

Data-space approaches for uncertainty quantification of CO₂ plume location in geological carbon storage

Wenyue Sun*, Louis J. Durlofsky

Department of Energy Resources Engineering, Stanford University, USA



ARTICLE INFO

Keywords:

Carbon storage
Sequestration
Data assimilation
Uncertainty quantification
Data-space inversion
Optimal monitoring

ABSTRACT

A data-space inversion (DSI) method is developed and applied to quantify uncertainty in the location of CO₂ plumes in the top layer of a storage aquifer. In the DSI procedure, posterior predictions of the CO₂ saturation distribution are generated using simulation results for prior geostatistical model realizations along with observed data, which in this case derive from observations at monitoring wells. Posterior (history-matched) geological models are not constructed in the DSI method, so many of the complications that arise in traditional data assimilation methods are avoided. The DSI method treats quantities of interest (QoI), such as the CO₂ saturation distribution in the top layer, as random variables. The posterior distribution for these QoI, conditioned to observed data, is formulated in the data space within a Bayesian framework. Samples from this posterior distribution are generated using the randomized maximum likelihood method. We also introduce a procedure to optimize the locations of monitoring wells using only prior-model simulation results. This approach is based on analytical DSI results, and determines monitoring well locations such that the reduction in expected posterior variance of a relevant quantity is maximized. The new DSI procedure is applied to three-dimensional heterogeneous aquifer models involving uncertainties in a wide range of geological parameters, including variogram orientation, porosity and permeability fields, and regional pressure gradient. Multiple monitoring scenarios, involving four to eight monitoring wells, are considered in the evaluation of our data space methodology. Application of DSI with optimal monitoring wells is shown to consistently reduce the posterior variance of average CO₂ saturation in the top layer, and to provide detailed saturation fields in reasonable correspondence with the ‘true’ saturation distribution. Finally, we demonstrate consistent improvement in DSI predictions as data are collected from an increasing number of (optimized) monitoring wells.

1. Introduction

The geological storage of CO₂ in deep saline aquifers represents a potential strategy for large-scale carbon mitigation. It is essential that the movement of the injected fluid be well understood, as such knowledge can be used to mitigate the risk of CO₂ leakage, which could otherwise lead to the contamination of fresh water resources. Because the flow models used for prediction include many highly uncertain parameters, such as porosity and permeability defined on simulation grid blocks, there will typically be large uncertainty associated with prior flow predictions. The incorporation of various types of monitoring data enables some amount of parameter calibration, leading to reduced uncertainty in predictions. There are however still challenges, both conceptual and computational, associated with conventional model-based inversion procedures. Our goal in this work is to present an alternative approach that circumvents many of the issues with model-based approaches.

In recent work, we developed data-space inversion (DSI) procedures to enable efficient predictions of well-based quantities, such as time-varying phase production and injection rates, in oil field problems (Sun and Durlofsky, 2017; Sun et al., 2017b). DSI differs considerably from traditional model-based inversion in that it directly samples the posterior distribution of quantities of interest (QoI), given measurements, within a Bayesian framework. DSI uses only prior-model simulation results and observed data to provide statistical predictions. It does not provide posterior (history-matched) models, which is a limitation in some contexts. In this work, we generalize the DSI methodology to enable effective CO₂ plume prediction under uncertainty. We also present an efficient procedure, within the DSI framework, to determine the (optimal) locations of monitoring wells such that expected uncertainty reduction in a relevant quantity is maximized.

Traditional model-based history matching has been actively studied in both oil field and groundwater applications. General discussion of various history matching algorithms can be found in Tarantola (2005), Oliver et al. (2008), Zhou et al. (2014), Oliver and Chen (2011),

* Corresponding author.

E-mail address: wenyuesunny@gmail.com (W. Sun).

Yeh (1986) and McLaughlin and Townley (1996). Ensemble-based history matching algorithms, such as ensemble Kalman filter and ensemble smoothers, are described in Evensen (2003), Anan센 et al. (2009), Emerick and Reynolds (2013), Chen and Zhang (2006) and González-Nicolás et al. (2015). Within the context of geological carbon sequestration, Sun and Nicot (2012) and Sun et al. (2013) used a probabilistic collocation method to assess the detectability of leakage using pressure data measured from monitoring wells, for an uncertain heterogeneous aquifer. González-Nicolás et al. (2015) compared the use of ensemble smoother and restart ensemble Kalman filter algorithms for the detection of potential CO₂ leakage pathways using pressure data. Espinet et al. (2013) investigated the use of a global optimization algorithm for model calibration (specifically the determination of facies permeability values) by matching a limited amount of pressure and gas monitoring data. Cameron et al. (2016) investigated the assimilation of pressure data in the zone above the storage aquifer, with the goal of locating and quantifying potential leaks during CO₂ storage operations. They used a particle swarm optimization algorithm, together with a Karhunen-Loève representation of porosity for model reduction, to find history-matched aquifer models. However, in both Espinet et al. (2013) and Cameron et al. (2016), the optimization procedures were very time consuming, and only a single history-matched model was constructed, which does not allow for uncertainty quantification.

In field application settings, several studies have applied history matching to improve geological models for prediction of CO₂ migration or leakage detection (Daley et al., 2008; Doughty et al., 2008; Ennis-King et al., 2011; Oladyshkin et al., 2011). However, many of these studies only considered simplified layered geological models with relatively few uncertain parameters. The observations were generally matched manually in these studies. For example, Daley et al. (2008) and Doughty et al. (2008) calibrated numerical models through a trial and error approach, with the target of matching the field pilot data, including seismic data, from the Frio formation in Dayton, Texas (USA). Ennis-King et al. (2011) manually matched downhole pressure measurements and the gas arrival time at observation wells through multiple stages.

Several approaches have been proposed recently to provide posterior forecasts without model updating. Scheidt et al. (2015) and Satija and Caers (2015) developed prediction-focused analysis (PFA) for predicting tracer flow at a few downstream locations. PFA aims to build a direct statistical relationship between the observations and QoI, using the simulated data from an ensemble of prior models. This statistical relationship is then used to provide predictions when actual observations are made. In the context of weather forecasting, Krishnamurti et al. (2000) and Mallet et al. (2009) linearly combined predictions from a set of prior models, with weights determined such that the predictions matched the historical data. The PFA approach and ensemble approaches described in Krishnamurti et al. (2000) and Mallet et al. (2009) may have some limitations including linear assumptions and the treatment of measurement error. Detailed discussion of these issues, and other related data-space approaches, can be found in Sun and Durlofsky (2017), Sun et al. (2017b) and Sun (2014). Our DSi procedure has been previously applied, within the context of oil field problems, to bimodal channelized systems (Sun and Durlofsky, 2017) and naturally fractured reservoirs (Sun et al., 2017b). In both studies, we observed reasonable agreement in forecasts between DSi and reference results obtained from a strict rejection sampling algorithm (Oliver et al., 2008) for well-based QoI. This demonstrates that DSi can provide appropriate posterior uncertainty quantification for the QoI considered.

In this study, we extend the DSi procedure to enable the prediction of the CO₂ saturation distribution in the top layer of a storage aquifer. The CO₂ distribution, which represents essential information for risk mitigation, is a spatial quantity, in contrast to the (local) time-varying well production quantities considered in Sun and Durlofsky (2017) and Sun et al. (2017b). The DSi methodology entails first generating and simulating a large ensemble of prior models (1000 models are used

for the examples in this paper; these are independent and can be run in parallel). This is the only time consuming step in DSi, and it provides a set of simulated ‘measurement’ and field CO₂ saturation data. We also present a data-space approach for the optimization of monitoring well placement. This procedure involves a multivariate Gaussian assumption on the statistics in data space, and the resulting expressions are consistent with the ensemble variance analysis algorithms applied in He et al. (2018). Given monitoring well data, posterior samples (distribution of CO₂ saturation in the top layer) are then constructed using a Bayesian formulation. These samples are generated using the randomized maximum likelihood method (Kitanidis, 1986; Oliver, 1996; Reynolds et al., 1999), which entails an inexpensive minimization involving a relatively small number of data-space variables.

This paper proceeds as follows. In Section 2, the DSi methodology for predicting the posterior distribution of CO₂ in the top layer is described. We then present the data-space procedure to optimize monitoring well locations. In Section 3, we give a detailed description of the three-dimensional uncertain heterogeneous aquifer model, present results for the optimal placement of monitoring wells, and apply our DSi procedure to provide posterior predictions. Several cases, involving different numbers of monitoring wells, are considered. The performance of DSi for ‘true’ models outside of the prior distribution, and the impact of grid refinement on CO₂ saturation fields and DSi results, are considered in Section 4. Concluding remarks and suggestions for future work are provided in Section 5.

2. Methodology

In this section, we introduce a data-space inversion (DSi) procedure for assimilation of measurement data. The DSi procedure is based on a Bayesian formulation in the data space, with the target of sampling the conditional/posterior distribution of quantities of interest (which in our case is the distribution of CO₂ in the top layer at a future time) given measurement data. In addition, based upon analytical DSi results, we then develop a data-space approach for a surveillance problem, in which the locations of monitoring wells are optimized to maximize the expected effectiveness of measurement data.

2.1. Data-space inversion formulation

We let \mathbf{m} denote the vector containing all unknown model parameters, including discrete (cell-by-cell) geological properties such as porosity and permeability. We express the forward model, which maps the input parameters \mathbf{m} to the output response, as

$$\mathbf{d}_f = \mathbf{g}(\mathbf{m}), \quad (1)$$

where $\mathbf{d}_f \in \mathbb{R}^{N_f \times 1}$ is a column vector that contains data variables corresponding to actual measurement data, denoted as $\mathbf{d}_{obs} \in \mathbb{R}^{N_{obs} \times 1}$, and to other quantities of interest (QoI). In this paper, the QoI are specified to be the CO₂ saturation in the top layer of a three-dimensional aquifer model. In our system, cap rock is not explicitly represented, but it would lie directly above the top layer of the model (see Section 3.1 for the detailed aquifer description). Thus, by quantifying the amount and distribution of CO₂ in the top layer of the model, we can assess the risk of leakage (which could occur through undetected fractures in the cap rock, for example). Related quantities based on the top-layer mass or mobility of CO₂ could also be used. We note that quantities of this type were used as measures of risk to be minimized in the optimizations performed in Cameron and Durlofsky (2012, 2014).

We write \mathbf{d}_f as $\mathbf{d}_f = [\mathbf{d}_h^T, \mathbf{d}_p^T]^T$, where $\mathbf{d}_h \in \mathbb{R}^{N_h \times 1}$ contains elements corresponding to \mathbf{d}_{obs} , and $\mathbf{d}_p \in \mathbb{R}^{N_p \times 1}$ contains the data variables we wish to predict (the subscripts ‘h’ and ‘p’ refer to the historical and prediction portions of the data vector, respectively). In this study, \mathbf{d}_h corresponds to pressure and CO₂ saturation data measured at monitoring wells over the 20-year CO₂ injection period, while \mathbf{d}_p represents the CO₂ saturation field in the top layer of the storage aquifer at 200 years.

Elements in \mathbf{d}_f , referred to as data variables, are uncertain due to the uncertainties associated with the parameters in \mathbf{m} . The forward model (Eq. (1)) is solved via numerical flow simulation. The ECLIPSE simulator (Schlumberger, 2013) with the CO2STORE option is used in this study since it has many specialized features useful for CO₂ storage problems.

Because \mathbf{d}_h , which is part of \mathbf{d}_f , represents the simulated data corresponding to \mathbf{d}_{obs} , we have $N_h = N_{obs}$, and can write

$$\begin{aligned} \mathbf{d}_{obs} &= \mathbf{d}_h + \boldsymbol{\epsilon} \\ &= H\mathbf{d}_f + \boldsymbol{\epsilon}, \end{aligned} \quad (2)$$

where $\boldsymbol{\epsilon}$ contains the measurement error associated with \mathbf{d}_{obs} , and $H \in \mathbb{R}^{N_h \times N_f}$ is an extraction matrix containing zeros and ones, which extracts the part of \mathbf{d}_f corresponding to \mathbf{d}_h . Measurement errors in $\boldsymbol{\epsilon}$ are assumed to be normally distributed with mean $\mathbf{0}$ and covariance $C_D \in \mathbb{R}^{N_h \times N_h}$. Eq. (2) provides the relationship between \mathbf{d}_f and \mathbf{d}_{obs} . Therefore, the posterior distribution of \mathbf{d}_f given \mathbf{d}_{obs} can be formulated using Bayes' rule if the prior probability density function (PDF) of \mathbf{d}_f is constructed. In the following, we first discuss the case where the prior PDF of \mathbf{d}_f is multivariate Gaussian, for which analytical solutions exist. Then we introduce approximate treatments for the case with a general distribution of \mathbf{d}_f .

When the model response \mathbf{d}_f is multivariate Gaussian, the posterior distribution of \mathbf{d}_f , conditioned to \mathbf{d}_{obs} , is given by (see the derivations by Tarantola (2005) and Oliver et al. (2008) for linear problems in traditional inverse modeling, and discussion in Sun and Durlofsky (2017) within the context of DSI)

$$\begin{aligned} p(\mathbf{d}_f | \mathbf{d}_{obs}) &= k_1 \exp \left(-\frac{1}{2} (H\mathbf{d}_f - \mathbf{d}_{obs})^T C_D^{-1} (H\mathbf{d}_f - \mathbf{d}_{obs}) \right. \\ &\quad \left. - \frac{1}{2} (\mathbf{d}_f - \boldsymbol{\mu}_{d_f})^T C_{d_f d_f}^{-1} (\mathbf{d}_f - \boldsymbol{\mu}_{d_f}) \right), \end{aligned} \quad (3)$$

where k_1 is a normalization constant, and $\boldsymbol{\mu}_{d_f} \in \mathbb{R}^{N_f \times 1}$ and $C_{d_f d_f} \in \mathbb{R}^{N_f \times N_f}$ are the prior mean and covariance matrix for \mathbf{d}_f , respectively. The prior mean and covariance matrix can be estimated numerically from the simulation results (Eq. (1)) for an ensemble of prior realizations of \mathbf{m} . On the right-hand side of Eq. (3), the function in the exponent is quadratic with respect to the variables in \mathbf{d}_f . Therefore, $p(\mathbf{d}_f | \mathbf{d}_{obs})$ is again multivariate Gaussian, with analytical expressions for the corresponding posterior mean and covariance matrix as given in Sun and Durlofsky (2017).

As mentioned earlier, \mathbf{d}_p , which is part of \mathbf{d}_f , contains the QoI we seek to predict. The corresponding posterior mean, $\boldsymbol{\mu}_{d_p | d_{obs}}$, and covariance, $C_{d_p | d_{obs}}$, are computed as

$$\boldsymbol{\mu}_{d_p | d_{obs}} = \boldsymbol{\mu}_{d_p} + C_{d_p d_h} (C_{d_h d_h} + C_D)^{-1} (\mathbf{d}_{obs} - \boldsymbol{\mu}_{d_h}), \quad (4)$$

$$C_{d_p | d_{obs}} = C_{d_p d_p} - C_{d_p d_h} (C_{d_h d_h} + C_D)^{-1} C_{d_h d_p}, \quad (5)$$

where $C_{d_p d_h} \in \mathbb{R}^{N_p \times N_h}$ denotes the cross-covariance between \mathbf{d}_p and \mathbf{d}_h , $C_{d_h d_p} = C_{d_p d_h}^T$, and $C_{d_p d_p} \in \mathbb{R}^{N_p \times N_p}$ and $C_{d_h d_h} \in \mathbb{R}^{N_h \times N_h}$ denote the covariance of \mathbf{d}_p and \mathbf{d}_h . These covariance matrices can be easily computed once the simulation results from a set of prior models are obtained. Interestingly, the right-hand side of Eq. (5) is independent of the actual measurements \mathbf{d}_{obs} , which means that, under the multivariate Gaussian condition, the posterior uncertainty of \mathbf{d}_p is always the same regardless of the values of the actual measurements (though it does depend on the amount of observed data since $N_h = N_{obs}$). This is also a well-known property of ordinary kriging methods. He et al. (2018) showed that Eq. (5) can be a useful metric to evaluate the effectiveness of possible future ‘measurements’ from a pilot program. In Section 2.3, we will use Eq. (5), with some rearrangement, to determine the optimal placement of monitoring wells.

Eq. (3) is limited to systems where \mathbf{d}_f is multivariate Gaussian, or nearly so. In our problem \mathbf{d}_f is in general non-Gaussian, so the problem is

more complicated. In such cases, it is useful to introduce an intermediate low-dimensional Gaussian variable ξ . The mapping between ξ and \mathbf{d}_f , which is referred to as a reparameterization, is expressed as $\mathbf{d}_f = \mathbf{f}(\xi)$. Because the flow response is in general non-Gaussian, the relationship between \mathbf{d}_f and ξ is nonlinear. We emphasize that the flow response is non-Gaussian, regardless of the structure of the permeability field (this is the case even for homogeneous permeability descriptions), as a result of the high degree of nonlinearity in the equations governing multiphase flow. This issue was discussed in detail in Sun and Durlofsky (2017) and Sun et al. (2017b), and will be illustrated again in Section 3.4.

With the reparameterization $\mathbf{d}_f = \mathbf{f}(\xi)$, we can write the posterior distribution of \mathbf{d}_f in terms of ξ as

$$p(\xi | \mathbf{d}_{obs}) = k_2 \exp \left(-\frac{1}{2} (H\mathbf{f}(\xi) - \mathbf{d}_{obs})^T C_D^{-1} (H\mathbf{f}(\xi) - \mathbf{d}_{obs}) - \frac{1}{2} \xi^T \xi \right), \quad (6)$$

where k_2 is a normalization constant. The detailed reparameterization, which involves principal component analysis (PCA) and histogram transform, is described in the next section. The PCA representation enables the prior PDF of the data variables to be expressed in terms of $\xi^T \xi$, as discussed in Sun and Durlofsky (2017). Importantly, with the representation in Eq. (6), an analytical expression of the posterior statistics of ξ (and thus \mathbf{d}_f), as given in Eqs. (4) and (5), does not exist. Thus the posterior distribution $p(\xi | \mathbf{d}_{obs})$ must be sampled using an appropriate sampling procedure.

