ENSEMBLE-BASED ASSIMILATION OF NON-LINEARLY RELATED DYNAMIC DATA IN RESERVOIR MODELS EXHIBITING NON-GAUSSIAN CHARACTERISTICS

A Dissertation in Energy and Mineral Engineering by Devesh Kumar

© 2019 Devesh Kumar

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

May 2019
The dissertation of Devesh Kumar was reviewed and approved* by the following:

Sanjay Srinivasan  
Professor of Petroleum and Natural Gas Engineering  
Dissertation Advisor, Chair of Committee

Mort Webster  
Associate Professor of Energy Engineering

Gregory King  
Professor of Petroleum and Natural Gas Engineering

Steven Greybush  
Assistant Professor of Meteorology

Luis Ayala H.  
Professor of Petroleum and Natural Gas Engineering  
Associate Department Head for Graduate Education

*Signatures are on file in the Graduate School.
Abstract

Inverse modeling techniques for estimating reservoir parameters (e.g., Transmissivity, Permeability, etc.) utilize some secondary information (e.g., hydraulic head or production data at well locations) to estimate reservoir parameters. Ensemble-based data assimilation methods are one such class of inverse modeling techniques. Ensemble Kalman filters (EnKF) in specific are built around the basic framework where modeling parameters such as transmissivity, permeability, storativity, porosity, hydraulic head, phase-saturation are included within a state vector $\psi^f$ that are updated to $\psi^a$, based on the available dynamic data. Although EnKF presents the ability to update a large number of parameters successively as data becomes available, it suffers from some major drawbacks. It is optimal only in the case when the multivariate joint distribution describing the state vector is multi-Gaussian. Also, a linear update equation comprised of covariance values between the observed variables and update parameter and covariance between the different observed variables are used in EnKF. These assumptions and simplifications result in models that yield inaccurate predictions of reservoir performance.

The aim of this research work is to propose a novel method for data assimilation which is free from the Gaussian and linear transfer function assumptions. This new method can be used to sequentially assimilate dynamic data into reservoir models using an ensemble based approach. Updating is performed in the indicator space where modeling is performed non-parametrically and the indicator transform is insensitive to non-linear operations. It is demonstrated that this indicator transform helps us achieve the desired generality which is a shortcoming of EnKF. Because the expected value of indicators directly yield the probability corresponding to an outcome, the method can be used to quantify the residual uncertainty in spatial description of reservoir properties. Because at all steps of the process an ensemble of models is available, so quantification of residual uncertainty in prediction forecasts is possible. Another advantage is that the data assimilation is sequential in nature implying that the updates can be performed in a quasi-real time sense as data becomes available.
Table of Contents

List of Figures vi
List of Tables xi
List of Symbols xii
Acknowledgments xv

Chapter 1
Introduction 1

Chapter 2
Proposed non-Gaussian ensemble-based data assimilation 6
2.1 EnKF as classical kriging 6

Chapter 3
Indicator-based data assimilation 11
3.1 Markov-Bayes assumption 13
3.2 The full indicator approach 18

Chapter 4
Data assimilation using full indicator approach - a synthetic example 22
4.1 Primary indicator definition 22
4.2 Secondary indicator threshold definition 23
4.3 Model Implementation 24

Chapter 5
Data assimilation using full indicator approach - a realistic example 28
5.1 Description of reference reservoir 29
5.2 Generation of initial ensemble of models ......................... 31
5.3 Assimilation of secondary data to update ensemble of permeability models ............................................. 35
5.4 Analysis of results ............................................................. 36
5.5 Comparison of indicator-based data assimilation (InDA), ensemble Kalman filter (EnKF) and normal score ensemble Kalman filter (NS-EnKF) results .................................................. 39
5.6 Forecasting future reservoir performance using InDA-based updated models .................................................... 53

Chapter 6
Indicator-based data assimilation with localization 60
6.1 Localization method used to modify EnKF ......................... 61
6.2 Localization method used to modify InDA .......................... 63
6.3 Determination of the localization parameter $L$ ...................... 64
6.4 Data assimilation with localization ...................................... 72
6.5 Comparison of EnKF with and without localization ............... 73
6.6 Comparison of InDA with and without localization ............... 79

Chapter 7
Multiple-point extension to indicator-based data assimilation 85
7.1 Proposed multiple-point extension ...................................... 86
7.2 The SNESIM Algorithm for model updates ......................... 87
7.3 Secondary data integration for MPS updates ....................... 88
7.4 A realistic example demonstrating InDA-Multiple point statistic update ................................. 93
7.4.1 Description of reference reservoir ................................. 93
7.4.2 Generation of initial ensemble of models .......................... 94
7.4.3 Assimilation of secondary data to update ensemble of permeability models ............................................ 98
7.4.4 Analysis of results .......................................................... 99
7.4.5 Comparison of results from InDA-MPS (Smoother), InDA with localization (InDA-Loc) and EnKF with localization (EnKF-Loc) .................................................. 109

Chapter 8
Conclusions ................................................................. 118

References ................................................................. 121
List of Figures

3.1  (a) The reference model for permeability and (b) the well locations used in the example ............................................................ 15
3.2  Reference oil saturation values at the end of (a) 2 months and (b) 10 months .......................................................... 15
3.3  (a) The ensemble average of initial permeability and (b) ensemble average of final updated permeability models ................. 16
3.4  Reference value and initial uncertainty in pressure at the grid cell (39, 12) containing a well ...................................................... 16
3.5  (a) The reference model for permeability and, (b) the well locations used to sample conditioning data ............................... 17
3.6  The absolute average change in permeability compared to initial value after $1^{st}$ (a) and $5^{th}$ (b) update ......................... 17
3.7  Updating parameter (permeability) using updated cumulative density function ................................................................. 21
4.1  Indicator threshold definitions for permeability ......................... 23
4.2  $|\Delta P|$ distribution at well cell block (39, 12) ......................... 24
4.3  Reference reservoir permeability and its underlying non-Gaussian permeability distribution (inset) .......................... 25
4.4  Reference reservoir Well Locations ........................................... 26
4.5  5 percent Permeability data use as conditioning data to generate the ensemble of 500 models ................................. 27
4.6  Uncertainty in oil production of the ensemble of models and reference production data at the location of well-3(39, 12) (a) Initial Ensemble of models (b) Final ensemble of models updated using InDA .... 27

5.1  Permeability distribution of reference model layers of STANFORD-V 30
5.2  Permeability distribution of reference layers of STANFORD-V 3-D view .......................................................... 31
5.3 Training image of permeability distribution for all reference model layers ........................................... 32
5.4 Conditioning data locations for generation of initial ensemble of permeability models ................................... 33
5.5 Permeability distribution of four layers for ensemble member-23 ...................................................... 34
5.6 Cumulative probability density function for reference permeability model and initial ensemble of models ...................................................... 35
5.7 Primary indicator cut-offs for permeability distribution ................................................................. 36
5.8 Comparison of uncertainty in liquid rate forecast before (a, c, e) and after InDA update (b, d, f) ........................................................................... 37
5.9 Comparison of initial and InDA updated permeability distribution of layer-3 for ensemble members 23 ................................................................. 38
5.10 Comparison of c.d.f. of permeability for entire ensemble before and after InDA update ...................................................................................... 39
5.11 Comparison of uncertainty in liquid rate forecast of wells for EnKF (a, c, e) and InDA updated (b, d, f) models ........................................................................... 41
5.12 Comparison of uncertainty in liquid rate forecast of wells for NS-EnKF (a, c, e) and InDA updated (b, d, f) models ........................................................................... 42
5.13 Comparison of initial ensemble average of permeability with EnKF and InDA updated ensemble average of permeability for layer-3 ...................................................... 43
5.14 Comparison of initial ensemble average of permeability with EnKF and InDA updated ensemble average of permeability for layer-1 ...................................................... 44
5.15 Comparison of initial ensemble average of permeability with NS-EnKF and InDA updated ensemble average of permeability for layer-3 ...................................................... 45
5.16 Comparison of initial ensemble average of permeability with NS-EnKF and InDA updated ensemble average of permeability for layer-1 ...................................................... 46
5.17 Comparison of initial ensemble variance of permeability with EnKF and InDA updated ensemble variance of permeability for layer-3 ...................................................... 49
5.18 Comparison of initial ensemble variance of permeability with EnKF and InDA updated ensemble variance of permeability for layer-1 ...................................................... 50
5.19 Comparison of initial ensemble variance of permeability with NS-EnKF and InDA updated ensemble variance of permeability for layer-3 ...................................................... 51
5.20 Comparison of initial ensemble variance of permeability with NS-EnKF and InDA updated ensemble variance of permeability for layer-1 ...................................................... 52
5.21 Comparison of permeability cumulative density function for the entire ensemble of EnKF, NS-EnKF and InDA updated realizations ................................. 53
5.22 Forecast wells along with existing wells . . . . . . . . . . . . . . . . 54
5.23 Comparison of uncertainty in liquid rate forecast of existing wells
for initial and indicator-based data assimilation updated models . . . 55
5.24 Comparison of uncertainty in oil rate forecast of existing wells for
initial and indicator-based data assimilation updated models . . . . 57
5.25 Comparison of uncertainty in liquid rate forecast of new wells for
initial and indicator-based data assimilation updated models . . . . 58
5.26 Comparison of uncertainty in liquid rate forecast of new wells for
initial and indicator-based data assimilation updated models . . . . 59

6.1 Cross-correlation as a function of distance between observational
location 2 (well-Inj1-2) and locations in the reservoir . . . . . . . . 65
6.2 Automatic estimation of $L$ from cross-correlation data as a function
of distance away from observational location 2 (well-Inj1-2) used in
EnKF . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 67
6.3 Automatic estimation of $L$ from cross-variance data as a function of
distance away from observational location 2 (well-Inj1-2) used in InDA 68
6.4 L values for different time steps used in EnKF and InDA localization 69
6.5 (a) Liquid rate derivative with respect to time for initial ensemble
members at well-Inj-2, (b) median of liquid rate derivative with
respect to time for initial ensemble members at all well locations . . 71
6.6 Localization function as a function as a function of spatial lag
between data and parameter locations . . . . . . . . . . . . . . . . 72
6.7 (a) The reference permeability distribution on one slice of the syn-
thetic reservoir example, (b) the initial ensemble average of perme-
ability, (c) the updated ensemble average of permeability obtained
using EnKF without localization, and, (d) the updated ensemble
average of permeability obtained using EnKF with localization . . 74
6.8 (a) The reference permeability distribution on another slice of the
synthetic reservoir example, (b) the initial ensemble average of per-
meability, (c) the updated ensemble average of permeability obtained
using EnKF without localization, and, (d) the updated ensemble
average of permeability obtained using EnKF with localization . . 75
6.9 (a) The initial ensemble variance of permeability, (b) the updated
ensemble variance of permeability obtained using EnKF without
localization, and, (c) the updated ensemble variance of permeability
obtained using EnKF with localization . . . . . . . . . . . . . . . . 76
6.10 Comparison of uncertainty in liquid rate prediction for existing and
and forecast wells for EnKF updates with and without localization . 78
6.11 (a) The reference permeability distribution on one slice of the synthetic reservoir example, (b) the initial ensemble average of permeability, (c) the updated ensemble average of permeability obtained using InDA without localization, and, (d) the updated ensemble average of permeability obtained using InDA with localization.

6.12 (a) The reference permeability distribution on another slice of the synthetic reservoir example, (b) the initial ensemble average of permeability, (c) the updated ensemble average of permeability obtained using InDA without localization, and, (d) the updated ensemble average of permeability obtained using InDA with localization.

6.13 (a) The initial ensemble variance of permeability, (b) the updated ensemble variance of permeability obtained using InDA without localization, and, (c) the updated ensemble variance of permeability obtained using InDA with localization.

6.14 Comparison of uncertainty in liquid rate prediction for existing and forecast wells for InDA updates with and without localization.

7.1 Flowchart of InDA-MPS update method.

7.2 Permeability distribution of reference model.

7.3 Training image and conditioning data locations for generation of initial ensemble of permeability models.

7.4 Permeability distribution of initial ensemble member models: 10, 20, 50 and reference permeability distribution.

7.5 Cumulative probability density function for reference permeability model and initial ensemble of models.

7.6 Evolution of tau values with time.

7.7 (a) $P(A=\text{pay-zone}|B)$ (b) $P(A=\text{pay-zone}|C)$ and (c) $P(A=\text{pay-zone}|B, C)$ for assimilation step of month-2.

7.8 Primary indicator cut-offs for permeability distribution.

7.9 Comparison of uncertainty in liquid rate forecast before (a, c, e) and after InDA-MPS (Smother) update (b, d, f) (wells under consideration are shown in the inset of (b)).

7.10 Permeability distribution of InDA-MPS (Smother) updated ensemble members models 10, 20, 50 and reference permeability distribution.

7.11 Comparison of c.d.f. of permeability for entire ensemble before and after InDA-MPS (Smother) update.

7.12 Selective templates used to update global multiple-point statistics.

7.13 Initial and updated non-pay and pay category frequencies at the center of selected templates.
7.14 Initial and only TI-statistics updated ensemble members models 20
(a and b), 50 (c and d) ................................................. 108
7.15 Comparison of uncertainty in liquid rate forecast of wells for InDA-
Loc (a, c, e) and InDA-MPS (Smootheh) updated (b, d, f) models
(wells under consideration are shown in the inset of (b)) ........ 110
7.16 Comparison of uncertainty in liquid rate forecast of wells for EnKF-
Loc (a, c, e) and InDA-MPS (Smootheh) updated (b, d, f) models
(wells under consideration are shown in the inset of (b)) ........ 111
7.17 Comparison of initial ensemble average (b) of permeability with
InDA-Loc (c) and InDA-MPS (Smootheh) d) updated ensemble
average of permeability ............................................. 113
7.18 Comparison of initial ensemble average (b) of permeability with
EnKF-Loc (c) and InDA-MPS (Smootheh) (d)updated ensemble
average of permeability ............................................. 114
7.19 Comparison of initial variance of permeability with InDA-MPS
(Smootheh) (b), InDA-Loc (c) and EnKF-Loc (d) updated ensemble
variance of permeability ............................................. 117
List of Tables

6.1  $L$ values $^1$ for all observational locations for EnKF and InDA calculated for data assimilation at the end of 2-months . . . . . . . . . . 70
List of Symbols

$\psi^I$ Initial State Vector
$\psi^a$ Updated State Vector
$h_{jk}$ Lag between points in space $u_j$ and $u_k$
$C_z(h_{jk})$ Autocovariance of primary data ‘z’ separated by lag $h_{jk}$
$C_p(h_{jk})$ Autocovariance of secondary data ‘p’ separated by lag $h_{jk}$
$C_{zp}(h_{jk})$ Cross-covariance between primary data ‘z’ and secondary data ‘p’ separated by lag $h_{jk}$
$C_{pz}(h_{jk})$ Cross-covariance between secondary data ‘p’ and primary data ‘z’ separated by lag $h_{jk}$, here ($C_{zp}(h_{jk})=C_{pz}(h_{jk})$)
$\psi(u)$ Primary variable at spatial location $u$
$\psi_k$ $k^{th}$ cut-off for Primary variable ($\psi(u)$)
$I(\psi(u),\psi_k)$ Indicator definition for primary variable $\psi(u)$ given a cut-off of $\psi_k$
$S(u)$ Secondary variable at spatial location $u$
$y(u,\psi_k)$ Indicator definition for secondary variable $S(u)$ conditioned to a given primary variable ($\psi(u)$) cut-off ($\psi_k$) as per Markov-Bayes assumption
$F(u_0,\psi_k)$ Prior Cumulative probability density of the primary variable ($\psi(u)$) corresponding the $k^{th}$ cut-off ($\psi_k$)
\( F(u_o, \psi_k| (m, n)) \) Cumulative probability density of the primary variable \((\psi(u))\) corresponding the \(k^{th}\) cut-off \((\psi_k)\) conditioned to \((m)\) primary data points and \((n)\) secondary data point

\( C_1(h_{jk}) \) Indicator autocovariance of primary indicator data ‘I’ separated by lag \(h_{jk}\)

\( C_{Y1}(h_{jk}) \) Indicator cross-covariance between secondary indicator data ‘Y’ and primary indicator data ‘I’ and separated by lag \(h_{jk}\)

\( m^1(\psi_k) \) Expected value of secondary indicator ‘Y’ given primary indicator ‘I=1’ calculated by pooling all points in space. All for a given primary variable cut-off \((\psi_k)\)

\( m^0(\psi_k) \) Expected value of secondary indicator ‘Y’ given primary indicator ‘I=0’ calculated by pooling all points in space. All for a given primary variable cut-off \((\psi_k)\)

\( B(\psi_k) \) Measure of information content in the secondary data called "Beez" given by Difference \(( m^1(\psi_k) - m^0(\psi_k) \) ). Used to scale the indicator autocovariance to indicator cross-covariance. All for a given primary variable cut-off \((\psi_k)\)

\( \psi^f_\alpha \) Initial State Vector corresponding to ensemble member \(\alpha\)

\( \psi^a_\alpha \) Updated State Vector corresponding to ensemble member \(\alpha\)

\( P_{observed} \) Observed secondary data

\( P_{observed, \alpha} \) Observed secondary data with observation error

\( P_{simulated, \alpha} \) Simulated secondary data corresponding to ensemble member \(\alpha\)

\( \Delta P^f_\alpha \) Absolute difference in observed and simulated values of secondary variable corresponding to ensemble member \(\alpha\)

\( \Delta P_p \) Secondary variable \((P)\) optimum cut-off for indicator definition as per Full Indicator paradigm

\( Y_p \) Secondary variable \((P)\) indicator definition for cut-off \(\Delta P_p\) as per Full Indicator paradigm

\( N(\mu, \sigma) \) Uni-variate Gaussian distribution with mean \(\mu\) and standard deviation \(\sigma\)
$L$ Localization length parameter used for the definition of localization function

$f_{xi}$ Localization weight of observed data $i$ for update location $x$ calculated using localization function

$P(A = k)$ Also written as $P(A)$, is the prior probability of occurrence of category $k$ of a categorical random variable $A$

$P(A = k|B)$ Also written as $P(A|B)$, is the conditional probability of occurrence of category $k$ of a categorical random variable $A$ given some information $B$ (conditional data, training image etc.)

