Math Geosci (2017) 49:679–715
DOI 10.1007/s11004-016-9672-8

A New Data-Space Inversion Procedure for Efficient Uncertainty Quantification in Subsurface Flow Problems

Wenyue Sun[1] · **Louis J. Durlofsky**[1]

Received: 27 June 2016 / Accepted: 27 December 2016 / Published online: 30 January 2017

Abstract Uncertainty quantification for subsurface flow problems is typically accomplished through model-based inversion procedures in which multiple posterior (history-matched) geological models are generated and used for flow predictions. These procedures can be demanding computationally, however, and it is not always straightforward to maintain geological realism in the resulting history-matched models. In some applications, it is the flow predictions themselves (and the uncertainty associated with these predictions), rather than the posterior geological models, that are of primary interest. This is the motivation for the data-space inversion (DSI) procedure developed in this paper. In the DSI approach, an ensemble of prior model realizations, honoring prior geostatistical information and hard data at wells, are generated and then (flow) simulated. The resulting production data are assembled into data vectors that represent prior 'realizations' in the data space. Pattern-based mapping operations and principal component analysis are applied to transform non-Gaussian data variables into lower-dimensional variables that are closer to multivariate Gaussian. The data-space inversion is posed within a Bayesian framework, and a data-space randomized maximum likelihood method is introduced to sample the conditional distribution of data variables given observed data. Extensive numerical results are presented for two example cases involving oil–water flow in a bimodal channelized system and oil–water–gas flow in a Gaussian permeability system. For both cases, DSI results for uncertainty quantification (e.g., P10, P50, P90 posterior predictions) are compared with those obtained from a strict rejection sampling (RS) procedure. Close agreement

✉ Wenyue Sun
wenyue@stanford.edu

Louis J. Durlofsky
lou@stanford.edu

[1] Department of Energy Resources Engineering, Stanford University, Stanford, CA 94305-2220, USA

between the DSI and RS results is consistently achieved, even when the (synthetic) true data to be matched fall near the edge of the prior distribution. Computational savings using DSI are very substantial in that RS requires $O(10^5$–$10^6)$ flow simulations, in contrast to 500 for DSI, for the cases considered.

Keywords Data-space inversion · Uncertainty quantification · History matching · Model-inversion · Data assimilation · Subsurface flow · Reservoir simulation

1 Introduction

Reservoir performance forecasting is a key component in the management of oil and gas assets. Forecasts are typically generated by solving a forward flow problem for some number of possible reservoir/geological models. Because geological parameters such as permeability are highly uncertain, production (flow) data are used to infer possible values for these parameters. This process is referred to as inverse modeling, data assimilation or, in the context of reservoir simulation, history matching. Although there is a very wide body of literature on the development and application of inverse modeling for subsurface flow problems, the resulting procedures remain computationally expensive, and not entirely robust, for realistic systems with a large number of unknown parameters.

The goal in this paper is to introduce a new data-space inversion (DSI) procedure to efficiently generate multiple forecasts, within a Bayesian framework, which are conditioned to flow-based observations. In the DSI method, the quantities of interest, which in this study are water injection rates and production rates for each fluid phase, are treated as random data variables with a prior probability density function (PDF). This prior PDF is estimated from data forecasts (i.e., flow simulations) performed for an ensemble of prior model realizations. In the examples here, 500 prior flow simulations are used, though these can all be performed at once, in parallel. Reservoir predictions are then generated by directly sampling the posterior PDF of data variables given observed data. This data-space sampling is very inexpensive computationally, which enables the efficient generation of multiple forecasts as required for uncertainty quantification.

Most previous data-assimilation approaches entail model-based inversion, where the goal is to find geological models which, when used as input to a flow simulator, provide flow results in agreement with observed data. A model in this context can be viewed as the porosity and permeability values for each block in a simulation grid. Because the DSI methodology is not a model-based method, the discussion of this extensive literature will be fairly brief. A comprehensive description of general inverse modeling is provided by Tarantola (2005). Data assimilation for history-matching oil/gas reservoirs is described in the book by Oliver et al. (2008) and, more recently, in the review paper by Oliver and Chen (2011). A number of different model-based formulations have been proposed. Gradient-based approaches (Sarma et al. 2006; Gao and Reynolds 2006) seek to obtain models by minimizing an objective function that includes data and model mismatch terms (the latter is essentially a regularization). Adjoint procedures are typically applied to construct the required gradients. Sampling-

based approaches (Mosegaard and Tarantola 1995; Park et al. 2013) consider a very large set of prior models that honor all available prior information and then retain, with higher probability, those models that best fit the observations.

Ensemble-based approaches such as ensemble Kalman filtering (Evensen 2003; Aanonsen et al. 2009) and ensemble smoothers (Evensen and van Leeuwen 2000; Emerick and Reynolds 2013) use a set of models, which are updated to improve the data match when new observations are available. There are some similarities between the DSI formulation and ensemble smoothers. For example, in both approaches, an initial ensemble of models and associated forecasts are needed to estimate the distribution of data and/or model variables. However, by inferring data variables directly, and only in the data space, DSI avoids the challenging model calibration step that arises with ensemble smoothers (as well as with other history-matching approaches) when the relationship between the model and data variables is highly nonlinear.

A number of techniques have been developed to generate forecasts without requiring posterior physical models (posterior models are models that provide simulation results that fit the observations). Artificial intelligence approaches, such as artificial neural networks, have been applied for reservoir forecasting (Mohaghegh 2005). These methods focus more on prediction accuracy rather than uncertainty quantification and often require a large amount of data to train a reliable model. They are therefore not well suited for the production forecasting problem considered here, in which only a limited amount of observed production data are available and the goal is to quantify uncertainty. In the context of weather forecasting, Krishnamurti et al. (2000) proposed multimodel ensemble forecast analysis. This approach was further refined by Pagowski et al. (2005) and later extended by Mallet et al. (2009) to allow sequential data assimilation. The basic idea of ensemble forecast analysis is to construct the forecast as a linear combination of simulation results from an ensemble of prior models. The weightings used in the linear combination are determined by constraining the forecast to match observations.

The procedures mentioned above, however, only produce a single forecast and are thus not suitable for uncertainty quantification. Scheidt et al. (2015) proposed a prediction-focused analysis (PFA) approach for uncertainty assessment in a subsurface solute-transport problem. They first projected, separately, the forecast and historical data responses (pollutant concentration at multiple time steps), from an ensemble of prior models, to very low-dimensional spaces. The relationship between historical data and predictions was then constructed by applying a kernel smoothing algorithm in the joint space of the two low-dimensional spaces. Scheidt et al. (2015) showed that PFA provided reasonable uncertainty quantification for the problem considered. However, the kernel smoothing algorithm used in that work is only appropriate for projected spaces of low dimension. This requirement will likely become problematic when the number of observed and forecast data increases. Satija and Caers (2015) later modified the PFA approach by introducing a linear regression to address the issue of dimension reduction. This approach, however, attempts to linearize the relationship between historical and forecast data responses, which may not always be appropriate for the types of reservoir forecasting problems considered in this paper.

In our DSI procedure, mapping operations are introduced to transform the non-Gaussian data variables to mapped variables that are close to Gaussian. Principal

component analysis is also applied in the mapped data space both for dimension reduction and to avoid the calculation of a potentially poorly conditioned pseudo-inverse matrix. To generate multiple samples from the posterior distribution of data variables given observed data, we apply the randomized maximum likelihood (RML) method (Kitanidis 1986; Oliver et al. 1996; Reynolds et al. 1999). Because the DSI procedure with mapping operations relaxes the Gaussian assumption on the distribution of data variables, it can be applied for cases with realistic phase-rate and/or bottom-hole pressure data from a number of wells. In addition, the incorporation of observation data error, in the form of the data covariance matrix C_D, can be readily achieved within the DSI formulation. Once simulation results from prior models are obtained, the DSI procedure can be performed efficiently (in a matter of minutes for the problems considered here) over a range of C_D, which allows for the efficient assessment of the impact of this important quantity on predictions. In model-based procedures, by contrast, the use of different C_D would require one to repeat the modeling calibration procedure, which is very time-consuming. It is important to emphasize, however, that DSI does not provide posterior geological models—only posterior forecasts/data vectors.

This paper is organized as follows. In Sect. 2, the basic data-space inversion (DSI) formulation is introduced under the assumption that the prior distribution of data variables is Gaussian. Results for a simple test case are then presented. Next, in Sect. 3, we present mapping operations to transform the non-Gaussian data variables into more nearly Gaussian variables, and we introduce PCA to reduce the dimension of the data space. The use of this extended DSI for generating multiple (RML) posterior data samples is then described. In Sect. 4, the extended DSI procedure is applied for a bimodal channelized system. We assess the accuracy of DSI by providing detailed comparisons between DSI results for the quantification of uncertainty in production forecasts to results from an exhaustive (and very expensive) rejection sampling procedure. A similar assessment is performed for an oil–water–gas system, with a three-dimensional Gaussian geological model, in Sect. 5. A detailed mathematical description of the mapping operations is presented in the Appendix.

2 Basic Data-Space Inversion Formulation

In this section, we introduce a new data-space inversion (DSI) formulation within a Bayesian framework. In the initial DSI formulation, the prior probability density function (PDF) of the data variables is assumed to be Gaussian. In this case, analytical solutions for the posterior PDF of the data variables conditioned to observed data can be constructed. The basic DSI formulation is then applied for a reservoir performance forecasting problem that displays a simple production response. In the following section, we will describe the extended DSI formulation to deal with cases when the prior PDF of the data variables is non-Gaussian.

2.1 Mathematical Formulation

In this paper, data variables refer to the uncertain quantities of interest, which in this case are well production data at different time steps. The data variables are contained in an N_d-dimensional column vector of the form

$$\mathbf{d}_{\text{full}} = [\mathbf{d}_1^{\text{T}},\ \mathbf{d}_2^{\text{T}}, \ldots, \mathbf{d}_k^{\text{T}}, \ldots, \mathbf{d}_{N_{\text{h}}}^{\text{T}}, \ldots, \mathbf{d}_{N_{\text{t}}}^{\text{T}}]^{\text{T}}, \tag{1}$$

where $\mathbf{d}_k$ is the data vector at time step t_k, N_{h} is the total number of time steps during the history-matching period, and N_{t} is the total number of time steps until the end of the forecast period. In the cases considered in this paper, $\mathbf{d}_{\text{full}}$ contains different types of well production data variables (e.g., oil and water production rates and bottom-hole pressure, or BHP, at different time steps) from multiple wells. The data variables during the history-matching period are observed (measured) and assembled into an N_{obs}-dimensional column vector $\mathbf{d}_{\text{obs}}$ given by

$$\mathbf{d}_{\text{obs}} = [\mathbf{d}_{\text{obs},1}^{\text{T}},\ \mathbf{d}_{\text{obs},2}^{\text{T}}, \ldots, \mathbf{d}_{\text{obs},k}^{\text{T}}, \ldots, \mathbf{d}_{\text{obs},N_{\text{h}}}^{\text{T}}]^{\text{T}}, \tag{2}$$

where $\mathbf{d}_{\text{obs},k}$ represents the observation vector at time step k.

Our goal is to predict quantities of interest, that are part of the data vector $\mathbf{d}_{\text{full}}$, conditioned to observations $\mathbf{d}_{\text{obs}}$. The approach used to estimate $\mathbf{d}_{\text{full}}$ conditioned to $\mathbf{d}_{\text{obs}}$ is referred to as a data-space inversion (DSI) procedure. The data space is here the space of all possible data vectors in the form of Eq. (1). Note that this data space is different from the data space referred to by Tarantola (2005) and Oliver et al. (2008), as their data space is of the same dimension as the observation vector shown in Eq. (2).

We adopt a Bayesian framework and treat $\mathbf{d}_{\text{full}}$ as a vector of random variables. Thus the conditional PDF of $\mathbf{d}_{\text{full}}$ conditional to $\mathbf{d}_{\text{obs}}$ can be written as

$$p(\mathbf{d}_{\text{full}}|\mathbf{d}_{\text{obs}}) = \frac{p(\mathbf{d}_{\text{obs}}|\mathbf{d}_{\text{full}})\,p(\mathbf{d}_{\text{full}})}{p(\mathbf{d}_{\text{obs}})} \propto p(\mathbf{d}_{\text{obs}}|\mathbf{d}_{\text{full}})\,p(\mathbf{d}_{\text{full}}), \tag{3}$$

where $p(\mathbf{d}_{\text{obs}}|\mathbf{d}_{\text{full}})$ is the conditional PDF of observing $\mathbf{d}_{\text{obs}}$ given $\mathbf{d}_{\text{full}}$, and $p(\mathbf{d}_{\text{full}})$ represents the prior PDF of the data variables, which is specified independent of observed data. Because observed data are obtained by measuring the data variables during the historical period, the relationship between $\mathbf{d}_{\text{obs}}$ and $\mathbf{d}_{\text{full}}$ can be expressed as

$$\mathbf{d}_{\text{obs}} = H\mathbf{d}_{\text{full}} + \boldsymbol{\epsilon}, \tag{4}$$

where H is simply an $N_{\text{obs}} \times N_{\text{d}}$ matrix that selects (extracts) the elements in $\mathbf{d}_{\text{full}}$ corresponding to $\mathbf{d}_{\text{obs}}$, and $\boldsymbol{\epsilon}$ represents the vector of observation (measurement) errors in the data. The observation errors are assumed to be Gaussian random variables with mean $\mathbf{0}$ and covariance matrix C_{D}. The conditional probability of observing $\mathbf{d}_{\text{obs}}$, given $\mathbf{d}_{\text{full}}$, is then equal to the probability of $\boldsymbol{\epsilon}$, that is

$$\begin{aligned} p(\mathbf{d}_{\text{obs}}|\mathbf{d}_{\text{full}}) &= p(\boldsymbol{\epsilon} = \mathbf{d}_{\text{obs}} - H\mathbf{d}_{\text{full}}) \\ &\propto \exp\left(-\frac{1}{2}(H\mathbf{d}_{\text{full}} - \mathbf{d}_{\text{obs}})^{\text{T}} C_{\text{D}}^{-1} (H\mathbf{d}_{\text{full}} - \mathbf{d}_{\text{obs}})\right). \end{aligned} \tag{5}$$

If all data variables are initially uncertain and the uncertainty can be represented by a Gaussian PDF with mean $\mathbf{d}_{\text{prior}}$ and covariance matrix C_{pd}, the prior PDF of the data variables becomes

$$p(\mathbf{d}_{\text{full}}) \propto \exp\left(-\frac{1}{2}(\mathbf{d}_{\text{full}} - \mathbf{d}_{\text{prior}})^{\text{T}} C_{\text{pd}}^{-1}(\mathbf{d}_{\text{full}} - \mathbf{d}_{\text{prior}})\right). \tag{6}$$

