Computer Science Technical Report

CSTR-1/2016

January 5, 2016

Ahmed Attia, Razvan Ştefănescu, and Adrian Sandu

“The Reduced-Order Hybrid Monte Carlo Sampling Smoother”

Computational Science Laboratory
Computer Science Department
Virginia Polytechnic Institute and State University
Blacksburg, VA 24060
Phone: (540)-231-2193
Fax: (540)-231-6075
Email: attia@vt.edu, sandu@cs.vt.edu
Web: http://csl.cs.vt.edu

Computational Science Laboratory
Innovative Computational Solutions

VirginiaTech
Invent the Future

The Reduced-Order Hybrid Monte Carlo Sampling Smoother

A. Attia*, R. Ştefănescu, and A. Sandu

Computational Science Laboratory,
Department of Computer Science,
Virginia Polytechnic Institute and State University,
Blacksburg, Virginia, 24060, USA

*Corresponding author
Email address: attia@vt.edu,rstefane@vt.edu, sandu@cs.vt.edu (A. Attia*, R. Ştefănescu, and A. Sandu)

The Reduced-Order Hybrid Monte Carlo Sampling Smoother

A. Attia*, R. Ştefănescu, and A. Sandu

Computational Science Laboratory,
Department of Computer Science,
Virginia Polytechnic Institute and State University,
Blacksburg, Virginia, 24060, USA

Abstract

Hybrid Monte-Carlo (HMC) sampling smoother is a fully non-Gaussian four-dimensional data assimilation algorithm that works by directly sampling the posterior distribution formulated in the Bayesian framework. The smoother in its original formulation is computationally expensive due to the intrinsic requirement of running the forward and adjoint models repeatedly. Here we present computationally efficient versions of the HMC sampling smoother based on reduced-order approximations of the underlying model dynamics. The schemes developed herein are tested numerically using the shallow-water equations model on Cartesian coordinates. The results reveal that the reduced-order versions of the smoother are capable of accurately capturing the posterior probability density, while being significantly faster than the original full order formulation.

Keywords: Data Assimilation, Hamiltonian Monte-Carlo, Smoothing, Reduced-Order Modeling, Proper Orthogonal Decomposition.

Contents

*Corresponding author
Email address: `attia@vt.edu,rstefane@vt.edu, sandu@cs.vt.edu` (A. Attia*, R. Ştefănescu, and A. Sandu)

1. Introduction

Many large-scale prediction problems such as atmospheric forecasting are formulated as initial value problems. The uncertainty of the associated model initial conditions can be decreased by combining imperfect forecasts produced by propagating the model dynamics with real measurements collected at discrete times over an assimilation window. The model state and the observations are both uncertain and can be modeled as random variables. The probability distribution describing the knowledge about the initial state system, before incorporating observations information, is known as the prior distribution. The likelihood function accounts for the discrepancies between the measurements and model-predicted observations. Data assimilation applies Bayes' theorem to obtain a posterior probability distribution, named the analysis, that characterizes the knowledge about the system state given the observed measurements. In practice, due to the state space high-dimensionality of many realistic models, it is impossible to exactly describe the posterior distribution and several assumptions and approximations are necessary. Widely accepted assumptions are that the background and the observation errors are characterized by Gaussian distributions, with no correlations between observation errors at different time instances.

Two families of methodologies are generally followed in order to generate accurate estimates of the true system state. Ensemble-based statistical methods seek to approximate the posterior probability density function (PDF) based on an ensemble of model states, while the variational approaches estimate the true state of the system by searching for the state that maximizes the posterior PDF. Among the variational techniques, the 4D-Var method achieves this goal by searching for a local minimum of an objective function corresponding to the negative logarithm of the posterior distribution. 4D-Var finds the maximum posterior (MAP) estimate of the true state and does not directly estimate the uncertainty associated with the analysis state. For scenarios including nonlinear observation operators and state models, the analysis distribution is not Gaussian, and the 4D-Var algorithm may be trapped in a local minimum of the cost function leading to incorrect conclusions.

Accurate solution of the non-Gaussian data assimilation problems requires accounting for all regions of high probability in the posterior distribution. The recently developed hybrid Monte-Carlo (HMC) sampling smoother [4, 5] is a four dimensional data assimilation scheme designed to solve the non-Gaussian smoothing problem by sampling from the posterior distribution. It relies on an accelerated Markov chain Monte-Carlo (MCMC) methodology where a Hamiltonian system is used to formulate proposal densities. Two issues are important when using HMC sampling strategy. First, it requires the formulation of the posterior negative-log function gradient. Secondly, the involved Hamiltonian system is propagated using a symplectic integrator whose parameters must be carefully tuned in order to achieve good performance.

While producing consistent description of the updated system uncertainty (e.g., the analysis error covariance matrix), the original formulation of the HMC sampling smoother is computationally expensive when compared to the 4D-Var approach. This is due to the large number of gradient evaluations required, which translates into many forward and adjoint models runs. In the case of large scale problems the computational cost becomes prohibitive. In this present study we propose a practical solution by approximating the gradient using information obtained from lower-dimensional subspaces via model reduction.

Reduced order modeling refers to the development of low-dimensional systems that represent the important characteristics of a high-dimensional or infinite dimensional dynamical system. Typically this is achieved by projecting model dynamics onto a lower dimensional spaces. Construction of low relevant manifolds can be achieved using the reduced basis method [8, 26, 43, 47, 18, 36], dynamic mode decomposition [46, 52, 62, 10] and Proper Orthogonal Decomposition (POD) [32, 37, 30, 38]. The latest is the most prevalent basis selection method for nonlinear problems. Data analysis using POD and method of snapshots [54, 55, 56] is conducted to extract basis functions, from experimental data or detailed simulations of high-dimensional systems, for subsequent use in Galerkin projections that yield low dimensional dynamical models. By coupling POD with empirical interpolation method (EIM) [8], discrete variant DEIM [14, 16, 15, 20] or best points interpolation method [41], one can obtain fast approximations of reduced order nonlinear terms independent of the dimension of high-fidelity space. Other such approaches include missing point estimation [3] and Gauss-Newton with approximated tensors [12, 13] relying upon the gappy POD technique [23].

The present manuscript develops practical versions of the HMC sampling smoother using approximate dynamical information obtained via POD/DEIM reduced order models [57, 59, 19]. Two reduced order variants are proposed differentiated by the choice of the sampling space used to generate the proposals. In the first case we are sampling in a reduced order space while the second approach samples directly from the high fidelity space. For both scenarios the negative logarithm of the posterior distribution reassembles one of the flavors of reduced order 4D-Var objective functions [24, 11, 63], where the prior term is either estimated in the reduced or full space respectively. While reduced order models are employed to estimate the posterior distribution negative logarithm and it's gradient, the associated Hamiltonian system generates proposals in a reduced or high-fidelity space depending on the nature of the sampling space.

The choice of reduced order manifolds is crucial for the accuracy of the reduced samplers. The smoothers may suffer from the fact that the basis elements are computed from a reference trajectory containing features which are quite different from those of the current proposal trajectory. To overcome the problem of unmodelled dynamics in the POD-basis we propose to update the basis from time to time according to the current state similarly as in the case of adaptive reduced order optimization [1, 45, 35]. Inspired by the recent advances in reduced order data assimilation field where it was shown that accurate reduced order Karush-Kuhn-Tucker conditions with respect to their full order counterparts highly increase the accuracy of the obtained analysis [61], we chose to update the reduced order manifolds based on high-fidelity forward and adjoint trajectories corresponding to current proposal as well as the associated gradient of the negative logarithm of the high-fidelity posterior distribution. The basis is refreshed once after several HMC smoothers iterations. The numerical results using swallow-water equations model on cartesian coordinates reveal that the reduced-order versions of the smoother are accurately capturing the posterior probability density, while being significantly faster than the original full order formulation.

The paper is organized as follows. Section 2 reviews the four-dimensional data assimilation problem and the original HMC smoother. Section 3 reviews reduced-order modeling, and introduces the reduced order 4D-Var data assimilation framework. Section 4 presents a rigorous description of the proposed versions of the HMC smoother. In section 5 several theoretical properties of the distributions sampled with reduced order models are derived. Numerical results are presented in Section 6 while conclusions are drawn in Section 7.

2. Data Assimilation

One can solve a data assimilation (DA) problem by describing the posterior distribution in the Bayesian formalism given the prior and the likelihood function. In the four-dimensional DA (4DDA) context, the main goal is to describe the posterior distribution of system state at the initial time of a specific assimilation window. The assimilation window is defined based on the availability of observations at specific discrete time instances. The knowledge about the system at the initial time t_0 defines the distribution (the prior) $\mathcal{P}^{\mathrm{b}}(\mathbf{x}_0)$ of the state $\mathbf{x}_0 \in \mathbb{R}^{\mathrm{N_{VAR}}}$

before absorbing knowledge presented by the observations captured over the assimilation window. Based on a set of observations $\{\mathbf{y}_k = \mathbf{y}[t_k] \in \mathbb{R}^m\}_{k=0,1,\dots,\textsc{Nobs}}$, at the discrete time points $\{t_k\}_{k=0,1,\dots,\textsc{Nobs}}$ in the interval $[t_0\,, t_F]$, the sampling distribution (likelihood function) is defined as $\mathcal{P}(\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{\textsc{Nobs}}|\mathbf{x}_0)$, and the posterior distribution describing the updated information about the model state at the initial time, given the measurements, takes the general form

$$\mathcal{P}(\mathbf{x}_0|\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{\textsc{Nobs}}) = \frac{\mathcal{P}^{\rm b}(\mathbf{x}_0)\,\mathcal{P}(\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{\textsc{Nobs}}|\mathbf{x}_0))}{\mathcal{P}(\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{\textsc{Nobs}})}\,. \tag{1}$$

Fully and accurately describing this general posterior probability density function in DA literature is an intractable problem, and usually simplifying assumptions are generally made. As we mentioned in Section 1, a frequent supposition is that the background and observation errors are characterized by Gaussian distributions. If the observations are assumed to be independent from the model states, and the associated error characteristics of these observations are not correlated in time, the posterior distribution takes the form

$$\mathcal{P}^{\rm a}(\mathbf{x}_0) = \mathcal{P}(\mathbf{x}_0|\mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{\textsc{Nobs}}) \propto \exp\left(-\mathcal{J}(\mathbf{x}_0)\right), \tag{2a}$$

$$\mathcal{J} : \mathbb{R}^{\textsc{Nvar}} \to \mathbb{R}, \quad \mathcal{J}(\mathbf{x}_0) = \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}_0^{\rm b}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\sum_{k=0}^{\textsc{Nobs}} \|\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}}\,, \tag{2b}$$

where $\mathbf{x}_0^{\rm b} = \mathbf{x}^{\rm b}[t_0]$ is a background/forecast state, the matrices $\mathbf{B}_0$ and $\mathbf{R}_k,\ k = 0, 1, \dots, \textsc{Nobs}$, are the covariance matrices associated with the background and measurement errors respectively. The observation operator $\mathcal{H}_k : \mathbb{R}^{\textsc{Nvar}} \to \mathbb{R}^m$, at time instance t_k, maps a given state $\mathbf{x}_k \in \mathbb{R}^{\textsc{Nvar}}$ to the observation space. The dimension of the observation space is usually much smaller than the size of the state space, that is $m \ll \textsc{Nvar}$. The associated norms over the state and observation spaces are defined as:

$$\|\mathbf{a} - \mathbf{b}\|^2_{\rm C} = (\mathbf{a} - \mathbf{b})^T \mathbf{C}(\mathbf{a} - \mathbf{b}), \tag{3}$$

where $\mathbf{a},\ \mathbf{b}$ belong to either $\mathbb{R}^{\textsc{Nvar}}$ or $\mathbb{R}^m$ and $\mathbf{C}$ is a matrix of $\mathbb{R}^{\textsc{Nvar}\times\textsc{Nvar}}$ or $\mathbb{R}^{m\times m}$ dimensions.

