INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS
Int. J. Numer. Meth. Fluids 2013; **71**:135–150
Published online 18 January 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/nmf). DOI: 10.1002/fld.3650

Hessian-based model reduction: large-scale inversion and prediction

C. Lieberman[1,*,†], K. Fidkowski[2], K. Willcox[1] and B. van Bloemen Waanders[3]

[1]*Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2]*Aerospace Engineering, University of Michigan, Ann Arbor, MI 48109, USA*
[3]*Optimization and Uncertainty Estimation Department, Sandia National Laboratories, Albuquerque, NM 87185, USA*

SUMMARY

Hessian-based model reduction was previously proposed as an approach in deriving reduced models for the solution of large-scale linear inverse problems by targeting accuracy in observation outputs. A control-theoretic view of Hessian-based model reduction that hinges on the equality between the Hessian and the transient observability gramian of the underlying linear system is presented. The model reduction strategy is applied to a large-scale ($\mathcal{O}(10^6)$ degrees of freedom) three-dimensional contaminant transport problem in an urban environment, an application that requires real-time computation. In addition to the inversion accuracy, the ability of reduced models of varying dimension to make predictions of the contaminant evolution beyond the time horizon of observations is studied. Results indicate that the reduced models have a factor $\mathcal{O}(1000)$ speedup in computing time for the same level of accuracy. Copyright © 2012 John Wiley & Sons, Ltd.

Received 1 July 2011; Revised 3 October 2011; Accepted 4 January 2012

KEY WORDS: model reduction; optimization; inverse problems

1. INTRODUCTION

An inverse problem is posed to estimate unknown parameters of a system given indirect observations. These parameter estimates can then be used in simulations to make predictions of output quantities of interest. For large-scale problems, when the parameter is high-dimensional and simulations are computationally expensive, state-of-the-art algorithms on parallel clusters may not be sufficient in scenarios requiring real-time computations. One example is response to a contaminant release, where the goal is to use sparse sensor data to determine an estimate of the initial release. Once the contaminant's initial release is ascertained, prediction of the evolution of the contaminant beyond the current observations informs evacuation and decontamination procedures. The timescale of these events necessitates real-time computations.

Instead of solving the inversion and prediction problems by using a high-fidelity model that resolves the system dynamics everywhere in space and time, we propose to exploit the low-dimensionality of the input–output maps of interest by generating reduced models. These reduced models are constructed to be accurate for the outputs of interest and can be simulated in seconds or minutes on a laptop computer in the field. This dramatic speedup in computing time makes possible in-the-field analysis of inversion and subsequent prediction to inform real-time decision making.

*Correspondence to: C. Lieberman, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Room 37-431, Cambridge, MA 02139, USA.

[†]E-mail: celieber@mit.edu

Consider the linear time-invariant unforced discrete-time forward model

$$x(k) = Ax(k-1), \quad x(0) = x_0, \quad y(k) = Cx(k), \quad k = 1, 2, \ldots, \tag{1}$$

where $x(k) \in \mathbb{R}^n$ and $y(k) \in \mathbb{R}^q$ are the system state and output at time step k, respectively. Our system has n state dimensions, q outputs of interest, and we consider K time steps. The initial condition is x_0. The model dynamics and spatiotemporal discretization determine the full-rank constant matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, whereas the outputs of interest define the constant matrix $\mathbf{C} \in \mathbb{R}^{q\times n}$. Our goal will be to utilize output observations from time steps $k = 1, 2, \ldots, K$ to infer the initial condition x_0 in order to make predictions for $k > K$. We assume that (1) is stable. Dynamical models of the form (1) describe many physical processes, such as contaminant transport, diffusion processes, and linearized fluid dynamics.

Projection-based model reduction seeks a low-dimensional subspace of the state space, range$(V) \subset \mathbb{R}^n$, defined by a matrix $\mathbf{V} \in \mathbb{R}^{n\times m}$, $m \ll n$, such that we approximate $x \approx Vx_r$, where $x_r \in \mathbb{R}^m$ is the reduced state. We obtain the reduced model that describes the evolution of x_r by projecting the governing equations onto a subspace, range$(W) \subset \mathbb{R}^n$, where $W \in \mathbb{R}^{n\times m}$. For the system (1), the reduced model is

$$x_r(k) = \mathbf{A}_r x_r(k-1), \quad x_r(0) = \mathbf{W}^T x_0, \quad y_r(k) = C_r x_r(k), \quad k = 1, 2, \ldots, K, \tag{2}$$

where the projection matrices $\mathbf{W}$ and $\mathbf{V}$ are defined such that $\mathbf{W}^T\mathbf{V} = I$. The reduced system matrices are $\mathbf{A}_r = \mathbf{W}^T\mathbf{A}\mathbf{V} \in \mathbb{R}^{m\times m}$ and $\mathbf{C}_r = \mathbf{C}\mathbf{V} \in \mathbb{R}^{q\times m}$.

Several methods exist for determining the bases W and V, including the proper orthogonal decomposition (POD) [1–3], approximate balanced truncation [4–8], and Krylov-subspace methods [9–11]. In this work, we use the POD with Galerkin projection (i.e., orthonormal $W = V$) combined with a Hessian-based model reduction approach that provides a basis tailored specifically for initial value problems [12]. In the Hessian-based model reduction approach, a set of seed initial conditions is computed as the dominant eigenvectors of the Hessian matrix, which encapsulates the input–output map. These eigenvectors can be computed efficiently using iterative matrix-free approaches [13]. State trajectories are computed for each seed initial condition and then combined via the POD method of snapshots to compute the basis V [3].

Projection-based model reduction methods have been employed to a limited extent in the inverse problem setting. Examples include a radiative source statistical inversion problem [14], the history matching context in computational geoscience [15], and identification of reaction parameters [16]. The control-theoretic concepts that underly model reduction have also been used to define problem-specific regularization for identification and control of a reservoir [17].

In this paper, we show why the Hessian-based model reduction method demonstrated for the forward problem in [12] is appropriate in the inverse setting. We demonstrate the approach for the solution of a large-scale initial condition inversion problem: convective–diffusive transport of a contaminant in an urban canyon. In [12], the Hessian-based model reduction was motivated by the solution to inverse problems but was applied only to forward predictions. In this work, we utilize the same Hessian-based model reduction algorithm and also apply the reduced model to the convection–diffusion equation modeling a contaminant transport problem. Unique to this work is the solution of the inverse problem using the reduced model, discussion of the regularization effects in the inverse problem for both full and reduced models, the presentation of a control-theoretic motivation for Hessian-based model reduction including a discussion of a compelling equality between the Hessian of the inverse problem and the transient observability gramian of the discrete-time forward model, the large-scale application whose full model has more than one million degrees of freedom, and the analysis of prediction accuracy on the basis of the inverted initial condition.

The paper is organized as follows. The state-space model, inverse problem, and a review of Hessian-based model reduction are presented in Section 2. In Section 3, we describe the control-theoretic motivation for the use of Hessian-based model reduction in the solution of large-scale inverse problems. We establish the equality between the transient observability gramian and the Hessian of the inverse problem. In Section 4, we describe the implementation of the model reduction

DOI: 10.1002/fld

algorithm for a large-scale application in convective–diffusive transport. The implementation includes parallel discontinuous Galerkin (DG) finite elements and large-scale eigenvalue solvers. In Section 5, we apply the algorithm to an inverse and a prediction problem in the convective–diffusive transport of a contaminant in an urban setting. We demonstrate the efficiency of the reduced models and analyze the convergence of the approximation with the size of the reduced model. Finally, we state conclusions in Section 6.

