Efficient Characterization of Uncertain Model Parameters with a Reduced-Order Ensemble Kalman Filter

SIAM J. SCI. COMPUT.
Vol. 36, No. 2, pp. B198–B224

EFFICIENT CHARACTERIZATION OF UNCERTAIN MODEL PARAMETERS WITH A REDUCED-ORDER ENSEMBLE KALMAN FILTER*

BINGHUAI LIN† AND DENNIS MCLAUGHLIN†

Abstract. Spatially variable model parameters are often highly uncertain and difficult to observe. This has prompted the widespread use of Bayesian characterization methods that can infer parameter values from measurements of related variables, while explicitly accounting for uncertainty. Ensemble versions of Bayesian characterization are particularly convenient when uncertain variables have complex spatial structures that do not conform to Gaussian descriptions. However, ensemble methods can be time consuming for high-dimensional problems. This paper describes a reduced-order approach to ensemble characterization that is particularly well suited for subsurface flow and transport problems. It uses a truncated discrete cosine transform to reduce the dimensionality of spatially variable time-invariant model parameters and a nonlinear extension of principle orthogonal decomposition to reduce the dimensionality of dynamic model states. The resulting nonlinear reduced-order model can be included in the forecast step of a reduced-order ensemble Kalman filter. These concepts are illustrated in a subsurface solute transport problem using ensembles produced by full- and reduced-order models. These ensembles are very similar when there are no measurement updates. When the forecast ensemble is recursively updated with measurements the reduced-order Kalman filter does at least as well as the full-order filter in characterizing a dynamic solute plume, even though its augmented state dimension is only 2% of the dimension of the full-order state. This substantial increase in efficiency implies that a reduced-order filter with the same ensemble size as its full-order counterpart can give comparable performance for orders of magnitude less computational effort or it can use a much larger ensemble for the same computational effort. The possibility of substantial increases in ensemble size could lead to performance improvements through reductions in sampling error and in the rank of the ensemble null space. Also, a reduced-order model similar to the one described here could be used in ensemble real-time control applications, where it can decrease the effort required for both characterization and control.

Key words. model order reduction, ensemble Kalman filter, proper orthogonal decomposition, discrete cosine transform

AMS subject classifications. 93E11, 65C60, 93B11

DOI. 10.1137/130910415

1. Introduction.

One of the most challenging problems in predictive modeling is to characterize spatially variable physical properties that are uncertain and difficult to observe. Uncertainties in these properties lead to prediction uncertainties that can reduce a model's usefulness for both forecasting and control applications. Ensemble versions of Bayesian estimation provide a convenient way to infer spatially distributed parameters from noisy measurements of related variables. Ensemble methods are especially helpful when the uncertain parameters have complex spatial structures that do not conform to Gaussian descriptions. Such situations are often encountered in geoscience applications, including assessment of water and petroleum resources, contaminant remediation, and carbon sequestration. Unfortunately, ensemble methods can be time consuming for high-dimensional problems since they require repeated solution of the forward models that relate parameters to measured variables; one so-

*Submitted to the journal's Computational Methods in Science and Engineering section February 21, 2013; accepted for publication (in revised form) December 19, 2013; published electronically March 20, 2014. This work was supported by Shell and Eni S.p.A.
http://www.siam.org/journals/sisc/36-2/91041.html
†Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge MA (blin@mit.edu, dennism@mit.edu).

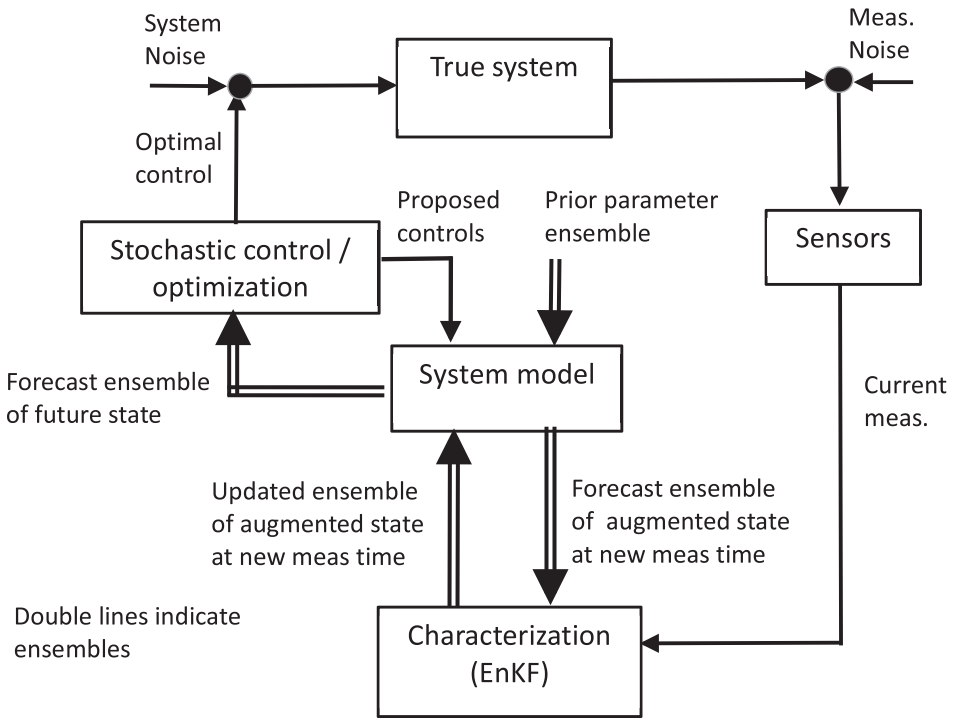

FIG. 1. *Basic components of an ensemble feedback control procedure, showing the role of real-time characterization.*

lution for each ensemble member. Since the performance of these methods depends (in part) on the size of the ensemble, computational limitations can have an adverse impact on the characterization.

In this paper we focus on real-time Bayesian characterization procedures that continually update a parameter ensemble as new measurements become available. This sequential approach to characterization is particularly useful for real-time control applications. Ensemble methods for addressing real-time characterization problems include various forms of the ensemble Kalman filter [20, 21] and the particle filter [28]. These techniques are suitable for many nonlinear problems and also offer considerable flexibility for describing prior uncertainty.

To provide context, Figure 1 shows how an ensemble real-time characterization strategy could fit into a larger framework that includes real-time control. Double lines indicate ensembles of many replicates while single lines indicate deterministic variables. It is convenient in this setting to define an augmented state vector that appends the time-invariant input parameters to the vector of dynamic states. The characterization procedure in the lower portion of the diagram uses the new measurement to update an ensemble of augmented states forecast from the previous measurement time. The system model uses the updated ensemble of current states to predict an ensemble of future states throughout a particular control horizon. In each control iteration the predicted ensemble depends on a specified control strategy. The control algorithm searches until it finds the strategy that optimizes a probabilistic measure of performance (usually a cost function averaged over the ensemble of predicted system

responses). This optimal control is applied to the true system, which generates a new set of measurements at the next update time. The entire process is then repeated.

Here we focus on the characterization portion of this real-time control procedure. We use an ensemble Kalman filtering algorithm to characterize the augmented state because it is convenient and relatively efficient compared to the alternatives [1, 2, 19, 22, 25, 30, 38, 46]. Efficiency is especially important in a high-dimensional distributed parameter ensemble control application, where the filter must be rerun in each iteration of the control optimization. The ensemble Kalman filter naturally divides into a forecast step carried out between two measurement times and an update step carried out sequentially at measurement times. The entire procedure is initialized with a prior ensemble of replicates and states at the beginning of the first forecast interval.

Even an efficient ensemble Kalman filter can be computationally costly if the forecast model is high dimensional. In such cases the ensemble filter can be expensive because (1) the cost of a single replicate forecast is high and (2) many replicates are needed to extract the greatest possible amount of information from a given set of measurements. When the ensemble size is less than the dimension of the augmented state the ensemble Kalman filter updates at any given measurement time are confined to a subspace spanned by the ensemble of forecast vectors [31]. In most geoscience applications computational limitations require an ensemble size that is much smaller than the state dimension. From a geometric perspective this constrains the filter update to lie in a reasonably small subspace of the complete state space. Errors orthogonal to this subspace cannot be corrected. Consequently, there is considerable incentive to select the ensemble subspace in a way that effectively captures dominant errors. One approach to this problem is to keep the state space large and identify the best possible subspace spanned by a small number of replicates. An alternative is to reduce the size of the state space so that the null space orthogonal to the ensemble subspace has a relatively low dimension. In this case, measurements will be better able to correct estimation errors. Both options involve approximations and it is possible to combine them.

This paper describes a new model reduction approach that can be used to construct a low-dimensional filter suitable for ensemble characterization. We use different methods for parameter and state reduction, extending the discrete cosine transform (for parameters) and principal orthogonal decomposition (for states) so they are suitable for ensemble forecasting. Our state reduction method can be applied to nonlinear models and is especially well suited for quadratic models. We illustrate the concepts with a virtual subsurface solute transport example. In this example a reduced-order filter performs comparably to or even better than the full-order filter while requiring much less computational effort. Conversely, for a fixed computational effort, the reduced-order filter can use a much larger ensemble than the full-order filter. This latter possibility has the advantage of reducing sampling errors in the ensemble-based covariances used to compute the filter update [6, 47]. The benefits of a reduced-order characterization in other applications will depend on how well the reduced-order model captures the behavior of the true system. This is an issue that we consider as part of our discussion.

We describe our general approach in section 2. Section 3 uses the solute transport example to illustrate how this approach is applied. Section 4 describes results from the example, comparing ensembles generated with full and reduced-order models, with and without measurement updates. We conclude in section 5 with a discussion of extensions and other applications.

2. Problem formulation and methodology. The spatially distributed models of most interest here are formulated in terms of partial differential equations derived from conservation laws. We assume here that these equations have been discretized over space, but not time, to give a set of nonlinear equations having the following form:

$$\frac{dx}{dt} = F(x, w, \alpha, u), \quad x(t_0) = x_0 \text{ specified}, \tag{2.1}$$

$$G(x, w, \alpha, u) = 0. \tag{2.2}$$

Here we have explicitly distinguished two types of spatially discretized equations, both derived from the original partial differential equations. These are a set of N_x ordinary differential equations (2.1) derived from equations with both time and space derivatives and a set of N_w algebraic equations (2.2) derived from equations with only space derivatives. In both cases the spatial discretization accounts for the effect of known boundary conditions and is carried out over a specified computational grid. We also distinguish an N_x-dimensional vector of prognostic variables (or states) x, an N_w-dimensional vector of diagnostic variables w, an N_α-dimensional vector of time-invariant but spatially variable uncertain model parameters α, and an N_u-dimensional vector u of specified time-varying control variables defined at discrete locations or along boundaries. In a stochastic formulation x and w are considered to be uncertain by virtue of their dependence on the uncertain parameters. The characterization process uses measurements of the states (or related variables) to reduce uncertainty in both parameters and states, with the ultimate goal of improving the model's predictions.

When the modeled system exhibits variability over a wide range of scales the parameter and state vectors are often very large, easily reaching sizes of the order 10^6 or more in problems with three spatial dimensions. As mentioned earlier, parameter characterization is very difficult in such large problems unless we replace the full-order model of (2.1) and(2.2) with a reduced-order model that captures the essential features of the system but is much faster and easier to work with. This reduced-order model should be formulated in terms of low-dimensional state and parameter vectors but should accommodate the same control vector as the full-order model.

The most common approach to model reduction relies on projection-based methods that derive the reduced-order model by projecting the high-dimensional spatially discretized governing equations onto the subspace spanned by a set of basis vectors [7]. The basis can be generated in various ways, including Krylov subspace methods [8, 27, 29], balanced truncation and its variants [41, 43], and proper orthogonal decomposition (POD) [9, 40, 45]. Considerable effort has been devoted to the problem of identifying basis vectors that work well over a range of possible inputs. Examples include POD and Krylov-based sampling methods [16, 35], greedy sampling approaches [12, 13], and interpolation methods [4, 17]. Note that the reduced-order model subspace generally contains the ensemble subspace, which may be significantly smaller. The use of these two different subspaces adds flexibility to the approximations introduced in a reduced-order characterization.