The model state $\mathbf{x}_k = \mathbf{x}[t_k]$, is obtained from the initial state $\mathbf{x}_0$, by propagating the model dynamics to time point t_k, i.e.

$$\mathbf{x}_k = \mathcal{M}_{t_0 \to t_k}(\mathbf{x}_0) = \mathcal{M}_{0,k}(\mathbf{x}_0)\,, \tag{4}$$

where, $\mathcal{M}_{t_0 \to t_k} : \mathbb{R}^{\textsc{Nvar}} \to \mathbb{R}^{\textsc{Nvar}},\ k = 1, \dots, \textsc{Nobs}$ represents the discretized mathematical model reflecting the state dynamics. The function $\mathcal{J}$ given in (2) is quadratic, and consequently the distribution (2) is Gaussian distribution, only if the states $\mathbf{x}_k$ are linearly related to the system initial condition $\mathbf{x}_0$, and the observations $\mathbf{y}_k$ are linearly related to the model states $\mathbf{x}_k$. In virtually all practical settings, the time dependent models are complicated resulting in a nonlinear bond between states at different time instances. More difficulty is added since for example, in the field of atmospheric sciences, highly-nonlinear observation operators are often constructed to relate new types of measurements to state variables.

2.1. 4D-Var data assimilation

The strongly-constrained 4D-Var DA scheme is an optimization algorithm that searches for the maximum aposteriori estimate (MAP) by seeking a *local* minimizer of the cost function (2b), constrained by the model equation (4). Precisely, the constrained optimization problem is defined as

$$\begin{aligned} \min_{\mathbf{x}_0} \mathcal{J}(\mathbf{x}_0) &= \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\sum_{k=0}^{\textsc{Nobs}} \|\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}}\,, \\ \mathbf{x}_k &= \mathcal{M}_{t_0\rightarrow t_k}(\mathbf{x}_0) \;\; k = 1,\dots,\textsc{Nobs}. \end{aligned} \tag{5}$$

Gradient-based schemes are generally employed to find the corresponding initial conditions which in the control theory language are known as control variables. Using the Lagrange multiplier technique the constrained optimization problem (5) is redesigned as an unconstrained minimization problem with the associated first order necessary optimality conditions:

$$\begin{aligned} &\textit{Forward model}: && \mathbf{x}_k = \mathcal{M}_{t_0\rightarrow t_k}(\mathbf{x}_0),\; k = 1,\dots,\textsc{Nobs}, && \text{(6a)} \\ &\textit{Adjoint model}: && \boldsymbol{\lambda}_{\textsc{Nobs}} = \mathbf{H}^T_{\textsc{Nobs}}\mathbf{R}^{-1}_{\textsc{Nobs}}\left(\mathbf{y}_{\textsc{Nobs}} - \mathcal{H}_{\textsc{Nobs}}(\mathbf{x}_{\textsc{Nobs}})\right), && \text{(6b)} \\ & && \boldsymbol{\lambda}_k = \mathbf{M}^T_{t_0\rightarrow t_{k+1}}\boldsymbol{\lambda}_{k+1} + \mathbf{H}_k^T\mathbf{R}_k^{-1}\left(\mathbf{y}_k - \mathcal{H}(\mathbf{x}_k)\right), \quad k = \textsc{Nobs}-1,..,0, && \text{(6c)} \\ &\textit{Cost function gradient:} && \nabla_{\mathbf{x}_0}\mathcal{J}(\mathbf{x}_0) = -\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\mathrm{b}} - \mathbf{x}_0\right) - \boldsymbol{\lambda}_0 = 0. && \text{(6d)} \end{aligned}$$

The Jacobians of the model and observation operators are denoted by $\mathbf{M}_{t_0\rightarrow t_{k+1}},\; k = 0,1,\dots,\textsc{Nobs}-1$ and $\mathbf{H}_k,\; k = 0,1,\dots,\textsc{Nobs}$ while the adjoint solution $\boldsymbol{\lambda}_k \in \mathbb{R}^{\textsc{Nvar}},\; k = 0,1,\dots,\textsc{Nobs}$, provides an efficient way to compute the gradient (6d). The nonlinear optimization procedure is computationally expensive and either low-resolution models (incremental 4D-Var [58]), or alternatively reduced-order models are used [61] to alleviate this drawback. It is well-known that 4D-Var does not inherently provide a measure of the uncertainty about the updated state (e.g., analysis error covariance matrix) and usually hybrid methods are considered to account for this type of information. This approach results in some inconsistency between the analysis state and the analysis error covariance matrix especially when they are obtained using different algorithms.

2.2. Smoothing by sampling and the HMC sampling smoother

Monte-Carlo smoothing refers to the process of representing/approximating the posterior distribution (2) using an ensemble of model states sampled from that posterior. Ensemble Kalman smoother (EnKS) [22] is an extension of the well-known ensemble Kalman Filter [31] to the case where observation are assimilated simultaneously. EnKS produces a minimum-variance unbiased estimate (MVUE) of the system state by estimating the expectation of the posterior $\mathbb{E}_{\mathcal{P}^{\mathrm{a}}}[\mathbf{x}_0]$ using the mean of an ensemble of states. The strict Gaussianity and linearity assumptions imposed by EnKS, usually result in poor performance of the smoother.

Pure sampling of the posterior distribution (1) using a Markov Chain Monte-Carlo (MCMC) [39, 40] technique is known in theory to provide more accurate estimates without strictly imposing linearity or Gaussianity constraints. MCMC is a family of Monte-Carlo schemes tailored to sample a given distribution (up to a proportionality constant), by constructing a Markov chain whose stationary distribution is set as the target distribution. By design, an MCMC sampler is guaranteed to converge to its stationarity. However, the choice of the proposal density, the convergence rate, the acceptance rate, and the correlation level among sampled points are the main building blocks of MCMC responsible for its performance and efficiency. Practical application of MCMC requires developing accelerated chains those can attain stationarity fast, and then explore the state space efficiently in very few steps. One of the MCMC samplers mainly designed for complicated PDFs and large dimensional spaces is the Hamiltonian/Hybrid Monte-Carlo sampler (HMC). HMC was firstly presented in [21] as an accelerated MCMC sampling algorithm. The sampler mainly uses information about the geometry of the posterior to guide its steps in order to avoid random walk behaviour and visit more frequently regions with high probability with the capability of jumping between separated modes of the target PDF.

HMC sampling. As all other MCMC samplers, HMC samples from a PDF

$$\pi(\mathbf{x}) \propto \exp\left(-\mathcal{J}^N(\mathbf{x})\right);\ \mathbf{x} \in \mathbf{R}^{\mathrm{Nvar}}\,, \tag{7}$$

where $\exp\left(-\mathcal{J}^N(\mathbf{x})\right)$ is the shape function of the distribution, and $\mathcal{J}^N : \mathbb{R}^{\mathrm{Nvar}} \to \mathbb{R}$ is the PDF negative-log. The power of MCMC, and consequently HMC, is that only the shape function, or alternatively the negative-log, is needed, while the scaling factor is not strictly required as in the case of standard application of Bayes' theorem. HMC works by viewing $\mathbf{x}$ as a position variable in an extended phase space consisting of points $(\mathbf{p}, \mathbf{x}) \in \mathbf{R}^{2\mathrm{Nvar}}$, where $\mathbf{p} \in \mathbb{R}^{\mathrm{Nvar}}$ is an auxiliary momentum variable. The Hamiltonian dynamics is modeled by the set of ordinary differential equations (ODEs):

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \nabla_{\mathbf{p}} H\,, \\ \frac{d\mathbf{p}}{dt} &= -\nabla_{\mathbf{x}} H\,, \end{aligned} \tag{8}$$

where $H = H(\mathbf{p}, \mathbf{x})$ is the constant total energy function of the system, known as the the Hamiltonian function or simply the Hamiltonian. A standard formulation of the Hamiltonian in the context of HMC is:

$$H(\mathbf{p}, \mathbf{x}) = \underbrace{\frac{1}{2}\,\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}}_{\text{kinetic energy}} + \underbrace{\mathcal{J}^N(\mathbf{x})}_{\text{potential energy}} \quad , \tag{9}$$

where $\mathbf{M} \in \mathbb{R}^{\mathrm{N_{VAR}}\times\mathrm{N_{VAR}}}$ is a positive definite matrix known as the mass matrix. This particular formulation leads to a canonical distribution of the joint state ($\mathbf{p}$, $\mathbf{x}$) proportional to:

$$\exp\left(-H(\mathbf{p},\, \mathbf{x})\right) = \exp\left(-\frac{1}{2}\,\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}\right)\, \pi(\mathbf{x})\,. \tag{10}$$

