarized as followed:

(i) After all the state vectors are transferred from ensemble machines to the Kalman gain master processor, the water saturations of all ensemble members are searched through for the water front domain, which is the start and end gridblocks of water front among all ensemble members (denoted by N_1 and N_2) in 1-D problem. The water front arrival time is defined as the time when the water saturation reaches 0.40 at a certain gridblock. For example, in Figure 7.3(a), N_1 and N_2 are 11 and 32 respectively.
(ii) In order to obtain the water arrival time of all ensemble members for each grid block in the water front area, simulation beyond the current assimilation time is needed to obtain the saturation *vs.* time relationship (called time window) for each gridblock in front area. One more task farming is needed to accomplish the extra simulations for water front arrival time information.
(iii) The time window for each ensemble member is extracted from the simulation output and GridFTPed back from distributed machines to Kalman gain processor for Kalman gain computation (Eq. 7.15).
(iv) The time window array is saved in the Kalman gain processor. After EnKF correction, the updated water front arrival time would be transformed back to saturation by the stored time window.

Figure 7.6 shows the time windows for gridblocks 2, 6, 20 and 32. Gridblocks 2 and 6 (Figure 7.6(a) and 7.6(b)) are behind the water front domain and water saturation is used in the state vector for assimilating data at 65 days. Realizations of the front arrival time at the gridblocks 20 and 32 can be read through the time window by drawing a horizontal line at saturation 0.4 in Figures 7.6(c) and 7.6(d). It is obvious that the distribution of water front arrival time in Figure 7.7 is more Gaussian than water saturation distribution in Figure 7.4.



(a) 2^{nd} Grid block (b) 6^{th} Grid block



(c) 20^{th} Grid block (d) 32^{nd} Grid block

Figure 7.6: The water saturation versus time at initial condition with red curve as referece

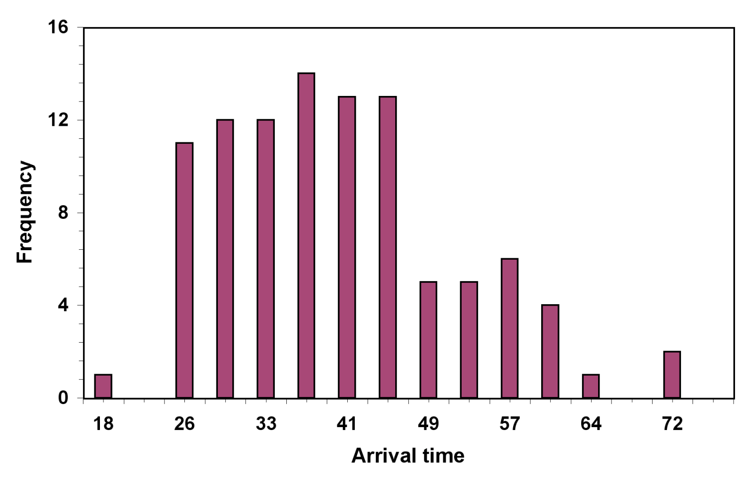
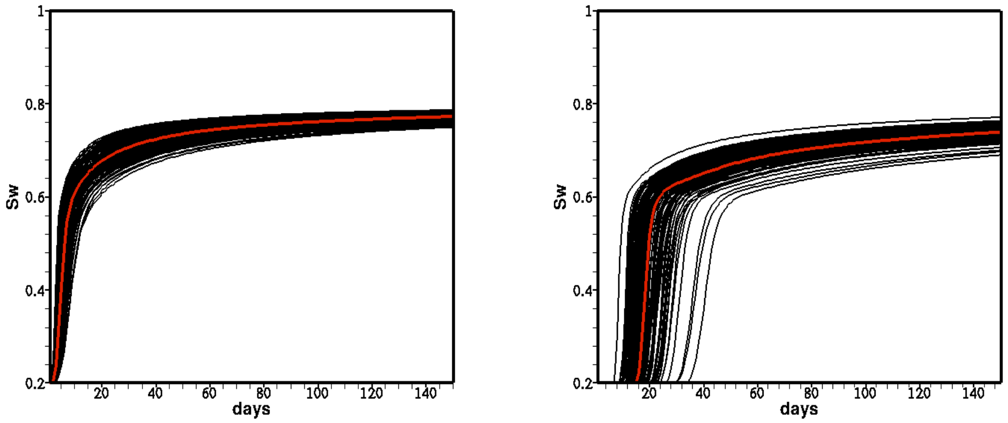


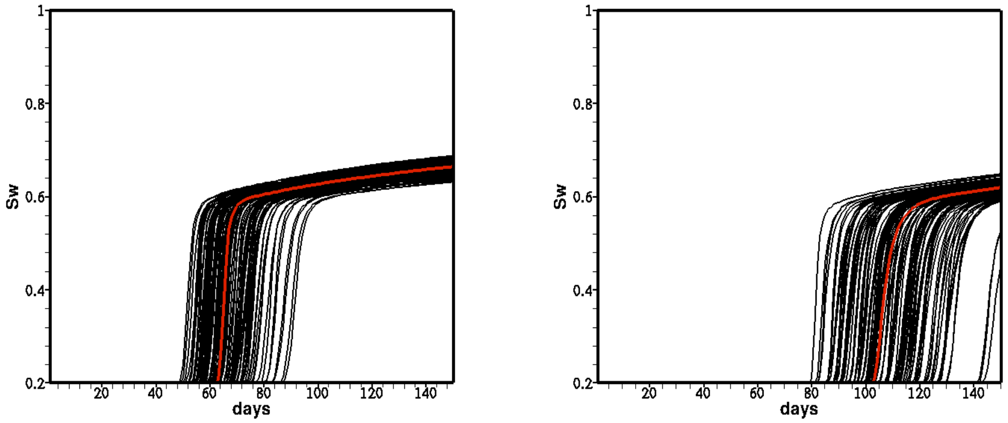
Figure 7.7: Distribution of the water front arrival time at gridblock 20 (Compared with bimodal distribution of water saturation at Figure 7.4

7.2.2 EnKF Assimilation

To obtain the updated saturation ensemble, the EnKF with reparameterization needs an extra step, since water front arrival times are included in the state vector and updated for the water front area. Figure 7.8 is the time window after an assimilation. The time windows remain unchanged if Figure 7.6(a) is compared with Figure 7.8(a), and Figure 7.6(b) is compared with Figure 7.8(b) because the saturations of the two gridblocks are updated instead of arrival times. Figures 7.8(c) and 7.8(d) show less fluctuation and get closer to the truth compared with Figures 7.6(c) and 7.6(d). Then the saturations of the water front area at the current assimilation time for each member can be interpolated from the newly updated water breakthrough curve in Figure 7.8(c) and Figure 7.8(d). The assimilation steps for EnKF with reparameterization and the traditional EnKF are the same. The updated saturation profiles from the traditional EnKF are shown in Figure 7.9(c). The fluctuation is substantial and many updated saturation values are beyond the physical bounds. Figure 7.9(b) is the result of EnKF with reparameterization. An improved match is noted as the ensemble members are getting closer to the truth. The advantage of EnKF with reparameterization is the improved water saturation estimate to honor future observation. But the extra computation effort for

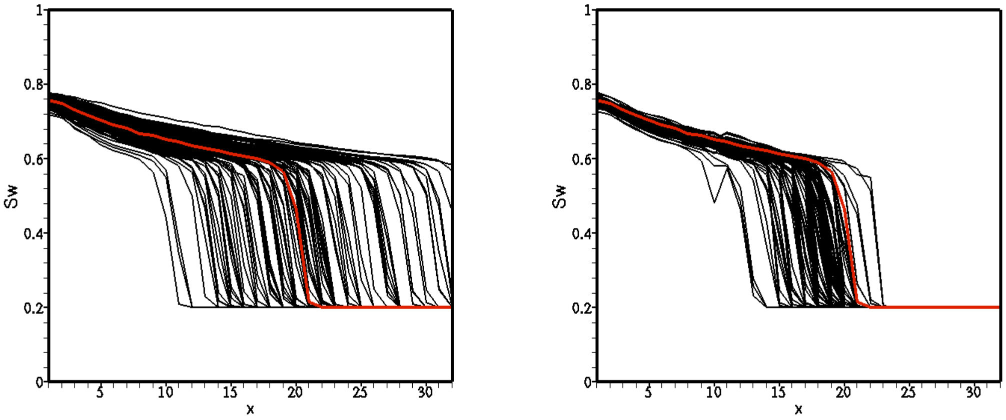


(a) 2^{nd} Grid block (b) 6^{th} Grid block



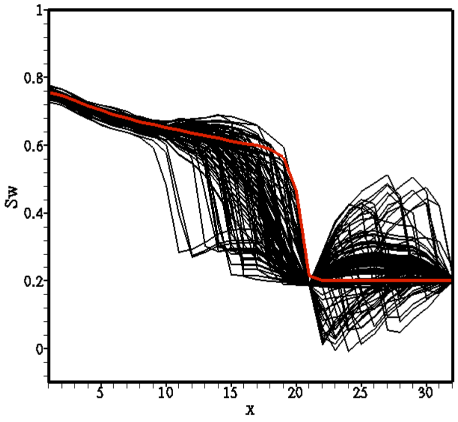
(c) 20^{th} Grid block (d) 32^{nd} Grid block

Figure 7.8: The water saturation versus time after EnKF with red curve as referece



(a) Before EnKF

(b) After EnKF with reparameterization



(c) After traditional EnKF

Figure 7.9: The water saturation versus time with red curve as truth

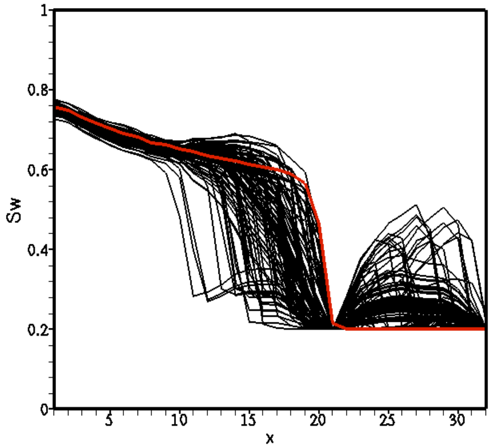


Figure 7.10: Truncated water saturation profile.

simulation, data transfer by the network and storage requirement make it difficult to apply for dealing with a 2-D or 3-D reservoir model. Next, an alternative EnKF with nonphysical values truncation, is discussed and compared with EnKF with reparameterization.

7.3 EnKF with Truncation

EnKF with reparameterization loses the simplicity and efficiency of traditional EnKF because of the extra computation, data transfer and storage burden. To avoid the problem, when the updating step generates unreasonable values for time dependent state solutions, simple truncation of nonphysical values to physical ones not only honors the observations, but also make the work flow simpler to implement. The distribution of water saturation is always nonlinear and/or non-Gaussian and easily has nonphysical values after updating. If an updated saturation is below connate, then replace it by connate water saturation as in Figure 7.10 (similarly, $S'_w = \min(1 - S_{or}, S_w)$).

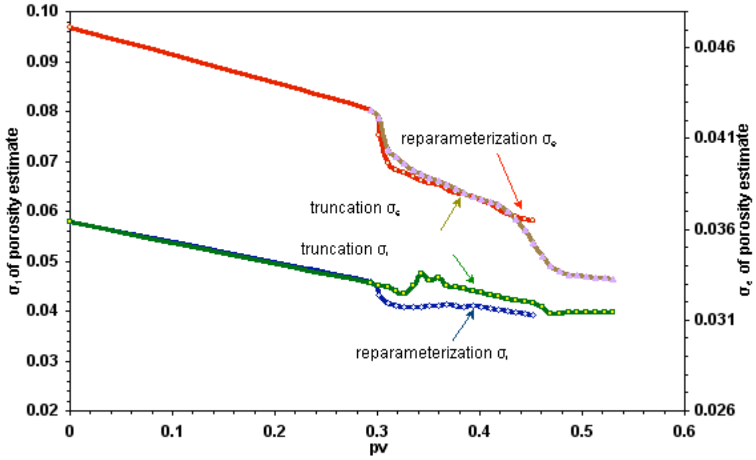
7.4 Comparison of EnKF with Reparameterization and Truncation

In the following section, the two methods are compared using the 1-D waterflooding case. Whereas 1-D water flooding is like a streamline problem, the water front is generally smoother in 2-D or 3-D due to additional dispersion in higher dimensions. Therefore, a 1-D problem is the most difficult test for non-Gaussianity and nonlinearity using the EnKF.

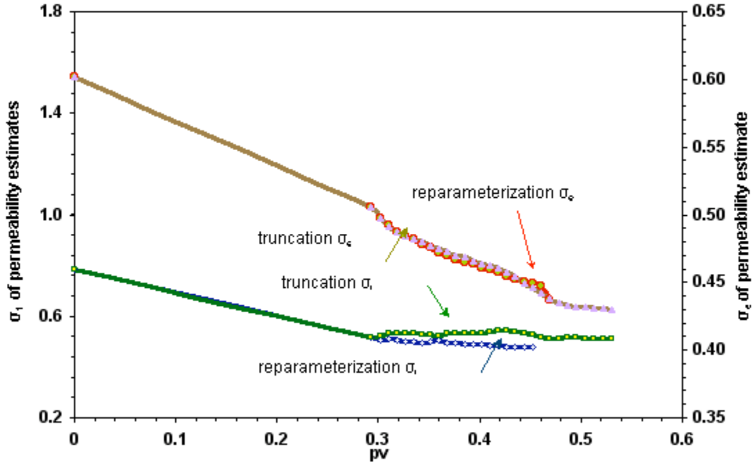
The "truth" case takes 65 days (0.281 pv) to reach gridblock 21 where the observation well is and breaks through the producer in 104 days (0.44 pv). The first assimilation is taken at 65 days when the water front reaches the observation well. The perturbed observation data are assimilated every two days. The assimilation stops after water breakthrough for EnKF with the reparameterization method. After the water breakthrough, the distribution of water saturation is no longer strongly non-linear and/or non Gaussian. Therefore, the total assimilation steps for EnKF with reparameterization and truncation are 20 and 29, respectively.