In this paper, the prior mean $\mathbf{d}_{\text{prior}}$ and covariance matrix C_{pd} are computed numerically from the simulated data vectors corresponding to an ensemble of prior reservoir model realizations. Given an ensemble of N_{r} independently generated prior models $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_{N_{\text{r}}}$, the simulated data are computed via

$$(\mathbf{d}_{\text{full}})_i = \mathbf{g}(\mathbf{m}_i), \quad i = 1, 2, \ldots, N_{\text{r}}, \tag{7}$$

where $\mathbf{g}(\cdot)$ represents the forward model, which relates model parameters to dynamic data, and $(\mathbf{d}_{\text{full}})_i$ is an N_{d}-dimensional vector containing the simulated data corresponding to the data variables in $\mathbf{d}_{\text{full}}$. It is assumed that there is no modeling error associated with the forward model. Thus the data vectors $(\mathbf{d}_{\text{full}})_i$ $(i = 1, 2, \ldots, N_{\text{r}})$ can be treated as independent realizations of the random vector $\mathbf{d}_{\text{full}}$ before conditioning to observed data. The prior mean for $\mathbf{d}_{\text{full}}$ is then computed as

$$\mathbf{d}_{\text{prior}} = \frac{1}{N_{\text{r}}} \sum_{i=1}^{N_{\text{r}}} (\mathbf{d}_{\text{full}})_i, \tag{8}$$

and the prior covariance matrix is given by

$$C_{\text{pd}} = \frac{1}{N_{\text{r}} - 1} \sum_{i=1}^{N_{\text{r}}} \left((\mathbf{d}_{\text{full}})_i - \mathbf{d}_{\text{prior}}\right)\left((\mathbf{d}_{\text{full}})_i - \mathbf{d}_{\text{prior}}\right)^{\text{T}}. \tag{9}$$

Combining Eqs. (3), (5) and (6), the posterior PDF of a data vector $\mathbf{d}_{\text{full}}$ given a vector of observations $\mathbf{d}_{\text{obs}}$ is

$$\begin{aligned} p(\mathbf{d}_{\text{full}}|\mathbf{d}_{\text{obs}}) = k \exp\Bigg(&-\frac{1}{2}(H\mathbf{d}_{\text{full}} - \mathbf{d}_{\text{obs}})^{\text{T}} C_{\text{D}}^{-1}(H\mathbf{d}_{\text{full}} - \mathbf{d}_{\text{obs}}) \\ &-\frac{1}{2}(\mathbf{d}_{\text{full}} - \mathbf{d}_{\text{prior}})^{\text{T}} C_{\text{pd}}^{-1}(\mathbf{d}_{\text{full}} - \mathbf{d}_{\text{prior}})\Bigg), \end{aligned} \tag{10}$$

where k is the normalization constant. The function in the exponent in Eq. (10) is of quadratic form with respect to the data vector $\mathbf{d}_{\text{full}}$. Therefore, the posterior PDF of $\mathbf{d}_{\text{full}}$ is Gaussian with mean $\widetilde{\mathbf{d}}$, which is given by either of the following two equivalent expressions

$$\begin{aligned} \widetilde{\mathbf{d}} &= \mathbf{d}_{\text{prior}} + \left(H^{\text{T}} C_{\text{D}}^{-1} H + C_{\text{pd}}^{-1}\right)^{-1} H^{\text{T}} C_{\text{D}}^{-1}(\mathbf{d}_{\text{obs}} - H\mathbf{d}_{\text{prior}}) \\ &= \mathbf{d}_{\text{prior}} + C_{\text{pd}} H^{\text{T}}(H C_{\text{pd}} H^{\text{T}} + C_{\text{D}})^{-1}(\mathbf{d}_{\text{obs}} - H\mathbf{d}_{\text{prior}}). \end{aligned} \tag{11}$$

The covariance $\widetilde{C}$ is expressed as

$$\widetilde{C} = \left(H^{\mathrm{T}} C_{\mathrm{D}}^{-1} H + C_{\mathrm{pd}}^{-1}\right)^{-1}$$
$$= C_{\mathrm{pd}} - C_{\mathrm{pd}} H^{\mathrm{T}} (H C_{\mathrm{pd}} H^{\mathrm{T}} + C_{\mathrm{D}})^{-1} H C_{\mathrm{pd}}. \quad (12)$$

Detailed proofs to obtain the posterior mean and covariance matrix for linear Gaussian models of this type were given by Tarantola (2005) and Oliver et al. (2008).

With the basic DSI method, reservoir predictions conditioned to observations can be generated directly by sampling the Gaussian distribution with mean $\widetilde{\mathbf{d}}$ and covariance $\widetilde{C}$. This basic DSI method is expected to remain approximately valid for data vectors that are 'close' to Gaussian, though as will be seen, the method is not directly applicable when the data vectors are strongly non-Gaussian.

2.2 Test Case

The basic DSI procedure will now be applied for production forecasting in an oil–water problem. The reservoir model covers an area of 1 km × 1 km and is of thickness 50 m. Four water injection wells operate near each of the corners, and there is a single producer at the center of the model (see Fig. 1a). The wells operate under fixed bottom-hole pressure (BHP) controls of 200 bar for the producer and 250 bar for all injectors. The relative permeability curves are shown in Fig. 1b. Capillary pressure effects are ignored. Oil and water viscosities at standard conditions are 1.16 and 0.31 cp, respectively. The initial oil and water saturations are 0.9 and 0.1.

The reservoir is modeled on a 50 × 50 × 5 simulation grid. The log-permeability field is assumed to be Gaussian with a mean of 5, which corresponds to a permeability of around 148 md and standard deviation ($\sigma_{\ln k}$) of 2.5. We generate an ensemble of $N_{\mathrm{r}} = 100$ prior models that are conditioned to hard data using the sequential Gaussian simulation algorithm within the SGeMS (Remy et al. 2009) geostatistical toolbox. The log-permeability field of one prior model is shown in Fig. 1a. An exponential variogram is used, with correlation lengths in terms of number of grid blocks specified as $l_1 = 25$ (in the x–y plane, along the I2–P1 direction in Fig. 1a), $l_2 = 5$ (along

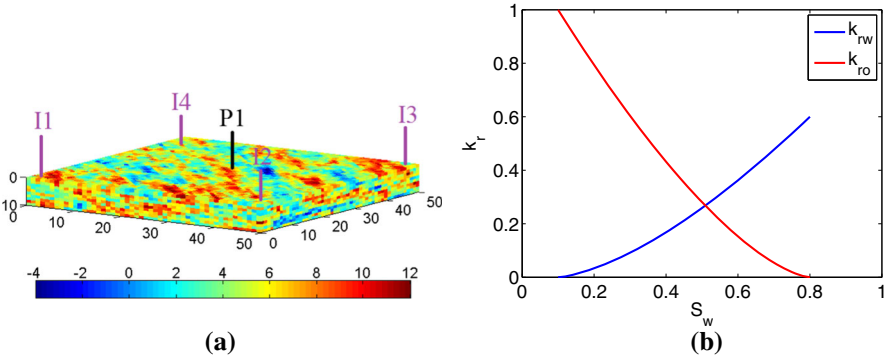

Fig. 1 Reservoir model and relative permeability curves: **a** log-permeability field of one prior model (conditioned to hard data), **b** oil and water relative permeability curves

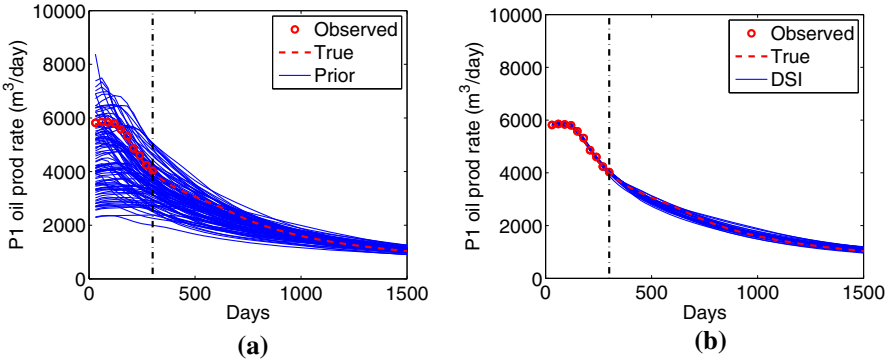

Fig. 2 DSI for oil production rate forecasts (model in Fig. 1a). **a** OPR from prior models, **b** OPR from DSI

the I1–P1 direction), and $l_z = 3$. Porosity is assumed to be constant at 0.2. Flow simulations are performed using Stanford's Automatic Differentiation-based General Purpose Research Simulator, AD-GPRS (Zhou 2012).

The history-matching period is the first 300 days. During this period, oil production rate (OPR) data at P1 are measured. The goal is then to predict OPR until 1500 days. For validation purposes, a 'true' model is generated and simulated, against which the prediction results will be compared. This model is consistent with the prior geological description, but it is not included in the set of 100 prior models used within the DSI procedure. Figure 2a shows the simulated data from all prior models. The vertical dashed line separates the history-matching and forecasting periods. The observed data, shown as red circles in Fig. 2a, are generated by adding random Gaussian noise to OPR results using the 'true' model. The measurement errors are assumed to be independently Gaussian distributed, with zero mean and covariance matrix C_{D}. The standard deviation of the measurement error is specified to be 2% of the corresponding rate data. Simulated data are reported every 30 days. We thus have $\mathbf{d}_{\mathrm{obs}} \in \mathbb{R}^{10\times 1}$ and $\mathbf{d}_{\mathrm{full}} \in \mathbb{R}^{50\times 1}$.

To generate predictions that are conditioned to observations using the DSI method, we assemble the simulated data into data vectors $(\mathbf{d}_{\mathrm{full}})_i$ $(i = 1, 2, \ldots, N_{\mathrm{r}})$ and compute the prior mean $\mathbf{d}_{\mathrm{prior}}$ and covariance matrix C_{pd} using Eqs. (8) and (9). Then, the posterior mean $\widetilde{\mathbf{d}}$ and covariance matrix $\widetilde{C}$ are obtained from Eqs. (11) and (12). Finally, predictions are generated by sampling the multivariate Gaussian distribution with mean $\widetilde{\mathbf{d}}$ and covariance $\widetilde{C}$. Figure 2b shows 25 conditioned predictions (blue curves). The predictions are seen to closely match the observed data, and they display a much smaller range of uncertainty compared with predictions from the prior models (Fig. 2a). The true production data (dashed red curve) lie within the range of predictions obtained from DSI. These results demonstrate that the basic DSI method can provide reasonable uncertainty quantification for this simple case.

3 Data-Space Inversion for Non-Gaussian Data Variables

In the previous section, we described the basic DSI method and showed that it is able to provide an accurate forecast for the simple example considered. However, the basic

DSI method assumes that the data vector $\mathbf{d}_{\text{full}}$ is (multivariate) Gaussian a priori, and this is not always a valid assumption. In this section, we describe an extended DSI procedure that enables the method to treat cases in which the prior distribution of data variables is non-Gaussian. The first step in this extended DSI procedure is to reparameterize the data variables such that the new variables, denoted by $\boldsymbol{\xi}$, are approximately (multivariate) Gaussian a priori. This idea of transforming non-Gaussian variables to approximately Gaussian variables has been applied by Gu and Oliver (2006) and Chen et al. (2009) to improve the performance of EnKF procedures. However, these investigators only considered the transformation (reparameterization) of water saturation data. To the best of our knowledge, general techniques for reparameterizing non-Gaussian production data variables have not been presented. In this section, the reparameterization of data variables for a particular reservoir flow response will be illustrated. A detailed mathematical treatment, which is applicable to a wide range of production data variable types, is presented in the Appendix.

3.1 Transformation of Non-Gaussian Data Variables

The data transformation for a commonly observed reservoir flow response is now described. Figure 3a shows a typical example of simulated water cut data from an ensemble of prior reservoir models (these models were considered in Sun (2014); detailed descriptions of the problem setup are not provided here since the purpose is to illustrate data-variable transformations). Water cut (F_w) in a production well, which is the fraction of water in the produced fluid, is defined as $F_w = q_w/(q_w + q_o)$, where q_w and q_o are the flow rates of water and oil in the well. Similar water cut behavior will be seen in the examples considered in Sects. 4 and 5. It is assumed for now that the data vector $\mathbf{d}_{\text{full}}$ only includes water cut at different time steps. The target is to transform these data variables to new variables that are closer to Gaussian. The simulated data are reported every 10 days and the total simulation time is 3000 days, which results in $(\mathbf{d}_{\text{full}})_i \in \mathbb{R}^{300\times 1}$ $(i = 1, 2, \ldots, N_{\text{r}})$. Because all data vectors $(\mathbf{d}_{\text{full}})_i$ are generated independently (prior to conditioning to observations), they are viewed as prior samples of the data vector $\mathbf{d}_{\text{full}}$. The Gaussianity of $\mathbf{d}_{\text{full}}$ can thus be examined through the ensemble of simulated data vectors.

Figure 3c shows the histogram of water cut at 1000 days. This histogram is clearly non-Gaussian, as it displays an L-shape and peaks at a water cut value of zero. This non-Gaussian behavior also holds for water cut at other time steps (particularly before 1000 days). Figure 3e shows the cross-correlation between water cut at different time steps. The large number of zero water cut values at 1000 days renders this plot highly nonlinear overall. The behaviors observed in Fig. 3c, e indicate that a multi-Gaussian assumption for the prior distribution of the data variables is clearly inappropriate in this case.

It is evident, however, that the non-Gaussian character of the water cut data is largely due to the different water breakthrough times (breakthrough denotes the point at which water appears at the production well, which is when water cut becomes nonzero). Data behavior is actually quite similar for different realizations within the same stage, i.e., before breakthrough the water cut is zero, and after breakthrough the

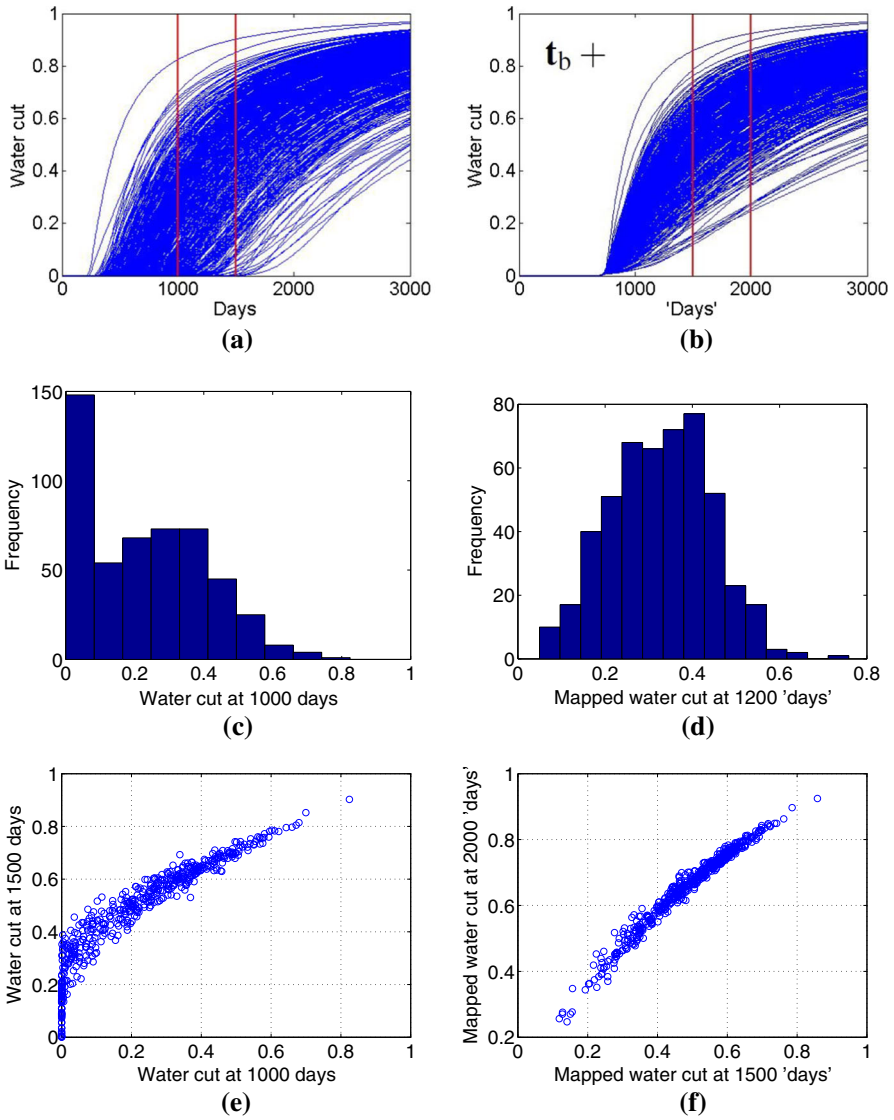

Fig. 3 Example illustrating impact of applying mapping operation: **a** simulated water cut from prior models, **b** water cut after mapping operation ($\mathbf{t}_b$ denotes the vector containing the water breakthrough times corresponding to all prior models, and '$\mathbf{t}_b+$' means that $\mathbf{t}_b$ is also included in the mapped data), **c** histogram of water cut at 1000 days, **d** histogram of mapped water cut at 1200 'days,' **e** cross-correlation between water cut at 1000 and 1500 days, **f** cross-correlation between mapped water cut at 1500 and 2000 'days'

water cut curves increase with a consistent concavity. These observations motivate a mapping of the column data vectors $(\mathbf{d}_{\text{full}})_i$ such that, in the mapped data vectors $\widehat{\mathbf{d}}_i$, all data in a given row correspond to the same stage. This mapping operation consists of time shifting and compressing/stretching operations (see the Appendix for the detailed mathematical formulation).