The exact flow $\Phi_T : \mathbb{R}^{2\mathrm{N_{VAR}}} \rightarrow \mathbb{R}^{2\mathrm{N_{VAR}}};\quad \Phi_T\big(\mathbf{p}[0], \mathbf{x}[0]\big) = \big(\mathbf{p}[T], \mathbf{x}[T]\big)$ describes the time evolution of the Hamiltonian system (8) and is practically approximated using a numerical integrator that is symplectic as $\phi_T : \mathbb{R}^{2\mathrm{N_{VAR}}} \rightarrow \mathbb{R}^{2\mathrm{N_{VAR}}};\quad \phi_T\big(\mathbf{p}[0], \mathbf{x}[0]\big) \approx \big(\mathbf{p}[T], \mathbf{x}[T]\big)$. The use of a symplectic numerical integrator to approximate the exact Hamiltonian flow results in changes of total energy. Traditional wisdom recommends splitting the pseudo-time step T into m smaller steps of size h in order to simulate the Hamiltonian trajectory between points $\big(\mathbf{p}[0], \mathbf{x}[0]\big)$, $\big(\mathbf{p}[T], \mathbf{x}[T]\big)$ more accurately. The symplectic integrator of choice is position (or velocity) Verlet. New higher order symplectic integrators were proposed recently and tested in the context of filtering for data assimilation [50, 6]. One step of size h of the position Verlet [51, 50] integrator is describes as follows

$$\mathbf{x}[h/2] = \mathbf{x}[0] + \frac{h}{2}\,\mathbf{M}^{-1}\,\mathbf{p}[0]\,, \tag{11a}$$

$$\mathbf{p}[h] = \mathbf{p}[0] - h\,\nabla_{\mathbf{x}}\mathcal{J}(\mathbf{x}[h/2])\,, \tag{11b}$$

$$\mathbf{x}[h] = \mathbf{x}[h/2] + \frac{h}{2}\,\mathbf{M}^{-1}\,\mathbf{p}[h]. \tag{11c}$$

While, the mass matrix is a user-defined parameter, it can be designed to enhance the performance the sampler [6]. The step parameters of the symplectic integrator $m,\ h$ can be empirically chosen by monitoring the acceptance rate in a preprocessing step. Specifically, the parameters of the Hamiltonian trajectory can be empirically adjusted such as to achieve a specific rejection rate. Generally speaking, the step size should be chosen to achieve a rejection rate between 25% and 30%, and the number of steps should generally be large [40].

Adaptive versions of HMC have been also proposed with the capability of adjusting it's step parameters. No-U-Turn sampler (NUTS) [29] is a version of HMC capable of automatically tuning its parameters to prohibit the sampler from retracing its steps along the constructed Hamiltonian trajectory. Another HMC sampler that tunes its parameters automatically using third-order derivative information is the Riemann manifold HMC (RMHMC) [25].

The intuition behind HMC sampler is to build a Markov chain whose stationary distribution is defined by the canonical PDF (10). In each step of the chain, a random momentum $\mathbf{p}$ is drawn from a Gaussian distribution $\mathcal{N}(0, \mathbf{M})$, and the Hamiltonian dynamics (8) at the final pseudo-time interval proposes a new point that is either accepted or rejected using a Metropolis-Hastings criterion. The two variables $\mathbf{p}$, and $\mathbf{x}$ are independent, so discarding the momentum generated at each step will leave us with sample points $\mathbf{x}$ generated from our target distribution.

In a previous work [5] we proposed using HMC as a pure sampling smoother to solve the nonlinear 4DDA smoothing problem. The method samples from the posterior distribution of the model state at the initial time on an assimilation window on which a set of observations are given at discrete times. Following general assumptions where the prior is Gaussian and the observation errors are normally distributed, the target distribution defined in (7) is identical with the posterior distribution associated with the smoother problem (2). Consequently the PDF negative-log $\mathcal{J}^N$ in (7) resembles the 4D-Var cost function $\mathcal{J}$ defined in (5) and the gradient of the potential energy required by the symplectic integrator is the gradient of the 4D-Var cost functional (6d). The main hindrance stems from the requirement of HMC to evaluate the gradient of the potential energy (target PDF negative-log) at least as many times as the symplectic integrator is involved, which is an expensive process. Despite the associated computational overhead, the numerical results presented in [5] show the potential of using HMC smoother to sample multi-modal, high-dimensional posterior distributions formulated in the smoothing problem.

3. Four-Dimensional Variational Data Assimilation with Reduced-Order Models

Optimization problems such as the one described in (5) for nonlinear partial differential equations often demand very large computational resources, so that the need for developing fast novel approaches emerges. Recently the reduced order approach applied to optimal control problems for partial differential equations has received increasing attention. The main idea is to project the dynamical system onto subspaces consisting of basis elements that represent the characteristics of the expected solution. These low order models serve as surrogates for the dynamical system in the optimization process and the resulting approximate optimization problems can be solved efficiently.

3.1. Reduced order modeling

Reduced order modeling refers to the development of low-dimensional models that represent desired characteristics of a high-dimensional or infinite dimensional dynamical system. Typically, models are constructed by projection of the high-order, high-fidelity model onto a suitably chosen low-dimensional reduced-basis [2]. Most reduced-bases for nonlinear problems are constructed from a collection of simulations (methods of snapshots [54, 55, 56])

The most popular nonlinear model reduction technique is Proper Orthogonal Decomposition (POD) and it usually involves a Galerkin projection with basis $V \in \mathbb{R}^{\mathrm{NVAR}\times\mathrm{NRED}}$ obtained as the output of Algorithm 1. Here NRED is the dimensional of the reduced-order state space spanned e.g., by the POD basis.

Algorithm 1 POD basis construction

```
Solve for the state variable solutions $\mathbf{x}_k,\ k = 1, .., \mathrm{NOBS}$ of (4). One can make use of more snapshots to construct the basis thus for example to consider a number of time steps larger than NOBS.
Compute the singular value decomposition (SVD) for the state variable snapshots matrix $[\mathbf{x}_0\, \mathbf{x}_1\ \dots\ \mathbf{x}_{\mathrm{NOBS}}] = \bar{\mathbf{V}}\Sigma\bar{W}^T$, with the singular vectors matrix $\bar{\mathbf{V}} = [\mathbf{v}_i]_{i=1,..,\mathrm{NVAR}}$.
Using the singular-values $\lambda_1 \geq \lambda_2 \geq \dots \lambda_n \geq 0$ stored in the diagonal matrix $\Sigma$, define $I(p) = (\sum_{i=1}^{p} \lambda_i)/(\sum_{i=1}^{\mathrm{NVAR}} \lambda_i)$.
Choose NRED, the dimension of the POD basis, such that $\mathrm{NRED} = \min_p\{I(p) : I(p) \geq \gamma\}$ where $0 \leq \gamma \leq 1$ is the percentage of total information captured by the reduced space $\mathcal{X}^{\mathrm{NRED}} = \mathrm{range}(\mathbf{V})$, usually $\gamma = 0.99$.
```

Assuming a POD expansion $\mathbf{x}_k \approx \mathbf{V}\tilde{\mathbf{x}}_k,\ \tilde{\mathbf{x}}_k \in \mathbb{R}^{\mathrm{NRED}},\ k = 0, .., \mathrm{NOBS}$ (for simplicity we neglected the centering trajectory, shift mode or mean field correction [42]) and making use of the basis orthogonality the associated POD-Galerkin model of (4) is obtained as

$$\tilde{\mathbf{x}}_{k+1} = \mathbf{V}^T \mathcal{M}_{t_k \to t_{k+1}}(\mathbf{V}\tilde{\mathbf{x}}_k),\ k = 0, .., \mathrm{N}_{\mathrm{NOBS}}. \tag{12}$$

The efficiency of POD - Galerkin technique is limited to the linear or bilinear terms [60] and strategies such as Empirical Interpolation Method (EIM) [7], Discrete Empirical Interpolation Method (DEIM) [15, 57] and tensorial POD [60]are usually employed to alleviate this deficiency.

3.2. Reduced order 4D-Var data assimilation

The "adjoint of reduced plus reduced of adjoint" approach (ARRA) leads to the construction of consistent feasible reduced first order optimality conditions [61] and this framework is employed to build the reduced POD manifolds for the reduced HMC samplers. In the case of Galerkin projection the POD reduced space is constructed based on sampling of both full forward and adjoint trajectories as well as the gradient of the cost function background term.

The reduced data assimilation problem minimizes the following reduced order cost function $\mathcal{J}^{\mathrm{POD}} : \mathbb{R}^k \to \mathbb{R}$

$$\mathcal{J}^{\mathrm{POD}}(\tilde{\mathbf{x}}_0) = \frac{1}{2}\|\mathbf{V}\tilde{\mathbf{x}}_0 - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\sum_{k=0}^{\mathrm{NOBS}} \|\mathbf{y}_k - \mathcal{H}_k(\mathbf{V}\tilde{\mathbf{x}}_k)\|^2_{\mathbf{R}_k^{-1}}\,, \tag{13a}$$

subject to the constraints posed by the ROM projected nonlinear forward model dynamics (12)

$$\tilde{\mathbf{x}}_{k+1} = \widetilde{\mathcal{M}}_{t_k \to t_{k+1}}(\tilde{\mathbf{x}}_k),\ \widetilde{\mathcal{M}}_{t_k \to t_{k+1}}(\tilde{\mathbf{x}}_k) = \mathbf{V}^T \mathcal{M}_{t_k \to t_{k+1}}(\mathbf{V}\tilde{\mathbf{x}}_k),\ k = 0, 1, \dots, \mathrm{N}_{\mathrm{NOBS}}\,. \tag{13b}$$

An observation operator that maps directly from the reduced model space to observations space may be introduced, however for this study the operator requires the projected states as input.

The associated reduced Karush-Kuhn-Tucker conditions [61] are:

ARRA reduced forward model: (14a)

$$\tilde{\mathbf{x}}_{k+1} = \widetilde{\mathcal{M}}_{t_k \to t_{k+1}}\left(\tilde{\mathbf{x}}_k\right), \quad k = 0, .., \mathrm{N}_{\textsc{Nobs}}\,,$$

ARRA reduced adjoint model: (14b)

$$\begin{aligned} \tilde{\boldsymbol{\lambda}}_{\textsc{Nobs}} &= \mathbf{V}^T \widehat{\mathbf{H}}^T_{\textsc{Nobs}} \mathbf{R}^{-1}_{\textsc{Nobs}} \left(\mathbf{y}_{\textsc{Nobs}} - \mathcal{H}_{\textsc{Nobs}}(\mathbf{V}\tilde{\mathbf{x}}_{\textsc{Nobs}})\right), \\ \tilde{\boldsymbol{\lambda}}_k &= \mathbf{V}^T \widehat{\mathbf{M}}^T_{t_k \to t_{k+1}} \mathbf{V} \tilde{\boldsymbol{\lambda}}_{k+1} + \mathbf{V}^T \widehat{\mathbf{H}}_k^T \mathbf{R}_k^{-1} \left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{V}\tilde{\mathbf{x}}_k)\right), \quad k = \textsc{Nobs} - 1, .., 0, \end{aligned}$$

ARRA cost function gradient : (14c)

$$\nabla_{\tilde{\mathbf{x}}_0} \mathcal{J}^{\textsc{pod}} = -\mathbf{V}^T \mathbf{B}_0^{-1} \left(\mathbf{x}_0^{\mathrm{b}} - \mathbf{V}\tilde{\mathbf{x}}_0\right) - \tilde{\boldsymbol{\lambda}}_0 = 0.$$

The operators $\widehat{\mathbf{H}}_k,\ k = 0, .., \textsc{Nobs}$ and $\widehat{\mathbf{M}}_{t_k \to t_{k+1}},\ k = 0, .., \textsc{Nobs}$ are the Jacobians of the high-fidelity observation operator $\mathcal{H}_k$ and model $\mathcal{M}_{t_k \to t_{k+1}}$ evaluated at $V\tilde{\mathbf{x}}_k$.