2. BACKGROUND

To develop the theory for the inverse problem and our model reduction approach, we rewrite the forward model (1) in block matrix form,

$$\mathbf{A}\mathbf{x} = \mathbf{F}x_0, \qquad \mathbf{y} = \mathbf{C}\mathbf{x}, \tag{3}$$

where $\mathbf{x} = [x(1) \cdots x(K)]^T$, $\mathbf{y} = [y(1) \cdots y(K)]^T$, and the matrices $\mathbf{A} \in \mathbb{R}^{nK \times nK}$ and $\mathbf{C} \in \mathbb{R}^{qK \times nK}$ are given by

$$\mathbf{A} = \begin{bmatrix} I & 0 & 0 & 0 & \cdots & 0 \\ -A & I & 0 & 0 & & \vdots \\ 0 & -A & I & 0 & \ddots & \\ 0 & 0 & -A & I & \ddots & \\ & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & & 0 & 0 & -A & I \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} C & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & C & 0 & & & \vdots \\ \vdots & 0 & C & 0 & & \\ & & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & & & 0 & 0 & C \end{bmatrix}, \tag{4}$$

and $\mathbf{F} = \begin{bmatrix} A^T & 0 & \cdots & 0 \end{bmatrix}^T$.

2.1. Inverse problem

We consider the inverse problem that seeks to reconstruct the initial condition x_0, given sparse measurements of the state from distributed sensors over a specified time horizon. The inversion task can be written as an optimization problem,

$$\min_{x_0} \mathcal{J} = \frac{1}{2}(\mathbf{y} - \mathbf{y}^*)^T(\mathbf{y} - \mathbf{y}^*) + \mathcal{R}(x_0), \tag{5}$$

$$\text{where} \quad \mathbf{A}\mathbf{x} = \mathbf{F}x_0,$$
$$\mathbf{y} = \mathbf{C}\mathbf{x},$$

where the vector $\mathbf{y}^*$ contains the space-time sensor readings, and $\mathbf{y}$ contains the corresponding model predictions of concentration over time at those sensor locations. The first term in the objective function $\mathcal{J}$ seeks to minimize the misfit between the predicted and the sensed data. The second term $\mathcal{R}(x_0)$ is a regularization term that is required because, in general, the entire system state cannot be uniquely identified from sparse observations [18]. This means that the inverse problem is ill-posed: many initial conditions x_0 may lead to identical observations $\mathbf{y}^*$. Regularization is a means of selecting a single initial condition estimate out of many candidates that are consistent with the observations, by incorporating prior knowledge about the problem. This can be achieved by penalizing states that do not exhibit certain desirable characteristics; for example, regularization may be used to preferentially select only smooth initial conditions by increasing the objective function cost of states with sharp peaks and troughs.

DOI: 10.1002/fld

With the elimination of the state and output equations, the inverse problem can be written as an unconstrained optimization problem,

$$\min_{x_0} \mathcal{J} = \frac{1}{2}(\mathbf{C}\mathbf{A}^{-1}\mathbf{F}x_0 - \mathbf{y}^*)^T(\mathbf{C}\mathbf{A}^{-1}\mathbf{F}x_0 - \mathbf{y}^*) + \mathcal{R}(x_0). \tag{6}$$

Our theoretical analysis focuses on the unregularized problem—just the data misfit portion of the objective function. The Hessian matrix, $\mathbf{H} \in \mathbb{R}^{n\times n}$, of this unregularized optimization problem is given by

$$\mathbf{H} = (\mathbf{C}\mathbf{A}^{-1}\mathbf{F})^T(\mathbf{C}\mathbf{A}^{-1}\mathbf{F}). \tag{7}$$

The regularization term will reappear in Section 5 when we solve the inverse problem.

The ill-posedness of the unregularized inverse problem manifests itself through singularity of the Hessian. The null space of the Hessian describes the initial conditions about which the observed data provide no information—initial conditions that are unobservable with respect to the available outputs $\mathbf{y}$. In the following, we show how this ill-posedness may be exploited to construct a reduced model by defining a projection subspace on the basis of the eigenvectors of the Hessian with the largest nonzero eigenvalues. We also describe the connection between the Hessian and the transient observability gramian of the forward model and show how a reduced model, constructed using dominant eigenvectors of the Hessian, represents states in the observable subspace of the forward model.

2.2. *Hessian-based model reduction*

We define v_i to be the ith eigenvector of the Hessian, $Hv_i = \lambda_i v_i$, with corresponding eigenvalue λ_i. We order the eigenvalues of the Hessian, $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n \geqslant 0$. The Hessian-based model reduction approach from [12] then proceeds as described in Algorithm 1.

Algorithm 1
Hessian-based model reduction

1. For the Hessian matrix H as defined in (7), find the r eigenvectors $v_1, v_2, \ldots, v_r$ with largest eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_r \geqslant \lambda_{r+1} \geqslant \ldots \geqslant \lambda_n \geqslant 0$. Select r on the basis of the decay of the Hessian eigenvalues.
2. For $i = 1, \ldots, r$, set $x_0 = v_i$ and compute the corresponding solution $\mathbf{x}^i$ by using (3).
3. Form the reduced basis as the span of the snapshots $x^i(k)$, $i = 1, 2, \ldots, r$, $k = 0, 1, \ldots, K$.

A common method for approximately solving the inverse problem is to form a low-rank approximation of the Hessian by using its dominant eigenvectors [18]. This method takes advantage of the compact eigenvalue spectrum of the Hessian [19]. Our model reduction approach also uses the dominant eigenvectors of the Hessian to construct the reduced basis; however, it differs in that we first derive a reduced model for the forward problem (3) and then solve the inverse problem by using the reduced model. Approximating directly the Hessian will likely yield a lower complexity online inversion; however, reducing the forward problem adds considerable flexibility because our reduced model may be used for the forward prediction of state evolution beyond the observation horizon for a much lower computational cost than that of the full model. Although we do not consider it here, it is, at least in principle, possible to treat variation in velocity fields or sensor locations with a nonlinear model reduction approach (e.g., empirical interpolation [20, 21] and missing point estimation [22]). Time-varying and space-varying velocity fields would affect the system matrix $\mathbf{A}$, whereas time-varying sensor locations would affect the observation matrix $\mathbf{C}$.

3. CONTROL-THEORETIC VIEW OF HESSIAN-BASED MODEL REDUCTION

In this section, we present a control-theoretic motivation for the use of Hessian-based model reduction in the solution of large-scale inverse problems. This is one example of the many connections that

DOI: 10.1002/fld

exist between concepts found in the controls community and those found in the PDE-constrained optimization and inverse problem communities.

We first recall the transient observability gramian, its properties, and how it can be used to define the output energy associated with a given initial condition. We then show that the unregularized Hessian of the inverse problem is equal to the transient observability gramian of the forward model (1), implying that the Hessian-based model reduction targets initial condition modes (and their state evolutions, via POD) that are informed by the sensor data.

3.1. Transient observability gramian

Consider the forward model (1). For a given initial condition $x(0) = v$, we define the total output energy of the forward model

$$\sum_{k=1}^{K} \|y(k)\|_2^2 = \sum_{k=1}^{K} \|CA^k v\|_2^2 = v^T \left(\sum_{k=1}^{K} {A^k}^T C^T C A^k \right) v = v^T G_K v, \quad K \geqslant 1,$$

where $G_K = \sum_{k=1}^{K} {A^k}^T C^T C A^k$ is the transient observability gramian [23].