$P(A = k|C)$ Also written as $P(A|C)$, is the conditional probability of occurrence of category $k$ of a categorical random variable $A$ given some secondary information $C$ (liquid rate data, well bottom hole pressure etc.)

$P(A = k|B,C)$ Also written as $P(A|B,C)$, is the conditional probability of occurrence of category $k$ of a categorical random variable $A$ given the combined information of $B$ and $C$

$\tau_1$ Measure of information redundancy in data source $B$ towards combined probability of $P(A|B,C)$

$\tau_2$ Measure of information redundancy in data source $C$ towards combined probability of $P(A|B,C)$

$a$ Inverse odds ratio representing a measure of information towards the occurrence of an event $A = k$, contained in the prior probability $P(A = k)$

$b$ Inverse odds ratio representing a measure of information towards the occurrence of an event $A = k$, contained in the conditional probability $P(A = k|B)$

$c$ Inverse odds ratio representing a measure of information towards the occurrence of an event $A = k$, contained in the conditional probability $P(A = k|C)$

$x$ Inverse odds ratio representing a measure of information towards the occurrence of an event $A = k$, contained in the conditional probability $P(A = k|B,C)$
Acknowledgments

I would like to thank my adviser, Dr. Sanjay Srinivasan, for providing me the opportunity to pursue my doctoral research under his guidance. I would also like to thank my doctoral committee members, Dr. Steven Greybush, Dr. Mort Webster and Dr. Gregory King for their valuable suggestions and insights. I am also grateful to my brother and parents for their continued support throughout the course of my graduate education.

Finally, the funding received as a part of the John and Willie Leone family chair is gratefully acknowledged.
Chapter 1  Introduction

Model parameters in petroleum reservoir applications are typically, spatial and temporal properties describing reservoir characteristics. These properties are generally not directly measurable in the field; they are, therefore, inferred by indirect methods such as model inversion. These inversions aim to accurately estimate the model parameters, such that model predictions of certain variables made with a forward model using these parameters, are consistent with the observed data for the same variables. Sequential data assimilation techniques like EnKF have been largely used for updating model parameters using secondary information. EnKF is a linear update scheme where a state vector containing primary variables of interest is updated sequentially when some secondary information is available. The main drawback with EnKF is that the evolution of the filter is governed by the second-order moments describing the multivariate and non-linear relationship between the update parameters and the data mismatch. Furthermore, the update is obtained as a weighted linear combination of the data mismatches. This proves to be sub-optimal in case of either non-Gaussian parameter distributions or non-linear relationship between the data mismatch and the parameter updates.

The initial guesses of the primary variables of interest are usually known with very high uncertainty, thus an ensemble of state variable realizations is generated using an appropriate algorithm. The EnKF updates are designed such that with each EnKF update the uncertainty in variables are reduced. Often the association between the reservoir parameters and secondary variable, such as the mismatch in flow predictions are extremely complicated, therefore, the EnKF method utilizes the first and the second order statistic calculated from the ensemble of realizations to approximate it.
Several methods have been proposed to overcome these shortcomings of EnKF. One such solution is the normal score ensemble Kalman filter (NS-EnKF), (Zhou et al. 2011; Schoniger et al. 2012). In this method, the probability distribution of the ensemble of variables at each location and at all time steps is transformed into a normal distribution by applying a normal transform. EnKF updates are then performed on these transformed variables with the assumption that the updates made using the covariance-driven Kalman gain term would be optimal. Finally, the updated variables are transformed back to their non-Gaussian distribution. However, the univariate Gaussian transformation performed on the state variables does not necessarily ensure that the joint distribution of the variables is multi-Gaussian. Consequently, the NS-EnKF updates are again sub-optimal. In addition, the drawback due to the representation of a complex non-linear relationship between the updated variable and the mismatch of flow response by a pair-wise covariance function remains in this approach.

Some other limitations of EnKF include a deteriorating quality of analyzed covariance matrix during real-time data assimilation. Because of this deterioration, the ensemble variance is increasingly underestimated over time, in an effect termed the filter-inbreeding problem (Hendricks Franssen and Kinzelbach 2008). Decrease in filter inbreeding can be achieved by inflating the covariance of the predicted state covariance matrix on the basis of a comparison between model uncertainty and observed errors at the measured points. To deal with this, a constant (Anderson and Anderson 1999) or adaptive (Anderson 2009; Miyoshi 2011) covariance matrix multipliers are applied. In essence, these multipliers are aimed at inflating or increasing the prior variance of update parameters, thus giving more weight to the information obtained from observations and subsequently inhibiting the separation of ensemble responses from truth. This separation is an unwanted process called filter divergence and is triggered by overconfidence in prior uncertainty of update variables. Although a possible solution, there is usually no specific basis for the choice of the multiplier and optimizing this multiplier could be an expensive exercise itself. Additionally, the finite size of the ensemble, the Gaussian parameter distribution assumption and linear forward flow function assumptions collectively impede the performance of EnKF. They can lead to extreme, nonphysical values for petrophysical properties, ultimately leading to the realizations departing significantly from prior information (Heidari et al. 2013).
Particle filters, in the form of sequential Bayesian filtering, possess the capability to represent the full probability distribution of predictive uncertainties, not just the first two moments (Moradkhani et al. 2005). Thus, particle filters can be used as tools to estimate posterior distributions of parameters. Subsequently, realizations from the posterior distribution can be obtained by Monte Carlo sampling. If the prior distribution of the parameters is known at least up to a proportionality, the posterior distribution can also be derived up to a proportionality. A proposal distribution is chosen to represent the approximate posterior distribution. This proposal distribution can be used to generate some initial samples (particles). The ratio of the probabilities of the particles from the approximate posterior and the proposal distribution is used to define weights corresponding to each particle. A new set of particles is then resampled with the probabilities equal to their weights. This process is called sampling importance resampling (SIR) (Rubin et al. 1988). In the limit after multiple resampling steps, the samples should closely resemble the real posterior distribution. The entire procedure of getting samples that represent the posterior distribution better after each filtering is called particle filtering (PF). For PF to work, the parameter probability distribution must be known at least up to a proportionality. In many cases, especially in geosciences, knowledge about the posterior parametric probability distribution may be elusive. Additionally, the proposal distribution has to be carefully chosen in order for the process to converge, this could be very difficult for complex parameter probability densities.

Another interesting approach aimed towards estimation of Gaussian parameter fields is Gradual Deformation (Hu 2000, LeRavalec-Dupin 2002). In this method, an initial suite of Gaussian models is generated first. Next, a new model is proposed, which is represented as the linear combination of the initial Gaussian models. The coefficients defining the linear combination are defined as a deformation parameter that is established by implementing a one parameter optimization scheme that minimizes the mismatch between the simulated and observed responses. Since a linear combination of Gaussian distributions is always Gaussian, the resultant model is assured to be multi-Gaussian. Any convenient single-parameter optimization scheme such as the Dekker-Brent method (Press et al. 1987, Chapter 9) can be carried out to minimize the objective function. The method has been demonstrated to result in a gradual deformation of the Gaussian field between an initial state (defined by a deformation parameter of 0) and a new state (corresponding to a
parameter value of 1), with intermediate values of the deformation parameter resulting in a morphing between the two end states. The method is theoretically accurate and practically effective for multi-Gaussian fields, but a problem arises if the fields are non-Gaussian, as linear combination of non-Gaussian fields might not necessarily preserve the characteristics of the original field. A proposed solution for taking care of the issue arising due to the non-Gaussian distribution, is to apply a Gaussian transformation to the existing non-Gaussian field and applying gradual conditioning to the transformed data, as discussed before. Although the non-Gaussian nature of the marginal distributions can be preserved by this technique, the spatial joint distribution of parameters may be non-Gaussian and hence not preserved by the process.

Another extension to handle non-Gaussian distributions is the probability perturbation (Tarun and Sanjay 2003; Hoffman and Caers 2005) scheme in which the information in production data pertaining to reservoir heterogeneity is calibrated in a probabilistic manner. The conditional probability (Gaussian or otherwise) representing the uncertainty in permeability at a location is iteratively updated using a deformation parameter that accounts for the additional information contained in the dynamic response data. Like the gradual deformation approach, a single deformation parameter is optimized that minimizes the deviation of the simulated response from the observation. This dynamic perturbation parameter controls the acceptance or rejection of a parameter variable category. In that sense, the parameter controls the transition from a prior probability distribution at an iteration step to a posterior probability distribution represented by a new realization of the permeability field. It is possible that an extremely large number of iterations may be required to achieve a match because the perturbation of the permeability field is controlled by a single parameter. To accelerate convergence, Yadav et al. (2005), proposed a domain decomposition scheme, using which deformation parameters for several reservoir regions could be optimized simultaneously using parallel CPUs.

In the following sections, a new method called indicator-based data assimilation (InDA), is introduced for assimilating dynamic observed data into reservoir models using an ensemble of realizations. It is shown that the proposed method can handle marginal non-Gaussian distributions of the parameters being updated and applies a non-linear parameter update equation. The mentioned generalities are achieved simultaneously using the indicator transformations of the variables involved in
the assimilation scheme. The following sections proceed by first presenting a link between EnKF and geostatistical data assimilation within a indicator framework. The description is followed by the definition of indicator transformations used in the study, and, then, the results obtained by the application of the proposed method on synthetic and realistic reservoirs are discussed. Next, the proposed method is compared with other methods such as EnKF and NS-EnKF. The results of EnKF and InDA updated models are analyzed, which gives rise to the need for integration of localization techniques into the update equation in both the aforementioned methods. As a result, a modified indicator-based data assimilation with localization (InDA-Loc) is proposed, and the results are compared to those from a localized EnKF scheme. Analysis of models updated from InDA-Loc are shown to yield reliable forecasts, of future state of the ensemble of reservoir models. However, some limitations of this extension are demonstrated in terms of lack of reproduction of the multiple-point statistics describing the spatial connectivity of the parameter models. To solve this issue, a modified version of indicator-based data assimilation with localization is proposed, which performs updates to multiple point statistics describing the spatial connectivity of models instead of just performing covariance (two-point statistics) based updates used in EnKF and InDA. We refer to this method as indicator-based data assimilation with multiple-point statistics (InDA-MPS). Finally, the results from the updates using InDA-Loc, EnKF-Loc and InDA-MPS are analyzed, and some recommendations for future work are offered.
Chapter 2  
Proposed non-Gaussian ensemble-based data assimilation

The proposed scheme for the ensemble-based data assimilation is based on the recognition that EnKF is a regression scheme in which the required statistics are established using the ensemble. This allows us to propose a non-parametric regression scheme based on the concept of indicators. Such a scheme has the advantage that it is optimal regardless of whether the underlying random function model is non-Gaussian and regardless of whether the relationship between the state variables is controlled by a non-linear transfer function model. The development of such a non-parametric ensemble based scheme is described in the subsequent sections.

2.1 EnKF as classical kriging

EnKF has gained a lot of attention lately as a scheme for parameter estimation in petroleum reservoir modeling applications (Haugen et al. 2008). This is because of its ability to handle large number of reservoir parameters and the use of a computationally inexpensive updating scheme. This method computes updates to an initial ensemble of models recursively using a scheme based on only the first and second order moments of the multivariate probability distribution characterizing the relationship between the state variables such as the reservoir rock properties and the flow responses such as pressures, rates, saturation etc. These moments are used to compute a Kalman gain that in turn is used to compute the updated increments to parameters corresponding to an ensemble of reservoir model realizations. EnKF
is advantageous over other stochastic inversion techniques as it does not use an objective function approach to perturb any parameter, and thereby avoids getting trapped in a local minima which is a major advantage over other techniques. In order to gain insight into the working of EnKF we can understand it in the context of simple kriging. Simple kriging is a well-known geostatistical interpolation tool that uses linear combination of conditioning data in order to calculate the best linear unbiased estimate (BLUE) (Isaaks and Srivastava 1989, Chapter 9). This estimate can be interpreted as the conditional expectation of the attribute being calculated, and since that expectation is expressed as a linear combination of the available data, the estimate is optimal for a Gaussian random function. The estimate is unbiased and additionally, the error variance is minimized in order to give us a system of equation that can be solved to establish the weights for the linear combination. If there are 'm' spatially distributed primary conditioning data annotated by 1, 2, 3, ..., m separated by varying lags \( h_{jk} (j, k = 1, 2, 3, ..) \) and the unknown location is designated as 'o', then the resulting simple kriging system of equations is:

\[
\sum_{k=1}^{m} v_k C_z (h_{jk}) = C_z (h_{jo}) \text{ where } j = 1, 2, 3, .. \quad (2.1)
\]

The estimate can be evaluated as:

\[
z_o = z_{\text{mean}_o} + \sum_{k=1}^{m} v_k (Z_{\text{observed}_k} - Z_{\text{mean}_k}) \quad (2.2)
\]

Now if we were to additionally consider secondary data in our estimation scheme, then the modified equations assuming 'n' spatially distributed secondary data would look like:

\[
\sum_{k=1}^{m} v_k C_z (h_{jk}) + \sum_{i=1}^{n} \lambda_i C_p (h_{ji}) = C_z (h_{jo}) \text{ where } j = 1, 2, 3, .. m \quad (2.3)
\]

\[
\sum_{k=1}^{m} v_k C_p (h_{jk}) + \sum_{l=1}^{n} \lambda_l C_p (h_{jl}) = C_p (h_{jo}) \text{ where } j = 1, 2, 3, .. n
\]

\( C_z (h) \) and \( C_p (h) \), are cross-covariance between the hard primary data \( (z) \) and the secondary data \( (p) \) at spatial \( h \) lag while \( C_z (h) \) and \( C_z (p) \) are the auto-covariance.
of primary and secondary data respectively. This system of equations Eq. (2.3) is known as co-kriging (Isaaks and Srivastava 1989, Chap. 17). The co-efficient $v_k's$ and $\lambda_i's$, weigh the contributions of primary and secondary conditioning data respectively to get a “BEST” estimate (Eq. (2.4)):

$$ z_o = z_{\text{mean}_o} + \sum_{k=1}^{k=m} v_k(Z_{\text{observed}_k} - Z_{\text{mean}_k}) + \sum_{l=1}^{l=n} \lambda_l(p_{\text{observed}_l} - p_{\text{mean}_l}) $$ (2.4)

If we imagine a scenario where the conditional expectation of the primary variable at the observed location is informed by secondary data alone, the co-kriging system of equation simplifies to:

$$ \sum_{l=1}^{l=n} \lambda_l C_p(h_{jl}) = C_{pz}(h_{jo}) \quad \text{where} \quad j = 1, 2, 3, \ldots n $$ (2.5)

In matrix form it can be written as:

$$ \begin{bmatrix} C_p(h_{11}) & C_p(h_{12}) & \ldots & C_p(h_{1n}) \\ C_p(h_{21}) & C_p(h_{22}) & \ldots & C_p(h_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ C_p(h_{n1}) & C_p(h_{n2}) & \ldots & C_p(h_{nn}) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C_{pz}(h_{1o}) \\ C_{pz}(h_{2o}) \\ \vdots \\ C_{pz}(h_{no}) \end{bmatrix} $$ (2.6)

Or Solving for $\lambda$ we can simply write:

$$ [\lambda]_{n\times 1} = [C_p(h_{jl})]_{n\times n}^{-1} [C_{pz}(h_{jo})]_{n\times 1} $$ (2.7)

And the estimated value becomes:

$$ z_o = z_{\text{mean}_o} + \sum_{l=1}^{l=n} \lambda_l(p_{\text{observed}_l} - p_{\text{mean}_l}) $$ (2.8)

$$ z_o = z_{\text{mean}_o} + [\lambda^T]_{1\times n} p_{\text{observed}} - p_{\text{mean}} $$ (2.9)

Replacing in terms of co-variances

$$ z_o = z_{\text{mean}_o} + [C_{pz}(h_{jo})]_{n\times 1} [C_p(h_{lj})]^{-1}_{n\times n} p_{\text{observed}} - p_{\text{mean}} $$ (2.10)
Under stationarity, the requisite covariances are calculated by pooling all data that are separated by the same spatial lag. However, instead of pooling together the data over the entire domain, the ensemble of models generated by stochastic simulation techniques affords us an opportunity to compute covariances by looking at repeat observations at specific pairs of locations. This becomes especially advantageous when computing covariance of flow variables such as pressure that can exhibit systematic variability over the entire domain in response to specified boundary conditions. In this case we alter the above equation in the following way and use the same to evaluate estimates at all locations $o = 1, 2, 3, ..$

$$z_a^o = z_f^o + [C_{pz}(h_{jo})]_{1 \times n}[[C_p(h_{lj})]_{n \times n} + \text{diag}[C_{\epsilon \epsilon}(h_{lj})]_{n \times n}]^{-1}[p_{\text{observed}} - p_{\text{simulated}}]_{n \times 1}$$

where $o = 1, 2, 3, ..$ and $j = l = 1, 2, 3, .., n$ \quad (2.11)

Comparing the co-kriging form of the update equation (Eq. (2.9)) with the EnKF version (Eq. (2.11)), it can be seen that at a given location 'o' the primary variable mean ($z_{\text{mean}_o}$) in co-kriging formulation is replaced by the present variable value ($z_f^o$) in the EnKF formulation. The updated and present values of 'z' are denoted by the superscript 'a' and 'f' respectively. Additionally in the EnKF formulation the simulated values of the secondary variable at the locations of observation replace the mean of the secondary data values ($p_{\text{mean}_j}$) in the co-kriging formulation. Samples from the distribution $\mathcal{N}(0, C_{\epsilon \epsilon})$ is added to each of the observations to account for error in the observation data. As $\epsilon$ is a vector of uncorrelated observation errors the variance of the observation errors $C_{\epsilon \epsilon}$ ($C_{\epsilon \epsilon} = \epsilon \epsilon^T$) is also uncorrelated making $C_{\epsilon \epsilon}$ a diagonal matrix.

