

RESEARCH

Open Access

Model order reduction for Bayesian approach to inverse problems

Ngoc-Hien Nguyen^{1*}, Boo Cheong Khoo² and Karen Willcox³

*Correspondence:

mpennh@nus.edu.sg

¹Singapore-MIT Alliance, National University of Singapore, Singapore 117576, Singapore

Full list of author information is available at the end of the article

Abstract

This work presents an approach to solve inverse problems in the application of water quality management in reservoir systems. One such application is contaminant cleanup, which is challenging because tasks such as inferring the contaminant location and its distribution require large computational efforts and data storage requirements. In addition, real systems contain uncertain parameters such as wind velocity; these uncertainties must be accounted for in the inference problem. The approach developed here uses the combination of a reduced-order model and a Bayesian inference formulation to rapidly determine contaminant locations given sparse measurements of contaminant concentration. The system is modelled by the coupled Navier-Stokes equations and convection-diffusion transport equations. The Galerkin finite element method provides an approximate numerical solution—the ‘full model’, which cannot be solved in real-time. The proper orthogonal decomposition and Galerkin projection technique are applied to obtain a reduced-order model that approximates the full model. The Bayesian formulation of the inverse problem is solved using a Markov chain Monte Carlo method for a variety of source locations in the domain. Numerical results show that applying the reduced-order model to the source inversion problem yields a speed-up in computational time by a factor of approximately 32 with acceptable accuracy in comparison with the full model. Application of the inference strategy shows the potential effectiveness of this computational modeling approach for managing water quality.

Keywords: Bayesian; Convection-diffusion equation; Navier-Stokes equations; Markov chain Monte Carlo; Inverse problem; Proper orthogonal decomposition; Reduced-order model

Background

Hydrodynamic processes such as contaminant transport in lakes and reservoirs have a direct impact on water quality. The contaminants will appear, spread out, and decrease in concentration, etc. because of some processes such as convection, diffusion, time rate release of contaminants, and distance of travel. To simulate such processes, a coupled system of partial differential equations (PDEs) including the Navier-Stokes equations (NSEs) and contaminant transport equations needs to be solved. A better understanding of these processes is important in managing water resources effectively.

The direct or forward problems compute the distribution of contaminant directly from given input information such as contaminant location, contaminant properties, fluid flow

properties, boundary conditions, initial conditions, etc. On the contrary, the inverse problems infer the unknown physical parameters, boundary conditions, initial conditions, or geometry given a set of measured data. These known data can be obtained experimental or computational. In realistic applications, data are not perfect because of error due to sensor noise. In addition, the model may contain some uncertain parameters such as wind velocity. Thus, inverse problems are generally stochastic problems.

A Bayesian inference approach to inverse problems is one way to deal with the stochastic problem. This approach takes into account both the information of the model parameters with uncertainty and the inaccuracy of data in terms of a probability distribution [1,2]. The Bayesian approach provides a general framework for the formulation of wide variety of problems such as climate modeling [3], contaminant transport modeling [4-6], and heat transfer [7]. Under the Bayesian framework, simulated solutions need to be evaluated repeatedly over different samples of the input parameters. There are available sampling strategies associated with Bayesian computation such as Markov chain Monte Carlo (MCMC) methods [8-11]. However, if we use traditional PDE discretization methods, such as finite element or finite volume methods, the resulting numerical models describing the system will be very large and expensive to solve.

This paper presents an efficient computational approach to solve the statistical inverse problem. The approach uses the combination of a reduced-order model and a Bayesian inference formulation. The Galerkin finite element method [12] provides an approximate numerical solution - the 'full model' or 'forward model'. The MCMC method is applied to solve the Bayesian inverse problems. In particular, we are interested in inferring an arbitrary source location in a time-dependent convection-diffusion transport equation, given a velocity field and a set of measured data of the concentration field at sparse spatial and temporal locations. We obtain the velocity field by solving the Navier-Stokes equations.

Solution of large-scale inverse problems can be accelerated by applying model order reduction [13-15]. The idea is to project the large-scale governing equations onto the subspace spanned by a reduced-space basis, yielding a low-order dynamical system. Proper orthogonal decomposition (POD) is the most popular method to find the reduced basis for a given set of data. The snapshot POD method, which was proposed by Sirovich [16], provide an efficient way for determining the POD basis vectors. Snapshots are solutions of a numerical simulation at selected times or choices of the parameters. The choice of snapshots should ensure that the resulting POD basis captures the most important characteristics of the system.

Since the convection term in contaminant transport equations is a velocity-dependent term, we cannot apply the standard POD reduction framework. The evaluation of the velocity-dependent convection term in the reduced system still depends on the full finite element dimension and has the same complexity as the full-order system. Thus, a system with velocity-dependent convection term needs an additional treatment to obtain an efficient reduced-order model.

This paper is outlined as follows. In the section 'Mathematical model and numerical methods', we introduce the mathematical model, the finite element approximate technique, and the Bayesian inference formulation and its solution using an MCMC method. In section 'Model order reduction', we describe the model order reduction technique and the treatment of the velocity-dependent convection term. In the section 'Numerical example', we use the numerical example to demonstrate the solution of the statistical

inverse problem and the reduced-order model performance. We provide some concluding remarks in the final section.

Methods

In this section, the mathematical model, finite element approximation technique, Bayesian inference formulation, and MCMC method are briefly introduced.

Transport model

Consider the fluid flow through a physical domain $\mathcal{D} \subset \mathbb{R}^d, d = 1, 2, 3$ with boundary Γ . The contaminant concentration $c := c(\mathbf{x}, t; \theta)$ satisfies the dimensionless parabolic PDEs, boundary conditions, and initial conditions as follows:

$$\frac{\partial c}{\partial t} + \mathbf{u} \cdot \nabla c - \frac{1}{Pe} \Delta c = f \quad \text{in } \mathcal{D} \times [t_0, t_f], \quad (1)$$

$$c = g \quad \text{on } \Gamma_D \times [t_0, t_f], \quad (2)$$

$$\frac{\partial c}{\partial \mathbf{n}} = 0 \quad \text{on } \Gamma_N \times [t_0, t_f], \quad (3)$$

$$c(\mathbf{x}, t_0) = c_0(\mathbf{x}) \quad \text{in } \mathcal{D}, \quad (4)$$

where $\mathbf{x} \in \mathcal{D}$ denotes the spatial coordinates, $t \in [t_0, t_f]$ denotes time, and c_0 is the given initial condition. The inlet boundary Γ_D is subjected to a Dirichlet condition, while the remainder of the boundary $\Gamma_N = \Gamma \setminus \Gamma_D$ satisfies Neumann conditions. The external source $f(\mathbf{x}, t; \theta)$ in Equation 1 is used here as a input parameter, where $\theta \in \mathcal{D}$ is the source location. In this study, we suppose that the contaminant source drives the transport process, given a velocity field $\mathbf{u} \in \mathbb{R}^d$. Thus, the Péclet number $Pe = \frac{\|\mathbf{u}\|L}{\kappa}$, where κ is the diffusivity and L is the characteristic length. The velocity field is a function of \mathbf{x} and t , $\mathbf{u}(\mathbf{x}, t)$, which can be obtained by solving the Navier-Stokes equations, as follows:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} - \frac{1}{Re} \nabla^2 \mathbf{u} + \nabla p = \mathbf{f}_c \quad \text{in } \mathcal{D} \times [t_0, t_f], \quad (5)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \mathcal{D} \times [t_0, t_f], \quad (6)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{on } \Gamma \times [t_0, t_f], \quad (7)$$

$$\mathbf{u}(\mathbf{x}, t_0) = \mathbf{u}_0(\mathbf{x}) \quad \text{in } \mathcal{D}, \quad (8)$$

where Re is the Reynolds number, p is the pressure field, and \mathbf{f}_c is the body force. So, for any given f and \mathbf{u} , one can solve for the solution $c(\mathbf{x}, t; \theta)$.

Finite element approximations

The finite element method [12] associated with the stabilized second-order fractional-step method is applied to solve the incompressible Navier-Stokes equations (5 to 8). The detailed discussion and derivation of this method is beyond the scope of this paper. For more details, please refer to [17,18].