The transient observability gramian is the finite-time analog of the infinite horizon observability gramian $G = \sum_{k=1}^{\infty} {A^k}^T C^T C A^k$. The transient gramian has several useful properties:

Recurrence relation The transient observability gramian is given by the recurrence relation

$$G_K = A^T C^T C A + A^T G_{K-1} A, \qquad G_0 = 0, \qquad K \geqslant 1.$$

Proof

Let $K \geqslant 1$. We have $G_K = A^T C^T C A + \sum_{k=1}^{K-1} {A^{k+1}}^T C^T C A^{k+1} = A^T C^T C A + A^T \sum_{k=1}^{K-1} {A^k}^T C^T C A^k A = A^T C^T C A + A^T G_{K-1} A$. □

Symmetric positive semidefiniteness The transient observability gramian is trivially symmetric, and its positive semidefiniteness results directly from the nonnegativity of the norm.

Gramian limit If the forward model (1) is stable, then $G = \lim_{K \to \infty} G_K$ exists and is the observability gramian of (1), satisfying the Lyapunov equation $G = A^T C^T C A + A^T G A$ [23].

3.2. Hessian equality

Before we establish the equality between the transient observability gramian and the Hessian of the inverse problem, we first reason that the eigenvectors of G_K corresponding to nonzero eigenvalues form an appropriate basis for the solution of the inverse problem. Let $Z_K = \text{span}(v_1, v_2, \ldots, v_r)$, where $G_K v_i = \lambda_i v_i$ with $\lambda_i > 0$ for $i = 1, 2, \ldots, r$ and $\lambda_{r+1} = 0$. Let x_0 be an arbitrary initial condition for the forward model (1). Let $x_0^{Z_K}$ be the component of x_0 in Z_K and $x_0^{\perp}$ be the component in the orthogonal complement of Z_K. We may write $x_0 = x_0^{Z_K} + x_0^{\perp}$.

Because the model is linear and $x_0^{\perp} \notin Z_K$, we only observe the output because of $x_0^{Z_K}$, that is, $y(k) = CA^k x_0^{Z_K} + CA^k x_0^{\perp} = CA^k x_0^{Z_K}$. Therefore, when solving the inverse problem, the sensor data will not reveal any information regarding the component of the initial condition in the orthogonal complement of Z_K. We cannot determine $x_0^{\perp}$. In the setting of inverse problems, this is the mathematical manifestation of one type of ill-posedness—there are an infinity of initial conditions that produce the same sensor data.

The classical inverse problem approach is to solve the optimization problem (6) with a low-rank approximation of the Hessian. This approach approximates $x_0^{Z_K}$ and disregards $x_0^{\perp}$. The Hessian-based model reduction approach utilizes Z_K to form a basis for the solution of the inverse problem but also projects the governing equations onto the subspace defined by that basis. The additional error incurred by the projection of the equations is compensated by our ability to use the resulting reduced model for forward predictions past the observation period at a low computational cost.

DOI: 10.1002/fld

We now show the equality between the transient observability gramian and the Hessian of the inverse problem. In light of this, the previous discussion can be seen as an alternative motivation for the Hessian-based model reduction for the solution of inverse problems of the form (6).

Proposition 1
The Hessian H is equal to the transient observability gramian G_K.

Proof
Note that

$$\mathbf{A}^{-1} = \begin{bmatrix} I & 0 & 0 & 0 & \cdots & 0 \\ A & I & 0 & 0 & & \vdots \\ A^2 & A & I & 0 & \ddots & \\ \vdots & A^2 & A & I & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ A^{K-1} & \cdots & A^3 & A^2 & A & I \end{bmatrix},$$

so that

$$\mathbf{A}^{-1}\mathbf{F} = \begin{bmatrix} A \\ A^2 \\ \vdots \\ A^K \end{bmatrix}$$

and

$$\mathbf{C}\mathbf{A}^{-1}\mathbf{F} = \begin{bmatrix} CA \\ CA^2 \\ \vdots \\ CA^K \end{bmatrix}.$$

Therefore, $H = (\mathbf{C}\mathbf{A}^{-1}\mathbf{F})^T(\mathbf{C}\mathbf{A}^{-1}\mathbf{F}) = \sum_{k=1}^{K} \left(CA^k\right)^T CA^k = G_K$. □

4. APPLICATION TO ATMOSPHERIC CONTAMINANT TRANSPORT

We consider a physical process modeled by the convection–diffusion equation,

$$\frac{\partial w}{\partial t} + \vec{v} \cdot \nabla w - \kappa \nabla^2 w = 0 \text{ in } \Omega \times (0, t_f), \tag{8}$$

$$w = 0 \text{ on } \Gamma_D \times (0, t_f), \tag{9}$$

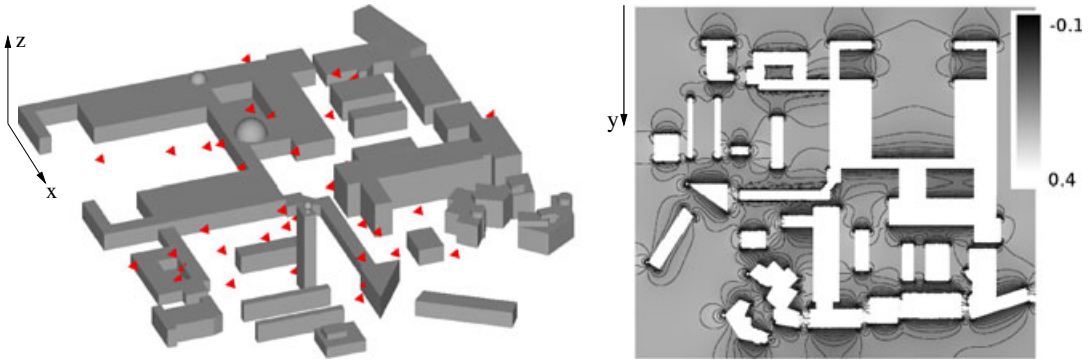
$$\frac{\partial w}{\partial n} = 0 \text{ on } \Gamma_N \times (0, t_f), \tag{10}$$

$$w = w_0 \text{ in } \Omega \text{ for } t = 0, \tag{11}$$

where w is the contaminant concentration (which varies in time and over the domain Ω), $\vec{v}$ is the velocity vector field, κ is the diffusivity, t_f is the time horizon of interest, and w_0 is the given initial condition. Homogeneous Dirichlet boundary conditions are applied on the inflow boundary Γ_D, whereas homogeneous Neumann conditions are applied on the other boundaries Γ_N.

Figure 1 shows our three-dimensional domain representing the Massachusetts Institute of Technology campus. The convection–diffusion equation for contaminant transport is discretized using the DG finite element method. The approximation basis is an order-p Lagrange on a fixed mesh of 251,853 tetrahedral elements. The degree of freedom count is therefore $251,853 \times 4 = 1,007,412$ for $p = 1$ and $251,853 \times 10 = 2,518,530$ for $p = 2$. For the DG discretization, pure upwinding is

DOI: 10.1002/fld



(a) Domain and sensor locations (b) Velocity field (y-component)

Figure 1. (a) Computational domain with sensor locations (triangles). (b) The y-component of the velocity field on a horizontal slice ($z = 0.09$) of the domain. The freestream velocity is in the y-direction.

used for the convection term, and the second form of Bassi and Rebay is used for the diffusive term [24]. For the results in this paper, the Peclet number based on the domain length (L), freestream velocity magnitude ($|\vec{v}_\infty|$), and diffusivity is $Pe = |\vec{v}_\infty| L/\nu = 50$.