Many model reduction studies focus only on the system states. However, reduction of the number of parameters is also important, especially in spatially distributed applications. Lieberman, Willcox, and Ghattas [36] describe an approach that separates the parameter and state reduction processes. The reduced parameter basis is derived by solving an optimization problem that minimizes output errors between full

order and reduced-order models, while penalizing low probability parameters. This method appears to work well in a subsurface flow application for relatively smooth conductivity fields. Kaleta et al. [34] use a linear reduced-order model derived from POD to estimate uncertain parameters with a variational procedure. In this paper we use different model reduction methods for both parameter and state reduction. The parameter reduction relies on linear POD concepts while the state reduction uses a nonlinear extension of the POD approach. For these reasons it is convenient to derive our reduced-order model by considering parameter and state reduction separately.

2.1. Parameter reduction with the discrete cosine transform. Compression of spatially variable parameter vectors is often feasible, especially in geoscience applications where there can be high spatial correlation (or redundancy) between parameter values in nearby grid cells. In such cases the complex spatial variability often has structure that can be concisely described with an appropriate parameterization. This involves replacement of the original N_α dimensional parameter vector α with a smaller $N_{\hat{\alpha}} \ll N_\alpha$ dimensional alternative $\hat{\alpha}$. A reduced-order parameterization has the advantage of making the characterization (or inverse) problem better posed as well as more computationally efficient.

Truncated linear transforms provide a convenient and flexible approach for developing reduced-order descriptions of time-invariant parameters. These transforms include the Karhunen-Loève (KL) expansion, which relies on basis functions derived from the prior parameter ensemble, and the discrete cosine transform (DCT), which relies on prespecified basis functions. The DCT used here works especially well for subsurface problems with channelized structures but the KL expansion could also be an attractive option in some problems [32, 10]. The DCT can be expressed as:

$$\alpha \approx \Theta\hat{\alpha}, \tag{2.3}$$

where Θ is an orthonormal N_α by $N_{\hat{\alpha}}$ matrix whose columns are a selected set of standard DCT basis vectors. These can be defined over one, two, or three-dimensional spatial grids. When $N_{\hat{\alpha}} = N_\alpha$ the DCT representation of α is exact and $\alpha = \Theta\hat{\alpha}$. When $N_{\hat{\alpha}} < N_\alpha$ the DCT representation is an approximation. Once the basis vectors are specified, (2.3) can be used to replace the elements of α with linear combinations of $\hat{\alpha}$ elements, wherever they appear. Then $\hat{\alpha}$ can be estimated from measurements as part of the characterization process.

The primary challenge involved in using a truncated DCT expansion is to decide which DCT basis functions to include in Θ. In an ensemble characterization application one option is to first derive a complete DCT representation (i.e., an $\hat{\alpha}$ vector with $N_{\hat{\alpha}} = N_\alpha$) for each replicate in the prior parameter ensemble [33]. A small set of leading basis vectors (those with the largest $\hat{\alpha}$ magnitudes) is identified for each replicate. The final truncated set of DCT basis vectors is then obtained by taking the union over all replicates of the leading replicate-specific basis vectors. This approach provides a concise but accurate overall representation of the prior parameter ensemble (especially when there is significant overlap in the dominant basis vectors of different replicates). The retained $N_{\hat{\alpha}} \ll N_\alpha$ basis vectors form the columns of Θ. Section 3.3.1 discusses methods for generating physically realistic prior ensembles.

2.2. State reduction with proper orthogonal decomposition. POD is a model reduction technique that is widely encountered in control and estimation applications because of its ability to generate accurate low-dimensional representations of very complex systems. This approach is most often used with linear (or linearized)

equations but can be extended to second-order expansions of nonlinear equations such as those given in (2.1) and (2.2). The following paragraphs briefly summarize how this extended version of POD can provide a convenient reduced-order approximation for the state variables in these equations.

In our application the POD method approximates the full-order variables x and w with the following truncated linear expansions [45]:

$$x \approx \Phi_x \hat{x}, \qquad w \approx \Phi_w \hat{w}. \tag{2.4}$$

Here Φ_x and Φ_w are orthonormal N_x by $N_{\hat{x}}$ and N_w by $N_{\hat{w}}$ matrices, respectively, with columns corresponding to the POD basis vectors for x and w. The $N_{\hat{x}}$- and $N_{\hat{w}}$-dimensional reduced-order vectors $\hat{x}$ and $\hat{w}$ serve as weighting coefficients for the basis vectors. Note that the idea of replacing the full-order variables with linear expansions over a truncated set of basis vectors is similar to the DCT approach described above. However, the POD basis vectors are derived from the full-order flow and transport model while the DCT basis vectors are derived from a set of prior geological replicates.

The POD basis vectors are connected to the full-order model through a set of "snapshot" matrices constructed from a relatively small number of full-order model "training runs." Each snapshot matrix corresponds to one of the dependent scalar variables in the original set of model partial differential equations. Each training run contributes one column of discretized dependent variable values to the associated snapshot matrix. The POD basis vectors for the various dependent variables are selected to be the leading singular vectors of the corresponding snapshot matrices. A total of $N_{\hat{x}} \ll N_x$ and $N_{\hat{w}} \ll N_w$ of these singular vectors form the columns of Φ_x and Φ_w. The POD basis vectors define a subspace of the full-order state space spanned by x and w. By construction, this subspace accounts for most of the variability (across different times, locations, parameter replicates, and controls) revealed in the training runs. The reduced-order model describes the evolution of $\hat{x}$ and $\hat{w}$ in this subspace.

The POD approach provides a flexible and general way to identify the most important features of the full-order model and capture them in the reduced-order model. As might be expected, its effectiveness depends greatly on the way that the training runs are designed. Section 4 provides more detail on the procedure used to formulate training runs and construct snapshot matrices for our subsurface transport example.

Once the POD matrices are defined the reduced-order model may be derived. To simplify notation we define augmented full- and reduced-order state vectors as follows:

$$y = \begin{bmatrix} x \\ \alpha \end{bmatrix}, \qquad \hat{y} = \begin{bmatrix} \hat{x} \\ \hat{\alpha} \end{bmatrix}.$$

In this case the full-order model may be written as

$$\frac{dy}{dt} = \begin{bmatrix} F(x, w, \alpha, u) \\ 0 \end{bmatrix} = F_a(y, w, u), \quad y(t_0) = y_0 \text{ specified}, \tag{2.5}$$

$$G_a(y, w, \alpha) = G(x, w, \alpha, u) = 0. \tag{2.6}$$

Note that (2.5) assumes that the parameter vector is time invariant. The reduced-order transformation for the augmented state is:

$$y = \Psi \hat{y} = \begin{bmatrix} \Phi_x & 0 \\ 0 & \Theta \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{\alpha} \end{bmatrix}. \tag{2.7}$$

In order to use the POD expansion we expand the nonlinear functions in (2.1) and (2.2) to second order around a nominal reduced-order parameter vector $\alpha_{nom} = \Theta\hat{\alpha}_{nom}$ and vectors x_{nom} and w_{nom} that are the solutions to (2.1) and (2.2) when $\alpha = \alpha_{nom}$. This gives the following quadratic approximation, expressed in terms of the perturbation vector $\delta y = y - y_{nom}$ and Kronecker products:

$$\frac{d\delta y}{dt} \approx F_{ay}\delta y + F_{aw}\delta w + \frac{1}{2}F_{ayy}(\delta y \otimes \delta y), \tag{2.8}$$

$$G_{ay}\delta y + G_{aw}\delta w + \frac{1}{2}G_{ayy}(\delta y \otimes \delta y) \approx 0, \tag{2.9}$$

where

$$y_{nom} = \begin{bmatrix} x_{nom} \\ \alpha_{nom} \end{bmatrix}, \quad \delta y = \begin{bmatrix} \delta x \\ \delta\alpha \end{bmatrix}.$$

The gradient (first derivative) matrices in (2.8) and (2.9) are defined as

$$F_{ay} = \frac{\partial F_a}{\partial y^T}, \quad G_{ay} = \frac{\partial G_a}{\partial y^T}, \quad F_{aw} = \frac{\partial F_a}{\partial w^T}, \quad G_{aw} = \frac{\partial G_a}{\partial w^T},$$

and the Hessian (second derivative) matrices are

$$F_{ayy} = \frac{\partial F_a}{\partial y^T \partial y^T}, \quad G_{ayy} = \frac{\partial G_a}{\partial y^T \partial y^T}.$$

In the examples considered here F_a and G_a are linear in w so their higher-order derivatives with respect to w are all zero. Note that the Kronecker product $\Gamma \otimes \Pi$ for m by n matrix Γ and an r by w matrix Π is an mr by nw matrix defined as

$$\Gamma \otimes \Pi = \begin{bmatrix} \Gamma_{11}\Pi & \cdots & \Gamma_{1n}\Pi \\ \vdots & \ddots & \vdots \\ \Gamma_{m1}\Pi & \cdots & \Gamma_{mn}\Pi \end{bmatrix}.$$

For example, $\delta y \otimes \delta y$ in (2.8) is an N_y^2 by 1 dimensional column vector composed of products of individual elements of δy. The scalar first and second derivatives that form the elements of the gradient and Hessian matrices are arranged to be consistent with the ordering implied by the Kronecker products. For example, the dimension of F_{ayy} is N_y by N_y^2.

We are most interested in cases where the expansions given in (2.8) and (2.9) are exact. These cases include certain applications of the Navier–Stokes equations or the subsurface flow and transport equations considered later in this paper. If the discretization technique is chosen appropriately the quadratic structure of the original partial differential equations can be preserved in the discretized equations given in (2.1) and (2.2).

We can derive a nonlinear reduced-order model from the quadratic approximation of the full-order model by projecting each equation onto the approximate subspace (premultiplying by Ψ^T or Φ_w^T) and invoking the properties of the Kronecker product.

$$\frac{d\delta\hat{y}}{dt} \approx \hat{F}_{\hat{y}}\delta\hat{y} + \hat{F}_{\hat{w}}\delta\hat{w} + \frac{1}{2}\hat{F}_{\hat{y}\hat{y}}(\delta\hat{y} \otimes \delta\hat{y}), \tag{2.10}$$

$$\hat{G}_{\hat{y}}\delta\hat{y} + \hat{G}_{\hat{w}}\delta\hat{w} + \frac{1}{2}\hat{G}_{\hat{y}\hat{y}}(\delta\hat{y} \otimes \delta\hat{y}) \approx 0, \tag{2.11}$$

where:

$$\hat{F}_{\hat{y}} = \Psi^T F_{ay} \Psi, \qquad \hat{F}_{\hat{w}} = \Psi^T F_{aw} \Phi_w, \qquad \hat{G}_{\hat{y}} = \Phi_w^T G_{ay} \Psi, \qquad \hat{G}_{\hat{w}} = \Phi_w^T G_{aw} \Phi_w,$$
$$\hat{F}_{\hat{y}\hat{y}} = \Psi^T F_{ayy} (\Psi \otimes \Psi), \qquad \hat{G}_{\hat{y}\hat{y}} = \Phi_w^T G_{ayy} (\Psi \otimes \Psi).$$

Equations (2.10) and (2.11) define a nonlinear reduced-order model with $N_{\hat{y}} + N_{\hat{w}}$ unknowns.

The coefficient matrices in the reduced-order model equations can be precomputed offline, before a real-time characterization and/or control process begins. In the general case these matrices all depend on the nominal variables. In the special quadratic case the gradient matrix elements are linear functions of the nominal variables and the Hessian matrix elements are constants. The number of elements in the Kronecker product matrix $\Psi \otimes \Psi$ appearing in the Hessian matrix expressions is $N_y^2 N_{\hat{y}}^2$ in the most general case. This could impose a burdensome storage requirement for large N_y. However, in the special cases of most interest here several blocks of this matrix are zero and the storage requirement is less. Also, the Hessian matrices are sparse in many of these problems. The transport example discussed in section 3 is a case in point.