We also use the finite element method for spatial discretization of the contaminant transport equations in Equations 1 to 4. This leads to the semi-discrete equations

$$\mathbf{M}\dot{\mathbf{c}} + (\mathbf{C}(\mathbf{u}) + \mathbf{K})\mathbf{c} = \mathbf{F}(t; \theta), \quad (9)$$

$$\mathbf{c}(t_0) = \mathbf{c}_0, \quad (10)$$

with the output of contaminant solution at sensor locations in the domain

$$\mathbf{y} = \mathbf{B}\mathbf{c}. \quad (11)$$

Here, $\mathbf{c}(t) \in \mathbb{R}^N$ is the discretized approximation of $c(\mathbf{x}, t)$ and contains N state unknowns where N is the number of grid points in the spatial discretization. $\dot{\mathbf{c}}$ is the derivative of \mathbf{c} with respect to time. The vector $\mathbf{y} \in \mathbb{R}^{N_o}$ contains the N_o outputs of the system related to the state through the matrix $\mathbf{B} \in \mathbb{R}^{N_o \times N}$, $N_o \ll N$. The matrices $\mathbf{M}, \mathbf{K}, \mathbf{C}(\mathbf{u}) \in \mathbb{R}^{N \times N}$ and the vector $\mathbf{F}(t; \theta) \in \mathbb{R}^N$ are defined in the finite element space as follows:

$$\mathbf{M}_{i,j} = \int_{\mathcal{D}} \phi_i(\mathbf{x}) \phi_j(\mathbf{x}) \, d\mathbf{x}, \quad (12)$$

$$\mathbf{K}_{i,j} = \int_{\mathcal{D}} \frac{1}{Pe} \nabla \phi_i(\mathbf{x}) \cdot \nabla \phi_j(\mathbf{x}) \, d\mathbf{x}, \quad (13)$$

$$(\mathbf{C}(\mathbf{u}))_{i,j} = \int_{\mathcal{D}} \mathbf{u} \cdot \nabla \phi_i(\mathbf{x}) \phi_j(\mathbf{x}) \, d\mathbf{x}, \quad (14)$$

$$(\mathbf{F}(t; \theta))_j = \int_{\mathcal{D}} f(t; \theta) \phi_j(\mathbf{x}) \, d\mathbf{x}. \quad (15)$$

Here ϕ_i with $i = 1, \dots, N$ is the finite element basis and $\theta \in \mathbb{R}^d$ is the coordinate vector of the source location.

Bayesian inference to inverse problems

The Bayesian formulation of an inference problem is derived from Bayes' theorem [1]:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}, \quad (16)$$

where $\mathbf{y} \in \mathbb{R}^{N_o}$ is the observed data and $\theta \in \mathbb{R}^d \subseteq \mathcal{D}$ is the vector of input parameters. $p(\theta)$ denotes the prior distribution of the parameters, $p(\mathbf{y}|\theta)$ is the likelihood function, and $p(\theta|\mathbf{y})$ is the posterior distribution of parameters given the observed data. $p(\mathbf{y})$ is the prior predictive distribution of \mathbf{y} and may be set to an unknown constant that is not needed in the MCMC solution method. In this case, we write Bayes's formula as

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta). \quad (17)$$

A simple model for the likelihood function can be obtained from the following relationship,

$$\mathbf{y} = G(\theta) + \eta, \quad (18)$$

where η is the noise which we assume to be white Gaussian noise $\eta \sim N(0, \sigma^2 I)$. The input-output relation in Equations 1 to 4 is denoted as the forward model $G(\theta)$. Subsequently, the likelihood can be written as

$$p(\mathbf{y}|\theta) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - G(\theta)\|^2\right). \quad (19)$$

There are many approaches to define the prior information, such as Gaussian Markov random fields or uniform distribution models, etc. In this work, we assume that our only prior information on the source location is given by the bounds on the domain. Thus, using the principle of maximum entropy [19], we take our prior to be a uniform distribution. Therefore, Equation 17 becomes

$$p(\theta|\mathbf{y}) \propto \prod_{i=1}^K \exp\left[-\frac{1}{2\sigma^2} (\mathbf{y}_i - G(\theta)_i)^T (\mathbf{y}_i - G(\theta)_i)\right]. \quad (20)$$

Here, K is the number of time steps over which we collect the output data.

Markov chain Monte Carlo for posterior sampling

MCMC simulation provides a sampling strategy from a proposal distribution to the target distribution using the Markov chain mechanism. Hastings introduced the Metropolis-Hastings (MH) algorithm [10], which is a generalized version of the Metropolis algorithm [8,9], in which we can apply both symmetric and asymmetric proposal distributions. In this work, the MH algorithm is used to solve the Bayesian inverse problem. The MH algorithm is summarized as follows:

1. Initialize the chain θ^0 and set $n = 0$
2. Repeat
 - $n = n + 1$
 - Generate a proposal point $\theta^* \sim q(\theta^*|\theta)$
 - Generate U from a uniform $U(0, 1)$ distribution
 - Update the state to θ^{n+1} as

$$\theta^{n+1} = \begin{cases} \theta^*, & \text{if } \beta < U \\ \theta^n, & \text{otherwise} \end{cases}$$
3. Until $n = N_{\text{mcmc}} \rightarrow \text{stop}$.

Here, β is the acceptance-rejection ratio, given by

$$\beta = \min \left(1, \frac{\pi(\theta^*) q(\theta^{n-1}|\theta^*)}{\pi(\theta^{n-1}) q(\theta^*|\theta^{n-1})} \right). \quad (21)$$

N_{mcmc} is the total number of samples, $\pi(\theta)$ is the target distribution, $q(\theta^*|\theta)$ is a proposal distribution, and U is a random number.

The discretized model in the form of ordinary differential equations (ODEs) (Equations 9 to 11) may have very large dimensions and be expensive to solve. The MCMC method requires evaluating repeatedly the solution of the forward model (many thousands or even millions of times). Hence, these simulations can be a computationally expensive undertaking. In such situations, the reduced-order model is needed to approximate the large-scale model, which allows efficient simulations.

Model order reduction

This section presents the model order reduction framework. This includes the reduction via projection, the proper orthogonal decomposition method, the additional treatment for the velocity-dependent convection term, and the error estimation.

Reduction via projection

A reduced-order system approximating the ODEs (9 to 11) can be obtained by approximating the full state vector \mathbf{c} as a linear combination of m basis vectors as follows:

$$\mathbf{c} \approx V \mathbf{c}_r, \quad (22)$$

where $\mathbf{c}_r \in \mathbb{R}^m$ is the reduced order state and $V = [v_1 \ v_2 \ \dots \ v_m] \in \mathbb{R}^{N \times m}$ is an orthonormal basis, i.e., $V^T V = I$. The low-order model can be derived by projecting the governing system (9 to 11) onto the reduced space formed by the column span of basis V . This yields

$$\mathbf{M}_r \dot{\mathbf{c}}_r + (\mathbf{C}_r(\mathbf{u}) + \mathbf{K}_r) \mathbf{c}_r = \mathbf{F}_r(t; \theta), \quad (23)$$

$$\mathbf{c}_r(t_0) = \mathbf{c}_{0r}, \quad (24)$$

$$\mathbf{y}_r = \mathbf{B}_r \mathbf{c}_r, \quad (25)$$

where

$$\mathbf{M}_r = V^T \mathbf{M} V, \quad (26)$$

$$\mathbf{K}_r = V^T \mathbf{K} V, \quad (27)$$

$$\mathbf{C}_r(\mathbf{u}) = V^T \mathbf{C}(\mathbf{u}) V, \quad (28)$$

$$\mathbf{F}_r(t; \theta) = V^T \mathbf{F}(t; \theta), \quad (29)$$

$$\mathbf{B}_r = \mathbf{B} V, \quad (30)$$

$$\mathbf{c}_{0r} = V^T \mathbf{c}_0. \quad (31)$$

The model reduction task is to find a suitable basis V so that $m \ll N$ and the reduced-order model yields accurate results. This study will consider POD as the method to compute the basis.