The DG method was chosen for its stability properties for convection-dominated flows, for its arbitrary high-order capability on general unstructured meshes around complex geometries, and for a compact stencil that simplifies the solver algorithm and parallelization. We note that although $p = 1$ DG yields the same formal second-order accuracy as standard reconstruction-based finite-volume methods, the methods are different in that DG places all high-order degrees of freedom within an element and, hence, maintains an element-wise compact stencil.

Second-order backwards-differencing is used for temporal discretization. With this choice of time discretization, state $x(k)$ depends on states $x(k-1)$ and $x(k-2)$. However, the system can still be written as in (3), with the appropriate modification of the subcomponent matrices to account for the two-step recurrence. The specific form (1) could also be recovered by redefining the state (e.g., by coupling adjacent states). We use 60 time steps for the inversion over time in which the contaminant traverses about one-third of the computational domain.

A temporal convergence study was performed with the forward model to determine an adequate number of time steps. This number was one for which the temporal error was on the order of the spatial discretization error. We note that both the spatial and temporal schemes are formally second-order accurate, and 60 time steps over the time interval considered yield a CFL number of approximately unity.

At each time step, the linear system of equations is solved using a line-preconditioned generalized minimal residual algorithm. To enable large-scale simulations, we performed the DG forward solver and the offline model reduction in parallel on a distributed-memory cluster with a message-passing interface communication. Specifically, 200 processors (quad-core AMD Opteron2354, 2.3 GHz, 2 GB random-access memory per core (Advanced Micro Devices, Sunnyvale, CA, USA)) are used for the model reduction calculations

Reduced models are constructed offline for $p = 1$ solution approximation of the inverse and the prediction problems. For us to construct these reduced models, eigenvalues and eigenvectors of the Hessian matrix are calculated using an implicitly restarted Lanczos iteration based on the method implemented in ARPACK [13]. Each iteration requires a forward solve and an adjoint solve, as the Hessian matrix is too large to form or store explicitly. For the inverse problem, a fixed number of 250 eigenvectors are calculated to 10^{-6} relative tolerance in the Ritz values in about 5000 CPU hours or roughly a day on 200 processors. We select the first 250 eigenvectors because they meet our user-specified tolerance that the ratio of the smallest to the largest eigenvalue is less than 5×10^{-5}.

Next, the Hessian eigenvectors are used to seed forward simulations, and basis vectors are computed using the POD of the forward simulation snapshots. These snapshots are taken every three iterations. A continuous vector inner product corresponding to the integration over the computational

DOI: 10.1002/fld

domain is used in the POD. The full-order sparse matrices associated with the DG discretization and the output calculation are projected onto the basis vectors to produce the reduced model matrices. All calculations involving full-order vectors and matrices are performed in parallel. All reduced model computations are performed in MATLAB on an AMD Athlon64 X2 Dual Core Processor operating at 1 GHz core (Advanced Micro Devices, Sunnyvale, CA, USA).

Two types of regularization terms $\mathcal{R}(x_0)$ in (5) are used in the full and reduced inversion calculations. These are continuous Tikhonov regularization $\mathcal{R}_M(x_0)$ and Laplace regularization $\mathcal{R}_L(x_0)$ given by

$$\mathcal{R}_M(x_0) = \frac{\beta_M}{2} x_0^T M x_0, \quad \mathcal{R}_L(x_0) = \frac{\beta_L}{2} x_0^T L x_0, \tag{12}$$

where $M, L \in \mathbb{R}^{n\times n}$ are, respectively, the mass matrix and the discrete Laplace operator arising from the DG discretization. β_M, β_L are the regularization parameters. The mass matrix is included in the Tikhonov case to yield basis-independent regularization with a continuous interpretation as the square integral of the initial concentration. On the other hand, Laplace regularization penalizes spatial oscillations in the initial condition. In our numerical experiments, both Tikhonov and Laplace regularizations are used for the full-order inversions, that is, $\mathcal{R}(x_0) = \mathcal{R}_M(x_0) + \mathcal{R}_L(x_0)$, whereas for the reduced-order inversions, Laplace regularization $\mathcal{R}(x_0) = \mathcal{R}_L(x_0)$ alone is found to be sufficient. The regularization parameters, β_M and β_L, are determined empirically by monitoring the convergence of the inverse solve, which is performed using the conjugate gradient method, and by inspecting the inverse solutions and computed outputs. The full-order results are shown for $\beta_M = 10$ and $\beta_L = 0.5$, whereas the reduced-order results use $\beta_L = 0.1$.

5. RESULTS AND DISCUSSION

The outputs in this example are obtained from 36 sensors with locations as shown in Figure 1(a) and with a prescribed velocity field shown in Figure 1(b). The sensors are spread evenly in x and y (away from the farfield boundaries) to span the range $z = 0.05$–0.12 in the vertical direction. The sensors were chosen in a pseudorandom fashion by a user without iteration or tuning. For reference, the dimensions of the right rectangular prism enclosing the domain are $4.7 \times 4.1 \times 0.85$, where one model unit corresponds to 100 m in the real urban domain. The velocity field is governed by potential flow with isopotential planes at the inflow (minimum y) and outflow (maximum y) boundaries and flow tangency on all other boundaries. The potential difference was set to unity, resulting in a freestream velocity magnitude of $1/4.1$. No physical significance is ascribed to this velocity magnitude, as comparisons with reality are made on the basis of the Peclet number. The sensors provide an estimate of the local contaminant concentration at every time step. Experimental observations are generated synthetically by using a forward simulation run with finite element order $p = 2$ approximation and a prescribed initial condition as shown in Figure 2. The data are

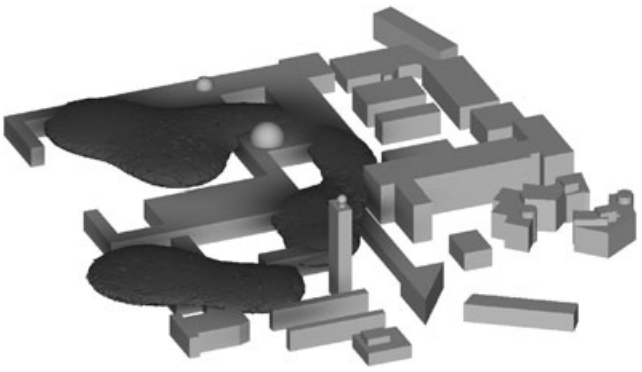


Figure 2. Prescribed (actual) initial condition used for the experiment. This plot depicts an isosurface of the contaminant at a nondimensional concentration of 0.75.

DOI: 10.1002/fld

corrupted by an additive, unbiased normal error with standard deviation $\sigma = 0.005$, which amounts to approximately 5% error for many of the sensors. The observation period consists of time steps $k = 1, 2, \ldots, 60$, and the prediction period is $k = 61, 62, \ldots, 120$.