4. Reduced-Order HMC Sampling Smoothers

One of the key features of HMC sampler is the clever exploration of the state space with guidance based on the distribution geometry. For this the HMC sampling smoother requires not only forward model propagation to evaluate the likelihood term, but also the evaluation of the gradient of the negative-log of the posterior PDF using the adjoint model. This gradient can be approximated using information from the reduced space as obtained from the reduced order model (13b).

The HMC algorithm requires that both the momentum **p** and the target state **x** are vectors of the same dimension. There are two ways to achieve this while using reduced order information:

i) sample the reduced-order subspace only, i.e., collect samples from (15a), or

ii) sample the full space, i.e., collect samples from (17a) but use an approximate gradient of the posterior negative-log likelihood function obtained in the reduced space.

We next discuss each of these options in detail.

4.1. Sampling in the reduced-order space

In this approach the model states are fully projected in the reduced space, $\tilde{\mathbf{x}} \in \mathbb{R}^{\textsc{Nred}}$. The target posterior distribution, and the potential energy are given by:

$$
\begin{aligned}
\pi(\tilde{\mathbf{x}}_0) &\propto \exp\big(-\widetilde{\mathcal{J}}(\tilde{\mathbf{x}}_0)\big)\,,\\
\widetilde{\mathcal{J}}(\tilde{\mathbf{x}}_0) &= \frac{1}{2}\left\|\tilde{\mathbf{x}}_0 - \mathbf{V}^T\mathbf{x}_0^{\rm b}\right\|^2_{(\mathbf{V}^T\mathbf{B}_0\mathbf{V})^{-1}} + \frac{1}{2}\sum_{k=0}^{\textsc{Nobs}} \left\|\mathbf{y}_k - \mathcal{H}_k(\mathbf{V}\tilde{\mathbf{x}}_k)\right\|^2_{\mathbf{R}_k^{-1}}\,, \\
\tilde{\mathbf{x}}_k &= \widetilde{\mathcal{M}}_{k-1,k}(\tilde{\mathbf{x}}_{k-1}), \quad k = 1, 0, \ldots, \textsc{Nobs}\,; \quad \tilde{\mathbf{x}}_0 = \mathbf{V}^T\mathbf{x}_0\,,
\end{aligned}
\tag{15a}
$$

and the gradient of the potentail energy reads:

$$
\nabla_{\tilde{\mathbf{x}}_0}\widetilde{\mathcal{J}}(\tilde{\mathbf{x}}_0) = -\left(\mathbf{V}^T\mathbf{B}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{x}_0^{\rm b} - \tilde{\mathbf{x}}_0\right) - \tilde{\boldsymbol{\lambda}}_0\,,
\tag{15b}
$$

where $\tilde{\boldsymbol{\lambda}}_0$ is the solution of the ARRA reduced adjoint model (14c).

The momentum variable is defined in the reduced space $\tilde{\mathbf{p}}_0 \in \mathbb{R}^{\textsc{Nred}}$, and the Hamiltonian reads:

$$
\widetilde{H}(\tilde{\mathbf{p}}_0,\, \tilde{\mathbf{x}}_0) = \frac{1}{2}\,\tilde{\mathbf{p}}_0^T\,\widetilde{\mathbf{M}}^{-1}\,\tilde{\mathbf{p}}_0 + \widetilde{\mathcal{J}}(\tilde{\mathbf{x}}_0).
\tag{16}
$$

Following [6, 5], the mass matrix can be chosen as the diagonal matrix $\widetilde{\mathbf{M}} = \mathrm{diag}(\mathbf{V}^T\mathbf{B}_0\mathbf{V})^{-1}$. Note that no further approximations are introduced to the numerical flow produced by the symplectic integrator because all calculations involving models states are calculated in the reduced space.

4.2. Sampling in the full space using approximate gradients

Here the model states are initially in the projected subspace defined by $\mathrm{P_v} = \mathbf{V}\mathbf{V}^T$ while the momentum is kept in the full space. The hope is that since the synthetic momentum is drawn at random from the full space for each proposed state, the symplectic integrator will help the sampler jump between slices of the full space rather that sampling a single subspace, leading to a better ensemble of states obtained from the original target posterior.

The target posterior distribution and the potential energy and its gradient are given by

$$
\begin{aligned}
\pi(\mathbf{x}_0) &= \exp\big(-\widehat{\mathcal{J}}(\mathbf{x}_0)\big)\,,\\
\widehat{\mathcal{J}}\big(\mathbf{x}_0\big) &= \frac{1}{2}\|\widehat{\mathbf{x}}_0 - \mathbf{x}_0^{\rm b}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\sum_{k=0}^{\textsc{Nobs}} \|\mathbf{y}_k - \mathcal{H}_k(\widehat{\mathbf{x}}_k)\|^2_{\mathbf{R}_k^{-1}}\,,\\
\widehat{\mathbf{x}}_0 &= \mathbf{x}_0; \quad \widehat{\mathbf{x}}_k = \mathbf{V}\widetilde{\mathcal{M}}_{k-1,k}(\mathbf{V}^T\widehat{\mathbf{x}}_{k-1})\,, \quad k = 1, 2, \ldots, \textsc{Nobs}.
\end{aligned}
\tag{17a}
$$

with gradient of the potential energy given by

$$\nabla_{\mathbf{x}_0} \widehat{\mathcal{J}}\left(\mathbf{x}_0\right) = -\mathbf{B}_0^{-1}(\mathbf{x}_0^{\mathrm{b}} - \mathbf{x}_0) - \widehat{\boldsymbol{\lambda}}_0, \tag{17b}$$

where $\widehat{\boldsymbol{\lambda}}_0$ is the solution of the following adjoint model

$$\begin{aligned} \widehat{\boldsymbol{\lambda}}_{\mathrm{N}_{\mathrm{OBS}}} &= \mathbf{H}_{\mathrm{N}_{\mathrm{OBS}}}^{T} \mathbf{R}_{\mathrm{N}_{\mathrm{OBS}}}^{-1} \left(\mathbf{y}_{\mathrm{N}_{\mathrm{OBS}}} - \mathcal{H}_{\mathrm{N}_{\mathrm{OBS}}}(\widehat{\mathbf{x}}_{\mathrm{N}_{\mathrm{OBS}}})\right), \\ \widehat{\boldsymbol{\lambda}}_{k-1} &= \mathbf{V}\widetilde{\mathbf{M}}_{k-1,k}^{T} V^{T} \widehat{\boldsymbol{\lambda}}_{k} + \mathbf{H}_{k-1}^{T} \mathbf{R}_{k-1}^{-1} \left(\mathbf{y}_{k-1} - \mathcal{H}_{k-1}(\widehat{\mathbf{x}}_{k-1})\right), \quad k = \mathrm{N}_{\mathrm{OBS}}, .., 1. \end{aligned} \tag{18a}$$

Here $\mathbf{H}_k$ represents the observation operator Jacobian linearized at $\widehat{\mathbf{x}}_k,\ k = 0, .., \mathrm{N}_{\mathrm{OBS}}$, and $\widetilde{\mathbf{M}}_{k-1,k}$ is the Jacobian of the reduced order model evaluated at $V^T \widehat{\mathbf{x}}_k,\ k = 0, .., \mathrm{N}_{\mathrm{OBS}}$. The Hamiltonian in this case takes the form:

$$\widehat{H}(\mathbf{p}_0,\, \mathbf{x}_0) = \frac{1}{2}\, \mathbf{p}_0^T \mathbf{M}^{-1} \mathbf{p}_0 + \widehat{\mathcal{J}}(\mathbf{x}_0)\,. \tag{19}$$

An additional approximation is introduced to the numerical flow produced by the symplectic integrator by the approximation of the gradient of the potential energy. This may require more attention to be paid to the process of parameter tuning especially in the case of very high dimensional spaces.

Algorithm 2 summarizes the sampling process that yields an ensemble of states $\{\tilde{\mathbf{x}}_0(e) \in \mathbb{R}^{\mathrm{N}_{\mathrm{RED}}}\}_{e=1,2,\dots,\mathrm{N}_{\mathrm{ENS}}}$ in the reduced space, or ensemble of states $\{\widehat{\mathbf{x}}_0(e), \in \mathbb{R}^{\mathrm{N}_{\mathrm{VAR}}}\}_{e=1,2,\dots,\mathrm{N}_{\mathrm{ENS}}}$ sampled from the high-fidelity state space with approximate gradient information, respectively

Note that in Algorithm 2, $\mathbf{x}_0^{(i)}$ refers to the model state at the initial time of the assimilation window (or models initial conditions) generated in step i of the Markov chain.