7.4.1 Comparison of Model Parameter Estimates

The error relative to the "truth" (σ_t) and standard error of ensemble (σ_e) for permeability and porosity estimates for the two methods are presented in Figures 7.11(a) and 7.11(b). After the first assimilation, both σ_t and σ_e decrease dramatically from their initial values. That means the objective function has gone down from 2.4 to 1.1 for permeability and from 0.164 to 0.086 for porosity. Compared with true porosity and permeability, σ_t of the EnKF with reparameterization demonstrates better match than truncation. However, the t-tests between the EnKF with reparameterization and with truncation are not significantly different statistically in Table 7.1. The F-tests indicate no difference in variance estimates. The maximum relative error of σ_t for porosity and $\ln k$ (Fig. 7.12) is 16 percent and 0.12 percent, respectively.



(a) Porosity results



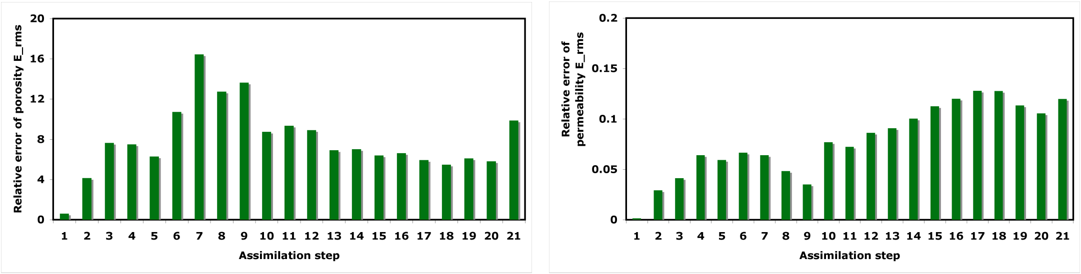
(b) Permeability results

Figure 7.11: σ_ts and σ_es of porosity and logarithm permeability estimates from truncation and reparameterization

Table 7.1: t-test and F-test for σ_t between the EnKF with reparameterization and with truncation

.

Statistical test	p-value
$t - test$	0.6924
$F - test$	8.82E-06



(a) Porosity results

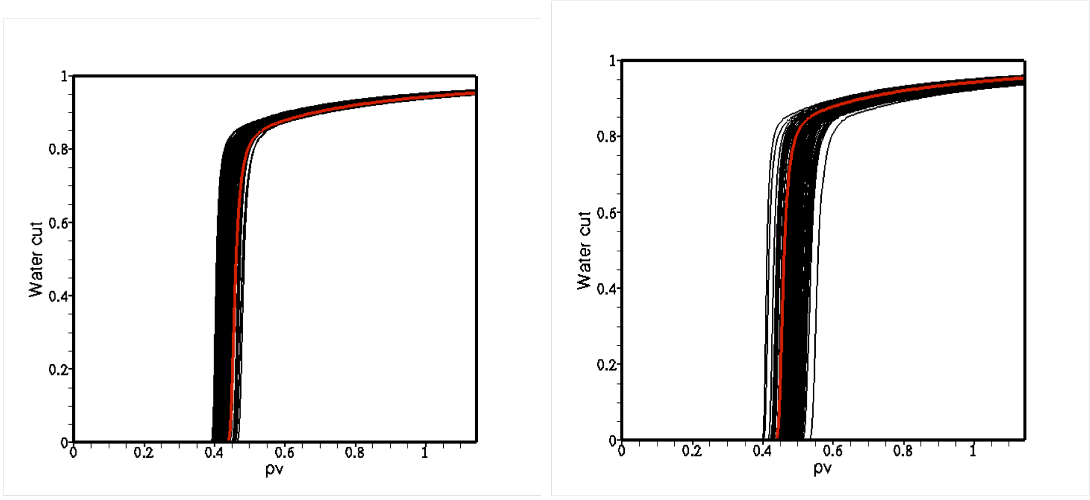
(b) Permeability results

Figure 7.12: Relative error of the porosity and logarithm permeability estimates between the EnKF with truncation and EnKF with reparameterization.

7.4.2 Comparison of Forecast Mismatches

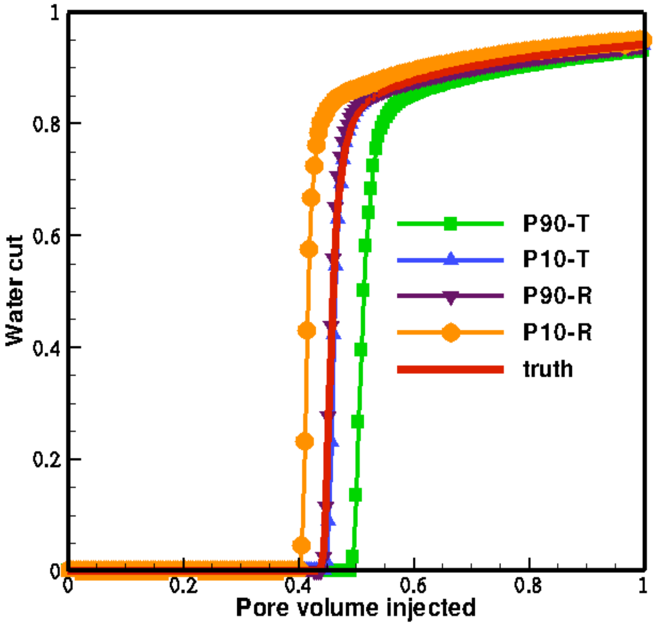
A concern with EnKF is that the permeability and porosity fields update with time as more data are assimilated. It seems somewhat likely that the final permeability and porosity fields resulting from 200 days assimilation would no longer honor the observation data at earlier times. The comparison of the “truth” (red line) with the simulated values of the observations (black lines) from the 0.45 pv injection (water breakthrough) assimilated ensemble are taken back to time 0 and rerun (Figures 7.13(a) and 7.13(b)). The forecast results of the two methods are presented in terms of a “P_{10}” and “P_{90}” uncertainty band (Figure 7.13(c)). The lowest water cut (circle line) represents the ninth lowest value predicted with one hundred ensemble of reparameterization method, same as the “P_{10}” of truncation method (upper-triangle line). The square line and lower-triangle line denote the “P_{90}” of truncation and reparameterization method, respectively. The uncertainties of the truncation method is a little bit larger than the reparamterization. “Truth” is not within the band of “P_{10}-P_{90}”.

To quantitatively compare the forecast results of the two method, the σ_ts are calculated based on forecasting on “truth” (Figure 7.14). The errors occur at water breakthrough time for both methods and the magnitudes of σ_t are similar. Then, we continue the assimilations to 0.56 pv injection using traditional EnKF to update the reparamterization model and truncation models. After the two models are rerun, they match the “truth” closely (Figure



(a) Forecast by the EnKF with reparameterization

(b) Forecast by the EnKF with truncation



(c) Forecast P_{10} and P_{90} comparison between the EnKF with reparameterization

Figure 7.13: P4 water cut forecast comparisons between truncation and reparameterization, 0.45 pv injection assimilation

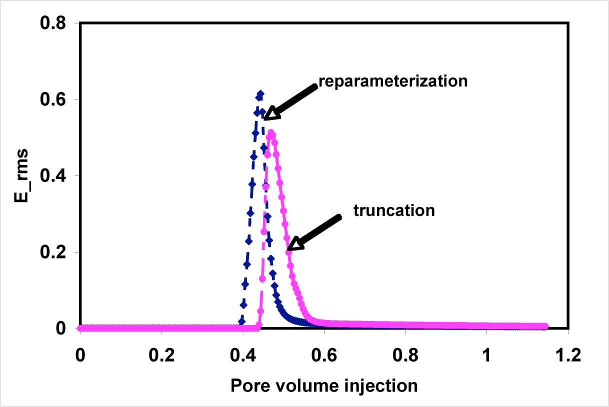
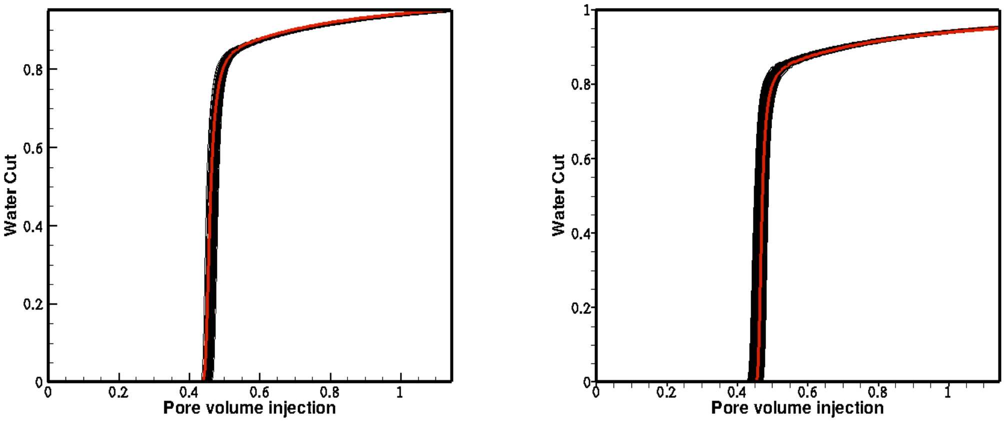


Figure 7.14: P4 water cut forecast σ_t comparison between the EnKF with truncation and reparameterization.



(a) Forecast comparison between the EnKF with reparamterization

(b) Forecast comparison between the EnKF with truncation

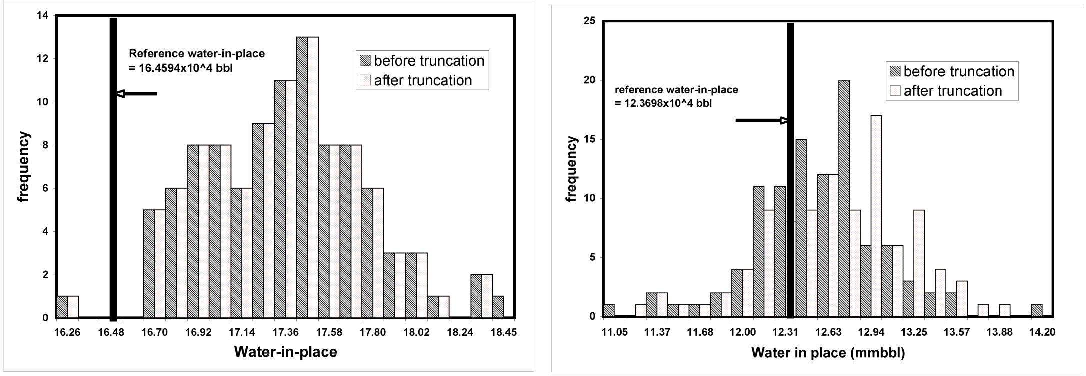
Figure 7.15: P4 water cut forecast by the EnKF with truncation, assimilation 0.56 pore volume injection.

7.15) and fewer difference after more flow information has been integrated. There is no large difference on the model parameter estimates by the two methods.

7.4.3 Consistency Check for the EnKF with Truncation

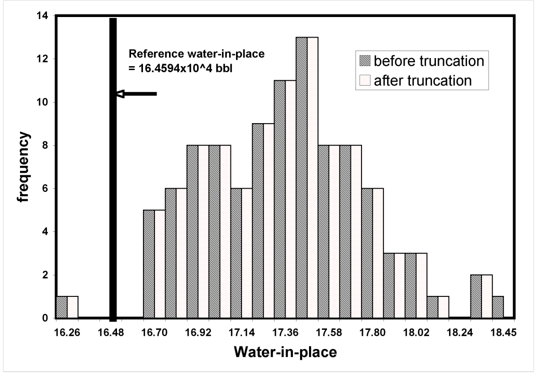
In traditional history matching, only the permeability and porosity fields would be updated by an optimization method. The pressure and saturation fields that are consistent with the permeability and porosity fields would be computed by running the reservoir simulator. In the EnKF method, the permeability, porosity, pressure, and saturation are all updated simultaneously at updated steps. Then a linearized approximation to the simulator is used to make a prediction of the saturation and pressure changes that should result from the porosity and permeability changes.