For generality a state vector ($\psi$) is defined, with components as static and dynamic variables like permeability, porosity, pressure and saturation at all locations in the spatial domain under consideration. The elements in ($\psi$) can be arranged in a way that the first 'm' values are primary data values at all locations and last 'n' elements are secondary variables. An operator $H$ is defined that relates the state vector to secondary data (Haugen et al. 2008), by the operation $H\psi$. Now the EnKF update equation (Eq. (2.11)) can be written in terms of the following
notation:

\[ \psi^a = \psi^f + C_{\psi\psi} H^T [H C_{\psi\psi} H^T + C_{\epsilon\epsilon}]^{-1} [p_{\text{observed}} - H \psi^f] \]  

(2.12)

The operation \( HC_{\psi\psi} H^T \) serves to retrieve the variance of the secondary data from the entire auto-covariance matrix \( C_{\psi\psi} \), while the operation \( C_{\psi\psi} H^T \) retrieves the cross-covariance (relationship) between the parameters to be update and the secondary data. As we are dealing with an ensemble of model realizations rather than '1' model, we can append a subscript 'j' representing the index of the realization in the ensemble:

\[ \psi^a_j = \psi^f_j + C_{\psi\psi} H^T [H C_{\psi\psi} H^T + C_{\epsilon\epsilon}]^{-1} [p_{\text{observed}} - H \psi^f_j] \]  

(2.13)

It is worthwhile noticing that this formulation (Eq. (2.13)) derived from modifying the co-kriging (Eq. (2.9)) equation is precisely of the same form as the EnKF assimilation equation in Haugen et al. (2008), thus establishing that EnKF is nothing but a form of co-kriging that uses the ensemble members to deduce the required statistics. This use of kriging in data assimilation has also been shown in some recent literature (Tolosana-Delgado et al. 2011). As co-kriging and subsequently EnKF utilizes a linear estimate of the conditional expectation, they are optimal in the case of multivariate Gaussian distributions defining the relationship between the primary parameters and the secondary data. Thus it may not be able to handle non-Gaussian variable distributions. This leads us towards geostatistical tools like indicator kriging, tailored to handle non-Gaussian distributions. These tools can be augmented accordingly to achieve generalized assimilation of data.
Chapter 3  
Indicator-based data assimilation

As mentioned in the last chapter (Chap. 2), EnKF is optimal only when the relationship between the state variables follow a multivariate Gaussian distribution and the relationship between the estimator for the model parameter and the pressure response is linear. These constraints are seldom realistic and rather restrictive. Indicator-based approaches in geostatistics allow us to address both these issues. In geostatistics, indicator based kriging and simulation are often used. This is performed after both the primary and secondary data are indicator coded.

**Indicator Coding of primary variable:** The value of Primary variable $\psi$ at a given location $u$ is $\psi(u)$, and by specifying the $k$ threshold values ($\psi_1, \psi_2, \psi_3, ..., \psi_k$), $\psi(u)$ can be written into an indicator vector $I(\psi(u), \psi_k)$ defined as follows:

$$I(\psi(u), \psi_k) = I(u, \psi_k) = \begin{cases} 
1, & \text{if } \psi(u) \leq \psi_k \\
0, & \text{otherwise}
\end{cases} \quad k = 1, 2, 3, ..., k,$$

So $I(\psi(u), \psi_k) \in \{0, 1\}$.  \hspace{1cm} (3.1)

**Indicator definition for secondary variable:** The value of secondary variable $(s)$ at a given location $u$ is $s(u)$. The indicator definitions for the secondary variables are often referred to ‘Soft Indicators’ ($y(u, \psi_k)$) and are defined as follows:

$$y(u, \psi_k) = \text{Prob}(\psi(u) \leq \psi_k | S = s(u)),$$
$\text{So } y(u, \psi_k) \in [0, 1].$  

Although the above indicator definitions are suitable for continuous variables, the transforms for discreet variables can also be defined with a slight modification. The expectation of the indicator vector $I(u, \psi_k)$ of a continuous variable directly yields the cumulative probability corresponding to the threshold used to define the indicator variable. Using the indicator transformed data, indicator co-kriging yields the conditional expectation of the indicator variable ($I^*(u_o; \psi_k | (m, n))$) at the estimation node ($u_o$), given '$m$' primary data and '$n$' secondary data. In other words it yields the conditional cumulative probability $F(u_o; \psi_k | (m, n))$ corresponding to the indicator threshold at the estimation node. This can be written as:

$$F(u_o; \psi_k | (m, n)) = I^*(u_o; \psi_k | (m, n)),$$

$$= F(u_o; \psi_k) + \sum_{i=1}^{i=m} \lambda_i[I(u_i, \psi_k) - F(u_i, \psi_k)],$$

$$+ \sum_{j=1}^{j=n} v_j[y(u_j; \psi_k) - E(Y(u_j, \psi_k))].$$  

(3.3)

In presence of secondary data only the above equation Eq. (3.3) becomes:

$$F(u_o; \psi_k | n) = F(u_o; \psi_k) + \sum_{j=1}^{j=n} v_j[y(u_j; \psi_k) - E(Y(u_j, \psi_k))].$$  

(3.4)

The system of equations to be solved are:

$$\sum_{j=1}^{j=n} v_j C_Y(h_{ij}) = C_{Yf}(h_{jo}) \text{ where } i = 1, 2, 3, .., n.$$  

(3.5)

The alteration required in a setting, where ensemble of models are available:

$$F(u_o; \psi_k | n) = F(u_o; \psi_k) + \sum_{j=1}^{j=n} v_j[y(u_j; \psi_k) - E(Y(u_j, \psi_k))].$$  

(3.6)

As can be seen clearly the definition of the system of equations (Eq. (3.5)), require indicator cross-covariance values between primary-secondary $C_{Yf}(h_{jo})$ data and
the indicator auto-covariance values of the secondary data $C_Y(h_{ij})$. The task of achieving a positive definite matrix for these covariance values might be difficult, so a general ‘trick’ in geostatistics is to resort to a ‘Markov-Bayes’ assumption.

### 3.1 Markov-Bayes assumption

The Markov-Bayes assumption is essentially a data screening hypothesis (Zhu and Journel 1993), that allows us to calculate the auto-covariance of secondary data and its cross-covariance with primary data simply as a function of the auto-covariance of primary data.

\[
C_Y(h, \psi_k) = \begin{cases} 
|B(\psi_k)|C_I(h, \psi_k) & \text{if } ||h|| = 0, \\
B(\psi_k)^2C_I(h, \psi_k) & \text{if } ||h|| > 0,
\end{cases}
\]

\[
C_{IY}(h, \psi_k) = B(\psi_k)C_I(h, \psi_k),
\]

\[
m^1(\psi_k) = E(Y(u, \psi_k)|I(u, \psi_k) = 1),
\]

\[
m^0(\psi_k) = E(Y(u, \psi_k)|I(u, \psi_k) = 0),
\]

\[
B(\psi_k) = m^1(\psi_k) - m^0(\psi_k), \quad \epsilon [-1, 1].
\]

The “Beez” ($B(\psi_k)$) used to scale the primary indicator covariances, measure the information content in the secondary data. As a result of the assumptions only the secondary data collocated at the estimation nodes are required for variable estimation. The indicator estimates can be solved using Eq. (3.6), the weights $v'_j$s can be calculated by solving Eq. (3.5). The advantage under the ‘Markov-Bayes’ assumption is that only the indicator auto-covariance values of the secondary data are required, as the cross-covariance values can be calculated using the “Beez’ values, this results in simplifications making it easy to calculate the estimates(Eq. (3.6)). As the final estimate at any location is a posterior probability corresponding a particular threshold, repeating the estimation process for all thresholds yields a posterior cumulative density function(cdf) of the parameter at that location.

**Example application of data assimilation using ‘Markov-Bayes’ assumption:** The aforementioned ‘Markov-Bayes’ method (Eq. (3.7)) is used for data assimilation using an ensemble of permeability model for two different synthetic cases:
a.) Oil Saturation data known at each grid location as conditioning data. The saturation data are available for the time step when models are updated.

b.) Well block pressure for the grid locations containing wells as conditioning data.

Fig. 3.1a shows the reference permeability model. It is evident from the given variation in permeability values that the flow of fluids in this system is likely to be localized along the high permeability pathways. For both cases, the ‘Markov-Bayes’ model for data integration within an indicator framework was implemented. The saturation map used as secondary data source for the assimilation is shown in Fig. 3.2. The results obtained by integrating the saturation data are shown in Fig. 3.3. The initial ensemble of models, (Fig. 3.3a) tend to over-estimate the E-W continuity of the models. The final set of models (Fig. 3.3b) seem to restrict the continuity of the models. However, it is also clear that the strong screening assumptions underlying the ‘Markov-Bayes’ algorithm tends to result in the models maintaining continuity only in the regions close to the “hard” conditioning data. Furthermore, the assumption of having saturation data at each and every location is rather unrealistic but necessary to satisfy the screening assumptions. In order to demonstrate the application for the second case where the pressure values in the grid blocks containing wells, are used to condition the models (Fig. 3.4), a modified case was considered. The reference model and the location of the wells for this case are shown in Fig. 3.5. The perturbations to the permeability field obtained using the ‘Markov-Bayes’ process are depicted in Fig. 3.6. These results are the ensemble average of model update magnitudes after assimilating the data corresponding to the 1\text{st} update (end of Month-2) and the 5\text{th} update (end of Month-10). As can be seen clearly in this case, the permeability updates are localized to the regions surrounding the wells. As noted before, this is directly attributable to the screening hypothesis employed in the ‘Markov-Bayes’ algorithm.
Figure 3.1 (a) The reference model for permeability and (b) the well locations used in the example

Figure 3.2 Reference oil saturation values at the end of (a) 2 months and (b) 10 months
Figure 3.3 (a) The ensemble average of initial permeability and (b) ensemble average of final updated permeability models.

Figure 3.4 Reference value and initial uncertainty in pressure at the grid cell (39, 12) containing a well.
Figure 3.5 (a) The reference model for permeability and, (b) the well locations used to sample conditioning data

Figure 3.6 The absolute average change in permeability compared to initial value after 1st (a) and 5th (b) update
3.2 The full indicator approach

The proposed approach is free from the assumptions of Gaussianity and linearity which may limit EnKF in some cases. It relies on updating parameter distributions directly estimated using indicator data obtained by applying thresholds. In order to integrate all sources of information in this framework, full co-indicator kriging is performed.

Defining a primary variable indicator:

\[
I(\psi^f_\alpha(u), \psi_k) = \begin{cases} 
1 & \text{if } \psi^f_\alpha(u) \leq \psi_k, \\
0 & \text{otherwise}
\end{cases}
\]  

(3.8)

The parameters to be updated \((\psi^f_\alpha(u))\) is a set \(\psi^f_\alpha(u) \forall u \in \text{spatial domain}\). The subscripts \(\alpha = 1, 2, 3, \ldots, n_{en}\) correspond to each of the \(n_{en}\) ensemble members. All operations and updates are performed using these indicators and then the updates are mapped back to the actual parameter values.

Defining a secondary variable indicator: If a forward model, typically a flow simulation model is run on the ensemble of model realizations then in addition to an ensemble of parameters values \((\psi^f_\alpha(u))\), a suite of simulated values of the secondary data \(P^f_\text{simulated-}\alpha\) is also generated. At the same time observed flow responses \((P^f_\text{observed})\) are also recorded; then \(n_{en}\) number of samples are drawn from a suitable observation error distribution, usually Gaussian \((N(0, C_\epsilon))\), as discussed in Sect. 2.1. These errors are added to the actual recordings to get observed responses with recording error \((P^f_\text{observed-}\alpha)\). With the availability of observed secondary data responses \(P^f_\text{observed-}\alpha\) at \('n' locations, a measure of distance between simulated and observed response \(|\Delta P^f_\alpha|\) can be defined (Eq. (3.9)). This measure can now be used as secondary data for updating parameters.

\[
|\Delta P^f_\alpha| = \text{abs}(P^f_\text{observed-}\alpha - P^f_\text{simulated-}\alpha),
\]

(3.9)

Where \(\alpha = 1, 2, 3, \ldots, n_{en}\).
The required indicator definitions for $|\Delta P^f_\alpha|$ are defined as $Y_p$ at all 'n' locations of observations as Eq. (3.10):

$$Y_p = Y_p(|\Delta P^f_\alpha|, |\Delta P|_p) = \begin{cases} 1, & \text{if } |\Delta P^f_\alpha| \leq |\Delta P|_p \\ 0, & \text{otherwise} \end{cases},$$

Where $|\Delta P|_p$ is a cut-off for the $p^{th}$ secondary data available, and $p = 1, 2, 3, ..., n$. \hspace{1cm} (3.10)

**Parameter update using full indicator approach:** As all the indicator definitions are binary (1 or 0), the conditional cumulative probability ($F(\psi^f_\alpha(u) \leq \psi_k | Y_1, Y_2, Y_3, ..., Y_n)$) at a location $u$ given $n$ secondary data observations can be written in the indicator update framework as Eq. (3.11).

$$F(\psi^f_\alpha \leq \psi_k | Y_1, Y_2, Y_3, ..., Y_n) = E(I(\psi^f_\alpha(u), \psi_k) | Y_1, Y_2, Y_3, ..., Y_n). \hspace{1cm} (3.11)$$

We define new variables $Z_i, i \epsilon \{o, 1, 2, ..., n\}$ such that:

$Z_o = 1$ and $Z_i = Y_i - E(Y_i), \; i \neq o$.

Then, the conditional expectation Eq. (3.11) can be written in terms of $Z_i$'s as:

$$E(I(\psi^f_\alpha(u), \psi_k) | Y_1, Y_2, Y_3, ..., Y_n) = E(I(\psi^f_\alpha(u), \psi_k) | Z_o, Z_1, Z_2, Z_3, ..., Z_n),$$

$$= I(\psi^f_\alpha(u), \psi_k)^*,$$

$$= \sum_{i=0}^{i=n} \lambda_i Z_i,$$

$$= \lambda_0 \cdot 1 + \sum_{i=1}^{i=n} \lambda_i Z_i,$$

$$= E(I(\psi^f_\alpha(u), \psi_k)) + \sum_{i=1}^{i=n} \lambda_i (Y_i - E(Y_i)). \hspace{1cm} (3.12)$$

Eq. (3.12) gives the formulation of calculating posterior probability of a parameter at a desired location conditioned to secondary indicator data ($Y_p^n$). The secondary
indicator definition \((Y_p)\) is equal to one for cases matching reference data closely and zero otherwise. The goal is to calculate probabilities conditioned to models closer to the reference values \((Y_p = 1)\), which can be achieved as described in Eq. (3.13):

\[
E(I(\psi_f, \psi_k)|Y_1 = 1, Y_2 = 1, Y_3 = 1, ..., Y_n = 1) = \sum_{i=1}^{n} \lambda_i (1 - E(Y_i)).
\]  

(3.13)

Solving Eq. (3.13) for all parameter thresholds results in an updated cumulative density (cdf) function at a given location. To get an updated value of parameter (e.g., permeability), the initial cdf value corresponding the quantile equal to the initial value of permeability is read, next the new quantile (updated value) corresponding this cdf value is read from the updated cdf (Fig. 3.7). The weights \(\lambda_i's\) can be obtained by solving the system of equations Eq. (3.14) derived from projection theorem (Luenberger 1997, pp. 49-50):

\[
\begin{bmatrix}
C_{y_1 y_1} & C_{y_1 y_2} & \cdots & C_{y_1 y_n} \\
C_{y_2 y_1} & C_{y_2 y_2} & \cdots & C_{y_2 y_n} \\
\vdots & \vdots & \ddots & \vdots \\
C_{y_n y_1} & C_{y_n y_2} & \cdots & C_{y_n y_n}
\end{bmatrix}
\begin{bmatrix}
\lambda_1 \\
\lambda_2 \\
\vdots \\
\lambda_n
\end{bmatrix}
= \begin{bmatrix}
C_{IV_1} \\
C_{IV_2} \\
\vdots \\
C_{IV_n}
\end{bmatrix}.
\]

(3.14)

The requisite indicator covariances used in Eq. (3.14) can again be computed from the ensemble of indicator transformed models. Combining Eq. (3.13) and Eq. (3.14), the cumulative density update of a parameter variable for a given parameter cut-off can be written as shown in Eq. (3.15):

\[
E(I(\psi_f, \psi_k)|Y_1 = 1, Y_2 = 1, Y_3 = 1, ..., Y_n = 1) =
\]

\[
E(I(\psi_f, \psi_k)) + \begin{bmatrix}
C_{IV_1} & C_{IV_2} & \cdots & C_{IV_n}
\end{bmatrix}
\begin{bmatrix}
\begin{bmatrix}
C_{y_1 y_1} & C_{y_1 y_2} & \cdots & C_{y_1 y_n} \\
C_{y_2 y_1} & C_{y_2 y_2} & \cdots & C_{y_2 y_n} \\
\vdots & \vdots & \ddots & \vdots \\
C_{y_n y_1} & C_{y_n y_2} & \cdots & C_{y_n y_n}
\end{bmatrix}
^{-1}
\begin{bmatrix}
1 - E(Y_1) \\
1 - E(Y_2) \\
\vdots \\
1 - E(Y_n)
\end{bmatrix}
\]

(3.15)
Figure 3.7 Updating parameter (permeability) using updated cumulative density function
As discussed in the previous Sect. 3.2, in order to apply the full indicator data assimilation approach we need the indicator transforms for both the primary and the secondary data.