2.3. Ensemble Kalman filtering with full- and reduced-order models. The ensemble Kalman filter is a sequential characterization procedure that uses randomly generated replicates to describe the distribution of uncertain states. When the filter is used in real-time applications the replicates are periodically updated with measurements so that the characterization can adapt to new information. This is especially useful in real-time control since it allows controls (e.g., well pumping rates) to respond to unanticipated changes in the system. Discrete time ensemble Kalman filtering algorithms divide naturally into a two step recursion: (1) a forecast step that uses a dynamic model to transform an ensemble of uncertain states $y^j_{t-1|t-1}$ updated at the previous measurement time $t-1$ into a new forecast ensemble $y^j_{t|t-1}$ at measurement time t, and (2) an analysis step that computes at time t an ensemble of updated replicates $y^j_{t|t}$ that incorporate new information from the noisy N_z-dimensional measurement vector z_t. Here $j = 1, \ldots, N_r$ is the replicate index and N_r is the number of replicates in the ensemble. The forecast and update steps are alternated for $t = 1, \ldots, N_t$ and the recursion is initialized with a specified ensemble $y^j_{0|0}$ of prior replicates at $t = 0$. Here we have expressed the filter steps in terms of the full-order augmented state y but they could just as well have been expressed in terms of reduced-order state $\hat{y}$.

In high-dimensional geophysical applications such as our subsurface transport example the forecast step often accounts for most of the computational time required by the ensemble Kalman filter. In such cases it may be very beneficial to work with an efficient reduced-order model such as the one derived in section 2.2. The reduced-order model needs to generate forecast ensembles that are sufficiently similar to those generated by a full-order model to give comparable overall filter performance. This can be tested by comparing full- and reduced-order characterization results. We assume that the measurements used in the filter update step can be expressed in terms of either the full- or reduced-order states as follows:

$$z_t = M_t y_t + v_t, \tag{2.12}$$

$$z_t \approx M_t \Psi \hat{y}_t + v_t, \tag{2.13}$$

where v_t is a zero-mean temporally uncorrelated N_z-dimensional random noise vector with a specified covariance matrix R_t. The specified N_z by N_y-dimensional measurement matrix M_t relates the measurements to the individual states in the vector y_t. Equation (2.13) is an approximation of (2.12) that uses (2.7) to express the measurement in terms of the reduced-order state.

The forecast step of the full- and reduced-order ensemble Kalman filters solves the appropriate augmented state equations for each replicate over the time interval between successive measurements. The dynamic forecast equations are initialized at the beginning of each interval with the current updated state:

Step 0. Initialization: At time $t = 0$ specify initial states $y^j_{0|0}$ or $\hat{y}^j_{0|0}$ for $j = 1, \ldots, N_r$.

Step 1. Forecast: For each $j = 1, \ldots, N_r$ solve the following replicate equations over the forecast interval $(t-1, t)$ to obtain the forecast $y^j_{t-1|t}$ or $\hat{y}^j_{t-1|t}$ (for $t = 1, \ldots, N_t$).

Full-order model:

$$\frac{dy^j}{dt} = F_a(y^j, w^j, u), \quad y^j_{t-1|t-1} \text{ specified}, \tag{2.14}$$

$$G_a(y^j_t, w^j, u) = 0. \tag{2.15}$$

Reduced-order model:

$$\frac{d\delta\hat{y}^j}{dt} = \hat{F}_{\hat{y}}\delta\hat{y}^j + \hat{F}_{\hat{w}}\delta\hat{w}^j + \frac{1}{2}\hat{F}_{\hat{y}\hat{y}}(\delta\hat{y}^j \otimes \delta\hat{y}^j), \quad \hat{y}^j_{t-1|t-1} \text{ specified}, \tag{2.16}$$

$$\hat{G}_{\hat{y}}\delta\hat{y}^j_t + \hat{G}_{\hat{w}}\delta\hat{w}^j_t + \frac{1}{2}\hat{G}_{\hat{y}\hat{y}}(\delta\hat{y}^j_t \otimes \delta\hat{y}^j_t) = 0, \tag{2.17}$$

$$\hat{y}^j_{t|t-1} = \Psi^T y_{nom,t} + \delta\hat{y}^j_{t|t-1}. \tag{2.18}$$

Note that the reduced-order state vector is reconstructed in (2.18) from the full-order nominal and the reduced-order perturbation, using the orthogonality of Ψ.

There are many ways to derive a set of updated replicates from the forecast replicates. It is often convenient to update the ensemble mean first and then construct each updated replicate by adding a perturbation to the mean, an approach commonly described as the ensemble square root Kalman filter [21]. Variants of this approach include the ensemble transform Kalman filter [11], the ensemble adjustment filter [5], and the local ensemble filter [42]. The ensemble square root variants proposed by Livings, Dance, and Nichols [39] and Sakov and Oke [44] have the advantage of ensuring that the sample posterior ensemble mean is equal to the actual posterior ensemble mean (i.e., that the sample mean is unbiased). In this work, we use the Livings, Dance, and Nichols [39] version of the ensemble square root Kalman update for both the full- and reduced-order filters.

Step 2. Update: For each $j = 1, \ldots, N_r$ update the forecast replicates at time t (for $t = 1, \ldots, N_t$).

Full-order filter:

$$\bar{y}_{t|t} = \bar{y}_{t|t-1} + \Lambda_t \left[z_t - M_t \bar{y}_{t|t-1} \right], \tag{2.19}$$

$$Y_{t|t} = Y_{t|t-1} T, \tag{2.20}$$

$$y^j_{t|t} = \bar{y}_{t|t} + \tilde{y}^j_{t|t}. \tag{2.21}$$

Here $\bar{y}_{t|t-1}$ and $\bar{y}_{t|t}$ are the arithmetic averages of the forecast and updated state replicates $y^j_{t|t-1}$ and $y^j_{t|t}$, while $Y_{t|t-1}$ and $Y_{t|t}$ are N_y by N_{rep} matrices with columns

equal to the ensemble perturbations $\tilde{y}^j_{t|t-1} = y^j_{t|t-1} - \bar{y}_{t|t-1}$ and $\tilde{y}^j_{t|t}$, respectively. The N_y by N_{rep}-dimensional matrix Λ_t is the Kalman gain at measurement time t, and the N_{rep} by N_{rep} matrix T is a weighting matrix that transforms forecast ensemble perturbations into updated ensemble perturbations. Both of these matrices are computed from $Y_{t|t-1}$ and the measurement error covariance matrix R_t [39].

Reduced-order filter:

$$\bar{\hat{y}}_{t|t} = \bar{\hat{y}}_{t|t-1} + \hat{\Lambda}_t \left[z_t - M_t \Psi \bar{\hat{y}}_{t|t-1} \right], \tag{2.22}$$

$$\hat{Y}_{t|t} = \hat{Y}_{t|t-1} \hat{T}, \tag{2.23}$$

$$\hat{y}^j_{t|t} = \bar{\hat{y}}_{t|t} + \tilde{\hat{y}}^j_{t|t}. \tag{2.24}$$

Here all hatted variables dependent on forecasts are computed from the reduced-order model and $\hat{\Lambda}_t$ has dimension $N_{\hat{y}}$ by N_{rep}. Note that (2.7) can be used to construct approximations of the full-order replicates from the reduced-order replicates whenever desired. This gives a complete real-time characterization of the uncertain parameters and states, using the reduced-order model to continually process new measurements.

3. Application to groundwater flow and transport.

3.1. Formulation of the full-order model. Although the approach presented above is relevant to a range of applications we focus here on a subsurface solute transport problem with special characteristics that make reduced-order modeling particularly attractive. The only source of uncertainty in this problem is a spatially variable hydraulic conductivity that yields spatially variable heads, velocities, and solute concentrations that are also uncertain. We seek to characterize all uncertain quantities from limited point observations of hydraulic head and concentration. This model characterization problem is relevant to pump-and-treat remediation of contaminant plumes, where periodically updated estimates of uncertain aquifer properties and concentrations are used to derive updated control actions (e.g., remedial well pumping rates).

The subsurface system of interest here is a thin isotropic confined groundwater aquifer with a constant layer thickness. We suppose that all measurements are from fully screened wells and that a two-dimensional vertically averaged description is appropriate. The corresponding flow and transport model can be written as

$$s \frac{\partial h(x,t)}{\partial t} = -\nabla \cdot q - \sum_{i=1}^{N_w} Q_i \delta(x - x_i), \tag{3.1}$$

$$q(x,t) = -K(x) d \nabla h(x,t), \tag{3.2}$$

$$\theta \frac{\partial c(x,t)}{\partial t} = -\nabla \cdot (qc) + \theta D \nabla^2 c - \sum_{i=1}^{N_w} Q_i c \delta(x - x_i). \tag{3.3}$$

Here $h(x,t)$ is the piezomeric head, $q(x,t)$ is the two-dimensional Darcy velocity, $c(x,t)$ is the solute concentration, s is the specific storage, $Q_i(t)$ is the volumetric flux pumped from the aquifer at well i, $\delta(x - x_i)$ is a two-dimensional spatial Dirac delta function, $K(x)$ is the isotropic hydraulic conductivity, d is the constant aquifer thickness, and θ is the porosity. To simplify the transport equation we assume the dispersion coefficient D is constant and isotropic in all directions. The head and concentration equations are accompanied by initial and boundary conditions that are discussed in more detail below. Note that the only nonlinearities in these equations are

products of the unknown dependent variables h, q_x, q_y, and c, the unknown parameter K, and their respective spatial gradients. That is, these equations are bilinear.

Spatial discretization of (3.1), (3.2), and (3.3) with traditional numerical schemes generally yields a set of coupled nonlinear ordinary differential equations such as (2.1) and (2.2), with the form of the nonlinearity depending on the discretization procedure. For example, a traditional finite difference technique that uses the harmonic or geometric means of grid cell conductivities to compute velocities replaces the bilinear relationships of the original partial differential equations with more complex nonlinearities that are less convenient for model reduction. Here we seek a numerical scheme that preserves, as much as possible, the bilinearity of the original equation.

If we discretize the flow equations of (3.1) and (3.2), with a mixed finite element approach [23, 24], a uniform partition of the computational domain, and lowest-order Raviart–Thomas subspaces on rectangles [26] the result is the following set of bilinear spatially discretized equations:

$$s\frac{\partial h(t)}{\partial t} - Bq(t) = R_1(t), \tag{3.4}$$

$$B^T h(t) + A(\eta)q(t) = R_2(t), \tag{3.5}$$

where h is an N_h dimensional column vector containing the unknown grid cell pressure heads and

$$q = \begin{bmatrix} q_x \\ q_y \end{bmatrix}$$

is an $N_q = N_{qx} + N_{qy}$-dimensional column vector containing x and y components of the unknown grid cell Darcy velocities. The components of the N_h by N_q matrix B are constant, while the components of the N_q by N_q matrix $A(\eta)$ depend linearly on the components of the N_α dimensional hydraulic resistivity vector η, which is defined by $\eta_i = 1/K_i$ for each grid cell i. In this application the resistivity is the only parameter included in the generic parameter vector α. Although our model and Kalman filtering algorithms are formulated in terms of resistivity we report results in terms of the more familiar log hydraulic conductivity. The N_h- and N_q-dimensional right-hand-side vectors R_1 and R_2 depend on the boundary conditions, including the flow conditions imposed at the pumping wells. Equations (3.4) and (3.5) are a set of two coupled equations in the unknowns h and q, for any given value of η.