We plot the mapped data vectors, in which the breakthrough 'times' are now identical for each realization, in Fig. 3b (quotes are used to indicate the time variable after mapping). In this figure, $\mathbf{t}_\mathrm{b}$ denotes a vector containing the actual breakthrough times for water cut shown in Fig. 3a. The mapped water cut data variables $\widehat{\mathbf{d}}$ are then expressed as

$$\widehat{\mathbf{d}} = [\, t_b,\ \hat{d}(\tau = \tau_1), \ldots, \hat{d}(\tau = \tau_k), \ldots, \hat{d}(\tau = \tau_{\widehat{N}_\mathrm{t}})]^\mathrm{T}, \tag{13}$$

where t_b denotes breakthrough time, τ is the 'time' variable for the mapped data (x-axis in Fig. 3b), $\widehat{N}_\mathrm{t}$ is the total number of 'time' steps, and $\hat{d}(\tau = \tau_k)$ represents the data variable at 'time' τ_k. Note that t_b must be included in $\widehat{\mathbf{d}}$. Otherwise, $\widehat{\mathbf{d}}$ could not be mapped back to the original data space. In this work, the 'time' steps τ_1 to $\tau_{\widehat{N}_\mathrm{t}}$ are set to be uniformly distributed from the start to the end of the simulation period (interpolation is applied when necessary), and the number of 'time' steps $\widehat{N}_\mathrm{t}$ is set equal to N_t.

Figure 3d shows the histogram of mapped data at 1200 'days,' and we see that it is close to Gaussian. The histograms of mapped data at other post-breakthrough 'times' (though not shown) were also found to be close to Gaussian. Furthermore, the cross-correlation between mapped water cut at 1500 and 2000 'days,' presented in Fig. 3f, is now nearly linear. This level of linearity is also observed in the cross-correlations between mapped water cut at other 'times.' Though it is not straightforward to determine just how close the mapped variables $\widehat{\mathbf{d}}$ are to multivariate Gaussian, such an assessment does not appear to be necessary, as it will be shown in Sect. 4.2 that the performance of DSI is much improved through use of these $\widehat{\mathbf{d}}$.

3.2 Data-Space Formulation with Reparameterization

In this section, we describe the extended DSI formulation with reparameterization of the original non-Gaussian data vector $\mathbf{d}_\mathrm{full}$. We let $\boldsymbol{\eta}$ denote the reparameterized data vector, which is multivariate Gaussian a priori. The relationship between $\mathbf{d}_\mathrm{full}$ and $\boldsymbol{\eta}$ is then expressed as

$$\mathbf{d}_\mathrm{full} = \mathbf{f}(\boldsymbol{\eta}). \tag{14}$$

In the previous section, mapping operations were introduced to transform $\mathbf{d}_\mathrm{full}$ to the mapped data vector $\widehat{\mathbf{d}}$. Thus, we can set $\boldsymbol{\eta} = \widehat{\mathbf{d}}$ if the prior distribution of $\widehat{\mathbf{d}}$ is nearly Gaussian.

In Sect. 2, we derived the conditional PDF of $\mathbf{d}_\mathrm{full}$ given observed data $\mathbf{d}_\mathrm{obs}$ within a Bayesian framework. Similarly, the conditional PDF of $\boldsymbol{\eta}$ can be written as

$$\begin{aligned} p(\boldsymbol{\eta}|\mathbf{d}_\mathrm{obs}) &\propto p(\mathbf{d}_\mathrm{obs}|\boldsymbol{\eta})\, p(\boldsymbol{\eta}) \\ &\propto \exp\left(-\frac{1}{2}(H\mathbf{f}(\boldsymbol{\eta}) - \mathbf{d}_\mathrm{obs})^\mathrm{T} C_\mathrm{D}^{-1} (H\mathbf{f}(\boldsymbol{\eta}) - \mathbf{d}_\mathrm{obs})\right. \\ &\quad \left. -\frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}_\mathrm{prior})^\mathrm{T} C_\eta^{-1} (\boldsymbol{\eta} - \boldsymbol{\eta}_\mathrm{prior})\right), \end{aligned} \tag{15}$$

where $\boldsymbol{\eta}_{\text{prior}}$ and C_η are the prior mean and prior covariance matrix of random vector $\boldsymbol{\eta}$. Equation (15) gives the PDF to be sampled to generate conditional realizations of $\boldsymbol{\eta}$, from which the corresponding data vector $\mathbf{d}_{\text{full}}$ can be predicted through Eq. (14).

We emphasize that the function in Eq. (14) must be nonlinear given the condition that $\mathbf{d}_{\text{full}}$ is non-Gaussian, but $\boldsymbol{\eta}$ is Gaussian. This can be proven through contradiction: Assuming $\mathbf{f}(\cdot)$ is linear, we can then write $\mathbf{d}_{\text{full}} = A\boldsymbol{\eta} + \mathbf{b}$, where $\boldsymbol{\eta} \in \mathbb{R}^{N_l}$, $A \in \mathbb{R}^{N_d \times N_l}$ and $\mathbf{b} \in \mathbb{R}^{N_d}$. If $\boldsymbol{\eta}$ is Gaussian, $\mathbf{d}_{\text{full}}$ is then also Gaussian, which contradicts the condition that $\mathbf{d}_{\text{full}}$ is non-Gaussian. Thus $\mathbf{f}(\boldsymbol{\eta})$ is a nonlinear function. In this case, the misfit function in the exponent in Eq. (15) is no longer quadratic. Thus, the posterior distribution of $\boldsymbol{\eta}$ cannot be expressed as a simple Gaussian distribution that can be easily sampled. In Sect. 3.4, we will describe the application of the randomized maximum likelihood (RML) method to sample this posterior PDF.

We see from Eq. (15) that the posterior PDF $p(\boldsymbol{\eta}|\mathbf{d}_{\text{obs}})$ requires the inverse of the prior covariance matrix C_η. It was found, however, that C_η can be poorly conditioned (and not invertible) when we set $\boldsymbol{\eta} = \widehat{\mathbf{d}}$, as in the development above. This ill-conditioning occurs when the number of mapped data vectors N_r used to construct C_η is less than the dimension of the mapped data vector $\widehat{N}_d$. In addition, the strong correlations between mapped variables in $\widehat{\mathbf{d}}$ (this correlation is evident in Fig. 3f) may render C_η low rank. To address this issue, principal component analysis (PCA) will be applied to reparameterize $\widehat{\mathbf{d}}$, which allows us to avoid computing C_η^{-1}. PCA also acts to reduce the number of variables to be estimated, and this makes the generation of posterior predictions even more efficient.

3.3 Reparameterization of Mapped Data Variables Using PCA

In this section principle component analysis (PCA) is introduced in order to effectively treat C_η^{-1}. PCA is a widely applied statistical procedure that enables the representation of a set of correlated variables using a set of linearly uncorrelated variables. Because the leading principal components are associated with the largest variance, PCA is highly effective for dimension reduction. See Shlens (2005) for a tutorial on PCA and Liang et al. (2002) for theory and proofs. The application of PCA to reparameterize reservoir model parameters has been discussed by many authors; see, e.g., Oliver (1996), Reynolds et al. (1996), Sarma et al. (2006) and Vo and Durlofsky (2014).

Here PCA is used to reparameterize the mapped data variables in $\widehat{\mathbf{d}}$. We define $X \in \mathbb{R}^{\widehat{N}_d \times N_r}$ as the following centered matrix

$$X = [\,\widehat{\mathbf{d}}_1 - \widehat{\mathbf{d}}_{\text{prior}} \quad \widehat{\mathbf{d}}_2 - \widehat{\mathbf{d}}_{\text{prior}} \quad \ldots \quad \widehat{\mathbf{d}}_{N_r} - \widehat{\mathbf{d}}_{\text{prior}}], \tag{16}$$

where $\widehat{\mathbf{d}}_{\text{prior}} = (1/N_r)\sum_{i=1}^{N_r} \widehat{\mathbf{d}}_i$ is the mean of the mapped data from all prior models. Singular value decomposition of the matrix $X/\sqrt{N_r - 1}$ is then performed, which gives

$$X = \sqrt{N_r - 1}\, U \Sigma V^T = \sqrt{N_r - 1}\, \Phi V^T, \tag{17}$$

where U is an $\widehat{N}_d \times \widehat{N}_d$ unitary matrix, Σ is an $\widehat{N}_d \times N_r$ diagonal matrix containing the nonnegative singular values of $X/\sqrt{N_r - 1}$, V is an $N_r \times N_r$ unitary matrix, and

$\Phi = U\Sigma$ is the basis matrix. The components of Φ are ordered by their corresponding singular values. There exist a maximum of $[\min(N_r, \widehat{N}_d) - 1]$ nonzero singular values. Only N_l columns in Φ are retained, where $N_l \leq [\min(N_r, \widehat{N}_d) - 1]$. Thus, $\Phi \in \mathbb{R}^{\widehat{N}_d \times N_l}$. The value of N_l is often determined by applying an 'energy' criterion (Cardoso et al. 2009). The relative energy in the largest N_l eigenvalues is given by

$$E_{N_l} = \frac{\sum_{i=1}^{N_l} \lambda_i}{\sum_{i=1}^{\min(N_r, \widehat{N}_d)-1} \lambda_i}, \tag{18}$$

where λ_i is the square of diagonal element i in matrix Σ. The value of N_l can be determined by specifying a value for E_{N_l} (e.g., 0.995). The cumulative 'energy loss' from retaining only the N_l largest eigenvalues is given by $1 - E_{N_l}$. In practice, N_l is often much smaller than $\widehat{N}_d$, as will be seen in an example in Sect. 4.

The mapped variables in $\widehat{\mathbf{d}}$ can then be represented in terms of $\boldsymbol{\xi}$ through use of

$$\widehat{\mathbf{d}} \approx \Phi\boldsymbol{\xi} + \widehat{\mathbf{d}}_{\text{prior}}, \tag{19}$$

where $\boldsymbol{\xi} \in \mathbb{R}^{N_l \times 1}$ is the reduced-space variable. Equation (19) will be an equality if N_l is determined by setting $E_{N_l} = 1$. The relationship between $\widehat{\mathbf{d}}$ and $\mathbf{d}_{\text{full}}$ will be denoted as

$$\widehat{\mathbf{d}} = \mathscr{F}(\mathbf{d}_{\text{full}}), \qquad \mathbf{d}_{\text{full}} = \mathscr{F}^{-1}(\widehat{\mathbf{d}}), \tag{20}$$

where $\mathscr{F}(\cdot)$ represents the forward mapping operation and $\mathscr{F}^{-1}(\cdot)$ the backward mapping operation. The relationship between $\boldsymbol{\xi}$ and $\mathbf{d}_{\text{full}}$ can then be expressed as

$$\mathbf{d}_{\text{full}} = \mathbf{f}(\boldsymbol{\xi}) = \mathscr{F}^{-1}(\widehat{\mathbf{d}}) = \mathscr{F}^{-1}(\Phi\boldsymbol{\xi} + \widehat{\mathbf{d}}_{\text{prior}}). \tag{21}$$

In this paper, we choose $\boldsymbol{\xi}$ to be the reparameterized vector representing the data variables $\widehat{\mathbf{d}}$ in Eq. (14), that is, $\boldsymbol{\eta} = \boldsymbol{\xi}$. From a property of PCA, the distribution of random vectors $\boldsymbol{\xi}$ will be approximately multivariate standard normal if the $\widehat{\mathbf{d}}$ vectors are nearly Gaussian. Thus, from Eq. (15), we have

$$p(\boldsymbol{\xi}|\mathbf{d}_{\text{obs}}) \propto \exp\left(-\frac{1}{2}(H\mathbf{f}(\boldsymbol{\xi}) - \mathbf{d}_{\text{obs}})^{\mathrm{T}} C_{\mathrm{D}}^{-1}(H\mathbf{f}(\boldsymbol{\xi}) - \mathbf{d}_{\text{obs}}) - \frac{1}{2}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi}\right), \tag{22}$$

where $\mathbf{f}(\boldsymbol{\xi})$ is defined by Eq. (21). Equation (22) characterizes the posterior distribution of $\boldsymbol{\xi}$ given $\mathbf{d}_{\text{obs}}$. This equation is in a reduced space relative to Eq. (15). Note that, in this subspace, the $(\boldsymbol{\eta} - \boldsymbol{\eta}_{\text{prior}})^{\mathrm{T}} C_{\eta}^{-1}(\boldsymbol{\eta} - \boldsymbol{\eta}_{\text{prior}})$ term in Eq. (15) now appears as $\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi}$, i.e., C_{η}^{-1} no longer appears directly.

The steps required to arrive at Eq. (22) will now be summarized. We first transform the prior data (generated by performing flow simulations for an ensemble of prior models) to mapped data that are more nearly multi-Gaussian. PCA is then applied to the mapped data to enable the representation of data variables $\mathbf{d}_{\text{full}}$ by new variables $\boldsymbol{\xi}$. Under a Bayesian framework, the posterior distribution of $\boldsymbol{\xi}$ is then given by Eq. (22). In the following section, we will introduce the RML method to sample this posterior

distribution. The resulting posterior samples of $\boldsymbol{\xi}$ will then be used to obtain the production forecasts.

3.4 Randomized Maximum Likelihood Method in Data Space

The RML method (Kitanidis 1986; Oliver et al. 1996) is a procedure for sampling a posterior PDF. It has been shown that RML can sample this PDF correctly if the response is linearly related to model parameters. Gao et al. (2006) applied RML for quantifying uncertainty for the PUNQ-S3 reservoir model (Floris et al. 2001). They concluded that RML was able to give reasonable uncertainty quantification even though the responses (predicted production data) were nonlinearly related to the model parameters. Vo and Durlofsky (2015) described a subspace RML method applicable with a parameterized (PCA-type) permeability representation. In this section, RML will be implemented within the context of DSI. We begin by introducing RML within the traditional model-inversion context.