To sample the distribution in Eq. (6), we use the randomized maximum likelihood (RML) method, originally developed by Kitamidis (1986) and Oliver (1996) for traditional model inversion. In our data-space procedure, the objective function for the RML algorithm is

$$S(\xi) = (H\mathbf{f}(\xi) - \mathbf{d}_{obs}^*)^T C_D^{-1} (H\mathbf{f}(\xi) - \mathbf{d}_{obs}^*) + (\xi - \xi^*)^T (\xi - \xi^*), \quad (7)$$

where \mathbf{d}_{obs}^* denotes perturbed measurements (obtained by adding noise, consistent with C_D , to ‘true’ data), and $\xi^* \in \mathbb{R}^{l \times 1} \sim N(\mathbf{0}, I)$ is a random sample from the multivariate normal distribution with mean $\mathbf{0}$ and identity covariance matrix. A set of posterior samples of ξ can be generated by minimizing $S(\xi)$ multiple times, each time with randomly sampled \mathbf{d}_{obs}^* and ξ^* . Posterior samples of \mathbf{d}_f are then obtained by evaluating $\mathbf{f}(\xi)$ on the generated samples of ξ . These samples of \mathbf{d}_f should closely match the observations \mathbf{d}_{obs} during the historical period (\mathbf{d}_h), but they will deviate from one another in the prediction period (\mathbf{d}_p), which provides an assessment of prediction uncertainty. We now introduce the low-dimensional representation of data variables, $\mathbf{d}_f = \mathbf{f}(\xi)$.

2.2. Reparameterization of data variables

As noted earlier, the reparameterization of data variables is accomplished by applying PCA in conjunction with a histogram transformation. We first generate the PCA representation of the correlated variables in \mathbf{d}_f . This is expressed as

$$\mathbf{d}_f \approx \mathbf{d}_{f,PCA} = \Phi \xi + \boldsymbol{\mu}_{d_f}, \quad (8)$$

where $\xi \in \mathbb{R}^{l \times 1}$ contains the PCA parameters, and $\Phi \in \mathbb{R}^{N_f \times l}$ is the basis matrix, whose columns correspond to the first l left singular vectors, multiplied by associated singular values, of the covariance matrix $XX^T/(N_r - 1)$. Here N_r denotes the number of prior realizations of \mathbf{m} that are simulated, and the matrix $X = [\mathbf{d}_f]_1 \ [\mathbf{d}_f]_2 \ \dots \ [\mathbf{d}_f]_{N_r}$, $X \in \mathbb{R}^{N_f \times N_r}$, assembles all simulated model responses. See Sun and Durlofsky (2017) for more details on the construction of Φ , and on the use of an energy criterion to determine l (which allows us to take $l \ll N_f$).

The PCA parameterization (Eq. (8)) enables us to represent the data variables using ξ , which contains parameters that are uncorrelated (this, along with the fact that $l \ll N_f$, greatly simplifies the minimization of $S(\xi)$ in Eq. (7)). However, ξ is still linearly related to \mathbf{d}_f . This linear relationship indicates that Eq. (8) should not be used to generate new realizations of \mathbf{d}_f when \mathbf{d}_f is non-Gaussian. This issue was discussed in detail, in the context of oil-field applications, in Sun and Durlofsky (2017) and Sun et al. (2017b).

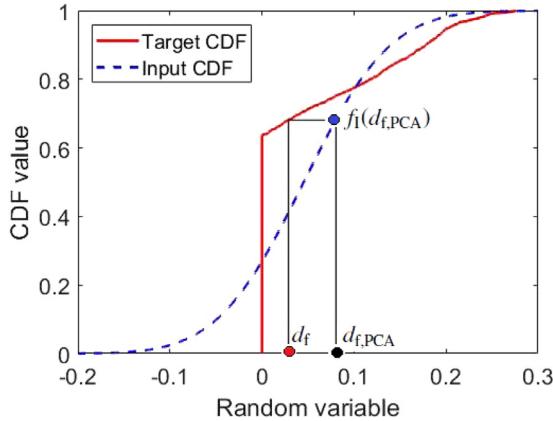


Fig. 1. Cumulative distribution function (CDF) transform used in DS. Note that the input and target CDF curves will be different for different random variables. The CDF curves in this figure correspond to the CO₂ saturation data variable, at 10 years, shown in Fig. 10b. The black point denotes the value from a PCA realization, the blue point is the corresponding CDF value from the initial CDF curve, and the red point represents the value after histogram transformation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Consistent with Sun et al. (2017b), in this work a histogram transformation step is applied to provide an improved approximation of \mathbf{d}_f , with respect to the histogram or marginal distribution. This procedure is depicted in Fig. 1. This transform, performed independently on each data variable, is written as $d_f^j = f_T^{-1} f_I(d_{f,PCA}^j)$, $j = 1, \dots, N_f$, where f_I is the input (or initial) CDF (dashed blue curve in Fig. 1) and f_T is the target CDF (red curve in Fig. 1). The input CDF for d_f^j is Gaussian, since it derives from $\mathbf{d}_{f,PCA}$, with mean μ_{df}^j and variance $(\Phi\Phi^T)^{jj}$ (the j th diagonal element), and the target CDF is constructed from the histogram of the simulated data for d_f^j . It is important to note that histogram transformation directly impacts only the marginal distribution of components in \mathbf{d}_f . We cannot claim that the cross-correlations in \mathbf{d}_f , as defined in prior simulation results, are strictly captured following the histogram transformation procedure. Nonetheless, as we will see in Section 3, this approach is effective for the QoI considered in this work.

Mathematically, the reparameterization step used in this study can be written as

$$\mathbf{d}_f = \mathbf{f}(\xi) = \mathbf{h}_T(\Phi\xi + \boldsymbol{\mu}_{df}), \quad (9)$$

where \mathbf{h}_T denotes the histogram transformation operation. Combining Eqs. (7) and (9), we can now generate multiple posterior predictions through use of RML. The overall DS procedure applied in this study is presented in Algorithm 1 (in the Appendix).

2.3. Optimization of monitoring well placement

Our intent in performing data assimilation is to reduce the prediction uncertainty for QoI. The actual reduction in uncertainty, however, depends on the predictive ability of the measurement data. Since measurement data directly relate to the location of monitoring wells, optimal placement of monitoring wells is clearly of interest. In carbon storage applications, Magnant (2011) used the Shannon entropy, a measure from information theory, to optimize the placement of downhole gauges for cross-well seismic analysis. Cameron and Durlofsky (2014) described a simplified Bayesian experimental design procedure to find improved monitoring well locations. Jeong et al. (2018) presented a binary integer programming technique for the cost-optimal design of a pressure monitoring network for leakage detection under geological uncertainty. Chen et al. (2018) described a filter-based data assimilation procedure for designing monitoring systems for geologic CO₂ sequestration. They used a machine learning algorithm to reduce the computational cost,

though their method is limited to cases where only a few uncertain parameters are involved. In more general cases, the training of the machine learning model may require very large numbers of simulations. In this section, we present our data-space approach for determining the optimal placement of monitoring wells, such that the resulting data are maximally effective in reducing the posterior uncertainty in a key QoI.

It is important to emphasize that the locations of monitoring wells are determined before any data are actually measured. The values of the measured data \mathbf{d}_{obs} will depend on the ‘true’ (but unknown) reservoir model \mathbf{m} , the monitoring well locations \mathbf{y} , and measurement noise ϵ . In this work, $\mathbf{y} \in \mathbb{Z}^{N_y \times 1}$ contains the i and j locations (integer values designating the x and y locations on a grid) for all monitoring wells. Thus, $N_y = 2N_w$, where N_w is the number of monitoring wells. Here the monitoring wells are taken to be vertical wells that extend through the top three layers of the formation and provide data in each layer. Note that \mathbf{d}_{obs} , as a function of \mathbf{m} , \mathbf{y} , and ϵ , now contains random variables instead of fixed historical observations as in Section 2.1.

We let J denote the QoI considered in the optimization of monitoring well locations. The QoI used in this optimization must be a scalar quantity; otherwise, the optimization formulation described below must be modified. The quantity used here should correspond closely to the main quantities of interest, which in our case are the amount and distribution of CO₂ in the top layer. We thus take J to be the average CO₂ saturation in the top layer of the model, which is given by

$$J = \frac{1}{N_{grid}} \sum_{i \in \text{Top Layer}} s_i, \quad (10)$$

where s_i is the CO₂ saturation in grid block i , N_{grid} is the number of grid blocks in the top layer, and the sum is over all blocks in the top layer. In general, smaller values of J correspond to less CO₂ in the top layer, and thus less leakage risk, and higher values to more leakage risk, so the distribution of J is clearly of interest.

Our goal is to find monitoring well locations \mathbf{y} such that the expected variance of J conditioned to measurement data, over all possible realizations of the model and measurement error, is minimized. This optimization problem can be stated as:

$$\mathbf{y}_{opt} = \underset{\mathbf{y}}{\operatorname{argmin}} \left[\int \int \sigma_{J|d_{obs}(\mathbf{m}, \mathbf{y}, \epsilon)}^2 p(\mathbf{m}) p(\epsilon) d\mathbf{m} d\epsilon \right], \quad (11)$$

where $\sigma_{J|d_{obs}(\mathbf{m}, \mathbf{y}, \epsilon)}^2$ denotes the posterior variance of J given ‘measurements’ $\mathbf{d}_{obs}(\mathbf{m}, \mathbf{y}, \epsilon)$. Other statistics, such as the difference between the 90th and 10th percentile ($P_{90} - P_{10}$) in the distribution of J values, could also be used to represent the posterior uncertainty of J . The reason for choosing variance is that the resulting minimization problem (Eq. (11)) can be solved very efficiently, as discussed below.

For now we consider the case where multivariate Gaussian statistics are preserved in the data space (discussed in Section 2.1). Then, by setting $\mathbf{d}_f = J$ in Eq. (5), we have

$$\sigma_{J|d_{obs}(\mathbf{m}, \mathbf{y}, \epsilon)}^2 = \sigma_{JJ}^2 - C_{Jd_h}(C_{d_hd_h} + C_D)^{-1} C_{d_h J}. \quad (12)$$

As noted earlier, the expression on the right-hand side of Eq. (12) is independent of the actual measurement values. More specifically, this expression is independent of the ‘true’ model and the actual measurement errors. Therefore, we must have

$$\int \int \sigma_{J|d_{obs}(\mathbf{m}, \mathbf{y}, \epsilon)}^2 p(\mathbf{m}) p(\epsilon) d\mathbf{m} d\epsilon = \sigma_{JJ}^2 - C_{Jd_h}(C_{d_hd_h} + C_D)^{-1} C_{d_h J}, \quad (13)$$

where the covariance matrices are estimated using the simulated data from an ensemble of prior models ($\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_{N_r}$). Specifically,

$$\sigma_{JJ}^2 = \frac{1}{N_r - 1} \sum_{i=1}^{N_r} (J_i - \mu_J)^2, \quad (14)$$

$$C_{Jd_h} = C_{d_h J}^T = \frac{1}{N_r - 1} \sum_{i=1}^{N_r} (J_i - \mu_J)[(\mathbf{d}_h)_i - \boldsymbol{\mu}_{d_h}], \quad (15)$$

$$C_{d_h d_h} = \frac{1}{N_r - 1} \sum_{i=1}^{N_r} \left[(\mathbf{d}_h)_i - \mu_{d_h} \right] \left[(\mathbf{d}_h)_i - \mu_{d_h} \right]^T, \quad (16)$$

where J_i and $(\mathbf{d}_h)_i$ denote the simulated value corresponding to prior model \mathbf{m}_i , and μ_J and μ_{d_h} denote the mean, over all prior models, of J and \mathbf{d}_h . We emphasize that $(\mathbf{d}_h)_i$ in Eqs. (15) and (16) is a function of \mathbf{y} since it corresponds to data measurements at monitoring wells.

Combining Eqs. (11) and (13), the optimization problem becomes

$$\mathbf{y}_{\text{opt}} = \underset{\mathbf{y}}{\operatorname{argmin}} \left[\sigma_{J,J}^2 - C_{J d_h} (C_{d_h d_h} + C_D)^{-1} C_{d_h J} \right], \quad (17)$$

where $\sigma_{J,J}^2$, $C_{d_h d_h}$, $C_{J d_h}$, and $C_{d_h J}$ are computed using Eqs. (14)–(16) for each different \mathbf{y} . Though these quantities must be evaluated at different monitoring well locations, the simulated data at all required locations for all models are stored, so there is no need to re-run any flow simulations. Because the prior variance $\sigma_{J,J}^2$ does not vary with \mathbf{y} , Eq. (17) can be written as

$$\mathbf{y}_{\text{opt}} = \underset{\mathbf{y}}{\operatorname{argmin}} [-U_R(\mathbf{y})], \quad \text{where } U_R(\mathbf{y}) = \frac{C_{J d_h} (C_{d_h d_h} + C_D)^{-1} C_{d_h J}}{\sigma_{J,J}^2}. \quad (18)$$

The term $C_{J d_h} (C_{d_h d_h} + C_D)^{-1} C_{d_h J}$ essentially represents the correlation between J and \mathbf{d}_h , and $U_R(\mathbf{y})$ can be interpreted as the expected uncertainty reduction in J given measurements collected at monitoring well locations \mathbf{y} . Therefore, Eq. (18) can be interpreted as finding the \mathbf{y}_{opt} that provides measurements enabling the maximum variance reduction in the specified QoI (J).

For arbitrary distributions of J and $\mathbf{d}_{\text{obs}}(\mathbf{m}, \mathbf{y}, \epsilon)$, it can be proven (Harville, 2003; He et al., 2018) that, for any \mathbf{y} ,

$$\int \int \sigma_{J|d_{\text{obs}}(\mathbf{m}, \mathbf{y}, \epsilon)}^2 p(\mathbf{m}) p(\epsilon) d\mathbf{m} d\epsilon \leq \sigma_{J,J}^2 - C_{J d_h} (C_{d_h d_h} + C_D)^{-1} C_{d_h J}, \quad (19)$$

with the equality holding when J and $\mathbf{d}_{\text{obs}}(\mathbf{m}, \mathbf{y}, \epsilon)$ are jointly multivariate Gaussian for any \mathbf{y} . Thus, the right-hand side of Eq. (19) always provides an upper bound estimate of the expected posterior variance (left-hand side of Eq. (19)). This observation indicates that monitoring wells placed at the optimized locations are able to provide prediction variance (uncertainty) less than, or at least equal to, the associated minimum value in the right hand side of Eq. (17). We reiterate that the monitoring well location optimization is based on all prior samples of \mathbf{d}_h (which in general display non-Gaussian statistics), and not on the permeability field itself. Because this optimization is performed before any data are collected, it involves flow-based quantities computed over all prior-model simulation results.

Note that the integration in Eq. (11) can also be estimated using Monte Carlo methods (Robert, 2004), with the variance $\sigma_{J|d_{\text{obs}}(\mathbf{m}, \mathbf{y}, \epsilon)}^2$ estimated using the posterior samples generated by performing the DSI procedure described in Section 2.1. Though this approach is able to incorporate non-multivariate Gaussian statistics, it is more time consuming and complex to apply than using Eq. (17). Therefore, in this paper, Eq. (17) is used to optimize monitoring well placement. As will be shown in Section 3, the optimal well placement found through Eq. (17) is able to achieve significant uncertainty reduction in the QoI.

The workflow to optimize the monitoring well locations using the data-space approach is summarized in Algorithm 2 (in the Appendix). For the optimization, we apply a genetic algorithm (called using the command `ga`) in Matlab. This appears to be adequate for current purposes, though it might be useful to test other optimizers.

3. Results for a three-dimensional heterogeneous aquifer model

In this section, we present detailed DSI results for CO₂ storage in a three-dimensional heterogeneous aquifer. The basic model setup is modified from that used in Cameron and Durlofsky (2012) and Jin and Durlofsky (2018). We consider uncertainties in the porosity and permeability fields and in the direction of a regional pressure gradient. Multiple model scenarios, represented by different variograms, are consid-

ered when generating prior geological models. Note that model structural uncertainty (e.g., uncertainty in the conceptual model), which is not included in this study, can be readily incorporated by considering an ensemble of prior realizations drawn from different geological concepts or training images. This was in fact done in Sun et al. (2017b) for a discrete-fracture-matrix model in which DSI was applied for production forecasting.

Multi-Gaussian log-permeability fields, as are considered here, have been widely used in previous studies related to forward modeling and data assimilation in CO₂ storage and related problems. For example, Farajzadeh et al. (2011) described flow regimes in geologic storage operations using two-dimensional multi-Gaussian models characterized by exponential variograms with different correlation lengths and Dykstra-Parsons coefficients. Similar models were used in the upscaling studies of Rabinovich et al. (2015). Sun et al. (2017a) considered the identification of leaks, with source-term attributes treated as uncertain. The true fields used in that work were in all cases multi-Gaussian. Cameron et al. (2016) also used multi-Gaussian permeability models for a leak-detection study in the context of CO₂ storage. Yoon et al. (2017) studied data assimilation for a problem involving similar physics to that in CO₂ storage. They considered the injection of freshwater into a saline aquifer for purposes of freshwater storage. Two-dimensional multi-Gaussian models were considered, and an ensemble Kalman filtering procedure was used for data assimilation. Finally, Chen et al. (2018) used multi-Gaussian models, with a spherical variogram, in their monitoring-well optimization study. We note that, in the great majority of previous investigations involving data assimilation with multi-Gaussian models, prior models were all characterized by a single variogram. Our treatment here is much more general, as our prior includes models generated using a wide range of variogram (and other) parameters, as described below.