If we discretize the transport equation in (3.3) on the same grid as the mixed finite element flow grid using an upwind finite difference method [49] the result can be written as

$$\theta\frac{\partial c(t)}{\partial t} + E(q)c(t) + Lc(t) = R_3(t), \tag{3.6}$$

where c is an N_c-dimensional vector containing unknown grid cell solute concentrations. The matrix $E(q)$ depends nonlinearly on the velocity vector q when upwinding is used to mitigate numerical dispersion. The matrix L is constant if the dispersion coefficients are assumed to be constants that do not depend on velocity. If the solute can only exit through the wells and the flow rates at the wells are prescribed the pumping term in (3.3) can be incorporated into the matrix E and $R_3 = 0$.

The discretized equation system given in (3.4), (3.5), and (3.6) can be further simplified if we suppose that the pumping rates are constant between measurement or control times and that the specific storage is small enough for the flow system to reach an approximate steady state between these times. In this case we only need to solve a steady state version of (3.4) for constant h and q over each interval.

The POD and DCT transformations relating the full- and reduced-order model states may be arranged to show the basis function matrices associated with each of the four dependent variable vectors and the parameter vector:

$$y = \begin{bmatrix} h \\ q_x \\ q_y \\ c \\ \eta \end{bmatrix} \approx \begin{bmatrix} \Phi_h & 0 & 0 & 0 & 0 \\ 0 & \Phi_{qx} & 0 & 0 & 0 \\ 0 & 0 & \Phi_{qx} & 0 & 0 \\ 0 & 0 & 0 & \Phi_c & 0 \\ 0 & 0 & 0 & 0 & \Theta \end{bmatrix} \begin{bmatrix} \hat{h} \\ \hat{q}_x \\ \hat{q}_y \\ \hat{c} \\ \hat{\eta} \end{bmatrix} = \Psi \hat{y}. \tag{3.7}$$

In this formulation the prognostic variables in the vector $\hat{x}$ are $\hat{h}$ and $\hat{c}$, the diagnostic variables in $\hat{w}$ are $\hat{q}_x$ and $\hat{q}_y$, and the parameter vector $\hat{\alpha}$ is $\hat{\eta}$.

The gradients and Hessians that appear in the quadratic full-order model approximation given in (2.10) and (2.11) may be computed directly from (3.4), (3.5), and (3.6). In this particular example, the second-order expansion is exact and many of the Hessian terms are zero. Although the expansion is exact the reduced-order model derived from it is still approximate since the basis function set is truncated (i.e., the full-order state space is projected onto a lower-dimensional subspace).

3.2. Derivation of the reduced-order model. A quadratic reduced-order transport model may be constructed by substituting the basis function expansions into the full-order model equations, following the general procedure outlined in section 2.2. The parameter basis function matrix Θ is constructed from standard DCT basis functions selected in accordance with the procedure summarized in section 2.1. Each of the remaining POD basis function matrices is constructed from a separate snapshot matrix compiled from training runs of the full-order model.

A POD basis derived from a small number of training runs normally works well only for conditions similar to those simulated in these runs. When the geological parameters and possibly the controls of the full-order model are uncertain the snapshot matrices should adequately reflect model behavior over the full range of possible conditions. In an ensemble characterization procedure, this range is described by the prior parameter (e.g., resistivity) ensemble. In order to keep the POD derivation manageable, we include training runs from a small subset of randomly selected parameter replicates rather than the entire ensemble. This approach works well if the snapshot matrices also include sensitivities of the states to the full-order model parameters [14]. Parameter sensitivities improve the quality of the reduced-order model by providing information on the effects of moderate deviations from the particular ensemble members used to construct the snapshot matrices. Our numerical tests have shown that such sensitivities are more helpful for flow than transport. If the model is to be used to test different candidate control strategies, as is required in a real-time remediation procedure, it is important that the snapshots also include a representative set of possible well pumping histories [37]. However, in this paper we only consider a single prespecified pumping history.

The parameter sensitivity derivatives included in the snapshot matrix for our example may be readily computed when the flow problem is piecewise steady state, as discussed above. In this case the full-order head and velocity sensitivities are obtained from

$$-B\frac{\partial q}{\partial \eta_i} = 0, \tag{3.8}$$

$$B^T \frac{\partial h}{\partial \eta_i} + A(\eta)\frac{\partial q}{\partial \eta_i} = -u_i, \tag{3.9}$$

where u_i is the ith column of the matrix $U(q)$ defined as

$$U(q) = \frac{\partial A(\eta)q}{\partial \eta^T}. \tag{3.10}$$

Note that $U(q)$ does not depend on η since the elements of $A(\eta)$ depend linearly on the elements of η. Equations (3.8) and (3.9) can be solved iteratively for the desired sensitivities $\partial h/\partial \eta_i$ and $\partial q/\partial \eta_i$.

In our subsurface transport example we identify four different snapshot matrices S_h, S_{qx}, S_{qy}, and S_c. The columns of S_h are composed of selected head solutions as well as selected head sensitivity derivatives $\partial h/\partial \eta_i$ and the columns of S_{qx} and S_{qy} are composed of selected velocity solutions as well as selected velocity sensitivity derivatives $\partial q/\partial \eta_i$. The columns of S_c only include selected concentration solutions. The basis functions that form the columns of the Φ_h, Φ_{qx}, Φ_{qy}, and Φ_c matrices are taken to be the leading orthonormal singular vectors of the corresponding snapshot matrices. The dimension of each reduced-order state vector is determined by a corresponding singular vector energy threshold, which is illustrated here for the head $\hat{h}$:

$$\varepsilon_h = \frac{\sum_{j=1}^{N_{\hat{h}}} \lambda_{hj}^2}{\sum_{j=1}^{N_h} \lambda_{hj}^2}.$$

Here ε_h is the head energy threshold and λ_{hj} is the jth singular value of the head snapshot matrix Φ_h, where the index j increases in order of decreasing singular values. The reduced-order head dimension $N_{\hat{h}}$ is the smallest value that gives $\varepsilon_h \approx 95 - 99\%$. A similar approach is used for the other dependent variable vectors.

We can now substitute the POD basis function approximations and gradient and Hessian matrices derived from (3.4), (3.5), and (3.6) into (2.10) and (2.11) to obtain a bilinear reduced-order subsurface transport model:

$$s\frac{\partial(\delta\hat{h})}{\partial t} - \hat{B}\delta\hat{q} = 0, \tag{3.11}$$

$$\hat{B}^T\delta\hat{h} + \hat{A}\delta\hat{q} + \hat{U}\delta\hat{\eta} + \hat{\eta}\,(\delta\hat{\eta} \otimes \delta\hat{q}) = 0, \tag{3.12}$$

$$\theta\frac{\partial\delta\hat{c}}{\partial t} + \hat{E}(\hat{q})\delta\hat{c} + \hat{L}\delta\hat{c} = 0, \tag{3.13}$$

$$\hat{h} = \Phi^T h(\eta_{nom}) + \delta\hat{h}, \quad \hat{q} = \Phi^T q(\eta_{nom}) + \delta\hat{q}, \quad \hat{c} = \delta\hat{c}, \tag{3.14}$$

where

$$\hat{B} = \Phi_h^T B \Phi_h, \quad \hat{A} = \Phi_q^T A(\eta_{nom})\Phi_q, \quad \hat{U} = \Phi_q^T U[q(\eta_{nom})]\Phi_h,$$

$$\hat{H} = \Phi_q^T H(\Phi_\eta \otimes \Phi_q), \quad H = \frac{\partial^2[A(\eta)q]}{\partial\eta^T\partial q^T},$$

$$\Phi_q = \begin{bmatrix} \Phi_{qx} & 0 \\ 0 & \Phi_{qy} \end{bmatrix},$$

$$\hat{E}(\hat{q}) = \Phi_c^T E(\Phi_q\hat{q})\Phi_c, \quad \hat{L} = \Phi_c^T L \Phi_c.$$

Since the quadratic expansion used to obtain the flow equations is exact, these equations hold for any nominal parameter value η_{nom}. Moreover, the full-order concentration equation is linear in the single unknown c since q is known (it is independently determined by the flow equation, which does not depend on c). Since the full-order

equation is linear in c the reduced-order concentration equation is linear in $\delta\hat{c}$. For numerical convenience, we select a typical nonzero nominal resistivity value for η_{nom} but define the concentration perturbation to be around the nominal $c_{nom} = 0$ so $\hat{c} = \delta\hat{c}$. Note that the control u enters the reduced-order flow equations through the functions $h(\eta_{nom})$ and $q(\eta_{nom})$ and it enters the concentration equation through $\hat{E}(\hat{q})$.

All of the coefficient matrices in the reduced-order model can be precomputed offline, before a real-time characterization and/or control process begins, except for $\hat{E}(\hat{q})$. The matrix $\hat{E}(\hat{q})$ (and therefore $E(q)$) must be recomputed online for every replicate whenever $\hat{q}$ changes. If the quasi-steady-state assumption made earlier holds then $\hat{q}$ only changes when the well pumping rate is modified in response to a control decision. If such changes are relatively infrequent the computational burden of evaluating $E(q)$ and $\hat{E}(\hat{q})$ online is relatively small.

Since $A(\eta)q$ is bilinear in the elements of q and η, the matrix $\hat{H}$ is constant and sparse. The matrices $\hat{B}$ and $\hat{L}$ are also constant while $\hat{A}$ depends linearly on the elements of η_{nom} and $\hat{U}$ depends linearly on the elements of $q(\eta_{nom})$ (which can be computed offline). The dimension of the Kronecker product $\Phi_\eta \otimes \Phi_q$ appearing in the $\hat{H}$ computation is $N_q N_{\hat{q}} N_\eta N_{\hat{\eta}}$. Storage of this product can require a significant amount of memory since Φ_η and Φ_q are not sparse.

Equations (3.11), (3.12), (3.13), and (3.14) constitute a nonlinear reduced-order transport model expressed in terms of the parameter $\hat{\eta}$ and the states $\hat{h}$, $\hat{q}$, and $\hat{c}$. If this model adequately reproduces the response of the original full-order model over a range of parameter realizations and controls it can greatly improve the efficiency of the site characterization process.

3.3. Implementation issues.

3.3.1. Prior ensemble. Ensemble filters designed for geological property characterization, such as our example, need to use prior ensembles that are physically realistic and appropriate for the site under investigation. One way to approach this goal is to generate replicates from training images that portray, in an approximate and qualitative way, the geological facies (i.e., high versus low conductivity zones) believed to apply at the study site. These training images can be derived from site-specific field observations (e.g., a seismic survey) or can rely on observations from other sites thought to have similar geological origins. In any case, if we have training images we can generate an ensemble of random prior replicates with a multipoint geostatistical procedure [18]. Multipoint generation procedures use training image pattern matching techniques rather than two-point statistics such as spatial covariances or variograms to identify structures that need to be captured in the replicates. The patterns are identified in various ways from the training images, depending on the specific method used.

We generate our prior full-order resistivity fields with the multipoint geostatistical algorithm FILTERSIM [48]. This replicate generator has the ability to identify and count the occurrence of patterns in a template scanned across the training image. It reproduces within-template structures in the replicates it generates. In our example FILTERSIM only generates categorical replicates that delineate boundaries between permeable and less permeable facies. When resistivity values are assigned to each of the two facies and the resulting binary resistivity images are approximated with a truncated DCT expansion the resulting replicates are no longer binary but exhibit gradual transitions between higher and lower resistivity areas. Examples are shown (in terms of log conductivity) in the figures presented in section 4.