Proper orthogonal decomposition

POD provides a method to compute the reduced-order basis V and construct the low-order system. Here, we briefly describe the general POD method (more details may be found in [16]).

Supposed that we have the snapshot matrix $X \in \mathbb{R}^{N \times Q}$, where Q is the number of snapshot solutions, which are built from the instantaneous solution $\mathbf{c}^s(t_k)$ of the forward model, corresponding to $s = 1, \dots, S$ random input source locations, picked up at $k = 1, \dots, T$ different time steps (so that $Q = ST$). The POD basis vectors are the m left singular vectors of X corresponding to the largest singular values ($m \leq Q$). Let $\sigma_i, i = 1, 2, \dots, Q$ be the singular values of X in decreasing order. We choose the number of POD basis vectors to retain in the reduced-order model as $m \leq Q$ so that

$$\sum_{i=1}^m \sigma_i^2 / \sum_{j=1}^Q \sigma_j^2 \geq \epsilon_E, \quad (32)$$

where $\epsilon_E(\%)$ is the required amount of energy, typically taken to be 99% or higher.

After obtaining the POD basis vectors for the contaminant, we have defined our reduced-order system. However, as mentioned earlier, the dimension of the system of Equations 23 to 25 is reduced only in state (concentration). The reduced convection matrix $(\mathbf{C}_r(\mathbf{u}))$ in Equation 28 still depends on the full dimension of the velocity field as in Equation 14. This means that we have to re-compute the full convection matrix $(\mathbf{C}(\mathbf{u}))$ at each time-step before projecting it onto the reduced-space basis to obtain the reduced convection matrix $(\mathbf{C}_r(\mathbf{u}))$. Hence, we need an additional treatment to avoid this computational cost.

Linear expansion techniques for velocity-dependent term

We proceed by representing the velocity field as a linear combination of POD velocity basis vectors. The velocity-dependent convection matrix can then be expanded based on this velocity expansion. In particular, let $\mathbf{u}(\mathbf{x}, t)$ be a given velocity field and $\{\mathbf{u}(\mathbf{x}, t_k)\}_{k=1}^T$ be a set of ‘snapshots’ of velocities which are obtained from Equations 5 to 8 and taken over T time steps. The velocity field is decomposed as

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{u}_m(\mathbf{x}) + \mathbf{u}'(\mathbf{x}, t), \quad (33)$$

where $\mathbf{u}_m(\mathbf{x}) = \frac{1}{T} \sum_{k=1}^T \mathbf{u}(\mathbf{x}, t^k)$ is the mean velocity field and $\mathbf{u}'(\mathbf{x}, t)$ is the fluctuating velocity field. The fluctuating velocity field is then represented by

$$\mathbf{u}'(\mathbf{x}, t) = \sum_{k=1}^T \alpha_k(t) \Psi_k(\mathbf{x}), \quad (34)$$

where $\Psi_k(\mathbf{x})$ is the k th POD velocity basis and $\alpha_k(t)$ is the corresponding time dependent amplitude. We now consider the expansion of the velocity field as

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{u}_m(\mathbf{x}) + \sum_{k=1}^{N_u} \alpha_k(t) \Psi_k(\mathbf{x}), \quad (35)$$

where $N_u \ll T$ is the number of POD basis vectors use to represent the velocity.

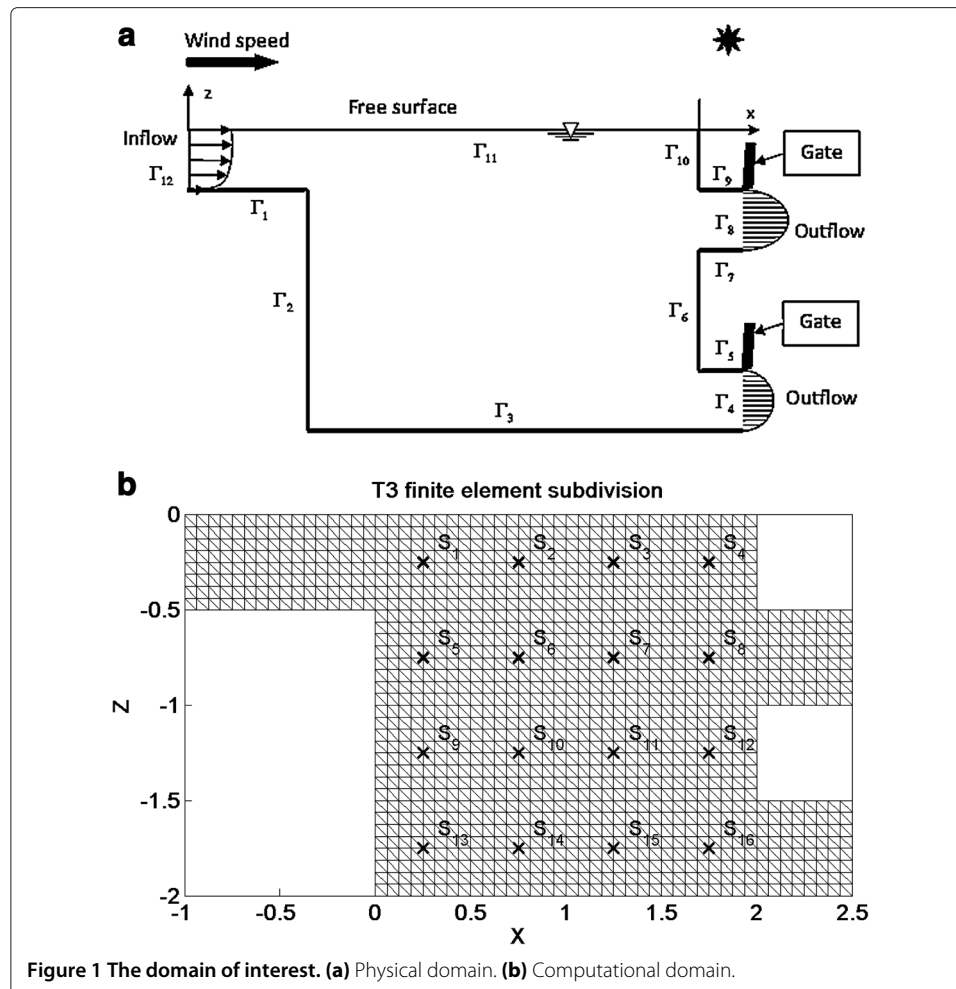


Figure 1 The domain of interest. (a) Physical domain. (b) Computational domain.

Galerkin projection of the time-dependent velocity term in Equation 1 has the form as

$$\begin{aligned} (\mathbf{u} \cdot \nabla c, V_i) &= \left(\left[\mathbf{u}_m + \sum_{k=1}^{N_u} \alpha_k \Psi_k \right] \cdot \nabla c, V_i \right) \\ &= [\mathbf{u}_m \cdot \nabla c, V_i] + \left(\sum_{k=1}^{N_u} \alpha_k \Psi_k \cdot \nabla c, V_i \right). \end{aligned} \quad (36)$$