Our objective is to utilize the corrupted data to invert for the initial condition by solving the inverse problem (6). Once we have an estimate of the initial condition, the goal is to predict the future evolution of the contaminant field, forward through time step $k = 120$. We study the inversion, prediction, and combined performance of the Hessian-based reduced models of varying dimensions for the initial condition and noisy data set described previously. To generate the reduced models, we sample 250 initial conditions (eigenvectors of the Hessian), which includes all eigenvectors up to our user-specified tolerance on the basis of the eigenvalues of the Hessian (see Figure 3). From each of these initial conditions, we generate a state trajectory. The POD basis vectors are then computed using a singular value decomposition of the complete set of snapshots. Figure 4 shows the resulting POD

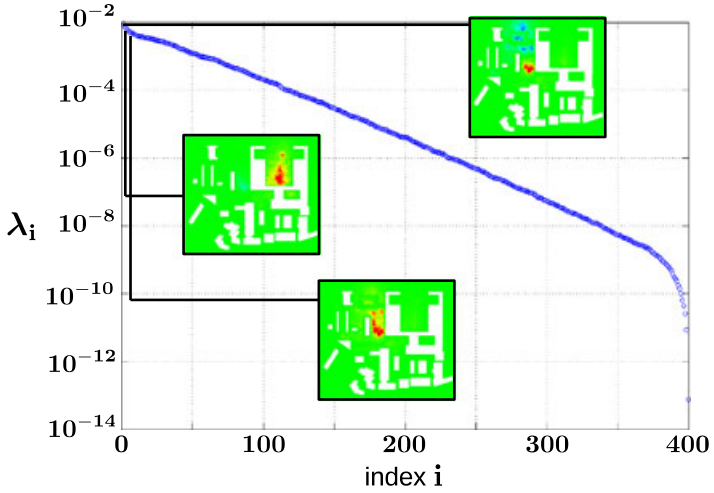


Figure 3. Eigenvalue spectrum of the Hessian of the inverse problem. The first 250 eigenvectors (i.e., those with nonnegligible eigenvalues) are used as seeds for the reduced model construction. The first (inset, top), second (inset, left), and third (inset, bottom) Hessian eigenvectors are also presented.

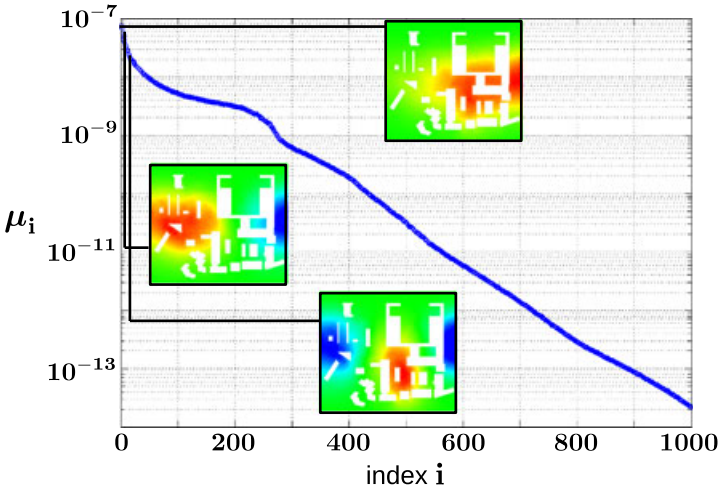


Figure 4. The first 1000 POD singular values, μ_i, for the snapshot set resulting from the use of 250 Hessian eigenvectors to generate state trajectories. The first (inset, top), second (inset, left), and third (inset, bottom) proper orthogonal decomposition modes are also presented.

DOI: 10.1002/fld

singular values. We obtain a reduced model of size m by using the first m POD basis vectors in the columns of the basis matrix $\mathbf{V}$.

For the entirety of this section, we use the following notation. We use x_0 for the actual initial condition, x_{0_F} for the inversion estimate of the initial condition using the full finite element model, and x_{0_m} for the inversion estimate of the initial condition using a reduced model of dimension m. For full finite element model prediction runs, we denote the state at time step k starting from the initial condition estimate u by $x_F(k;u)$. Likewise, reduced model (of dimension m) state predictions are denoted by $x_m(k;u)$. Further in this section, we study the ℓ_2 output error over the prediction period. The full model prediction resulting from the initial condition estimate u is written $y_F(61-120;u)$; the reduced model prediction is $y_m(61-120;u)$.

Before investigating the convergence of the reduced models with increasing dimension, we first present a side-by-side comparison of the state at time steps $k = 0, 60, 120$ (initial condition, end of observation period, end of prediction period). In Figure 5, we show a horizontal slice of the state at each time for three cases: (i) the actual initial condition; (ii) inversion and prediction using the full finite element model with $n = 1,007,412$ ($p = 1$); and (iii) inversion and prediction using a reduced model of dimension $m = 800$. Although the full and reduced model inversions do not recover the initial condition exactly, they do locate the most concentrated regions with the contaminant. Note that we do not expect either model to be able to recover the initial condition exactly, unless it happens to reside only within the observable subspace of the full model. However, the figure shows that the reduced model has picked up all of the features of the full inversion with only 800 basis vectors, whereas the full model contains over 1 million degrees of freedom. This large reduction in dimension translates directly to computational savings in the inversion

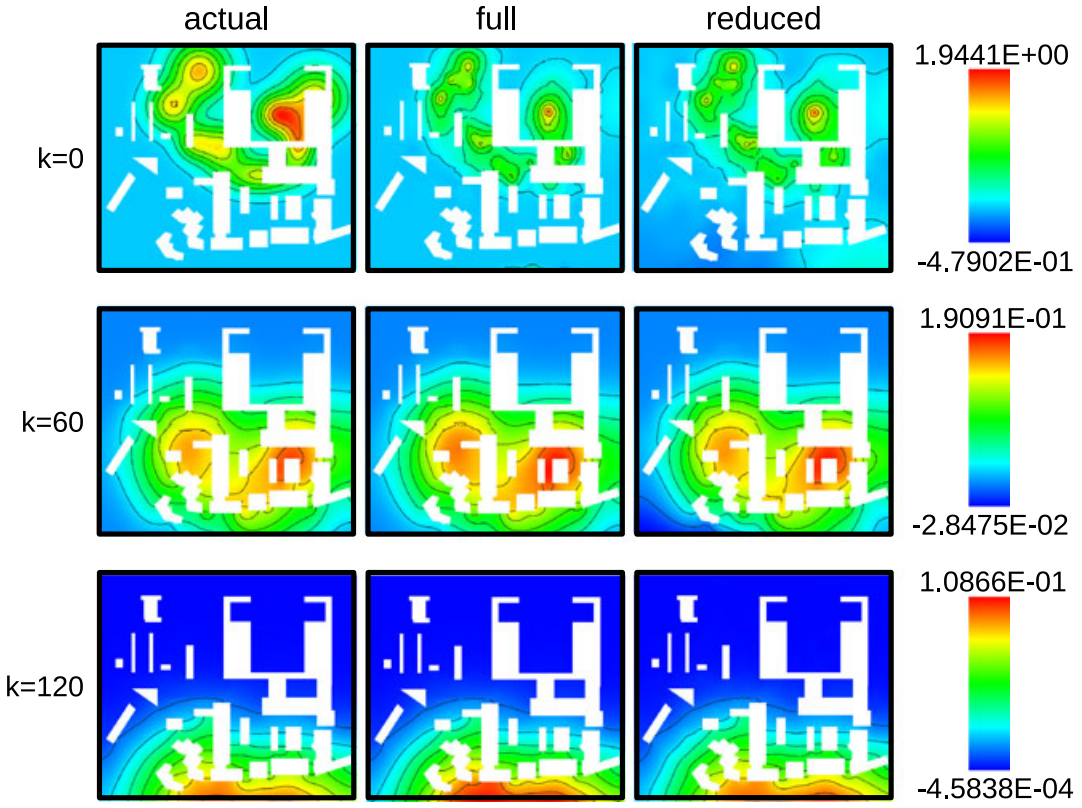


Figure 5. Comparison between the actual contamination event (left column), inversion and prediction by the full model (center column), and inversion and prediction by the reduced model of dimension $m = 800$ (right column) at the initial condition $k = 0$ (top row), at the end of the observation period $k = 60$ (middle row), and at the end of the prediction period $k = 120$ (end row). All plots are cuts through $z = 0.1$, a representative slice.