5. Properties of the Distributions Sampled with Reduced-Order Models

As explained above, our main goal in this work is to explore the possibility of lowering the computational expense posed by the original HMC smoother [5] by following a reduced-order modeling approach. In the previous Section 5, we mensioned that the use of HMC sampling smoother with reduced order models requires following either of two alternatives, namely sampling the posterior distribution fully projected in the lower dimensional subspace, or sampling the high fidelity distribution with gradients approximated using information obtained from the reduced space. In both cases, some amount of information will be lost due to either projecting the posterior PDF, or approximating the components appearing in the lielihood term. More specifically, in the latter case, approximating the negative-log lielihood terms can lead to samples collected from a totaly different distribution than the true posterior distribution. In the rest of this section, we discuss the properties of the probability distributions resulting

Algorithm 2 HMC Sampling [5].

```
1: Initialize the mass matrix: $\widetilde{\mathbf{M}} \in \mathbf{R}^{\textsc{Nred}\times\textsc{Nred}}$ for sampling from (15a), and $\mathbf{M} \in \mathbf{R}^{\textsc{Nvar}\times\textsc{Nvar}}$ for sampling from (17a).
2: Initialize the chain. Preferably, the initial pair should be as close as possible to the target distribution.
3: At each step $i$ of the Markov chain draw a random auxiliary momentum: $\tilde{\mathbf{p}}_0^{(i)} \sim \mathcal{N}(\mathbf{0}_{\textsc{Nred}}, \widetilde{\mathbf{M}})$ for sampling from (15a), and $\mathbf{p}_0^{(i)} \sim \mathcal{N}(\mathbf{0}_{\textsc{Nred}}, \mathbf{M})$ for sampling from (17a).
4: Use a symplectic numerical integrator (e.g., position Verlet) to advance the current state by a pseudo-time increment $T$ to obtain a *proposal* state :

       $$\begin{aligned} &\text{For sampling from (15a): } (\tilde{\mathbf{p}}_0^*, \tilde{\mathbf{x}}_0^*) = \widetilde{\phi}_T(\tilde{\mathbf{p}}_0^{(i)}, \tilde{\mathbf{x}}_0^{(i)}). \\ &\text{For sampling from (17a): } (\mathbf{p}_0^*, \mathbf{x}_0^*) = \widehat{\phi}_T(\mathbf{p}_0^{(i)}, \mathbf{x}_0^{(i)}), \end{aligned} \tag{20}$$

   where $\widehat{\Phi}_T$ indicates the flow approximation resulting from approximation of the gradient of the potential energy.
5: For sampling from (15a), use the Hamiltonian (16) to evaluate the loss of energy‘ $\Delta\widetilde{H} = \widetilde{H}\big(\tilde{\mathbf{p}}_0^*, \tilde{\mathbf{x}}_0^*\big) - \widetilde{H}\big(\tilde{\mathbf{p}}_0^{(i)}, \tilde{\mathbf{x}}_0^{(i)}\big)$. For sampling (17a), use the Hamiltonian (19) to approximate the energy loss $\Delta\widehat{H} = \widehat{H}\big(\mathbf{p}_0^*, \mathbf{x}_0^*\big) - \widehat{H}\big(\mathbf{p}_0^{(i)}, \mathbf{x}_0^{(i)}\big)$.
6: Calculate the acceptance probability:

       $$\begin{aligned} &\text{For sampling from (15a): } a^{(i)} = 1 \wedge e^{-\Delta\widetilde{H}}, \\ &\text{For sampling from (17a): } a^{(i)} = 1 \wedge e^{-\Delta\widehat{H}}. \end{aligned} \tag{21}$$

7: Discard both current and proposed momentum.
8: **(Acceptance/Rejection)** Draw a uniform random variable $u^{(i)} \sim \mathcal{U}(0,1)$:
       i-  If $a^{(i)} > u^{(i)}$ accept the proposal as the next sample;
       ii- If $a^{(i)} \le u^{(i)}$ reject the proposal and continue with the current state;
9: Repeat steps 2 to 7 until Nens distinct samples are drawn.
10: Project the ensemble to the full space.
```

from projection or due to approximaton of the negative-log likelihood terms making use of information coming only from a reduced-order subspace.

In the direct case where the posterior distribution is fully projected to the lower dimensional subspace, little can be said about the resulting distribution unless if the true posterior is Gaussian. We explore this case in details in what follows.

5.1. Projection of the posterior distribution for linear model and observation operators

In this case the full distribution is projected into the lower-dimensional subspace by approximating both background and observation terms in Equation (2b). This projection leads to ensembles generated only in the reduced-space, and are then projected back to the high-fidelity space by left multiplication with $\mathbf{V}$. Projecting the ensembles back to the full space will not change their mass distribution in the case of a linear model and observation operators, and will just embed the ensembles in the full space.

If both the model and the observation operator are linear operators, the posterior (2) is a Gaussian distribution $\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0^{\mathrm{a}}, \mathbf{A}_0)$, with a posterior (analysis) mean $\mathbf{x}_0^{\mathrm{a}}$, and an analysis error covariance matrix $\mathbf{A}_0$, i.e.

$$\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0) = \frac{(2\pi)^{\frac{-\textsc{Nvar}}{2}}}{\sqrt{|\det(\mathbf{A}_0)|}} \exp\left(-\frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{a}}\|^2_{\mathbf{A_0}^{-1}}\right). \tag{22}$$

The mean and the covariance matrix of the Gaussian posterior (22) are given by

$$\begin{aligned}
\mathbf{A}_0^{-1} &= \mathbf{B}_0^{-1} + \sum_{k=0}^{\textsc{Nobs}} \mathbf{M}_{0,k}^T\, \mathbf{H}_k^T\, \mathbf{R}_k^{-1}\, \mathbf{H}_k\, \mathbf{M}_{0,k}\,, \\
\mathbf{x}_0^{\mathrm{a}} &= \mathbf{A}_0 \cdot \left(\mathbf{B}_0^{-1}\, \mathbf{x}_0^{\mathrm{b}} + \sum_{k=0}^{\textsc{Nobs}} \mathbf{M}_{0,k}^T\, \mathbf{H}_k^T\, \mathbf{R}_k^{-1}\, \mathbf{y}_k\right).
\end{aligned} \tag{23}$$

Projecting this PDF onto the subspace spanned by columns of the matrix $\mathbf{V}$ (e.g., POD basis) results in a projected PDF $\widetilde{\mathcal{P}}^{\mathrm{a}}(\tilde{\mathbf{x}}_0) = \mathcal{N}(\mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}, \mathbf{V}^T\mathbf{A}_0\mathbf{V})$; $\tilde{\mathbf{x}}_0 \in \mathbb{R}^{\textsc{Nred}}$, i.e.,

$$\widetilde{\mathcal{P}}^{\mathrm{a}}(\tilde{\mathbf{x}}_0) = \frac{(2\pi)^{\frac{-\textsc{Nred}}{2}}}{\sqrt{|\det(\mathbf{V}^T\mathbf{A}_0\mathbf{V})|}} \exp\left(-\frac{1}{2}\|\tilde{\mathbf{x}}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}\|^2_{(\mathbf{V^T A_0 V})^{-1}}\right). \tag{24}$$

The linear transformation, of the analysis state, with the orthogonal projector $\mathrm{P_v} = \mathbf{V}\mathbf{V}^T$, results as well in the Gaussian distribution $\widehat{\mathcal{P}}^{\mathrm{a}}(\widehat{\mathbf{x}}_0) = \mathcal{N}(\mathrm{P_v}\mathbf{x}_0^{\mathrm{a}}, \mathrm{P_v}\mathbf{A}_0\mathrm{P_v}) \equiv \mathcal{N}(\widehat{\mathbf{x}}_0^{\mathrm{a}}, \widehat{\mathbf{A}}_0)$, $\widehat{\mathbf{x}}_0 \in \mathbb{R}^{\textsc{Nvar}}$. The covariance matrix $\widehat{\mathbf{A}}_0$ however is not full rank, and the Gaussian distribution is degenerate. The density function of this singular distribution can be rigourously formulated by defining a restriction of Lebesgue measure to the affine subspace of $\mathbb{R}^{\textsc{Nvar}}$ whose dimension is limited to rank($\widehat{\mathbf{A}}_0$). The Gaussian (singular) density then formula takes the form [34, 44]

$$\widehat{\mathcal{P}}^{\mathrm{a}}(\widehat{\mathbf{x}}_0) = \frac{(2\pi)^{\frac{-\textsc{Nred}}{2}}}{\sqrt{|\det{}^*(\widehat{\mathbf{A}}_0)|}} \exp\left(-\frac{1}{2}\|\widehat{\mathbf{x}}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}}\|^2_{\widehat{\mathbf{A}}_0^{\dagger}}\right)\cdot\delta_{(\mathbf{I}-\mathrm{P_v})\widehat{\mathbf{x}}_0} \tag{25}$$

where det* is the pseudo determinant, and † refers to the matrix pseudo inverse. Of course, $\tilde{\mathbf{x}}_0 \in \mathbb{R}^{\textsc{Nred}}$, $\widehat{\mathbf{x}}_0 \in \mathbb{R}^{\textsc{Nvar}}$. One can think of the PDF (25) as a version of (24) embeded in the high-fidelity state space.

Theorem 5.1. *If* $\widetilde{\mathcal{P}}^{\mathrm{a}}(\tilde{\mathbf{x}}_0)$ *and* $\widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{V}\tilde{\mathbf{x}}_0)$ *are the distributions defined in* (24) *and* (25)*, respectively, then the following result holds true for a given reduced basis* $\mathbf{V}$

$$\widetilde{\mathcal{P}}^{\mathrm{a}}(\tilde{\mathbf{x}}_0) = \widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{V}\tilde{\mathbf{x}}_0),\ \forall \tilde{\mathbf{x}}_0 \in \mathbb{R}^{\textsc{Nred}}. \tag{26}$$

Proof. For this purpose it is sufficient to prove that

$$\|\widehat{\mathbf{x}}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}}\|^2_{\widehat{\mathbf{A}}_0^{\dagger}} = \|\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}\|^2_{(\mathbf{V^T A_0 V})^{-1}} \tag{27}$$

Assume the relation given by Equation (27) is correct, we get the following:

$$\begin{aligned} \|\widehat{\mathbf{x}}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}}\|^2_{\widehat{\mathbf{A}}_0^{\dagger}} &= \|\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}\|^2_{(\mathbf{V}^T\mathbf{A}_0\mathbf{V})^{-1}}\,, \\ (\widehat{\mathbf{x}}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}})^T\widehat{\mathbf{A}}_0^{\dagger}(\widehat{\mathbf{x}}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}}) &= (\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}})^T\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}) \\ (\mathrm{P_v}\mathbf{x}_0 - \mathrm{P_v}\mathbf{x}_0^{\mathrm{a}})^T\widehat{\mathbf{A}}_0^{\dagger}(\mathrm{P_v}\mathbf{x}_0 - \mathrm{P_v}\mathbf{x}_0^{\mathrm{a}}) &= (\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}})^T\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}) \end{aligned} \tag{28a}$$