### 4.1 Primary indicator definition

Primary indicator thresholds are defined such that they are consistent with the cdf of the primary variable (e.g., permeability) to the best possible level of accuracy. As indicator-based data assimilation (InDA) involves cumulative probability updates at the thresholds, it is very important to define them in a way such that they capture the cumulative probability variation across the entire range of the primary variable. A complex cdf with several sharp changes in the cumulative probability value is indicative of multiple modes in the corresponding histogram and is best described using multiple thresholds around the modes. An example of threshold definitions is (Fig. 4.1, dark dots), with a total of 38 thresholds. It is evident that the number of thresholds are higher at around 34md in order to capture the sharp change in the cdf around that value of permeability (Fig. 4.1).
4.2 Secondary indicator threshold definition

The full indicator algorithm (Sect. 3.2) outlined above utilizes $|\Delta P|$ data as conditioning data.

$$|\Delta P| = |P^{\text{Observed}} - P^{\text{Simulated}}| \quad (4.1)$$

In order to code this information threshold values are required. To keep the analysis more tractable, it would be preferred to specify a single threshold ($|\Delta P|_p$) for secondary data that can yield an optimum update. An ensemble of simulated values of secondary data like well block pressures are generated after a conducting a flow simulation on the ensemble of parameter variables like permeability. With the observed values of well block pressures at the well locations an ensemble of absolute $|\Delta P|$ values (Fig. 4.2) are generated. An optimum threshold value $|\Delta P|_p$ can now be defined using the $|\Delta P|$ value distribution. The idea is to choose $|\Delta P|_p$ as a quantile of the $|\Delta P|$ distribution, such that it yields an updated cdf for permeability that is optimal. The values of $|\Delta P|$ corresponding a model realization represents a measure of distance between the model and reality. If the values of $|\Delta P|$ lesser than or equal to $|\Delta P|_p$ are indicator coded 1 and the ones greater than the threshold are
coded as 0, then the indicator transformed values of $|\Delta P|$ would provide a binary measure of distance between the models and reality (1 being close and 0 being far). These secondary indicator values can now be used as conditioning data in the full indicator algorithm (Sect. 3.2). Conditioning the permeability models to this indicator definition should yield permeability models closer to the reality. The median ($Q_{50}$) of the $|\Delta P|$ distribution is chosen as a threshold $|\Delta P|^p$; it gives a good measure of the most likely deviation of model realizations from reality.

### 4.3 Model Implementation

The ‘Indicator-based Data Assimilation’ (InDA) is applied on a synthetic reservoir example with 2601 grid blocks ($51 \times 51$). A ($51 \times 51$) reference reservoir representing a non-Gaussian permeability field (Fig. 4.3) is generated using unconditional Sequential Indicator Simulation (SISIM) (Journel 2013; Journel and Alabert 1989). This reservoir has 7 well locations (Fig. 4.4) with 3 producers and 4 water injectors. Flow simulation is run on this reservoir to generate reference well block pressure values. Next, five percent of the permeability values are randomly selected (Fig. 4.5)
and an ensemble of 500 models are generated using the 5 percent of the data as conditioning data. These models were also generated using the SISIM method that avoids making the assumption about the permeability distribution being Gaussian. The uncertainty in the oil production response from the flow simulation runs on the initial model realizations (Fig. 4.6a) is represented by a wide spread of ensemble forecast values around the reference values. The reference pressure data at the well locations are used as secondary data to update the permeability values. ‘Indicator-based Data Assimilation’ of pressure data at all 3 producer well locations is performed on a 2 month interval over a total duration of 18 months, resulting in 9 assimilation steps. An uncorrelated Gaussian error with 0 mean and a standard deviation of 10 psia was used for all the observations. The final updated permeability models are used to rerun the entire ensemble of models for a period of 47 months from the initial time (Fig. 4.6b). This gives a forecast of 29 months beyond the assimilation period of 18 months. As can be seen in Fig. 4.6, the proposed ‘Indicator-based Data Assimilation’ (InDA) performs well in assimilating the observed pressure data resulting in a reliable uncertainty evolution.
with time. The final updated models yield a narrower band of oil production values around the reference oil production values (Fig. 4.6b), compared to the initial ensemble of oil production response (Fig. 4.6a).
Figure 4.5 5 percent Permeability data use as conditioning data to generate the ensemble of 500 models

Figure 4.6 Uncertainty in oil production of the ensemble of models and reference production data at the location of well-3(39, 12) (a) Initial Ensemble of models (b) Final ensemble of models updated using InDA
In Chap. 4 the proposed method was successfully applied to a synthetic reservoir which exemplified the benefits of this technique over traditional methods like EnKF. The reservoir used in that example was a synthetic reservoir. In order to better understand the usefulness of the proposed method a realistic reservoir with features replicating fluvial depositional environment is used. The reservoir is called STANFORD-V (Mao and Journel 1999) and is an exhaustive reference data set which was generated for the general purpose of testing any proposed algorithm for reservoir characterization. The reservoir has channel like features with very high permeability values that serve as pay zones (areas where a hydrocarbon fluids can flow easily and hence can be produced easily) while the non-pay zones (areas where a hydrocarbon fluids cannot flow easily and hence cannot be produced easily) are characterized by very low permeability values. As a result the permeability distribution is highly non-Gaussian. This makes it an excellent candidate for the proposed indicator based data assimilation technique, in which the ability to handle non-Gaussian distribution of parameters being estimated is inherently embedded. Additionally as characterization of channel like features requires information of higher order statistics of the spatial permeability distribution, this exercise would also help in bringing out the limitations of two point update schemes which is a common signature of majority of ensemble-based update methods. This would motivate a more generalized multipoint extension of the proposed method.
5.1 Description of reference reservoir

The original ‘STANFORD-V’ reservoir has a total of 10 layers with 100 and 130 grid blocks in the ‘x’ and ‘y’ direction respectively. In this exercise only the top 4 layers are used for simplicity. This results in a reservoir with a total of 52,000(100×130×4) grid cells. Although the InDA technique can efficiently handle updates of 10 layers; the numerical simulations of the ensemble of models involved in the forecast step can be time consuming with all 10 layers. The use of a 4 layer case provides required insight to the problem without any loss of generality. The reference reservoir has channels (pay zones) with permeability (md) value taken from the distribution \( \mathcal{N}(500, 10^2) \), while the non-pay zones are taken from the distribution \( \mathcal{N}(10, 5^2) \). The large differences in the permeability (md) values of pay and non-pay zones are not only realistic in a fluvial depositional environment but also result in a non-Gaussian permeability distribution of the reservoir. Figure 5.1 shows the permeability distribution of the four reference layers used in this exercise. The bottom most layer is annotated as 1 and the number increases as we move upwards. The liquid rate (production/injection) data from the well locations shown in Fig. 5.1 will be used for data assimilation.
Figure 5.1 Permeability distribution of reference model layers of STANFORD-V
5.2 Generation of initial ensemble of models

The ensemble of initial models were generated using the Single Normal Equation Simulation (SNESIM) (Strebelle 2002) provided in the Stanford Geostatistical Modeling Software (SGEMS). The SNESIM algorithm produces realizations which are modeled by multipoint statistics which enables representation of channel-like features that are signature characteristics of a fluvial reservoirs. The algorithm relies on a training images (Maharaja 2008) in order to calculate the higher order statistics. The training image generated for the purposes of this exercise is shown in Fig. 5.3. In order to replicate a realistic scenario, the conditioning data used for the generation of the initial ensemble of 100 permeability models were restricted only to the well locations (Fig. 5.4). The layer wise permeability distribution of an arbitrary model (Ensemble member 23) is shown in Fig. 5.5. The cumulative probability density of permeability values of the entire reference reservoir shows a non-Gaussian characteristic (Fig. 5.6a). The same characteristic is well captured by the initial ensemble of permeability models (Fig. 5.6b).
Figure 5.3 Training image of permeability distribution for all reference model layers
Figure 5.4 Conditioning data locations for generation of initial ensemble of permeability models
Figure 5.5 Permeability distribution of four layers for ensemble member-23
5.3 Assimilation of secondary data to update ensemble of permeability models

As evident from Sect. 5.2 the initial ensemble of models have initial uncertainty associated with them. The liquid rate (production/injection) data from the reference models can be used as a secondary data for updating the ensemble of models. The choice of liquid rate ensures that association of permeability with flow rate is clearly captured. Using oil or water rate only might result in the association to be highly dependent upon saturation and not permeability alone. All the well locations are used for conditioning. The primary data cut-off is defined as shown in Fig. 5.7, it is evident that more cut-off values are provided in the parts of the c.d.f. where there is a curvature change. For the calculation of the secondary thresholds the procedure introduced in Sect. 4.2 will be used. An uncorrelated Gaussian error with 0 STB/day mean and a standard deviation of 100 STB/day is used for all observations. The liquid rate data is assimilated every 2 months for 20 months resulting in a total of 10 permeability update steps.
5.4 Analysis of results

The ensemble of models once updated is then flow simulated for a period of 47 months from the initial time. As the models are calibrated only for the first 20 months (Sect. 5.3) the ability of the models to match future rates for the next 27 months is tested. Comparisons of liquid rate forecasts with respect to reference rate values are made both for the initial ensemble of models and the final InDA based updated models (Fig. 5.8). This brings out the effectiveness of the proposed data assimilation technique. It is evident that the uncertainty in the permeability distribution has evolved reliably as a result of updates, making the ensemble of permeability models closer to the reality. This is clearly visible in Fig. 5.9 where the the final updated model (Fig. 5.9c) is comprised of channel descriptions comparable to the reference model (Fig. 5.9a). The comparison of the initial model (Fig. 5.9b) with the reference model (Fig. 5.9a) on the other hand brings out the difference in
Figure 5.8 Comparison of uncertainty in liquid rate forecast before (a, c, e) and after InDA update (b, d, f)
Figure 5.9 Comparison of initial and InDA updated permeability distribution of layer-3 for ensemble members 23
the initial channel descriptions from reality. In addition to matching liquid rates the indicator based update scheme preserves the initial non-Gaussian distribution of the permeability (Fig. 5.10). This is a direct implication of the threshold wise treatment of the probability update. This approach also results in honoring the range of permeability within desired or realistic limits. The representation of the spatial distribution of permeability in the updated models is reasonably close to the reference distribution as evident from the aforementioned analysis in this paragraph.

5.5 Comparison of indicator-based data assimilation (InDA), ensemble Kalman filter (EnKF) and normal score ensemble Kalman filter (NS-EnKF) results

The proposed indicator-based data assimilation provides generality over non-Gaussian permeability distribution and non-linear forward models both of which offer challenges to data assimilation techniques like ensemble Kalman filter. In this section the results obtained from both update methods are analyzed and compared. Although EnKF handles non-linearity of flow models to an extent if data assimilation exercise is performed after short forecast steps; longer forecast steps may result in growth of the non-linear error introduced by non-linear forward models. This might pose a deterrent to the accuracy of the assimilation process.
In this section the results obtained from EnkF, NS-EnKF and InDA assimilation methods are analyzed and compared.

**Comparison of flow rates:** The comparison of flow rates from InDA and EnKF updates models (Fig. 5.11) suggests that the performance of both methods are comparable with respect to the ability to forecast liquid rates of wells. The same holds true for comparison of flow rates from InDA and NS-EnKF updated models (Fig. 5.12)

**Comparison of ensemble mean:** Noticeable and important differences become evident when the ensemble mean of the updated permeability models from the different methods are compared.

**InDA and EnKF:** Fig. 5.13 shows the comparison of average of all realizations of permeability for layer-3. The initial average map (Fig. 5.13b) shows that the ensemble of permeability models capture the reference permeability field (Fig. 5.13a) moderately. The average of the final EnKF updated models (Fig. 5.13c) do not represent the reference distribution clearly, while the InDA updated models (Fig. 5.13d) closely represent the reference model. Figure 5.14 shows a similar comparison but for a different layer (Layer-1). In this case the initial average map (Fig. 5.14b) shows that the ensemble of permeability models do not capture the reference permeability field (Fig. 5.14a). As before the average of the final EnKF updated models (Fig. 5.14c) do not represent the reference distribution clearly and most permeability updates are concentrated locally, close to the conditioning well location. The InDA updated models (Fig. 5.14d) do not represent the reference models closely but the permeability updates extend along the large-scale spatial variability of the the reference model (approximately in the $45^\circ$ from the positive y-axis). Further the updates appear as straight channel-like features rather than curved channel-like features which is a signature short coming of a two-point updating scheme like EnKF or InDA when applied to initial models generated from from multi-point algorithms using less conditioning data.

**InDA and NS-EnKF:** Fig. 5.15 shows the comparison of average of all realizations of permeability for layer-3. The initial average map (Fig. 5.15b) shows that the ensemble of permeability models capture the reference permeability field (Fig. 5.15a) moderately. The average of the final NS-EnKF updated models (Fig. 5.15c) do represent the reference distribution to a certain extent and is considerably better than the EnKF updates (Fig. 5.13c). The InDA updated models (Fig. 5.15d) also
Figure 5.11 Comparison of uncertainty in liquid rate forecast of wells for EnKF (a, c, e) and InDA updated (b, d, f) models
Figure 5.12 Comparison of uncertainty in liquid rate forecast of wells for NS-EnKF (a, c, e) and InDA updated (b, d, f) models
Figure 5.13 Comparison of initial ensemble average of permeability with EnKF and InDA updated ensemble average of permeability for layer-3
Figure 5.14 Comparison of initial ensemble average of permeability with EnKF and InDA updated ensemble average of permeability for layer-1
Figure 5.15 Comparison of initial ensemble average of permeability with NS-EnKF and InDA updated ensemble average of permeability for layer-3
Figure 5.16 Comparison of initial ensemble average of permeability with NS-EnKF and InDA updated ensemble average of permeability for layer-1
represent the reference model closely. The results from InDA are closely comparable to NS-EnKF in terms of quality. The additional benefit of InDA because of its better treatment of non-linearity can be visualized in its ability to represent the spatial proportions of the channel features slightly better than NS-EnKF. Figure 5.16 shows a similar comparison but for a different layer (Layer-1). In this case the initial average map (Fig. 5.16b) shows that the initial ensemble of permeability models do not capture the reference permeability field (Fig. 5.16a). In this case the average of the final NS-EnKF updated models (Fig. 5.16c) is closely comparable to InDA updated models (Fig. 5.16d) and both do not represent the reference models closely. The permeability updates in both cases extend along the large-scale spatial variability of the the reference model(approximately in the 45° from the positive y-axis). Further the updates appear as straight channel-like features rather than curved channel-like features which is a signature short coming of a two-point updating scheme like NS-EnKF or InDA when applied to initial models generated from from multi-point algorithms using less conditioning data.

**Comparison of ensemble variance:** Differences were also noted when the ensemble variance of the updated permeability models from the different methods are compared.

**InDA and EnKF:** The comparison of the ensemble variance of the permeability values for layer-3 using the two schemes (Fig. 5.17) shows how the uncertainty space evolves with different update techniques. As the permeability distribution of reference and initial ensemble of models are non-Gaussian, the InDA updates of permeability successfully conserve the underlying statistical distribution, on the other hand the EnKF updates fail to conserve the distribution due to the underlying Gaussian assumption of the update equations. Both of these impacts can be easily seen in the variance maps. The updates from EnKF Fig. 5.17b results in considerable uncertainty reduction in isolated pockets while the uncertainty is high, away form these regions. The spatial variation of uncertainty for the InDA (Fig. 5.17c) is roughly along the channel features of the reference reservoir. The variance updates of a different layer (Layer-1)(Fig. 5.18) results in an inference which is coherent with the ensemble average analysis for layer-1. The variance reduction follows the same spatial trend as the mean showing the combined effect of the scarce initial conditioning data with scattered sparse secondary information.
InDA and NS-EnKF: The comparison of the ensemble variance of the permeability values for layer-3 using the two schemes (Fig. 5.19) shows how the uncertainty space evolves with different update techniques. As the permeability distribution of reference and initial ensemble of models are non-Gaussian, both NS-EnKF and InDA updates of uncertainty successfully conserve the underlying statistical distribution and the impact is represented accordingly in the variance maps. The updates from both NS-EnKF Fig. 5.19b and InDA (Fig. 5.19c) is roughly along the channel features of the reference reservoir. Again better treatment of non-linearity by InDA can be visualized in its ability to represent the spatial proportions of uncertainty evolution along channel features slightly better than NS-EnKF. The variance updates of a different layer (Layer-1) (Fig. 5.20) results in an inference which is coherent with the ensemble average analysis for layer-1. The variance reduction follows the same spatial trend as the mean showing the combined effect of the scarce initial conditioning data with scattered sparse secondary information.