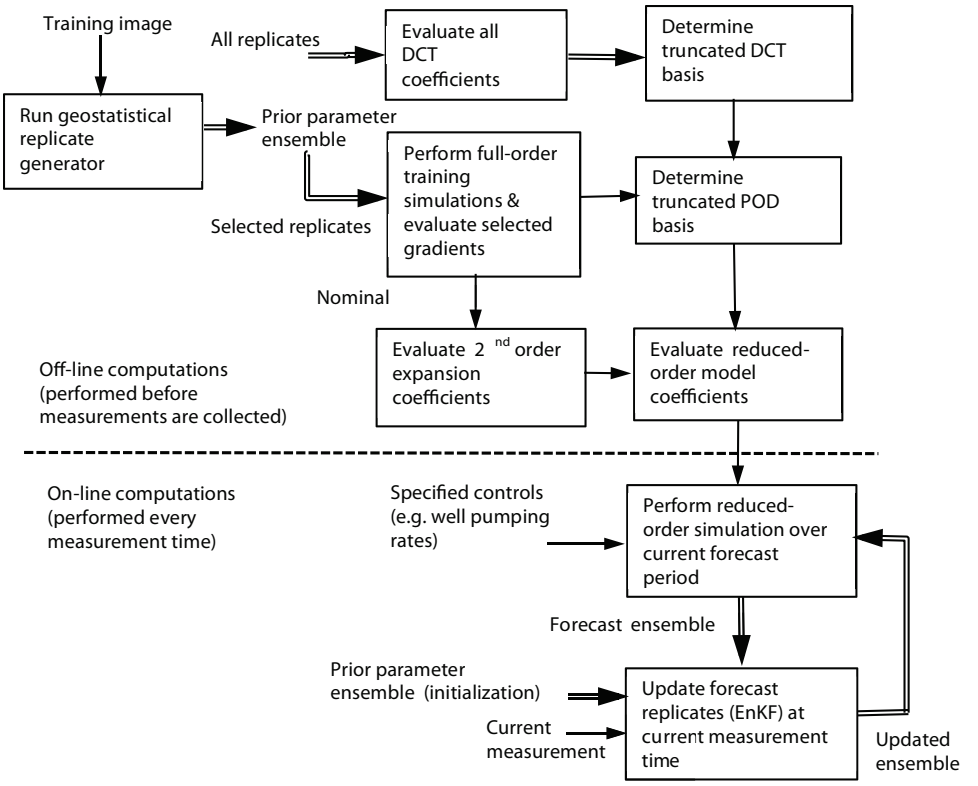

FIG. 2. *Offline and online computations required in the ensemble real-time characterization procedure.*

3.3.2. Offline versus online computation. To some extent reduced-order modeling is a trade-off between the offline computation required to derive a robust reduced-order model that works well over a range of conditions versus the online computation required to use the model in a real-time setting, as in an ensemble Kalman filter. Figure 2 distinguishes the offline operations needed to derive the reduced-order model from the online computations performed by the ensemble Kalman filter. The amount of offline computation that is acceptable is clearly problem dependent but it is generally true that more offline time is required to derive more accurate reduced-order models. The decrease in computational effort provided by the model reduction procedure should be large enough to justify the offline computation.

The primary factor that increases offline computation is the number of training runs included in the snapshots, since each training run requires a simulation of the full-order model. The likelihood of obtaining a robust reduced-order approximation that provides significant online savings increases with more training runs, especially if these runs are carefully selected. Online computational effort increases with the dimensionality of the reduced-order model and with the number of replicates used in the ensemble Kalman filter. Our experience suggests that the savings provided by model reduction is greater when the original full-order model is larger. This reflects the observation that the number of variables needed to capture the dominant aspects of a spatially discretized model's dynamic behavior tends to be a smaller fraction of the total model size as the model is more finely discretized and its dimension increases.

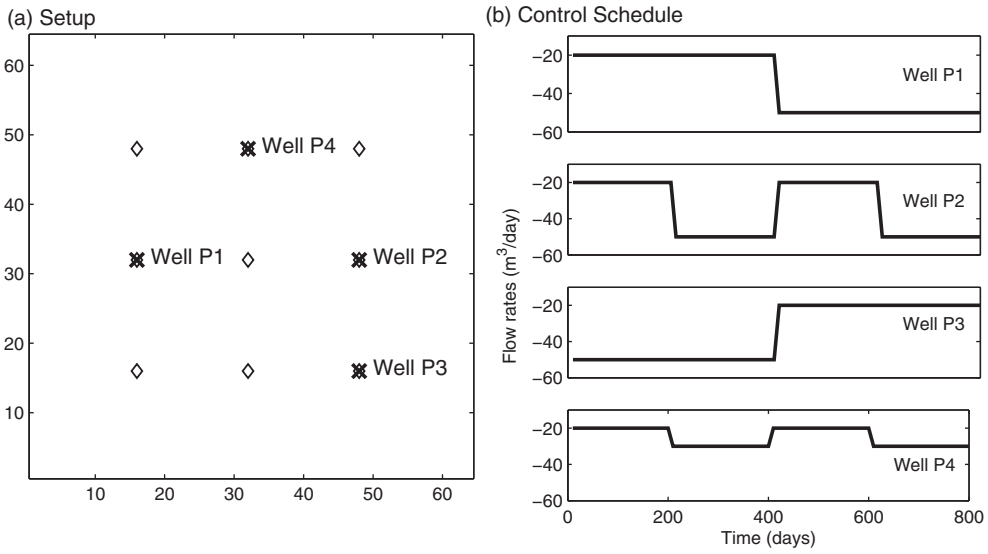

FIG. 3. *Experimental setup:* (a) *simulation domain with* 9 *measurement locations (black diamonds) and* 4 *pumping wells (black crosses) at* $P1$, $P2$, $P3$, *and* $P4$; (b) *predetermined pumping rates at the four wells.*

When this occurs reduced-order modeling may provide dramatic computational improvements with only minor decreases in performance.

4. Numerical experiment.

4.1. Experimental setup. The model reduction concepts described above can be tested by comparing the performance of ensemble Kalman filters derived from full- and reduced-order models. The perspective taken here is different than in some other model reduction studies since we are primarily interested in the reduced-order model's ability to reproduce an ensemble of characterization results derived from real-time measurements rather than its ability to match a single deterministic prediction for one replicate. A time series of Kalman filter ensembles continually benefits from new measurement information that compensates for model simplifications by adjusting the reduced-order states. Although measurement feedback cannot transform a poor model into an accurate one it can help keep a reasonably accurate approximate model on track. The numerical experiment described here examines this process.

Our experiment considers subsurface flow and transport in the vertically averaged rectangular domain shown in Figure 3(a). This domain is assumed to be a confined aquifer that is pumped at the four wells indicated on the figure. The only source of uncertainty is the hydraulic conductivity, which varies over space in a structured manner. The general characteristics of the uncertain conductivity field are conveyed by a specified training image that is used to generate a prior ensemble of possible geological facies replicates. Point measurements of head and concentration are used to characterize the resistivity and, by implication, the log-conductivity fields. The $640 \times 640 \times 1$-m computational domain is discretized into $64 \times 64 \times 1$ uniform computational grids of $10 \times 10 \times 1$-m cells. The spatially uniform porosity has a known value of 0.2. The specific storage is assumed to be approximately zero relative to the time scales of interest so that the flow system responds immediately to changes

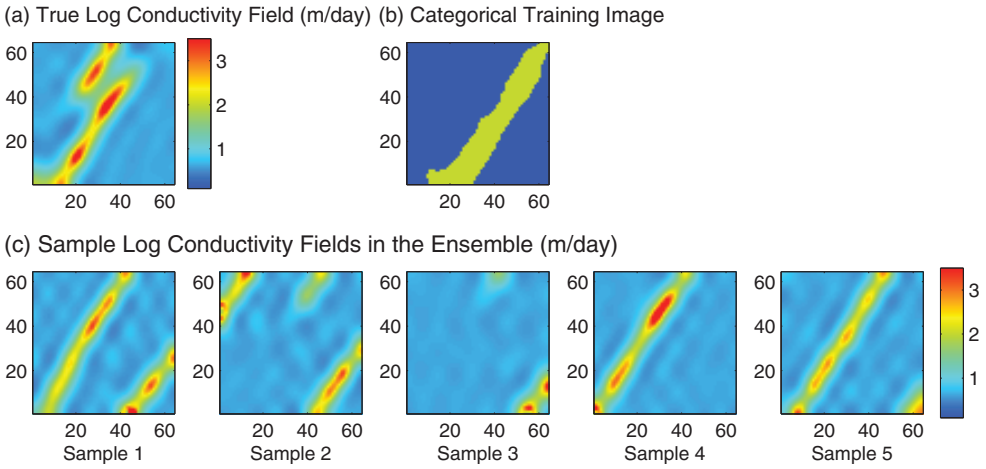

FIG. 4. *Samples of hydraulic conductivity:* (a) *the true conductivity field;* (b) *the categorical training image used for generating the prior ensemble;* (c) *sample prior replicates after DCT truncation.*

in well pumping rates. The dispersion coefficient is assumed to be an isotropic flow-independent constant over the whole domain.

The flow boundary conditions are zero water flux on the east and west sides and constant head boundaries of 30 m and 0 m on the south and north sides, except during solute injection (see below). These conditions impose a regional flow from south to north with local fluctuations induced by well pumping and log hydraulic conductivity variations. The transport boundary conditions are zero solute flux on the east and west sides and zero solute concentration on the south and north sides, except during injection at the solute source located in the middle of the southern boundary. Solute is injected at a constant rate of $5\,m^3$/day and constant concentration of 50 mg/L for the first 400 days. The initial concentration is zero everywhere in the domain.

The simulation period is 800 days, divided into 4 control intervals of 200 days each. The pumping rates at the four wells follow the specified control schedule shown in Figure 3(b). In a real-time control setting the pumping rates would be adjusted in response to measurements so would not be known in advance. Here we do not adjust pumping rates because we want to focus on the characterization performance of the reduced-order model. We plan to examine reduced-order ensemble real-time control in future work. Since the dispersion tensor is isotropic and constant and the flow system is at steady state over each control interval, the transport equation is linear, with coefficients that change instantly when the pumping rates change. This simplification facilitates the reduced-order modeling process.

In this experiment we suppose that heads and concentrations are measured at the nine locations indicated by the black diamonds in Figure 3(a). These include the four pumping wells. The measurements are taken at the end of the four control steps so there are four EnKF updates during the simulation. We perform a virtual characterization study by generating synthetic noisy measurements from the full-order model, using a "true" hydraulic conductivity derived by applying FILTERSIM and a DCT approximation to a facies-based training image, as described in section 3.3.1. The true log-conductivity replicate is shown in Figure 4(a). The training image used to generate the true log-conductivity field and also all of the prior Kalman filter replicates is shown in Figure 4(b). This image is characterized by a single high conductivity

diagonal channel with a kink in the middle of the domain. The diagonal feature appears, in various degrees, in most of the prior replicates. Note that the true log conductivity is not included in the prior ensemble.

Figure 4(c) shows five typical prior log-conductivity replicates from two 50 member ensembles generated from the training image. The training image provides valuable prior information about the existence of a channel (or channels) trending in the southwest to northeast direction. One might argue that the qualitative facies information conveyed in the training image is likely to be available in many field studies. Even with this information, there remains considerable uncertainty about the location and geometry of possible channel structures. This is the uncertainty that we seek to reduce with head and concentration measurements.

In our experiment the full-order and reduced-order prior conductivity ensembles are obtained from the same truncated DCT representation of the FILTERSIM facies replicates. This representation uses only $N_\eta = 30$ DCT coefficients, about 3% of the number of pixels. Our use of a DCT approximation in the full-order characterization partly reflects the observation that some form of geological property parameterization is commonly adopted in most practical subsurface modeling studies. In addition, a characterization performance comparison of a full-order dynamic model with highly correlated pixel-based conductivities versus a reduced-order dynamic model with a compressed DCT representation of conductivity penalizes the full-order approach by forcing it to estimate many more parameters from the same limited number of measurements. We have confirmed this in experiments not shown here. So we use a DCT parameterization in both the full- and reduced-order models because it provides a fairer assessment of the benefits achieved with our POD model reduction approach.