Or equivalently:

$$\begin{aligned} 0 &= \left(\mathrm{P_v}\mathbf{x}_0 - \mathrm{P_v}\mathbf{x}_0^{\mathrm{a}}\right)^T\widehat{\mathbf{A}}_0^{\dagger}\left(\mathrm{P_v}\mathbf{x}_0 - \mathrm{P_v}\mathbf{x}_0^{\mathrm{a}}\right) - (\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}})^T\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}) \\ &= \left(\mathbf{V}(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}})\right)^T\widehat{\mathbf{A}}_0^{\dagger}\left(\mathbf{V}(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}})\right) - \left(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}\right)^T\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}\right) \\ &= \left(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}\right)^T\mathbf{V}^T\widehat{\mathbf{A}}_0^{\dagger}\mathbf{V}\left(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}\right) - \left(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}\right)^T\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}\right) \\ &= (\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}})^T\left(\mathbf{V}^T\widehat{\mathbf{A}}_0^{\dagger}\mathbf{V} - \left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\right)(\mathbf{V}^T\mathbf{x}_0 - \mathbf{V}^T\mathbf{x}_0^{\mathrm{a}})\,. \end{aligned} \tag{28b}$$

This holds true if the matrix $\mathbf{V}^T\widehat{\mathbf{A}}_0^{\dagger}\mathbf{V} - \left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}$ is equal to a zero matrix. The matrix $\mathbf{V}$ has orthonormal columns and consequently $(\mathrm{P_v}\mathbf{A}_0\mathrm{P_v})^{\dagger} = \left(\mathbf{V}\mathbf{V}^T\mathbf{A}_0\mathrm{P_v}\right)^{\dagger} = \left(\mathbf{V}^T\mathbf{A}_0\mathrm{P_v}\right)^{\dagger}\mathbf{V}^{\dagger}$. Since the pseudo inverse and the transpose operations are commutative, we get the following:

$$\begin{aligned} \left(\mathbf{V}^T\mathbf{A}_0\mathrm{P_v}\right)^{\dagger} &= \left(\left(\mathrm{P_v}\mathbf{A}_0^T\mathbf{V}\right)^T\right)^{\dagger}\,, \\ &= \left(\mathbf{V}^T\right)^{\dagger}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{\dagger}\,, \end{aligned} \tag{28c}$$

and consequently:

$$\begin{aligned} \mathbf{V}^T\widehat{\mathbf{A}}_0^{\dagger}\mathbf{V} = \mathbf{V}^T(\mathrm{P_v}\mathbf{A}_0\mathrm{P_v})^{\dagger}\mathbf{V} &= \mathbf{V}^T\left(\mathbf{V}^T\right)^{\dagger}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{\dagger}\mathbf{V}^{\dagger}\mathbf{V}\,, \\ &= \mathbf{V}^T\left(\mathbf{V}^T\right)^{\dagger}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^{\dagger}\mathbf{V}\,, \\ &= (\mathbf{V}^T\mathbf{V})\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}(\mathbf{V}^T\mathbf{V})\,, \\ &= \left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\,, \end{aligned} \tag{28d}$$

where $\mathbf{V}^{\dagger} = \mathbf{V}^T$, and $\left(\mathbf{V}^T\right)^{\dagger} = \mathbf{V}$ since $\mathbf{V}$ has orthonormal columns. This means that the relation (27) holds, and the equivalence between (24) and (25) follows immediately.

This result suggests that sampling from the distribution (25) can be carried out efficiently by sampling the distribution (24), then projecting the ensembles back to the full space using $\mathbf{V}$.

By determining the Kullback Leibler (KL) [17] divergence measure between the high fidelity distribution $\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)$ and the probability distribution $\widehat{\mathcal{P}}^{\mathrm{a}}(\widehat{\mathbf{x}}_0)$, one can estimate the error between the projected samples obtained using distribution (24) and those sampled from the high fidelity distribution $\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)$.

Theorem 5.2. *The KL divergence measure between the Gaussian distribution $\widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{x}_0)$ given by* (25)*, and the probability distribution $\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)$ defined in* (22)*, is given as*

$$D_{\mathrm{KL}}\left(\widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{x}_0)||\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)\right) = \frac{1}{2}\left((\textsc{Nvar} - \textsc{Nred}) \ln{(2\pi)} + \ln\left(\frac{|\det(\mathbf{A}_0)|}{|\det^{*}(\widehat{\mathbf{A}}_0)|}\right) + ||\widehat{\mathbf{x}}_0^{\mathrm{a}} - \mathbf{x}_0^{\mathrm{a}}||_{\mathbf{A}_0^{-1}} + \mathrm{trace}\Big(\left(\mathbf{A}_0^{-1} - \widehat{\mathbf{A}}_0^{\dagger}\right) \widehat{\mathbf{A}}_0\Big)\right) , \tag{29}$$

where $\mathbf{V} \in \mathbf{R}^{\textsc{Nvar}\times\textsc{Nred}}$ and $\textsc{Nred} < \textsc{Nvar}$.

Proof. The KL measure is obtained as

$$D_{\mathrm{KL}}\left(\widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{x}_0)||\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)\right) = \mathrm{E}_{\widehat{\mathcal{P}}^{\mathrm{a}}}\left[\ln\left(\frac{\widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{x}_0)}{\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)}\right)\right] , \tag{30a}$$

$$\ln\left(\frac{\widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{x}_0)}{\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)}\right) = \ln\left(\frac{(2\pi)^{\frac{-\textsc{Nred}}{2}}}{\sqrt{|\det^{*}(\widehat{\mathbf{A}}_0)|}}\right) + \ln\left(\frac{\sqrt{|\det(\mathbf{A}_0)|}}{(2\pi)^{\frac{-\textsc{Nvar}}{2}}}\right) + \ln\left(\frac{\exp\left(-\frac{1}{2}||\mathbf{x}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}}||^2_{\widehat{\mathbf{A}}_0^{\dagger}}\right)}{\exp\left(-\frac{1}{2}||\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{a}}||^2_{\mathbf{A_0}^{-1}}\right)}\right) \tag{30b}$$

$$= \frac{(\textsc{Nvar} - \textsc{Nred}) \ln{(2\pi)}}{2} + \frac{1}{2}\ln\left(\frac{|\det(\mathbf{A}_0)|}{|\det^{*}(\widehat{\mathbf{A}}_0)|}\right) + \frac{1}{2}\Big(||\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{a}}||^2_{\mathbf{A_0}^{-1}} - ||\mathbf{x}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}}||^2_{\widehat{\mathbf{A}}_0^{\dagger}}\Big) \tag{30c}$$

$$\mathrm{E}_{\widehat{\mathcal{P}}^{\mathrm{a}}}\left[\ln\left(\frac{\widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{x}_0)}{\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)}\right)\right] = \frac{(\textsc{Nvar} - \textsc{Nred}) \ln{(2\pi)}}{2} + \frac{1}{2}\ln\left(\frac{|\det(\mathbf{A}_0)|}{|\det^{*}(\widehat{\mathbf{A}}_0)|}\right) + \frac{1}{2}\mathrm{E}_{\widehat{\mathcal{P}}^{\mathrm{a}}}\left[\Big(||\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{a}}||^2_{\mathbf{A_0}^{-1}} - ||\mathbf{x}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}}||^2_{\widehat{\mathbf{A}}_0^{\dagger}}\Big)\right] , \tag{30d}$$

where $\ln\left(\frac{|\det(\mathbf{A}_0)|}{|\det^{*}(\widehat{\mathbf{A}}_0)|}\right)$ is the sum of logarithms of eigenvalues of $\mathbf{A}_0$ lost due to projection. This value can be also replaced with $ln\Big(\frac{|\det(\mathbf{A}_0)|}{|\det{(\mathbf{V}^T\mathbf{A}_0\mathbf{V})}|}\Big)$ due to the nature of the matrix $\mathbf{V}$. The expectation of the quadratic terms in Equation (30d) can be obtained as follows:

$$\begin{aligned}\mathrm{E}_{\widehat{\mathcal{P}}^{\mathrm{a}}}\left[\Big(||\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{a}}||^2_{\mathbf{A}_0^{-1}} - ||\mathbf{x}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}}||^2_{\widehat{\mathbf{A}}_0^{\dagger}}\Big)\right] &= \mathrm{E}_{\widehat{\mathcal{P}}^{\mathrm{a}}}\left[\Big(||\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{a}}||^2_{\mathbf{A}_0^{-1}}\right] - \mathrm{E}_{\widehat{\mathcal{P}}^{\mathrm{a}}}\left[||\mathbf{x}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}}||^2_{\widehat{\mathbf{A}}_0^{\dagger}}\Big)\right] && \text{(31a)}\\ &= \mathrm{E}_{\widehat{\mathcal{P}}^{\mathrm{a}}}\left[(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{a}})^T\mathbf{A}_0{}^{-1}(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{a}})\right] - \mathrm{E}_{\widehat{\mathcal{P}}^{\mathrm{a}}}\left[(\mathbf{x}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}})^T\widehat{\mathbf{A}}_0^{\dagger}(\mathbf{x}_0 - \widehat{\mathbf{x}}_0^{\mathrm{a}})\right] && \text{(31b)}\\ &= (\widehat{\mathbf{x}}_0^{\mathrm{a}} - \mathbf{x}_0^{\mathrm{a}})^T\mathbf{A}_0^{-1}(\widehat{\mathbf{x}}_0^{\mathrm{a}} - \mathbf{x}_0^{\mathrm{a}}) + \mathrm{Tr}\Big(\mathbf{A}_0^{-1}\,\widehat{\mathbf{A}}_0\Big) - \mathrm{Tr}\Big(\widehat{\mathbf{A}}_0^{\dagger}\,\widehat{\mathbf{A}}_0\Big) && \text{(31c)}\end{aligned}$$

from Equations (31), and (30), we obtain:

$$
\begin{aligned}
D_{\mathrm{KL}}\left(\widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{x}_0)\|\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)\right) &= \frac{(\textsc{Nvar} - \textsc{Nred})\ln(2\pi)}{2} + \frac{1}{2}\ln\left(\frac{|\det(\mathbf{A}_0)|}{|\det^{*}(\widehat{\mathbf{A}}_0)|}\right) + \frac{1}{2}\|\widehat{\mathbf{x}}_0^{\mathrm{a}} - \mathbf{x}_0^{\mathrm{a}}\|_{\mathbf{A}_0^{-1}} \qquad (32a)\\
&+ \frac{1}{2}\mathrm{Tr}\left(\mathbf{A}_0^{-1}\,\widehat{\mathbf{A}}_0\right) - \frac{1}{2}\mathrm{Tr}\left(\widehat{\mathbf{A}}_0^{\dagger}\,\widehat{\mathbf{A}}_0\right)\\
&= \frac{(\textsc{Nvar} - \textsc{Nred})\ln(2\pi)}{2} + \frac{1}{2}\ln\left(\frac{|\det(\mathbf{A}_0)|}{|\det^{*}(\widehat{\mathbf{A}}_0)|}\right) + \frac{1}{2}\|\widehat{\mathbf{x}}_0^{\mathrm{a}} - \mathbf{x}_0^{\mathrm{a}}\|_{\mathbf{A}_0^{-1}} \qquad (32b)\\
&+ \frac{1}{2}\mathrm{Tr}\left(\mathbf{A}_0^{-1}\,\widehat{\mathbf{A}}_0 - \widehat{\mathbf{A}}_0^{\dagger}\,\widehat{\mathbf{A}}_0\right)\\
&= \frac{1}{2}\left((\textsc{Nvar} - \textsc{Nred})\ln(2\pi) + \ln\left(\frac{|\det(\mathbf{A}_0)|}{|\det^{*}(\widehat{\mathbf{A}}_0)|}\right) + \|\widehat{\mathbf{x}}_0^{\mathrm{a}} - \mathbf{x}_0^{\mathrm{a}}\|_{\mathbf{A}_0^{-1}} + \mathrm{trace}\left(\left(\mathbf{A}_0^{-1} - \widehat{\mathbf{A}}_0^{\dagger}\right)\,\widehat{\mathbf{A}}_0\right)\right), \qquad (32c)
\end{aligned}
$$

which completes the proof.