When the model variables $\mathbf{m}$ are Gaussian, and the observations $\mathbf{d}_{\text{obs}}$ are related to model variables through $\mathbf{d}_{\text{obs}} = \mathbf{g}(\mathbf{m}) + \boldsymbol{\epsilon}$, the model-space RML method, as described by Oliver et al. (2008), proceeds as follows:

1. Sample $\mathbf{m}^*$ from the prior Gaussian distribution $N[\boldsymbol{\mu}_{\text{m}}, C_{\text{m}}]$, where $\boldsymbol{\mu}_{\text{m}}$ and C_{m} are the mean and covariance of the model parameters.
2. Sample perturbed observations $\mathbf{d}^*_{\text{obs}}$ from the Gaussian distribution $N[\mathbf{d}_{\text{obs}}, C_{\text{D}}]$.
3. Compute maximum likelihood model variables $\mathbf{m}_{\text{rml}}$ through use of

$$\mathbf{m}_{\text{rml}} = \arg\min_{\mathbf{m}} \Big[(\mathbf{g}(\mathbf{m}) - \mathbf{d}^*_{\text{obs}})^{\text{T}} C_{\text{D}}^{-1} (\mathbf{g}(\mathbf{m}) - \mathbf{d}^*_{\text{obs}}) + (\mathbf{m} - \mathbf{m}^*)^{\text{T}} C_{\text{m}}^{-1} (\mathbf{m} - \mathbf{m}^*) \Big]. \tag{23}$$

4. Repeat steps 1–3 to generate additional posterior samples of $\mathbf{m}$.

In the context of this DSI formulation, the observed data $\mathbf{d}_{\text{obs}}$ are now related to 'model' parameters $\boldsymbol{\xi}$ through Eq. (21). Through use of PCA, the prior distribution of $\boldsymbol{\xi}$ is essentially multivariate standard normal. The RML method in data space then proceeds as follows:

1. Sample $\boldsymbol{\xi}^*$ from the standard normal distribution $N[\mathbf{0}, I]$.
2. Sample perturbed observations $\mathbf{d}^*_{\text{obs}}$ from the Gaussian distribution $N[\mathbf{d}_{\text{obs}}, C_{\text{D}}]$.
3. Compute maximum likelihood 'model' parameters $\boldsymbol{\xi}_{\text{rml}}$ through use of

$$\boldsymbol{\xi}_{\text{rml}} = \arg\min_{\boldsymbol{\xi}} \Big[(H\mathbf{f}(\boldsymbol{\xi}) - \mathbf{d}^*_{\text{obs}})^{\text{T}} C_{\text{D}}^{-1} (H\mathbf{f}(\boldsymbol{\xi}) - \mathbf{d}^*_{\text{obs}}) + (\boldsymbol{\xi} - \boldsymbol{\xi}^*)^{\text{T}} (\boldsymbol{\xi} - \boldsymbol{\xi}^*) \Big]. \tag{24}$$

4. Repeat steps 1–3 to continue sampling the posterior distribution of $\boldsymbol{\xi}$. Compute corresponding reservoir forecasts through Eq. (21).

In this work, the minimization of the objective function in Eq. (24) is accomplished using the quasi-Newton line search algorithm within Matlab. Because objective function evaluations are very fast, this minimization can usually be accomplished within a few seconds. Thus the total computation time to generate multiple forecasts with this data-space RML procedure is on the order of minutes. The time-consuming step is the simulation of the prior models, though these computations can all be done in parallel. We emphasize that if new observed data are incorporated, or the covariance matrix C_D is varied, new forecasts can be efficiently generated without resimulating the prior models.

Finally, we note that Tarantola (2005) showed that if $\mathbf{g}(\mathbf{m})$ is linear (i.e., $\mathbf{g}(\mathbf{m}) = G\mathbf{m}$, where G is the sensitivity matrix), then the minimum of the objective function in Eq. (23) has a Chi-squared distribution with expectation equal to N_{obs}, the number of observed data, and variance equal to $2N_{obs}$. This conclusion also applies to the minimization problem in Eq. (24). For the cases considered in this paper, the acceptable offset is restricted to be five standard deviations (as was done in Gao and Reynolds (2006)), i.e., we require

$$N_{obs} - 5\sqrt{2N_{obs}} \leq S_{\xi}(\boldsymbol{\xi}_{rml}) \leq N_{obs} + 5\sqrt{2N_{obs}}, \tag{25}$$

where $S_{\xi}(\cdot)$ is the objective function (right-hand side) in Eq. (24). If Eq. (25) is not satisfied, the generated estimate is simply discarded. For all cases presented in this paper, more than 90% of the generated estimates satisfied Eq. (25) and were thus accepted.

3.5 Pre-selection of the Prior Models

In practice, it is possible that the simulated data from many of the prior models are far from the observed data (e.g., predicted water breakthrough time is significantly later than that observed in the production data, or target rate controls are not met). For such cases, we have experimented with a pre-selection procedure in which the prior-model simulation data that are 'far' from the observed data are eliminated when constructing the basis matrix Φ in Eq. (17). This results in a PCA reparameterization that provides a better representation of the region of interest in data space. The mismatch function

$$S_d(\mathbf{d}_i) = (\mathbf{g}(\mathbf{m}_i) - \mathbf{d}_{obs})^T C_D^{-1} (\mathbf{g}(\mathbf{m}_i) - \mathbf{d}_{obs}), \quad i = 1, 2, \ldots, N_r, \tag{26}$$

where $\mathbf{g}(\mathbf{m}_i)$ denotes the simulation results during the historical period for model $\mathbf{m}_i$, is used to determine which prior-model data are retained and which are eliminated. Specifically, only the N_u models ($N_u < N_r$) corresponding to the smallest data mismatches are retained.

This pre-selection procedure, with $N_u/N_r \approx 0.2$, was found to provide slightly improved forecasting results relative to results in which all prior-model simulation data are used. This observation is based on comparisons of DSI results to reference results obtained using full rejection sampling (described below). In a typical case, however, rejection sampling results will not be available, and in their absence, a method

for determining N_u/N_r (or a cutoff value for S_d in Eq. (26)) has yet to be devised. Therefore, in the DSI results presented below, the pre-selection procedure will not be applied. A systematic assessment of this technique and the determination of appropriate (case-specific) criteria for its use are subjects for future study.

3.6 Summary of the Extended DSI Procedure

In this section, the extended DSI formulation was presented for cases when data variables are non-Gaussian a priori. The overall procedure can be summarized as follows:

1. Generate an ensemble of N_r prior model realizations that are conditioned to all prior information including hard data (i.e., property values at well locations).
2. Perform flow simulation on all prior models to obtain an ensemble of simulated data vectors $(\mathbf{d}_{\text{full}})_i$ $(i = 1, 2, \ldots, N_r)$. As an optional step, apply pre-selection to eliminate some $(\mathbf{d}_{\text{full}})_i$ (pre-selection is not applied for the results in this paper).
3. Transform the simulated data vectors to mapped data vectors $\widehat{\mathbf{d}}_i$ that are closer to multi-Gaussian. The detailed mappings are presented in the Appendix.
4. Apply PCA on the mapped data vectors $\widehat{\mathbf{d}}_i$ to obtain the basis matrix Φ. Represent the data variables $\mathbf{d}_{\text{full}}$ using new (reduced) variables $\boldsymbol{\xi}$ (Eq. (21)).
5. Sample the conditional distribution of $\boldsymbol{\xi}$ given $\mathbf{d}_{\text{obs}}$ using the RML method (Eq. (24)). Then, generate the corresponding predictions of $\mathbf{d}_{\text{full}}$ using Eq. (21).

The most time-consuming component of the DSI procedure is the generation and simulation of a relatively large set of prior models. For the examples considered in this paper, limited numerical experimentation indicated that flow results from around 200–1000 prior models provided acceptable DSI accuracy. Because these simulations can all be performed in parallel, the elapsed computational time will be on the order of only 1–10 simulations if a reasonable number (O(100–1000)) of computational cores are available. Once the prior-model simulations are performed, generating forecasts is very efficient as the data-space function evaluations are very fast.

In the examples presented here, all prior models represent realizations sampled from a particular geostatistical ensemble, i.e., all models are based on the same training image or variogram and are conditioned to the same set of well data. This is not a requirement of the DSI procedure, however, and it is possible to treat data variables originating from a range of prior models. Thus, one could incorporate, for example, uncertainty in geological style (training image) or relative permeability functions into the DSI treatment.

4 Numerical Example: Case 1

In this section, the DSI procedure described in the previous section is applied to a two-dimensional channelized system. DSI results for uncertainty quantification will be compared to those obtained from a rejection sampling (RS) procedure. Additional DSI results are presented in Sun (2014).

4.1 Model Setup and Predictions from Prior Models

Four prior models for the two-dimensional channelized system considered here are shown in Fig. 4. The model is represented on a 60 × 60 grid, with each grid block of size 25 m × 25 m × 10 m. The permeability fields are generated using a 'cookie-cutter' approach (Castro 2007). The binary facies (sand and mud) model is generated using the 'snesim' geostatistical algorithm within SGeMS (Strebelle 2002). The heterogeneity structure within a particular facies is modeled as Gaussian. The Gaussian realizations ($\ln k$) with different means (1 for mud and 5 for sand) and standard deviation (0.5 for mud and 1.5 for sand) are simulated independently using the sequential Gaussian simulation algorithm within SGeMS. All prior models are conditioned to hard data at five well locations. The porosity is assumed to be constant at 0.2. A total of three producers and two injectors are drilled in sand, as shown in Fig. 4.

An oil–water system is considered. Fluid properties are the same as in the test case described in Sect. 2.2. Injectors operate at fixed BHPs of 550 bar (I1) and 600 bar (I2), and all producers operate at fixed BHPs of 200 bar. Initial reservoir pressure is 325 bar. The flow simulation period is 3000 days, with production data reported every 30 days. Simulated data include water injection rate (WIR) for injectors and water production rate (WPR) and oil production rate (OPR) for producers. For the DSI method, a total

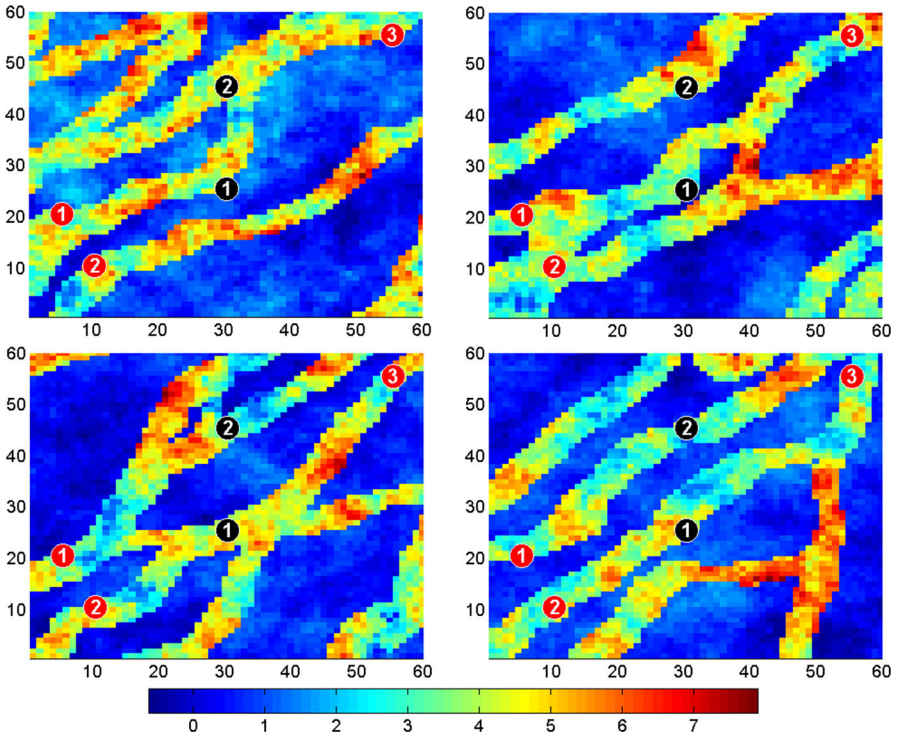

Fig. 4 Log-permeability (in md) maps for four prior channelized models. *Red circles* denote production wells, and *black circles* indicate injection wells

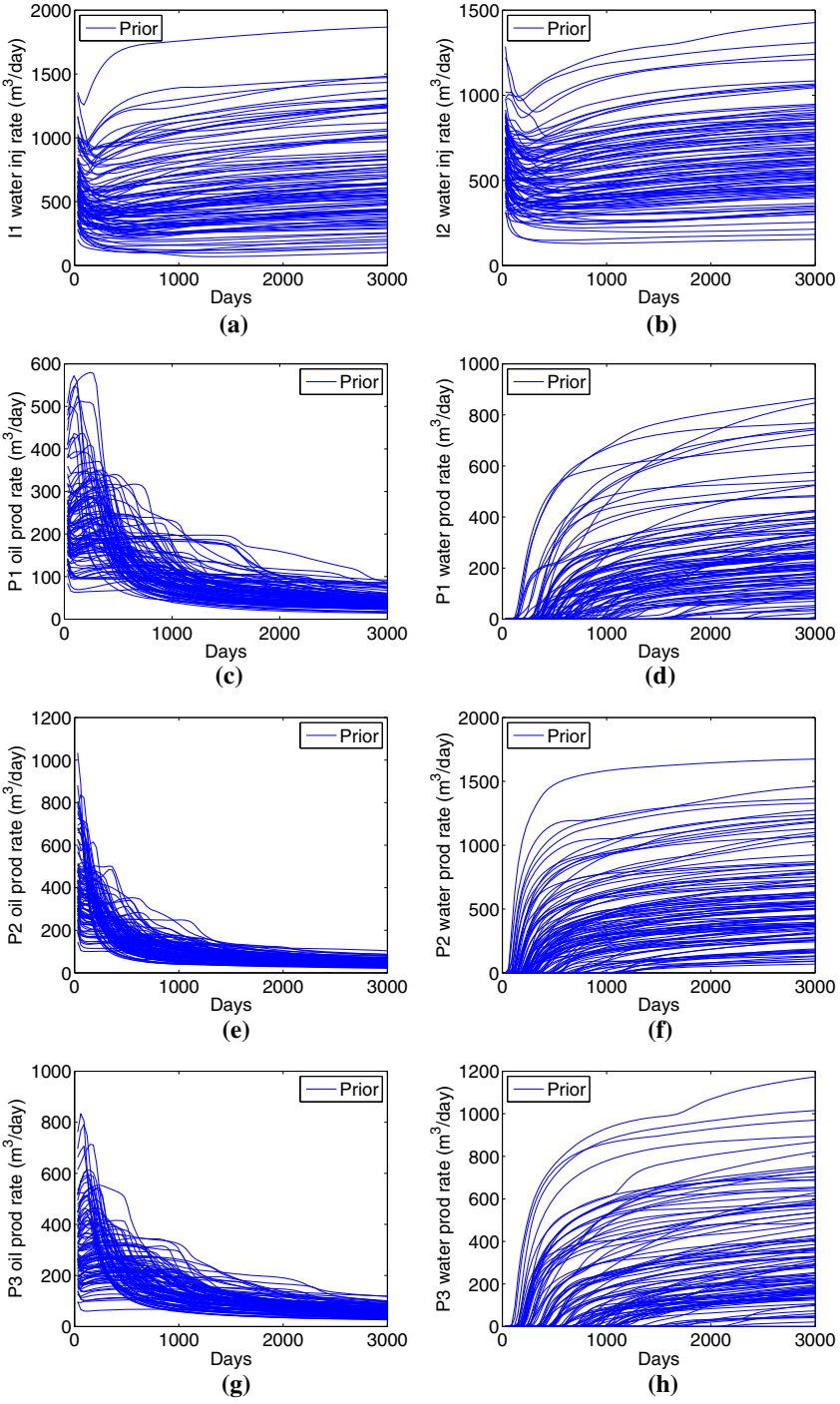

Fig. 5 Production forecasts from prior models (Case 1). **a** Water rate (I1), **b** water rate (I2), **c** oil rate (P1), **d** water rate (P1), **e** oil rate (P2), **f** water rate (P2), **g** oil rate (P3), **h** water rate (P3)

of $N_r = 500$ prior models are generated. Figure 5 shows the simulation results for 100 prior models generated using the AD-GPRS simulator (Zhou 2012). The predictions from these prior models display considerable uncertainty, especially for WIR and WPR, even though all models are conditioned to hard data.