It is not clear, however, that material balance is honored after the truncation of non-physical water saturation. In order to check the global material balance error, we computed the total water-in-place ($\sum_{i=1}^{N_m} V_i \phi_i S_{w,i}$) for the ensemble for the first assimilation step and 0.45 pv assimlation: (1) comparing saturations and porosity before the truncation, and (2) comparing with the “true” water-in-place. If the differences in the computed water-in-place values, for each ensemble member, for the two comparisons are small, it implies that the truncations are valid and that the material balance is generally honored in truncation method. In the first assimilation step, the mean of water-in-place distribution for 100 ensemble members before truncation is around 12.78×10^4 bbl. After the truncation of non-physical water saturations, the mean of water-in-place distribution is 12.94×10^4 bbl. The relative error is approximately 1.3 percent. The 0.54 pv assimilation, the distributions of water-in-place before and after truncation are the same. In this case, the truncation method seems to honor the material balance. The mass conservation is kept by EnKF with truncation method. Note the increased relative errors between the updated model from the first assimilation step (Figure 7.17(a)) to 0.54 pv assimilation (Figure 7.17(b)). The mean value goes up from 4.15 to 5.23 percent. The relative errors of some models have reached 11 percent. The EnKF reduces a nonlinear minimization problem in a huge parameter space, involving the minimization of an objective function with multiple local minima, to a statistical minimization problem in the ensemble space. Thus, by searching for the mean rather than the mode of the posterior pdf. Pore volume is indeed uncertain, and we are adjusting it. We would do the same with a traditional gradient approach. The difference here as that the gradient method uses the PDE only to evolve, and the EnKF uses covariances as well. Thus, whereas 11 percent would be a huge error for a simulator, it is probably acceptable for the EnKF.



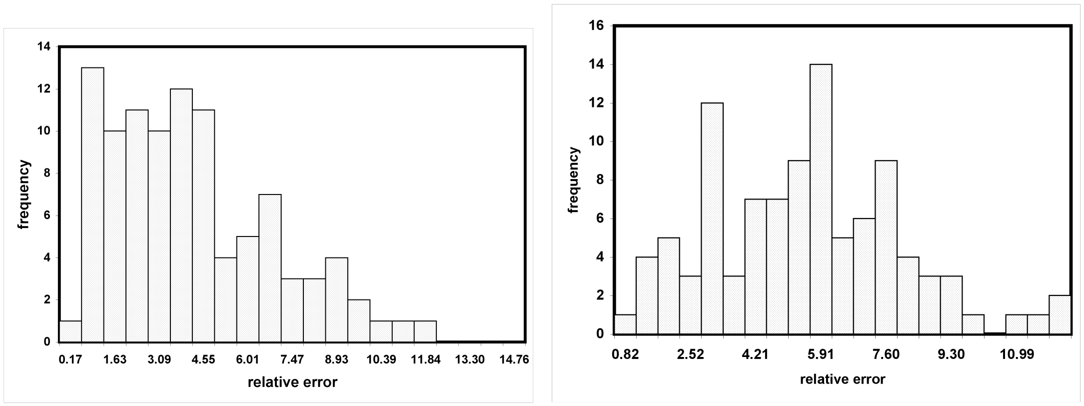
(a) Initial models

(b) The first assimilation



(c) 0.54 pv assimilation

Figure 7.16: Distribution of water-in-place of the EnKF with truncation method



(a) The first assimilation

(b) 0.54 pv assimilation

Figure 7.17: Relative error of water-in-place between truncation and “truth”

7.5 Discussion

When the distribution of the state vector, *i.e.*, saturation, is not normally distributed, non-physical values may be generated by the EnKF. The problem cannot be resolved by EnKF with iteration because the reduced error covariance often results in a filter divergence problem.

EnKF with reparameterization can reduce the non-Gaussian effects in the presence of the sharp water front as illustrated by the one-dimensional example. EnKF with reparameterization better adjusts the water saturations compared with the traditional EnKF, as a result of the strongly non-Gaussian distribution of water saturation around the front area. Since the EnKF is a sequential method, the improved estimate of the state variables improves the ability to assimilate future observations for EnKF with reparameterization.

EnKF with reparameterization has a high demand for computation time and storage space, which hinders its application (especially in 2 or 3 dimensions). The large amount of time needed for time window data transfer poses a challenge for the grid environment. From this research, we found that model parameter estimates of a simple EnKF with truncation are similar to EnKF with reparameterization. Therefore, the EnKF with reparameterization can be used for the initial assimilation steps when the nonlinear and non-Gaussian conditions are strong. The EnKF with truncation can also be used to continue the process (even if one chooses reparameterization for earlier steps).

Chapter 8
Discussions

This chapter gathers several topics of particular interest, highlighting the theoretical and practical aspects of the EnKF in a Grid environment and pointing out remaining challenges. In Section 8.1, a proposed sampling method to efficiently select representative ensemble in small size is introduced. Section 8.2 provides some thoughts about extending distributed EnKF to ensemble-based closed-loop work flow. Resource pricing and run strategy at a Grid environment is discussed in Section 8.3. Some suggestions about the distributed EnKF improvement are introduced in the section. The last section is the discussion of optimal EnKF process based on work done in Chapter 6 and 7.

8.1 Improved Ensemble Selection Using Many Secondary Responses

EnKF, as a stochastic inversion methods, provides multiple reservoir models and forecasts, and integrates production and geologic data. The geologic knowledge encapsulated in the prior ensemble of reservoir models determines the diversity of the posterior ensemble. Because the error covariance is estimated by Monte Carlo methods in traditional EnKF, the estimation might suffer from the spurious correlation when the size of the ensemble is small. We propose a method to choose essentially diverse yet manageably small prior ensembles to improve the efficiency of EnKF.

To make prior ensembles manageably small, a subsample is commonly drawn using a univariate ranking of an easy-to-estimate secondary flow model response. The secondary response should correlate with full-physics responses such as cumulative oil recovery. However, there are two complications. First, the rank depends on the analyzed response; for example, using recovery efficiency and breakthrough time will give different ranks. Second, model rank may depend on engineering factors such as well spacing or completion geometry.These complications are addressed by embracing the full joint, multivariate distribution of many secondary responses. The secondary responses are chosen to reflect important reservoir characteristics such as connectivity, pore volume, anisotropy, and conductivity; and to reduce dependence on specific development scenarios. Models that are diverse in these secondary responses should be diverse in the full-physics responses to be simulated later. To create the prior ensemble, we choose realizations from this multivariate secondary response (M2R) space and use a quasirandom method to ensure the sample is both diverse and representative.

It seems reasonable to use single-phase tracer simulations for screening; the secondary responses include injectivity, Lorenz coefficients, and residence time statistics. Although the screening simulation simplifies the physics and operational constraints, tracer simulations include full geomodels. Various injector-producer pairs sample geomodel heterogeneity and anisotropy. This provides hundreds of secondary responses to describe the geomodels. Principal component analysis reduces secondary response dimensionality and diagonalizes the response covariance matrix, which simplifies sampling. Then, low-discrepency quasi-Monte Carlo Hammersley sequences samples the secondary response principal component space using a nearest neighbor approach.

This work is being pursued currently , but is not included in this dissertation. In the ongoing work, the multivariate secondary response (M2R) method for full simulations, a naive random sample and two different univariate random samples are used to select 100 realizations from a suite of 1000 geomodels separately. These 100-member ensembles (M2R, two univariate, and one naive) are used as prior ensembles for a waterflood EnKF inversion, and convergence of each ensemble to the reference model is evaluated. The comparison shows that M2R provides a larger uncertainty space to avoid filter divergence and improves inversion performance because of more careful sampling of the prior, geomodel space.

8.2 Ensemble-Based Closed-Loop Work Flow

"Closed-loop reservoir management" is focused on the development of concepts and algorithms to improve ultimate oil recovery though the use of measurement and control. It consists of two parts: geological model updating and production optimization. The main sources of inspiration are data assimilation techniques and optimization. Chen, Oliver, and Zhang (2008) combines the EnKF for continuous model updating with ensemble-based optimization method (EnOpt) to form a real-time closed-loop reservoir management. The sensitivities needed in the data assimilation and production optimization are approximated from the ensemble in a straightforward manner without the need for adjoint computation. Therefore the EnOpt can be extended based on our distributed EnKF work flow (Figure 8.1). Following the work of Chen, Oliver, and Zhang (2008), the ensemble-based closed-loop optimization using the EnOpt starts from Step 10.

Step 10a Initialize EnOpt step. When $\ell = 1$ (the first iteration), generate initial control variables $\boldsymbol{x}_1$ and initial ensemble of control variables $\boldsymbol{x}_{1,j},\ (j = 1, 2, \ldots, N_e)$. $\boldsymbol{x}$ represents the vector of control variables that contains all the well constraints at different control steps, $\boldsymbol{x} = [x_1, x_2, x_3, \ldots, x_{N_x}]$. $\boldsymbol{x}_{1,j},\ (j = 1, 2, \ldots, N_e)$ are realizations of control variables used to approximate $\boldsymbol{C}_x \boldsymbol{G}_\ell^T$, where $\boldsymbol{C}_x$ is the prior covariance of control variables $\boldsymbol{x}$, $\boldsymbol{G}_\ell$ is the sensitivity of objective function to the control variables evaluated at the ℓth iteration.

(i) If $k = 1$, $\boldsymbol{x}_{1,j}$ are generated in two steps. First, a mean control of all wells is sampled from a uniform distribution with the upper and lower limits equal to the maximum and minimum possible well constraints for each realization of each well. Second, a temporally correlated Gaussian random field (GRF) with zero mean is generated for each realization and added to the mean control. Set $\boldsymbol{x}_1 = 1/N_e \sum_{j=1}^{N_e} \boldsymbol{x}_{1,j}$.

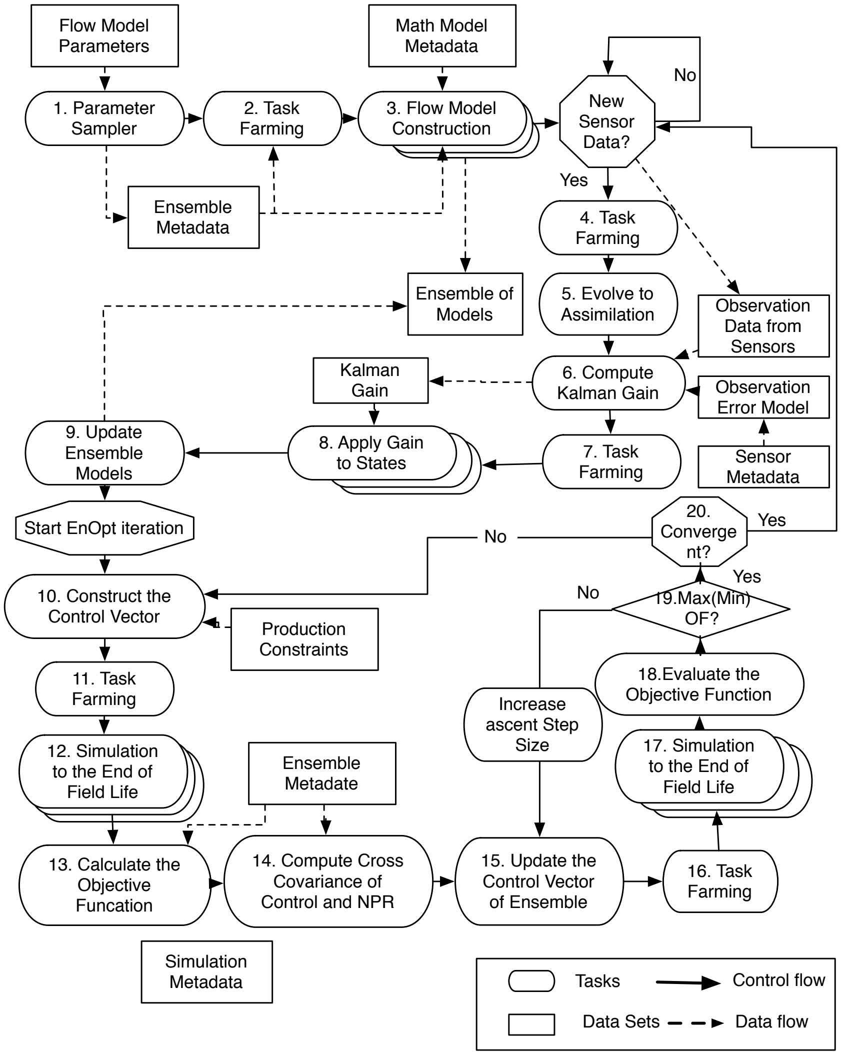


Figure 8.1: Work flow for closed-loop management on Grid.

(ii) if $k \neq 1$, $\boldsymbol{x}_1$ is set equal to $\boldsymbol{x}$. then a GRF is generated for each realization of each well and added to $\boldsymbol{x}$ to form $\boldsymbol{x}_{1,j}$.

Step 10b Start the EnOpt loop. If $\ell \neq 1$, a GRF is generated for each realization of each well and added to the control variable $\boldsymbol{x}_\ell$.

Step 11 & 12 Task farming and run the simulations from the current time step to the end of field life.

Step 13 Use the simulation results to calculate the optimized objective function, *e.g.*, Net present value for water flood problem.

$$g(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{N_t} \frac{v_o Q_{oi}(\boldsymbol{x}, \boldsymbol{y}) - v_w Q_{wi}(\boldsymbol{x}, \boldsymbol{y})}{(1 + r_\tau)^{t_i/\tau}} \tag{8.1}$$

where i is the time step; N_t is the total number of time steps; r_τ is the discount rate in terms of time span τ and t_i is the accumulative time since the start of production. v_o and v_w are the price of oil and the cost of water disposal, respectively. Q_{oi} and Q_{wi} are the total oil and water production over time step Δt_i. y is the reservoir properties.