**Comparison of cumulative density function:** Finally the comparison of the cumulative density functions of the permeability distributions for the entire ensemble, before and after updating shows the ability of InDA to conserve the non-Gaussian distribution of permeability (Fig. 5.21c), while the EnKF updated permeability realizations have a much smoother Gaussian like distribution (Fig. 5.21a). The negative values yielded by the EnKF updates are replaced by a small positive value to ensure proper functioning of numerical simulation, this causes the distribution to be truncated at the lower end. The threshold wise indicator updates of permeability in InDA leads to updates being correctly constrained within the desired or relevant range of permeability values (Fig. 5.21c), while the EnKF based updates lead to unrealistic range of values (Fig. 5.21a). Although in this case NS-EnKF (Fig. 5.21a) updates preserve the non-Gaussian shape slightly better than InDA, the cdf reproduction can easily be achieved by having higher number of thresholds closer to the modes when applying InDA.

48
Figure 5.17 Comparison of initial ensemble variance of permeability with EnKF and InDA updated ensemble variance of permeability for layer-3
Figure 5.18 Comparison of initial ensemble variance of permeability with EnKF and InDA updated ensemble variance of permeability for layer-1
Figure 5.19 Comparison of initial ensemble variance of permeability with NS-EnKF and InDA updated ensemble variance of permeability for layer-3
Figure 5.20 Comparison of initial ensemble variance of permeability with NS-EnKF and InDA updated ensemble variance of permeability for layer-1
Figure 5.21 Comparison of permeability cumulative density function for the entire ensemble of EnKF, NS-EnKF and InDA updated realizations

5.6 Forecasting future reservoir performance using InDA-based updated models

To test the ability of updated models to predict the future flow responses of new wells and their impacts on the flow responses of the existing wells, four new wells are opened for production. The locations of these new wells are shown in Fig. 5.22. In this section the liquid and oil rates of the ensemble of InDA-based updated models are compared with their respective reference values. To analyze the incremental benefit in ensemble forecasting accuracy of InDA-based updated models, flow responses corresponding to the initial ensemble of models are also
Figure 5.22: Forecast wells along with existing wells

Figure 5.23 shows the liquid rate responses of some existing producer wells. The initial ensemble liquid rate responses show a significant amount of uncertainty when compared to their respective reference rates (Figs. 5.23a, 5.23c and 5.23e). It is clear that the wells show an acceptable liquid rate match in the InDA-based calibration period, which is the time on and before the twenty-month mark on Figs. 5.23b, 5.23d and 5.23f. A spike in the liquid rates can be noticed in some of the wells around the twenty-month time mark, as a flow response to the event of opening new wells in the reservoir. The flow responses after this event also shows an acceptable match with the reference values. This comparison shows that the liquid rate responses of the updated models hold good beyond the assimilation period. This means the changes in the state of the reference reservoir are captured reliably by the ensemble of models.
Figure 5.23 Comparison of uncertainty in liquid rate forecast of existing wells for initial and indicator-based data assimilation updated models
The Fig. 5.24 shows the oil rate responses of the existing producer wells that were discussed in the last paragraph. Analogous to the their corresponding liquid rates the initial ensemble oil rate responses also show a significant amount of uncertainty when compared to their respective reference rates (Figs. 5.24a, 5.24c and 5.24e). Although the ensemble liquid rates were used for the model updates, coherence between the reference and the simulated rates are seen in both calibration (on and before 20 months) and forecasting (beyond 20 months) periods.

Figure 5.25 shows the comparison of the ensemble liquid rates of the updated and initial models, for the new producer wells that were opened after the calibration period. In this case, again an acceptable liquid rate match can be observed. This bolsters the ability of the InDA-based updated models to capture the spatial distribution of the reference reservoir, that in turn results in a reliable liquid rate forecast even for wells that were not used in the model update process. A similar comparison of the oil rates from the new wells Fig. 5.26, shows that the updated models not only forecast liquid rates with accuracy but also have a reliable oil rate response.

In summary InDA-based updated models provide reliable spatial distributions of reservoir properties like permeability, that are close to the real or reference distribution. A realistic reservoir description helps to forecast the future state of the reservoir with a greater accuracy. This also provides an opportunity to study the impacts of possible changes in the state of the reservoir, that might occur in the future. In addition to these mentioned benefits, the ensemble nature of the updates makes it easy to quantify the uncertainty associated with both the reservoir parameters and the flow responses at any given time.
Figure 5.24 Comparison of uncertainty in oil rate forecast of existing wells for initial and indicator-based data assimilation updated models
Figure 5.25 Comparison of uncertainty in liquid rate forecast of new wells for initial and indicator-based data assimilation updated models
Figure 5.26 Comparison of uncertainty in liquid rate forecast of new wells for initial and indicator-based data assimilation updated models
Chapter 6  
Indicator-based data assimilation with localization

Data assimilation methods like ensemble Kalman filter or the indicator-based data assimilation (discussed in Chap. 3), rely on the usage of statistics inferred from an ensemble of models to perform updates. As a result of this dependency, the final updated models are sensitive to the size of the ensemble used in the update process. Although a large ensemble size is preferred for extracting reliable statistics (Yin et al. 2015) it is often impractical to perform numerical simulations for a large number of models because of computational and time-related constraints. Both EnKF and InDA methods, update parameters using the information contained in the observed data and the spatial correlation between the observed data and the update parameter. A simple difference between the observed and the simulated values of the variable is computed and used for data integration (in the case of ensemble Kalman filter) or is transformed to an indicator and subsequently used for data assimilation (in the case of indicator-based data assimilation). Additionally, a measure of spatial correlation between the observed data and the update parameter can be calculated in the form of a spatial covariance (Eq. (2.13)) or indicator covariance (Eq. (3.15)) depending upon the method being used. The suitable covariance is calculated by pooling the data and the parameter variables at their respective locations over the entire ensemble. As a result of this pooling, the accuracy of the updates rely on the size of the ensemble. For practical reasons, smaller ensemble sizes are often used, resulting in limited accuracy of spatial correlations, specially when the lag distance between the data and parameter variables is large. In those cases, spurious correlations result in spurious updates over large lags (Houtekamer and Zhang 2016).
Localization techniques (Houtekamer and Mitchell, 2001; Greybush et al., 2011) can be used in the update steps of data assimilation methods to mitigate this problem. These methods weigh the information obtained from different observational data points based on their respective spatial distances to the location where updates are performed. Because of the localization scheme, the observations at larger distances from the update locations, have smaller contributions to the final updates. In the following sections a comparison of EnKF and InDA methods with and without localization are shown in order to exemplify its benefits.

### 6.1 Localization method used to modify EnKF

In fields like meteorology and oceanography, various kinds of localization techniques have been used (Houtekamer and Mitchell, 2001; Greybush et al., 2011). As mentioned at the beginning of the chapter, localization techniques are used to weigh the information obtained from different observational data points, based on their respective spatial distances to the location where updates are performed. In the case of r-localization for EnKF (Greybush et al. 2011), the weighting is done according to a function, which has the spatial distance between the observed data and the location of the parameter update as its argument. The EnKF update equation, when written for the entire state vector, is shown in Eq. (2.13). For simplicity, we can extract a single component, ’x’ (e.g., x being the permeability value at a certain spatial location). The EnKF update equation for this component without any localization can be written as shown in Eq. (6.1):

\[
x_j^a = x_j^f + \left[ C_{xp1} C_{xp2} \ldots C_{xp_n} \right] \times 
\left[ \begin{array}{ccc}
C_{p1p1} & C_{p1p2} & \ldots & C_{p1p_n} \\
C_{p2p1} & C_{p2p2} & \ldots & C_{p2p_n} \\
\vdots & \vdots & \ddots & \vdots \\
C_{pn1p1} & C_{pn1p2} & \ldots & C_{pn1p_n}
\end{array} \right]
\left[ \begin{array}{ccc}
C_{\epsilon_1\epsilon_1} & 0 & \ldots & 0 \\
0 & C_{\epsilon_2\epsilon_2} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & C_{\epsilon_n\epsilon_n}
\end{array} \right]^{-1}
\left[ \begin{array}{c}
p_{obs1} - p_1 \\
p_{obs2} - p_2 \\
p_{obs3} - p_3 \\
\vdots \\
p_{obsn} - p_n
\end{array} \right]_j,
\]

where \( p_i \forall i = 1, 2, 3, \ldots, n \) are simulated \( p \) values from the \( j^{th} \) ensemble member.

(6.1)
When comparing state vector formulation to the scalar form written for a single variable equation, it should be noted that neither auto-covariance terms for the variable, \( x' \), nor the mapping operator, \( H' \), from Eq. (2.13) appear in Eq. (6.1). The term \( H' \) is a operator that maps the parameter (variable to be updated) space to the data (variable used for data assimilation) space. The process of running a full scale numerical simulation serves the same exact purpose as operator \( H' \), where high-dimensional parameter variables like spatially distributed permeability, are mapped to low-dimensional variables like flow responses at well locations. Simply put, the matrices \( C_{\psi \psi} H^T, H C_{\psi \psi} H^T, H \psi_j \) in Eq. (2.13) are analogous to the matrices \( C_{xp}, C_{pp}, p_j \) in Eq. (6.1). Following, Greybush et al. (2011), the modified EnKF with localization can be applied as shown in Eq. (6.2):

\[
x^*_j = x_j^f + \left[ C_{xp_1} C_{xp_2} ... C_{xp_n} \right] \times \left[ \begin{array}{cccc}
C_{p_1p_1} & C_{p_2p_1} & \cdots & C_{p_np_1} \\
C_{p_1p_2} & C_{p_2p_2} & \cdots & C_{p_np_2} \\
\vdots & \vdots & \ddots & \vdots \\
C_{p_1p_n} & C_{p_2p_n} & \cdots & C_{p_np_n}
\end{array} \right] \left[ \begin{array}{cccc}
f_{x1} C_{\epsilon_1 \epsilon_1} & 0 & \cdots & 0 \\
0 & f_{x2} C_{\epsilon_2 \epsilon_2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & f_{xn} C_{\epsilon_n \epsilon_n}
\end{array} \right]^{-1} \left[ \begin{array}{c}
p_{obs1} - p_1 \\
p_{obs2} - p_2 \\
p_{obs3} - p_3 \\
\vdots \\
p_{obsn} - p_n
\end{array} \right]_j
\]

where \( p_i, \forall i = 1, 2, 3, ..., n \) are simulated \( p \) values from the \( j^{th} \) ensemble member.

In Eq. (6.2), the terms \( f_{xi} \) impose the required localization of the conditioning data \((i = 1, 2, ..., n)\), while updating a parameter \( x \). The definition of terms \( f_{xi} \) are shown in the equation:

\[
f_{xi} = e^{(d_{xi}^2/2L^2)} \forall i = 1, 2, 3, ..., n
\]

where \( d_{xi} \) is the spatial lag between the update variable and the \( i^{th} \) observation variable and \( L \) is a localization parameter.

As evident from Eq. (6.3), the function \( f_{xi} \) is exponential, with increasing values as the distance lag \( d_{xi} \) increases. The values of \( f_{xi} \) act as multipliers to the diagonal elements of the observation error matrix, each of which corresponds to an observed data point. As the aforementioned weighted error matrix appears under the inverse operator in the update equation, it is ensured that the results in highly weighted
data points will exert less influence on the update at that location. Naturally, the
updates with localization are most influenced by observational data points that are
located in close spatial proximity, with $L$ determining the length scales for which
the localization effect is pronounced.

## 6.2 Localization method used to modify InDA

While the EnKF updates are performed on parameters directly, the InDA updates
are performed first on the cumulative density function (cdf) of the parameter and
then transforming the cdf updates to parameter updates as discussed in Chap. 3.
It should be noted that the form of equation for cdf update for a certain parameter
cut-off (Eq. (3.15)) used in InDA is very similar to that of the parameter update
equation (Eq. (6.1)) used in EnKF. This analogy can be used to propose a modified
form of InDA update equation as shown in Eq. (6.4):

$$E(I(\psi^f(u), \psi_k)|Y_1 = 1, Y_2 = 1, Y_3 = 1, ..., Y_n = 1) = E(I(\psi^f(u), \psi_k)) +$$

$$\begin{bmatrix}
    f_{x_1}C_{Y_1Y_1} & C_{Y_1Y_2} & \cdots & C_{Y_1Y_n} \\
    C_{Y_1Y_2} & f_{x_2}C_{Y_2Y_2} & \cdots & C_{Y_2Y_n} \\
    \vdots & \vdots & \ddots & \vdots \\
    C_{Y_1Y_n} & C_{Y_2Y_n} & \cdots & f_{x_n}C_{Y_nY_n}
\end{bmatrix}^{-1}
\begin{bmatrix}
    1 - E(Y_1) \\
    1 - E(Y_2) \\
    \vdots \\
    1 - E(Y_n)
\end{bmatrix}$$

(6.4)

It should be noted that, although analogous, the localization technique used in
InDA (Eq. (6.4)) is not exactly equivalent to that used in EnKF (Eq. (6.2)),
but serves a very similar purpose as demonstrated in later sections. The fact
that the localization techniques are slightly different is because, unlike the EnKF
update equation where the observed data error matrix appears explicitly, the InDA
update equation uses the observation error implicitly. This implicit inclusion of
observation errors into the InDA method is clearly mentioned in Sect. 4.2, where
the secondary indicator definitions are defined using the absolute difference between
the simulated observed variable from the ensemble and the perturbed observed
data. The perturbed observations are obtained by sampling values from a Gaussian
distribution with the observed data as the mean and variance of observation error as
a variance. The number of samples is equal to the size of the ensemble being used;
thus each perturbed observation sample can be paired with a simulated observation variable from the ensemble. The ensemble of absolute difference of these pairs can then be used to define secondary indicator definitions $Y$, which in turn can be used in the update equation (Eq. (6.4)), where a covariance matrix of these indicators appears under the inverse operator. Application of localization to the diagonal elements of this matrix results in updates being influenced more by the observed data that are close to the update location than the ones that are far away.

### 6.3 Determination of the localization parameter $L$

Using localization with either EnKF or InDA (Eqs. (6.2) and (6.4) respectively) involves the localization function given by Eq. (6.3). This equation evaluates a localization weight based on the spatial distance ($d_{xi}$) between the location of update parameter ($x$) and the observational variable ($i$). The additional parameter which is required for this calculation is $L$. Although it is possible to use a suitable value of $L$ based on experience and the spatial dimensions of the reservoir being under consideration, this section introduces a more quantitative method to estimate this value using the ensemble of reservoir model.

The exponential localization function Eq. (6.3) used in this case is sensitive to the value of $L$ being used. The value of the function is 1 for a distance equal to zero ($d_{xi} = 0$) and increases as the distance ($d_{xi}$) increases. Further, because of the exponential nature of the localization function, for the same reservoir and data observations, a very small value of $L$ would result in larger localization function values than when $L$ is large. This in turn means that on one hand, very small values of $L$ may result in observed data having a very small impact on majority of parameter updates in the reservoir, while very large values of $L$ would result in parameter updates over a larger volume of the reservoir, similar to the case without any localization. It is therefore, conjectured that an optimum value of $L$ should be based on the distance beyond which the cross-correlation between observed data and update parameter is negligible.

For EnKF used in this paper, at any assimilation time step, the observed variable ($P$, liquid rate at a well location ($STB$/Day)) is perturbed with an observation error distribution ($\mathcal{N}(0, 100STB$/Day)). Then, $e_n$ number of samples (number equal to the size of the ensemble) of these perturbed samples are then used to
Figure 6.1 Cross-correlation as a function of distance between observational location 2 (well-Inj1-2) and locations in the reservoir.

update $e_n$ parameter values (e.g., $K$, permeability) at each location of the reservoir. For any location where production data is available and any location with an update parameter ($x$), the $e_n$ pairs of parameters and observed values can be used to calculate a cross-correlation value. This calculation can be repeated for all locations in the reservoir, giving an estimate of cross-correlation with respect to distance away from the observed data location. In this example, all the observation data locations are well locations shown in Fig. 5.1. Table 6.1 tabulates these wells with an assigned serial number (first column of Table 6.1) associated with each well. The aforementioned cross-correlation calculations for observational location 2 (well-Inj1-2) is shown in Fig. 6.1. It can be ascertained visually that the correlation decreases with increase in spatial lag from the observational location and levels off some where between 2000 ft and 4000 ft, at which point the values seem to randomly vary around around a central measure of zero correlation, signifying little or no spatial correlation. Thus choosing, a localization parameter $L$ some where between 2000 ft and 4000 ft can be a good approximation for this case. As evaluating these
$L$ values manually for each observational data point can be an arduous process, an automatic method is proposed, that processes the cross-correlation vs. spatial lag information (Fig. 6.1) to estimate a value of $L$.

The first step is to make the data less noisy. Computing the covariance at a lag increment of 82.021 ft (length of grid block in $x$ or $y$ direction), the median cross-correlation corresponding to each lag is considered. To this median relationship, a second-order spline function is fitted and the spatial lag where the derivative of the spine is very close to zero ($< 10^{-5}$) is considered to yield the $L$ value. The aforementioned process is explained in Fig. 6.2, which approximates an $L$ of 2349 ft.

The same exact process can be applied for determination of $L$ for InDA. In case of InDA, at any assimilation time step, the observed variable ($P$, liquid rate at a well location ($STB/Day$)) is perturbed with an observation error, ($\mathcal{N}(0, 100STB/Day)$), as discussed in Sect. 6.6. Then, $e_n$ number of samples (number equal to the size of the ensemble) drawn from this distribution and the indicator transform ($Y$) of their differences from the simulated values of the same variables are then used to update the cumulative density function of the parameter values (e.g., permeability) at each location of the reservoir. Such an update also requires the indicator transformation ($I$) of the parameters to be updated. For an observed indicator variable location, the $e_n$ pairs of parameters and observed values can be used to calculate a cross-covariance value, and the same covariance calculation can be repeated for all locations away from the observed data location. Whereas in the case of EnKF, cross-correlation values are used to avoid data unit-related complications, the cross-covariance can be generally used in case of InDA as indicator transformed values are either zeros or ones and have no associated units. Then, the same approach as in the case of EnKF is used to first find an median cross-covariance vs. lag relationship for a step length of 82.021 ft and then fitting a second order spline to the relationship. Finally, the spatial lag where the derivative of the spine is very close to zero ($< 10^{-6}$) is considered to be the $L$ value. The aforementioned process is explained in Fig. 6.3, which approximates an $L = 2989$ ft.