In this case, we have multiple types of data at multiple wells. Data variables are assembled as

$$\mathbf{d}_{\text{full}} = [\mathbf{d}_{\text{I1}}^{\text{T}}, \mathbf{d}_{\text{I2}}^{\text{T}}, \mathbf{d}_{\text{OPR, P1}}^{\text{T}}, \mathbf{d}_{\text{WPR, P1}}^{\text{T}}, \mathbf{d}_{\text{OPR, P2}}^{\text{T}}, \mathbf{d}_{\text{WPR, P2}}^{\text{T}}, \mathbf{d}_{\text{OPR, P3}}^{\text{T}}, \mathbf{d}_{\text{WPR, P3}}^{\text{T}}]^{\text{T}}, \quad (27)$$

where $\mathbf{d}_{\text{I1}}$ and $\mathbf{d}_{\text{I2}}$ denote column vectors containing injection rate data at all time steps for wells I1 and I2, $\mathbf{d}_{\text{OPR, P1}}$ contains oil production rate data at all time steps for well P1, etc. Because the total simulation period is 3000 days, and uniform time steps of size 30 days are used, each component on the right hand of Eq. (27) is of dimension 100, and $\mathbf{d}_{\text{full}} \in \mathbb{R}^{800\times 1}$.

The mapping operations (described in Sect. 3.1 and in the Appendix) are applied for each data type on a well-by-well basis. The mapped set of data variables is thus expressed as

$$\widehat{\mathbf{d}} = [\widehat{\mathbf{d}}_{\text{I1}}^{\text{T}}, \widehat{\mathbf{d}}_{\text{I2}}^{\text{T}}, \widehat{\mathbf{d}}_{\text{OPR, P1}}^{\text{T}}, \widehat{\mathbf{d}}_{\text{WPR, P1}}^{\text{T}}, \widehat{\mathbf{d}}_{\text{OPR, P2}}^{\text{T}}, \widehat{\mathbf{d}}_{\text{WPR, P2}}^{\text{T}}, \widehat{\mathbf{d}}_{\text{OPR, P3}}^{\text{T}}, \widehat{\mathbf{d}}_{\text{WPR, P3}}^{\text{T}}]^{\text{T}}. \quad (28)$$

No mapping is applied for WIR (for I1 and I2), so $\widehat{\mathbf{d}}_{\text{I1}} = \mathbf{d}_{\text{I1}}$ and $\widehat{\mathbf{d}}_{\text{I2}} = \mathbf{d}_{\text{I2}}$. The mapping operation for WPR was described in Sect. 3.1. Because OPR declines at a particular well when water breaks through at that well, the transition time for well OPR (explained in the Appendix) coincides with water breakthrough time. The mapped data for P1 are shown in Fig. 6, and we can see that the mappings act to align the pre- and post-breakthrough stages. We reiterate that, with these mapping operations, the breakthrough times are treated as new variables in $\widehat{\mathbf{d}}$ and need to be predicted explicitly. Mapping operations for P2 and P3 are analogous to those for P1.

Figure 7 shows the cumulative energy loss versus the number of principal components retained in Φ (Eq. (17)). This plot corresponds to the mapped data variables defined by Eq. (28). It is evident that most of the energy is carried by the first few

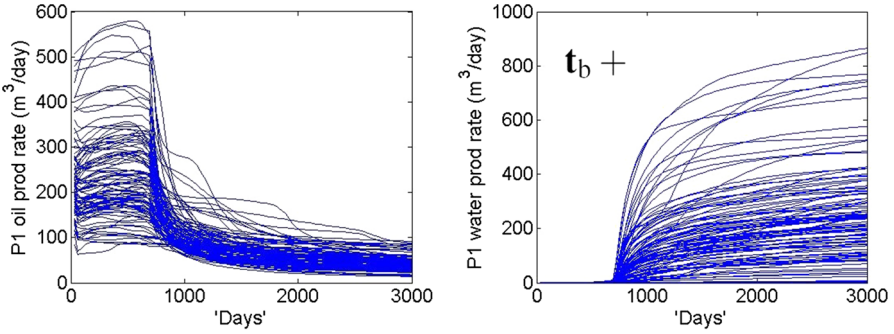

Fig. 6 Production forecasts at P1 after mapping operations. Note that the mapping for oil rate also involves water breakthrough times $\mathbf{t}_{\text{b}}$. **a** Mapped oil rate (P1), **b** mapped water rate (P1)

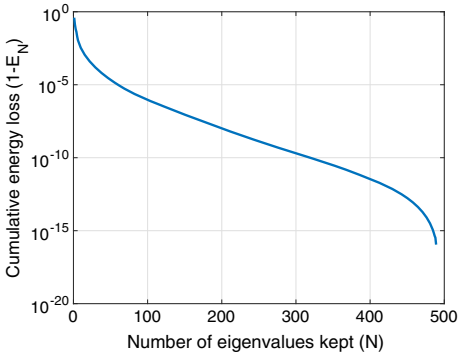

Fig. 7 Cumulative energy loss for principal components

components. We seek to form a basis such that the fraction of energy ignored is less than 0.005, which requires us to retain the first $N_l = 9$ principal components (in this paper, unless otherwise indicated, cumulative energy loss is always set to be 0.005).

4.2 Production Forecasts with Data-Space Inversion

The impact of the mapping operations on DSI results will now be demonstrated. The historical period is taken to be the first 480 days. A 'true' model that is consistent with the training image and geostatistics is generated and used to provide the true data (this model is not included in the set of prior models used within DSI). The observed data are generated from the true data by adding measurement error, which is assumed to follow a multivariate independent Gaussian distribution with zero mean and covariance matrix C_D (Gao et al. 2006). In the first set of results (shown in Fig. 8), the standard deviation of measurement error is set to be equal to 10% of the corresponding true data. Observed data at all wells are used, which results in a total of $16 \times 8 = 128$ observations.

Figure 8 shows the ensemble of predictions (blue curves) from the DSI procedure with and without the mapping operations. Observed data for WPR at P1 and P3 are shown as red circles, and the true data are indicated by red dashed lines. A total of 50 predictions are generated for each case. In Fig. 8a, the true data fall outside of the predictions if mapping operations are not applied. In Fig. 8c, unphysical (negative) water rates are observed in the forecasts when mapping is not applied. In addition, water breakthrough time, which is often an important quantity in practice, cannot be obtained directly from the forecasts.

When mapping operations are applied, however, the DSI procedure provides improved predictions that capture the true data for WPR at both P1 and P3 (Fig. 8b, d). Similar improvements over the results without mapping are also observed at other wells. Note finally that the reduction in uncertainty, relative to predictions from the prior models, is very significant (compare Fig. 5d, h to Fig. 8b, d), even though the measurement errors are set to be relatively large in this case.

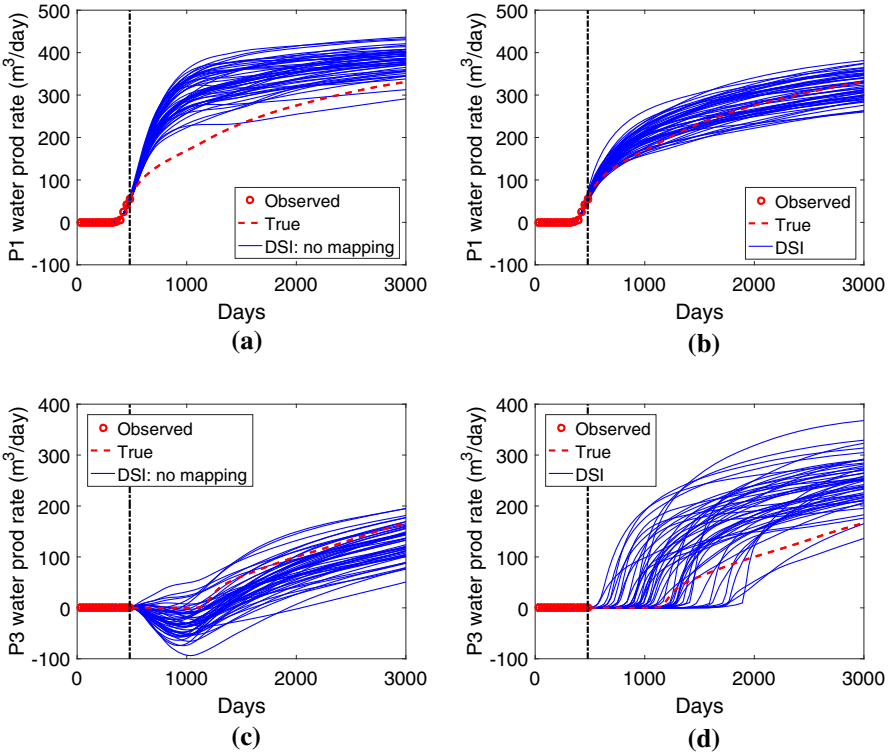

Fig. 8 Production forecasts from DSI without (*left plots*) and with (*right plots*) mapping operations. **a** Water rate (P1), **b** water rate (P1), **c** water rate (P3), **d** water rate (P3)

4.3 Rejection Sampling

The results presented above show that prediction uncertainty is reduced substantially through application of DSI. However, unless these DSI results are compared to some benchmark, there is no way of knowing whether the resulting uncertainty assessment is accurate. The rejection sampling (RS) approach provides a reference estimate for production forecast uncertainty, as RS performs a proper sampling of the posterior distribution. RS is very computationally intensive, but it is able (in concept) to generate samples from complex distributions. The basic idea in RS is to first generate a prior sample and to then decide whether to accept or reject this sample based on a test. This test is designed to be independent of previously generated samples, which guarantees that all accepted samples are independent from each other (Oliver et al. 2008).

The RS approach in this paper proceeds as follows:

1. Sample $\mathbf{m}$ from its prior distribution $N[\boldsymbol{\mu}_{\mathrm{m}}, C_{\mathrm{m}}]$.
2. Sample a variable p from a uniform distribution over [0, 1].
3. Accept $\mathbf{m}$ as a posterior sample if $p \leq L(\mathbf{m})/S_L$, where $L(\mathbf{m})$ is the likelihood function and S_L is taken to be the maximum likelihood value corresponding to all models considered for RS.

The likelihood function $L(\mathbf{m})$ is defined as

$$L(\mathbf{m}) = c \exp\left(-\frac{1}{2}(\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\text{obs}})^{\text{T}} C_{\text{D}}^{-1}(\mathbf{g}(\mathbf{m}) - \mathbf{d}_{\text{obs}})\right), \tag{29}$$

where c is a normalization constant. In this paper, S_L is chosen to be the maximum of the likelihood function values over all prior models considered in RS.

The rejection sampling approach requires a very large number of prior simulations as the amount of observed data increases. Therefore, in examples involving RS, only a small amount of observed data will be considered, and a larger standard deviation for measurement error will be specified. With these treatments, a reasonable fraction of the models considered will be accepted by the RS algorithm, which enables it to provide reference uncertainty quantification results.

4.4 Comparison of DSI and Rejection Sampling Results

Observed data in this case are obtained from wells I2, P1 and P3 at times of 180, 360 and 540 days. The total number of observed data is thus 15. In this example, the standard deviation of measurement error is set to be equal to 15% of the corresponding data. In the RS procedure, 10^6 equally probable prior models are generated and simulated. A total of 242 models are accepted. Thus, with RS, $O(5 \times 10^5)$ simulations are needed in order to generate around 100 posterior models.

Figures 9 and 10 show the forecasting results from the prior models, RS and DSI, for wells I2, P1 and P3 (note that P1 experiences early water breakthrough and P3 late breakthrough). For DSI, a total of 100 forecasts are generated. In these plots, the gray regions indicate the P10–P90 interval for the prior models, the solid black curves denote the P10, P50 and P90 DSI results, the dashed blue curves are the analogous RS results, the dashed red curves represent the true model response, and the red points are the observed data. Immediately evident is the large amount of uncertainty reduction accomplished, even though the amount of assimilated data in this case is limited. Of most interest here, however, is the generally close agreement between DSI results and those from the much more expensive RS procedure (though some discrepancies are evident, such as in P3 WPR in Fig. 10c). The CDFs in the plots in the right columns quantify both the level of agreement between DSI and RS, and the amount of uncertainty reduction in the various well quantities.

In Fig. 10a, the true P3 OPR data are seen to be outside of the P10–P90 range of predictions from the prior models at around 1000 days. The DSI procedure is nonetheless able to provide accurate uncertainty quantification results, with the true data falling very near the P90 curve. Also of interest are the results for P3 WPR (Fig. 10c). Here, although water breakthrough is not observed during the history-matching period, prediction uncertainty is still reduced significantly. In addition, the CDFs at 1500 days (Fig. 10d) display a close match. Though not shown here, for P3 WPR we achieved DSI results that matched the RS results very closely when the pre-selection step, described in Sect. 3.5, was applied with $N_u = 100$. Taken in total, the general agreement between the DSI and RS results, in terms of P10, P50 and P90 predictions and posterior CDFs, is quite acceptable.