In the reduced-order model derivation we use 10 resistivity replicates to make 10 corresponding training runs for the POD approximation of model states. Snapshots of head and velocity are taken at each of the four measurement times and concentration snapshots are taken at each of 80 time steps (every 10 days). Head and velocity sensitivities with respect to all the elements of the full-order resistivity vector are also computed from (3.8) and (3.9) for each of the 10 replicates and each of the four measurement times. For each of the three head and velocity snapshot matrices this gives 40 columns for the states and 16,384 for the sensitivities. The concentration snapshot matrix has 800 columns for states only.

The singular value decompositions of the snapshot matrices are truncated at the number of basis function terms that account for 99%, 93%, 93%, and 99% of the energy in the head, x velocity, y velocity, and concentration, respectively. This gives reduced-order dimensions of 69, 113, 84, and 184, for these four states, as compared to the original full-order dimension of about 4096 for each state. We determined how many modes to keep in the truncated approximation by looking at the complete spectrum of the eigenvalues (ordered by decreasing energy). Normally there is a sharp drop in eigenvalue energy at some point where truncation has a minimal effect on model accuracy. However, this point may be different for different states. The reduced-order flow approximation is more sensitive to high frequency modes than the transport approximation. Therefore, we truncate at 99% of the energy for the flow equations and at 93% of the energy for the transport equation. This choice gives good accuracy as well as significant computational savings.

The DCT reduced-order descriptions of the parameters and the POD reduced-order description of the states together provide the information needed to construct the complete reduced-order model summarized in (3.11) through (3.14).

4.2. Experimental results. We begin our discussion of experimental results by examining the spatial structure of the updated ensemble mean log-conductivity and concentration fields. The ensemble means provide convenient point estimates of the true state, although they are typically smoother since they are obtained by averaging over replicates that have variability comparable to the truth. Figure 5(b) shows the updated ensemble mean log-conductivity fields obtained from the full- and reduced-order filters at the four measurement times. These can be compared to the true field shown in Figure 5(a). It is clear that the channelized structures of the log-conductivity field are gradually recovered for both the full- and reduced-order cases. The full-order filter does somewhat better delineating both the shape and interior values of the dominant diagonal channel. Both filters have difficulty capturing the kink in the northern portion of the channel. This reflects the fact that this feature is not present in most of the prior replicates and that there are not enough measurements in the vicinity to identify such detail.

Figure 5(c) shows snapshots of the true concentration field simulated from the full-order model with the true resistivity field, plotted at the four measurement times. The four crosses with different colors indicate the locations of the four pumping wells. The true plume is pulled to the four pumping wells as it travels from south to north under the influence of the regional head gradient. Concentration measurements at each of the nine sampling points become informative as soon as the plume breaks through. Figure 5(d) shows the corresponding ensemble mean concentrations obtained from the full-order and reduced-order cases.

For this particular example, the reduced-order filter seems to do better at capturing the plume shape, especially after the final update on day 800. It is possible that this reflects the fact that the full-order filter must estimate 40 times more dynamic states than the reduced-order filter from the same number of measurements. Similar results have been reported by [3]. However, it is also possible that the result observed in Figure 5(d) is an anomaly that might only occur rarely in a more extensive set of experiments with many different true conductivity fields and well pumping schedules. For present purposes it is sufficient to note that the reduced-order filter captures true plume behavior at least as well as the full-order model with much less effort. In fact, the effort required to run many different full-order (rather than reduced-order) filters is the primary factor limiting the scope of our computational tests.

The effects of model reduction on the properties of the state ensembles can be conveniently investigated by comparing the full- and reduced-order time series for head and concentration at the four pumping wells. Figure 6 examines the head series, which responds almost instantly to pumping changes at the four measurement/control times The four columns of plots show, from left to right, full- and reduced-order model ensembles with no measurement updates and full- and reduced-order Kalman filter ensembles that are updated at the four measurement times. Individual ensemble replicates are shown in gray, the true head is shown in dash-dot lines, and the ensemble average is shown in dashed lines. The reduced-order and full-order model ensembles are almost (but not exactly) the same, indicating that the reduced-order model gives an excellent match to the full-order forecast. This reflects, in part, the fact that the same control history is used in the reduced-order model runs and the full-order training simulations. This may not be possible in practice since the control history may not be known in advance. Then we can expect more difference between the full- and reduced-order model forecasts.

It is apparent from Figure 6 that the updated ensembles are generally narrower than the no-update ensembles, reflecting the additional information provided by the

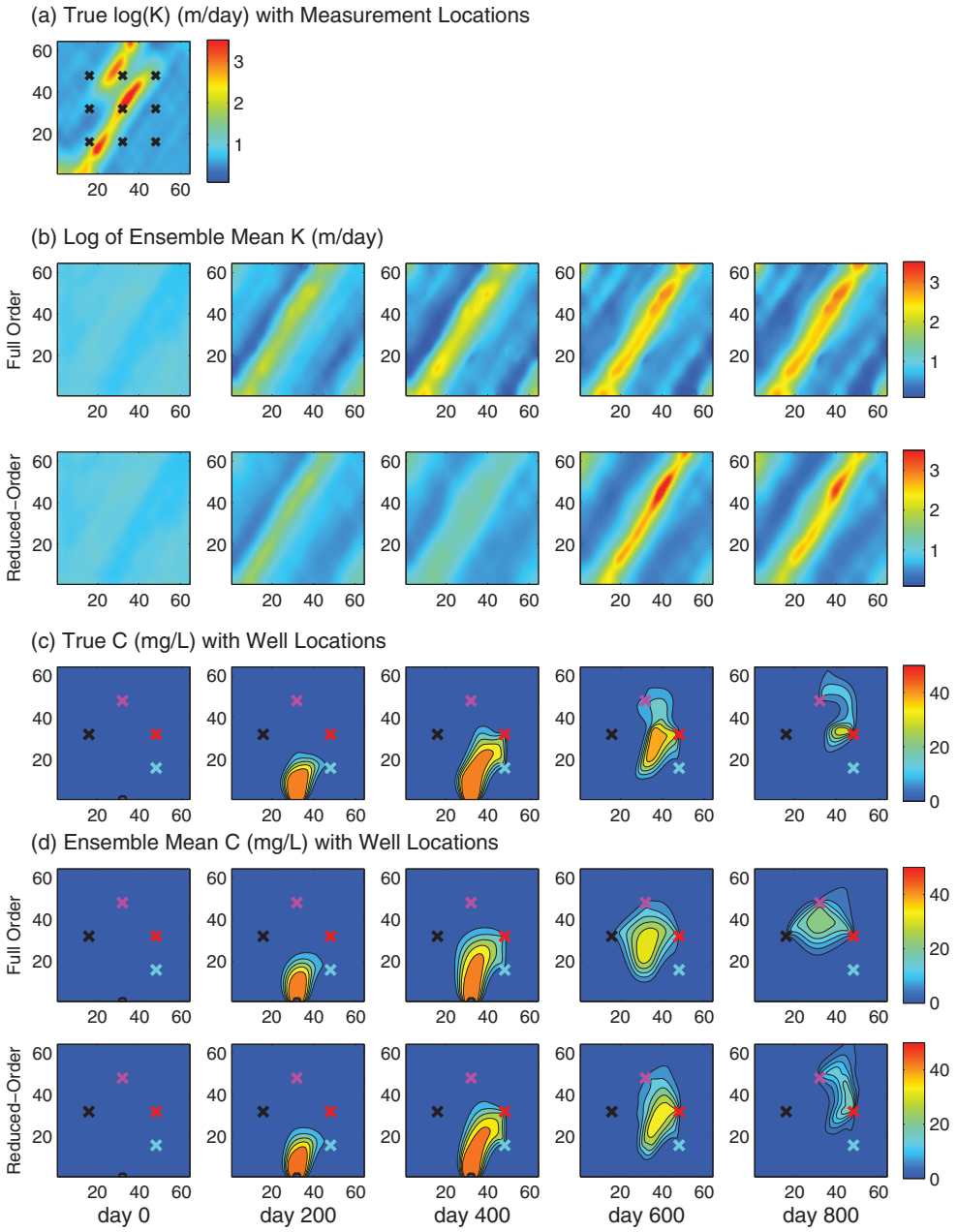

FIG. 5. *Comparison of true and updated log conductivity and concentration:* (a) *true log-conductivity field;* (b) *ensemble mean log conductivity using the full- and reduced-order filters with measurement updating;* (c) *true concentration with well locations indicated by black (P1), red (P2), cyan (P3), and magenta (P4) crosses;* (d) *ensemble mean concentration using the full- and reduced-order filters with measurement updating, well locations indicated by crosses.*

measurements. The reduced-order model gives results that are generally similar to the full-order model, except at well P4 (the well furthest from the source), where the reduced-order ensemble is noticeably different and generally wider than the full-order ensemble. At the other wells the updated ensembles generally narrow and become

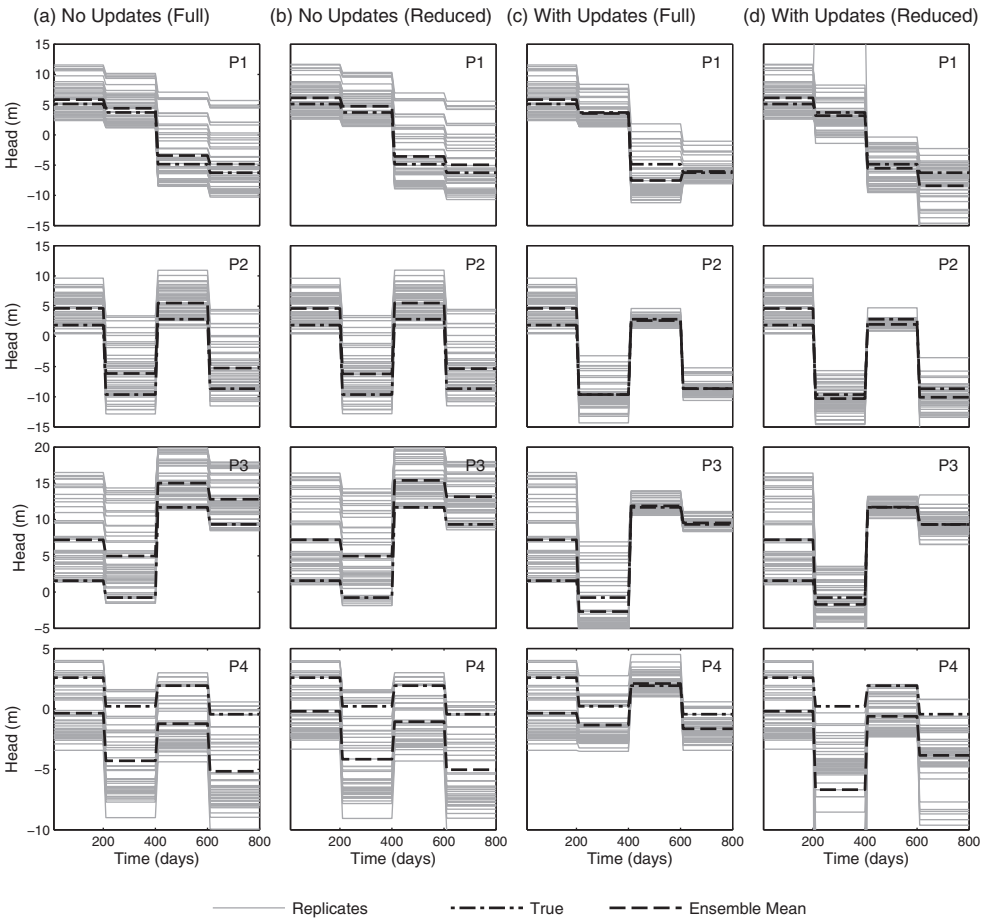

FIG. 6. *Head time series at the four well locations for replicates (gray lines), the true conductivity field (black dashed-dot lines), and the ensemble mean (black dashed lines):* (a) *predictions using the full-order model without measurement updating;* (b) *predictions using the reduced-order model without measurement updating;* (c) *predictions using the full-order filter with the measurement updating at the end of each control step;* (d) *predictions using the reduced-order filter with measurement updating at the end of each control step.*

more centered on the true value as time progresses and more measurements become available. This is illustrated by the convergence of the updated ensemble mean to the true value, except at P4.