Step 14 Compute the cross covariance $\boldsymbol{C}_{x,g(x)}$ using

$$\boldsymbol{C}_{\boldsymbol{x},g(\boldsymbol{x})} \approx \frac{1}{N_e - 1} \sum_{j=1}^{N_e} \left(\boldsymbol{x}_{\ell,j} - \bar{\boldsymbol{x}}_\ell\right) \left(g(\boldsymbol{x}_{\ell,j}, \boldsymbol{y}_j) - \bar{g}(\boldsymbol{x}_\ell, \boldsymbol{y})\right) \tag{8.2}$$

where $\bar{\boldsymbol{x}} = \frac{1}{N_e} \sum_{j=1}^{N_e} \boldsymbol{x}_{\ell,j}$ and $\bar{g}(\boldsymbol{x}_\ell, \boldsymbol{y}) = \frac{1}{N_e} \sum_{j=1}^{N_e} g(\boldsymbol{x}_{\ell,j}, \boldsymbol{y}_j)$

Step 15 Compute the updated control variables $\boldsymbol{x}_{\ell+1}$ using

$$\boldsymbol{x}_{\ell+1} = \frac{1}{\alpha} \boldsymbol{C}_{\boldsymbol{x}} \boldsymbol{C}_{\boldsymbol{x},g(\boldsymbol{x})} + \boldsymbol{x}_\ell \tag{8.3}$$

where α is the tuning parameter for ascent step size.

Step 16 & 17 Task farming and run the simulations with the new $\boldsymbol{x}_{\ell+1}$.

Step 18 Calculate the objective function with the simulation results.

Step 19 If $g(\boldsymbol{x}_{\ell+1}) > g(\boldsymbol{x}_\ell)$, overwrite $\boldsymbol{x}_\ell$ by $\boldsymbol{x}_{\ell+1}$ and $\ell = \ell + 1$; otherwise increase the step size α_ℓ and go to step 15.

Step 20 Check if the iteration tolerance and the increases of the tuning parameter α_ℓ is greater than two, then set $\boldsymbol{x} = \boldsymbol{x}_\ell$ and iteration stops. Otherwise, go back to Step 10b.

8.3 Grid Computing Issues

The utilization of Grid computing in industrial research projects has been limited because of issues, *i.e.*, application softwares, financial cost and data security. In the following sections, we discuss how to enhance reliability and usability of Grid computing environment to make a contribute to decreased computing costs via a future market in secure, distributed Grid computing.

8.3.1 Reservoir Simulation Capabilities

From our experience, one of the issues restricting widespread utilization of the ResGrid is the limited set of suitable simulators amenable to Grid environments. The current used open source simulator (UTCHEM) is easy to compile and suitable to different architect. However, the simulation is time consuming by using UTCHEM even solving a small problem because its sparse linear solver does not optimized to shorten the computation time. As we know, solving linear equations account for eighty percent of computation time for simulator. Therefore the future work is to involve massively parallel reservoir simulators into the distributed EnKF work flow. Cactus BlackOil is an implicit-pressure, explicit saturation simulator with two hydrocarbon components, a water component, and 3 phases; it exploits mature, widely used toolkits for utilities and parallelization (Cactus) and a linear solver (PETSc). Cactus also provides notification methods (*e.g.*, SMS to mobile phones), extensive I/O (*e.g.*, streaming HDF5), and mechanisms for parameter steering. Continued work on larger problems will focus on coupling the EnKF and using more processors, such as solving one million gridblocks model inversion problem; this work is aided by TeraGrid (30,000 SU) and LONI (50,000 SU) supercomputing allocations (White et al. 2008).

8.3.2 Resource Pricing and Run Strategy

Another issue is how the Grid computing can expand the end users' computation resources with the intention of minimizing financial expenditure. A common criticism of Grid computing is that, with the falling prices of commodity hardware, one could purchase servers and clusters to obtain an equivalent amount of processing power. As large corporations purchase machines that sit mainly idle, using computational Grid at minimal cost makes good business sense assuming the infrastructure complexity can be solved. Indeed, the corporate user greatly benefits in this scenario. Economical models help researchers manage and evaluate resource allocations to user communities. It supports Deadline and Budget Constrained (DBC; Buyya et al. 2002) scheduling algorithms for executing task-farming applications on large scale distributed systems depending on their cost, power, and availability and users quality of service requirements.

(i) Optimize for Time: This strategy produces results as early as possible, but before a deadline and within a budget limit.

(ii) Optimize for Cost: This strategy produces results by deadline, but reduces cost within a budget limit.

In these scheduling processes, the resource broker employed the economical market model for establishing a service access price. It used grid resource trading services for establishing connection with the Grid Trader (which is a trader sever decides access costs based on resource-owner defined charging algorithms/policies and interacts with accounting system for recording usage details and billing as per negotiation.) running on resource providers machines and obtained service prices accordingly. The broker architecture is generic enough to use any protocols used in real world market for negotiating access to resources and choosing appropriate ones. The access price varies from one consumer to another and from time to time, as defined by the resource owners. Depending on the deadline and the specified budget, the broker develops a plan for assigning jobs to resources. While doing so it does dynamic load profiling to learn the ability of resources to execute jobs.

8.3.3 Process Monitoring

One improvement of ResGrid should focus on troubleshooting and monitoring in a Grid environment, which can also be more complicated relative to a heterogeneous cluster or large server environment. Dealing with varying operating systems, different and frequently changing patch levels, and the impact of security updates potentially resulting in network misconfigurations can be difficult to diagnose and resolve for the Grid administrator. There are several aspects to monitoring that are important to consider. It is imperative to monitor the availability and status of the various compute nodes. Monitoring also includes the ability of the user to see the status of their job, view its progress and potentially even peek at the output as it is being generated. This is a capability that users are accustomed to the cluster environments, but it becomes much more difficult when the remote machine is an unknown entity, potentially running a different operating system and even cases where the user does not have login privileges. Grids can be notorious for having machines drop out or jobs never returning. Our suggestion is to use Message Passing Interface (MPI) or other software with enhanced error handling function can provide more extensive diagnostic information, such as the ability to distinguish calculation- related issues (*e.g.*, timeout, convergence problems) from infrastructure- related issues (*e.g.* machines getting turned off, disks filling up, machine owner putting the job to sleep). There are two primary approaches to this type of fault tolerance, one is to run multiple replicas of each process simultaneously and the other is to resubmit processes when they fail. They result in increased performance for the Grid. This, in turn, results in more compute capacity for the Grid users without any increased cost.

8.4 EnKF Application Suggestions

Chapter 7 and 6 address two problems essential to the mechanics of the EnKF. After the closer investigation and testing, some suggestions of the EnKF usage can be summarized.

Although there are a number of errors permeating the EnKF, such as naturally inaccuracies in the prior sample covariance, measurement error or other approximations, which can cause the filter divergence, it is not difficult to achieve the successful updating process if measures have been taken. If the prior sampling has the best spread of model parameter space and the measurement errors are correct, then "covariance inflation factor" is a simple and efficient remedy to overcome filter divergence although it is an empirical method. Optimal magnitude of the inflation is function of the ensemble size; but the 1% inflation factor is nearly optimal for different ensemble sizes.

The research explore into the distance-dependent filtering of covariance estimates generated by a finite ensemble by Schur product. Some more work is needed to understand why such filtering may be beneficial, how much improvement may be expected from filtering, how this may change with the size of the ensemble, how to choose the correct critical distance and how to involve the anisotropy. An understanding of the dual effects of localization and inflation is another interesting topic. Therefore, it is too earlier to provide some suggestions for localization method now.

The simple EnKF with truncation works as well as reparameterization. More cautious steps would be: the EnKF with reparameterization can be used for the initial assimilation steps when the nonlinear and non-Gaussian conditions are strong. The EnKF with truncation can also be used to continue the process.

8.5 Variance Trade-Offs in the EnKF

It is known that increasing the number of model parameters has involved more detailed geological models in model inversion. It has two effects in history matching: it makes the problem more rank-deficient because of increased unknowns and the variance in the model parameters, $\boldsymbol{m}$. However, it *invariably* decreases the model mismatch $\boldsymbol{d} - \boldsymbol{g}(\boldsymbol{m})$ in history match phase and more precisely matches the observation data.Thus, the modelers must confront a choice: (a) improve fit and decrease bias with a large parameter space, or (b) sacrifice fit but improve parameter estimation — and usually prediction reliability – by using a smaller parameter space. This is the classic bias-variance tradeoff (Hastie, Tibshirani, and Friedman 2003, p. 37).

To chose the most appropriate number of model parameters, a model selection method must distinguish between competing choices. The goal of model selection is to balance the quality of fit to observational data and model parameters against the complexity, or predictiveness, of the model achieving fit. The Bayesian Information Criterion (BIC) is one of the criteria based on Bayesian statistics (Liddle 2007; Burnham and Anderson 2005).

The BIC was introduced as the following equation,

$$I_{B,\mu} = -N_d \ln\left(\frac{S_{obs,err}}{N_d}\right) - N_m \ln\left(\frac{S_{m,err}}{N_m}\right) + N_m \ln(N_d) \tag{8.4}$$

where $S_{obs,err} = \sum_{i=1}^{N_d} \left(g(m_i) - d_{obs}\right)^2$, $S_{m,err} = \sum_{i=1}^{N_m}(m_i - \bar{m}_i)^2$, where $\mu \in \{1 \ldots N_e\}$ is an ensemble member, and all terms on the right of this equation refer to a particular member. As N_m increases, the $S_{obs,err}$ decreases because of improved fit, but the $S_{m,err}$ will increase by more parameters, therefore the $I_{B,\mu}$ does not increase significantly compared with less model parameter scenario. Especially if the observation data are limited, there is not sufficient flow information. The rank-deficient problem caused by increased degree of freedom may generate bias estimation for models with large N_m.

Chapter 9
Summary and Conclusions

In this dissertation, ResGrid and its extended version were developed with the help from researchers at CCT (Center for Computation and Technology) as the toolkits for designed reservoir simulation studies and real-time reservoir monitoring with measurement data from permanent down-hole sensors.

9.1 Summary

ResGrid was applied to simulate, analyze and discriminate the most frequently used geostatistical algorithms, *i.e.*, LUSIM and SGSIM. The suite of 1,920 simulations shows that the overestimate of permeability by SGSIM has significant effect on flow responses, *e.g.*, breakthrough time and upscaled permeability.

The distributed EnKF work flow is applied to “synthetic” 1-D and 2-D, two-phase, water flood examples. A prototype, one linux processor EnKF work flow is used for a “semisynthetic” field model, PUNQS3, a 3-D, three-phase case. The results show that the work flow can successfully track the measurement data and tune the porosity and permeability, and as more measurements are assimilated, the forecasts are improved. Moreover, many characteristics of EnKF are especially desirable for continuous model updating. It preserves the underlying geology, and estimates the uncertainty in rock properties and distribution of fluids, which is essential for updating development plans. The continuous updating ensures that the predictions and optimizations always start from a solution that matches the observed production data. However, although new measurements are added continuously, the error in the estimated porosity and permeability increases late in time. Sometimes, the estimated errors of rock properties increase because of the incorrect assimilation frequency. The reasons for these phenomena were investigated and discussed. The remedy methods, such as prior covariance inflation, Schur product and selection of suitable assimilation frequency are investigated. We have focused on another problem where EnKF encounters a difficulty, that is, the inability sample the posterior pdf correctly in non-Gaussian, multimodal problems. A reparameterization and simple truncation methods are applied to solve the problem. The estimated errors of permeability from the two methods are compared, and the simpler and more efficient truncation method is found to be adequate in the test case.

9.2 Conclusions

Specific conclusions from the research are summarized as following:

(i) The developed EnKF on Grid software is computationally robust and efficient. Because EnKF repeats the model propagations a large number of times, the software considerably alleviate the engineering tasks. It is a general and portable tool that can be widely applied because of its desirable attributes. For example, it is user-friendly, any reservoir engineers who have reservoir simulation experiences but few exposure of data assimilation and grid computing knowledge can use the toolkit without much training; easily couple with any open-source simulators.

(ii) One notable difference that appears when comparing the EnKF with other Kalman filter implementations is the way covariance is computed. The EnKF computes covariances directly from the ensemble when they are required, not evolving with time using functions. Therefore the sequential algorithm only uses fairly small matrices at any particular time, so that the memory requirement is fairly modest.

(iii) Although the EnKF has many advantages and flexible abilities compared with other traditional history matching methods, it provides incorrect characterization of model parameters when filter divergence occurs. This is a potential problem for EnKF and needs attention. Many possible reasons are interwoven and make the problem complicated.

- An empirical inflation factor to avoid collapse of the ensemble members around the ensemble mean efficiently diminishes one source of filter divergence. We found that the optimal magnitude of the inflation is a function of ensemble size; the smaller the size of the ensemble, the larger the inflation. Another issue with covariance inflation is that a single inflation factor may not be optimal over all parts of model all the time. This issue deserves more exploration.

- We demonstrate a Schur product of the ensemble covariance with a correlation function with local support can provide a notable benefit to EnKF process for reservoir simulation, as well as for the meteorology problems for which it has already been applied. In our study, the inversion is improved when the noise (the error) in a covariance estimate is smaller than the signal (the true magnitude of the covariance) for a small ensemble size. However, more research on the relationships between the correlation function with ensemble size, observation type, and observation density are required.