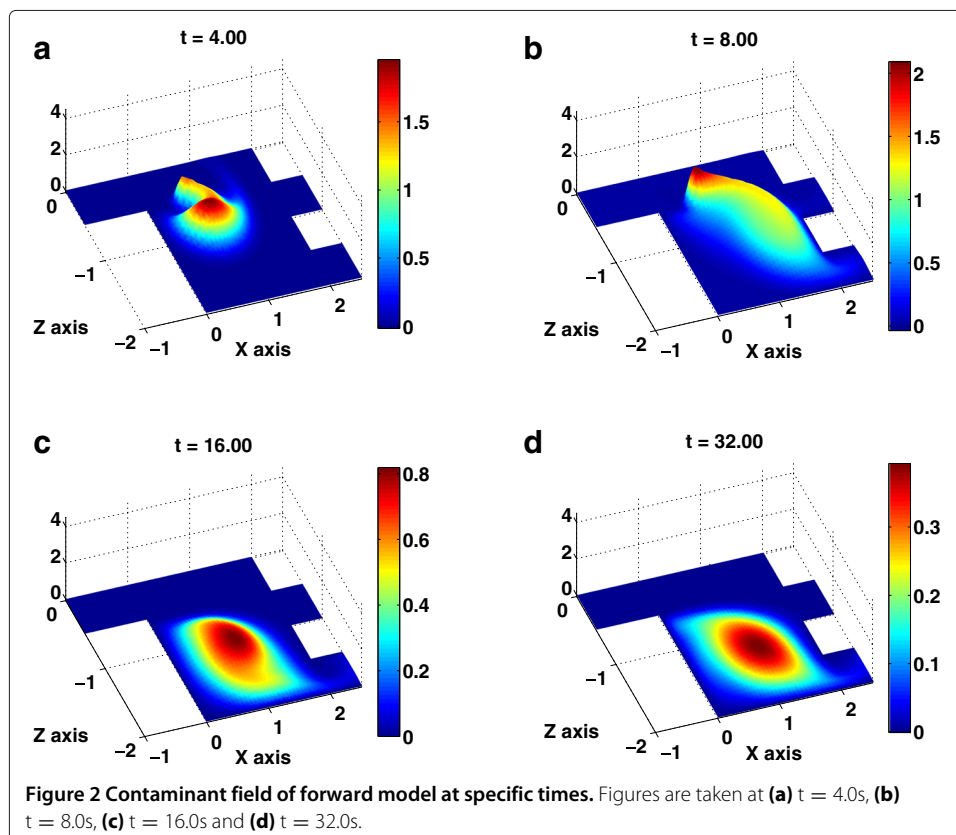
The discrete form of Equation 36 then becomes

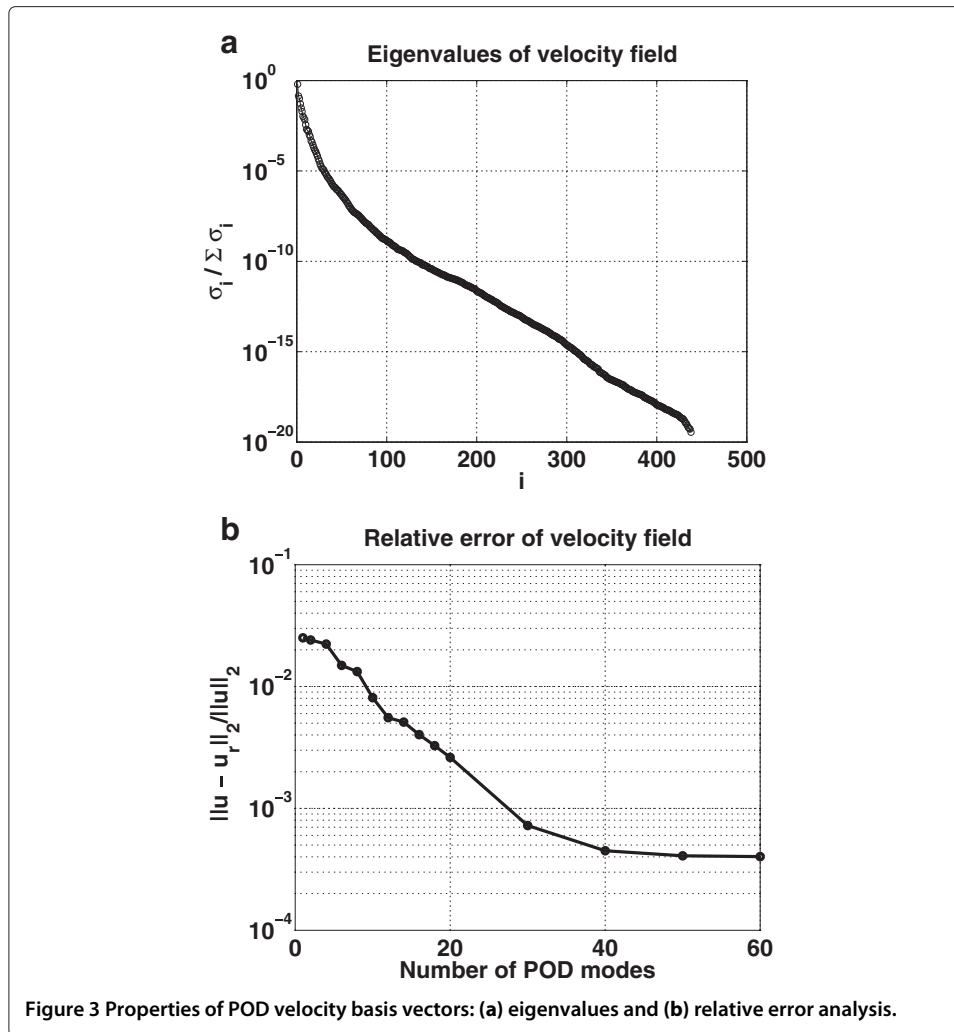
$$\begin{aligned} \mathbf{C}_r(\mathbf{u}) &= [V_i, \mathbf{u}_m \cdot \nabla V_j] + \sum_{k=1}^{N_u} \alpha_k [V_i, \Psi_k \cdot \nabla V_j] \\ &= \mathbf{C}_{rm}(\mathbf{u}_m) + \sum_{k=1}^{N_u} \alpha_k \mathbf{C}_{rk}(\Psi_k). \end{aligned} \quad (37)$$

Here, $\mathbf{C}_{rm}(\mathbf{u}_m(\mathbf{x}))$ and $\mathbf{C}_{rk}(\Psi_k(\mathbf{x}))$, $k = 1, \dots, N_u$ are the reduced-order forms of the full-order convection matrices $\mathbf{C}(\mathbf{u}_m(\mathbf{x}))$ and $\mathbf{C}(\Psi_k(\mathbf{x}))$, respectively. They are computed only once in the offline step. Thus, the first case of the reduced-order model (with Equation 28) is only reduced in state (concentration), but the second case (with Equation 37) is now reduced in both state (concentration) and parameter (velocity).

Error estimation

In order to estimate the efficiency of the reduced model relative to the full model, we use the relative errors of state solutions and relative errors of outputs. These errors are defined as follows:





$$\varepsilon_F(t^k, \theta) = \frac{\|\mathbf{c}(t^k; \theta) - V\mathbf{c}_r(t^k; \theta)\|_{L_2(\mathcal{D})}}{\|\mathbf{c}(t^k; \theta)\|_{L_2(\mathcal{D})}}, \quad (38)$$

$$\varepsilon_s(t^k, \theta) = \frac{\|\mathbf{y}(t^k; \theta) - \mathbf{y}_r(t^k; \theta)\|_{L_2(\mathcal{D})}}{\|\mathbf{y}(t^k; \theta)\|_{L_2(\mathcal{D})}}. \quad (39)$$

Here, $\mathbf{c}(t^k; \theta)$, $\mathbf{c}_r(t^k; \theta)$, $1 \leq k \leq T$ are the full and reduced solutions, while $\mathbf{y}(t^k; \theta)$ and $\mathbf{y}_r(t^k; \theta)$, $1 \leq k \leq T$ are the full and reduced outputs of interest, respectively.

Results and discussion

We consider a numerical example based on a 2D mathematical model. The velocity field in the reservoir is obtained by solving a 2D laterally averaged Navier-Stokes model. The Bayesian formulation of the inverse problem is then solved to determine an uncertain contaminant source location. We compare the effectiveness of solving the inverse problem using both the full model and reduced model techniques.

Model setup

The physical domain is illustrated in Figure 1a, which represents a simplified model of a 2D laterally averaged reservoir system. The reservoir system includes a main reservoir

section and the river connections. We assume that the contaminant is present within the main reservoir section and that the contaminant transport processes are mainly affected by the inflow and wind velocity. We assume that other factors (such as heat exchange, etc.) have little influence on the processes. The spatial domain is discretized with a finite element mesh, as shown in Figure 1b. The total number of grid points is $N = 1,377$ and the total number of elements is $N_E = 2,560$. The simulation is run from initial time $t_0 = 0$ to final time $t_f = 40$, with time-step size $\Delta t = 0.08$. Thus, the number of time steps is $T = 500$.