DOI: 10.1002/fld

Table I. CPU timings and percent relative errors for the full and reduced model inversions and predictions. For the sake of brevity, we record the timings only for the reduced models of dimension $m = 200, 400, 600, 800$. The percent relative errors are provided for the inversion (Figure 7) and prediction (Figure 10). For the prediction, the relative errors are computed using the same model for inversion and prediction. Full-order computations are performed in parallel to a distributed-memory cluster with MPI communication by using 200 processors (quad-core AMD Opteron2354, 2.3 GHz, 2 GB RAM per core), and the CPU time is shown aggregated over all cores used in the parallel runs. The reduced model computations are performed in MATLAB on an AMD Athlon64 X2 Dual Core Processor operating at 1 GHz. The offline computational cost associated with the reduced-order model is 5000 CPU hours.

	Inversion		Prediction	
Model	CPU time	Error (%)	CPU time	Error (%)
Full $n = 1\,007\,412$	3300 h	57	7.8 h	5.3
Reduced $m = 800$	68 min	62	3.7 s	3.1
Reduced $m = 600$	47 min	55	1.7 s	3.9
Reduced $m = 400$	3.5 min	58	1.4 s	11
Reduced $m = 200$	16 s	171	0.6 s	58

(see Table I). Figure 6 shows the observations and predictions at six sensor locations in the domain. The noise-corrupted data are shown in the observation period followed by the actual contamination in the prediction period. The results demonstrate that the reduced model is able to replicate and, in some cases, outperform the full model in the predictions for both growing and decaying concentration profiles.

The authors believe that the better performance of the reduced model may be due to the effects of regularization in the full inversion. The reduced model is a natural regularizer because the governing equations are projected onto a low-dimensional subspace defined by basis vectors that tend to be smooth. We find that difficulties with choosing regularization type and magnitude are often diminished when using the reduced model. The type of regularization for the reduced model is different from that used for the full model because the reduced model already incorporates a certain subspace regularization because the initial condition is obtained as a linear combination of POD modes.

We now turn our attention to the inversion performance. In particular, we are interested in the convergence of the reduced model initial condition estimate as the model dimension m increases. In Figure 7, we plot the relative error of the inversion estimate of the initial condition for models of dimension varying from $m = 50$ to $m = 800$. We present the error relative to the L_2-norm of the actual initial condition. Also shown in the figure is the inverted field for the full model and three of the reduced models. The dashed line indicates the relative error of the full model inversion estimate. For reduced models of dimension smaller than $m = 350$, the inversion is inaccurate. For reduced models of dimension $m \geqslant 400$, the accuracy of the reduced model inversion is similar to that of the full model inversion. We have not tied this model size to an energy criterion; in general, it will depend on the problem and the regularization used in the inversion.

Another way of measuring the performance of the inversion is to propagate the inverted initial condition estimates forward by using the full finite element model. In Figure 8, we plot the relative error of the full model prediction stemming from initial conditions estimated via inversions with reduced models of varying dimension. In particular, we measure the error at the midpoint of the prediction period, time step $k = 90$. The results are similar to Figure 7. Once the reduced model becomes large enough ($m \geqslant 400$), the relative error falls below 8%. The full model prediction of the initial condition estimated using the full model to solve the inverse problem is also shown for reference; it has approximately 9% relative error.

In Figure 9, we consider the prediction performance of the reduced models using the inversion estimate of the reduced model of dimension $m = 800$. This singles out the prediction performance, relegating lossy projections only to that of the initial condition from the $m = 800$ model to a smaller reduced model. Note that the transformation from the $m = 800$ inverted initial condition estimate to a smaller reduced model of dimension r involves only retaining the first r reduced states because

DOI: 10.1002/fld

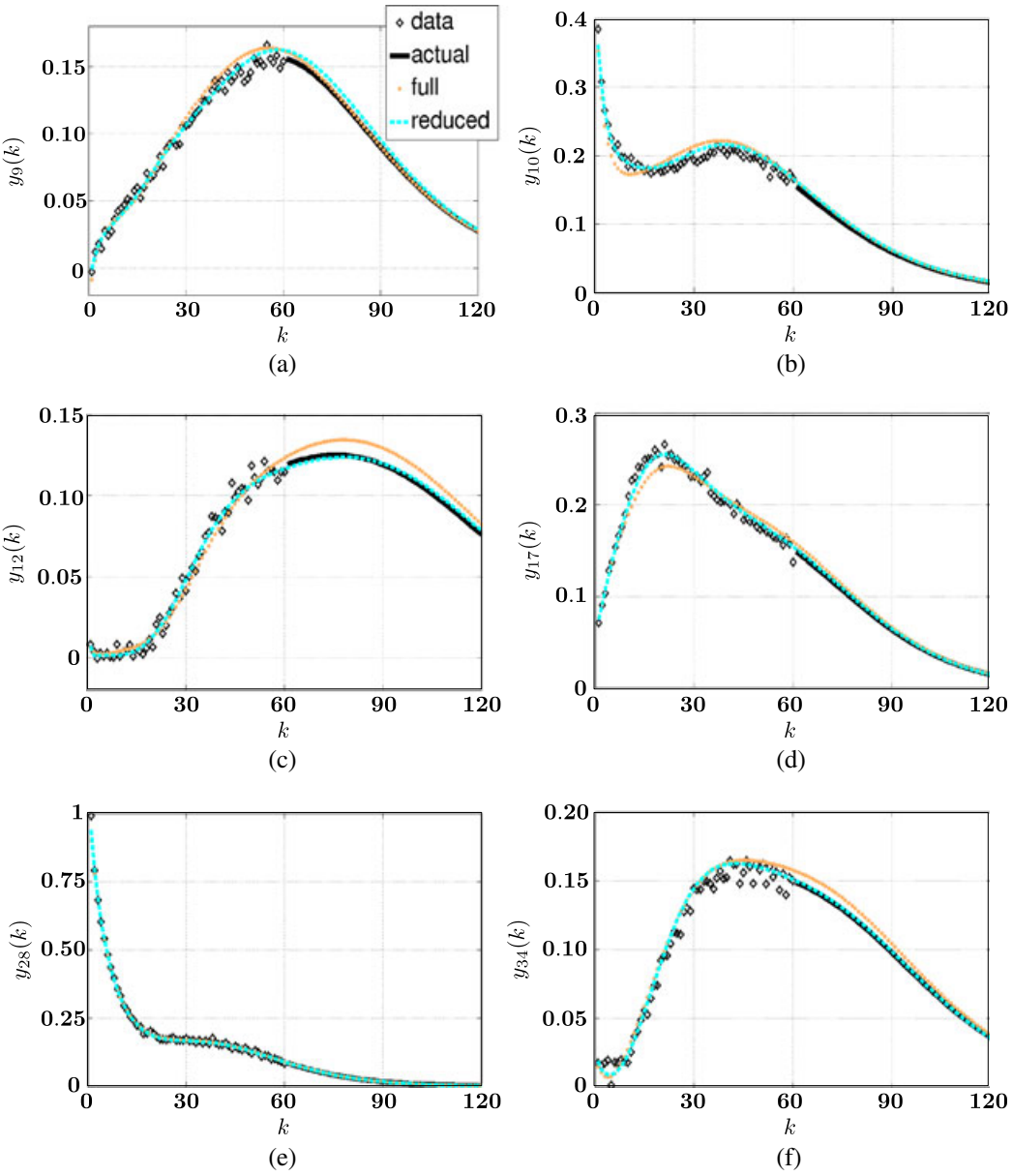


Figure 6. Sensor readings and predictions at six representative sensors. The diamonds show the noisy observation data for $k = 1, 2, \ldots, 60$, whereas the solid black line is the prediction given the actual initial condition for $k = 61, \ldots, 120$. The dotted and dashed lines show the prediction from the inverted initial condition ahead for $k = 1, 2, \ldots, 120$ using the full and $m = 800$ reduced models, respectively.

of the orthogonality of the basis vectors. At time step $k = 90$, we measure the error relative to the actual concentration field. Using the reduced model for both inversion and prediction is the fastest method for obtaining a prediction. However, if there is extra time available to carry out the computations, it is possible to use the reduced model for the more costly inversion, followed by the full model to predict forward. The figure reveals, once again, that the combined performance is adequate for a reduced model of modest dimension.