The concentration time series shown in Figure 7 give another perspective on the effects of model reduction. Here again, the full- and reduced-order ensembles are nearly the same when there are no measurement updates. In all cases the concentrations at each well are small until the plume arrives. The no-update replicates vary significantly from one another in both arrival times and magnitudes. Many replicates reach well P1 even though the true plume does not arrive there during the simulation period. The full- and reduced-order ensembles both initially diverge between measurements but they narrow dramatically when they are updated with new measurements. In wells P2 and P3 the updated ensemble means capture the recession of the plume reasonably well and even in P4, where there is more variability at later times, the updated ensemble means eventually converge towards the true concentration.

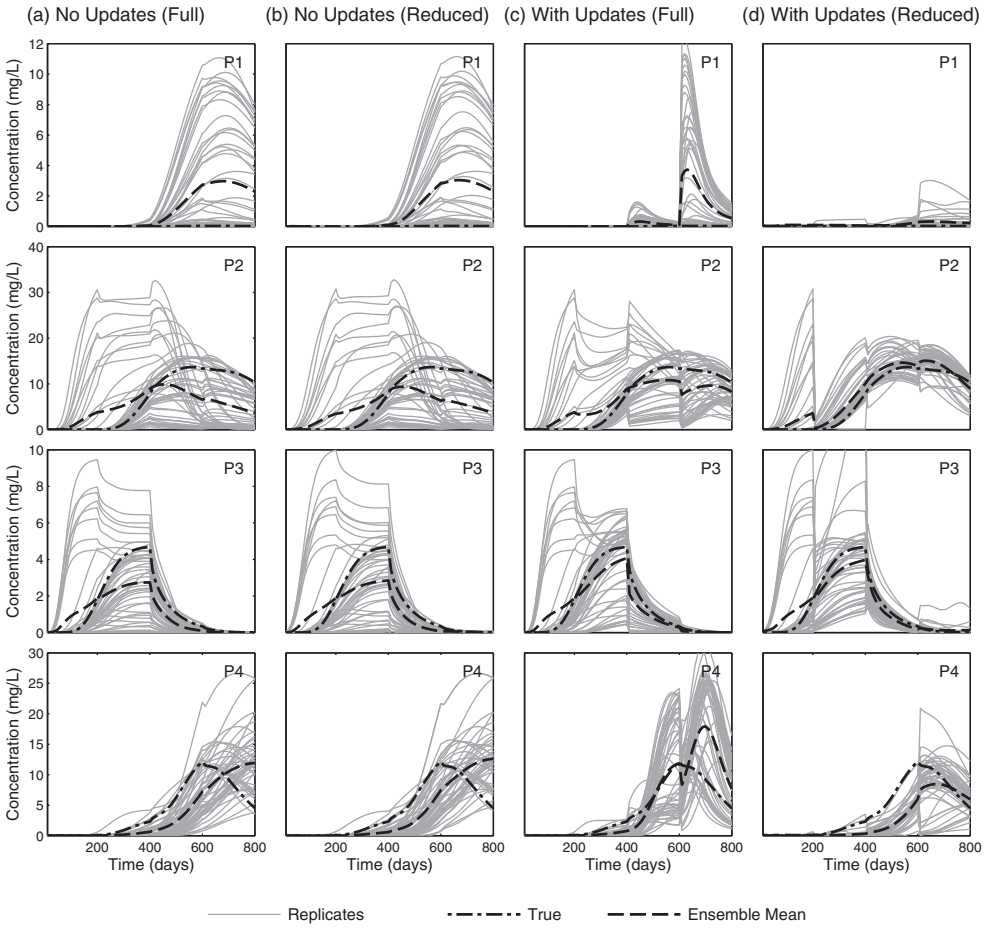

FIG. 7. *Concentration time series at the four well locations for replicates (gray lines), the true conductivity field (black dashed-dot lines), and the ensemble mean (black dashed lines):* (a) *predictions from the full-order model without measurement updating;* (b) *predictions from the reduced-order model without measurement updating;* (c) *predictions from the full-order filter with the measurement updating at the end of each control step;* (d) *predictions from the reduced-order filter with measurement updating at the end of each control step. Note that the true solute plume does not reach well* $P1$.

Overall, the reduced-order updated ensemble tends to be closer to the true concentration and more responsive to new measurements than the full-order alternative. This result is consistent with our earlier observation that the full-order filtering problem is more challenging because many more states must be estimated from the same limited number of measurements. The concentration results summarized here suggest that a reduced-order model may perform as well as or even better than a full-order model when incorporated into a characterization procedure with continual measurement feedback.

Figure 8 illustrates spatial variability among different conductivity and concentration replicates, with and without measurement updating, at the final measurement time at day 800. Figures 8(a) and 8(b) show the true log-conductivity field and solute plume. Figure 8(c) shows five typical prior log-conductivity replicates while Figure 8(d) shows the corresponding five updated log-conductivity replicates obtained

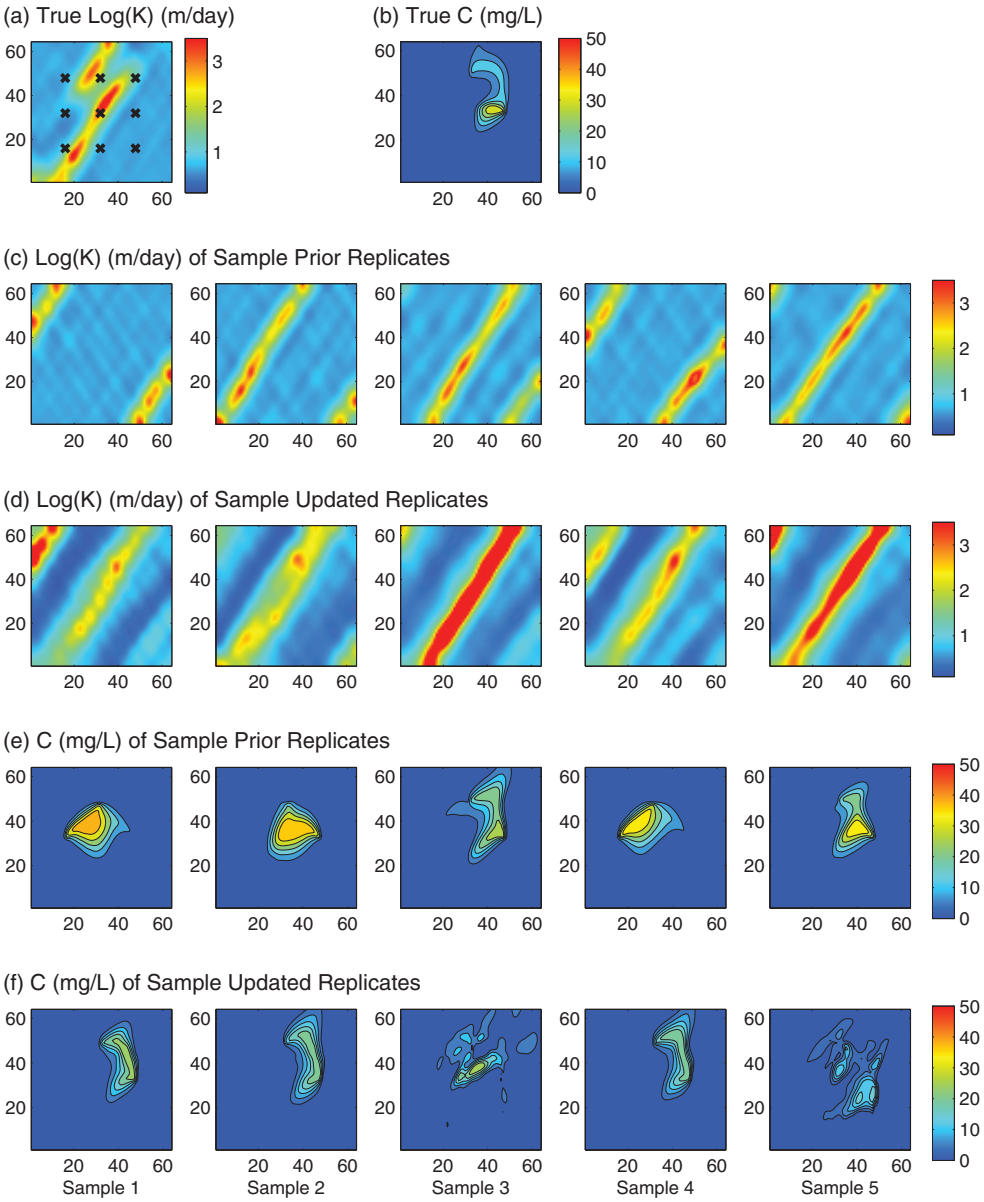

FIG. 8. *Representative conductivity and concentration replicates:* (a) *true log-conductivity field;* (b) *true concentration field, on day* 800*;* (c) *five prior log-conductivity replicates;* (d) *corresponding five log-conductivity replicates from reduced-order filter with measurement updating, on day* 800*;* (e) *corresponding five concentration replicates from reduced-order model without measurement updating, on day* 800*;* (f) *corresponding five concentration replicates from reduced-order filter with measurement updating, on day* 800.

from the reduced-order filter at day 800. Although there is considerable variability among the updated replicates they tend to more consistently identify the location of the dominant conductive channel than the prior replicates. Figure 8(e) shows the solute plumes obtained at day 800 from a no-update reduced-order forecast based on the five prior log-conductivity replicates from Figure 8(c). These concentration plots

give an indication of our prior uncertainty about the plume location and shape and are the spatial counterparts to the second column in Figure 7. Figure 8(f) shows the updated solute plumes obtained from the reduced-order filter at day 800. The individual replicates in this figure vary less in size, shape, and orientation than the no-update forecast, illustrating the value of the measurements. The results from a full-order model forecast and filter update (not shown here) are qualitatively similar.

5. Discussion and conclusions. Reduced-order modeling offers some valuable benefits for ensemble characterization studies. Since the reduced-order model requires much less computation for each replicate forecast it is possible to either include many more replicates for a given amount of total computational effort or to greatly reduce effort for a given ensemble size. The opportunity to increase ensemble size makes it possible to decrease the dimension of the estimation null space or even to use a full-rank filter. This enhances the filter's ability to respond to informative measurements. It also reduces the effect of sampling errors that lead to suboptimal updates.

In the subsurface characterization example considered here the reduced-order filter performs as well as, if not better than, a full-order filter, especially in terms of its ability to identify the location and shape of a solute plume. It is especially satisfying to see that the reduced-order filter generally duplicates the ensemble spread of the full-order filter, indicating that the reduced-order model is able to provide a reasonable description of uncertainty as well as average conditions. The reduced-order model displays this capability both with and without measurement updates.

More generally, the performance of a reduced-order filter can be expected to depend on a number of problem-specific factors. First, the training snapshots used to construct the model's reduced-order basis vectors should be representative of the conditions the model is likely to encounter during real-time operations. Although this is difficult to determine in advance, the reduced-order modeling process is more likely to be successful if the training runs include a representative sample of different parameter replicates as well as snapshots of sensitivities of states with respect to parameters. In our application, sensitivities of heads and velocities with respect to resistivities are especially helpful in improving the performance of the reduced-order model. In fact, we were able to obtain better results with only a few prior training run replicates if we use sensitivities than we obtain with many more training runs without sensitivities. Overall, our method is very efficient not only during the online simulation stage, but also for the training runs.

While sensitivities certainly improve performance, in our application the accuracy of the reduced-order model still depends on the representativeness of the prior replicates used in the training runs. The reduced-order model may closely approximate the full-order model for a particular prior ensemble but if the range of conditions portrayed in this prior does not include the true system both models will give poor estimation results. Further discussion of this point, with some examples, is provided in [33]. Care should be taken in any Bayesian estimation application to ensure that the prior ensemble is reasonable and sufficiently diverse to capture real-world conditions.