A time-dependent velocity field is obtained from the 2D laterally averaged system, which is given in Equations 5 to 8, where the body force $\mathbf{f}_c = [f_{cx}; f_{cz}]^T = [0; g]^T$ with g the gravitational acceleration.

The boundary conditions are set up as follows (refer to Figure 1a for the boundary definitions):

$$(u, w) = (1, 0) \quad \text{on } \Gamma_{12}, \quad (40)$$

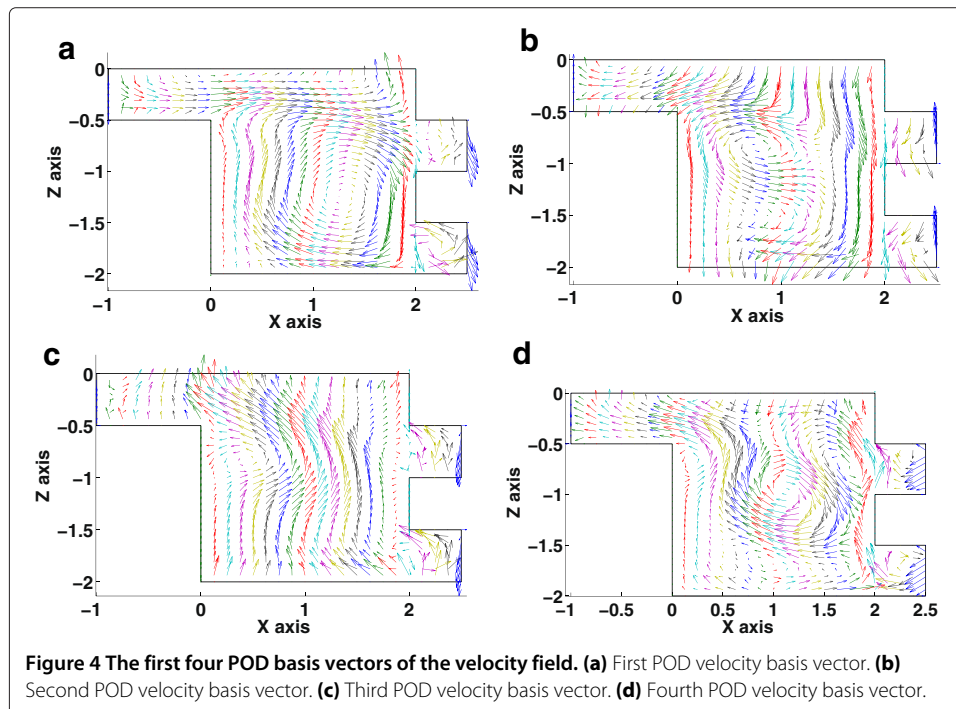
$$(u, w) = (-16 * (2.0 + z) * (1.5 + z), 0) \quad \text{on } \Gamma_4, \quad (41)$$

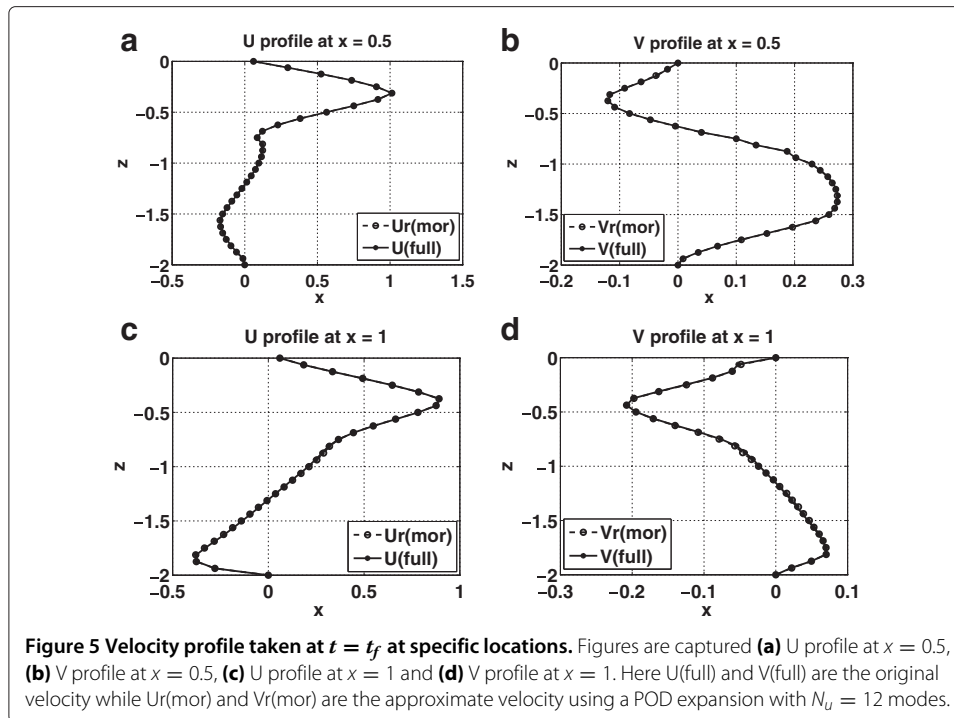
$$(u, w) = (-16 * (1.0 + z) * (0.5 + z), 0) \quad \text{on } \Gamma_8, \quad (42)$$

$$(u, w) = (0.03V_a, 0) \quad \text{on } \Gamma_{11}, \quad (43)$$

$$p = 0 \quad \text{on } \Gamma_{11}. \quad (44)$$

The velocity on remaining boundaries are set to zero. Here, V_a is the wind speed at 10 m above the water surface. In this example, we assumed that $V_a = 2$ m/s for the entire simulation time. The Reynolds number is set to $Re = 5.0e^5$, and a mixing length turbulence model is used [20]. The full model system is given in Equations 9 to 10. The Crank-Nicolson method [21] is used to discretize the system in time.





The setup for the contaminant transport part of the problem is as follows. We use a contaminant source that is defined as the superposition of Gaussian sources, each one active on the time interval $t_{0k} \in [t_0, t_{\text{off}}]$ and centered at $\theta_k \in \mathcal{D}$, with strength h_k and width σ_{sk} :

$$f(\mathbf{x}, t; \theta) = \sum_{k=1}^{n_s} \frac{h_k}{2\pi\sigma_{sk}^2} \exp\left(-\frac{|\theta_k - \mathbf{x}|^2}{2\sigma_{sk}^2}\right) \delta(t - t_{0k}). \quad (45)$$

Here, to simplify the problem, we choose the number of sources to be $n_s = 1$, located at $\theta_1 = (x_c, z_c)$, with strength $h_1 = 0.2$ and width $\sigma_{s1} = 0.05$. The active time of the source is $t_{01} \in [0, t_{\text{off}}]$ with $t_{\text{off}} = 10$. The inflow boundary and other solid boundaries satisfy a homogeneous Dirichlet condition on the contaminant concentration; the outflow boundaries and free surface boundary satisfy a homogeneous Neumann condition. The diffusivity coefficient is assumed to be constant, $\kappa = 0.005$. Thus, the Péclet number $Pe = \frac{\|\mathbf{u}\|L}{\kappa} = 100$, where the length of the inflow section is used as the characteristic

Table 1 The properties of various reduced-order models

| N_s | m | $\epsilon_{\text{ROM-1}}$ | $\epsilon_{\text{ROM-2}}$ | $\frac{t_{\text{full}}}{t_{\text{ROM-1}}}$ | $\frac{t_{\text{full}}}{t_{\text{ROM-2}}}$ |
|-------|-----|---------------------------|---------------------------|--|--|
| 10 | 153 | 3.25e-1 | 4.32e-1 | 0.801 | 111 |
| 20 | 220 | 9.81e-2 | 9.74e-2 | 0.799 | 56 |
| 30 | 257 | 8.03e-3 | 6.75e-3 | 0.795 | 32 |
| 40 | 284 | 7.53e-3 | 7.08e-3 | 0.785 | 26 |
| 50 | 308 | 3.39e-3 | 2.12e-2 | 0.776 | 22 |
| 60 | 328 | 5.83e-3 | 6.08e-3 | 0.767 | 19 |
| 70 | 342 | 1.41e-2 | 1.21e-2 | 0.764 | 18 |
| 80 | 355 | 9.09e-3 | 8.21e-3 | 0.763 | 15 |

length $L = 0.5$. The contaminant concentration is assumed to be zero at initial time $t_0 = 0$.