- The dual effects of Schur product and inflation on the covariance can be understood: the inflation increase the variance but Schur product tends to generate sparseness of variance. The two methods seem not conflict with each other, but the combined effects require further investigation. It is important to find the source of error and explore which methods are most suitable and effective to treat it, improving on the current, regrettably *ad hoc* approaches.

(iv) In general, new features of heterogeneity in the reservoir model can be revealed with reduced uncertainty by assimilating more production data, resulting in more accurate predictions. However, if the available production data are the same type of data and from the same wells for a long time, the value of production data is diminishing with time resulting in less updating at later times. The early observations give the largest contribution to the reduction in estimate error. In most cases, the transient flow at the early time observations provide more information than the observations from a steady state flow. It appears that the early assimilation frequency should be higher than at later times. We demonstrated that the need for small assimilation steps at any assimilation step when the dynamics of reservoir system have significant changes. In general, it appears that selection of assimilation intervals can be improved or automated by

- more carefully analyzing redundancy (or correlation) of observations,
- recognizing significant events in the observed data (such as breakthroughs), and
- monitoring the standard error of ensemble (σ_e) or the Ensemble gain ($\boldsymbol{K}_k$) to detect problems and take appropriate action.

(v) The Gaussian assumption of EnKF is critical. The EnKF using the full nonlinear forecast model and Monte Carlo sampling in the forecast step simplifies both the algorithm and makes it more robust. However, simplification causes the EnKF to misadjust the models if the dynamics of a flow system become sufficiently nonlinear and the distribution of model responses is non-Gaussian. A more general approach is required.

- Although the EnKF with reparameterization outperforms the traditional EnKF in time-dependent variable updating, it is not suitable for grid environment because of the large amount of time window data transfer load for network and difficult implementation in more than one dimension.
- Model parameter estimated by the simple, easy and efficient EnKF with truncation are similar to EnKF with reparameterization.
- The EnKF with reparameterization can be used selectively, only at significant flow changes at early measurement times. This would yield significant gains in efficiency.

(vi) It appears a relatively large ensemble size is required to span the space of model solutions and avoid the filter suffer the spurious correlation. Therefore, the variability of initial ensemble is critical for successful EnKF process because EnKF is a best linear unbiased estimation only if the neighborhood of the truth is adequately sampled. More studies are needed, such as multivariate secondary response (M2R) method, which can be used to choose essentially diverse yet manageably small prior ensembles to improve the efficiency of EnKF.

(vii) The loss of some ensemble members because of shut-in/reboot of some systems appears not to hurt the results much, especially after several steps of assimilation. After that, the modal parameters become stable.

9.3 Future Work

The future work should focus on two points. One is the development work on computation efficiency and functional versatility of the work flow. Another is to solve the large model inversion problem to integrate the EnKF and optimization together. More detailed are as following,

(i) To build up the distributed EnKF work flow by using Cactus BlackOil as simulator to improve computational efficiency based on its robust parallel functionality and advanced linear solver.

(ii) Continued work on larger problems should focus on coupling the EnKF and using more processors, such as solving one million gridblocks model continuous updating problem; this work may use TeraGrid (30,000 SU) and LONI (50,000 SU) supercomputing allocations (White et al. 2008).

(iii) To keep the work flow for the continuously updated geological model remains unchanged. To develop the ensemble-based closed-loop optimization using the EnOpt (or other ensemble-based optimization methods) to extend the workflow to closed-loop reservoir management work flow.

Bibliography

Addiego-Guevara, E. A., M. D. Jackson, and M. A. Giddins. 2008, 20-23 April. "Insurance Value of Intelligent Well Technology Against Reservoir Uncertainty." *SPE/DOE Symposium on Improved Oil Recovery.* Tulsa, Oklahoma.

Allen, Gabrielle, Kelly Davis, T. Goodale, A. Hutanu, H. Kaiser, T. Kielmann, A. Merzky, R. Van Nieuwpoort, A. Reinefeld, F. Schintke, T. Schuett, E. Seidel, and B. Ullmer. 2005. "The GridLab Grid Application Toolkit: Towards Generic and Easy Application Programming Interfaces for the Grid." *Proceedings of the IEEE* 93 (March): 534–550. No.3.

Anderson, J. L. 2007. "Exploring the Need for Localization in Ensemble Data Assimilation Using a Hierarchical Ensemble Filter." *Physica D* 230:99–111.

Anterion, F., B. Karcher, and R. Eymard. 1989. "Use of Parameter Gradients for Reservoir History Matching." *SPE paper 18433 presented at the SPE Symposima on Reservoir Simulation, Houston, TX.*

Aziz, K., and A. Settari. 1979. *Petroleum Reservoir Simulation.* Elsevier Applied Science Publishers.

Barker, J. W., M. Cuypers, and L. Holden. 2001. "Quantifying Uncertainty in Production Forecasts: Another Look at the PUNQ-S3 Problem." *SPEJ* 6 (4): 433–441.

Bear, Jacob. 1988. *Dynamics of Fluids in Porous Media.* Dover Publications, Inc. New York.

Bissel, R. 1994. "Calculating Optimal Parameters for History Matching." *4th European Conference on the Mathematics of Oil Recovery, Roros, Norway.*

Brouwer, D. R., J. D. Jansen, S. van der Starre, C. P. J. W. van Kruijsdijk, and C. W. J. Berentsen. 2001, 21-22 May. "Recovery Increase through water flooding with smart well technology." *SPE European formation damage conference.* Hague, Netherlands.

Brouwer, D. W., G. Navdal, J. D. Jansen, E. H. Vefring, and J. W. van Kruijsdijk. 2004, 26-29 September. "Improved Reservoir Management Through Optimal Control and Continuous Model Updating." *SPE Annual Technical Conference and Exhibition.* SPE 90149.

Brown, George A., and Arthur Hartog. 2002. "Optical Fiber Sensors in Upstream Oil & Gas." *Journal of Petroleum Technology* 54 (11): 63–65.

Burnham, Kenneth P., and David R. Anderson. 2005. *Model Selection and Multimodel Inference A Practical Information-Theoretic Approach.* Springer.

Buyya, Rajkumar, David Abramson, Jonathan Giddy, and Heinz Stockinger. 2002. “Economic Models for Resource Management and Scheduling in Grid Computing.” *Concurrency Comutat.: Pract. Exper.* 14:1507–1542.

Calhoun, T. G., and R.M. Tittle. 1968, September. “Use of Radioactive Isotopes in Gas Injection.” *SPE 43rd Annual Fall Meeing.* Houston, Texas. SPE 2277.

Carter, R. D., L. F. Kemp, Jr., A. C. Pierce, and D. L. Williams. 1974. “Performance Matching With Constraints.” *SPE Journal* 14:187–196.

Chen, W. H., G. R. Gavalas, John H. Seinfeld, and Mel L. Wasserman. 1974. “A New Algorithm for Automatic History Matching.” *SPE Journal*, pp. 593–608.

Chen, Yan, D. S. Oliver, and Dongxiao Zhang. 2007. “Data Assimilation for Nonlinear Problems by Ensemble Kalman Filter with Reparameterization.” *Elsevier Science*, March.

Chen, Yan, Dean S. Oliver, and Dongxiao Zhang. 2008. “Efficient Ensemble-Based Closed-Loop Production Optimization.” *SPE 112873 presented at 2008 SPE/DOE Improved Oil Recovery Symposium in Tulsa, Oklahoma.*

Dong, Yannong, and D. S. Oliver. 2005. “Quantitative Use of 4D Seismic Data for Reservoir Description.” *SPE Journal* 10(1):91–99.

ECLIPSE 100. 2008, July. `http://www.slb.com`. `URL` last accessed. Schlumberger Black-Oil Simulator.

Evensen, G. 1994. “Sequential Data Assimilation With a Nonlinear Quasi-Geostrophic Model Using Monte Carlo Methods to Forecast Error Statistics.” *J. Geophys. Res.* 99 (C5): 10143–10162.

———. 2003. “The Ensemble Kalman Filter: Theoretical Formulation and Practical Implementation.” *Ocean Dyn.* 53 (4): 334–367.

———. 2004. “Sampling Strategies and Square Root Analysis Schemes for the EnKF.” *Ocean Dynamics* 54:539–560.

Evensen, G., J. Hove, H.C. Meisingset, E. Reiso, K. S. Seim, and O. Espelid. 2007, 26-28 February. “Using the EnKF for Assisted History Matching of a North Sea Reservoir Model.” *2007 SPE Reservoir Simulation Symposium.* Texas, TX. SPE 106184.

Evensen, Geir. 2007. *Data Assimilation: The Ensemble Kalman Filter.* Edited by Verlag. Springer Science Business Media.

Fjerstad, P. A., A. S. Sikandar, H. Cao, J. Liu, and W. Da Sie. 2005, December 5-6. “Next Generation Parallel Computing for Large-Scale Reservoir Simulation.” *SPE International Improved Oil Recovery Conference in Asia Pacific.* Kuala Lumpur, Malaysia. SPE 97358.

Floris, F. J. T., M. D. Bush, M. Cuypers, and A. R. Syversveen. 2001. “Methods for Quantifying the Uncertainty of Production Forecasts: A Comparative Study.” *Petroleum Geoscience* 7 (Supp): 87–96.

Foster, Ian, and Carl Kesselman. 1998. “A Security Architecture for Computational Grids.” *5th ACM Conference on Computer and Communications Security.* San Francisco, CA.

———. 1999. *Globus: A Toolkit-Based Grid Architecture.* San Francisco, CA: Morgan Kaufmann Publishers.

Frey, James, Todd Tannenbaum, Miron Livny, Ian Foster, and Steven Tuecke. 2002. “Condor-G: A Computation Management Agent for Multi-Institutional Grids.” *Cluster Computing* 5 (3): 237–246 (July).

Gao, Guohua, Mohammad Zafari, and Albert C. Reynolds. 2006. “Quantifying Uncertainty for the PUNQ-S3 Problem in a Bayesian Setting with RML and EnKF.” *SPE Journal*, December, 506–515.

Gaspari, Gregory, and Stephen E. Cohn. 1999. “Construction of Correlation Functions in Two and Three Dimensions.” *Q. J. R. Meteorol. Soc.* 125:723–757.

Gelb, A. 1979. *Applied Optimal Estimation.* Cambridge, Massachusetts, and London, England.

Globus. 2008, July. `http://www.globus.org/`. `URL` last accessed. Globus Toolkit.

Goodale1, Tom, Shantenu Jha, Hartmut Kaiser, Thilo Kielmann, Pascal Kleijer, Gregor von Laszewski, Craig Lee, Andre Merzky, Hrabri Rajic, and John Shalf. 2006. “SAGA: A Simple API for Grid Applications. High-level Application Programming on the Grid.” *Computational Methods in Science and Technology* 12 (1): 7–20.

Goovaerts, Pierre. 1997. *Geostatistics for Natural Resources Evaluation.* Applied Geostatistics Series. Oxford: Oxford University Press.

GSLIB: Geostatistical Software Library. 2008, July. `http://www.gslib.com`. `URL` last accessed.

Gu, Yaqing. 2006. “History Matching Production Data Using the Ensemble Kalman Filter.” Ph.D. diss., Newbourne School of Petroleum and Geological Engineering, The University of Oklahoma.

Gu, Yaqing, and Dean S. Oliver. 2005. “History Match of the PUNQ-S3 Reservoir Model Using the Ensemble Kalman Filter.” *SPE Journal*, June, 217–224.

———. 2006. “The Ensemble Kalman Filter for Continuous Updating of Reservoir Simulation Models.” *J. of Energy Resources Technology* 128 (March): 79–87.

Haddad, Sammy, Eduardo Proano, and Yogesh Patel. 2004, 26-29 September. “A Method to Diagnose Depletion, Skin, kh, and Drive Mechanism Effects Using Reservoir Monitoring

Data." *SPE Annual Technical Conference and Exhibition.* Houston, Texas.

Hamill, Thomas M., and Jeffrey S. Whitaker. 2001. "Distance-Dependent Filtering of Background Error Covariance Estimates in an Ensemble Kalman Filter." *Mon. Weather Rev.* 129 (3): 2776–2790.

Hastie, T., R. Tibshirani, and J. H. Friedman. 2003. *The Elements of Statistical Learning.* Springer.

He, N., A. C. Reynolds, and D. S. Oliver. 1996, October. "Three-dimensional Reservoir Description from Multiwell Pressure Data and Prior Information." *SPE Annual Technical Conference and Exhibition.* Denver, CO. SPE 36509.

Houtekamer, P. L., and H. L. Mitchell. 1998. "Data Assimilation Using an Ensemble Kalman Filter Technique." *Mon. Weather Rev.* 126 (3): 796–811.

Hubbert, M. K. 1969. *The Theory of Groundwater Motion and Related Papers.* New York: Hafner Publishing Company.

IBM Redbooks. 2008. `http://www.redbooks.ibm.com/`. `URL` last accessed.

Jackson, D. D. 1979. "The Use of a Priori Data to Resolve Non-unique in Linear Inversion." *Geophys. J.R. Astr. Soc.* 57:137–157.

Jacquard, P., and C. Jain. 1965. "Permeability Distribution from Field Pressure Data." *SPE Journal* 5(4):281–294.

Jahns, Hans O. 1966. "A Rapid Method for Obtaining a Two-Dimensional Reservoir Description from Well Pressure Response Data." *SPEJ* 6(12):315–327.

Jansen, J. D., D. R. Brouwer, G. Naevdal, and J. W. van Kruijsdiik. 2005. "Closed-Loop Reservoir Management." *First Break* 23:43–48.