Finally, we consider the error in the prediction of outputs, as opposed to the state, over the prediction period (see Figure 10). The errors are measured in the discrete ℓ_2 sense relative to the outputs of the actual scenario. The inversion estimates and forward predictions are computed here with the same reduced model for each size m. For comparison, the prediction error resulting from the full

DOI: 10.1002/fld

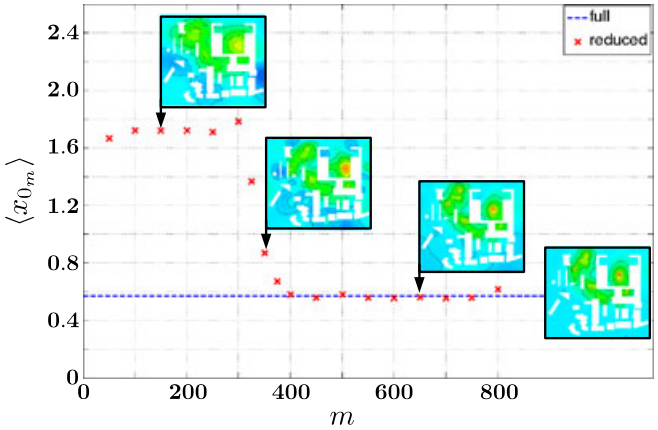


Figure 7. Convergence in the relative error of the reduced model inversion estimate with increasing dimension. The inversion estimate using a reduced model of dimension m has a relative error $\langle x_{0_m} \rangle = \|x_{0_m} - x_0\|_{L_2(\Omega)} / \|x_0\|_{L_2(\Omega)}$, where x_0 is the initial condition used to generate the data. For comparison, the full model inversion estimate x_{0_F} (inset, right) and its relative error (dashed line) are also shown.

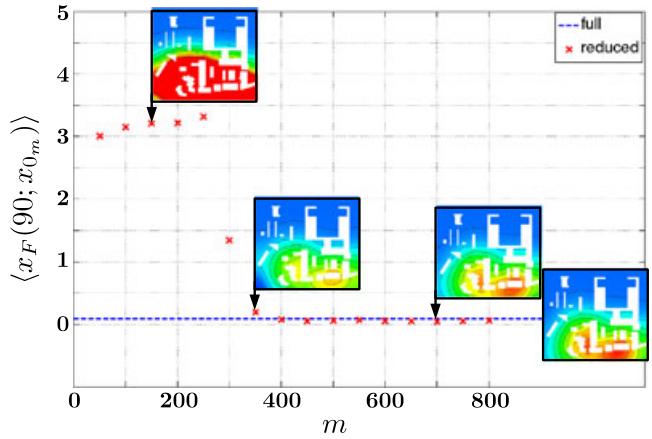


Figure 8. Convergence in the relative error, for the increasing dimension, of the full model prediction using the reduced model inversion estimate. The error is measured at the midpoint of the prediction period, time step $k = 90$. The relative error is given by $\langle x_F(90; x_{0_m}) \rangle = \|x_F(90; x_{0_m}) - x_F(90; x_0)\|_{L_2(\Omega)} / \|x_F(90; x_0)\|_{L_2(\Omega)}$, where x_0 is the initial condition used to generate the data. For comparison, the full model prediction using the full model inversion $x_F(90; x_{0_F})$ (inset, right) and its relative error (dashed line) are also shown.

model inversion is shown. These results demonstrate that we can predict the output of interest within approximately 5% error by using reduced models with as few as $m = 500$ dimensions.

It is important to note that the prediction capability of the reduced model is directly tied to the Peclet number, the size of the domain, and the length of the observation period. If the observation period is long enough such that an initial condition can travel across the entire domain, then that condition is included in the basis constructed by the Hessian-based model reduction. Because time $t = 0$ is arbitrarily assigned, this implies that the reduced model will be able to predict adequately forward as long as the state at the end of the observation period is properly reconstructed.

Our reduced models are able to deliver accurate prediction performance for a time horizon beyond the observation period even though the snapshot data do not extend into the prediction period. This occurs for our problem because prediction locations are identical to sensor locations and because the observation period is sufficiently long compared with the prediction period. Because of the time invariance of the model, the evolution of the state from the end of the observation period is the same

DOI: 10.1002/fld

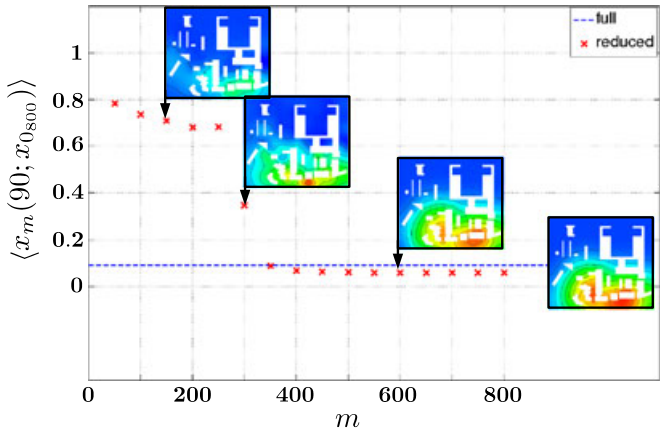


Figure 9. Convergence in the relative error, for increasing dimension, of the combined performance of the reduced models. Inversion estimates are all computed using the reduced model of dimension $m = 800$, and predictions are performed using the reduced models of varying dimension m. The relative error is given by $\langle x_m(90;x_{0_{800}})\rangle = \|x_m(90;x_{0_{800}}) - x_F(90;x_0)\|_{L_2(\Omega)}/\|x_F(90;x_0)\|_{L_2(\Omega)}$, where x_0 is the initial condition used to generate the data. For comparison, the full model prediction using the full model inversion estimate $x_F(90;x_{0_F})$ (inset, right) and its relative error (dashed line) are also shown.