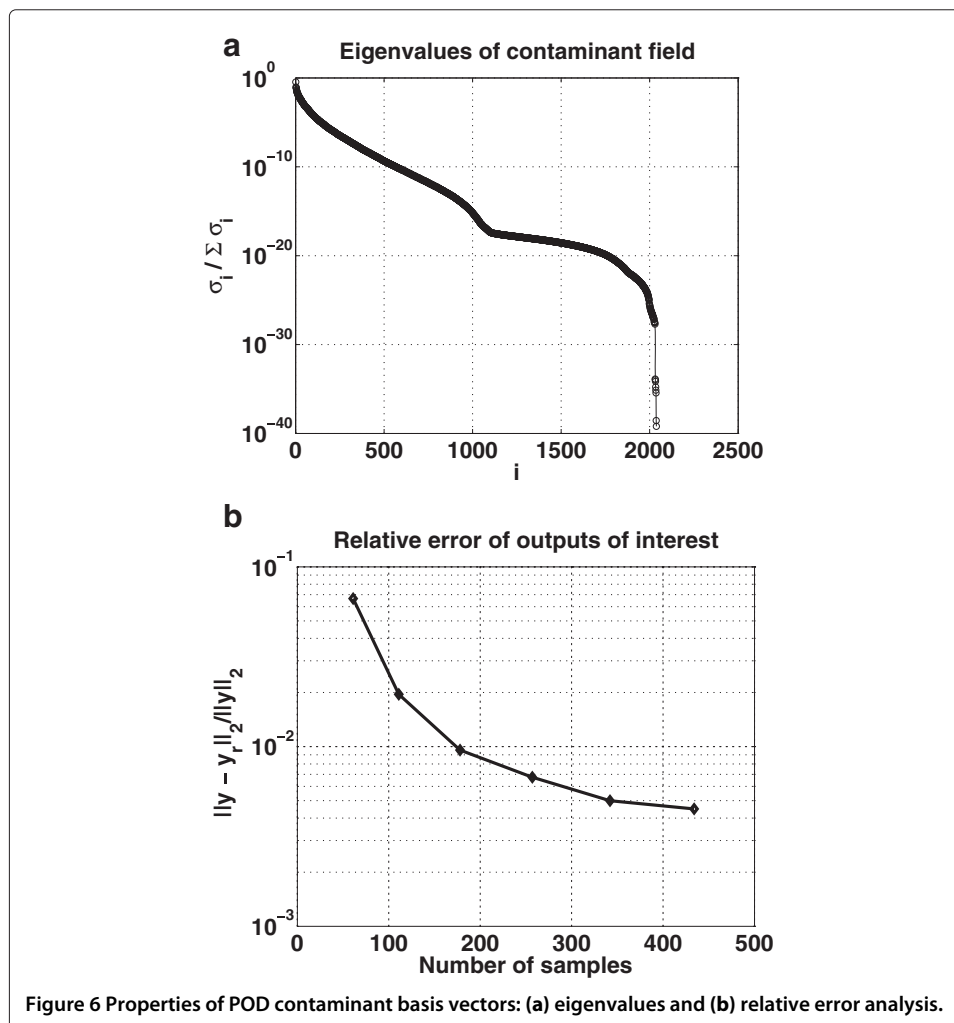
With this setup, we have specified our forward problem with the input parameter $\theta \in \mathbb{R}^2$ in the range $\Theta := [0, 2] \times [-2, 0]$. Figure 2 shows the contaminant solution $c(\mathbf{x}, t; \theta)$ of the forward model with $\theta_1 = (0.5, -0.5)$ at specific times. The contaminant field grows while the source is active. After the shutoff time of the source, the contaminant moves away, spreads out, and decreases in concentration due to convection and diffusion, until it flows out of the domain.

The outputs of interest are the values of contaminant solution $c(\mathbf{x}, t; \theta)$ at selected sensor locations in the computational domain. These sensors are located on a 4×4 uniform grid covering the reservoir domain as shown in Figure 1b.

Model order reduction

To perform the model order reduction, we need to obtain the POD expansion of velocity field.

Figure 3a shows all eigenvalues of the velocity field. As the number of eigenvalues increases, the significant of eigenvalues decreases, respectively. Figure 3b shows the relative errors of representing the velocity field with various numbers of POD velocity modes.



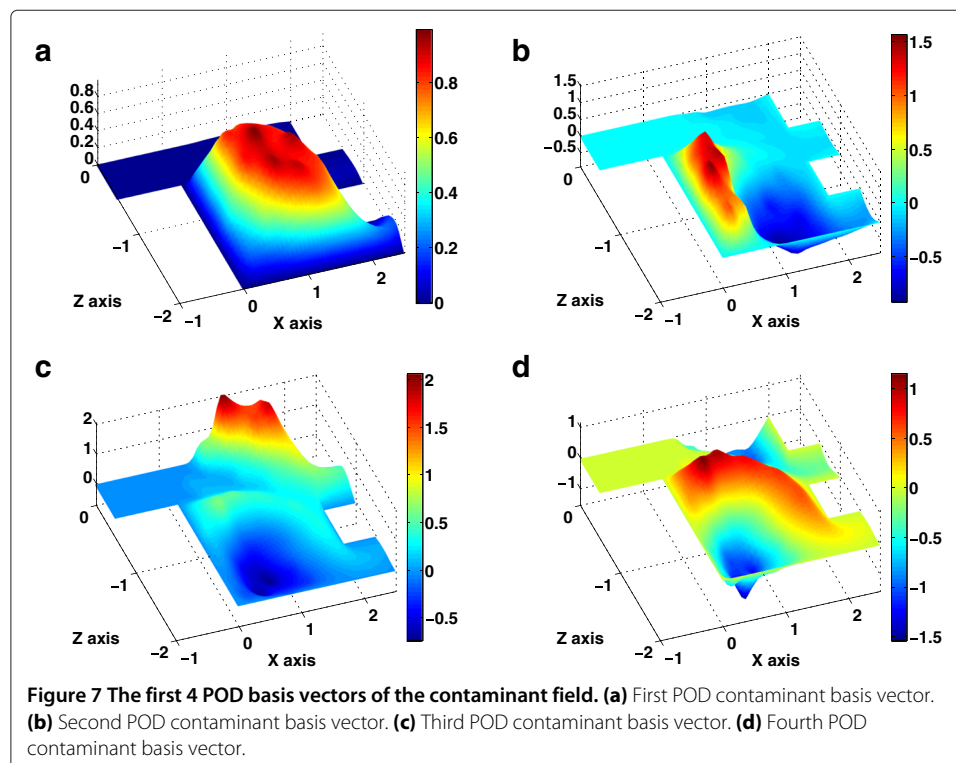
The goal of this step is to find a suitable N_u that can provide a fast computation of the velocity-dependent convection term and to ensure that the relative errors are acceptable. In this example, we choose the number of terms in the velocity expansion to be $N_u = 12$ yields the relative error around 5×10^{-3} .