Jazwinski, Andrew H. 1970. *Stochastic Process and Filtering Theory.* Volume 64 of *Mathematics in Science and Engineering.* New York: Academic Press.

Kalla, Subhash, and Christopher D. White. 2007. "Efficient Design of Reservoir Simulation Studies for Development and Optimization." *SPE Reservoir Evaluation & Engineering* 10 (6): 629–637 (December).

Kalman, R. E. 1960. "A New Approach to Linear Filter and Prediction Problems." *J. Basic. Eng.* 82:35–45.

Lake, L. W. 1989. *Enhanced oil recovery.* Old Tappan, NJ; Prentice Hall Inc.

Landa, J. L., and R. N. Horne. 1997, October. "A Procedure to Integrate Well Test Data, Reservoir Performance History and 4-D Seismic Information into a Reservoir Description." *SPE Annual Technical Conference and Exhibition.* San Antonio, TX. SPE 38653.

Lee, T. Y., and J. H. Seinfeld. 1987. “Estimation of Two-phase Petroleum Reservoir Properties by Regularization.” *Journal of Computational Physicas* 69:397–419.

Lei, Z., D. Huang, A. Kulshrestha, S. Pena, G. Allen, X. Li, R. Duff, S. Kalla, C. White, and J. Smith. 2006, 16-19 May. “ResGrid: A Grid-aware Toolkit For Reservoir Uncertainty Analysis.” *IEEE International Symposium on Cluster Computing and the Grid (CCGrid06).* Singapore.

Lei, Zhou, and Gabrielle Allen. 2007. “A Novel Mechanism for Application Execution in Multicluster Grids.” *International Conference on Parallel Processing.*

Li, Gaoming, and A. C. Reynolds. 2007, 11-14 November. “An Iterative Ensemble Kalman Filter for Data Assimilation.” *SPE Annual Technical Conference and Exhibition.* SPE 109808.

Li, Hongmei, and Christopher D. White. 2003. “Geostatistical Models for Shales in Distributary Channel Point Bars.” *AAPG Bulletin* 87 (December): 1851–1868. No.12.

Li, R., A. C. Reynolds, and D. S. Oliver. 2003. “Sensitivity Coefficient for Three-Phase History Matching.” *Journal of Canadian Petroleum Technology*, pp. 70–77.

Liddle, Andrew R. 2007. “Information criteria for astrophysical model selection.” *Monthly Notices of the Royal Astronomical Society: Letters* 377 (1): L74–L78.

Liu, Ning, Soraya Betancourt, and Dean S. Oliver. 2001. “Assessment of Uncertainty Assessment Method.” *SPE Annual Technical Conference and Exhibition.*

Liu, Ning, and Dean S. Oliver. 2005. “Critical Evaluation of the Ensemble Kalman Filter on History Matching of Geologic Facies.” *SPE Reservoir Evaluation and Engineering* 8(6):470–477.

Lorentzen, R., K. K. Fjelde, J. Froyen, and G. Navdal. 2001, October. “Underbalanced and Low-Head Drilling Operation: Real Time Interpretation of Measured Data and Operational Support.” *SPE Annual Technical Conference and Exhibition.* Denver, CO. SPE 71384.

Lorentzen, R. J., A. M. Berg, G. Navdal, and E. H. Vefreing. 2006. “A New Approach for Dynamic Optimization of Waterflooding Problems.” *SPE Intelligent Energy Conference and Exhibition.*

Manin, Y., M. Charara, and J. P. Delhomme. 2002, 29-31 October. “Characterization of Reservoir Fluid Dynamics from Multiple Downhole Permanent Sensors.” *European Petroleum Conference.* Aberdeen, United Kingdom.

Margulis, S. A., D. B. McLaughlin, D. Entekhabi, and S. Dunne. 2002. “Land Data Assimilation and Estimation of Soil Moisture Using Measurements from the Southern Great Plains 1997 Field Experiment.” *Water Resources Research* 38 (12): 103–114.

Marschall, Roland, and Donald Sherlock. 2002, 13-17 April. “Some Aspects of 4-D Seismics for Reservoir Monitoring.” *SPE/DOE Improved Oil Recovery Symposium.* Tulsa, Oklahoma.

Maybeck, Peter S. 1979. *Stochastic Models, Estimation, and Control.* Volume 1 of *Mathematics in Science and Engineering.* New York: Academic Press, Inc.

Miller, R. N., E. F. Carter, and S. T. Blue. 1999. “Data Assimilation into Nonlinear Stochastic Models.” *Dynamic Meteorology and Oceanography* 51(2):167–194.

Navdal, G., L. M. Johnsen, S. I. Aanonsen, and E. H. Vefring. 2005. “Reservoir Monitoring and Continuous Model Updating Using Ensemble Kalman Filter.” *SPE Journal* 10:66–74.

Navdal, G., T. Mannseth, and E. H. Vefring. 2002. “Near-Well Reservoir Monitoring through Ensemble Kalman Filter.” *2002 SPE/DOE Improved Oil Recovery Symposium.* SPE-75235.

Oglesby, K. D., P. Mehdizadeh, and G. J. Rodgers. 2006, 24-27 September. “Portable Multiphase Production Tester for High-Water-Cut Wells.” *SPE Annual Technical Conference and Exhibition.* San Antonio, Texas.

Oliver, D. S., B. Cunha Luciane, and Albert C. Reynolds. 1997. “Markov Chain Monte Carlo Methods for Conditioning a Permeability Field to Pressure Data.” *Math. Geology* 29(1):61–91.

Oliver, Dean S. 1994. “Incorporation of Transient Pressure Data into Reservoir Characterization.” *In Situ* 18(3):243–275.

Oliver, Dean S., Nanqun He, and Albert C. Reynolds. 1996. “Conditioning Permeability Fields to Pressure Data.” *European Conference for the Mathematics of Oil Recovery.*

PUNQ-S3 Model for Quantifying Uncertainty in Production Forecast. 2008, Aug. `http://www3.imperial.ac.uk/earthscienceandengineering/research/perm/punq-s3model`. `URL` last accessed.

Reichle, R. H., D. B. McLaughlin, and D. Entekhabi. 2002. “Hydrologic Data Assimilation With the Ensemble Kalman Filter.” *Mon. Weather Rev.* 130 (1): 103–114.

SAGA. 2008, July. `http://saga.cct.lsu.edu/`. `URL` last accessed. A Simple API for Grid Applications.

Sarma, P., L. J. Durlofsky, K. Aziz, and W. H. Chen. 2005. “Efficient Real-time Reservoir Management using Adjoint-base Optimal Control and Model Updating.” *paper SPE 94241 in the proceedings of the SPE Europec/EAGE Annual Conference, Madrid, Spain.*

Skjervheim, J.-A., G. Evensen, S. Aanonsenan, and T. Johansen. 2005. “Incorporating 4d Seismic Data in Reservoir Simulation Models Using Ensemble Kalman Filter.” *2005*

SPE Annual Technical Conference and Exhibition. SPE-95789.

Tarantola, Albert. 1997. *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation.* Amesterdam, Netherland: Elsevier.

Tarantola, Albert, and B. Valette. 1982. "Inverse Problems = Quest for Information." *J. Geophys* 50:159–170.

UTCHEM. 2008, July. `http://www.cpge.utexas.edu/utchem/source`. `URL` last accessed. UT Austin Chemical Compositional Simulator.

Wang, Chunhong, Gaoming Li, and Albert C. Reynolds. 2007. "Production Optimization in Closed-Loop Reservoir Management." *SPE 109805 presented at the 2007 SPE ATCE in Anaheim, California.*

Wasserman, M. L., A. S. Emanuel, and J. H. Seinfeld. 1974. "Practical Applications of Optimal Control Theory to History-Matching Multiphase Simulator Models." *SPE 5020 presented at SPE ATCE, Houston, TX.*

Watson, A. T., J. H. Seinfeld, G. R. Gavalas, and P. T. Woo. 1979. "History Matching in Two-Phase Petroleum Reservoirs." *SPE 8250 presented at the SPE Annual Technical Conference and Exhibition, Las Vegas, NV.*

Wen, X.-H., and W. H. Chen. 2005a. "Real-time Reservoir Model Updating Using Ensemble Kalman Filter." *2005 SPE Reservoir Simulation Symposium.* SPE-92991.

———. 2005b. "Some Practical Issues on Real-time Reservoir Model Updating using Ensemble Kalman Filter." *2005 International Petorleum Technology Conference.* IPTC-11024.

White, Christopher D. 2007. "PETE 7280 Mathematical Simulation of Petroleum Reservoir Performance." Informal Lecture Notes.

White, Christopher D., Gabrielle Allen, Tevfik Kosar, Juan Lorenzo, Mayank Tyagi, and Zhou Lei. 2008. "LSU UCoMS Project Annual Report." Technical Report, Louisiana State University.

White, Christopher D., and S. A. Royer. 2003, February. "Experimental Design as a Framework for Reservoir Studies." *SPE Reservoir Simulation Symposium.* Houston, Texas. SPE 79676.

White, Christopher D., Brian J. Willis, Keshav Narayanan, and Shirley P. Dutton. 2001. "Identifying and Estimating Significant Geologic Parameters with Experimental Design." *SPE Journal*, September.

Willis, Brian J., and Christopher D. White. 2000. "Quantitative Outcrop Data for Flow Simulation." *Journal of Sedimentary Research* 70:788–802.

Wu, Zan, and A. Datta-Gupta. 2002. "Rapid History Matching Using a Generalized Travel Time Inversion Method." *SPE Journal* 7:113–122.

Wu, Zan, Albert C. Reynolds, and Dean S. Oliver. 1999. "Conditioning Geostatistical Models to Two-Phase Production Data." *SPE Journal* 7:113–122.

Xiu, D., and G. E. Karniadaskis. 2003. "Modeling Uncertainty in Flow Simulations via Generalized Polynomial Chaos." *Journal of Computational Physics* 187:137–167.

Zafari, M.and Albert C. Reynolds. 2007. "Assessing the Uncertainty in Reservoir Description and Performance Predictions with the Ensemble Kalman Filter." *SPE 95750*, September, 382–387.

Zhang, D., Z. Lu, and Y. Chen. 2007. "Dynamic Reservoir Data Assimilation with an Efficient, Dimension-Reduced Kalman Filter." *SPE Journal* 12(1):108–117.

Zhang, F., and A. C. Reynolds. 2002. "Optimization Algorithms for Automatic History." *8th European Conference on the Mathematics of Oil Recovery.*

Zhang, F., Albert C. Reynolds, and Dean S. Oliver. 2002. "Evaluation of the Reduction in Uncertainty Obtained by Conditioning a 3D Stochastic Channel to Multi-well Pressure Data." *Mathematical Geology* 34:715–742.

Appendix A. Geostatistical Studies Tables

Table A-1: Parametric *t*-test and *F*-test of LUSIM and SGSIM model differences of upscaled permeability in maximum continuity direction (* significant at 95% level of confidence)

combination	*t*-test	relative Mean	*F*-test	std. ratio
0-0-1-0	0.000*	-16	0.000*	61
0-0-2-1	0.000*	-13	0.003*	45
0-1-0-3	0.000*	-22	0.003*	42
0-1-2-0	0.000*	-18	0.004*	38
0-2-1-2	0.000*	-12	0.064*	8
0-2-3-3	0.002*	-11	0.407	2
0-3-0-2	0.000*	-16	0.036*	12
0-3-2-1	0.001*	-13	0.224	4
1-0-2-2	0.002*	-12	0.034*	12
1-0-3-3	0.000*	-12	0.627	2
1-1-1-2	0.012*	-9	0.015*	18
1-1-1-3	0.016*	-9	0.733	1
1-2-0-1	0.002*	-8	0.063	8
1-2-2-0	0.009*	-14	0.069	8

Table A-2: Parametric *t*-test and *F*-test of LUSIM and SGSIM model differences of sweep efficiency in maximum continuity direction

combination	*t*-test	relative Mean	*F*-test	std. ratio
0-0-1-0	0.000*	8	0.018*	17
0-0-2-1	0.000*	7	0.005*	33
0-1-0-3	0.000*	9	0.051	9
0-1-2-0	0.000*	7	0.109	6
0-2-1-2	0.030*	4	0.006*	29
0-2-3-3	0.139	2	0.296	0
0-3-0-2	0.000*	6	0.984	1
0-3-2-1	0.002*	6	0.172	5
1-0-2-2	0.007*	4	0.329	3
1-0-3-3	0.051	4	0.873	1
1-1-1-2	0.068	3	0.014*	20
1-1-1-3	0.421	0	0.45	0
1-2-0-1	0.003*	4	0.246	4
1-2-2-0	0.193	2	0.72	1
1-3-0-0	0.014*	6	0.028*	13
1-3-3-1	0.496	0	0.723	1

Table A-3: Parametric *t*-test and *F*-test of LUSIM and SGSIM model differences of breakthrough time in maximum continuity direction.