The indicator transformation technique is explained in detail in the earlier chapters and requires the specification of cut-offs for accomplishing multiple indicator transformations. For the purposes of the calculation of indicator cross-covariance vs. spatial lag relationship, the cut-off value of 250 md is used as it partitions the bi-modal distribution of permeability roughly into non-channel and channel
Figure 6.2 Automatic estimation of $L$ from cross-correlation data as a function of distance away from observational location 2 (well-Inj1-2) used in EnKF permeability. Any other measure, like mean or median, of the prior permeability distribution can also be used for the same purpose. The estimation of $L$ values discussed in the previous section is specific to a observational data point. The $L$ value for another observational data point may be different. For the purpose of data assimilation by either using EnKF and InDA a single value of $L$ is desirable. For calculating a single value of $L$, the arithmetic mean of the $L$ values for all the observational data points is used. However, a new $L$ value is computed for every assimilation step. Table 6.1, shows the length parameters calculated for different observational data points for both EnKF and InDA. The last row of the table also shows the average values of $L$, which are used for data assimilation with localization at the end of the 2-month time period. The spline fitting method used in both the methods to evaluate a localization distance is an approximate method to automatize the calculation of a localization length for any given source of data. This method yields acceptable result in majority of cases but might yield some outliers, like the
Figure 6.3 Automatic estimation of $L$ from cross-variance data as a function of distance away from observational location 2 (well-Inj1-2) used in InDA

length of 11402 ft for observation number 29 (Prod5-2) calculated for InDA. The use of an average value minimizes the impact of such extreme calculations while providing a single measure of localization length that can be used in an assimilation step. A total of 10 assimilation steps are conducted at a 2 month interval and the final average $L$ values used for each of these time steps are shown in Fig. 6.4, which clearly shows a trend where the values of localization lengths are decreasing with time for both EnKF and InDA cases albeit with different rates. The reason for this is attributed to the fact that because of the changing fluid flow regimes in the reservoir with respect to time, the liquid rate responses achieve a stable state after the start of fluid flow in the reservoir at initial time. This is illustrated in (Fig. 6.5a), which shows the derivative of flow rate for all ensemble members with respect to time for observation location 2 (well-Inj-2), and (Fig. 6.5b) the median of these ensemble derivatives for all the observation locations (wells). The reduction of values of derivatives of liquid rates, which are getting closer to zero with time, means, as time passes the liquid rates converge towards a constant
Figure 6.4 L values for different assimilation steps used in EnKF and InDA localization

value, implying that its correlation with permeability values at any location in the reservoir decreases as time increases. As the $L$ values measure the spatial lag across which there is a strong correlation between liquid rates and permeability values, the decreasing strength of correlation with time results in decreasing values of $L$ with time. As far as the different rates of decreasing $L$ values for InDA and EnKF methods are concerned, no specific reason can be attributed to this difference.
<table>
<thead>
<tr>
<th>Observational data point (i)</th>
<th>Well Name</th>
<th>( L_i ) for EnKF (ft)</th>
<th>( L_i ) for InDA (ft)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Inj1-1</td>
<td>646</td>
<td>2746</td>
</tr>
<tr>
<td>2</td>
<td>Inj1-2</td>
<td>2349</td>
<td>2089</td>
</tr>
<tr>
<td>3</td>
<td>Inj1-3</td>
<td>680</td>
<td>3665</td>
</tr>
<tr>
<td>4</td>
<td>Inj1-4</td>
<td>2689</td>
<td>2107</td>
</tr>
<tr>
<td>5</td>
<td>Inj2-1</td>
<td>1107</td>
<td>3121</td>
</tr>
<tr>
<td>6</td>
<td>Inj2-2</td>
<td>1293</td>
<td>5890</td>
</tr>
<tr>
<td>7</td>
<td>Inj2-3</td>
<td>1482</td>
<td>2452</td>
</tr>
<tr>
<td>8</td>
<td>Inj2-4</td>
<td>1287</td>
<td>5504</td>
</tr>
<tr>
<td>9</td>
<td>Inj3-1</td>
<td>3339</td>
<td>3911</td>
</tr>
<tr>
<td>10</td>
<td>Inj3-2</td>
<td>857</td>
<td>4656</td>
</tr>
<tr>
<td>11</td>
<td>Inj3-3</td>
<td>955</td>
<td>8940</td>
</tr>
<tr>
<td>12</td>
<td>Inj3-4</td>
<td>4107</td>
<td>8641</td>
</tr>
<tr>
<td>13</td>
<td>Inj4-1</td>
<td>2340</td>
<td>3342</td>
</tr>
<tr>
<td>14</td>
<td>Inj4-2</td>
<td>859</td>
<td>3640</td>
</tr>
<tr>
<td>15</td>
<td>Prod1-2</td>
<td>1473</td>
<td>5740</td>
</tr>
<tr>
<td>16</td>
<td>Prod1-3</td>
<td>1009</td>
<td>4235</td>
</tr>
<tr>
<td>17</td>
<td>Prod1-4</td>
<td>2644</td>
<td>5348</td>
</tr>
<tr>
<td>18</td>
<td>Prod2-1</td>
<td>402</td>
<td>4411</td>
</tr>
<tr>
<td>19</td>
<td>Prod2-2</td>
<td>741</td>
<td>2221</td>
</tr>
<tr>
<td>20</td>
<td>Prod2-3</td>
<td>361</td>
<td>1761</td>
</tr>
<tr>
<td>21</td>
<td>Prod2-4</td>
<td>343</td>
<td>636</td>
</tr>
<tr>
<td>22</td>
<td>Prod3-2</td>
<td>1977</td>
<td>1226</td>
</tr>
<tr>
<td>23</td>
<td>Prod3-3</td>
<td>2151</td>
<td>8459</td>
</tr>
<tr>
<td>24</td>
<td>Prod3-4</td>
<td>2261</td>
<td>6112</td>
</tr>
<tr>
<td>25</td>
<td>Prod4-1</td>
<td>1199</td>
<td>1606</td>
</tr>
<tr>
<td>26</td>
<td>Prod4-2</td>
<td>2579</td>
<td>2819</td>
</tr>
<tr>
<td>27</td>
<td>Prod4-3</td>
<td>2611</td>
<td>5118</td>
</tr>
<tr>
<td>28</td>
<td>Prod5-1</td>
<td>3431</td>
<td>3571</td>
</tr>
<tr>
<td>29</td>
<td>Prod5-2</td>
<td>3168</td>
<td>11402</td>
</tr>
<tr>
<td>30</td>
<td>Prod5-3</td>
<td>1448</td>
<td>8531</td>
</tr>
<tr>
<td>31</td>
<td>Prod5-4</td>
<td>1440</td>
<td>6727</td>
</tr>
<tr>
<td>32</td>
<td>Prod6-1</td>
<td>2720</td>
<td>3716</td>
</tr>
<tr>
<td>33</td>
<td>Prod6-2</td>
<td>323</td>
<td>7649</td>
</tr>
<tr>
<td>34</td>
<td>Prod7-1</td>
<td>403</td>
<td>493</td>
</tr>
<tr>
<td>35</td>
<td>Prod7-2</td>
<td>2023</td>
<td>2099</td>
</tr>
<tr>
<td>36</td>
<td>Prod8-2</td>
<td>2525</td>
<td>3687</td>
</tr>
</tbody>
</table>

\[
L(\text{ft}) = \sum_i L_i
\]

All values of localization lengths are rounded to the nearest foot.
Figure 6.5 (a) Liquid rate derivative with respect to time for initial ensemble members at well-Inj-2, (b) median of liquid rate derivative with respect to time for initial ensemble members at all well locations.
6.4 Data assimilation with localization

For the demonstration of updates with localization, the same reservoir data and model is used as in Chap. 5. The impact of localization on both EnKF and InDA methods are compared. In this example, the grid block dimensions in $x$ and $y$ directions are equal ($\approx 82\, ft.$), while in the $z$ direction, the dimension varies because of the folded structure of the reservoir. The localization function used for this demonstration has an $L$ value calculated using the methods mentioned in Sect. 6.6. An example of a localization function with $L = 1640\, ft$ (approximately 20 grid block distances in the $x - y$ plane) can be visualized as shown in Fig. 6.6.

![Localization function](image)

**Figure 6.6** Localization function as a function as a function of spatial lag between data and parameter locations
6.5 Comparison of EnKF with and without localization

To appreciate the benefits of localization, this section compares the cases of EnKF updated models and related responses, with and without localization, for the model described above. The comparison of Fig. 6.7c and Fig. 6.7d clearly show the benefits of EnKF with localization, over EnKF without localization. On an average, the use of localization helps in retaining some of the channel-like features extending between wells 'Prod1-3' and 'Prod4-3', present in the initial ensemble of models (Fig. 6.7b). The same holds true for channel-like features extending between wells 'Prod2-3' and 'Inj2-3'. These features are completely lost, in case localization is not used. When a similar comparison is made for another slice of the reference model (Figs. 6.8c and 6.8d) we do not notice an appreciable difference between the localization and non-localization cases. Although some differences are evident with average channel connectivity between wells 'Prod6-1' and 'Prod2-1', and also between 'Inj1-1' and 'Prod4-1', these differences are not as dramatic. These observations suggest, that the accuracy of the final updates are highly impacted by the quality of the initial ensemble of models. The quality in this case of channelized reservoirs can be defined by how well the reference model channels are represented in the initial ensemble of models. For the first slice (Fig. 6.7b), the initial ensemble of models exhibit characteristics close to the reference model, whereas for the second slice (Fig. 6.8b), the initial ensemble does exhibit such characteristics. Accordingly, the incremental benefits of localization within EnKF are substantially more for the first slice than for the second slice.

Another benefit of using localization becomes evident when looking at the ensemble variance map of the updated permeability values (Fig. 6.9). The artificial reduction in variance in regions away from the well locations is a common problem in case of EnKF without any localization (Fig. 6.9b). In regions that are far away from the observation locations, the variance (uncertainty) in parameter values should remain high. Any reduction in the variance is artificial due to the inaccuracies in inferring the covariances for such large lags. This adverse effect is greatly minimized when localization is applied (Fig. 6.9c). It should be noted, however, that the unwanted uncertainty reductions in regions away from the wells are only minimized, but still
Figure 6.7 (a) The reference permeability distribution on one slice of the synthetic reservoir example, (b) the initial ensemble average of permeability, (c) the updated ensemble average of permeability obtained using EnKF without localization, and, (d) the updated ensemble average of permeability obtained using EnKF with localization.
Figure 6.8 (a) The reference permeability distribution on another slice of the synthetic reservoir example, (b) the initial ensemble average of permeability, (c) the updated ensemble average of permeability obtained using EnKF without localization, and, (d) the updated ensemble average of permeability obtained using EnKF with localization
Figure 6.9 (a) The initial ensemble variance of permeability, (b) the updated ensemble variance of permeability obtained using EnKF without localization, and, (c) the updated ensemble variance of permeability obtained using EnKF with localization
persist because, despite the localization, an update is calculated assuming a linear mapping between the data mismatch and the parameter updates. This assumed linear mapping spuriously leads to an underestimation of updated variance. The impact of localization can also be seen when comparing the liquid rate responses for cases with and without localization. The comparison of ensemble liquid rate responses for existing wells like 'Prod4-3' and 'Prod5-3' show, that as a result of artificial reduction in uncertainty/variance underestimation of parameters in the ensemble of models updated using EnKF with localization, the liquid rate responses do not collapse to a narrow spread (Figs. 6.10b and 6.10d). It can be seen that the ensemble spread of liquid rates covers the reference liquid rate responses slightly better than the case without localization (Figs. 6.10a and 6.10c). This observation is especially true in the history matching period (before 20 months). In case of forecast well liquid rates ('forecast1-3'), the reduction in uncertainty at the forecast well location is smaller because of the application of localization (Fig. 6.10f) as expected, when compared to the case without localization (Fig. 6.10e), as evident from the size of the ensemble spread of the liquid rate responses. The updates at the location of the well, in both cases with and without localization do not yield liquid rate responses that are consistently covering the reference liquid rates and are instead more closer to the initial ensemble spread.
Figure 6.10 Comparison of uncertainty in liquid rate prediction for existing and forecast wells for EnKF updates with and without localization
6.6 Comparison of InDA with and without localization

In Sect. 6.5, the comparison of the cases of updated models and related responses, with and without localization brings out the benefits of localization in the context of EnKF. Thus it is worthwhile investigating the benefits of localization when used with InDA. The comparison of Fig. 6.11c and Fig. 6.11d clearly shows, the benefits of InDA with localization over InDA without localization for the first slice through the 3D model. On an average, the use of localization helps in retaining some of the channel-like features extending between wells "Prod1-3" and "Prod4-3" present in the initial ensemble of models Fig. 6.11b which is also true for channel-like features extending between wells "Prod2-3" and "Inj2-3". Additionally, the InDA updates help in generating additional channel feature around "Prod1-3" in the north-south direction. Moreover a channel feature is also generated in the east-west direction connecting "Prod1-3" and "Prod2-3". This feature is especially important because the use of localization results in the generation of a fully connected channel feature between the wells (Fig. 6.11c), whereas without localization the feature appears to be broken (Fig. 6.12d). When a similar comparison is made for the second slice (Figs. 6.12c and 6.12d) we observe that just like in the case of EnKF, the benefit of localization is more marginal. This observation suggests that accuracy of the final updates are highly impacted by the quality of the initial ensemble models. The same observation about the quality of the initial ensemble of models in terms of their proximity to the reference first slice applies.

The benefits of using localization with InDA becomes even more evident when looking at the ensemble variance map of the updated permeability (Fig. 6.13) values. Spurious reduction in variance in regions away from the well locations is a common problem in case of InDA without any localization (Fig. 6.13b). This adverse effect is greatly minimized when localization is applied (Fig. 6.13c). Although the spurious reduction in variance in regions away from the wells is minimized, this reduction is not eliminated. However, it should be noted that the underestimations are greatly minimized (Fig. 6.13c) when compared to EnKF with localization (Fig. 6.13b) because the mismatch between the simulated and observed data are non-linearly mapped to new update value. This non-linear mapping leads to correctly updated
Figure 6.11 (a) The reference permeability distribution on one slice of the synthetic reservoir example, (b) the initial ensemble average of permeability, (c) the updated ensemble average of permeability obtained using InDA without localization, and, (d) the updated ensemble average of permeability obtained using InDA with localization.
Figure 6.12 (a) The reference permeability distribution on another slice of the synthetic reservoir example, (b) the initial ensemble average of permeability, (c) the updated ensemble average of permeability obtained using InDA without localization, and, (d) the updated ensemble average of permeability obtained using InDA with localization.
Figure 6.13 (a) The initial ensemble variance of permeability, (b) the updated ensemble variance of permeability obtained using InDA without localization, and, (c) the updated ensemble variance of permeability obtained using InDA with localization.
variance.
The impact of localization on the updated ensemble of models can be seen clearly when comparing the liquid rate responses from the updated models. The comparison of the ensemble flow responses of existing wells "Prod4-3" and "Prod5-3" show that, as a result of the variance being preserved in areas far away from well locations, the ensemble variance of liquid rate responses does not collapse, and additionally the ensemble spread covers the reference flow responses (Figs. 6.14b and 6.14d) better when compared to the case without localization (Figs. 6.14a and 6.14c). A more visible difference can be seen in case of the forecast ensemble liquid rates, where the updates at the location of the well "forecast1-3", are not only consistent with the reference model but the ensemble spread of flow responses with localization (Fig. 6.14f) also capture the reference flow responses better than the case without localization (Fig. 6.14e).

In summary, the use of localization in both EnKF and InDA prevents the spurious underestimation of uncertainty in parameter values, especially at locations far away from the observation well locations. The retention of this uncertainty prevents the model flow responses from collapsing. The real benefit of localization with InDA is the better preservation of variance at locations far away from the observation wells, because the mismatches between the simulated and observed data are non-linearly mapped to new update values. These benefits result in a robust assessment of uncertainty associated with production forecasts. Thus the models updated using InDA with localization are suitable for predicting flow rate responses for forecast wells, as shown in this chapter.
Figure 6.14 Comparison of uncertainty in liquid rate prediction for existing and forecast wells for InDA updates with and without localization
Chapter 7  
Multiple-point extension to indicator-based data assimilation

Most ensemble-based data assimilation methods including ensemble Kalman filter or indicator-based data assimilation (discussed in Chap. 3 and 5), utilize statistics in the form of covariances to perform updates. The covariances measure the dependence between the attributes at two locations in space. The parameter values at an arbitrary location are updated based on the secondary information observed at some other locations. The covariance values between the primary data and the secondary data are used to derive the weights attached to the data mismatches. In ensemble based methods the covariance of the primary and secondary data is calculated using the ensemble of models. In the case of indicator-based data assimilation the covariance of the indicator transform of the variables is calculated. The underlying assumption is that the multivariate probability distribution characterizing the state space is completely described by the covariance function. This is true only if the multivariate distribution is Gaussian. The spatial distribution of the primary data can be non-Gaussian as a result of the spatial continuity exhibited by channels (high permeability areas where hydrocarbons flow easily) and non-channels (high permeability areas where hydrocarbons do not flow easily). Additionally, the relationship between the primary data and the secondary data can be non-linear. Both these issues can result in sub-optimal EnKF updates. Although the indicator-based data assimilation provides aforementioned generalities, it does not take into account the multivariate interactions between the state variables beyond the covariances. For instance, in case of integrating dynamic response data, the well liquid rates or the well bottom hole pressures are not only
dependent on the permeability values at individual locations but also depend on the higher-order combinations of all permeability values in the reservoir that dictate the flow path for fluids in the reservoir. This leads to the following observations:

- The mismatch between the observed and simulated values should be associated to a pattern of permeability rather than the value at an individual location.