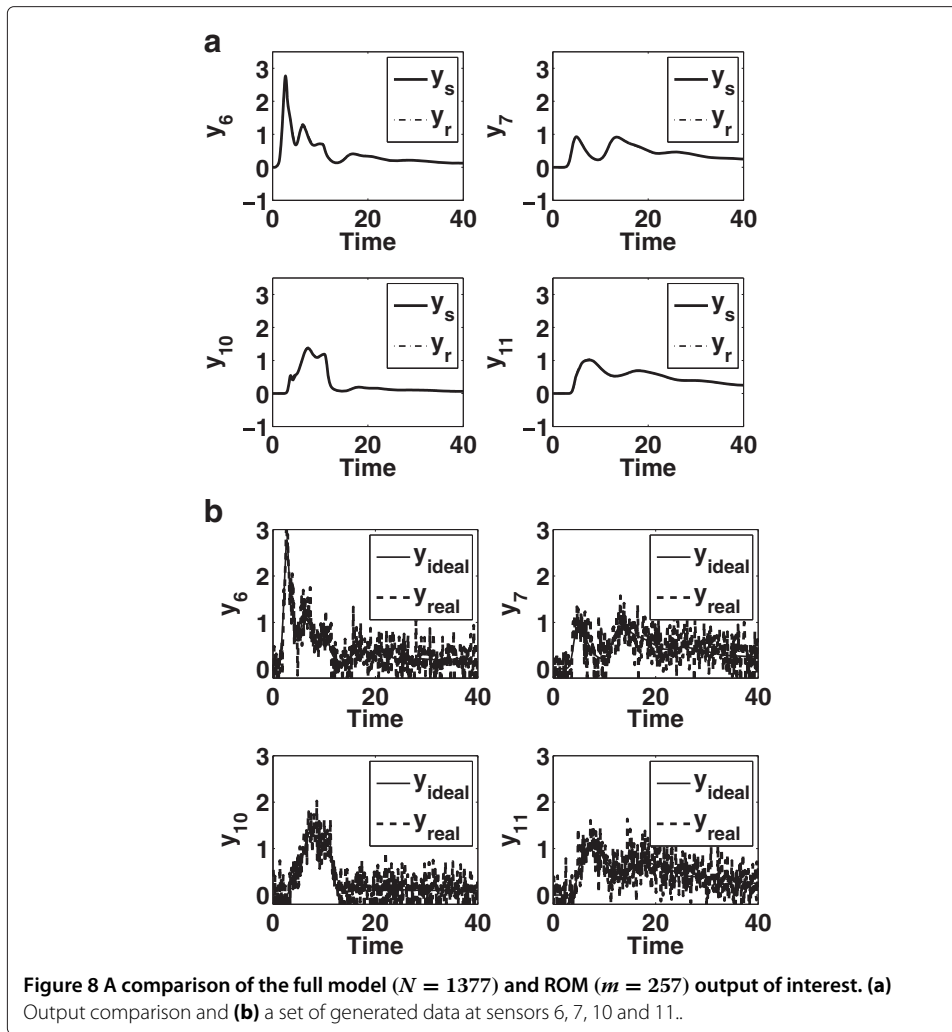
Figure 4 illustrates the first four POD basis vectors of the velocity field. Clearly, they represent the most important characteristics of the velocity field.

A comparison between the original velocity and the approximate velocity representation at final simulating time is given in Figure 5. We observe that the approximate velocity profiles at $x = 0.5$ (Figures 5a,b) and $x = 1.0$ (Figures 5c,d) are able to represent well the characteristics of the original velocity profiles.

To generate the snapshots needed for the POD basis to represent the contaminant concentration, we choose S location samples in the computational domain. In this example, we generate random input locations $\theta \in \mathcal{D} \subset \mathbb{R}^2$ for source term $f(\mathbf{x}, t; \theta)$.

Table 1 shows the relative errors and computational time ratio of the reduced-order models (ROM) for different sample sizes of snapshot matrices. Here, we use the notation ROM-1 to denote the reduced model mentioned in the section ‘Proper orthogonal decomposition’ and ROM-2 for the one mentioned in the section ‘Linear expansion techniques for velocity-dependent term’. The POD basis is chosen based on the energy as in Equation 32, with $\epsilon_E = 99.999\%$. We observe that while the relative errors in the outputs are similar for both cases, the ratio of computational time differs significantly. The case ROM-2 shows speedups from 15 to 111 times, while the case ROM-1 is in fact more expensive to solve than the full forward model ($t_{full} \approx 285$ seconds)¹. This is because the first case consumes a lot of time in re-computing the velocity-dependent convection





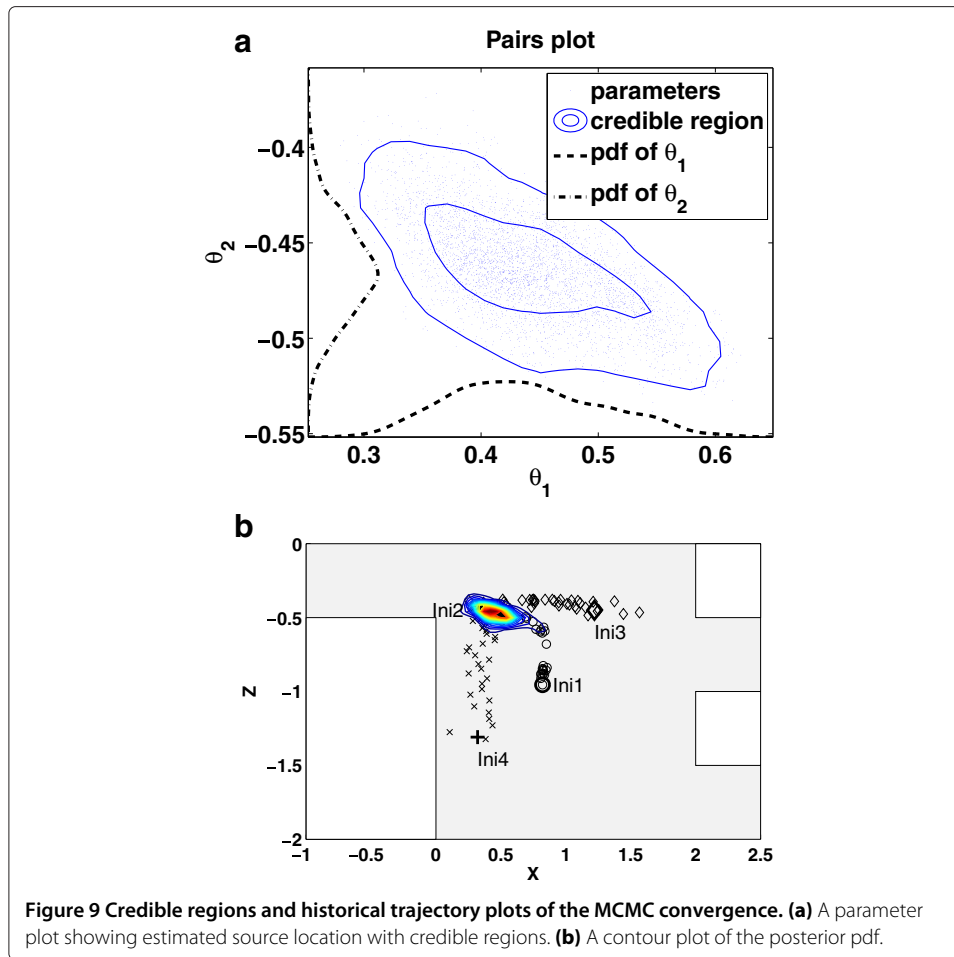
matrix and then projecting it onto the reduced space at each time step. We choose as our reduced-order model the case ROM-2 with sample size $S = 30$.

Figure 6a presents the eigenvalues of contaminant snapshot matrix for the sample test case $S = 30$. We observe that POD eigenvalues of the contaminant field are decayed very fast, and they are significantly important only at the first few hundred values. The first four POD contaminant basis vectors are shown in Figure 7.

Figure 6b shows relative errors of ROMs for different numbers of POD basis vectors. With the energy taken to 99.99999%, the ROM resulted in 434 POD basis vectors with relative error around 4×10^{-3} . In the trade-off between accuracy results and computational

Table 2 Estimated inverse solutions for different numbers of POD basis

| $\epsilon_E(\%)$ | m | θ_1 | θ_2 | ϵ_θ | $t_{mcmc}(s)$ |
|------------------|-----|------------|------------|-------------------|---------------|
| 99.0 | 61 | 0.788 | -0.595 | 5.31e-1 | 1.75e+3 |
| 90.9 | 111 | 0.484 | -0.491 | 6.95e-2 | 2.74e+3 |
| 99.99 | 178 | 0.451 | -0.471 | 1.05e-2 | 7.55e+3 |
| 99.999 | 257 | 0.443 | -0.467 | 2.42e-3 | 8.71e+4 |
| 99.9999 | 343 | 0.446 | -0.466 | 0 | 5.11e+5 |



efforts (please refer to Table 1), we can choose our ROM with 99.999% of energy captured which resulting in the size of $m = 257$ POD basis vectors and an acceptably small error.