This measure can be used to quantify the quality of POD basis given an estimation of the analysis error covariance matrix, e.g., based on an ensemble of states, sampled from the high fidelity distribution, or approximated based on statistics of the 4D-Var cost functional. Notice that the KL measure given in (29) is finite since $\widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{x}_0)$ is absolutely continuous with respect to $\mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)$ (and it is zero only if $\widehat{\mathcal{P}}^{\mathrm{a}}(\mathbf{x}_0) = \mathcal{P}^{\mathrm{a}}(\mathbf{x}_0)$). For this reason, we set the projected PDF as the reference density in the KL measure.

5.2. Approximating the likelihood function using reduced order models

In the latter approach, the background term is kept in the high fildelity space, while only the terms involving model propagations are approximated using reduced-order models. This means that the target distribution is the PDF give by (17a). The use of this approximation in the HMC algorithm results in samples collected from the distribution (17a). This approximation maintains the background term in the full space, while the model states involved in the observation term are approximated in the lower-dimensional subspace. This means that the posterior distribution is non-degenerate in the full space due to the background term. However, it is not immediately obvious which distributions samples will be collected from. In Theorem 5.3 we show the link between posterior distribution given by (17a) and the distribution defined in (2).

Theorem 5.3. *The posterior distribution* $\boldsymbol{\pi}$ *defined in* (2) *associated with the high-fidelity model* (4) *is proportional to the analysis posterior distribution* $\widetilde{\boldsymbol{\pi}}$ *introduced in* (17a) *associated with the reduced order model, by the ratio between joint likelihood functions given projected and high-fidelity states, i.e.*

$$
\widetilde{\boldsymbol{\pi}}(\mathbf{x}_0) = \boldsymbol{\pi}(\mathbf{x}_0)\cdot\prod_{k=0}^{\textsc{Nobs}}\frac{\mathcal{P}\left(\mathbf{y}_k|\mathbf{x}_k = \mathbf{V}\,\widetilde{\mathcal{M}_{0,k}}(\mathbf{V}^T\mathbf{x}_0)\right)}{\mathcal{P}\left(\mathbf{y}_k|\mathbf{x}_k = \mathcal{M}_{0,k}(\mathbf{x}_0)\right)}\,. \qquad (33)
$$

Proof. The exact and the approximate posterior distributions $\boldsymbol{\pi}(\mathbf{x}_0)$, $\widetilde{\boldsymbol{\pi}}(\mathbf{x}_0)$ are generally described as follows

$$\begin{aligned}\boldsymbol{\pi}(\mathbf{x}_0) = \mathcal{P}^{\rm a}(\mathbf{x}_0) &= \mathcal{P}^{\rm b}(\mathbf{x}_0) \cdot \mathcal{P}(\mathbf{y}_0|\mathbf{x}_0) \cdot \prod_{k=1}^{\textsc{Nobs}} \mathcal{P}(\mathbf{y}_k|\mathbf{x}_k) \cdot \mathcal{P}(\mathbf{x}_k|\mathbf{x}_{k-1}) \\ &= \mathcal{P}^{\rm b}(\mathbf{x}_0) \cdot \mathcal{P}(\mathbf{y}_0|\mathbf{x}_0) \cdot \prod_{k=1}^{\textsc{Nobs}} \mathcal{P}\big(\mathbf{y}_k|\mathbf{x}_k = \mathcal{M}_{0,k}(\mathbf{x}_0)\big)\,,\end{aligned} \tag{34a}$$

$$\begin{aligned}\widetilde{\boldsymbol{\pi}}(\mathbf{x}_0) &= \mathcal{P}^{\rm b}(\mathbf{x}_0) \cdot \mathcal{P}(\mathbf{y}_0|\tilde{\mathbf{x}}_0) \cdot \prod_{k=1}^{\textsc{Nobs}} \mathcal{P}(\mathbf{y}_k|\tilde{\mathbf{x}}_k) \cdot \mathcal{P}(\tilde{\mathbf{x}}_k|\tilde{\mathbf{x}}_{k-1}) \\ &= \mathcal{P}^{\rm b}(\mathbf{x}_0) \cdot \mathcal{P}(\mathbf{y}_0|\tilde{\mathbf{x}}_0) \cdot \prod_{k=1}^{\textsc{Nobs}} \mathcal{P}\big(\mathbf{y}_k|\mathbf{x}_k = \mathbf{V}\,\widetilde{\mathcal{M}}_{0,k}(\mathbf{V}^T\mathbf{x}_0)\big).\end{aligned} \tag{34b}$$

This leads to the following:

$$\begin{aligned}\frac{\widetilde{\boldsymbol{\pi}}(\mathbf{x}_0)}{\boldsymbol{\pi}(\mathbf{x}_0)} &= \frac{\prod_{k=0}^{\textsc{Nobs}} \mathcal{P}\big(\mathbf{y}_k|\mathbf{x}_k = \mathbf{V}\,\widetilde{\mathcal{M}}_{0,k}(\mathbf{V}^T\mathbf{x}_0)\big)}{\prod_{k=0}^{\textsc{Nobs}} \mathcal{P}\big(\mathbf{y}_k|\mathbf{x}_k = \mathcal{M}_{0,k}(\mathbf{x}_0)\big)}\,, \\ \widetilde{\boldsymbol{\pi}}(\mathbf{x}_0) &= \boldsymbol{\pi}(\mathbf{x}_0) \cdot \prod_{k=0}^{\textsc{Nobs}} \frac{\mathcal{P}\big(\mathbf{y}_k|\mathbf{x}_k = \mathbf{V}\,\widetilde{\mathcal{M}}_{0,k}(\mathbf{V}^T\mathbf{x}_0)\big)}{\mathcal{P}\big(\mathbf{y}_k|\mathbf{x}_k = \mathcal{M}_{0,k}(\mathbf{x}_0)\big)}\,.\end{aligned} \tag{34c}$$

This result suggests that the larger the distances $\|\mathbf{x}_k - \mathbf{V}\tilde{x}_k\|_2,\ k = 1, 2, .., \textsc{Nobs}$ are, the more different the distributions $\boldsymbol{\pi}$ and $\widetilde{\boldsymbol{\pi}}$ will be. By selecting appropriate reduced manifolds $\mathbf{V}$ and decreasing the error associated with the reduced order models, the ratio can be brought closer to 1.

Corollary 5.3.1. *The KL divergence measure between the original posterior* (2) *and the approximated distribution* (17a) *is:*

$$\begin{aligned}D_{\rm KL}(\widetilde{\boldsymbol{\pi}}\|\boldsymbol{\pi}) = \mathrm{E}_{\widetilde{\boldsymbol{\pi}}}\left[\ln(\widetilde{\boldsymbol{\pi}}) - \ln(\boldsymbol{\pi})\right] &= \mathrm{E}_{\widetilde{\boldsymbol{\pi}}}\left[\mathcal{J}(\mathbf{x}_0) - \widetilde{\mathcal{J}}(\mathbf{x}_0)\right] \\ &= \mathrm{E}_{\widetilde{\boldsymbol{\pi}}}\left[\mathcal{J}^{\rm obs}(\mathbf{x}_0) - \widetilde{\mathcal{J}}^{\rm obs}(\mathbf{x}_0)\right]\,,\end{aligned} \tag{35}$$

where $\mathcal{J}^{\rm obs}(\mathbf{x}_0)$, *and* $\widetilde{\mathcal{J}}^{\rm obs}(\mathbf{x}_0)$ *are the observation terms in the full and the approximate 4D-Var cost function.*

Corollary 5.3.2. *In the filtering case, where only one observation is assimilated, if the initial condition is projected on the columns of* $\mathbf{V}$ *to approximate the likelihood term, the posterior distribution is given by:*

$$\widetilde{\boldsymbol{\pi}}(\mathbf{x}_0) = \widetilde{\mathcal{P}^{\rm a}}(\mathbf{x}_0) \propto \frac{\boldsymbol{\pi}(\mathbf{x}_0^{\|}) \cdot \mathcal{P}^{\rm b}(\mathbf{x}_0)}{\mathcal{P}^{\rm b}(\mathbf{x}_0^{\|})}\,, \tag{36}$$

where $\mathbf{x}_0 = \mathbf{x}_0^{\|} + \mathbf{x}_0^{\perp}$, *with* $\mathbf{x}_0^{\|} \in$ *range*$(\mathbf{V})$ *and* $\mathbf{x}_0^{\perp} \in$ *null*$(\mathbf{V}^T)$

Proof. The two states $\mathbf{x}_0$ and $\mathbf{x}_0^{\|}$ differ only along a direction $\mathbf{x}_0^{\perp}$ orthogonal to the reduced space, that is $\mathbf{V}^T\mathbf{x}_0^{\perp} = 0$, and consequently

$$\mathbf{V}^T\,\mathbf{x}_0 = \mathbf{V}^T(\mathbf{x}_0^{\|} + \mathbf{x}_0^{\perp}) = \mathbf{V}^T\,\mathbf{x}_0^{\|}.$$