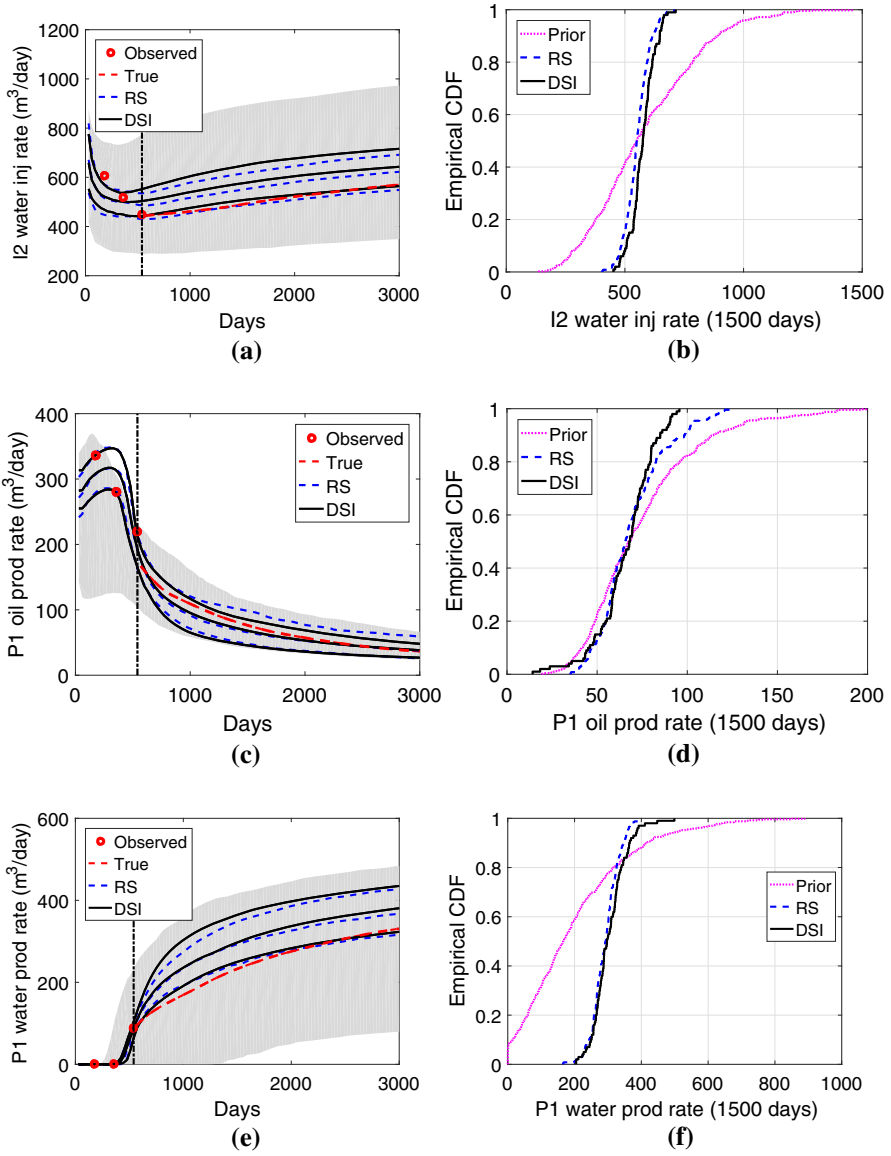

Fig. 9 Statistics of production forecasts, for I2 and P1, from prior models, RS and DSI. *Left plots* compare the P10, P50 and P90 results obtained from DSI (*black curves*) and RS (*dashed blue curves*). The *gray shaded areas* represent the P10–P90 range of predictions from prior models. *Right plots* compare the cumulative distribution function (CDF) of production forecasts at 1500 days from prior models, RS and DSI. **a** Water rate (I2), **b** water rate CDF (I2), **c** oil rate (P1), **d** oil rate CDF (P1), **e** water rate (P1), **f** water rate CDF (P1)

4.5 DSI Uncertainty Quantification with Different 'True' Models

In the previous section, we compared forecasts obtained from RS and DSI for a single selected 'true' model, from which the observed data were generated. Satija and Caers

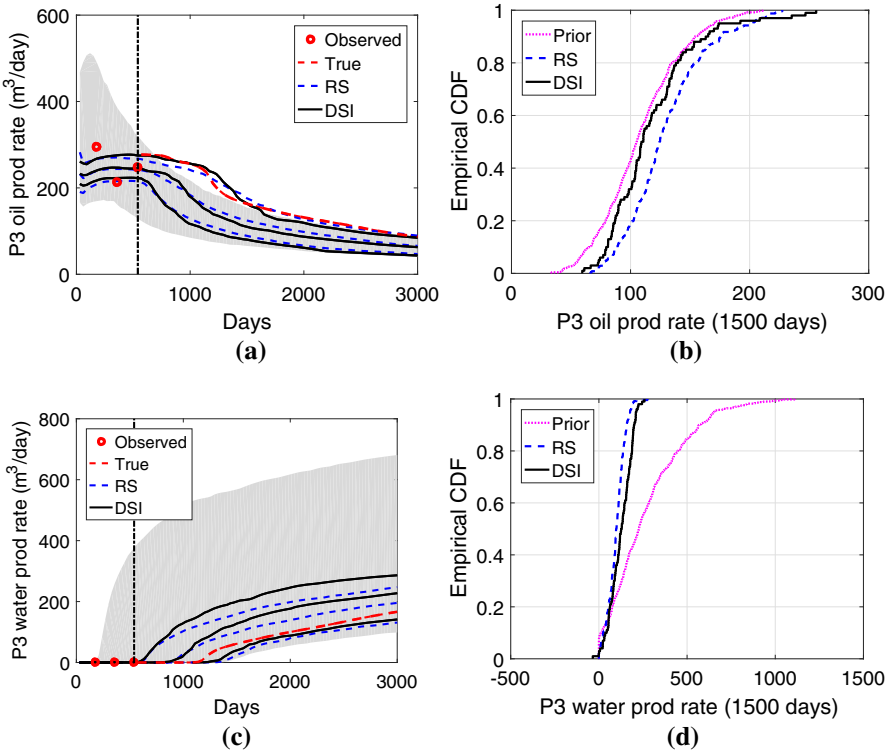

Fig. 10 Statistics of production forecasts for P3 from prior models, RS and DSI. *Curves* and *colors* in this figure have the same meaning as in Fig. 9. **a** Oil rate (P3), **b** oil rate CDF (P3), **c** water rate (P3), **d** water rate CDF (P3)

(2015) showed that their data-space approach was not applicable when the synthetically generated observed data were at the edge of the simulated data from the prior models. This observation indicates that the forecasting results depend on how the reference true data are distributed in the prior data space. In order to assess this issue within the context of DSI, the procedure will now be applied for ten different 'true' models.

RS is again applied to provide the reference forecasting results. We use the same set of 10^6 prior models as were used in the previous section when applying RS for the different 'true' models. Note that the ten 'true' models are not contained in the set of prior models used by either RS or DSI. In these examples, we include fewer observations to assure that a sufficient number of models (100 or more) are accepted in all cases. The observations are again at 180, 360 and 540 days, but now only include data from wells I1 and P3, which results in a total of nine observations. The standard deviation of the measurement error is again set to 15% of the corresponding data. Using the 15 observations considered in Sect. 4.4 resulted in RS acceptance of as few as ten models (out of 10^6) for some 'true' cases. A total of 100 predictions are again generated with DSI for each of the different 'true' models.

Figures 11, 12 and 13 present results, in the form of box plots, for cumulative water injection at I1, cumulative water production at P3 and cumulative oil production at P3,

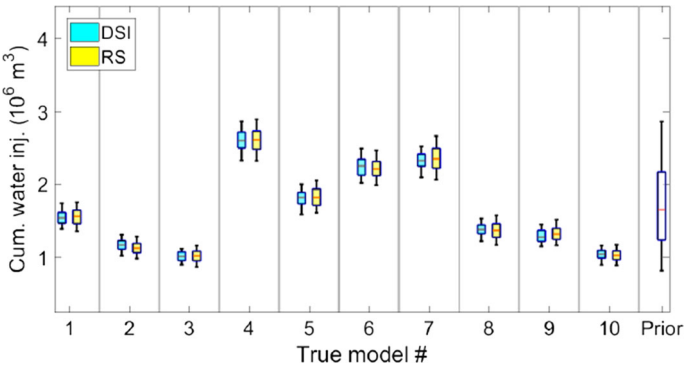

Fig. 11 Box plots of cumulative water injection at I2 obtained from DSI, RS and prior models, for ten different 'true' models. The *red line within each box* indicates the median, the *bottom* and *top of each box* denote the P25 and P75 results, and the *ends of the lines extending out from the boxes* correspond to the P5 and P95 results

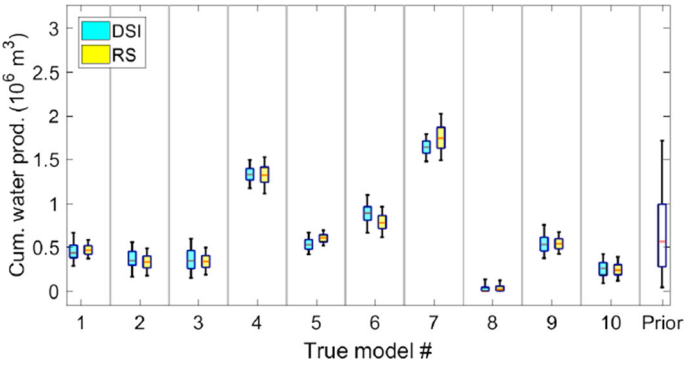

Fig. 12 Box plot of cumulative water production at P3. The *lines* and *colors* in this figure have the same meaning as in Fig. 11

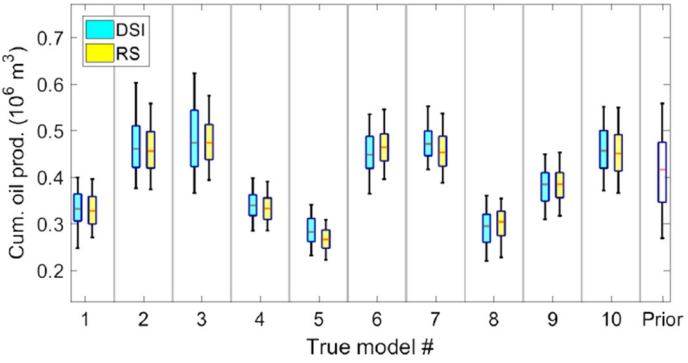

Fig. 13 Box plot of cumulative oil production at P3. The *lines* and *colors* in this figure have the same meaning as in Fig. 11

all after 3000 days (which is the simulation time frame). Box plots corresponding to predictions from the prior models are also shown. In the box plots, the red line within each box designates the median, the bottom and top of each box correspond to the P25 and P75 results, and the ends of the lines extending out from the boxes indicate the P5 and P95 results. The relatively close agreement between the RS and DSI results for all three quantities in all ten models is evident, as is the reduction in uncertainty relative to that for the prior models.

From the figures, it is apparent that the posterior distributions corresponding to some of the 'true' models (e.g., 'true' models 7 and 8 in Fig. 12) are at the edges of the prior distribution. However, DSI still provides posterior distributions that are reasonably close to those from the reference RS procedure. This observation indicates that DSI is able to, at least in these cases, provide reasonable posterior distributions even when the true data are far from most of the prior-model predictions.

4.6 DSI Uncertainty Quantification with More Observed Data

As a final test for this problem, DSI predictions are now generated for a case in which more observed data are available. RS results will not be shown for this example as it is computationally infeasible to accept enough models with the amount of data used here. The model setup is the same as described in Sect. 4.1. The observed data are measured every 60 days, from 60 to 600 days, for every well. Simulation data are reported every 30 days until 3000 days. Thus, we have $\mathbf{d}_{\text{obs}} \in \mathbb{R}^{80\times 1}$ (80 data points) and $\mathbf{d}_{\text{full}} \in \mathbb{R}^{800\times 1}$. The standard deviation of the measurement error is 10% of the corresponding data.

The same 500 prior models are used with DSI. Figure 14 shows the P10, P50 and P90 DSI results (black curves) along with the prior P10–P90 interval (gray shaded area). Uncertainty reduction after conditioning to observed data is significant for all quantities, especially for water injection and production rates. Note that the true data for some wells (Fig. 14a, f, g) are outside of the P10 to P90 range of predictions from the prior models, which means that the true data are not well captured by the prior models. However, the DSI approach is still able to provide posterior forecasts in which the true data largely fall within the P10–P90 interval, except at a few times in Fig. 14e, g.

5 Numerical Example: Case 2

In this section, the DSI procedure will be applied for a more realistic case. The system is now three-dimensional and contains oil, water and gas, and the field proceeds through primary recovery followed by waterflood, with well settings changing during these different stages of production. DSI results for this case will again be compared to those from RS.

5.1 Model Setup and Predictions from Prior Models

The reservoir now has an anticlinal (folded) structure, with the depth of the top layer ranging from 2400 m near the center to 2450 m at the flank. The geocellular model is

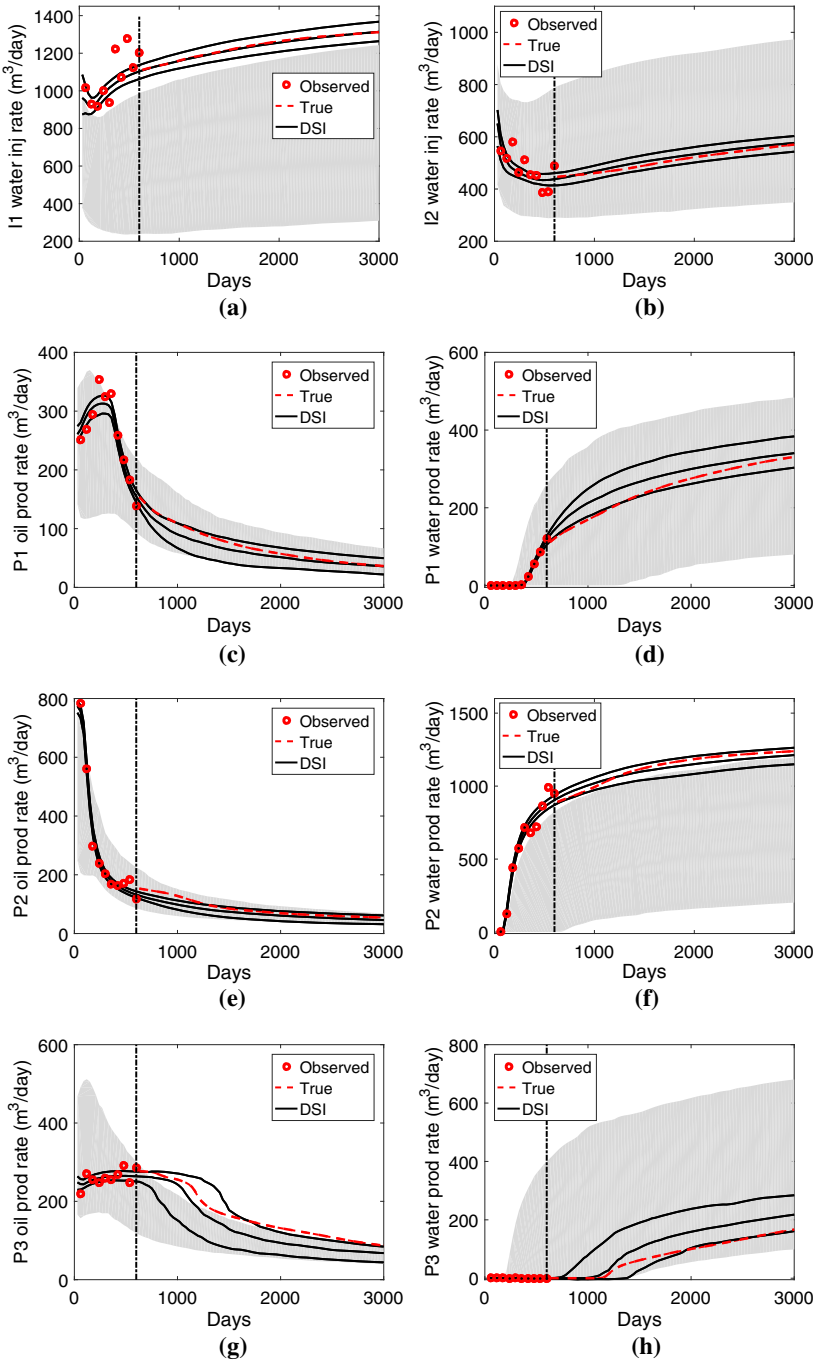

Fig. 14 DSI production forecasts based on more data. *Curves* in this figure have the same meaning as in Fig. 9. **a** Water rate (I1), **b** water rate (I2), **c** oil rate (P1), **d** water rate (P1), **e** oil rate (P2), **f** water rate (P2), **g** oil rate (P3), **h** water rate (P3)

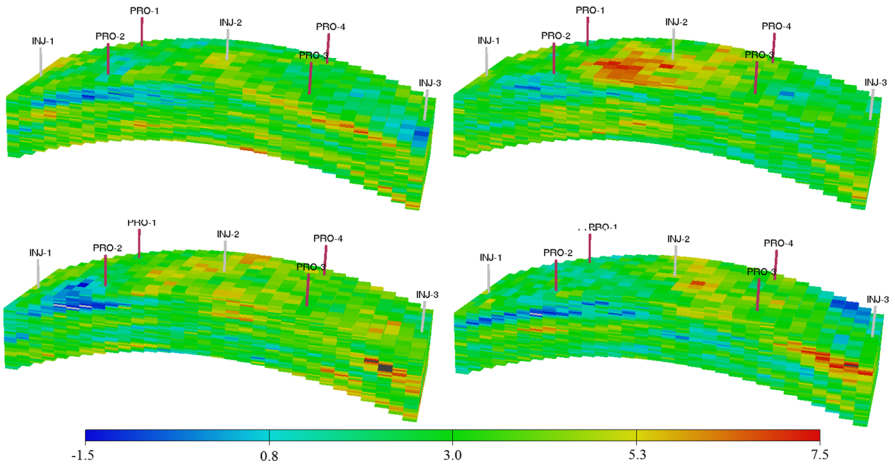

Fig. 15 Log-permeability (k, in md) maps for four prior models. Well locations are also shown ('INJ' and 'PRO' denote injector and producer; these are referred to as 'I' and 'P' in the text)

constructed on a $31 \times 11 \times 40$ grid with each grid block of size 50 m $\times$ 50 m $\times$ 2 m. The permeability field is Gaussian, and the parameter k in each grid block is generated using SGeMS (Remy et al. 2009) with an exponential variogram function.