combination	*t*-test	relative Mean	*F*-test	std. ratio
0-0-1-0	0.000*	32	0.024*	14
0-0-2-1	0.000*	27	0.113	6
0-1-0-3	0.000*	37	0.598	2
0-1-2-0	0.000*	29	0.487	2
0-2-1-2	0.403	7	0.8	1
0-2-3-3	0.134	8	0.386	0
0-3-0-2	0.001*	32	0.231	0
0-3-2-1	0.000*	27	0.495	0
1-0-2-2	0.006*	19	0.409	2
1-0-3-3	0.011*	17	0.652	1
1-1-1-2	0.054	16	0.004*	37
1-1-1-3	0.171	13	0.358	0
1-2-0-1	0.003*	29	0.077	8
1-2-2-0	0.392	13	0.546	2
1-3-0-0	0.048*	29	0.152	5
1-3-3-1	-0.591	6	0.672	2

Table A-4: Parametric *t*-test and *F*-test of LUSIM and SGSIM model differences of upscaled permeability in minimum continuity direction.

combination	*t*-test	relative Mean	*F*-test	std. ratio
0-0-1-0	0.007*	-10	0.000*	121
0-0-2-1	0.000*	-9	0.022*	15
0-1-0-3	0.004*	-17	0.000*	200
0-1-2-0	0.001*	-15	0.004*	36
0-2-1-2	0.001*	-12	0.168	5
0-2-3-3	0.001*	-10	0.777	1
0-3-0-2	0.001*	-14	0.084	7
0-3-2-1	0.001*	-11	0.127	0
1-0-2-2	0.007*	-10	0.040*	11
1-0-3-3	0.000*	-10	0.697	2
1-1-1-2	0.060	-7	0.032*	12
1-1-1-3	0.000*	-17	0.960	1
1-2-0-1	0.171*	-5	0.386	3
1-2-2-0	0.032*	-11	0.010*	23
1-3-0-0	0.042*	-11	0.465	2
1-3-3-1	0.631	-2	0.672	2

Table A-5: Parametric *t*-test and *F*-test of LUSIM and SGSIM model differences of sweep efficiency in minimum continuity direction

combination	*t*-test	relative Mean	*F*-test	std. ratio
0-0-1-0	0.000*	5	0.003*	2
0-0-2-1	0.010*	5	0.024*	3
0-1-0-3	0.006*	5	0.048*	5
0-1-2-0	0.004*	5	0.141	1
0-2-1-2	0.006*	5	0.094	2
0-2-3-3	0.315	2	0.080	0
0-3-0-2	0.004*	6	0.015*	6
0-3-2-1	0.008*	5	0.505	1
1-0-2-2	0.011*	3	0.130	2
1-0-3-3	0.015*	2	0.217	4
1-1-1-2	0.238	3	0.422	1
1-1-1-3	0.138	3	0.363	1
1-2-0-1	0.044*	2	0.939	1
1-2-2-0	0.434	1	0.228	1
1-3-0-0	0.022*	6	0.177	2
1-3-3-1	0.109	-4	0.062	0

Table A-6: Parametric *t*-test and *F*-test of LUSIM and SGSIM model differences of breakthrough time in minimum continuity direction.

combination	*t*-test	relative Mean	*F*-test	std. ratio
0-0-1-0	0.000*	15	0.000*	41
0-0-2-1	0.002*	17	0.002*	15
0-1-0-3	0.010*	16	0.010*	10
0-1-2-0	0.002*	22	0.002*	5
0-2-1-2	0.007*	20	0.007*	7
0-2-3-3	0.410	5	0.410	0
0-3-0-2	0.003*	21	0.003*	18
0-3-2-1	0.020*	19	0.020*	0
1-0-2-2	0.000*	15	0.000*	5
1-0-3-3	0.085	9	0.085	4
1-1-1-2	0.188	12	0.188	2
1-1-1-3	0.051	17	0.051	0
1-2-0-1	0.115	7	0.115	1
1-2-2-0	0.702	-2	0.702	4
1-3-0-0	0.031*	18	0.031*	4
1-3-3-1	0.852	-2	0.852	0

Table A-7: Parametric t-test and F-test of LUSIM and HYBRID model differences in maximum continuity direction

	upscaled permeability		sweep efficiency		breakthrough time	
combination	t-test	F-test	t-test	F-test	t-test	F-test
0-0-1-0	0.0010*	0.0003*	0.0133	0.0001*	0.0019*	0.1387
0-0-2-1	0.0003*	0.0025*	0.0281	0.0100*	0.0130	0.1925
0-1-0-3	0.0007*	0.0001*	0.0143	0.0407	0.0002*	0.4793
0-1-2-0	0.0001*	0.0029*	0.0063*	0.9461	0.0014*	0.8537
0-2-1-2	0.0005*	0.0355	0.0005*	0.2733	0.0018*	0.1859
0-2-3-3	0.0003*	0.2543	0.0540	0.2737	0.2700	0.5445
0-3-0-2	0.0005*	0.0254	0.0018*	0.0119	0.0012*	0.0032
0-3-2-1	0.0000*	0.0705	0.0515	0.1384	0.0021*	0.9387
1-0-2-2	0.0074*	0.0086*	0.0019*	0.0219	0.0038*	0.6731
1-0-3-3	0.0000*	0.8818	0.0004*	0.0394	0.1244	0.3258
1-1-1-2	0.0044*	0.0155	0.0138	0.2268	0.2005	0.8562
1-1-1-3	0.0001*	0.4783	0.0103	0.4426	0.0010*	0.2609
1-2-0-1	0.0150	0.2307	0.0003*	0.1757	0.4073	0.1973
1-2-2-0	0.0061*	0.0048*	0.0064*	0.0534	0.1753	0.6342
1-3-0-0	0.0154	0.3304	0.0159	0.1531	0.1291	0.9645
1-3-3-1	0.0777	0.3726	0.0039*	0.8771	0.3899	0.0797

Table A-8: Parametric t-test and F-test of LUSIM and HYBRID model differences in minimum continuity direction.

	upscaled permeability		sweep efficiency		breakthrough time	
combination	t-test	F-test	t-test	F-test	t-test	F-test
0-0-1-0	0.0010*	0.0003*	0.0133	0.0001*	0.0019*	0.1387
0-0-2-1	0.0003*	0.0025*	0.0281	0.0100*	0.0130	0.1925
0-1-0-3	0.0007*	0.0001*	0.0143	0.0407	0.0002*	0.4793
0-1-2-0	0.0001*	0.0029*	0.0063*	0.9461	0.0014*	0.8537
0-2-1-2	0.0005*	0.0355	0.0005*	0.2733	0.0018*	0.1859
0-2-3-3	0.0003*	0.2543	0.0540	0.2737	0.2700	0.5445
0-3-0-2	0.0005*	0.0254	0.0018*	0.0119	0.0012*	0.0032
0-3-2-1	0.0000*	0.0705	0.0515	0.1384	0.0021*	0.9387
1-0-2-2	0.0074*	0.0086*	0.0019*	0.0219	0.0038*	0.6731
1-0-3-3	0.0000*	0.8818	0.0004*	0.0394	0.1244	0.3258
1-1-1-2	0.0044*	0.0155	0.0138	0.2268	0.2005	0.8562
1-1-1-3	0.0001*	0.4783	0.0103	0.4426	0.0010*	0.2609
1-2-0-1	0.0150	0.2307	0.0003*	0.1757	0.4073	0.1973
1-2-2-0	0.0061*	0.0048*	0.0064*	0.0534	0.1753	0.6342
1-3-0-0	0.0154	0.3304	0.0159	0.1531	0.1291	0.9645
1-3-3-1	0.0777	0.3726	0.0039*	0.8771	0.3899	0.0797

Table A-9: Parametric t-test and F-test of LUSIM and SPECSIM model differences in maximum continuity direction.

	upscaled permeability		sweep efficiency		breakthrough time	
combination	t-test	F-test	t-test	F-test	t-test	F-test
0-0-1-0	0.0023*	0.0008*	0.0000*	0.0143	0.0007*	0.0081
0-0-2-1	0.0000*	0.0335	0.0005*	0.0044*	0.0000*	0.3224
0-1-0-3	0.0000*	0.0009*	0.0000*	0.0746	0.0009*	0.2803
0-1-2-0	0.0005*	0.0007*	0.0006*	0.0939	0.0011*	0.5525
0-2-1-2	0.0018*	0.0396	0.0037*	0.0064*	0.2449	0.6276
0-2-3-3	0.0025*	0.6590	0.2326	0.4607	0.4293	0.6590
0-3-0-2	0.0003*	0.0106	0.0001*	0.8025	0.0014*	0.1531
0-3-2-1	0.0019*	0.1023	0.0030*	0.1460	0.0026*	0.7430
1-0-2-2	0.0020*	0.0155	0.0013*	0.3186	0.0020*	0.3379
1-0-3-3	0.0033*	0.1582	0.0070*	0.8253	0.0041*	0.9983
1-1-1-2	0.1224	0.0067*	0.1013	0.0154	0.1487	0.0027
1-1-1-3	0.1099	0.3188	0.4060	0.4551	0.4844	0.3711
1-2-0-1	0.0032*	0.0325	0.0078*	0.1883	0.0026*	0.0547
1-2-2-0	0.0089*	0.1421	0.2701	0.6669	0.3980	0.4017
1-3-0-0	0.0042*	0.0825	0.0023*	0.0231	0.1459	0.0984
1-3-3-1	0.4942	0.7166	0.4027	0.7310	0.2968	0.5504

Table A-10: Parametric t-test and F-test of LUSIM and SPECSIM model differences in minimum continuity direction.

	upscaled permeability		sweep efficiency		breakthrough time	
combination	t-test	F-test	t-test	F-test	t-test	F-test
0-0-1-0	0.0014*	0.2837	0.0169	0.5412	0.3103	0.9786
0-0-2-1	0.2274	0.5754	0.0143	0.3343	0.2674	0.8123
0-1-0-3	0.0000*	0.2403	0.0099	0.2750	0.1811	0.8819
0-1-2-0	0.0561	0.4792	0.0282	0.9653	0.2596	0.8254
0-2-1-2	0.0849	0.3728	0.1065	0.3583	0.0081	0.6232
0-2-3-3	0.1515	0.6101	0.0137	0.3682	0.1388	0.9392
0-3-0-2	0.1826	0.7671	0.0270	0.4652	0.3338	0.7714
0-3-2-1	0.1824	0.6953	0.1013	0.7202	0.3455	0.9128
1-0-2-2	0.0085*	0.6857	0.0034	0.2145	0.1942	0.9821
1-0-3-3	0.0014*	0.8154	0.0013	0.0657	0.0886	0.9504
1-1-1-2	0.1626	0.7435	0.0429	0.6214	0.2511	0.9799
1-1-1-3	0.1974	0.7630	0.2641	0.8440	0.3704	0.8916
1-2-0-1	0.1956	0.7592	0.1220	0.4265	0.3634	0.9124
1-2-2-0	0.3176	0.9571	0.3175	0.8801	0.3472	0.9118
1-3-0-0	0.2635	0.9720	0.1178	0.9040	0.1168	0.9318
1-3-3-1	0.2020	0.9295	0.0851	0.7836	0.1995	0.8781

Appendix B. Build the Initial PUNQ-S3 Models

The initial models are constructed from well data combined with geological description from the PUNQ web page. The geological information is summarized as followed: Layers 1, 3, and 5 have two different facies, channel sand (facies 1) and shale (facies 2). From the geological information in Table B-1, the proportions of different facies at each layer are calculated by width and spacing ratios. For example, the width of channels in layer 3 is about 1,000 meters with 3,000 meters spacing between the centerlines of two channels. Therefore, the channel sand (facies 1) accounts for $p_1 = 1,000/3,000 = 0.33$, the proportion of shale is $p_2 = 0.67$ in Table B-2. Layers 1, 3, and 5 consist of two different facies, channel sand (facies 1) and shale (facies 2) have the same ratio as shown in Table B-2. A third facies (mouthbar) is present in layer 4 but not present in any other layers so the fraction of width occupied by facies 3 in layers 1, 2, 3 and 5 is given by $p_3 = 0$. These fractions will be used as the probability that a particular facies occupies a particular gridblock of a layer.

Table B-1: Sedimentary facies with estimates for width and spacing for major flow units of each layer.

Layer	Facies	W	Spacing
1	Channel Fill	800 m	2-5 km
2	Lagoonal Shale	-	-
3	Channel Fill	1,000m	2-5 km
4	Mouthbar	500-5,000m	10 km
5	Channel Fill	2,000m	4-10 km

In layers 1, 3 and 5, the principle direction of anisotropy was given as 110 to 170 degrees southeast which is equivalent to -20 $\sim$ -80 degrees measured from the x-axis. Actually, -60,-45 and -30 degrees are used as the principle direction of channels for layers 1, 3 and 5. The well data at Table B-3 are generated by adding Gaussian noises with standard deviation equal to 15 percent of their true values.

The porosity of channel sand is greater than 20 percent. If a well data for porosity at a point in layers 1, 2, 3 or 5 corresponds to a porosity value greater than 20 percent, the measured porosity was assumed to correspond to channel sand. Averaging all such values gives the mean values of porosity for channel sand. The variance of porosity can be estimated

Table B-2: Sedimentary facies with estimates of proportion for major flow units.