In this work pressure and CO₂ saturation data within the storage aquifer are considered to be available. In practice, pressure data can be obtained through embedded pressure transducers, and CO₂ saturation can be measured in situ using time-lapse sonic logs (Caspari et al., 2011), as well as by seismic and other monitoring technologies, as described in Zhang et al. (2018). In this study, synthetic pressure and CO₂ saturation data are collected at monitoring wells during the 20-year injection period, with monitoring well locations optimized using the data-space approach described in Section 2.3. Other data types could be readily incorporated if available. Given the monitoring data, the conditional predictions for the CO₂ plume at the top layer, after 180 years of post-injection equilibration (i.e., at the end of the 200-year time frame), are generated using the DSI algorithm presented in Section 2.1. We emphasize that no additional data are collected after the 20-year injection period. Note that this data assimilation could be performed more frequently, e.g., every year during the injection period, and this would be very inexpensive with DSI since no additional simulation runs are required.

3.1. Aquifer model

The storage aquifer covers an area of 8.5 km × 8.5 km, and is of a vertical thickness of 151 m. The aquifer initially contains only brine at a pressure of 15.5 MPa and a temperature of 55 °C. A total of 8 Mt/year of CO₂ is injected into the storage aquifer for 20 years. This CO₂ volume corresponds to approximately 2.5% of the aquifer pore volume (the aquifer is of an average porosity of about 0.2). The storage aquifer is modeled on a uniform grid of 35 × 35 × 11 blocks, each of size 243 m × 243 m × 13.7 m.

The basic aquifer model is presented in Fig. 2. One realization of the heterogeneous porosity field for the storage aquifer is shown in Fig. 2a. Around the storage aquifer a larger-scale regional aquifer is introduced, which essentially acts to dissipate pressure (Cameron et al., 2016). Fig. 2b shows the simulation grid for the full system, with the white central region representing the storage aquifer shown in Fig. 2a.

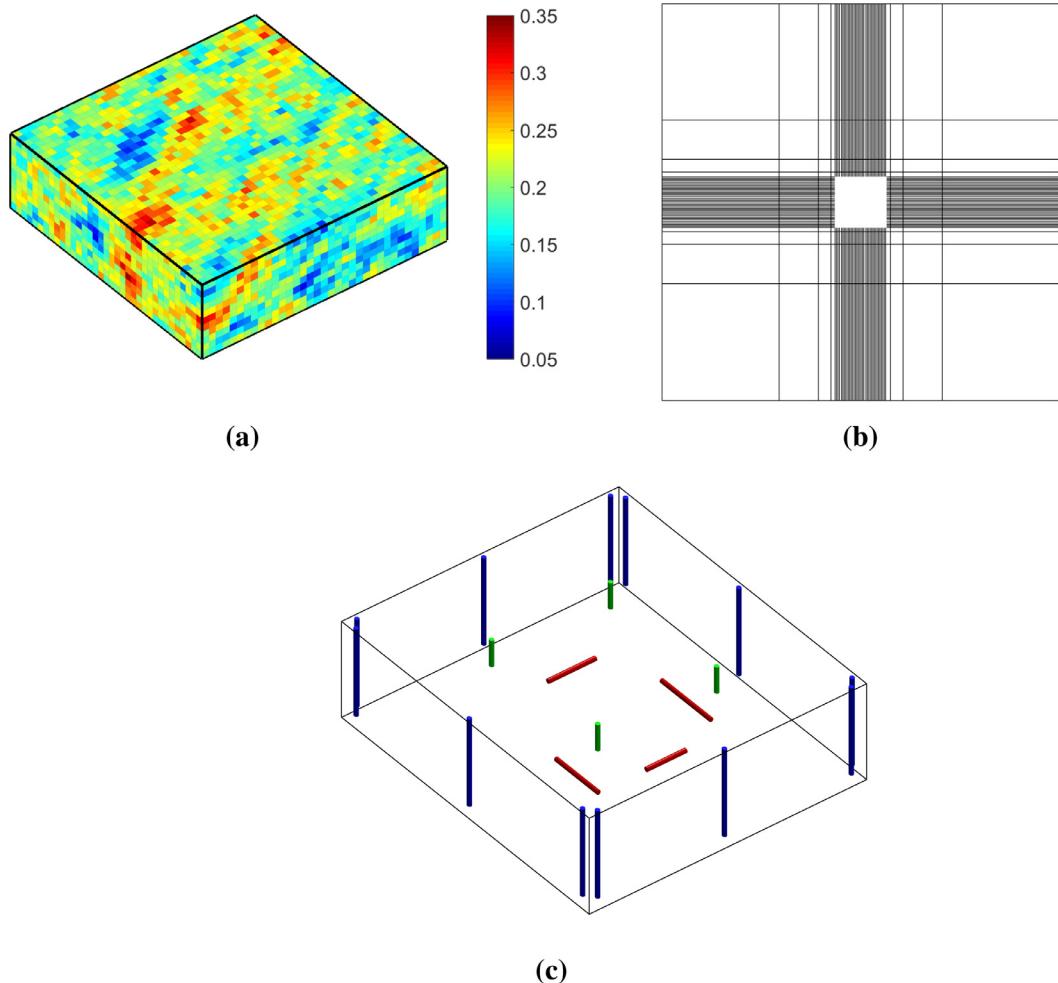


Fig. 2. Setup of aquifer model: (a) one geological realization of the porosity field of the storage aquifer, (b) simulation grid of the full model, (c) locations of four horizontal CO_2 injection wells (shown in red), possible locations of monitoring wells (shown in green), and pseudo-wells (shown in blue) used to create the regional pressure gradient. Note that all wells are placed within the storage aquifer. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

To save computational effort, the grid block size increases by a factor of three as we move out from the storage aquifer. The full system is modeled on a $43 \times 43 \times 11$ grid, and is of overall dimension $67 \text{ km} \times 67 \text{ km} \times 151 \text{ m}$. In the region outside the storage aquifer, porosity and permeability values are specified to be constant at 0.2 and 29 md, which correspond to the mean values in the storage aquifer. No-flow boundary conditions are applied at the boundaries of the full model.

Four horizontal injectors are located in layer 9 (third from the bottom) of the storage aquifer. Each injector is controlled with a constant CO_2 injection rate of 2 Mt/year. The four injectors, shown as red lines in Fig. 2c, are arranged in a square pattern. In addition, vertical monitoring wells (depicted as green lines in Fig. 2c) provide CO_2 saturation and pressure data at all of the grid blocks they contact (recall that monitoring wells penetrate the top three layers of the model). Because these wells penetrate the cap rock, in practice it will be important that proper well completion, isolation, and cementing procedures are applied.

The simulation model includes two fluid phases, referred to as water and gas, and three components – water, supercritical CO_2 and salt. Brine salinity is specified to be 10,000 mg/kg. The gas-water relative permeability curves, which include hysteresis (important for residual trapping), are taken from Cameron and Durlofsky (2012) and are shown in Fig. 3a. Capillary pressure curves are generated using the Brooks-Corey model (Saadatpoor et al., 2010), with specific model coefficients

of $p_e = 0.36$ (entry pressure) and $\lambda = 0.67$ (pore size distribution index). Capillary pressure curves, which vary from block to block and depend on porosity and permeability, are modeled using the Leverett J -function (see Eqs. (2.3) and (2.4) in Saadatpoor et al. (2010)). Fig. 3b shows the capillary pressure at reference porosity (0.2) and permeability (29 md). Given the large size of the grid blocks, the absolute permeability, and the relative permeability and capillary pressure curves, should be regarded as appropriately upscaled functions (this upscaling can be accomplished using, e.g., the methods described in Rabinovich et al. (2015)). Chemical reactions and mineralization are not included in our model, as these effects are not expected to be significant over the simulation time frame considered in this study. All flow simulations are performed using the ECLIPSE simulator (Schlumberger, 2013) with the CO2STORE option.

In this study, the porosity and log-permeability fields in the storage aquifer are represented using two-point geostatistical models. Sequential Gaussian simulation, available within the SGeMS toolbox (Remy et al., 2009), is used to generate multiple geological realizations of the porosity field. These realizations are not conditioned to any well hard data. The histogram of the porosity distribution is specified to be normal, with mean $\bar{\phi} = 0.2$ and standard deviation $\sigma_\phi = 0.04$. The correlation structure of the porosity field is characterized using an ellipsoid-shaped exponential variogram model, with the three correlation lengths of the principal axes, specified in terms of the number of grid blocks, set to be $l_1 = 40$, $l_2 = 5$, and $l_3 = 5$. We consider different

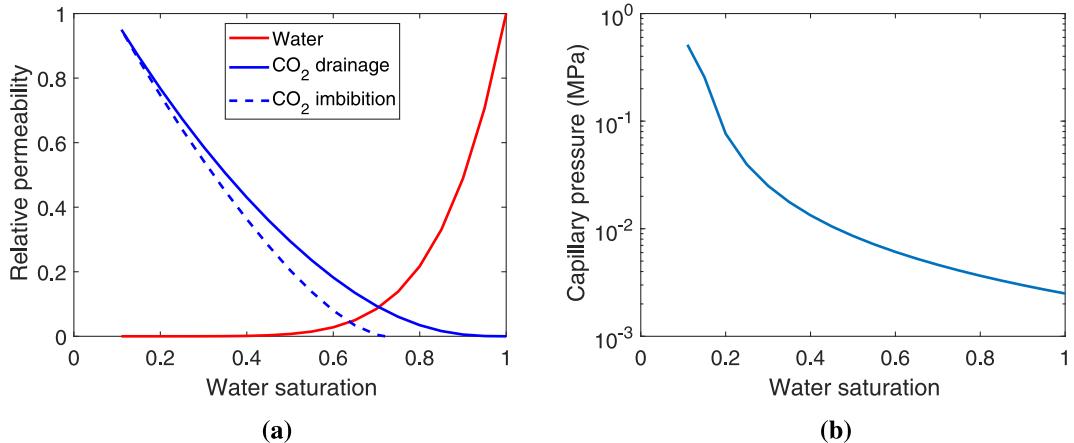


Fig. 3. Two-phase flow properties: (a) relative permeability curves for water and CO₂ (Cameron and Durlofsky, 2012); (b) reference capillary pressure at porosity and permeability values of 0.2 and 29 md. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1

Uncertain parameters, ranges, and values for three test cases. All parameters are assumed to be uniformly distributed a priori.

Parameter	Lower	Upper	Test Case 1	Test Case 2	Test Case 3	Test Case 4
Azimuth (°)	-30	30	-29	-17	9	45
Dip (°)	-30	30	22	-18	-20	45
k_z/k_x	0.05	0.2	0.08	0.1	0.19	0.3
q_w (m ³ /day)	0	55	0.1	12	36	75

variogram models (and thus different geological scenarios) by varying the orientations of the principal axes of the ellipsoid. The ranges of these parameters are shown in Table 1. Azimuth and dip refer to the two ellipsoid orientation angles (see Remy et al., 2009 for more details). The other parameters appearing in the table are described below.

The permeability value in the x -direction (k_x) is related to porosity via

$$\ln k_x = a + b \left(\frac{\phi - \bar{\phi}}{\sigma_\phi} \right) + e, \quad (20)$$

where $a = 3.5$, $b = 1$, and e represents random noise sampled from a normal distribution with mean $\bar{e} = 0$ and standard deviation $\sigma_e = 0.4$. Therefore, k_x is log-normally distributed with mean $\ln k_x = 3.5$ and standard deviation $\sigma_{\ln k_x} \approx 1.08$. We specify $k_y = k_x$ and $k_z = ck_x$, where c is randomly sampled from a uniform distribution within [0.05, 0.2] for each prior permeability realization (i.e., k_z/k_x is constant for a given model). Fig. 4 shows two random realizations of the porosity field, and the corresponding $\ln k_x$ fields. Clear differences in the correlation structure of the two porosity (or permeability) fields, due to the use of different variogram models, are evident.

In addition to uncertainties in the distribution of porosity and permeability, we also include an uncertain large-scale regional pressure gradient, which acts to move the CO₂ plume horizontally. In the west-to-east direction, the pressure gradient component is specified to be constant at 280 Pa/m, while in the south-to-north direction, the pressure gradient component is uncertain and ranges from 0 Pa/m to 280 Pa/m. The effect of this regional pressure gradient is captured in the simulations by specifying water rates for an array of 12 wells, penetrating all layers, distributed around the storage aquifer (shown as blue lines in Fig. 2c). Wells on the west and south sides are specified to be injectors, and producers are on the opposite sides. The injectors and producers on the west and east sides are assigned constant water rates of 55 m³/day, which results in a pressure gradient of ~280 Pa/m. The water rate q_w , however, is sampled uniformly from 0 m³/day to 55 m³/day for wells on the north and south sides (each well in a particular realization is assigned the same q_w). Therefore, the uncertainty related to the regional

pressure gradient is represented using the single random variable q_w (ranges for this parameter are given in Table 1).

In the results presented below, four different test cases will be considered in detail for the evaluation of our data-space approaches. Parameters for all test cases are provided in Table 1.

3.2. Simulation results from prior models

In this paper, a total of $N_r = 1000$ prior models are generated, over the parameter ranges given in Table 1, as described in Section 3.1. The duration of the injection period is 20 years (during which data are collected), and this is followed by 180 years of equilibration. We are interested in assessing the CO₂ saturation distribution at the top layer at the end of the equilibration period. Fig. 5 shows the CO₂ plume evolution in a cross-section containing an injection well for one randomly selected prior model. During the injection period, CO₂ accumulates in the system while moving toward the top of the formation due to buoyancy. From 140 years to 200 years, we see that the CO₂ distribution changes relatively little, indicating that much of the CO₂ is essentially trapped. This is caused both by residual trapping and trapping due to heterogeneous capillary pressure. The slight changes in the CO₂ distribution in the top layer at a late time (Fig. 5g and h) are due to spreading in the direction orthogonal to the image (in the y -direction).

Fig. 6 displays the CO₂ saturation distribution in the top layer (these are $x - y$ plots, in contrast to the $x - z$ plots in Fig. 5) corresponding to six different prior models. The results presented in this figure are selected based on the total amount of CO₂ in the top layer at 200 years (computed as J in Eq. 10). In Fig. 6, J_p^{prior} indicates the p th percentile of J among all prior simulation runs. There are large variations in the distribution and total amount of CO₂ in the top layer, which is evident by comparing Fig. 6a-f. More specifically, the P_{90} result in Fig. 6e ($J_{90}^{\text{prior}} = 0.056$) shows about twice the CO₂ compared with the P_{10} result in Fig. 6a ($J_{10}^{\text{prior}} = 0.027$). In addition, although there are essentially four CO₂ ‘lobes’ corresponding to the four injectors, the CO₂ plumes form into different shapes and move to different locations. This phenomenon can be caused by differences in the regional pres-

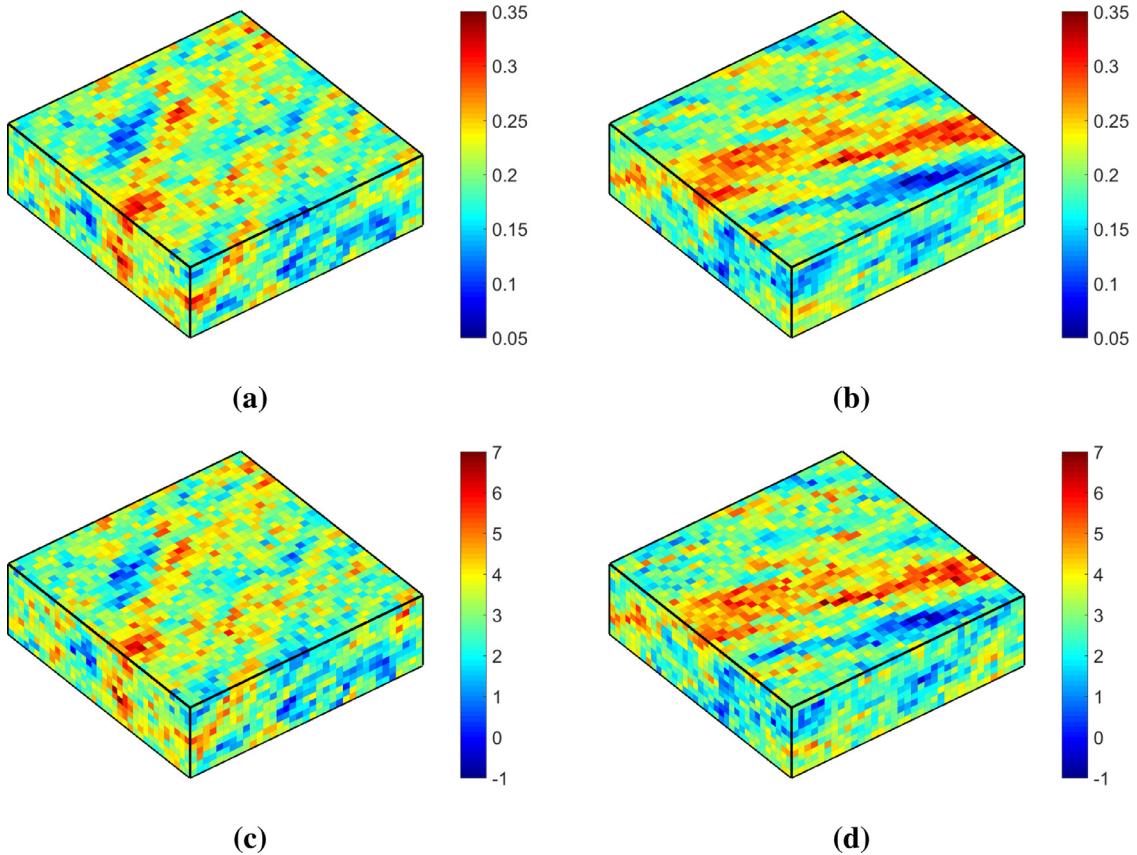


Fig. 4. Two geological realizations of porosity ((a) and (b)) and the associated $\ln k_x$ fields ((c) and (d)).