Four randomly generated realizations, all conditioned to honor the permeability (k) at well blocks, are shown in Fig. 15. The histogram of $\ln k$ is Gaussian with a mean of 3 and standard derivation of 1.5. The directional permeability for each block is specified as $k_x = k_y = k$ and $k_z = 0.3k$. The porosity is again constant (0.2) for all grid blocks. A total of four producers and three injectors, all fully penetrating, are introduced into the model.

The initial saturations of oil and water above the water–oil contact, at a depth of 2510 m, are 0.8 and 0.2, respectively. Only the water phase exists below the oil–water contact. All gas is initially dissolved in the oil phase. Fluid properties are the same as those used in the PUNQ-S3 simulation model (Barker et al. 2001; Floris et al. 2001). Capillary pressure effects are ignored. The initial reservoir pressure is 234 bar.

In this example, the simulations are performed using the commercial simulator ECLIPSE (Schlumberger 2013). The simulation period is 3000 days, with production data reported every 30 days. The first 900 days correspond to primary production, during which all producers operate at a fixed oil rate (200 m^3/day) subject to a minimum BHP of 100 bar. There is no water injection during this period. Water injection (from all three injectors) starts at 900 days, with the injection wells operated at a BHP of 500 bar. During the water injection phase, the producers operate at a fixed oil rate control of 250 m^3/day subject to a minimum BHP of 100 bar.

Figure 16 shows the simulation results, for wells I2 and P2, for 100 prior models. We see that the BHP in P2 (Fig. 16b) decreases monotonically during the primary production period (first 900 days) before reaching the lower limit (100 bar). Oil production (OPR) in P2 starts to decline once the minimum BHP is reached (Fig. 16c). After water injection starts (at 900 days, see Fig. 16a), OPR in P2 increases until reaching

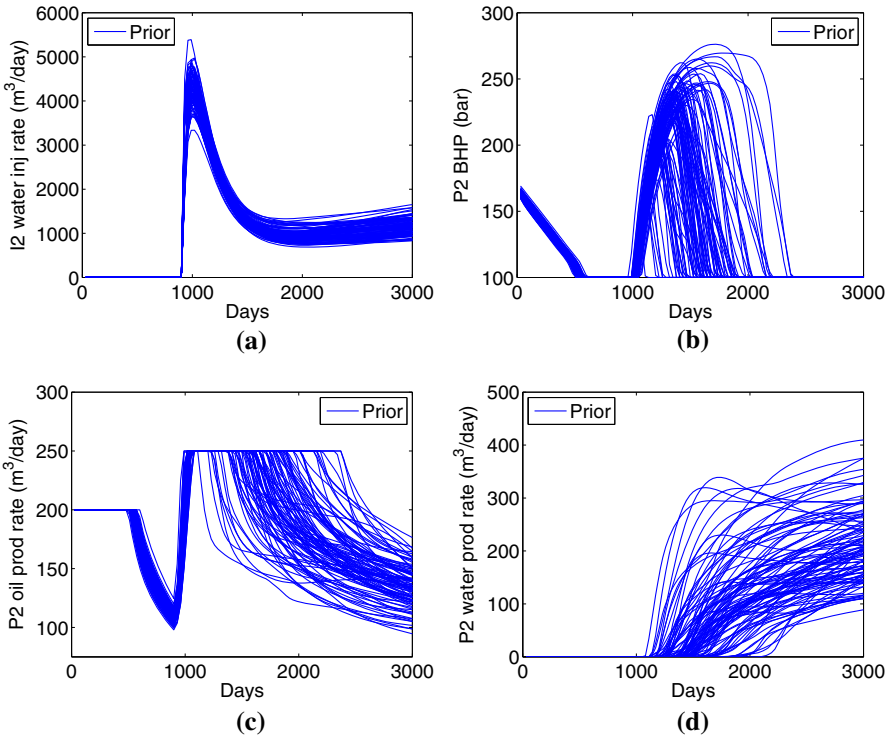

Fig. 16 Production forecasts from prior models (Case 2). **a** Water rate (I2), **b** bottom-hole pressure (P2), **c** oil rate (P2), **d** water rate (P2)

the target rate of 250 m^3/day. The BHP in P2 also increases as water is injected, but it later declines due to water breakthrough, which leads to a decrease in OPR. Water breakthrough is observed for all prior models (Fig. 16d). Free gas is produced in only a few of the prior models (most of the gas stays dissolved in oil, where it impacts oil phase properties).

5.2 Results Comparison

In this case, we generate observed data from 360 to 1800 days (measured every 360 days) at I2 and P2, which provides $\mathbf{d}_{\text{obs}} \in \mathbb{R}^{20\times 1}$. The measurement errors are again assumed to be independent Gaussian with zero mean and standard derivation of 10% of the corresponding true data. Relatively little data are again used to enable RS to accept a reasonable fraction of models. A total of $N_{\text{r}} = 500$ models are used for DSI, and 500,000 models are used for RS (which resulted in 218 accepted models).

Results for production data at I2 and P2 are predicted until 3000 days. Thus, we have $(\mathbf{d}_{\text{full}})_i \in \mathbb{R}^{400\times 1}$ $(i = 1, 2, \ldots, 500)$. With DSI, mapping operations are applied for the OPR, BHP and WPR data at well P2. No mapping operations are applied for the WIR (I2) data. Figure 17 shows the mapped data corresponding to the results in

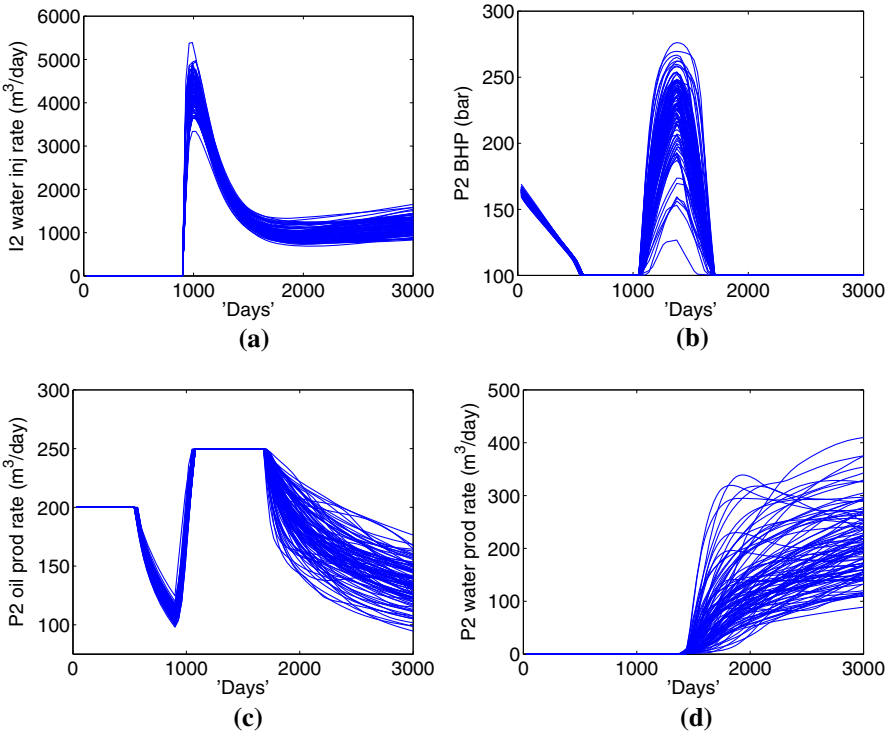

Fig. 17 Production forecasts after mapping operations. The transition times are included in the mapped data, but are not shown in the figure. **a** Water rate (I2), **b** bottom-hole pressure (P2), **c** oil rate (P2), **d** water rate (P2)

Fig. 16. The mapping for the WPR data is as described in Sect. 3.1. The mapping operations for the BHP and OPR data (detailed in the Appendix) are slightly more complicated, but the basic approaches are the same as for the WPR mapping.

Figures 18 and 19 show the posterior distributions obtained from DSI and RS. Close agreement in terms of both the P10, P50 and P90 results, and the empirical CDFs, is again observed. We see a decrease in prediction uncertainty after application of DSI even though there are relatively few observations and measurement error is relatively large. The true data essentially fall within the P10–P90 interval (for the I2 water rate, the interval is very narrow and the data lie right around the P90 result). It is significant that no unphysical DSI predictions are observed (i.e., no overshoot or undershoot of target rates or BHPs occurs), which indicates that the mapping operations and other DSI treatments are applicable to cases involving multiple changes in well controls.

6 Additional Issues

For practical subsurface problems, full model-inversion can be very challenging due to geological complexities and the nonlinear relationship between model parameters and the flow response. However, the distribution of flow responses in the data space may

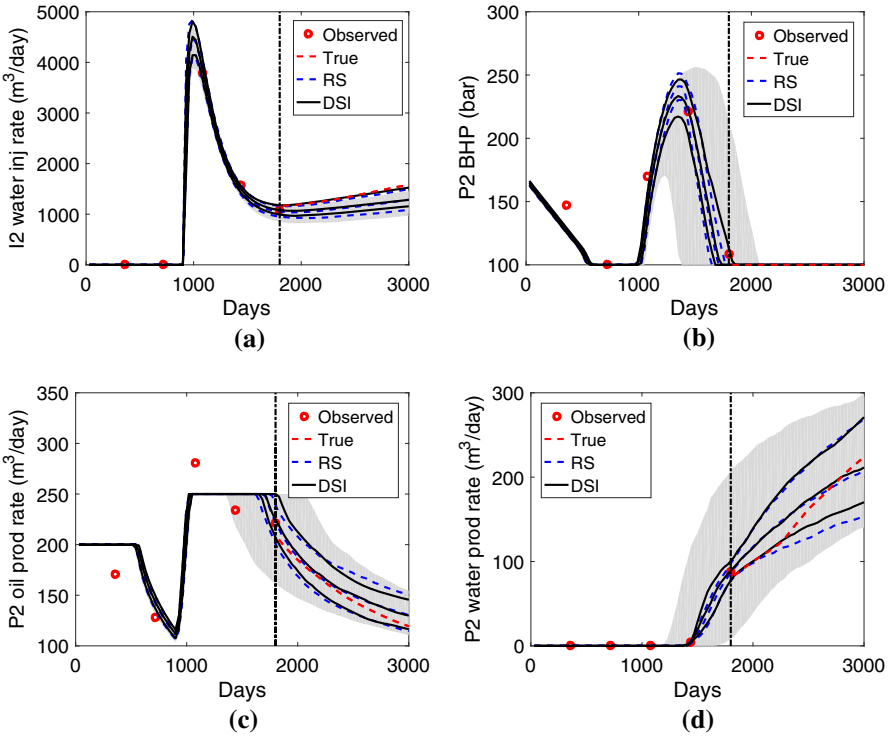

Fig. 18 Statistics of production forecasts from prior models, RS and DSI. *Curves* and *colors* in this figure have the same meaning as the *left plots* in Fig. 9. **a** Water rate (I2), **b** bottom-hole pressure (P2), **c** oil rate (P2), **d** water rate (P2)

be less complicated, which makes it possible (and appealing) to apply the DSI method when the forecasts and associated uncertainty are of main interest. The numerical results in this paper show that the DSI method is able to provide results similar to those obtained from the exhaustive rejection sampling approach. This suggests that DSI is indeed applicable for data inversion and uncertainty quantification.

The DSI method is formulated under a Bayesian framework, with the prior distribution of data variables estimated using the flow responses corresponding to an ensemble of prior models. The posterior/conditional distribution of data variables given observations is then directly sampled in the data space, without running additional flow simulations. Therefore, DSI should not be used to generate forecasts outside the bounds of forecasts from the prior models. A wider prior (e.g., realizations from additional geological scenarios) should be used in cases where the observed data are outside the forecasts from the existing prior models.

In the examples presented in this work, $N_r = 500$ prior models were used. The choice of N_r, however, depends on several factors, including available computational resources, dimension of the data space and variability of the data. It is expected that more prior models will be required as the dimension of the data space becomes larger, or when the data display large variability (in other words, when the distances between

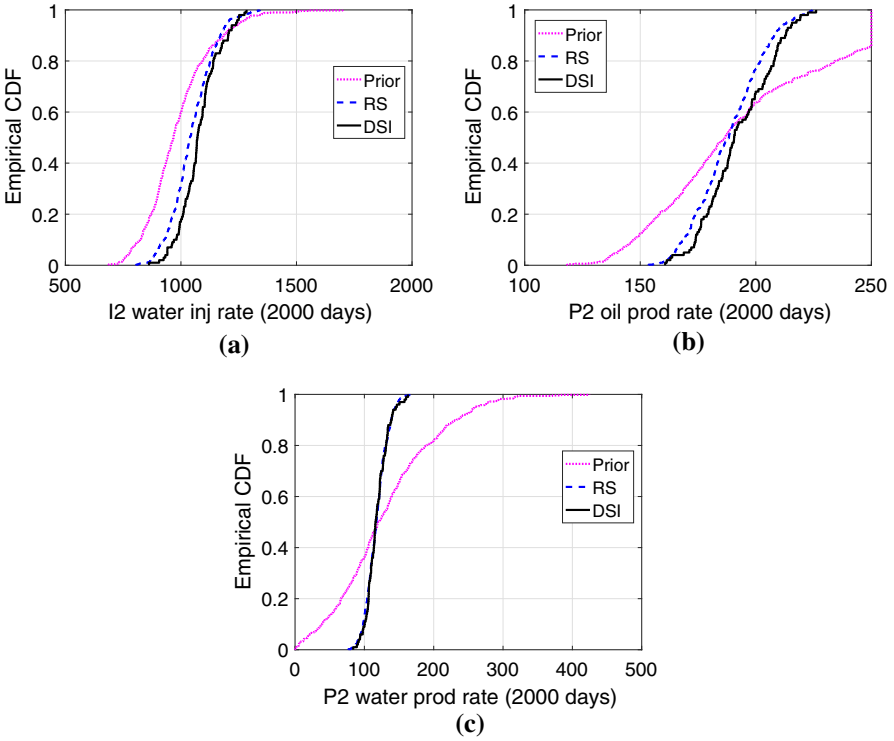

Fig. 19 Statistics of production forecasts from prior models, RS and DSI. *Curves* and *colors* in this figure have the same meaning as the *right plots* in Fig. 9. **a** Water rate CDF (I2), **b** oil rate CDF (P2), **c** water rate CDF (P2)

prior samples in the data space are large). In practice, Eq. (25) can be used to determine whether the number of prior models is sufficient. If a significant portion (e.g., 50%) of the generated DSI estimates does not satisfy Eq. (25) during the RML procedure, it might be necessary to consider a larger number of prior models. This will ensure that there are a sufficient number of samples in the data space that are 'close' to the observed data.