Well	Facies	p_1	p_2	p_3	Principle Dir
layer 1	Channel sand & shale	0.32	0.68	0	-60
layer 2	Lagoonal Shale	0	1.0	0	-60
layer 3	Channel sand & shale	0.33	0.67	0	-45
layer 4	Mouthbar	0	0.69	0.31	60
layer 5	Channel sand & shale	0.33	0.67	0	-30

Table B-3: Well property data

Well	PRO-1	PRO-4	PRO-5	PRO-11	PRO-12	PRO-15
layer 1-ϕ	0.0825	0.2298	0.2412	0.0807	0.0832	0.2535
layer 1-$\ln k$	3.6743	6.514	6.085	4.0519	3.5502	5.5903
layer 1-$\ln k_z$	2.2059	2.2298	6.1176	3.7693	3.0413	6.1981
layer 2-ϕ	0.0631	0.0684	0.0716	0.0867	0.0954	0.1044
layer 2-$\ln k$	3.2129	3.0137	2.5084	2.7883	4.2405	3.8122
layer 2-$\ln k_z$	1.2633	0.9875	1.0094	1.8578	3.4015	2.4284
layer 3-ϕ	0.1219	0.0995	0.2382	0.2887	0.0799	0.1521
layer 3-$\ln k$	4.6476	3.8683	5.7408	6.2875	4.2277	4.778
layer 3-$\ln k_z$	2.55	3.3104	5.4433	5.5541	3.2929	4.6214
layer 4-ϕ	0.1618	0.1504	0.166	0.1599	0.1484	0.1994
layer 4-$\ln k$	5.5268	6.533	4.8816	4.9159	5.6424	6.1363
layer 4-$\ln k_z$	4.3646	3.6502	4.168	3.1715	3.5704	3.8065
layer 5-ϕ	0.2383	0.1625	0.0987	0.1271	0.2909	0.2418
layer 5-$\ln k$	5.4004	6.1618	3.0654	4.8846	6.0195	7.3068
layer 5-$\ln k_z$	5.6233	5.9546	2.5284	2.9698	5.5591	6.0218

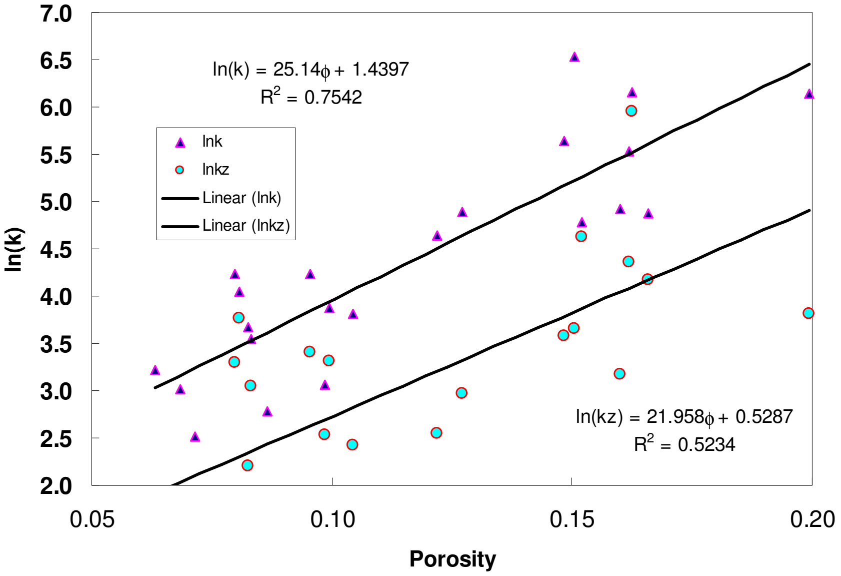


Figure B-1: $\ln k$, $\ln k_z$vs. porosity

directly. The porosity of mouthbar sand is greater than 15 percent. Repeat this process for the mouthbar and shale to obtain the mean and standard deviation of porosity. The means and standard deviations of $\ln k$ and $\ln k_z$ are computed by the same method. In Table B-5, we generated mean value of porosity based on the relative width of the layer that are occupied by each facies (Eq. B-2). The means of porosity, $\ln k$ and $\ln k_z$ of each layer j can be calculated using the following equations:

Table B-4: Mean and variance of sedimentary facies.

Facies	μ_ϕ	σ_ϕ	$\mu_{\ln k}$	$\sigma_{\ln k}$	$\mu_{\ln k_z}$	$\sigma_{\ln k_z}$
Channel Sand	0.2528	0.0238	6.118	0.603	5.536	0.77
Shale	0.0986	0.0289	3.905	0.93	2.758	1.28
Mouthbar	0.1643	0.0185	5.606	0.656	3.789	0.43

$$\bar{\phi}_j = p_{1,j}\bar{\phi}_{f,1} + p_{2,j}\bar{\phi}_{f,2} + p_{3,j}\bar{\phi}_{f,3} \tag{B-1}$$

$$\overline{\ln k}_j = p_{1,j}\overline{\ln k}_{f,1} + p_{2,j}\overline{\ln k}_{f,2} + p_{3,j}\overline{\ln k}_{f,3} \tag{B-2}$$

$$\overline{\ln k}_{zj} = p_{1,j}\overline{\ln k}_{zf,1} + p_{2,j}\overline{\ln k}_{zf,2} + p_{3,j}\overline{\ln k}_{zf,3} \tag{B-3}$$

Porosity and permeability are strongly and positively correlated , with $\rho_{\phi,\ln k} = 0.8$ for $\ln k$ *versus* ϕ (Fig. B-1). Table B-5 gives the statistics information needed to build the initial model. Layers 1, 3 and 5 have similar geostatistical properties. The following steps were taken to generate the PUNQ-S3 initial model:

Table B-5: Geostatistical data for all layers.

layers	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5
μ_ϕ	0.15	0.0986	0.15	0.11	0.15
σ_ϕ	0.10	0.0289	0.10	0.06	0.10
ϕ_{max}	0.324	0.185	0.324	0.220	0.324
ϕ_{min}	0.012	0.012	0.012	0.012	0.012
$\mu_{ln(k)}$	4.61	3.91	4.64	4.42	4.63
$\sigma_{ln(k)}$	2.62	0.73	2.51	1.52	2.51
$ln(k)_{max}$	7.927	6.695	7.927	7.574	7.927
$ln(k)_{min}$	1.115	1.115	1.115	1.115	1.115
$\mu_{ln(k_z)}$	3.65	2.76	3.68	3.07	3.67
$\mu_{ln(k_z)}$	2.29	0.63	2.20	1.33	2.20
$ln(k_z)_{max}$	7.85	6.60	7.846	5.079	7.846
$ln(k_z)_{min}$	-1.08	-1.08	-1.082	-1.082	-1.082
r_1	800	-	1000	2750	2000
r_2	5040	2520	5040	4125	5040
α	-60	-60	-45	60	-30

(i) Sequential Gaussian Simulate for porosity conditioned to well data which have noises added.

(ii) Cosimulate horizontal permeability based on the porosity field and well permeability with added noise.

(iii) Cosimulate vertical permeability based on the horizontal permeability field and conditioned to well horizontal permeability with added noise.

(iv) Transform all fields to squeeze the tails to acceptable physical values.

Squeezing is done using the following exponential transformation,

$$x_{new} = x_{maxup} + (x_{limup} - x_{maxup})\, exp(-\frac{x - x_{limup}}{x_{maxup} - x_{limup}})\; for\; x > x_{limup} \tag{B-4}$$

and similarly for $x < x_{limlow}$.

However, when a multivariate Gaussian model is used, the porosity and permeability field generated with the estimated geostatistical parameters listed in Table B-5 yield unrealistically high or low values (*e.g.*, negative porosities) because of the very large variances. We truncate such value by specifying upper and lower bounds for the rock property fields. The upper bounds and lower bounds of ϕ, $\ln k$ and $\ln k_z$ for each facies are set equal to their means plus or minus 3 times their standard deviations. The lower bounds of ϕ, $\ln k$ and $\ln k_z$ for all layers are determined by the lower bounds of shale facies estimated by this procedure. In layers 1, 3 and 5, the upper bounds of $\ln k$ and $\ln k_z$ are determined by channel sand facies. In layers 2 and 4, upper bounds are determined by shale and mouthbar facies, respectively. The estimated upper and lower bounds, denoted by subscripts of max and min, respectively, are listed in Table B-5.

Appendix C. Production Assimilation Data of PUNQ-S3

The kinds and numbers of observation data available at different times varies as listed in Table C-1. Shut-in pressures are marked in parentheses. Unmarked pressures are flowing pressures. The data consist of:

(i) Shut-in bottom-hole pressure (BHP) for each well with after 1, 4, 5, 6, 7, and 8 years of production.
(ii) Flowing BHPs for each well after 1 day and 3, 6, and 9 months, and immediately before the shut-in period after 5, 6, 7, and 8 years of production.
(iii) Gas-oil ratios (GOR) after 4.5, 5, 5.5, 6.5, and 7.5 years for each well, and after 5, 6, 7, and 8 years for the remaining wells.
(iv) Water cut after 7, 7.5, and 8 years for the one well that experienced water breakthrough, and after 8 years for the other wells.

The water cut and GOR measurements coinciding with the shut-in periods are taken immediately before shut-in.

Table C-1: The kinds and numbers of data available at different measurement times.

index	days	bottom-hole pressure	gas-oil ratio ratio	water cut	oil production rate
1	1.01	6	-	-	6
2	91	6	-	-	6
3	182	6	-	-	6
4	274	6	-	-	6
5	366	6(shut-in)	-	-	6
6	1461	6(shut-in)	-	-	6
7	1642	-	6	-	6
8	1826	6	6	-	6
9	1840	6(shut-in)	-	-	6
10	1841	-	6	-	6
11	2008	-	6	-	6
12	2192	6	6	-	6
13	2206	6(shut-in)	-	-	6
14	2373	-	6	-	6
15	2557	6	6	-	6
16	2571	6(shut-in)	-	-	6
17	2572	-	-	1	6
18	2738	-	6	-	6
19	2922	6	6	6	6
20	2936	6(shut-in)	-	-	6
Total	-	84	54	7	120

Appendix D. Nomenclature

Symbol	Description
a	range or critical distance
b	distance between grid block and observation location
$\boldsymbol{C}_D$	covariance matrix of measurement errors
$\boldsymbol{C}_M$	covariance matrix of model parameters
$\boldsymbol{C}_x$	prior covariance of control variables
$\boldsymbol{C}_Y$	covariance matrix of state vector
$\boldsymbol{D}$	matrix of data
$\boldsymbol{d}$	data vector
$E[.]$	expected value
$f(.)$	model solutions by model forward
$g(.)$	observation forecast by model forward
$\boldsymbol{G}_\ell$	sensitivity of objective function
$\boldsymbol{H}$	operator matrix or row vector
$\boldsymbol{I}$	unit matrix
$\boldsymbol{K}$	Kalman gain
$\bar{k}$	upscaled permeability
$\boldsymbol{L}$	"square root" of the covariance matrix C with Cholesky decomposition
$\ln k$	logarithm permeability
$\boldsymbol{M}$	matrix of model parameters
m	vector of model parameters
N	number of active cells
$\boldsymbol{N}_d$	number of data
$\boldsymbol{N}_e$	number of ensemble
$\boldsymbol{N}_m$	number of realization
$\boldsymbol{N}_t$	total number of time steps
N_{pD1}	fraction of the initial tracer free water recovered after 1 pore volume of injection
$\boldsymbol{O}(.)$	objective function
p	pressure
$p(.)$	probability density function (PDF)
$\tilde{q}$	mass depletion per unit volume per unit time, positive for production, negative for injection
$\boldsymbol{Q}$	covariance matrix
Q_{oi} and Q_{wi}	total oil and water production
r_τ	discount rate
R_s	solution gas-oil ratio
S_w	water saturation
$S_w c$	connate water saturation
S_g	gas saturation
T	water front arrival time
u_T	total velocity, $u_w + u_n$ in two-phase flow

Symbol	Description
$\boldsymbol{x}$	vector of control variables
$\boldsymbol{Y}$	matrix of ensembles
y	vector of state vector
$\bar{y}$	mean of vector of state variables
$\boldsymbol{Z}_j$	random vector

Greek

$\delta(.)$	updating coefficient matrix
ϵ	measurement error
γ_m	phase specific density
ω	model error vector
Ψ	state transient matrix
ϕ	porosity
σ^2	variance
σ_e	standard error of ensemble
σ_t	error between estimate mean and the true state vector for synthetic case
τ	dimensionless breakthrough time

Subscripts

j	ensemble member index
k	time index, absolute permeability
l,m	phases
o	oleic
obs	observed
$prior$	prior distribution
uc	unconditional realization
v	vapor
w	water
x	x direction
y	y direction

Superscripts

t or $true$	truth
p	predicted or prior
T	transpose
u	updated
-1	inverse

Abbreviations

Symbol	Description
1-D	one dimensional
2-D	two dimensional
3-D	three dimensional
BHP	Bottomhole pressure
KL	Karhunen-Loveve
GOR	gas-oil ratio
LBFGS	limited memory Broyden-Fletcher-Gikdfarb-Shanno
MCMC	Markov chain Monte Carlo
RML	Randomized Maximum Likelihood
WCT	water cut
WOR	water oil ratio
PDE	partial differential equation(s)

Miscellaneous

ℓ	iteration

Vita

Xin Li was born in March, 1969, at Shanghai, China. She finished her undergraduate studies at Southwest Petroleum University, Chendu, Sichuan province in July 1991. She earned a Master of Science degree in Petroleum Engineering there in July 1994. She worked for China National Offshore Oil Company from August 1994 to November 2003. In August 2004, she came to Louisiana State University to pursue graduate studies in Petroleum Engineering. She is currently a candidate for the degree of Doctor of Philosophy in Petroleum Engineering, which is expected to be awarded in December 2008.