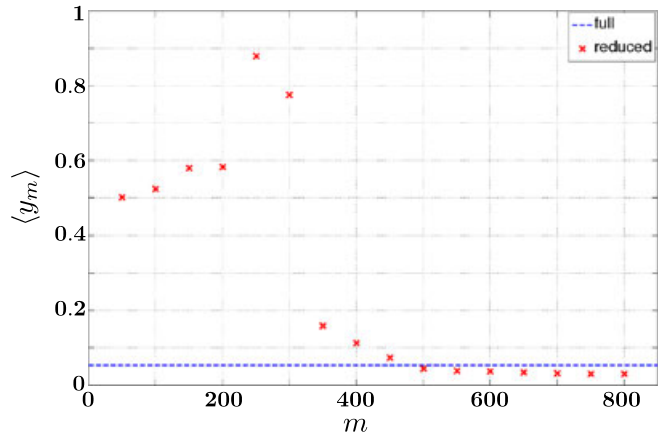


Figure 10. Convergence of the reduced model prediction errors with increasing dimension m. The prediction error is measured in discrete ℓ_2 over the sensors for the time steps $k = 61,\ldots,120$. That is, $\langle y_m\rangle = \|y_F(61-120;x_0) - y_m(61-120;x_{0_m})\|_{\ell_2}/\|y_F(61-120;x_0)\|_{\ell_2}$. Note that the forward prediction and inversions are computed here with the same reduced model for each size m. The dashed line is the relative error in the full model prediction.

trajectory that would result with that state as an initial condition, but shifted in time. If this trajectory lies within the observable subspace, its dynamics will have been sampled in the reduced model construction. It should be noted that for this problem, large inversion errors can lead to small prediction errors (see Table I). Only some components of the initial condition impact the prediction outputs of interest. Accurate predictions, therefore, require only an accurate estimation of those modes. In our problem, the prediction outputs are the same as the observations from sensors, just shifted in time. Because of the time invariance of the model and that the Hessian-based model reduction targets accuracy for the observation outputs, the modes needed for an accurate prediction are also accurately estimated in the inverse problem; this leads to small prediction errors.

6. CONCLUSIONS

The results demonstrate that the Hessian-based reduced models can be used for inversion and prediction in some contaminant transport scenarios. Control-theoretic concepts play an important role

DOI: 10.1002/fld

in our understanding of our ability to reconstruct certain modes of the initial condition and the effect of regularization on the solution. The application of similar concepts to prediction and control problems are subjects of ongoing research. Connections between control-theoretic concepts and PDE-constrained optimization approaches is a rich area that may lead to new insights and algorithms. For example, controllability and observability of the isothermal weakly compressible single-phase flow equations are studied for the identification and control of a reservoir in [17]. The findings lead to a new method for controllability-based and observability-based subspace regularization in the history-matching problem.

We present results for just one arbitrarily chosen initial condition and corrupted data, but we expect the results to be similar for initial conditions with the characteristics encoded in our regularization scheme (exhibiting some smoothness). With the dramatic reduction in model dimension from more than 1 million degrees of freedom to just a few hundred, we make possible the inversion and prediction of large-scale contaminant transport events on laptop computers in the field. Future applications of reduced models in this setting include the addition of optimal control scenarios to mitigate contamination in critical areas. The Hessian-based model reduction could also be applied in the nonlinear setting, but the explicit connection to the transient observability gramian is limited to the linear case.

ACKNOWLEDGEMENTS

The first and third authors acknowledge the support of the Department of Energy Applied Mathematics Program, Award Number DE-FG02-08ER25858 (Program Manager A. Landsberg), and the support of the Singapore-MIT Alliance Computational Engineering Programme.

This study was funded by the Computational Engineering Programme of the Singapore-MIT Alliance; the Air Force Office of Scientific Research under grant number FA9550-06-0271, program manager Dr Fariba Fahroo; the National Science Foundation under DDDAS grant CNS-0540186, program directors Dr Frederica Darema and Dr Anita Lasalle; and the Computer Science Research Institute at Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

REFERENCES

1. Deane A, Kevrekidis I, Karniadakis G, Orszag S. Low-dimensional models for complex geometry flows: application to grooved channels and circular cylinders. *Physics of Fluids* 1991; **3**(10):2337–2354.
2. Holmes P, Lumley J, Berkooz G. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press: Cambridge, UK, 1996.
3. Sirovich L. Turbulence and the dynamics of coherent structures. Part 1: coherent structures. *Quarterly of Applied Mathematics* 1987; **45**(3):561–571.
4. Moore B. Principal component analysis in linear systems: controllability, observability, and model reduction. *IEEE Transactions on Automatic Control* 1981; **AC-26**(1):17–31.
5. Gugercin S, Antoulas A. A survey of model reduction by balanced truncation and some new results. *International Journal of Control* 2004; **77**:748–766.
6. Li J, White J. Low rank solution of Lyapunov equations. *SIAM Journal on Matrix Analysis and Applications* 2002; **24**(1):260–280.
7. Penzl T. Algorithms for model reduction of large dynamical systems. *Linear Algebra and its Applications* 2006; **415**(2–3):322–343.
8. Sorensen D, Antoulas A. The Sylvester equation and approximate balanced reduction. *Linear Algebra and its Applications* 2002; **351–352**:671–700.
9. Feldmann P, Freund R. Efficient linear circuit analysis by Padé approximation via the Lanczos process. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 1995; **14**:639–649.
10. Gallivan K, Grimme E, Van Dooren P. Padé approximation of large-scale dynamic systems with Lanczos Methods. *Proceedings of the 33rd IEEE Conference on Decision and Control*, Lake Buena Vista, FL., December 1994; 443–448.
11. Grimme E. Krylov projection methods for model reduction. *PhD Thesis*, Coordinated-Science Laboratory, University of Illinois at Urbana-Champaign, 1997.
12. Bashir O, Willcox K, Ghattas O, van Bloemen Waanders B, Hill J. Hessian-based model reduction for large-scale systems with initial condition inputs. *International Journal for Numerical Methods in Engineering* 2008; **73**(6):844–868.
13. Lehoucq R, Maschhoff K, Sorensen D, Yang C. *ARPACK Users' Guide*. SIAM: Philadelphia, 1998.

DOI: 10.1002/fld

14. Wang J, Zabaras N. Using Bayesian statistics in the estimation of heat source in radiation. *International Journal of Heat and Mass Transfer* 2005; **48**:15–29.
15. Kaleta M, Hanea R, Heemink A, Jansen J. Model-reduced gradient-based history matching. *Computational Geosciences* 2011; **15**(1):135–153.
16. Galbally D, Fidkowski K, Willcox K, Ghattas O. Nonlinear model reduction for uncertainty quantification in large-scale inverse problems. *International Journal for Numerical Methods in Engineering* 2010; **81**(12):1581–1608.
17. Zandvliet M, van Doren J, Bosgra O, Jansen J, van den Hof P. Controllability, observability and identifiability in single-phase porous media flow. *Computational Geosciences* 2008; **12**(4):605–622.
18. Vogel C. *Computational Methods for Inverse Problems*. SIAM: Philadelphia, 2002.
19. Adavani S, Biros G. Multigrid algorithms for inverse problems with linear parabolic PDE constraints. *SIAM Journal on Scientific Computing* 2008; **31**:369–397.
20. Barrault M, Maday Y, Nguyen N, Patera A. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. *Numerical Analysis* 2004; **339**:667–672.
21. Chaturantabut S, Sorensen D. Nonlinear model reduction via discrete empirical interpolation. *SIAM Journal on Scientific Computing* 2010; **32**(5):2737–2764.
22. Astrid P, Weiland S, Willcox K, Backx T. Missing point estimation in models described by proper orthogonal decomposition. *IEEE Transaction on Automatic Control* 2008; **53**(10):2237–2251.
23. Georges D. The use of observability and controllability gramians or functions for optimal sensor and actuator location in finite-dimensional systems. *Proceedings of the 34th IEEE Conference on Decision and Control*, New Orleans, LA, December 1995; 3319–3324.
24. Bassi F, Rebay S. GMRES discontinuous Galerkin solution of the compressible Navier–Stokes equations. In *Discontinuous Galerkin Methods: Theory, Computation and Applications*, Cockburn K, Karniadakis G, Shu C-W (eds). Springer: Berlin, 2000; 197–208.

DOI: 10.1002/fld