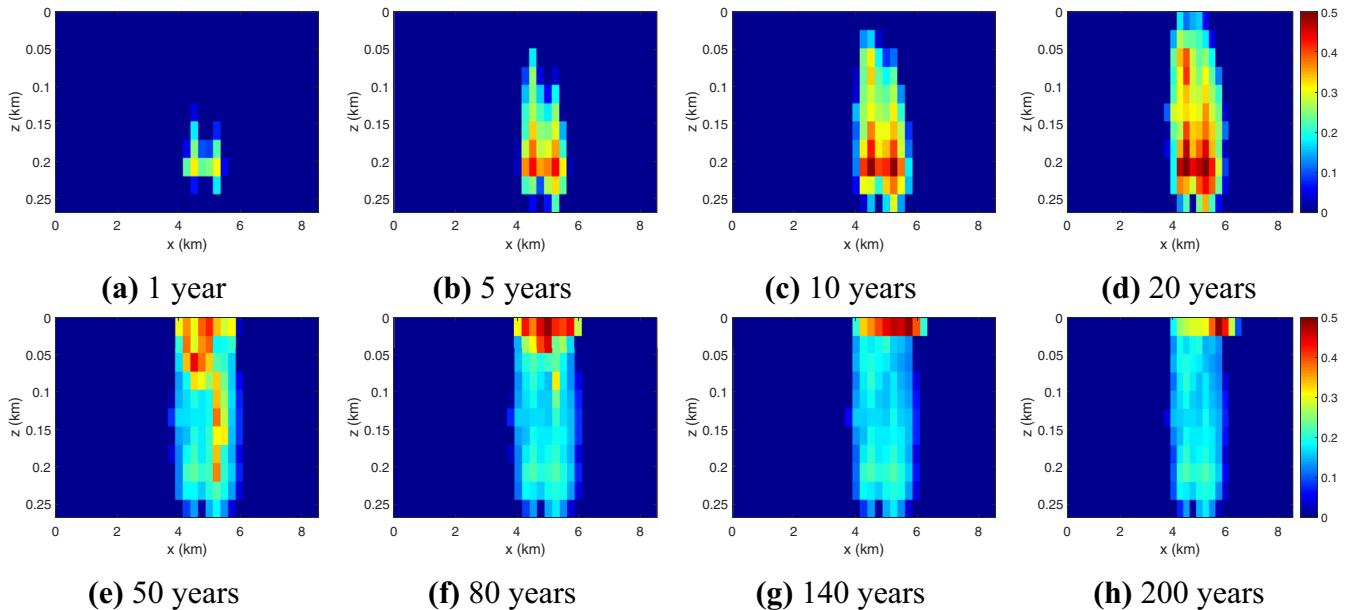


Fig. 5. Cross-sectional view of the CO₂ plume evolution over the full simulation time frame (200 years) for one random prior realization. This $x - z$ cross section contains a horizontal injection well, which extends in the x -direction. The first row shows CO₂ saturation during the injection period, and the second row during the remaining 180 years (equilibration period).

sure gradient and/or by differences in the porosity and permeability fields. As will be demonstrated below, data at the monitoring wells enable us to narrow our estimates for the location and shape of the CO₂ plume.

3.3. Optimization of monitoring well locations

We now present results for the optimization of a specified number of monitoring wells. The number of monitoring wells is denoted by N_w . Recall that this optimization is performed after simulating the N_r prior models, but before any data are measured.

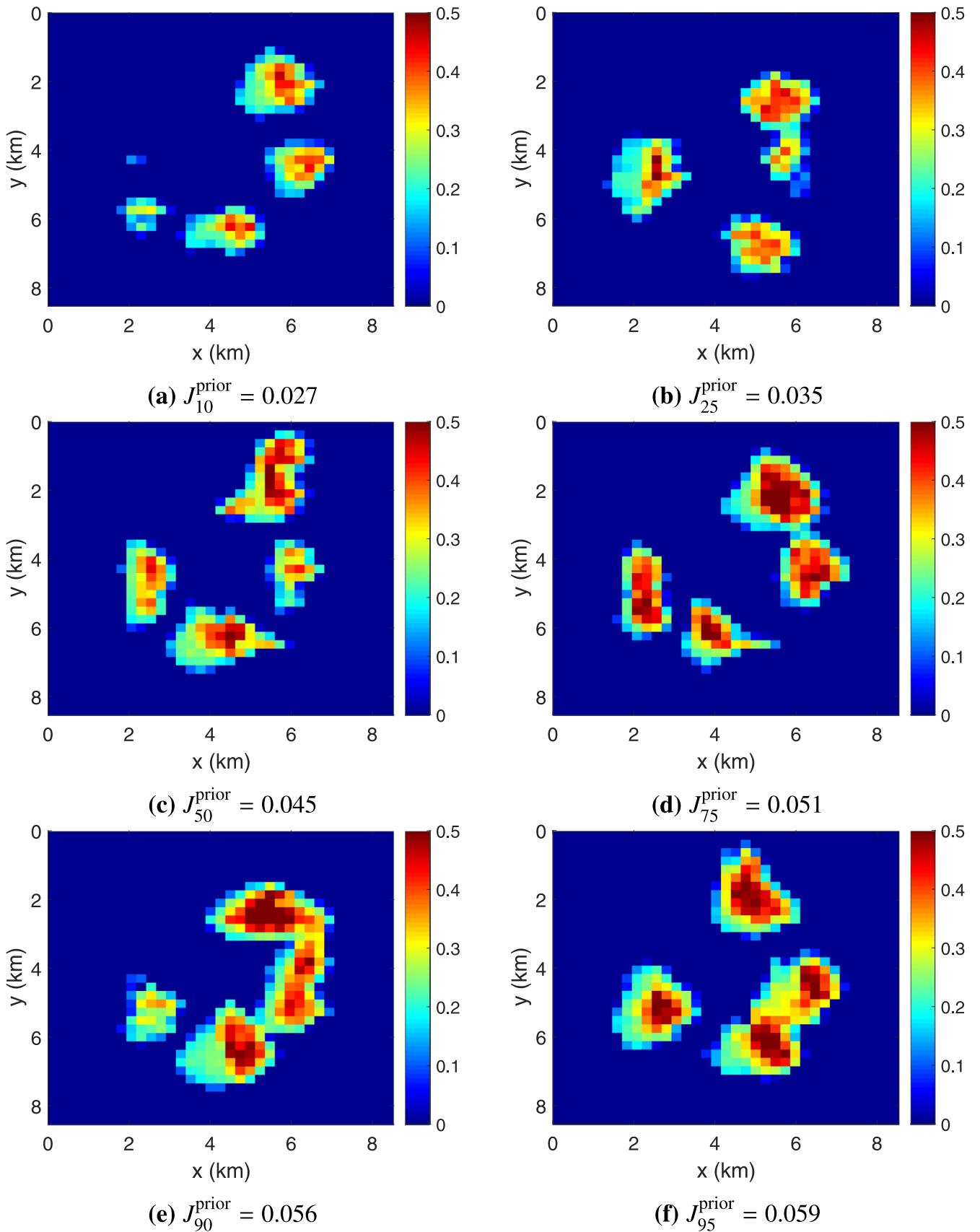


Fig. 6. CO₂ saturation distributions for different prior models at 200 years. Results in (a) to (f) correspond to models that lead to increasing amounts of CO₂ in the top layer (as quantified by J_p^{prior} , which represents the pth percentile of J over all prior realizations).

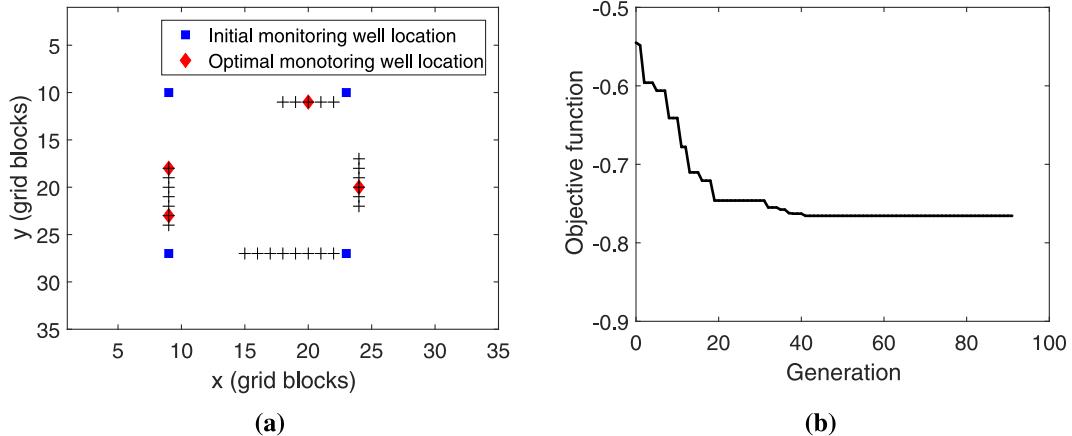


Fig. 7. Location optimization of four monitoring wells: (a) areal view of injection and monitoring well locations within the storage aquifer. The black cross-hatched lines represent the projected locations of the horizontal injections wells, and the blue squares and red diamonds show the initial guess and optimized monitoring well locations. (b) progression of the best *ga* solution. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

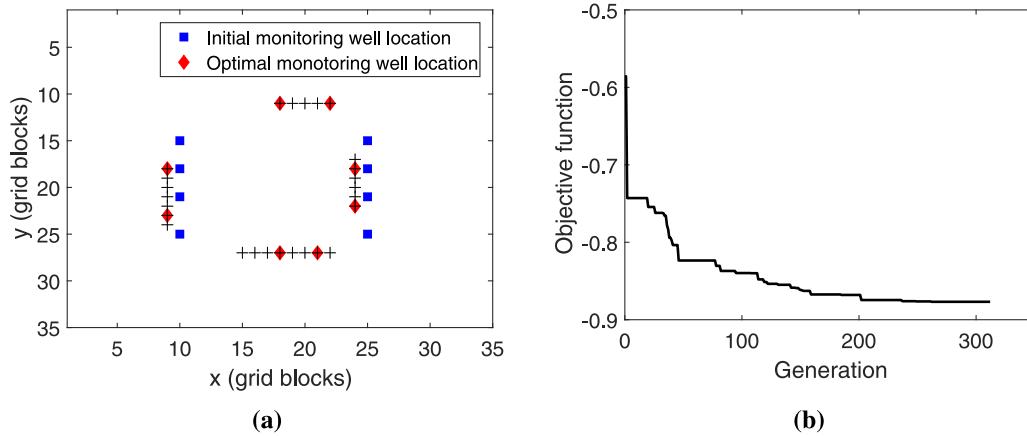


Fig. 8. Location optimization of eight monitoring wells. Lines and symbols have the same meaning as in Fig. 7.

As discussed earlier, the QoI J is specified to be the average CO_2 saturation in the top layer at 200 years (Eq. (10)), and we are interested in maximizing the expected uncertainty reduction for J after assimilating 20 years of measurement data. For each monitoring well, pressure and CO_2 saturation data are measured annually for 20 years in each of the top three layers. The standard deviation of the measurement error for pressure and CO_2 saturation data are specified to be 0.1 MPa and 0.02, respectively. In this section, we will consider two different cases, $N_w = 4$ and $N_w = 8$. Therefore, the dimension of \mathbf{d}_h is $N_h = N_w \times 3 \times 2 \times 20 = 480$ when $N_w = 4$, and $N_h = 960$ when $N_w = 8$. The number of parameters to determine in the optimization is 8 ($N_w = 4$) or 16 ($N_w = 8$). As mentioned in Section 2.3, the Matlab genetic algorithm (*ga*) optimizer (with default options) is used in this study.

The initial monitoring well locations, determined heuristically, are shown as blue squares in Fig. 7a for the case with $N_w = 4$. The computed (expected) uncertainty reduction for this configuration is $U_R(\mathbf{y}_{\text{init}}^4) = 0.34$, where $\mathbf{y}_{\text{init}}^4$ denotes the initial locations for the four monitoring wells, and the function $U_R(\mathbf{y})$ is defined in Eq. (18). This result means that the prediction uncertainty (represented as variance) of J is expected to be reduced by 34% compared with the prior variance $\sigma_{J,J}^2$, after assimilating all possible outcomes of measurement data collected at these well locations. Fig. 7b shows the progression of the best *ga* solution (the objective function is $-U_R(\mathbf{y})$), over the 91 generations of the optimization procedure. The best *ga* solution at generation 0 is -0.54 rather than -0.34 because (at least) one of the solutions in the initial population leads to more uncertainty reduction than the initial guess $\mathbf{y}_{\text{init}}^4$. The

elapsed time for this optimization is about five minutes with our Matlab implementation.

The optimal locations, $\mathbf{y}_{\text{opt}}^4$, are shown as red diamonds in Fig. 7a. This result corresponds to $U_R(\mathbf{y}_{\text{opt}}^4) = 0.77$, which is a clear improvement over the initial guess. It is noteworthy that the optimal solution tends to place the monitoring wells above the injectors (shown as black cross-hatched lines in Fig. 7a), as would generally be expected, though there are two monitoring wells above the left injector and none above the lower injector. We note that the optimized well locations will be different with different initial guesses. Based on limited experimentation, however, it does appear that monitoring wells tend to be placed above injectors.

Fig. 8 displays analogous results for the case of $N_w = 8$. A total of 312 *ga* generations are required in this case, and the total elapsed time is around 30 minutes. The initial and optimal expected uncertainty reduction are $U_R(\mathbf{y}_{\text{init}}^8) = 0.61$ and $U_R(\mathbf{y}_{\text{opt}}^8) = 0.87$. Interestingly, the optimal solution in this case corresponds to two monitoring wells placed above each injector (see Fig. 8a).

3.4. DSI predictions of CO_2 plume: Test Case 1

We now apply the DSI methodology to a series of synthetic test cases to demonstrate the reduction in uncertainty that can be achieved using this procedure. In all cases, a ‘true’ model, which is not included in the set of N_r prior models, is generated. The geological parameters for the three ‘true’ models considered in this section (Test Cases 1, 2 and 3) are given in Table 1. Flow simulation is then performed on the ‘true’ model

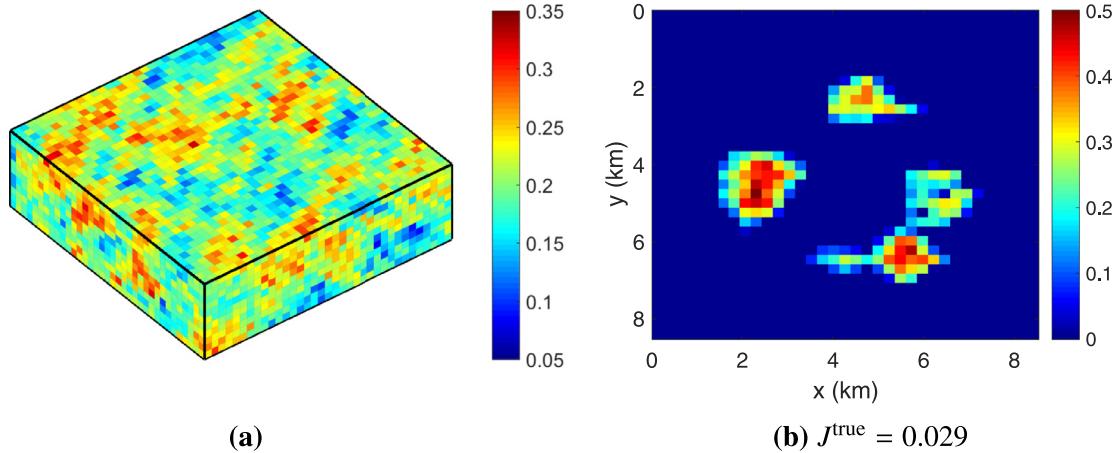


Fig. 9. True model for Test Case 1: (a) porosity field, (b) CO₂ saturation in top layer (200 years).

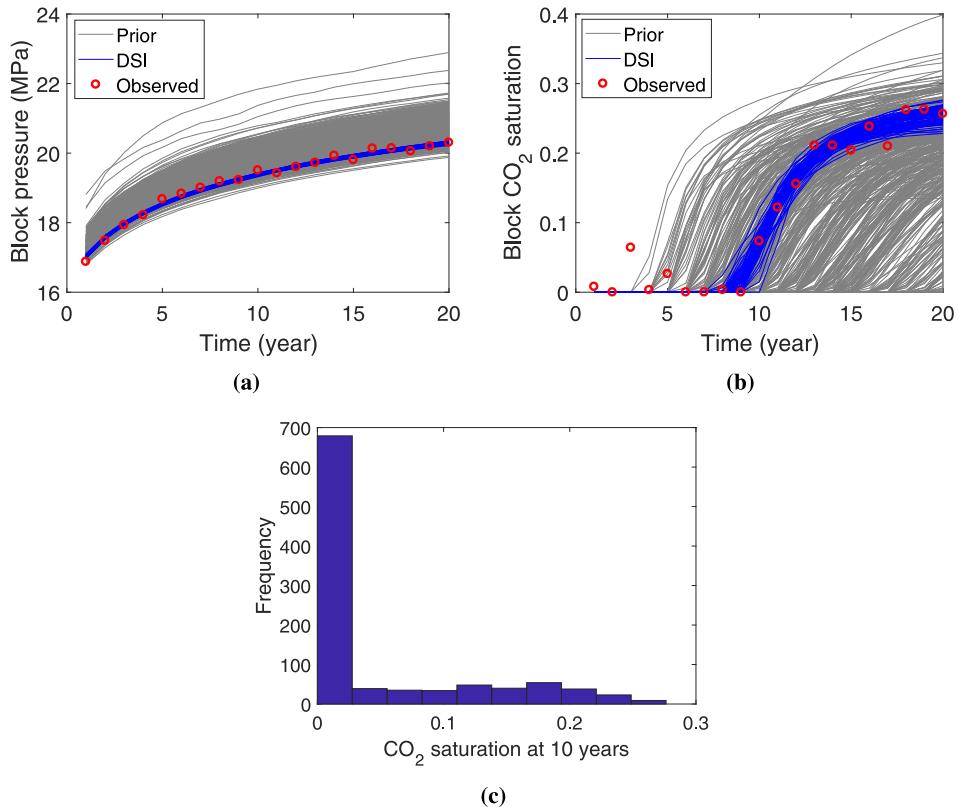


Fig. 10. DSI predictions for historical data at a single grid block containing a monitoring well (this grid block is located in the third layer of the model): (a) pressure, (b) CO₂ saturation, (c) histogram of CO₂ saturation at 10 years (Test Case 1). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to provide the true data. Measurement error, with mean and standard deviation as given in [Section 2.3](#), is added to the true data to provide the observed data \mathbf{d}_{obs} . Note that this measurement error could be used to account for the fact that monitoring wells are much smaller than the grid blocks they penetrate (i.e., because monitoring-well diameter is much less than the horizontal grid block dimension, measured data correspond to the grid block value plus some error). A grid-refinement assessment, such as that presented in [Section 4.3](#), could be used to quantify this error in practice.