- The update procedure should yield the updated higher-order statistics describing the pattern of variability of permeability.

The above observations motivate the discussion of multipoint/multivariate statistics beyond covariance values. In an indicator framework the required statistics are calculated on the indicator transform of the actual parameter values. In the following sections a method to perform multiple-point data assimilation is proposed.

### 7.1 Proposed multiple-point extension

Although the two-point InDA method provides some generalities for data assimilation, it still relies on the covariance measure to relate the observed data to the parameters being updated. An extension to InDA can be achieved by using the update procedure to yield updates of multiple-point statistics corresponding to pattern templates of permeability variations in the reservoir. As exhaustive calculation of higher-order statistics for an entire reservoir is computationally impractical for most real cases, the calculation of these statistics are restricted only to a few selective templates. In the two-point scheme the cumulative probabilities are updated one location at a time. In the multiple-point counter part, the updates of higher-order statistics will be made only for these selective templates. Then, the updated template statistics can serve as an input to a multiple-point simulation scheme like SNESIM. In the subsequent sections, we will present a method for isolating the most relevant templates in the training image that describe the channel connectivity in the ensemble of models. Updates to the statistics for those selective templates, can be performed using a method like InDA. It should be noted that the ensemble of permeability models generated using the the updated statistics will have different spatial connectivity, than the one generated using the initial TI-based statistics. We will use the updated statistics together with local update
information from two-point update methods like InDA in order to obtain reservoir models that reflect a range of uncertainty consistent with the information in the production data as well reflect the geological continuity of reservoir properties.

Prior to describing this process for updating and reflecting reservoir connectivity, a brief recap of the SNESIM algorithm is presented so that the process of updating the multiple-point statistics using dynamic data can be better understood.

### 7.2 The SNESIM Algorithm for model updates

As mentioned in Strebelle (2002), SNESIM is an algorithm to generate geostatistical models honoring multiple-point statistics of reservoir properties like permeability. In this algorithm, the variable being simulated is annotated as $A$. In the context of a categorical variable simulation, $A$ can take any $K$ number of categories ($A = 1, 2, 3, .., k, .., K$), with prior probabilities $P(A = k)$ or $P_A(k)$ for simplicity. These prior probabilities are often simple proportions from expert opinion or geological sources. The initial ensemble of models is generated using the $P(A)$ from expert opinion. Once the update process starts, this proportion value might undergo slight changes, estimate of $P(A = k)$ after each update step can be calculated by first pooling the parameter values across all the locations in the reservoir for all the ensemble members, and then calculating the proportion for which $A = k$ occurs. The information related to the spatial distribution of the variable is provided in form of a the training image and some conditioning data that are actually recorded values, typically at well locations, both these information are collectively annotated as $B$. Training images capture the plausible ergodic variations of categorical properties across the spatial domain under consideration. The statistics pertaining to the training image are retrieved using a spatial template $T$. The template consists of a central node and neighboring nodes surrounding that central node. For computational efficiency, the training image is scanned once at the beginning of the algorithm and the probability values for all possible combinations of outcomes in the surrounding nodes of the template $T$ as well as the outcome in the central node of the template are recorded and stored in a search tree. Subsequently during simulation, based on the pattern of outcomes in the neighboring nodes of the template $B$, the conditional probability $P(A|B)$ for outcomes in the central node are extracted. All the locations to be simulated in
the image are visited in a random order, considering both previously-simulated and observed values of reservoir variables as conditioning data. A location to be simulated serves as a central node while the previously simulated values and original conditioning data form a pattern around the central node. The probability of all possible categories at the central node \( P(A = k|B) \), where \( k = 1, 2, 3, \ldots, K \)) is extracted from the search tree, and then the category with the highest probability is assigned to the central node. The process is repeated until all locations are simulated.

Within this framework for multiple point simulation, the training statistics for some selective templates are updated in the search tree using the InDA data assimilation procedure. These selective templates are chosen based on their number of repetitions within the training image. The templates appearing more frequently in a training image are considered to have a greater impact on the ensemble of models generated using the training image. For the example case of data assimilation presented in some later sections of this chapter, templates appearing more than 200 times in a training image are considered as selective templates. The ensemble of models is re-simulated using the updated multiple point statistics. Using the ensemble, the updated probability \( P(A|B) \) at any spatial location can be estimated by pooling the categorical values of the variable at that location from the entire ensemble. As mentioned in Sect. 7.1 when the training image statistics are updated in the search tree the resultant simulations honor only the global multiple-point statistics. Thus when \( P(A|B) \) is updated, it does not ensure that the simulations reflect the local flow characteristics in the vicinity of the well. For local uncertainty updates, the integration of secondary flow response data \( C \) is required.

### 7.3 Secondary data integration for MPS updates

As mentioned in Sect. 7.2, the integration of secondary data \( C \) is required to achieve local uncertainty updates with final local probabilities annotated as \( P(A|B, C) \). In SNESIM the probability \( P(A|B, C) \) at any spatial location \( (x, y, z) \) is calculated by integrating \( P(A|B) \) and \( P(A|C) \) by using the tau \( (\tau) \) (Bordley 1982; Journel
2002; Krishnan 2008) model. According to this model:

\[
\frac{x}{a} = \left(\frac{b}{a}\right)^{\tau_1} \left(\frac{c}{a}\right)^{\tau_2},
\]

where:

\[
x = \frac{P(A|B,C)}{1 - P(A|B,C)},
\]

\[
b = \frac{P(A|B)}{1 - P(A|B)},
\]

\[
c = \frac{P(A|C)}{1 - P(A|C)},
\]

\[
a = \frac{P(A)}{1 - P(A)},
\]

\(\tau_1\) and \(\tau_2\) are parameters used in this equation.

In Eq. (7.1), the term \(x\) represents the inverse odds ratio of occurrence of an event \(A\) (occurrence of a variable category \((A = k)\) at a location \((x, y, z)\)) conditioned to the information \(B\) (in training image, multiple-point statistics and conditioning data) and \(C\) (flow response data). The terms \(b\) and \(c\) represent the inverse odds ratio of occurrence of the event \(A = k\) conditional to the data/information \(B\) and \(C\), respectively. The terms \(\tau_1\) and \(\tau_2\) account for the redundancy of information arriving from different data (Krishnan 2008) sources.

It is clear from Eq. (7.1) that secondary data integration at any location using this approach requires \(P(A)\), \(P(A|B)\), \(P(A|C)\), \(\tau_1\), and \(\tau_2\). \(P(A)\) is usually known in the form prior proportion of category \(k\) and the values for \(P(A|B)\) are calculated using the multiple point statistics and the search tree within the SNESIM algorithm (Sect. 7.2), so the only remaining unknown quantities are \(P(A|C)\), \(\tau_1\) and \(\tau_2\). One approach to compute the \(\tau\) values accounting for redundancy of different data sources is implemented within the SNESIM algorithm (Stanford Center for Earth Resources Forecasting 2015) and is given by the following equation:

\[
\tau_1(x, y, z) = \begin{cases} 
\frac{P(A(x,y,z)|B) - P(A)}{1 - P(A)} & \text{if } P(A(x,y,z)|B) \geq P(A), \\
\frac{P(A) - P(A(x,y,z)|B)}{P(A)} & \text{otherwise}
\end{cases}
\]

\[
\tau_2(x, y, z) = \begin{cases} 
\frac{P(A(x,y,z)|C) - P(A)}{1 - P(A)} & \text{if } P(A(x,y,z)|C) \geq P(A), \\
\frac{P(A) - P(C(x,y,z)|B)}{P(A)} & \text{otherwise}
\end{cases}
\]

\(7.2\)
If we take a natural logarithm of the first line in Eq. (7.1) and substitute for \( \tau_2 \) as a function of \( P(A(x, y, z)|C) \) as in Eq. (7.2), the resultant expression is:

\[
\ln \left( \frac{x}{a} \right) = \begin{cases} 
\tau_1 \ln \left( \frac{b}{a} \right) + \left( \frac{P(A(x, y, z)|C) - P(A)}{1 - P(A)} \right) \ln \left( \frac{c}{a} \right), & \text{if } P(A(x, y, z)|C) \geq P(A), \\
\tau_1 \ln \left( \frac{b}{a} \right) + \left( \frac{P(A) - P(A(x, y, z)|C)}{P(A)} \right) \ln \left( \frac{c}{a} \right), & \text{otherwise.}
\end{cases}
\] (7.3)

Further, if we replace \( c \) as a function of \( P(A(x, y, z)|C) \) then the following can be written:

\[
\ln \left( \frac{x}{a} \right) = \begin{cases} 
\tau_1 \ln \left( \frac{b}{a} \right) + \left( \frac{P(A(x, y, z)|C) - P(A)}{1 - P(A)} \right) \ln \left( \frac{P(A(x, y, z)|C)}{a} \right), & \text{if } P(A(x, y, z)|C) \geq P(A), \\
\tau_1 \ln \left( \frac{b}{a} \right) + \left( \frac{P(A) - P(A(x, y, z)|C)}{P(A)} \right) \ln \left( \frac{P(A(x, y, z)|C)}{a} \right), & \text{otherwise.}
\end{cases}
\] (7.4)

If we take all the terms to the L.H.S and write a non-linear objective function \( g \) as shown in Eq. (7.5):

\[
g(P(A(x, y, z)|C)) = \begin{cases} 
\ln \left( \frac{x}{a} \right) - \tau_1 \ln \left( \frac{b}{a} \right) - \left( \frac{P(A(x, y, z)|C) - P(A)}{1 - P(A)} \right) \ln \left( \frac{P(A(x, y, z)|C)}{a} \right), & \text{if } P(A(x, y, z)|C) \geq P(A), \\
\ln \left( \frac{x}{a} \right) - \tau_1 \ln \left( \frac{b}{a} \right) - \left( \frac{P(A) - P(A(x, y, z)|C)}{P(A)} \right) \ln \left( \frac{P(A(x, y, z)|C)}{a} \right), & \text{otherwise},
\end{cases}
\] (7.5)

In a root-finding sense, the objective would be to update guesses of \( \tau_1 \) and \( P(A(x, y, z)|C) \) in order to converge to a value close to 0 for \( g(\cdot) \) (Eq. (7.5)). With an initial guess value for \( P(A(x, y, z)|C)^o \), the converged value of \( P(A(x, y, z)|C) \) can be evaluated using an optimization scheme like Newton-Raphson over the function
Assuming $\tau_1 = \tau_2 = 1$, the $P(A(x, y, z)|C)^o$ values are evaluated from the $c$ values calculated from the first line of the Eq. (7.1). The assumption of $\tau_1 = \tau_2 = 1$ ensures that the initial guess for $P(A(x, y, z)|C)^o$ disregards data redundancy between information sources $B$ and $C$. After this stage, all further calculations are performed with $\tau_1(x, y, z)$ values calculated using first line of Eq. (7.2). The additional variable required for the calculation of $P(A(x, y, z)|C)^o$ is $x$, which is a function of $P(A(x, y, z)|B, C)$. The $P(A(x, y, z)|B, C)$ values calculated from two-point InDA updates are a good estimate of the actual $P(A(x, y, z)|B, C)$ values, and are used for the calculation $P(A(x, y, z)|C)^o$. Final converged values of $P(A(x, y, z)|C)$ are then used to evaluate final $\tau_2(x, y, z)$ values. For secondary data integration in SNESIM the $P(A(x, y, z)|C)$ values are required at all spatial locations $(x, y, z)$ and can be readily provided in form of the converged $P(A(x, y, z)|C)$ values. The SNESIM algorithm also requires scalar values of $\tau_1$ and $\tau_2$ values that are representative of all reservoir locations. The application of Eq. (7.2) results in a different value of $\tau_1$ and $\tau_2$ at each location $(x, y, z)$ in the reservoir. Average values of $\tau_1(x, y, z)$ and $\tau_2(x, y, z)$ respectively are calculated over the entire spatial domain $(x, y, z)$ and specified as the unique $\tau$ values for the entire simulation. The ensemble of models thus generated have updated global multiple-point statistics in form of the updated selective template statistics and reduced local uncertainty because of secondary data integration.

In summary, the two-point InDA-Loc provides approximate values of local probabilities for occurrence of a certain parameter category ($k$) at a given location ($P(A(x, y, z) = k|C)$) conditioned to the secondary data ($C$) and its redundancy ($\tau_1$ and $\tau_2$) to other existing information sources ($B$). The aforementioned probability and redundancy information is then combined using the SNESIM algorithm. A flow chart of data assimilation using InDA-MPS is shown in Fig. 7.1.
Start

Generate an initial ensemble of models

Run ensemble flow simulation until all observed secondary data are available

Perform InDA with localization update to get approximate $P(A(x, y, z)|B, C)$ and $x$ values (Eq. (7.1)). Also perform InDA update of training image statistics

Assume $\tau_1^o = \tau_2^o = 1$ to evaluate $P(A(x, y, z)|C)^o$ using tau model (Eq. (7.1))

Calculate $\tau_1(x, y, z)$ using Eq. (7.2)

Using $x$ in Eq. (7.1), $P(A(x, y, z)|C)^o$ and $\tau_1(x, y, z)$ calculate converged value of $P(A(x, y, z)|C)$ using Eq. (7.5)

With converged $P(A(x, y, z)|C)$ calculate $\tau_2(x, y, z)$ using Eq. (7.2)

Calculate $\tau_1 = \bar{\tau}_1(x, y, z)$ and $\tau_2 = \bar{\tau}_2(x, y, z)$

Generate new ensemble of models using initial conditioning data, updated training image statistics and soft data in form of $\tau_1$, $\tau_2$ and $P(A(x, y, z)|C))$

Figure 7.1 Flowchart of InDA-MPS update method.
7.4 A realistic example demonstrating InDA-Multiple point statistic update

The indicator-based data assimilation-multiple-point statistics (InDA-MPS) method has been explained in detail in the previous sections. In this section, an example of this data assimilation method is demonstrated with a realistic working example. The InDA-MPS requires $P(A|C)$ values for the generation of MPS updates (Sect. 7.3), and the $P(A|C)$ calculations in turn require InDA (with localization) updates. The localization parameter value is determined using the method described in Sect. 6.3.

7.4.1 Description of reference reservoir

This exercise uses only the third layer of the 4-layer model used in Chap. 5. This results in a reservoir with a total of 13,000 ($100 \times 130 \times 1$) grid cells. The reference reservoir has pay zones (high permeability areas where hydrocarbons can flow easily) with permeability (in millidarcies) values taken from the distribution $\mathcal{N}(500, 10^2)$, while the non-pay zones (low permeability areas where hydrocarbons cannot flow easily) are taken from the distribution $\mathcal{N}(10, 5^2)$, which is, again, same as the one used in Chap. 5. Figure 7.2 shows the reference permeability distribution of the model used in this exercise. The liquid rate (production/injection) data from the well locations shown in Fig. 7.2 are used for data assimilation.
7.4.2 Generation of initial ensemble of models

The ensemble of initial models is generated using the single normal equation simulation (SNESIM) (Strebelle 2002) available in GSLIB format (Stanford Center for Earth Resources Forecasting 2015). The SNESIM algorithm produces realizations that are modeled by multiple-point statistics which enables representation of channel-like features that are signature characteristics of a fluvial reservoir. The algorithm relies on a training image (Maharaja 2008) to calculate the higher order statistics. The training image generated for the purposes of this exercise is shown in Fig. 7.3a. In order to replicate a realistic scenario, the conditioning data used for the generation of the initial ensemble of 100 permeability models were restricted to the well locations for pay category with additional six non-pay locations which are assumed to be dry well locations (Fig. 7.3b). The permeability distribution of arbitrary models (ensemble members 10, 20, 50 and the reference permeability distribution) are shown in Fig. 7.4. The cumulative probability density of permeability values of the entire reference reservoir shows a non-Gaussian characteristic (Fig. 7.5a). The same characteristic is well captured by the initial ensemble of permeability models (Fig. 7.5b).
Figure 7.3 Training image and conditioning data locations for generation of initial ensemble of permeability models
Figure 7.4 Permeability distribution of initial ensemble member models: 10, 20, 50 and reference permeability distribution
Figure 7.5 Cumulative probability density function for reference permeability model and initial ensemble of models
7.4.3 Assimilation of secondary data to update ensemble of permeability models

The initial ensemble of permeability models represents the initial uncertainty associated with the prediction of permeability in the reservoir. The liquid rate (production/injection) data from the reference model can be used as a secondary data for updating the ensemble of models. The primary data cut-off is defined as shown in Fig. 7.8, it is evident that more cut-off values are provided in the parts of the cdf where there is a curvature change. For calculation of the secondary thresholds the procedure introduced in Sect. 4.2 is used. An uncorrelated Gaussian error with 0 STB/day mean and a standard deviation of 300 STB/day is used for all observations. The liquid rate data is assimilated every 2 months for 14 months, resulting in a total of 7 permeability update steps. As discussed in Sect. 7.3 each of these assimilation steps requires calculations for spatial distribution of $P(A|C)$ values and average $\tau$ values. The plot of $\tau_1$ and $\tau_2$ with respect to time is shown in (Fig. 7.6), where $\tau_1$ describes the data redundancy from the sources, like the training image and conditioning data ($B$), with other data sources towards the occurrence of any given category $k$ for the parameter variable $A$. The information from conditioning data is constant throughout the assimilation process but the information contained in the training image is updated at each assimilation step by updating the global training image statistics (Fig. 7.1). However the statistics of only a selective number of templates are updated which only changes the training image statistics by a small degree, as indicated by the $\tau_1$ value almost constant with time. The $\tau_2$ describes the data redundancy from the sources like liquid rate responses ($C$) with other data sources towards the occurrence of any given category $k$ for the parameter variable $A$. As the information contained in $C$ is changing throughout the history matching process, the $\tau_2$ values also change. The values increase with time at a decreasing rate, this means that the additional information contained in $C$ (liquid rate responses) with increasing number of assimilation is becoming smaller with each assimilation step. For the assimilation step of month-2, the maps of $P(A = \text{pay} - \text{zone}|B)$, $P(A = \text{pay} - \text{zone}|C)$ and the combined probability of $P(A = \text{pay} - \text{zone}|B, C)$ is shown in (Fig. 7.7), which clearly shows how the combined probability information weighs the two information sources ($B$ and $C$) with their respective $\tau$ values for that time step. The updated models...
are then run from the initial time to the next time step where observed data is available for assimilation.