Figure 8a shows the outputs of the full model (y) and ROM-2 (y_r) with size $m = 257$ at selected sensor locations. The selected sensors (sensor 6, 7, 10, and 11) capture more information of the contaminant than the others for the case studied. We observe that ROM-2 is able to capture well the behavior of the full model at the sensor locations. This reduced-order model will be utilized as the forward solver in solving the inverse problem.

Inverse problems

MCMC is now used to solve the inverse problem for a variety of source locations using the ROM-2 solver above. By accounting for both measured noise and uncertain information in model parameters, Bayesian inference to inverse problem leads to a well-posed problem resulting in a posterior distribution of the unknown [1,2]. Once obtaining the posterior

Table 3 Estimated inverse solutions for different MCMC starting points

| Initial | θ_1 -ini | θ_2 -ini | θ_1 -M | θ_2 -M | $t_{mcmc}(s)$ |
|---------|-----------------|-----------------|---------------|---------------|---------------|
| Ini1 | 0.821 | -0.955 | 0.451 | -0.467 | 8.71e+4 |
| Ini2 | 0.322 | -0.451 | 0.458 | -0.463 | 4.53e+4 |
| Ini3 | 1.222 | -0.451 | 0.452 | -0.464 | 9.80e+4 |
| Ini4 | 0.323 | -1.313 | 0.456 | -0.465 | 1.05e+5 |

distribution sample, all statistical questions related to the unknown could be answered with sample averages.

To represent the behavior of uncertain variables such as wind velocity, we generate noisy data at sensor locations. The 'real' data (y_{real}) is generated synthetically by adding noise to the ideal data (y_{ideal}) as in Equation 18. The noise is assumed to be Gaussian $\eta \sim N(0, \sigma^2 I)$. Figure 8b shows the noisy data at sensor locations with $\sigma = 0.3$.

We use the snapshots with $S = 30$ samples above to generate several reduced-order models with different energies. To estimate the relative error of the inverse problem solutions (ε_θ), we choose the solution corresponding to the largest order as a 'truth' solution. We do the MCMC simulation with the starting point $\theta_{\text{ini}} = (0.821; -0.955)^T$. The total number of MCMC samples is set to $N_{\text{mcmc}} = 5000$. The initial burn-in period is set to $N_{\text{burnin}} = 500$. After this stage, data is saved to compute summary statistics of source locations.

Table 2 shows the results of the inverse source problems with different numbers of POD basis. For the case with $m = 257$, the computational time is around 24 h to obtain the solution. If we use the full forward solver, with the speedup factor as given in Table 1, the computational time is estimated around 773 h or approximately 32 days.

Figure 9a shows the credible regions of the source locations θ . In this figure, both a pairwise scatter plot and 1D marginal distributions are displayed. The dash line (in θ_1 axis) and dash-dot line (in θ_2 axis) show the probability density function of each parameter while the solid blue regions show the 95% and 99% credible regions in which the parameters appeared the most (the blue dots).

To assess the convergence of the MCMC approach, we generate four different initial points for the MH sampler. Figure 9b shows the contour plot of the posterior probability density function (the contour plot corresponds to the result of Init1) and the historical trajectories of MCMC samplers for each starting point. We observe that for all four initial points, the MCMC simulations converge to the same final solution. However, depending on the position of the initial points, the computational time varies, as given in Table 3.

Conclusion

This study has applied successfully the combination of a model order reduction technique based on the POD and a Bayesian inference approach to solve an inverse problem that seeks to identify an uncertain contaminant location. Applying an additional POD expansion to approximate the velocity results in a reduced-order model in both state (concentration) and parameter (velocity). The resulting reduced-order model is efficient for solution of the forward problem. Solution of the Bayesian formulation of the inverse problem using the reduced-order solver is much more rapid than using the full model, yielding the probability density of the source location in reasonable computational times. The computational time of using a reduced-order model with $m = 257$ degrees of freedom is about a factor of 32 times lower than using the full model with size $N = 1,377$. This reduction is important in real-time water quality management applications because it reduces time cost and storage requirements.

Endnote

¹The simulations were performed on a personal computer (PC) with processor Intel(R) Core(TM)2 Duo CPU E8200 @2.66GHz 2.66GHz, RAM 3.25GB, 32-bit Operating System.

Author details

¹Singapore-MIT Alliance, National University of Singapore, Singapore 117576, Singapore. ²Department of Mechanical Engineering, National University of Singapore, Singapore 117576, Singapore. ³Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA.

Received: 9 October 2013 Accepted: 23 November 2013

Published: 29 April 2014

References

1. Tarantola A (2004) Inverse problem theory and methods for model parameter estimation. SIAM, Philadelphia
2. Sivia DS, Skilling J (2006) Data analysis: a Bayesian tutorial, Second edition. Oxford University Press Inc., New York
3. Jackson C, Sen MK, Stoffa PL (2004) An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions. *J Climate* 17: 2828–2841
4. Snodgrass MF, Kitanidis PK (1997) A geostatistical approach to contaminant source identification. *Water Resour Res* 34(4): 537–546
5. Woodbury AD, Ulrych TJ (2000) A full-Bayesian approach to the groundwater inverse problem for steady state flow. *Water Resour Res* 36(1): 159–171
6. Marzouk YM, Najm HN, Rahn LA (2007) Stochastic spectral methods for efficient Bayesian solution of inverse problems. *J Comput Phys* 224(2): 560–586
7. Wang J, Zabaras N (2005) Hierarchical Bayesian models for inverse problems in heat conduction. *Inverse Probl* 21: 183–206
8. Metropolis N, Ulam S (1949) The Monte Carlo method. *J Am Stat Assoc* 44: 335–341
9. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. *J Chem Phys* 21: 1087–1092
10. Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97–109
11. Gelman A, Carlin J, Stern H, Rubin D (2003) Bayesian data analysis. Second Edition, Chapman & Hall/CRC Texts in Statistical Science
12. Zienkiewicz OC, Morgan K (1983) Finite elements and approximation. Wiley, New York
13. Antoulas AC (2005) Approximation of large-scale dynamical systems. SIAM Advances in Design and Control, Philadelphia
14. Lieberman C, Willcox K, Ghattas O (2010) Parameter and state model reduction for large-scale statistical inverse problems. *SIAM J Sci Comput* 32(5): 2523–2542
15. Lieberman C, Willcox K (2012) Goal-oriented inference: approach, linear theory, and application to advection-diffusion. *SIAM J Sci Comput* 34(4): 1880–1904
16. Sirovich L (1987) Turbulence and the dynamics of coherent structures. Part 1: coherent structures. *Q Appl Math* 45(3): 561–571
17. Codina R, Blasco J (2000) Stabilized finite element method for the transient Navier-Stokes equations based on a pressure gradient projection. *Comput Methods Appl Mech Engrg* 182: 277–300
18. Ma X, Zabaras N (2008) A stabilized stochastic finite element second-order projection method for modeling natural convection in random porous media. *J Comput Phys* 227: 8448–8471
19. Jaynes ET (1957) Information theory and statistical mechanics. *Phys Rev* 104(4): 620–630
20. Wilcox CD (1993) Turbulence Modeling for CFD, DCW Industries, La Cañada, California
21. Crank J, Nicolson P (1996) A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Adv Comput Math* 6(1): 207–226

doi:10.1186/2196-1166-1-2

Cite this article as: Nguyen et al.: Model order reduction for Bayesian approach to inverse problems. *Asia Pacific Journal on Computational Engineering* 2014 **1**:2.

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com