The porosity field for Test Case 1 is shown in [Fig. 9a](#), and the top-layer CO₂ at 200 years is displayed in [Fig. 9b](#) ($J^{\text{true}} = 0.029$ for this case). [Fig. 10](#) shows the time-series for the observed data (red circles),

for a single grid block where data are measured annually. Note that the synthetic observed data are forced to be nonnegative in [Fig. 10b](#) when negative measurement error is added to zero CO₂ saturation data. The simulated data from $N_r = 1000$ prior models, for the selected grid block, are shown as the gray curves in [Fig. 10](#). [Fig. 10c](#) presents the histogram of the CO₂ saturation data at 10 years from the prior-model simulations. These results display strong non-Gaussian character, which illustrates the fact (noted in [Section 2.1](#) and discussed in detail in [Sun and Durlofsky \(2017\)](#) and [Sun et al. \(2017b\)](#)) that multiphase flow simulation results are non-Gaussian even when the underlying log-permeability field is multi-Gaussian.

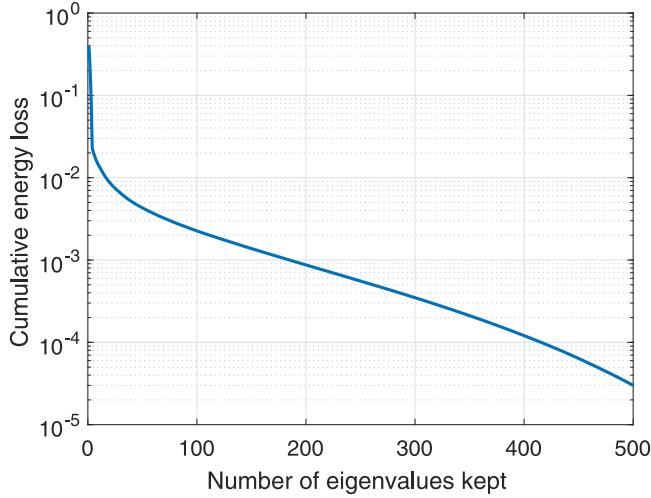


Fig. 11. Cumulative energy loss for principal components (Test Case 1).

We first consider the case with four optimized monitoring wells, which results in a total of $N_{\text{obs}} = 480$ observations (consistent with N_h in the previous section). Our intent is to predict the CO₂ saturation distribution in the top layer at 200 years (recall that data are collected only during the 20-year injection period). Thus, the number of variables to predict in \mathbf{d}_p is equal to the number of grid blocks in the top layer; i.e., $N_p = N_{\text{grid}} = 35 \times 35 = 1225$. The dimension for the full data vector, \mathbf{d}_f , in the DSI procedure is then $N_f = N_h + N_p = 1705$. Because \mathbf{d}_f contains both pressure and saturation data variables, which are of very different scales, we normalize the pressure data variables to be of mean 0 and standard deviation 1 before applying PCA.

Fig. 11 shows the cumulative energy loss (in other words, the variability that is not explained in the subspace) for different numbers of retained principal components, which is the dimension of ξ (denoted as l) in Eq. (8). We determine the number of principal components (l) by specifying the energy loss to be less than 0.005, which results in $l = 42$ in this case. Though the dimension of the data space, N_f , is fairly large, the reduced space with PCA parameterization is reasonably small. This observation indicates that the simulated data have strong cross-correlations.

Applying the DSI procedure (Algorithm 1), with $l = 42$, we generate a total of $N_{\text{post}} = 100$ posterior predictions. The DSI predictions for the historical period, \mathbf{d}_h , are shown in Fig. 10 as the blue curves. It is evident that the DSI predictions are able to essentially match the observed data

(red circles) for both quantities. Though our interest is not to predict the observed data, these results serve as a quality check of the application of the DSI method. If the DSI predictions are not able to match observed data, more prior runs might be required. It is also possible that additional geological scenarios (e.g., different variogram parameters in the current setting) should be considered. For more discussion of this issue, please see our earlier work (Sun and Durlofsky, 2017).

Fig. 12 shows three posterior realizations of the CO₂ saturation distribution obtained with the DSI method. These three realizations are selected based on the average top-layer CO₂ saturation at 200 years (J in Eq. (10)). We compute J for all 100 posterior DSI top-layer saturation fields, and then rank them from lowest to highest. Figs. 12a–c correspond to the results of rank 10, 50, and 90 (i.e., the P_{10} , P_{50} , P_{90} results). The P_{10} and P_{90} results can be viewed as optimistic and pessimistic predictions, respectively, of the top-layer CO₂ saturation at 200 years. It is evident that the amount of CO₂ in the DSI predictions is reasonably close to that from the true model (Fig. 9b), even though the amount of predicted CO₂ in the prior models varies substantially, as seen in Fig. 6. More specifically, J^{true} ($= 0.029$) is encompassed within the range of DSI predictions ($[0.026, 0.038]$ for J_{10} to J_{90}), and this range is significantly smaller than the range of prior predictions ($[0.027, 0.056]$ for J_{10}^{prior} to J_{90}^{prior}). In addition, the overall pattern of the true CO₂ plume (Fig. 9b) is captured better in the DSI predictions (Fig. 12) than in the prior predictions shown in Fig. 6, especially for the plume at the left side of the model.

3.5. DSI predictions for different monitoring scenarios

We next consider three other monitoring scenarios to investigate the effects of the number and placement (optimal versus heuristic) of monitoring wells. Parameters for all monitoring scenarios are shown in Table 2. With monitoring scenario 2, the location of monitoring wells is not optimized, though all other parameters are the same as in monitoring scenario 1 considered in the previous section. Monitoring scenarios 3 and 4 differ from monitoring scenario 1 in terms of the number of monitoring wells and the associated number of observations.

Fig. 13 shows the DSI predictions for CO₂ saturation distribution when monitoring well locations are not optimized. As in Fig. 12, the three saturation fields again correspond to P_{10} , P_{50} , P_{90} results. The predictions show clear improvement, in terms of matching the true plume, relative to the prior predictions (shown in Fig. 6). However, the variability of the average CO₂ among the three realizations in Fig. 13 is larger than in the results corresponding to optimal monitoring well placement (Fig. 12). This observation is also evident in Fig. 14, where we plot the empirical cumulative density function (CDF) for the top-layer average CO₂ saturation over the 100 posterior DSI predictions. The dashed

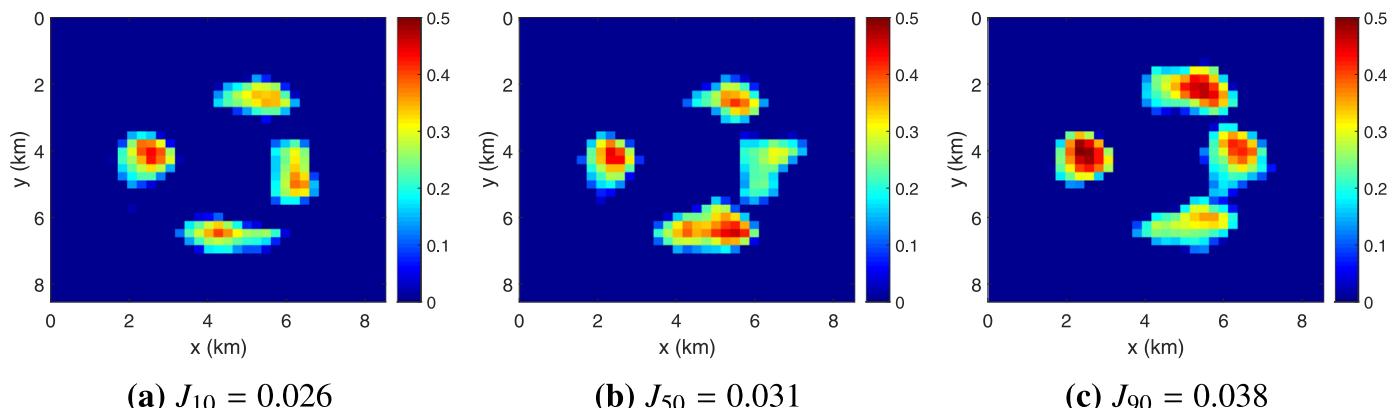


Fig. 12. Test Case 1: (a), (b) and (c) are the DSI predictions of rank 10, 50, and 90 for the CO₂ saturation distribution in the top layer at 200 years. The ranking is determined based on the average top-layer CO₂ saturation among all 100 posterior DSI predictions. For this case, $J^{\text{true}} = 0.029$. The measurement data are from four monitoring wells, with optimal placement.

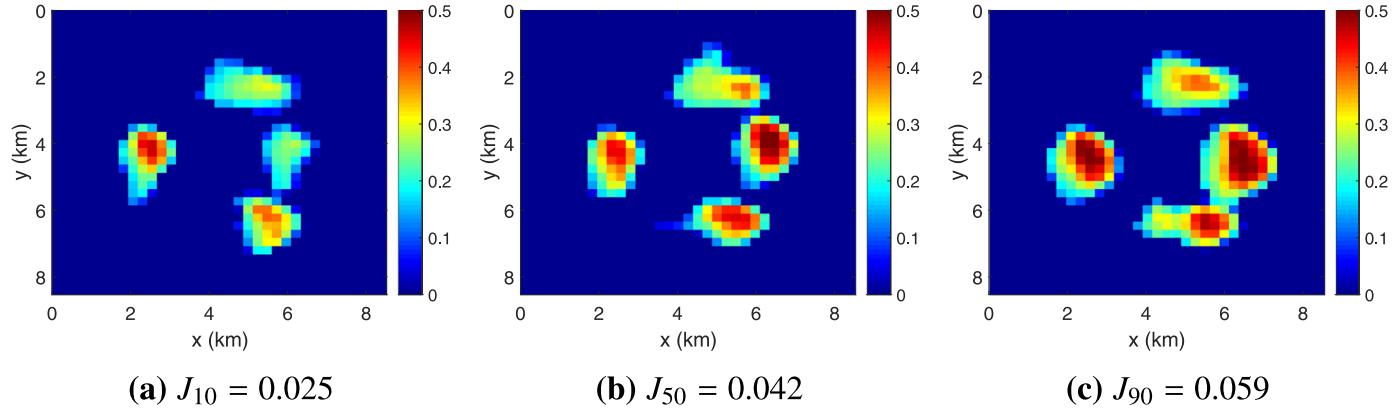


Fig. 13. Test Case 1, with (non-optimized) monitoring scenario 2 defined in Table 2. (a), (b) and (c) have the same meaning as in Fig. 12. For this case, $J^{\text{true}} = 0.029$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2

Parameters for four monitoring scenarios. MW: monitoring well; N_w : number of monitoring wells; N_{obs} : number of observations; N_f : dimension of data space; y : location of monitoring wells.

Monitoring scenario	Label	N_w	N_{obs}	N_f	y
1	MW-4-opt	4	480	1705	optimized
2	MW-4-init	4	480	1705	initial
3	MW-6-opt	6	720	1945	optimized
4	MW-8-opt	8	960	2185	optimized

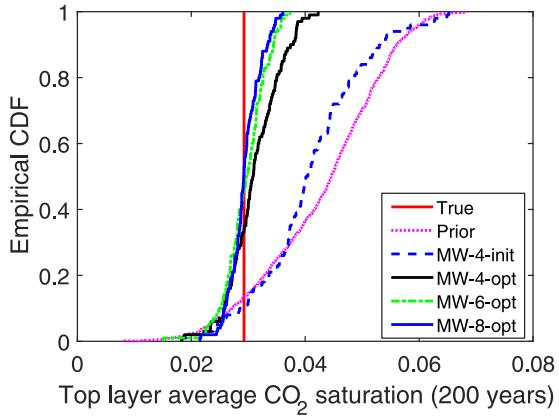


Fig. 14. Empirical CDFs computed from posterior DSI results for top-layer average CO_2 saturation (J). Four different monitoring scenarios, defined in Table 2, are considered. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

blue and the solid black lines represent results, for $N_w = 4$, without and with optimization of monitoring well locations. We see that the uncertainty range is much narrower when using data measured at optimized monitoring well locations. The dashed green and solid blue lines display results for six and eight (optimal) monitoring wells. We observe a consistent decrease in uncertainty as N_w is increased, consistent with expectations.

Fig. 15 displays CO_2 saturation predictions for monitoring scenario 4, which entails the use of eight (optimally placed) monitoring wells. These results match the true plume better than those for monitoring scenario 1, where $N_w = 4$. In particular, predictions of the lower plume in Fig. 15 are clearly more accurate than those in Fig. 12. This observation can be explained by the placement of monitoring wells, shown in Figs. 7a and 8a. In the case with $N_w = 4$, there are no monitoring wells above the lower injector, while in the case with $N_w = 8$, two monitoring

wells are placed above this injector. As a result, there are more observations corresponding to this CO_2 plume. In addition, with more monitoring wells, the connection between the lower plume and the plume on the right is also captured more accurately.

To better quantify the accuracy of DSI predictions for the CO_2 saturation distribution in the top layer, we compute the average saturation error for each posterior prediction, Δ^k , as follows:

$$\Delta^k = \frac{1}{N_{\text{grid}}} \sum_{i \in \text{Top Layer}} |s_{i,\text{pred}}^k - s_{i,\text{true}}| \quad (21)$$

Here Δ^k denotes the average pixel-wise prediction error in posterior prediction k , N_{grid} is the total number of top-layer grid blocks, $s_{i,\text{pred}}^k$ is the predicted CO_2 saturation at 200 years in grid block i from DSI posterior prediction k (e.g., Fig. 15a), and $s_{i,\text{true}}$ is the true CO_2 saturation at 200 years in grid block i (shown in Fig. 9b). Given Δ^k for all $N_{\text{post}} = 100$ DSI predictions, we can construct a CDF of the Δ^k for each monitoring scenario.

Fig. 16 displays the resulting empirical CDFs. The prediction error from the prior models is the largest as no additional measurement information is assimilated. We also see that optimizing monitoring well locations improves prediction accuracy (from the dashed blue line to the black line in Fig. 16), and that the use of more monitoring wells is also beneficial (from the solid black line to the solid blue line). Another interesting observation is that, comparing the solid blue (optimized, $N_w = 8$) and pink (prior) curves, the worst prediction from the blue curve has an average prediction error of 0.066, which is better than 98% of the predictions from the prior models. These results demonstrate that the DSI procedure is indeed able to provide improved predictions by assimilating measured data.

3.6. DSI predictions with different ‘true’ models

In the previous section, the DSI procedure was evaluated on a single ‘true’ model (Test Case 1), from which the synthetic measurements were generated. We now evaluate the methodology on two additional ‘true’ models, referred to as Test Cases 2 and 3 (we reiterate that neither true model was included in the set of $N_r = 1000$ prior models). In each of the test cases, (optimal) monitoring scenarios 1 and 4, defined in Table 2 and corresponding to $N_w = 4$ and $N_w = 8$, are considered. Since the prior models (and thus the prior simulation results) do not change between test cases, the locations of the monitoring wells are the same for all three cases.

Table 3 shows the prior results (which are the same as for Test Case 1) and DSI results, in terms of P_{10} , P_{50} , and P_{90} statistics, for the top-layer average CO_2 saturation at 200 years. Note that J^{true} varies considerably between test cases. For all three cases, the DSI predictions consistently

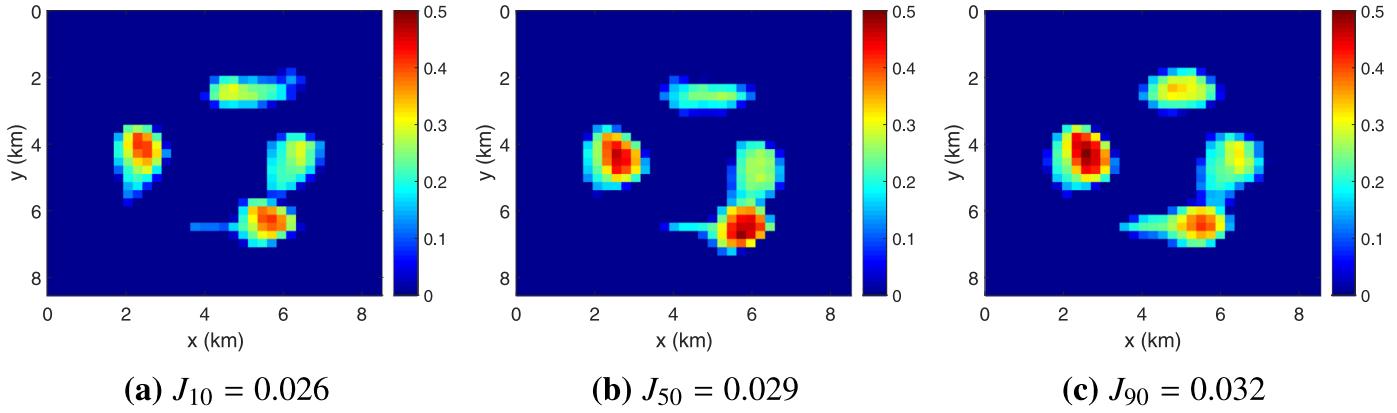


Fig. 15. Test Case 1, with monitoring scenario 4 ($N_w = 8$) defined in Table 2. (a), (b) and (c) have the same meaning as in Fig. 12. For this case, $J^{\text{true}} = 0.029$.