In this work, modeling error associated with the forward simulations was not considered (this error is commonly neglected in the literature). However, in actual applications, modeling error can be important and may even dominate measurement error. In that case, the posterior distribution in Eq. (22) should be modified to account for the impact of modeling error. Although discussion of specific representations of modeling error is beyond the scope of this paper, it is important to note that this effect may be incorporated into C_D, which appears in Eq. (22) (Tarantola 2005; Oliver et al. 2008). This means that DSI can be used to efficiently explore the impact of different modeling error treatments, since the posterior distribution in Eq. (22) can be evaluated without running any additional flow simulations. This is a clear advantage of DSI relative to traditional model-inversion approaches, as the latter would require repeatedly performing time-consuming inversions for each C_D considered.

7 Conclusions

In this paper, a novel data-space inversion (DSI) procedure was developed for reservoir performance forecasting given observed production data. In this procedure, only a set of prior reservoir models and corresponding flow simulation results are required. The predictions conditioned to observations are generated by appropriately combining the (mapped) simulation results from prior reservoir models. A randomized maximum likelihood (RML) algorithm in the data space was introduced to sample the conditional distribution of data variables given observed data. We also developed mapping operations and applied principal component analysis (PCA) to transform data variables to new variables that are closer to multivariate Gaussian. The performance of DSI was shown to improve when these mapping operations were applied.

Detailed results quantifying posterior production uncertainty were presented for two examples involving bimodal channelized geology and a Gaussian geological description. Rejection sampling (RS) was performed in both cases to provide reference results for uncertainty quantification. RS, however, requires $O(10^5–10^6)$ simulation runs for the cases considered, and even more simulations would be required with RS as the amount of observed data increases. DSI, using only 500 prior simulations, was shown to provide uncertainty quantification results in reasonable agreement with those from RS. The accuracy of DSI relative to RS was shown to hold over ten different 'true' models in the first example and for a complicated production scenario (primary production followed by waterflood, with several switches in well control) in the second example. These results suggest that DSI will be applicable for a range of challenging problems. A pre-selection procedure was suggested to improve DSI results by eliminating some of the prior simulation data. Although this treatment was not used in the results presented here, it appears to be useful and should be assessed and formalized in future work.

The DSI approach does not provide posterior reservoir models, and this would be a disadvantage in applications where such models are needed. However, for problems where the predicted reservoir response and the uncertainty in that response are of primary interest, DSI may be preferable to traditional model-based inversion methods. For some types of problems, such as those involving discrete fracture models, the grid is in general different for each geological realization. This poses challenges for many model-based approaches, which commonly assume that the grid is the same in all (prior and posterior) models. This is not an issue for DSI, since the method works only with observed production data and entails no assumptions regarding the grid. Along these lines, the successful application of DSI to a realistic naturally fractured system was recently presented by Sun et al. (2016).

Another application in which DSI could be very useful is in the design of monitoring/surveillance systems (Le and Reynolds 2014; He et al. 2015). In such problems, the goal is to, for example, determine the locations of monitoring wells such that the expected reduction in uncertainty is maximized. The inner loop in such an optimization is a data assimilation, and DSI appears to be well suited for this. In addition, because DSI can be run very quickly with different levels of measurement noise (or various treatments for modeling error), the impact of this effect on forecast uncertainty can be efficiently assessed. Finally, it will be of interest to apply DSI to more realistic

problems with many wells and large amounts of observed data. For such cases, prior models could include uncertainty in the training image/geological scenario or relative permeability functions, in addition to uncertainty in the permeability field.

Acknowledgements We thank Chevron ETC and the Stanford Smart Fields Consortium for financial support. We are grateful to Hai Xuan Vo, David Cameron, Celine Scheidt, Vladislav Bukshtynov and Oleg Volkov for useful discussions and assistance with simulation software.

Appendix: Detailed Mapping Operations

In this Appendix, the general pattern-based mapping operations applied in this study are described. For generality and simplicity of notation, we describe the mapping operations for an ensemble of time-series functions $y_i(t)$ $(i = 1, 2, \ldots, N_r)$. Note that the data values in $(\mathbf{d}_{\text{full}})_i$, discussed in Sect. 3.1, are simply the values of these time-series functions at different time steps.

It is assumed that the functions can all be separated into the same number of stages, with the same general behavior within each stage. For example, the functions in Fig. 20a can be separated into four stages: constant, decline, constant and decline. The times separating the different stages are referred to as transition times. The goal is then to map these functions to those shown in Fig. 20b, in which the corresponding transition 'times' for all functions are the same.

The overall starting and ending times, t_s and t_e, are assumed to be the same for all functions. A total of $M + 1$ stages are then identified ($M = 3$ for the cases shown in Fig. 20a), and the corresponding transition times are designated t_i^j $(i = 1, 2, \ldots, N_r;\ j = 1, 2, \ldots, M)$. Defining $t_i^0 = t_s$ and $t_i^{M+1} = t_e$, the mapped functions are then constructed as

$$\hat{y}_i(\tau) = y_i\left(t_i^j + (t_i^{j+1} - t_i^j)\frac{\tau - \tau^j}{\tau^{j+1} - \tau^j}\right), \quad \tau^j \le \tau \le \tau^{j+1}, \quad 0 \le j \le M, \tag{30}$$

where $\hat{y}_i(\tau)$ denotes the mapped function for member i in the ensemble, and τ^j is the transition 'time' for all mapped functions (note τ^j is the same for all functions).

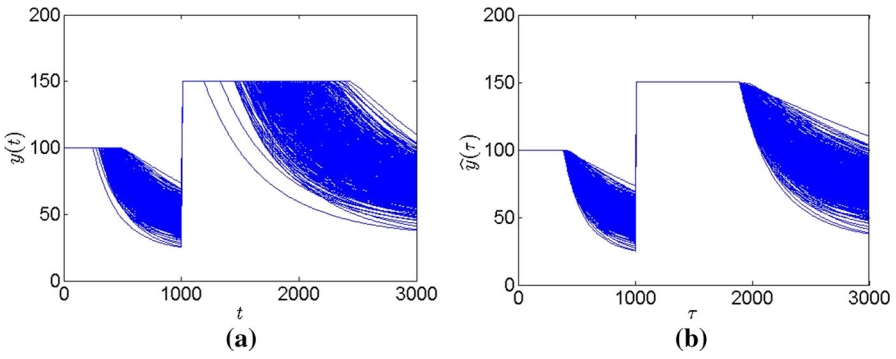

Fig. 20 Illustration of mapping operations. **a** Original functions, **b** functions after mapping

The values of τ^j ($j = 0, 1, \ldots, M+1$) must be predefined. In this paper, a particular transition 'time' is defined as the mean of the corresponding transition times for all of the original functions, i.e.,

$$\tau^j = \sum_{i=1}^{N_r} \frac{t_i^j}{N_r}, \quad 0 \le j \le M+1. \tag{31}$$

Figure 20b shows the mapped functions corresponding to the original functions in Fig. 20a, with transition 'times' defined by Eq. (31).

The forward mapping operation for $y_i(t)$ is expressed as

$$\mathscr{F} : y_i(t) \to \hat{y}_i(\tau),\ t_i^1,\ t_i^2, \ldots, t_i^M. \tag{32}$$

The transition times t_i^1 to t_i^M must also be included since they are needed for the backward mapping. Note that the values in the mapped data vectors $\widehat{\mathbf{d}}_i$, introduced in Sect. 3.1, are simply the values of $\hat{y}_i(\tau)$ at different 'time' steps, plus the identified transition times.

Once we construct a predicted mapped function, denoted by $\hat{y}_p(\tau)$, and its associated (predicted) transition times $t_p^1, t_p^2, \ldots, t_p^M$, the backward mapping is given by

$$\mathscr{F}^{-1} : \hat{y}_p(\tau),\ t_p^1,\ t_p^2, \ldots, t_p^M \to y_p(t), \tag{33}$$

where

$$y_p(t) = \hat{y}_p\left(\tau^j + (\tau^{j+1} - \tau^j)\frac{t - t_p^j}{t_p^{j+1} - t_p^j}\right), \quad t_p^j \le t \le t_p^{j+1},\ \ 0 \le j \le M. \tag{34}$$

Here τ^j are the pre-computed values from Eq. (31), and t_p^0 and t_p^{M+1} are the known start and end simulation times. This completes the description of the forward and backward mapping operations used to 'align' the transitions between the various stages.

An alternate mapping approach, based on the use of histogram transformations, was introduced in Sun et al. (2016). That procedure is applicable when the production data do not display clear patterns, as is the case when wells are abruptly shut and opened at different times. Because the histogram transformations are applied independently for each data variable, the nonlinear correlations between data variables (as shown in Fig. 3e) might not be as effectively mitigated with this treatment. In future work, it will be of interest to compare results with the two approaches for systems that display clearly identifiable stages.

References

Aanonsen SI, Nævdal G, Oliver DS, Reynolds AC, Vallès B (2009) The ensemble Kalman filter in reservoir engineering—a review. SPE J 14(3):393–412

Barker JW, Cuypers M, Holden L (2001) Quantifying uncertainty in production forecasts: another look at the PUNQ-S3 problem. SPE J 6(4):433–441

Cardoso MA, Durlofsky LJ, Sarma P (2009) Development and application of reduced-order modeling procedures for subsurface flow simulation. Int J Numer Methods Eng 77(9):1322–1350
Castro SA (2007) A probabilistic approach to jointly integrate 3D/4D seismic, production data and geological information for building reservoir models. PhD thesis, Stanford University
Chen Y, Oliver DS, Zhang D (2009) Data assimilation for nonlinear problems by ensemble Kalman filter with reparameterization. J Pet Sci Eng 66(1–2):1–14
Emerick AA, Reynolds AC (2013) Ensemble smoother with multiple data assimilation. Comput Geosci 55:3–15
Evensen G (2003) The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn 53(4):343–367
Evensen G, van Leeuwen PJ (2000) An ensemble Kalman smoother for nonlinear dynamics. Mon Weather Rev 128(6):1852–1867
Floris F, Bush M, Cuypers M, Roggero F, Syversveen AR (2001) Methods for quantifying the uncertainty of production forecasts: a comparative study. Pet Geosci 7(SUPP):87–96
Gao G, Reynolds AC (2006) An improved implementation of the LBFGS algorithm for automatic history matching. SPE J 11(1):5–17
Gao G, Zafari M, Reynolds AC (2006) Quantifying uncertainty for the PUNQ-S3 problem in a Bayesian setting with RML and EnKF. SPE J 11(4):506–515
Gu Y, Oliver DS (2006) The ensemble Kalman filter for continuous updating of reservoir simulation models. J Energy Resour Technol 128(1):79–87
He J, Xie J, Sarma P, Wen XH, Chen WH, Kamath J (2015) Model-based a priori evaluation of surveillance programs effectiveness using proxies. Paper SPE 173229 presented at the SPE reservoir simulation symposium, Houston, Texas, USA, 23–25 February
Kitanidis PK (1986) Parameter uncertainty in estimation of spatial functions: Bayesian analysis. Water Resour Res 22(4):499–507
Krishnamurti TN, Kishtawal C, Zhang Z, LaRow T, Bachiochi D, Williford E, Gadgil S, Surendran S (2000) Multimodel ensemble forecasts for weather and seasonal climate. J Clim 13(23):4196–4216
Le DH, Reynolds AC (2014) Estimation of mutual information and conditional entropy for surveillance optimization. SPE J 19(4):648–661
Liang Y, Lee H, Lim S, Lin W, Lee K, Wu C (2002) Proper orthogonal decomposition and its applications—part I: theory. J Sound Vib 252(3):527–544
Mallet V, Stoltz G, Mauricette B (2009) Ozone ensemble forecast with machine learning algorithms. J Geophys Res 114(D5):148–227
Mohaghegh SD (2005) Recent developments in application of artificial intelligence in petroleum engineering. J Pet Technol 57(04):86–91
Mosegaard K, Tarantola A (1995) Monte Carlo sampling of solutions to inverse problems. J Geophys Res Solid Earth 100(B7):12,431–12,447
Oliver DS (1996) Multiple realizations of the permeability field from well test data. SPE J 1(2):145–154
Oliver DS, Chen Y (2011) Recent progress on reservoir history matching: a review. Comput Geosci 15(1):185–221
Oliver DS, He N, Reynolds AC (1996) Conditioning permeability fields to pressure data. In: 5th European conference on the mathematics of oil recovery, Leoben, Austria, 3–6 September
Oliver DS, Reynolds AC, Liu N (2008) Inverse theory for petroleum reservoir characterization and history matching. Cambridge University Press, Cambridge
Pagowski M, Grell G, McKeen S, Dévényi D, Wilczak J, Bouchet V, Gong W, McHenry J, Peckham S, McQueen J (2005) A simple method to improve ensemble-based ozone forecasts. Geophys Res Lett. doi:10.1029/2004GL022305
Park H, Scheidt C, Fenwick D, Boucher A, Caers J (2013) History matching and uncertainty quantification of facies models with multiple geological interpretations. Comput Geosci 17(4):609–621
Remy N, Boucher A, Wu J (2009) Applied geostatistics with SGeMS: a user's guide. Cambridge University Press, Cambridge
Reynolds AC, He N, Chu L, Oliver DS (1996) Reparameterization techniques for generating reservoir descriptions conditioned to variograms and well-test pressure data. SPE J 1(4):413–426
Reynolds AC, He N, Oliver DS (1999) Reducing uncertainty in geostatistical description with well-testing pressure data. In: Schatzinger RA, Jordan JF (eds) Reservoir characterization—recent advances. American Association of Petroleum Geologists, Tulsa, pp 149–162

Sarma P, Durlofsky LJ, Aziz K, Chen W (2006) Efficient real-time reservoir management using adjoint-based optimal control and model updating. Comput Geosci 10(1):3–36
Satija A, Caers J (2015) Direct forecasting of subsurface flow response from non-linear dynamic data by linear least-squares in canonical functional principal component space. Adv Water Resour 77:69–81
Scheidt C, Renard P, Caers J (2015) Prediction-focused subsurface modeling: investigating the need for accuracy in flow-based inverse modeling. Math Geosci 47(2):173–191
Schlumberger, (2013) Eclipse reference manual. Version 2013.2
Shlens J (2005) A tutorial on principal component analysis. http://www.cs.cmu.edu/~elaw/papers/pca.pdf
Strebelle S (2002) Conditional simulation of complex geological structures using multiple-point statistics. Math Geosci 34(1):1–21
Sun W (2014) Data driven history matching for reservoir production forecasting. Master's thesis, Stanford University
Sun W, Durlofsky LJ, Hui MH (2016) Production uncertainty quantification for a naturally fractured reservoir using a new data-space inversion procedure. In: 15th European conference on the mathematics of oil recovery, Amsterdam, Netherlands, 29 August–1 September
Tarantola A (2005) Inverse problem theory and methods for model parameter estimation. SIAM, Philadelphia
Vo HX, Durlofsky LJ (2014) A new differentiable parameterization based on principal component analysis for the low-dimensional representation of complex geological models. Math Geosci 46(7):775–813
Vo HX, Durlofsky LJ (2015) Data assimilation and uncertainty assessment for complex geological models using a new PCA-based parameterization. Comput Geosci 19(4):747–767
Zhou Y (2012) Parallel general-purpose reservoir simulation with coupled reservoir models and multisegment wells. PhD thesis, Stanford University