### 7.4.4 Analysis of results

The ensemble of models, once updated, are then flow simulated for a period of 47 months from the initial time. As the models are calibrated only for the first 14 months (Sect. 7.4.3) the ability of the models to match future rates for the next 33 months is tested. In this section the smoother approach is used to employ the InDA-MPS method, thus after each update of permeability and porosity, the entire ensemble of models is rerun from the initial time step to the next time step, where data is available for assimilation. Comparisons of liquid rate forecasts with respect to reference rate values are made both for the initial ensemble of models and the final InDA-MPS (Smoother) updated models (Fig. 7.9). These comparison bring
Figure 7.7 (a) $P(A=\text{pay-zone}|B)$ (b) $P(A=\text{pay-zone}|C)$ and (c) $P(A=\text{pay-zone}|B, C)$ for assimilation step of month-2
out the effectiveness of the proposed data assimilation technique. It is evident that post assimilation, the updated models predict the liquid rate response to be very close to the reference.

The impact of the update process on the spatial distribution of permeability is clearly seen in Fig. 7.10 where the the final updated model (Figs. 7.10a to 7.10c) comprises of channel descriptions comparable to the reference model (Fig. 7.10d). The channels in the initial models (Figs. 7.4a to 7.4c) on the other hand are far from the reference model (Fig. 7.4d). In addition to matching liquid rates the InDA-MPS (Smooother) update scheme preserves the initial non-Gaussian distribution of the permeability (Fig. 7.11). This preservation is a direct implication of the indicator coding of the data and the subsequent probability update. This approach also results in honoring the range of permeability within desired or realistic limits.
Figure 7.9 Comparison of uncertainty in liquid rate forecast before (a, c, e) and after InDA-MPS (Smother) update (b, d, f) (wells under consideration are shown in the inset of (b))
Figure 7.10 Permeability distribution of InDA-MPS (Smother) updated ensemble members models 10, 20, 50 and reference permeability distribution
As discussed before in this chapter, the InDA-MPS method, updates the global training image statistics and the local permeability values. The update of training image statistics is a unique feature of the InDA-MPS method, which ensures that the initial uncertainty in the training image is addressed during the update process. Some templates (Fig. 7.12) are selected from all the possible templates in the search tree by using the criteria mentioned in Sect. 7.2. The maximum template size used in the example discussed in this chapter is 26, but as evident from Fig. 7.12 some templates of size 30 also appear. This is because the, "Multiple Grid Implementation' within SNESIM (Strebelle 2002), was used in this example. This implementation, under certain circumstances, allows template sizes slightly larger than the maximum template size, with different possible configurations of larger templates which is clearly evident in two different templates with size 30 (Figs 7.12c,7.12d and Figs 7.12e,7.12f). The frequency of occurrence of pay or non-pay categories at the center of these templates is then updated using InDA (Fig. 7.13). These new frequencies are then used to update the search tree, which is then used as an input to the SNESIM algorithm along with additional local conditioning information to generate a new set of ensemble (Sect. 7.3). Figure 7.14 shows the comparison of the initial permeability realizations with the realizations generated using only the updated training image statistics (after 7 update steps). This gives an idea about the contribution of global multiple-point statistics update alone towards the description of final models. As can be seen from the figure, the updating of the multiple point statistics results in better depiction of the curvilinear characteristics of the channel (especially in the regions highlighted in Fig. 7.14). However, the updated regions are not localized around the wells where the production information is available.
Figure 7.11 Comparison of c.d.f. of permeability for entire ensemble before and after InDA-MPS (Smooother) update
Figure 7.12 Selective templates used to update global multiple-point statistics
Figure 7.13 Initial and updated non-pay and pay category frequencies at the center of selected templates
Figure 7.14 Initial and only TI-statistics updated ensemble members models 20 (a and b), 50 (c and d)
7.4.5 Comparison of results from InDA-MPS (Smoother), InDA with localization (InDA-Loc) and EnKF with localization (EnKF-Loc)

The development of InDA-MPS was motivated by the additional benefits of including multiple-point statistics into data assimilation methods. This section investigates these benefits by comparing the models and the responses from these models, generated using different data assimilation methods like InDA-MPS, InDA using two-point covariance with localization (InDA-Loc) and EnKF with localization (EnKF-Loc). As is the case in Sect. 7.4, the two-point update encoded within InDA-MPS is InDA with localization which uses the localization implementation discussed in Chap. 6. The update is implemented in a smoother sense, meaning, after each update of model parameter like permeability the updated models are run from initial time step to the next update time step. Hence the method is named InDA-MPS (Smoother). The other methods which are being compared to InDA-MPS (Smoother) are InDA with localization (InDA-Loc) and EnKF with localization (EnKF-Loc), both using the localization method discussed in Chap. 6. Both these algorithms work like a filter, meaning each update step involves the update of both model parameter and state variables and models are run from one update step to the next update step.

Comparison of flow rates: The comparison of flow rates from InDA-MPS (Smoother) and InDA-Loc updates models (Fig. 7.15) suggests that the performance of both methods are comparable with respect to the ability to forecast liquid rates of wells. The same holds true for comparison of flow rates from InDA-MPS (Smoother) and EnKF-Loc updated models (Fig. 7.16).
Figure 7.15 Comparison of uncertainty in liquid rate forecast of wells for InDA-Loc (a, c, e) and InDA-MPS (Smotherer) updated (b, d, f) models (wells under consideration are shown in the inset of (b))
Figure 7.16 Comparison of uncertainty in liquid rate forecast of wells for EnKF-Loc (a, c, e) and InDA-MPS (Smooother) updated (b, d, f) models (wells under consideration are shown in the inset of (b))
Comparison of ensemble mean: Noticeable and important differences become evident when the ensemble mean of the updated permeability models from the different methods are compared. Fig. 7.17 shows the comparison of average of all realizations of permeability obtained by InDA-MPS(Smoother) compared to those obtained by InDA-Loc. The initial average map (Fig. 7.17b) shows that the ensemble of permeability models capture the variability exhibited by reference permeability field (Fig. 7.17a) only approximately. The average of the final InDA-Loc updated models (Fig. 7.17c) exhibit spatial variability more consistent with the reference model, but the continuity of high-permeability is disrupted because of the two-point covariance used for the update in InDA-Loc. The InDA-MPS (Smoother) updated models (Fig. 7.17d) represent the continuity of permeability exhibited by the reference most accurately.

Fig. 7.18 shows the comparison of average of all updated permeability models obtained using InDA-MPS (Smoother) against those obtained using EnKF-Loc. The average of the final EnKF-Loc updated models (Fig. 7.18c) does not represent the reference distribution as accurately as either InDA-MPS or InDA-Loc which is clearly evident by the poor connectivity of high permeability between wells "Inj3-3" and 'Prod3-3', and 'Inj3-3' and 'Prod2-3'. The connectivity in this region is better represented by InDA-MPS and InDA-Loc (Figs. 7.17c and 7.17d respectively). In fact the ensemble average of the EnKF-Loc models does not exhibit the crisp contrast between high and low permeability values that characterizes the non-Gaussian reference. This difference in contrast is a result of the inherent Gaussian assumption in EnKF. The InDA-MPS (Smoother) updated models (Fig. 7.18d) not only captures the reference distribution of permeability very well, but also represents the connectivity of the permeability field accurately.
Figure 7.17 Comparison of initial ensemble average (b) of permeability with InDA-Loc (c) and InDA-MPS (Smother) (d) updated ensemble average of permeability
Figure 7.18 Comparison of initial ensemble average (b) of permeability with EnKF-Loc (c) and InDA-MPS (Smother) (d) updated ensemble average of permeability
Comparison of ensemble variance: Additional important differences are noticed when the ensemble variance of the updated permeability models from the different methods is compared (Fig. 7.19). The comparison of the initial variance map of permeability (Fig. 7.19a) with the InDA-MPS updated variance map (Fig. 7.19b) reveals that the uncertainty associated with permeability predictions is reduced after data assimilation. When compared to the reference model (Fig. 7.17a), the low uncertainty areas show similar spatial extent and connectivity to the permeability channel distribution in the reservoir which can be attributed directly to the MPS based updates coupled with the local conditioning brought about by $P(A|C)$ in the InDA-MPS method. It should also be noted that, despite the reduction of uncertainty in the bulk of the channel structures the higher uncertainty still exists at the channel edges is evident in Fig. 7.19b. The knowledge of this lack of confidence of the exact channel edge locations is beneficial, as it serves as a possible warning against drilling new wells at these locations.

Another important observation is the reduction in uncertainty in areas away from the well locations, primarily in the south-east corner of the map, the corresponding updated values of permeability in this region primarily comprises of low permeability values (Fig. 7.18d), which is in accordance with the reference values (Fig. 7.18a) in the region. The reason behind the aforementioned uncertainty reduction lies in the inclusion of local probability values in form of $P(A|C)$. As mentioned in Sect. 7.3, the InDA-Loc is used to approximate the values of $P(A|C)$, that are required for the SNESIM simulation step involved in the InDA-MPS update. As the region under consideration is far from any data observation source (well), the InDA-Loc results in $P(A|C)$ values that are closer to $P(A)$ values as a result of localization embedded in the InDA-Loc algorithm. The proportions for non-pay ($P(A = non\,-\,pay)$) and pay ($P(A = pay)$) used in this example are 0.64 and 0.36 respectively which means that $P(A|C)$ values at any of these locations under consideration are more likely to be higher for the non-pay category than the pay category. When this information is integrated into the SNESIM algorithm in form of secondary data the non-pay category is assigned to all these locations under consideration resulting in the prevalence of non-pay category with high confidence, as can be seen in the final updated ensemble average and variance maps of permeability (Figs. 7.18b and 7.19b respectively).

In case of InDA-Loc the regions located away from any secondary data sources
undergo small or no updates; as a result of this aforementioned nature of updates the uncertainty reduction in noticed in the areas closer to the well locations and following channel features (Fig. 7.19c), that are consistent with the reference values of permeability (Fig. 7.17a). Although EnKF-Loc employs the same strategy of small or no updates at distances away from secondary data sources, the mismatch between the simulated and observed data are linearly mapped to new update values because of the Gaussian and linear assumptions involved in EnKF. This linear mapping leads to low updated variance very close to the well locations and a smoothly varying reduced variance in regions away from the well that are impacted by the update process (Fig. 7.18d).

The various comparisons made in this chapter bring out the benefits of InDA-MPS over two point methods like InDA-Loc and EnKF-Loc. The updates based on multiple-point statistics yield well-defined channel structures in the updated models better than the two-point update methods. Despite starting from initial ensemble of models that do not capture the reference distributions quite well the final ensemble of models result in reservoir models closely representing the reference models, which have well defined channel structures with crisp boundaries between high and low permeability regions.

While InDA-Loc updated models provide a good average description of channels, the boundaries of these structures are not well defined as in case of InDA-MPS. The lack of clear definition of boundaries is because the updates are performed using two-point statistics like indicator-covariances. In case of EnKF-Loc the channel descriptions are inferior than both InDA-Loc and InDA-MPS as its ability to generate new channel connectivity consistent with the reference model is not as good as the other two aforementioned methods. Although the channel bodies of the InDA-MPS updated models have reduced uncertainty, the edges of the well-defined channel features retain high uncertainty. This high uncertainty is beneficial as it provides a high confidence (low uncertainty) path way in the reservoir where new wells can be drilled, while at the same time having a low confidence (high uncertainty) edge regions which should be avoided for new exploration purposes or can also be used as targets for delineation wells. Finally, as a result of InDA-MPS updated models that are consistent with the reference model and an uncertainty in model distribution that is reliable, the ensemble flow responses from the wells can be used with high confidence for future forecasting purposes.
Figure 7.19 Comparison of initial variance of permeability with InDA-MPS (Smotherer) (b), InDA-Loc (c) and EnKF-Loc (d) updated ensemble variance of permeability
Chapter 8  |  Conclusions

Traditional EnKF has the capability to sequentially assimilate dynamic data in reservoir models. However, its performance may be suboptimal if used in cases where modeling parameters are non-Gaussian or the forward flow models are non-linear. As a result its applicability is limited to extremely specific and simple cases. Most of the real situations where we need to apply data assimilation are significantly complicated and may involve complex geologic features that exhibit non-Gaussian, multiple point spatial characteristics. These complexities might render the application of EnKF to be unreliable. Although there are certain techniques like Particle Filters which try to tackle the above problem, they still suffer from constraints that often results in their application being impractical. The proposed research provides generality over existing ensemble-based data assimilation algorithms like EnKF. The updates honor the prior model for uncertainty of parameters which has the advantage of ensuring that the updated models are geologically consistent, while at the same time the posterior suite of models exhibits the residual uncertainty reliably. Mapping of parameters into indicator space provides computational efficiency. The proposed method is called the indicator-based data assimilation (InDA) and preserves the underlying statistics of the data distribution allowing for data to be assimilated sequentially without having to run the models from the initial time for each update.

As InDA relies on the application of two-point statistics for performing updates, it has been shown that it suffers from issues related to the use of limited ensemble size for the inference of these statistics. A modified version of InDA with localization is proposed to handle the adverse impacts of limited ensemble size better. The localization method used with InDA requires the estimation of a correlation length parameter and an automatic method has been proposed for its
estimation. The application of InDA with localization to assimilate liquid rate data into a realistic reservoir shows its ability to produce calibrated models that can provide reliable reservoir forecasting and uncertainty quantification using a limited ensemble size. The proposed method prevents the underestimation of updated variable uncertainty, which prevents ensemble collapse and also preserves marginal non-Gaussian distribution of the variables being updated. The application of the method to assimilate liquid rate data into a realistic reservoir shows its ability to produce calibrated models that can provide reliable reservoir forecasting and uncertainty quantification using a limited ensemble size. In its current formulation the proposed method preserves marginal non-Gaussian distribution of the variables being updated.

Finally, a multiple-point statistics (MPS) based update method (InDA-MPS), is also proposed. The method relies on the use of secondary data to update the global multiple point statistics and local parameter uncertainty. These updates are used to generate new ensemble of models. The application of this method on a channel reservoir shows its ability to perform MPS updates and reproduce channel features in the updated models, that are closer to the reference model.

Despite the superior quality of updated models generated by InDA-MPS the major shortcoming of this method is the smoother approach that is used. The smoother approach essentially means that after each update step for parameter variables the models are required to be run from initial time step to the next update step. This approach imposes a large cost in terms of computation and time. The problem arises from the fact that a training image is used for updating a parameter like permeability at each assimilation step. In InDA-MPS a single training image is used for parameter variables like permeability and statistics for a training image undergo small updates at each assimilation step. This approach is practically possible and justified in the case of a variable like permeability for which training images are readily available. The problem with multiple-phase fluid systems like hydrocarbon reservoirs is that it also requires the update of state variables like pressure and saturation. With no available training images for the aforementioned state variables the smoother approach ensures that for each ensemble update of parameters a consistent set of updated state variables are generated by simply running a full-scale non-linear numerical simulation from initial time step using the latest ensemble of parameters. Future research is required to use the update
method with a filter approach where updated models are required to be run, simply from one update step to the next. One possible way to conduct a filter based update could be to simply use two-point update schemes like InDA and EnKF for the state variables while using a multiple-point update for the parameter variables. Another possible approach could be to extract all the required multiple-point statistics from the ensemble of state variables at a given assimilation step and use it to perform multiple-point updates for both parameter and state variables. Although these proposed methods might pose challenges of their own, it will they also provide exciting avenues for research in the field of generalized data assimilation techniques.
References


Vita
Devesh Kumar

Devesh Kumar was born in a small town in Jharkhand, a state in eastern India. He graduated with a bachelor’s degree in Petroleum Engineering from the Indian Institute of Technology (Indian School of Mines), Dhanbad in June of 2014. After graduating from college, he worked as a consultant Reservoir Engineer for Schlumberger at their Navi-Mumbai location in India, until July of 2015. In August of 2015, he joined the Department of Energy and Mineral Engineering at the Pennsylvania State University to pursue a doctoral degree in Petroleum and Natural Gas Engineering. He studied the usage of generalized data assimilation tools and it’s possible applications in history matching of hydrocarbon reservoirs. He worked on the development of ensemble methods for data assimilation that do not require assumptions of Gaussian distribution for the state variables or the linear mapping of innovations terms to the state variables being updated. He also developed data assimilation methods which perform spatially multivariate updates of state vectors. He also completed a PhD minor in Computational Science in the fall of 2017 for which he took relevant courses including courses in statistical computing and high-performance computing. Throughout the course of his graduate education, he has been advised by Prof. Sanjay Srinivasan.