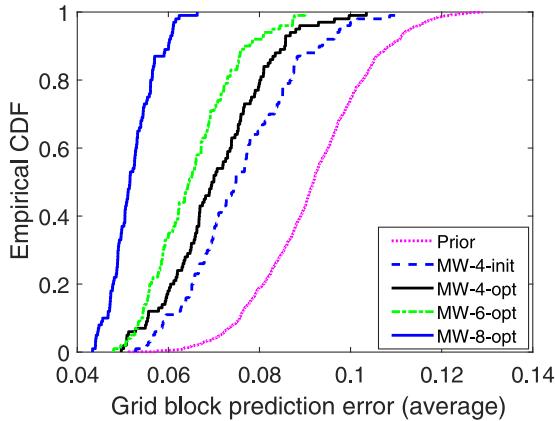


Fig. 16. Empirical CDFs computed from posterior DSIs results for pixel-wise error in saturation relative to true plume (Eq. (21)). Four different monitoring scenarios are considered, as defined in Table 2. Results for Test Case 1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3

Prior and DSIs predictions for different test cases under two monitoring scenarios (defined in Table 2). The P_{10} , P_{50} and P_{90} statistics are computed using prior or DSIs results for top-layer average CO₂ saturation at 200 years (Eq. (10)).

		True	P_{10}	P_{50}	P_{90}	$P_{90} - P_{10}$
Prior	–	–	0.027	0.045	0.056	0.029
DSI	Test Case 1 (MW-4-opt)	0.029	0.026	0.031	0.038	0.013
DSI	Test Case 1 (MW-8-opt)	–	0.026	0.029	0.032	0.006
DSI	Test Case 2 (MW-4-opt)	0.040	0.038	0.045	0.052	0.014
DSI	Test Case 2 (MW-8-opt)	–	0.033	0.038	0.041	0.008
DSI	Test Case 3 (MW-4-opt)	0.051	0.042	0.049	0.055	0.013
DSI	Test Case 3 (MW-8-opt)	–	0.044	0.048	0.052	0.008

encompass the true data within the P_{10} to P_{90} range, and show a narrower range of uncertainty (represented as $P_{90} - P_{10}$) when more monitoring wells are used.

Prediction results for the top-layer CO₂ saturation distribution, for $N_w = 4$, are shown in Fig. 17 for Test Cases 2 and 3. For Test Case 2, the true CO₂ distribution is displayed in Fig. 17d. There we see a very different pattern from the true plume for Test Case 1 (Fig. 9a). There is relatively little CO₂ at the left and bottom of the model, while the two CO₂ plumes toward the upper right display high CO₂ saturation and actually connect with each other. These phenomena are essentially captured in the DSIs results (Figs. 17e–g), even though this behavior is not directly evident in the prior results. The true plume for Test Case 3 is shown in Fig. 17h. Here we see four isolated CO₂ plumes and overall

high CO₂ saturation ($J^{\text{true}} = 0.051$). These features are largely captured in the corresponding DSIs results (Figs. 17i–k). Taken in total, the results for the three different test cases suggest that the DSIs procedure is indeed able to provide improved forecasts for top-layer CO₂ distribution at 200 years by assimilating 20 years of measurements from monitoring wells.

4. Additional considerations

In the previous section, we presented DSIs results for test cases in which all of the ‘true’ models were within the prior distribution. This means that the prior realizations used in DSIs were consistent with the underlying ‘true’ model. In this section, we evaluate DSIs performance for cases involving prior realizations that are inconsistent with the ‘true’ model. More specifically, we first consider a ‘true’ model characterized by parameters that fall outside the ranges defined in Table 1. We then describe a means for quantifying the consistency of observed data with the prior-model simulation results. Finally, we consider a ‘true’ model defined on a finer grid than that used for the prior simulations, which allows us to assess (for a particular example) the impact of model error on DSIs results.

4.1. DSIs predictions for true model parameters outside prior distributions

Table 1 displays the geological parameters for Test Case 4. We see that the ‘true’ model parameters for this example all lie outside the range of the prior distributions. Test Case 4 thus represents a potentially challenging case for DSIs due to this inconsistency between true model parameters and prior distributions. Monitoring scenario 1 (four wells, optimized) is considered in this case. Fig. 18 presents the (synthetic) observed data, shown as red circles, from a monitoring well at a single grid block (these data are for the same grid block as in Fig. 10). The gray curves represent the simulated results corresponding to the (same) $N_r = 1000$ prior models described in Section 3. It is clear that the observed data are near the edge of the prior distributions of simulated data for both pressure and CO₂ saturation.

Fig. 19 displays the distribution of top-layer average CO₂ saturation at 200 years from prior realizations (pink curve), along with the prediction corresponding to the ‘true’ model (vertical red line). We see that the true model result in this case is at the edge of the prior distribution. Specifically, $J^{\text{true}} = 0.064$ is above the 99th percentile of the prior CDF ($J_{99}^{\text{prior}} = 0.063$). This is likely due to the high vertical permeability for Test Case 4 ($k_z/k_x = 0.3$ for this case, as shown in Table 1).

Although the Test Case 4 parameters are all outside the prior ranges, the observed data are still somewhat consistent with the prior simulation results, as is evident in Fig. 18. If this were not the case; i.e., if the observed data fell well outside the prior simulation results, it would indicate that the prior was too narrow and that a broader prior should

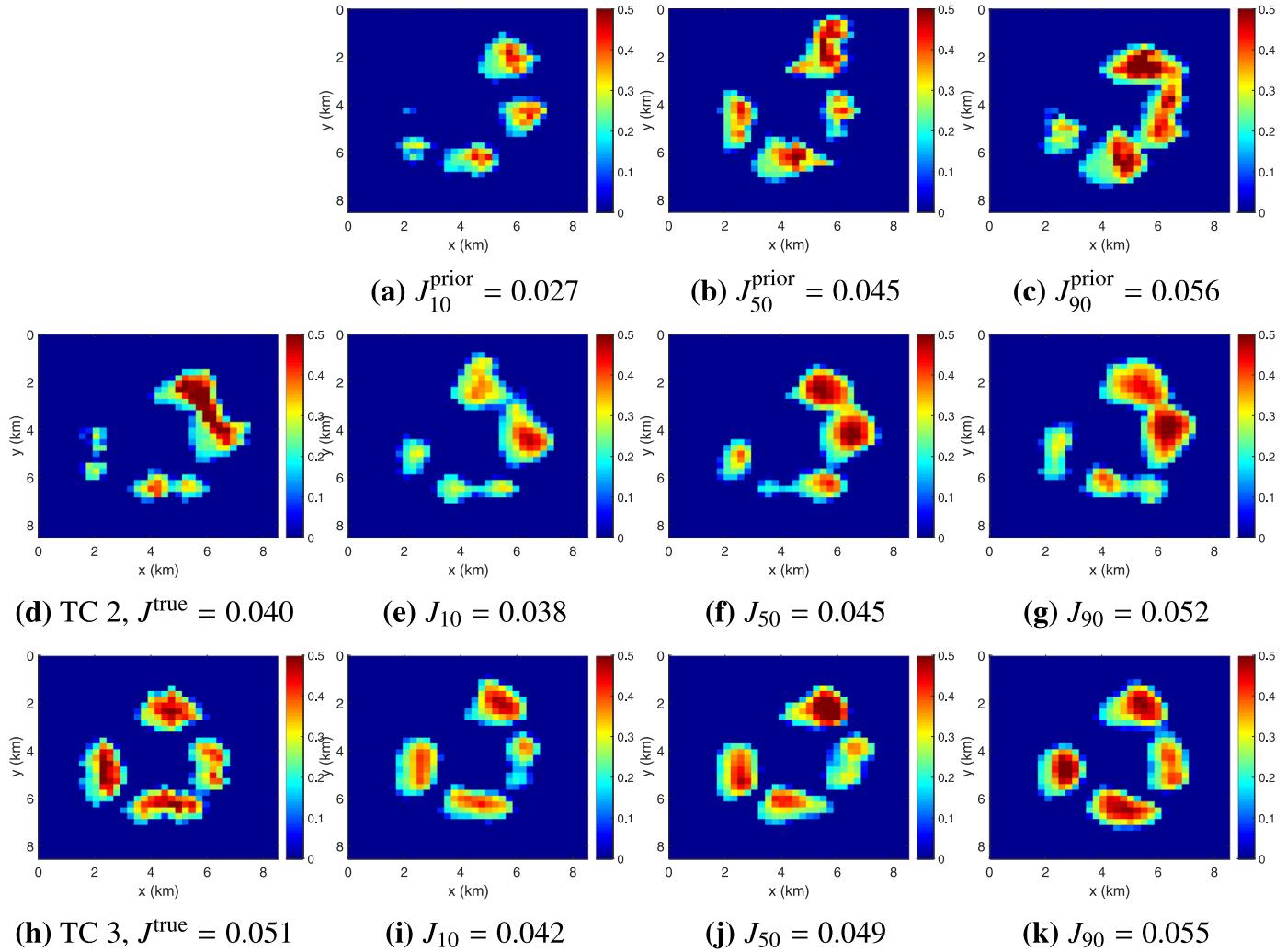


Fig. 17. Prediction of top-layer CO₂ saturation at 200 years. (a) to (c) are P_{10} , P_{50} and P_{90} predictions from prior models, (d) is the true result for Test Case (TC) 2, (e) to (g) are P_{10} , P_{50} and P_{90} DSIs predictions for Test Case 2 with four monitoring wells, (h) is the true result for Test Case (TC) 3, (i) to (k) are P_{10} , P_{50} and P_{90} DSIs predictions for Test Case 3 with four monitoring wells.

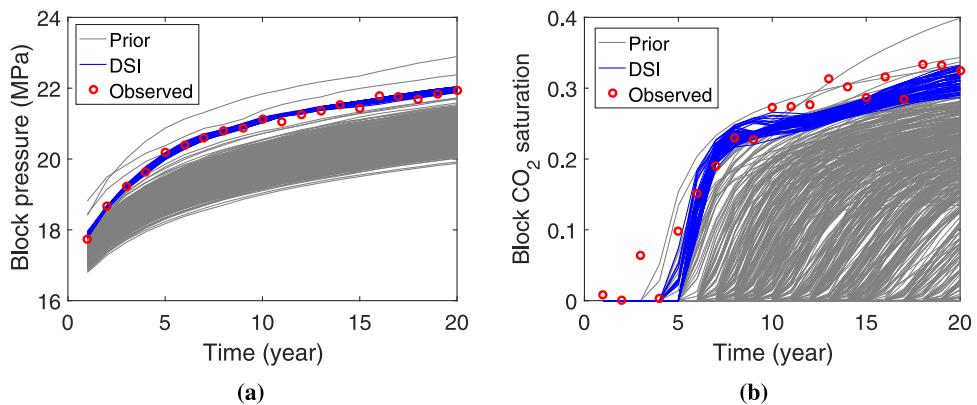


Fig. 18. DSIs predictions for historical data at a single grid block: (a) pressure, (b) CO₂ saturation (Test Case 4). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

be considered. The degree of consistency between the observed data and prior simulation results can be quantified using the Mahalanobis distance (He et al., 2017), which will be described and applied below. Within a DSIs setting, a lack of consistency here means that additional prior simulation runs must be performed. It is, however, straightforward to include simulation results from a wide range of priors in DSIs, as illustrated in Sun et al. (2017b).

We now apply the DSIs methodology with the original set of $N_r = 1000$ prior models. We reiterate that these prior models are not consistent with the underlying ‘true’ model. DSIs predictions of the grid block data are shown as blue curves in Fig. 18. It is evident that the DSIs results display a reasonable match to the observed data, though the predictions for CO₂ saturation at around 12 years somewhat underestimate the observations. The distribution of DSIs predictions for top-layer average CO₂

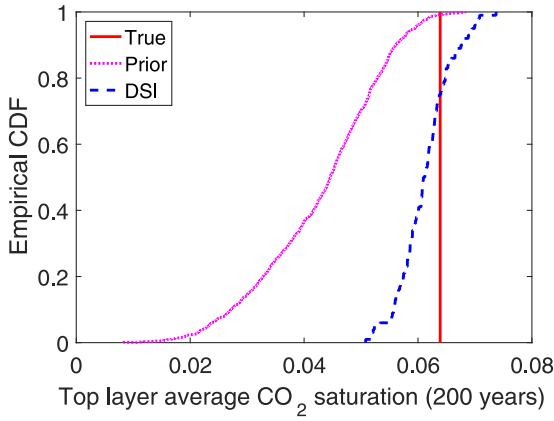


Fig. 19. Empirical CDF computed from posterior DSI results for top-layer average CO_2 saturation (J). Results are for Test Case 4. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

saturation is shown in Fig. 19 (dashed blue line). We see that the range of uncertainty is reduced significantly after assimilating the measured data, and that the true result (red line) is encompassed within the posterior CDF.

Fig. 20a presents CO_2 saturation predictions from the ‘true’ model along with P_{10} , P_{50} , and P_{90} DSI results. Some features in the ‘true’ CO_2 plumes are not fully captured in the DSI predictions. For example,

plume at the left side of the model has a more circular shape compared with the DSI posterior results, and the plume at the right has a wedge shape that is not captured by the three DSI predictions. These discrepancies likely result because the prior distributions of model parameters were specified too narrowly for this case.

It is noteworthy, however, that the general pattern and location of the true CO_2 plumes are still captured in the DSI predictions. In addition, the average CO_2 saturation obtained from the ‘true’ model ($J^{\text{true}} = 0.064$) is encompassed within the $P_{10}–P_{90}$ range of DSI results ($[J^{10}, J^{90}] = [0.056, 0.068]$). These results thus suggest that our DSI procedure displays a degree of robustness in terms of the prior parameter ranges. We reiterate that it is essential that the prior simulation results encompass the observations. If this is not the case, we would not expect DSI to provide reasonable posterior predictions. In the following section, we describe the use of Mahalanobis distance to quantify the consistency between observations and (prior) simulated data.

4.2. Use of Mahalanobis distance to assess consistency with the prior

Inversion algorithms based on Bayesian statistics rely on appropriately validated prior models (Oliver and Alfonzo, 2018). These algorithms cannot be expected to ‘work’ when the true model is outside of the prior distribution. Thus, in practice, it may be important to evaluate the consistency between the true model and the prior before performing data assimilation.

Within a DSI setting, the consistency between observations and simulated data can be assessed by, for example, visually comparing results in a low-dimensional space (Satija et al., 2017), or by using a

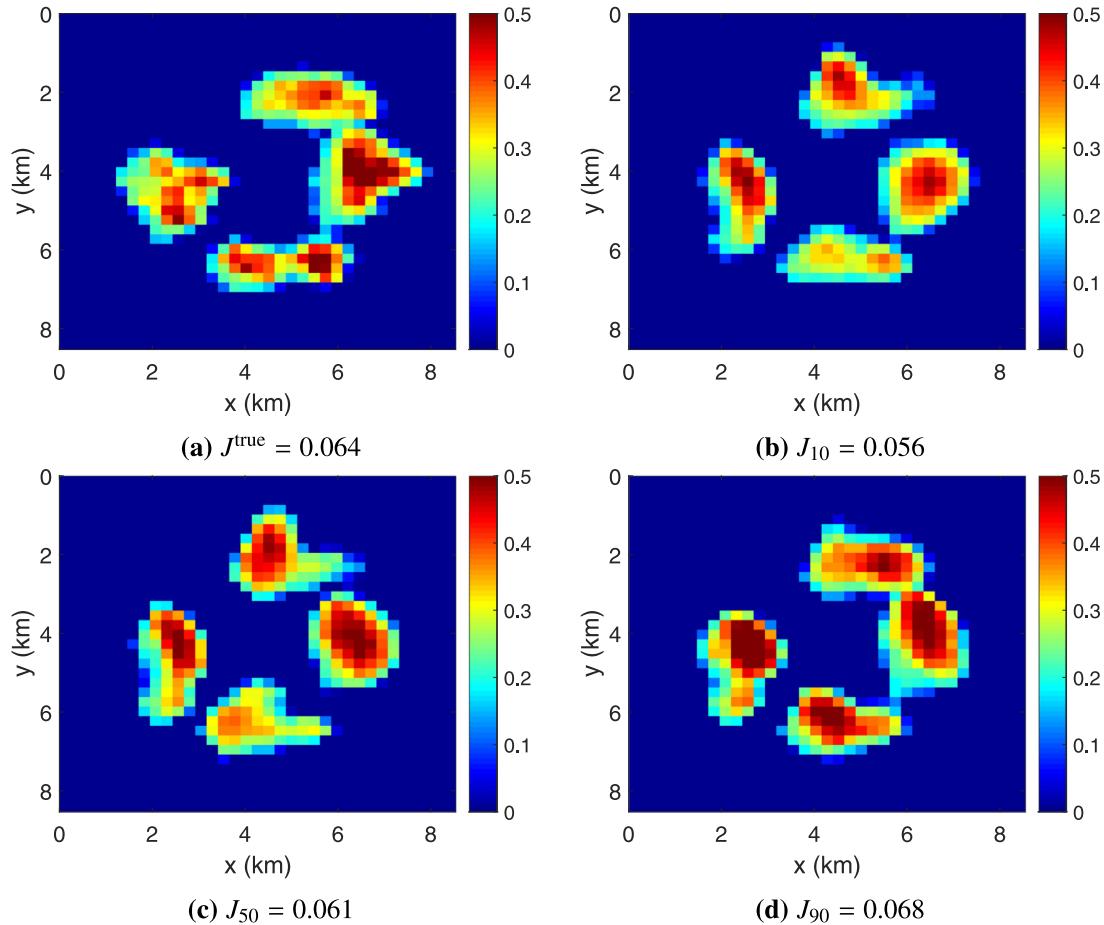


Fig. 20. Predictions for top-layer CO_2 saturation at 200 years. (a) prediction from true model, (b), (c) and (d) are P_{10} , P_{50} , and P_{90} DSI predictions (Test Case 4, monitoring scenario 1).