A second factor affecting reduced-order model performance is the extent and quality of the measurements used for ensemble updating. If the reduced-order model's predictions are somewhat off during the forecast step but the measurements are able to bring the state replicates closer to the true value at the update the model's forecast errors will have less long-term impact. This may be why our reduced-model does better, in some cases, at characterizing the true solute plume, even though its log-

conductivity characterization is somewhat less accurate than the one provided by the full-order model.

A third factor affecting the performance of a model reduction based on a POD approach is the accuracy of the approximations introduced in the POD projection. In our example a truncated POD expansion is substituted into a set of second-order (bilinear) state equations that exactly reproduce the nonlinear equations of the full-order model. This approach yields a low dimensional but still nonlinear reduced-order model. An exact expansion may not be feasible in other applications with more complex nonlinearities. In such cases, reduced-order modeling may still be feasible, using methods such as the discrete empirical interpolation method [15], but these alternatives may require more computation.

The approach described here derives the reduced-order model once, in a set of offline computations, and does not modify the model in realtime. This contributes to the substantial computational savings achieved during real-time operation. The offline computational requirements can be extensive, especially when second-order coefficient matrices must be evaluated and stored. So the approach is more attractive in real-time applications where it is desirable to transfer effort from online to offline computations. The use of parameter sensitivities and realistic replicates for the training runs significantly reduces the offline computational effort.

Our approach benefits from a distinction between reducing the dimension of the parameter vector (through a truncated DCT approximation) and reducing the dimensions of the dynamic variable vectors (through truncated POD approximations). This distinction reflects differences in the nature of the training information and performance standards in each case. Parameter reduction should capture dominant geological features, as reflected in a prior parameter ensemble generated from an appropriate training image. Dynamic variable reduction should capture essential dynamical behavior, as reflected in an ensemble of training runs obtained by running the full-order model. So it is reasonable to approach the two tasks somewhat differently.

One of the reasons to develop reduced-order models for real-time characterization is to facilitate other real-time activities. Real-time control can use reduced-order models in the search algorithms that identify optimal controls as well as in the characterization procedure that supports the control derivation. Reduced-order models can also be useful in real-time monitoring and targeted observation investigations where a decrease in computational effort can facilitate searches for the most informative observing strategy. Finally, reduced-order models offer benefits for screening and risk-assessment studies that rely on large numbers of simulations. These applications go beyond the characterization problem considered here but they all build on the concepts of reduced-order ensemble modeling and filtering. There appears to be much to be gained from further research on reduced-order modeling in a real-time context, especially where ensemble methods are appropriate.

Acknowledgments. We thank Ruben Juanes, Youssef M. Marzouk, and Karen Willcox for their comments.

REFERENCES

[1] S. I. Aanonsen, G. Naevdal, D. S. Oliver, A. C. Reynolds, and B. Valles, *The ensemble Kalman filter in reservoir engineering-a review*, SPE J., 14 (2009), pp. 393–412.

[2] J. I. Allen, M. Eknes, and G. Evensen, *An ensemble Kalman filter with a complex marine ecosystem model: Hindcasting phytoplankton in the Cretan sea*, Ann. Geophys., 21 (2003), pp. 399–411.

[3] M. U. Altaf, M. El Gharamti, I. Hoteit, and A. W. Heemink, *A reduced adjoint approach to variational data assimilation*, Comput. Methods Appl. Mech. Engrg., 254 (2013), pp. 1–13.
[4] D. Amsallem and C. Farhat, *Interpolation method for adapting reduced-order models and application to aeroelasticity*, AIAA Journal, 46 (2008), pp. 1803–1813.
[5] J. L. Anderson, *An ensemble adjustment Kalman filter for data assimilation*, Monthly Weather Rev., 129 (2001), pp. 2884–2903.
[6] J. L. Anderson and S. L. Anderson, *A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts*, Monthly Weather Rev., 127 (1999), pp. 2741–2758.
[7] A. C. Antoulas, *An overview of approximation methods for large-scale dynamical systems*, Ann. Rev. Control, 29 (2005), pp. 181–190.
[8] Z. J. Bai, *Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems*, Appl. Numer. Math., 43 (2002), pp. 9–44.
[9] M. Bergmann, L. Cordier, and J.-P. Brancher, *Optimal rotary control of the cylinder wake using proper orthogonal decomposition reduced-order model*, Phys. Fluids, 17 (2005), 097101.
[10] E. W. Bhark, B. Jafarpour, and A. Datta-Gupt, *A generalized grid connectivity–based parameterization for subsurface flow model calibration*, Water Res. Res., 47 (2011), W06517.
[11] C. H. Bishop, B. J. Etherton, and S. J. Majumdar, *Adaptive sampling with the ensemble transform Kalman filter. Part* I: *Theoretical aspects*, Monthly Weather Rev., 129 (2001), pp. 420–436.
[12] T. Bui-Thanh, K. Willcox, and O. Ghattas, *Model reduction for large-scale systems with high-dimensional parametric input space*, SIAM J. Sci. Comput., 30 (2007), pp. 3270–3288.
[13] T. Bui-Thanh, K. Willcox, and O. Ghattas, *Parametric reduced-order models for probabilistic analysis of unsteady aerodynamic applications*, AIAA J., 46 (2008), pp. 2520–2529.
[14] K. Carlberg and C. Farhat, *A low-cost, goal-oriented 'compact proper orthogonal decomposition' basis for model reduction of static systems*, Internat. J. Numer. Methods Engrg., 86 (2010), pp. 381–402.
[15] S. Chaturantabut and D. C. Sorensen, *Nonlinear model reduction via discrete empirical interpolation*, SIAM J. Sci. Comput., 32 (2010), pp. 2737–2764.
[16] L. Daniel, O. C. Siong, L. S. Chay, K. H. Lee, and J. White, *A multiparameter moment-matching model-reduction approach for generating geometrically parameterized interconnect performance models*, IEEE Trans. Comput.-Aided Design Integrated Circuits Systems, 23 (2004), pp. 678–693.
[17] J. Degroote, J. Vierendeels, and K. Willcox, *Interpolation among reduced-order matrices to obtain parameterized models for design, optimization and probabilistic analysis*, Internat. J. Numer. Methods Fluids, 63 (2010), pp. 207–230.
[18] C. V. Deutsch, *Geostatistical reservoir modeling*, Appl. Geostatist. Ser., Oxford University Press, Oxford, 2002.
[19] M. Eknes and G. Evensen, *An ensemble Kalman filter with a* 1*-d marine ecosystem model*, J. Marine Systems, 36 (2002), pp. 75–100.
[20] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics*, J. Geophys. Res. Oceans, 99 (1994), pp. 10143–10162.
[21] G. Evensen, *Sampling strategies and square root analysis schemes for the EnKF*, Ocean Dynam., 54 (2004), pp. 539–560.
[22] G. Evensen, *The ensemble Kalman filter for combined state and parameter estimation Monte Carlo techniques for data assimilation in large systems*, IEEE Control Systems Mag., 29 (2009), pp. 83–104.
[23] R. E. Ewing, *Finite element methods for nonlinear flows in porous media*, Comput. Methods Appl. Mech. Engrg., 51 (1985), pp. 421–439.
[24] R. E. Ewing and R. F. Heinemann, *Mixed finite element approximation of phase velocities in compositional reservoir simulation*, Comput. Methods Appl. Mech. Engrg., 47 (1984), pp. 161–175.
[25] D. Galbally, K. Fidkowski, K. Willcox, and O. Ghattas, *Non-linear model reduction for uncertainty quantification in large-scale inverse problems*, Internat. J. Numer. Methods Engrg., 81 (2010), pp. 1581–1608.
[26] I. Galligani and E. Magenes, *Mathematical Aspects of Finite Element Methods*, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977.
[27] K. Gallivan, E. Grimme, and P. Vandooren, *Pade approximation of large-scale dynamic systems with Lanczos methods*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Vol. 1, IEEE Control Systems Society, Piscataway, NJ, 1994, pp. 443–448.

[28] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation*, IEEE Proc. F Radar Signal Process., 140 (1993), pp. 107–113.
[29] E. Grimme, *Krylov Projection Methods for Model Reduction*, Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, 1997.
[30] V. Haugen, G. Naevdal, L. J. Natvik, G. Evensen, A. M. Berg, and K. M. Flornes, *History matching using the ensemble Kalman filter on a North Sea field case*, SPE J., 13 (2008), pp. 382–391.
[31] M. A. Iglesias, K. J. H. Law, and A. M. Stuart, *Ensemble Kalman methods for inverse problems*, Inverse Problems, 29 (2013), 045001.
[32] B. Jafarpour, *Oil reservoir characterization using ensemble data assimilation*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2008.
[33] B. Jafarpour and D. B. McLaughlin, *Estimating channelized-reservoir permeabilities with the ensemble Kalman filter: The importance of ensemble design*, SPE J., 14 (2009), pp. 374–388.
[34] M. P. Kaleta, R. G. Hanea, A. W. Heemink, and J.-D. Jansen, *Model-reduced gradient-based history matching*, Comput. Geosci., 15 (2011), pp. 135–153.
[35] K. Kunisch and S. Volkwein, *Control of the Burgers equation by a reduced-order approach using proper orthogonal decomposition*, J. Optim. Theory Appl., 102 (1999), pp. 345–371.
[36] C. Lieberman, K. Willcox, and O. Ghattas, *Parameter and state model reduction for large-scale statistical inverse problems*, SIAM J. Sci. Comput., 32 (2010), pp. 2523–2542.
[37] B. Lin and D. McLaughlin, *Real-time ensemble control with reduced-order modeling*, SIAM J. Sci. Comput., submitted.
[38] N. Liu and D. S. Oliver, *Critical evaluation of the ensemble Kalman filter on history matching of geologic facies*, SPE Reservoir Eval. Engrg., 8 (2005), pp. 470–477.
[39] D. M. Livings, S. L. Dance, and N. K. Nichols, *Unbiased ensemble square root filters*, Phys. D, 237 (2008), pp. 1021–1028.
[40] X. Ma and G. E. Karniadakis, *A low-dimensional model for simulating three-dimensional cylinder flow*, J. Fluid Mech., 458 (2002), pp. 181–190.
[41] B. C. Moore, *Principal component analysis in linear-systems - controllability, observability, and model-reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–32.
[42] E. Ott, B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. J. Patil, and J. A. Yorke, *A local ensemble Kalman filter for atmospheric data assimilation*, Tellus Ser. Dynam. Meteorol. Oceanogr., 56 (2004), pp. 415–428.
[43] C. W. Rowley, *Model reduction for fluids, using balanced proper orthogonal decomposition*, Internat. J. Bifur. Chaos, 15 (2005), pp. 997–1013.
[44] P. Sakov and P. R. Oke, *A deterministic formulation of the ensemble Kalman filter: An alternative to ensemble square root filters*, Tellus Ser. A Dynam. Meteorol. Oceanogr., 60 (2008), pp. 361–371.
[45] L. Sirovich, *Turbulence and the dynamics of coherent structures.* I: *Coherent structures.* II: *Symmetries and transformations.* III: *Dynamics and scaling*, Quart. Appl. Math., 45 (1987), pp. 561–590.
[46] J. A. Skjervheim, G. Evensen, S. I. Aanonsen, B. O. Ruud, and T. A. Johansen, *Incorporating 4d seismic data in reservoir simulation models using ensemble Kalman filter*, SPE J., 12 (2007), pp. 282–292.
[47] J. S. Whitaker and T. M. Hamill, *Ensemble data assimilation without perturbed observations*, Monthly Weather Rev., 130 (2002), pp. 1913–1924.
[48] T. Zhang, *Filter-based training pattern classification for spatial pattern simulation*, Ph.D. thesis, Stanford University, Stanford, CA, 2006.
[49] C. Zheng and G. D. Bennett, *Applied Contaminant Transport Modeling*, 2nd ed., Wiley-Interscience, New York, 2002.