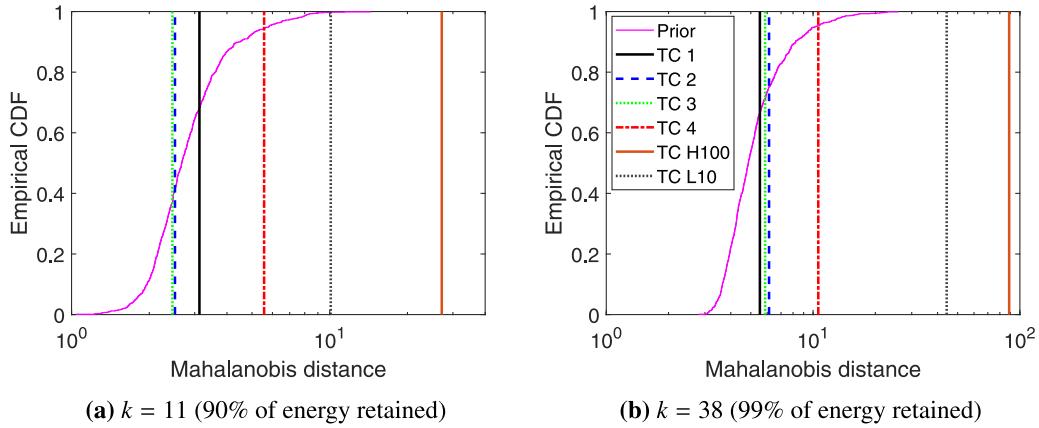


Fig. 21. Mahalanobis distance D_k for prior models (CDFs) and all test cases considered (vertical lines) with different numbers of principal components and percent of ‘energy’ retained. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

hypothesis test based on Mahalanobis distance (He et al., 2017). In this study, we consider the use of Mahalanobis distance to quantitatively gauge the consistency of the prior-model simulation results with the observed data. Mahalanobis distance is computed based on the first and second moments of the prior distribution, though it is nonetheless widely used in consistency assessments for general problems (De Maesschalck et al., 2000).

In accordance with He et al. (2017), we define the Mahalanobis distance as

$$D(\mathbf{d}) = \{(\mathbf{d} - \mu_{d_h})^T C_{d_h d_h}^{-1} (\mathbf{d} - \mu_{d_h})\}^{1/2}, \quad (22)$$

where $\mathbf{d} \in \mathbb{R}^{N_h \times 1}$ denotes a (high-dimensional) data ‘point’ (vector) and other quantities are as defined in Section 2. Eq. (22) measures the distance between a point \mathbf{d} and the mean of a multivariate distribution $N(\mu_{d_h}, C_{d_h d_p})$. In this paper, the mean and covariance matrix are computed using the simulated data from all $N_r = 1000$ prior models. There are, however, some limitations associated with computing Mahalanobis distance directly using Eq. (22). These include the fact that $C_{d_h d_h}$ is often not invertible (it is not invertible in our case), and that random noise may dominate the computation if a high number of singular vectors are retained in the approximation of $C_{d_h d_h}^{-1}$.

These two issues can be addressed by using a variation of Eq. (22) in which the Mahalanobis distance is recast as the Euclidean distance of the sum of the squares of the scores of principal components (Brereton, 2015; De Maesschalck et al., 2000). Specifically, we express \mathbf{d} as $\mathbf{d} = U\Sigma\omega + \mu_{d_h}$, where $U \in \mathbb{R}^{N_h \times k}$ corresponds to the left singular vectors of $C_{d_h d_h}$, $\Sigma \in \mathbb{R}^{k \times k}$ is a diagonal matrix containing the singular values, and $\omega \in \mathbb{R}^{k \times 1}$ contains the scores. Now, writing ω as $\omega = \Sigma^{-1}U^T(\mathbf{d} - \mu_{d_h})$ and $C_{d_h d_h}$ as $C_{d_h d_h} = U\Sigma^2U^T$, we can express Eq. (22) as

$$D_k(\mathbf{d}) = \{\omega^T \omega\}^{1/2}. \quad (23)$$

Note that this expression aligns directly with the PCA representation of \mathbf{d}_f , which is used in an intermediate step in our formulation in Section 2.

We will now assess the use of Eq. (23) in identifying the degree of consistency between observed data and prior simulated data. Observed data for Test Cases 1–4 will be assessed, as will data for two additional cases that are purposely set up to be inconsistent with the prior. These two new cases correspond to homogeneous models with high and low permeability values. Specifically, in Test Case H100, we set permeabilities to be $k_x = k_y = 100$ md and $k_z = 15$ md (these values are higher than the average permeability values in the prior models defined in Table 1), and in Test Case L10 we set $k_x = k_y = 10$ md and $k_z = 0.1$ md (which are lower than the average values in the prior models). For Test Cases H100 and L10, the porosity values are constant at $\phi = 0.2$, and no

large-scale regional pressure gradient is applied. Other properties are the same as described in Section 3.1. We again consider monitoring scenario 1. The settings for measurement error and the DSI treatments are as in Sections 3 and 4.1.

Results for Mahalanobis distance (Eq. (23)) are shown in Fig. 21. The plot in Fig. 21a ($k = 11$) corresponds to retaining 90% of the ‘energy’ in the nonzero singular vectors associated with $C_{d_h d_h}$, and the plot in Fig. 21b ($k = 38$) corresponds to retaining 99% of this energy. As noted earlier, if we retain very high fractions of energy (e.g., 99.99%), over-fitting to random measurement noise can occur, which complicates the interpretation of the computed distances. In Fig. 21, the pink curves represent D_k computed from the prior-model simulation results. To generate these curves, we take each prior-model simulation result in turn as \mathbf{d} , and then apply Eq. (23) to compute D_k . The CDF is then constructed from the full set of 1000 prior simulation results.

The values for D_k for Test Cases 1–4 and Test Cases H100 and L10 are shown in Fig. 21. It is apparent from the plots that results for Test Cases 1–3 fall well within the prior distribution, results for Test Case 4 fall toward the tail, and results for Test Cases H100 and L10 fall either outside the prior distribution entirely (Fig. 21b) or at the extreme edge (Fig. 21a for Test Case L10). These results are very encouraging and in line with our expectations, since Test Cases 1–3 are indeed from within the prior, Test Case 4 is similar to the prior models but lies outside of the prior distribution, and Test Cases H100 and L10 are quite inconsistent with the prior models. Thus we see that the Mahalanobis distance D_k , computed using Eq. (23), is very useful for identifying inconsistencies between observed data and prior-model simulation data for this problem.

If the observed data fall at the extreme edge, or outside, the prior distribution, DSI should not be immediately applied. Rather, the set of prior models should first be expanded such that the observed data fall within the prior distribution of simulated data. Unless the prior is extended in this manner, we would not expect DSI to ‘work’ for the case in question, and it would not make sense to apply it. The lack of applicability of DSI in such settings is illustrated in Fig. 22 for Test Cases L10 and H100, where we display DSI results for the CDF of top-layer average CO₂ for these cases (using the original set of 1000 simulation results as the prior). All curves are as defined previously. The fact that the true data (vertical red lines) do not intersect the DSI results indicates that the method fails in these cases. This failure is expected, however, as it is clearly predicted by the Mahalanobis distance plots in Fig. 21. Thus, a practitioner would know not to apply DSI for either of these cases without first extending the set of prior-model simulation results such that the observed data fall well within the prior distribution.

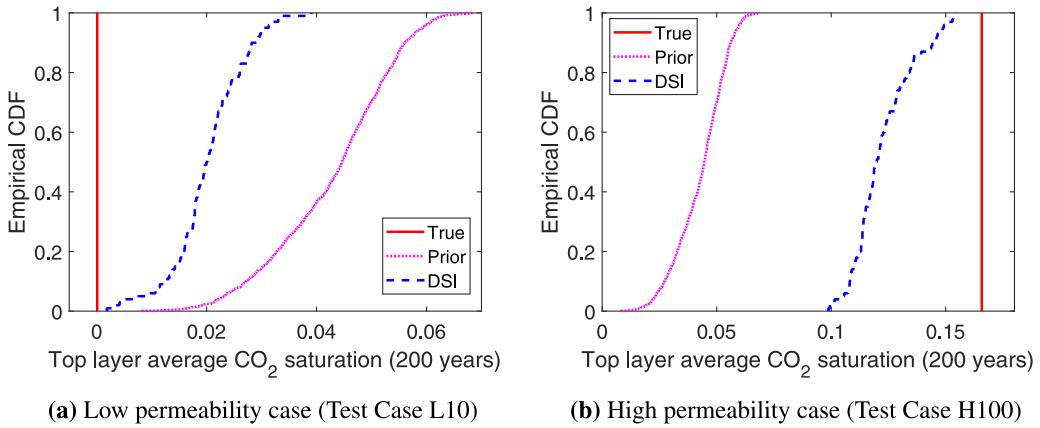


Fig. 22. Empirical CDFs computed from posterior DSI results for top-layer average CO_2 saturation (J) for homogeneous permeability test cases. DSI fails in these cases because the true models are inconsistent with the set of prior models.

4.3. Impact of grid resolution

Because our DSI procedure requires performing flow simulations for a large set of prior models, efficiency considerations motivate the use of relatively coarse models. In the results presented up to this point, the grid block size was $243 \text{ m} \times 243 \text{ m} \times 13.7 \text{ m}$ in the storage region (with grid blocks of this size, relative permeability and capillary pressure curves should, in practice, be upscaled from a finer-scale description). In addition, the observed data were generated from simulation results for a ‘true’ model with these large grid blocks. We now quantify the impact of grid refinement, which acts to better capture detailed flow behavior in the simulations. First, CO_2 saturation distributions for coarse and fine models are compared. Then, DSI is applied for a case in which the ‘true’ data are generated using a fine-scale simulation.

The higher-resolution (fine) models are generated by introducing $3 \times 3 \times 3$ refinement in each (original) block in the storage region. This results in a fine grid containing $105 \times 105 \times 33$ cells in the storage region. Additional heterogeneity in the porosity and permeability fields is not incorporated. This means that these properties are constant within $3 \times 3 \times 3$ regions in the fine model. Although finer-scale heterogeneity could also be introduced, this approach enables us to isolate and assess the impact of numerical error only, rather than the combined effect of numerical error and heterogeneity resolution level.

We first present saturation results for Test Cases 1, 2 and 3 simulated at the fine and coarse scales. The simulation setup is as described in Section 3. The CO_2 saturation distributions for the three cases are shown in Fig. 23. Original (coarse model) results are displayed in the left column, and fine-scale results appear in the right column. To enable direct comparison, the fine-model results are obtained by averaging the saturation data over the $3 \times 3 \times 3$ fine blocks corresponding to each coarse block. We see that the location and shape of the CO_2 plumes in the fine models are very similar to those in the corresponding coarse models. The average CO_2 saturations (listed in the figure captions as J_{coarse} and J_{fine}) are, however, somewhat higher in the fine models, suggesting that CO_2 vertical migration is somehow limited by the coarse discretization.

We next consider the correspondence in fine-to-coarse model results for a range of test cases. Fig. 24 displays a scatter plot of average CO_2 saturation from coarse models and the corresponding fine models for 20 different cases. Again, the coarse-model results are for the top-layer average CO_2 saturation, and the fine-model results are for CO_2 saturation averaged over the top three layers (in this way the results correspond to the same physical regions). We observe that the average CO_2 saturation is consistently higher for the fine models, with a shift of around 0.008 at

the lower saturation values, and around 0.02 at the highest saturation values. Nonetheless, a clear trend is evident, indicating that the error is systematic rather than random. We note that these observations are consistent with the results of Cameron and Durlofsky (2012), where coarse models were used for the optimization of CO_2 storage problems.

We now apply DSI for a case in which the observed data correspond to simulation results for a ‘true’ model that contains grid refinement (i.e., the storage region in the ‘true’ model is now simulated using $105 \times 105 \times 33$ blocks). In this assessment, the same $N_r = 1000$ coarse-scale prior realizations, as were used in all previous examples, will again be applied. The ‘true’ model in this case is the refined version of Test Case 1, which we refer to as Refined Test Case 1. The average CO_2 saturation for Refined Test Case 1 at 200 years is about 0.037, in contrast to 0.029 for the original (coarse) Test Case 1. This difference is consistent with the shifts noted above.

Four monitoring wells at optimized locations (monitoring scenario 1 in Table 2) are considered. Note that the monitoring wells now penetrate the top nine layers in the fine model (corresponding to the top three layers in the coarse model). The data used for DSI now correspond to pressure and CO_2 saturation averaged over the $3 \times 3 \times 3$ grid blocks around monitoring wells. This treatment maintains consistency with the coarse models used for DSI (and with Test Case 1 in Section 3). Noise is added as described earlier. The observed data (not shown) fall within the prior simulation results, so we expect DSI to provide sensible results for Refined Test Case 1.

DSI results, in terms of the empirical CDF for average CO_2 over the top three layers of the fine model, are shown in Fig. 25. The prior CDF, shown as the pink curve, is based on coarse-model (original) simulations, so it is identical to that for Test Case 1 in Fig. 14. The posterior DSI CDF, shown as the dashed blue curve, is however different from the posterior for monitoring well scenario 1 in Fig. 14 (black curve in that figure), since the data in the current case are different; i.e., they now derive from a refined-model simulation. More specifically, for the original Test Case 1 ($J^{\text{true}} = 0.029$), we had $J_{10} = 0.026$, $J_{50} = 0.031$, $J_{90} = 0.038$, while for Refined Test Case 1 ($J^{\text{true}} = 0.037$, indicated by the vertical red line in Fig. 25), we have $J_{10} = 0.029$, $J_{50} = 0.035$, $J_{90} = 0.042$. Thus we see that the DSI predictions shift towards higher saturation values, consistent with the data from the Refined Test Case 1 simulation. This is an encouraging result, since it demonstrates that DSI is able to provide reasonable predictions using coarse-scale prior simulations even when the ‘true’ data potentially contain finer-scale information. This suggests that DSI may be reasonably robust with respect to this type of model error. DSI performance will, however, need to be assessed for a range of cases to quantify its behavior in this regard.

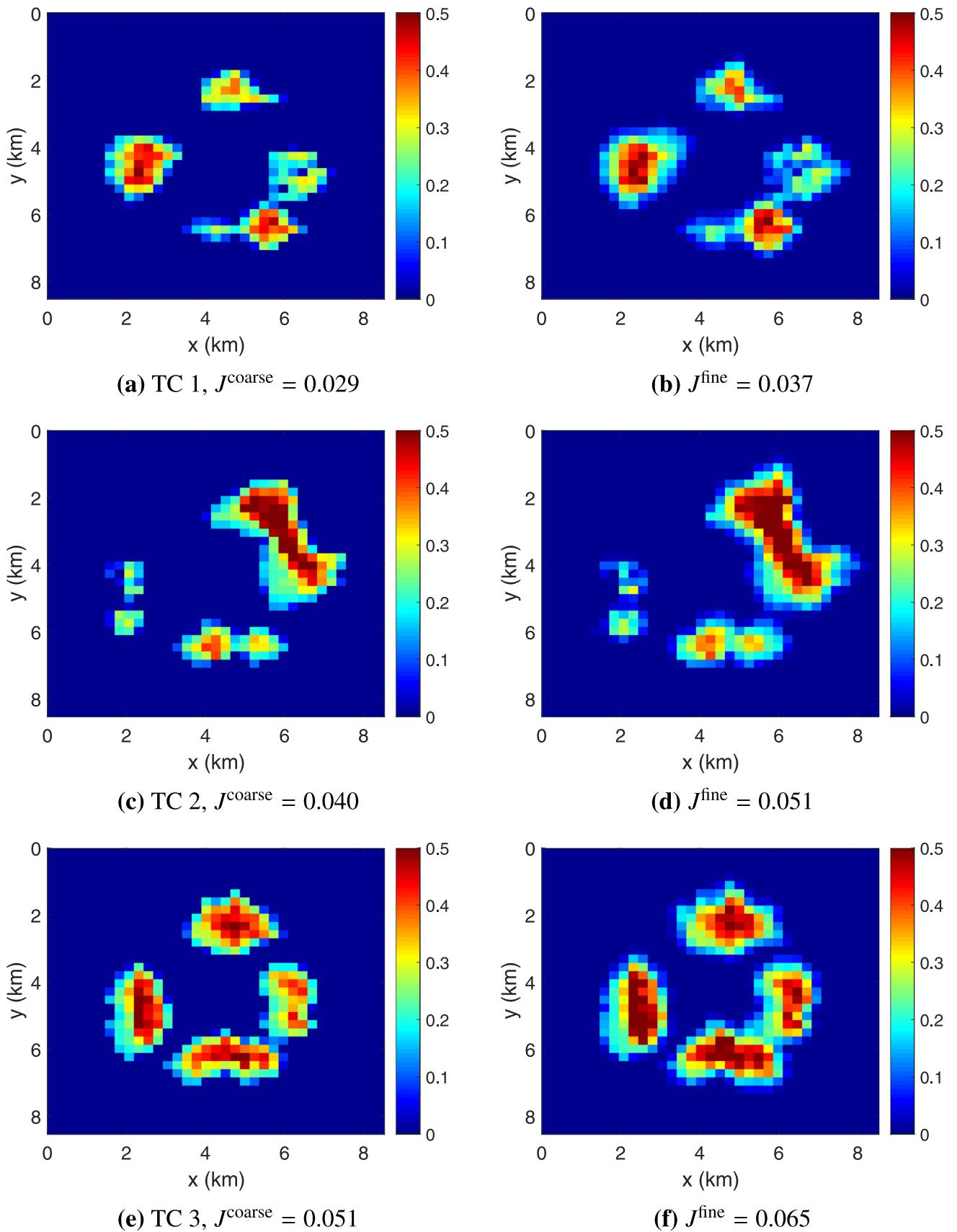


Fig. 23. CO₂ saturation distributions at 200 years. (a), (c), and (e) correspond to Test Cases 1, 2, and 3 with (the original) coarse grid of 35 × 35 × 11 blocks in the storage aquifer, and (b), (d), and (f) are the associated predictions for fine-grid models with 105 × 105 × 33 blocks in the storage region. Maps in (a), (c), and (e) are for the top layer, while those in (b), (d), and (f) are obtained by averaging fine-scale saturations over the 3 × 3 × 3 fine blocks corresponding to each coarse block.

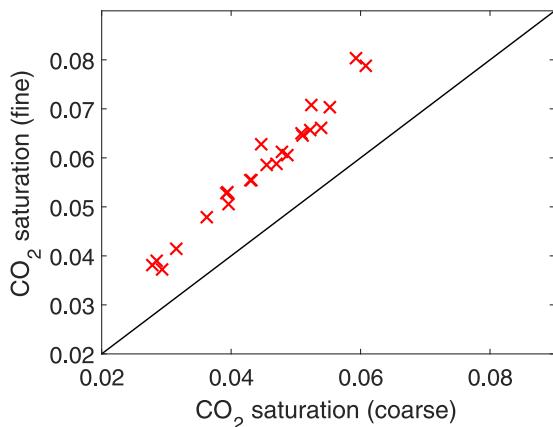


Fig. 24. Cross plot of the coarse-scale and fine-scale average CO_2 saturation for 20 realizations. Coarse-scale results are for the top layer, while fine-scale results correspond to averages over the upper three layers.

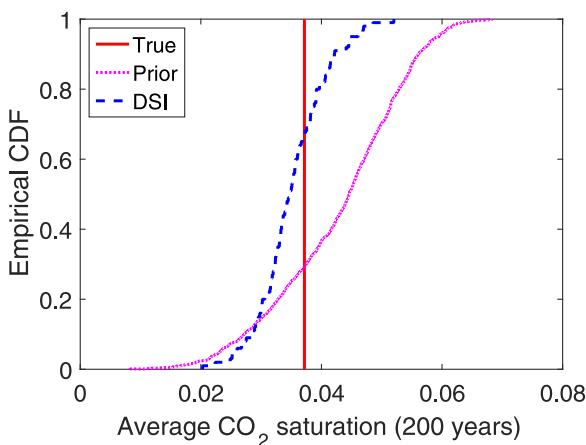


Fig. 25. Empirical CDF computed from posterior DS1 results for average CO_2 saturation (J).

5. Concluding remarks

In this work, we devised and applied a data-space inversion (DSI) procedure for the prediction of CO_2 saturation distribution in the top layer of a storage aquifer. The method provides posterior estimates based on prior-model simulation results and measurements at some number of monitoring wells. In contrast to traditional history matching approaches, the DSI procedure does not entail calibration of model parameters, and thus does not provide posterior models. In the DSI procedure, a set of realizations (1000 in the cases considered here), honoring prior geological information, is first generated and simulated to provide prior samples for variables corresponding to measurement data and prediction quantities of interest (QoI). A reparameterization based on principal component analysis and histogram transformation is applied to facilitate the sampling of posterior data variables. This sampling is accomplished using the randomized maximum likelihood (RML) method in the data space.

We also applied analytical results from the DSI formulation to quantify the effectiveness of measurements, over all prior models, for given

monitoring well locations. The analytical expressions used for this determination are consistent with the ensemble variance analysis formulations studied by He et al. (2018). Measurement effectiveness is defined here in terms of the expected reduction of posterior variance for a specified QoI. In this work, this QoI was taken to be the average CO_2 saturation in the top layer of the storage aquifer at 200 years. Given this analytical (and thus fast to compute) measure of effectiveness, we successfully optimized the monitoring well locations using a genetic algorithm.

The overall DSI procedure was then applied to synthetic three-dimensional heterogeneous aquifer systems. Uncertainties in the underlying variogram model, in the porosity and permeability distributions, and in the regional-scale pressure gradient were incorporated into the prior models. Detailed physics, including relative permeability hysteresis and capillary pressure heterogeneity, were included in the flow simulations. We evaluated DSI for three different ‘true’ models, with varying numbers of monitoring wells. DSI predictions were compared with true-model simulation results for the top-layer CO_2 saturation distribution. In all cases involving optimal monitoring wells, we observed that the DSI posterior predictions for the amount and general distribution of CO_2 in the top layer displayed considerably less uncertainty than estimates from the prior models. We further showed that DSI prediction accuracy was consistently enhanced as more monitoring wells were used, and that a higher degree of uncertainty reduction was achieved using optimally located (instead of heuristically placed) monitoring wells.

DSI was also shown to provide reasonable results for a ‘true’ model characterized by parameters that lie outside the prior distributions, and for a case in which the observed data were generated from a fine-grid simulation. We also considered two homogeneous test cases that led to observed data that were very inconsistent with the prior simulation data. We showed that an appropriately computed Mahalanobis distance is able to clearly identify this inconsistency, and that DSI will fail if directly applied for such cases.

There are several directions in which this work could be extended. It will be useful to apply our DSI procedures to more complex geological models such as bimodal channelized systems, which can be challenging for traditional (model-based) history matching methods. Although the prediction of posterior well-based quantities for such systems has been accomplished in a DSI setting in Sun and Durlofsky (2017), it will be interesting to assess the predictive ability of DSI for QoI of the type considered here (such as top-layer CO_2 saturation). The use of DSI for different problem setups, such as CO_2 leakage detection under uncertainty as described by Cameron et al. (2016) and Jeong et al. (2018), should also be considered. In these applications, leakage may have already occurred, which will change the formulation (objective) of the optimization problem. It may be possible to provide better uncertainty quantification for this problem using DSI than with model-based methods. Finally, some of the detailed treatments in our DSI procedure could be further studied, with the goal of reducing the number of prior runs required, or improving DSI predictions for a given number of prior runs. For example, it may be beneficial to apply functional PCA, as used by Satija and Caers (2015) for dimension reduction of tracer data, in place of the PCA parameterization used in this work.

Acknowledgments

We thank David Cameron and Zhaoyang Larry Jin for providing aquifer models and simulation files, and for useful discussions. We are grateful to Chevron ETC and the Stanford Smart Fields Consortium for financial support.

Appendix

Algorithm 1 Data assimilation using DSI (Sun and Durlofsky, 2017).

1. Generate a set of N_r prior realizations by sampling the prior distribution of model parameters (including geological parameters and other uncertain parameters)
2. Perform flow simulation to obtain an ensemble of simulated data vectors, denoted as $(\mathbf{d}_t)_i$, $i = 1, 2, \dots, N_r$. The simulated data vector \mathbf{d}_t includes data corresponding to both measurement data and prediction QoI
3. Compute the basis matrix Φ and construct the histogram transformation function \mathbf{h}_T (shown in Eq. (9))
4. **for** $n = 1$ to N_{post} (where N_{post} is the specified number of posterior forecasts) **do**
5. Sample $\mathbf{d}_{\text{obs}}^*$ from multivariate Gaussian distribution $N(\mathbf{d}_{\text{obs}}, C_D)$, and ξ^* from multivariate normal distribution
6. Generate ξ_{opt} by minimizing $S(\xi)$ in Eq. (7) using selected optimizer. Here the Matlab *fmincon* optimization function (without analytical gradients) is used
7. Compute normalized mismatch $S(\xi_{\text{opt}})/N_{\text{obs}}$. If this value is larger than specified threshold (a value of 10 is used in this study – see Sun and Durlofsky (2017) for discussion of this parameter) **do** restart from Step 5
8. Obtain corresponding posterior sample as $\mathbf{f}(\xi_{\text{opt}})$ (Eq. (9))
9. **end for**

Algorithm 2 Optimize placement of monitoring well locations using a data-space approach.

1. Generate a set of N_r prior realizations (in practice we use the realizations generated in Step 1 in Algorithm 1)
2. Perform flow simulation on all prior realizations, and record simulated data at all grid blocks (in practice we use the flow results generated in Step 2 in Algorithm 1)
3. Compute J_i , $i = 1, 2, \dots, N_r$ (Eq. 10) and σ_{JJ}^2 (Eq. (14))
4. Specify initial guess for monitoring well location \mathbf{y}
5. Repeat the following steps for each iteration, until optimal locations \mathbf{y}_{opt} in Eq. 18 are found
6. Compute $C_{dh dh}$, $C_{J dh}$, and $C_{dh J}$ (Eqs. (15) and (16)) using simulated data corresponding to monitoring well location \mathbf{y}
7. Compute $-U_R(\mathbf{y})$ in Eq. (18)
8. Obtain updated \mathbf{y} from selected optimizer. In this study, Matlab (2017a version) integer optimizer *ga*, with default options, is used

References

- Aanonsen, S.I., Nævdal, G., Oliver, D.S., Reynolds, A.C., Vallès, B., 2009. The ensemble Kalman filter in reservoir engineering—a review. *SPE J.* 14 (3), 393–412.
- Brereton, R.G., 2015. The Mahalanobis distance and its relationship to principal component scores. *J. Chemom.* 29 (3), 143–145.
- Cameron, D.A., Durlofsky, L.J., 2012. Optimization of well placement, CO₂ injection rates, and brine cycling for geological carbon sequestration. *Int. J. Greenhouse Gas Control* 10, 100–112.
- Cameron, D.A., Durlofsky, L.J., 2014. Optimization and data assimilation for geological carbon storage. In: Computational Models for CO₂ Sequestration and Compressed Air Energy Storage. CRC Press, pp. 357–388.
- Cameron, D.A., Durlofsky, L.J., Benson, S.M., 2016. Use of above-zone pressure data to locate and quantify leaks during carbon storage operations. *Int. J. Greenhouse Gas Control* 52, 32–43.
- Caspari, E., Müller, T., Gurevich, B., 2011. Time-lapse sonic logs reveal patchy CO₂ saturation in-situ. *Geophys. Res. Lett.* 38 (13).
- Chen, B., Harp, D.R., Lin, Y., Keating, E.H., Pawar, R.J., 2018. Geologic CO₂ sequestration monitoring design: a machine learning and uncertainty quantification based approach. *Appl. Energy* 225, 332–345.
- Chen, Y., Zhang, D., 2006. Data assimilation for transient flow in geologic formations via ensemble Kalman filter. *Adv. Water Resour.* 29 (8), 1107–1122.
- Daley, T.M., Myer, L.R., Peterson, J., Majer, E., Hoversten, G., 2008. Time-lapse crosswell seismic and VSP monitoring of injected CO₂ in a brine aquifer. *Environ. Geol.* 54 (8), 1657–1665.
- De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L., 2000. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* 50 (1), 1–18.
- Doughty, C., Freifeld, B.M., Trautz, R.C., 2008. Site characterization for CO₂ geologic storage and vice versa: the Frio brine pilot, Texas, USA as a case study. *Environ. Geol.* 54 (8), 1635–1656.
- Emerick, A.A., Reynolds, A.C., 2013. Ensemble smoother with multiple data assimilation. *Comput. Geosci.* 55, 3–15.
- Ennis-King, J., Dance, T., Xu, J., Boreham, C., Freifeld, B., Jenkins, C., Paterson, L., Sharma, S., Stalker, L., Underschultz, J., 2011. The role of heterogeneity in CO₂ storage in a depleted gas field: history matching of simulation models to field data for the CO2CRC Otway Project, Australia. *Energy Procedia* 4, 3494–3501.
- Espinet, A., Shoemaker, C., Doughty, C., 2013. Estimation of plume distribution for carbon sequestration using parameter estimation with limited monitoring data. *Water Resour. Res.* 49 (7), 4442–4464.
- Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.* 53 (4), 343–367.
- Farajzadeh, R., Ranganathan, P., Zitha, P.L.J., Bruining, J., 2011. The effect of heterogeneity on the character of density-driven natural convection of CO₂ overlying a brine layer. *Adv. Water Resour.* 34 (3), 327–339.
- González-Nicolás, A., Baú, D., Alzraiee, A., 2015. Detection of potential leakage pathways from geological carbon storage by fluid pressure data assimilation. *Adv. Water Resour.* 86, 366–384.
- Harville, D., 2003. The expected value of a conditional variance: an upper bound. *J. Stat. Comput. Simul.* (8) 609–612.
- He, J., Sarma, P., Bhark, E., Tanaka, S., Chen, B., Wen, X.-H., Kamath, J., 2018. Quantifying expected uncertainty reduction and value of information using ensemble-variance analysis. *SPE J.* 23 (02), 428–448.
- He, J., Tanaka, S., Wen, X.-H., Kamath, J., 2017. Rapid S-curve update using ensemble variance analysis with model validation. Paper SPE-185630-MS, presented at the SPE Western Regional Meeting, Bakersfield, California, USA, 23–27 April (2017).
- Jeong, H., Sun, A.Y., Zhang, X., 2018. Cost-optimal design of pressure-based monitoring networks for carbon sequestration projects, with consideration of geological uncertainty. *Int. J. Greenhouse Gas Control* 71, 278–292.
- Jin, Z.L., Durlofsky, L.J., 2018. Reduced-order modeling of CO₂ storage operations. *Int. J. Greenhouse Gas Control* 68, 49–67.
- Kitanidis, P.K., 1986. Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resour. Res.* 22 (4), 499–507.
- Krishnamurti, T.N., 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.
- Magnant, Z., 2011. Numerical Methods for Optimal Experimental Design of Ill-posed Problems. Emory University Ph.D. thesis.
- Mallet, V., Stoltz, G., Mauricette, B., 2009. Ozone ensemble forecast with machine learning algorithms. *J. Geophys. Res.* 114 (D5), 148–227.
- McLaughlin, D., Townley, L.R., 1996. A reassessment of the groundwater inverse problem. *Water Resour. Res.* 32 (5), 1131–1161.
- Oladyshkin, S., Class, H., Helmig, R., Nowak, W., 2011. An integrative approach to robust design and probabilistic risk assessment for CO₂ storage in geological formations. *Computat. Geosci.* 15 (3), 565–577.
- Oliver, D.S., 1996. On conditional simulation to inaccurate data. *Math. Geosci.* 28 (6), 811–817.
- Oliver, D.S., Alfonzo, M., 2018. Calibration of imperfect models to biased observations. *Computat. Geosci.* 22 (1), 145–161.
- Oliver, D.S., Chen, Y., 2011. Recent progress on reservoir history matching: a review. *Computat. Geosci.* 15 (1), 185–221.
- Oliver, D.S., Reynolds, A.C., Liu, N., 2008. Inverse Theory for Petroleum Reservoir Characterization and History Matching. Cambridge University Press.
- Rabinovich, A., Ithlisawatpan, K., Durlofsky, L.J., 2015. Upscaling of CO₂ injection into brine with capillary heterogeneity effects. *J. Pet. Sci. Eng.* 134, 60–75.
- Remy, N., Boucher, A., Wu, J., 2009. Applied Geostatistics with SGeMS: A User's Guide. Cambridge University Press.
- Reynolds, A.C., He, N., Oliver, D.S., 1999. Reducing uncertainty in geostatistical description with well-testing pressure data. In: Reservoir Characterization—Recent Advances. American Association of Petroleum Geologists, Tulsa, pp. 149–162.
- Robert, C.P., 2004. Monte Carlo Methods. Wiley Online Library.
- Saadatpoor, E., Bryant, S.L., Sepehrnoori, K., 2010. New trapping mechanism in carbon sequestration. *Transp. Porous Media* 82 (1), 3–17.
- 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.
- Satija, A., Scheidt, C., Li, L., Caers, J., 2017. Direct forecasting of reservoir performance using production data without history matching. *Computat. Geosci.* 21 (2), 315–333.
- 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.
- Sun, A.Y., Nicot, J., 2012. Inversion of pressure anomaly data for detecting leakage at geologic carbon sequestration sites. *Adv. Water Resour.* 44, 20–29.
- Sun, A.Y., Wheeler, M.F., Islam, A., 2017a. Identifying attributes of CO₂ leakage zones in shallow aquifers using a parametric level set method. *Greenhouse Gases Sci. Technol.* 7 (4), 649–664.
- Sun, A.Y., Zeidouni, M., Nicot, J., Lu, Z., Zhang, D., 2013. Assessing leakage detectability at geologic CO₂ sequestration sites using the probabilistic collocation method. *Adv. Water Resour.* 56, 49–60.
- Sun, W., 2014. Data Driven History Matching for Reservoir Production Forecasting. Stanford University Master's thesis.

- Sun, W., Durlofsky, L.J., 2017. A new data-space inversion procedure for efficient uncertainty quantification in subsurface flow problems. *Math. Geosci.* 49 (6), 679–715.
- Sun, W., Hui, M., Durlofsky, L.J., 2017b. Production forecasting and uncertainty quantification for naturally fractured reservoirs using a new data-space inversion procedure. *Computat. Geosci.* 21 (5–6), 1443–1458.
- Tarantola, A., 2005. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM.
- Yeh, W.W.-G., 1986. Review of parameter identification procedures in groundwater hydrology: the inverse problem. *Water Resour. Res.* 22 (2), 95–108.
- Yoon, S., Williams, J.R., Juanes, R., Kang, P.K., 2017. Maximizing the value of pressure data in saline aquifer characterization. *Adv. Water Resour.* 109, 14–28.
- Zhang, L., Dilmore, R., Namhata, A., Bromhal, G., 2018. Feasibility of CO₂ migration detection using pressure and CO₂ saturation monitoring above an imperfect primary seal of a geologic CO₂ storage formation: a numerical investigation. *Computat. Geosci.* 22 (3), 909–923.
- Zhou, H., Gómez-Hernández, J.J., Li, L., 2014. Inverse methods in hydrogeology: evolution and recent trends. *Adv. Water Resour.* 63, 22–37.