In the filtering case the cost function reads:

$$
\begin{aligned}
\widetilde{\mathcal{J}}(\mathbf{x}_0) &= \frac{1}{2}\|\mathbf{x}_0^{\|} + \mathbf{x}_0^{\perp} - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\left\|\mathbf{y}_k - \mathcal{H}_0\left(\mathbf{V}\mathbf{V}^T\mathbf{x}_0^{\|}\right)\right\|_{\mathbf{R}_0^{-1}} \\
&= \frac{1}{2}\|\mathbf{x}_0^{\|} + \mathbf{x}_0^{\perp} - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\left\|\mathbf{y}_k - \mathcal{H}_0(\mathbf{x}_0^{\|})\right\|_{\mathbf{R}_0^{-1}}, \\
\mathcal{J}(\mathbf{x}_0^{\|}) &= \frac{1}{2}\|\mathbf{x}_0^{\|} - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\left\|\mathbf{y}_k - \mathcal{H}_0\left(\mathbf{x}_0^{\|}\right)\right\|_{\mathbf{R}_0^{-1}}, \\
-\widetilde{\mathcal{J}}(\mathbf{x}_0) &= -\mathcal{J}(\mathbf{x}_0^{\|}) - \frac{1}{2}\|\mathbf{x}_0^{\|} + \mathbf{x}_0^{\perp} - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\|\mathbf{x}_0^{\|} - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}}.
\end{aligned}
\tag{37a}
$$

Exponentiating of both sides leads to the following:

$$
\begin{aligned}
\exp\left(-\widetilde{\mathcal{J}}(\mathbf{x}_0)\right) &= \exp\left(-\frac{1}{2}\|\mathbf{x}_0^{\|} + \mathbf{x}_0^{\perp} - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}}\right)\exp\left(-\mathcal{J}(\mathbf{x}_0^{\|})\right)\exp\left(\frac{1}{2}\|\mathbf{x}_0^{\|} - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}}\right), \\
\widetilde{\boldsymbol{\pi}}(\mathbf{x}_0) &\propto \frac{\boldsymbol{\pi}(\mathbf{x}_0^{\|}) \cdot \mathcal{P}^{\mathrm{b}}(\mathbf{x}_0)}{\mathcal{P}^{\mathrm{b}}(\mathbf{x}_0^{\|})}.
\end{aligned}
\tag{38}
$$

This completes the proof.

Corollary 5.3.3. *In the filtering case, if* $\mathbf{x}_0^{\perp} = 0$ *the two distributions* $\widetilde{\boldsymbol{\pi}}(\mathbf{x}_0)$*, and* $\boldsymbol{\pi}(\mathbf{x}_0)$ *coincide, and if* $\mathbf{x}_0^{\|} = 0$ *then the reduced distribution* $\widetilde{\boldsymbol{\pi}}(\mathbf{x}_0)$ *coincides with the background distribution* $\mathcal{P}^{\mathrm{b}}(\mathbf{x}_0)$*.*

In the general case we have:

$$
\begin{aligned}
\widetilde{\mathcal{J}}(\mathbf{x}_0) = \widetilde{\mathcal{J}}(\mathbf{x}_0^{\|} + \mathbf{x}_0^{\perp}) &= \frac{1}{2}\|\mathbf{x}_0^{\|} + \mathbf{x}_0^{\perp} - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\sum_{k=0}^{\mathrm{Nobs}}\left\|\mathbf{y}_k - \mathcal{H}_k\left(\mathbf{V}\widetilde{\mathcal{M}}_{0,k}\left(\mathbf{V}^T(\mathbf{x}_0^{\|} + \mathbf{x}_0^{\perp})\right)\right)\right\|_{\mathbf{R}_k^{-1}} \\
&= \frac{1}{2}\|\mathbf{x}_0^{\|} + \mathbf{x}_0^{\perp} - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\sum_{k=0}^{\mathrm{Nobs}}\left\|\mathbf{y}_k - \mathcal{H}_k\left(\mathbf{V}\widetilde{\mathcal{M}}_{0,k}\left(\mathbf{V}^T\mathbf{x}_0^{\|}\right)\right)\right\|_{\mathbf{R}_k^{-1}},
\end{aligned}
\tag{39a}
$$

$$
\mathcal{J}(\mathbf{x}_0^{\|}) = \frac{1}{2}\|\mathbf{x}_0^{\|} - \mathbf{x}_0^{\mathrm{b}}\|^2_{\mathbf{B}_0^{-1}} + \frac{1}{2}\sum_{k=0}^{\mathrm{Nobs}}\left\|\mathbf{y}_k - \mathcal{H}_k\left(\mathcal{M}_{0,k}\left(\mathbf{x}_0^{\|}\right)\right)\right\|_{\mathbf{R}_k^{-1}},
\tag{39b}
$$

Corollary 5.3.4. *The posterior* $\widetilde{\boldsymbol{\pi}}$ (17a) *is Gaussian with analysis covariance and mean:*

$$
\begin{aligned}
\widehat{\mathbf{A}}_0^{-1} &= \mathbf{B}_0^{-1} + \sum_{k=0}^{\mathrm{Nobs}}\mathbf{V}\,\widetilde{\mathbf{M}}_{0,k}^T\,\mathbf{H}_k^T\,\mathbf{R}_k^{-1}\,\mathbf{H}_k\,\widetilde{\mathbf{M}}_{0,k}\,\mathbf{V}^T \\
\widehat{\mathbf{x}}_0^{\mathrm{a}} &= \widehat{\mathbf{A}}_0 \cdot \left(\mathbf{B}_0^{-1}\,\mathbf{x}_0^{\mathrm{b}} + \sum_{k=0}^{\mathrm{Nobs}}\mathbf{V}\widetilde{\mathbf{M}}_{0,k}^T\,\mathbf{H}_k^T\,\mathbf{R}_k^{-1}\,\mathbf{y}_k\right).
\end{aligned}
\tag{40}
$$

From Equations (23) and (40) we conclude that the analysis mean and covariance associated with the distribution $\widetilde{\boldsymbol{\pi}}$ (17a) are not obtained simply by projecting the mean and covariance of the high fidelity distribution $\boldsymbol{\pi}$ (2), i.e., $\widehat{\mathbf{A}}_0 \neq \mathbf{V}\mathbf{V}^T\mathbf{A}_0\mathbf{V}\mathbf{V}^T$ and $\widehat{\mathbf{x}}_0^{\mathrm{a}} \neq \mathbf{V}\mathbf{V}^T\mathbf{x}_0^{\mathrm{a}}$.

Corollary 5.3.5. *For a constant model operator* $\mathbf{M}_{k-1,k} = \mathbf{M}$ *the mean and the covariance of the high-fidelity* *posterior* (2) *are*

$$
\begin{aligned}
\mathbf{A}_0^{-1} &= \mathbf{B}_0^{-1} + \sum_{k=0}^{\mathrm{N_{OBS}}} (\mathbf{M}^k)^T\, \mathbf{H}_k^T\, \mathbf{R}_k^{-1}\, \mathbf{H}_k\, (\mathbf{M}^k)\,, \\
\mathbf{x}_0^{\mathrm{a}} &= \mathbf{A}_0 \cdot \left(\mathbf{B}_0^{-1}\, \mathbf{x}_0^{\mathrm{b}} + \sum_{k=0}^{\mathrm{N_{OBS}}} (\mathbf{M}^k)^T\, \mathbf{H}_k^T\, \mathbf{R}_k^{-1}\, \mathbf{y}_k \right)\,,
\end{aligned}
\tag{41}
$$

while in the case of posterior (17a)*, the associated analysis covariance and mean are*

$$
\begin{aligned}
\widehat{\mathbf{A}}_0^{-1} &= \mathbf{B}_0^{-1} + \sum_{k=0}^{\mathrm{N_{OBS}}} \left((\mathrm{P_v} \mathbf{M} \mathrm{P_v})^k \right)^T\, \mathbf{H}_k^T\, \mathbf{R}_k^{-1}\, \mathbf{H}_k\, (\mathrm{P_v} \mathbf{M} \mathrm{P_v})^k \\
\widehat{\mathbf{x}}_0^{\mathrm{a}} &= \widehat{\mathbf{A}}_0 \cdot \left(\mathbf{B}_0^{-1}\, \mathbf{x}_0^{\mathrm{b}} + \sum_{k=0}^{\mathrm{N_{OBS}}} \left((\mathrm{P_v} \mathbf{M} \mathrm{P_v})^k \right)^T\, \mathbf{H}_k^T\, \mathbf{R}_k^{-1}\, \mathbf{y}_k \right)\,.
\end{aligned}
\tag{42}
$$

A closed form for the distribution (17a) can be obtained if a) the observation errors are defined given the state vectors in the lower-dimensional subspace embedded in the full space (the projected space), b) the observation errors at time t_k follow Gaussian distribution with zero mean and covariance matrix $\tilde{\mathbf{R}}_k$, that is

$$
\tilde{\mathbf{e}}_k^{\mathrm{obs}} = \mathbf{y}_k - \mathcal{H}(\mathbf{V}\tilde{\mathbf{x}}_k) \sim \mathcal{N}(0,\, \tilde{\mathbf{R}}_k)\,, \tag{43}
$$

and c) forcing the regular assumptions of time independence of observation errors, and independence from model background state (in the smaller space), once can obtain the posterior defined by (17a).

6. Numerical Results

In this section we test numerically the reduced order sampling algorithms using the shallow-water equations (SWE) model in Cartesian coordinates.

6.1. The SWE model

Many phenomena in fluid dynamics are characterized by horizontal length scale much greater than the vertical length, consequently when equipped with Coriolis forces, the shallow water equations model (SWE) becomes a valuable tool in atmospheric modeling, as a simplification of the primitive equations of atmospheric flow. Their solutions represent many of the types of motion found in the real atmosphere, including slow-moving Rossby waves and fast-moving gravity waves [28]. The alternating direction fully implicit finite difference scheme [27] was considered in this paper and it is stable for large CFL condition numbers.

The SWE model using the β-plane approximation on a rectangular domain is introduced (see [27])

$$
\frac{\partial w}{\partial t} = A(w)\frac{\partial w}{\partial x} + B(w)\frac{\partial w}{\partial y} + C(y)w, \quad (x, y) \in [0, L] \times [0, D], \quad t \in (0, t_{\mathrm{f}}], \tag{44}
$$

where $w = (u, v, \phi)^T$ is a vector function, u, v are the velocity components in the x and y directions, respectively, h is the depth of the fluid, g is the acceleration due to gravity, and $\phi = 2\sqrt{gh}$.

The matrices A, B and C have the form

$$A = -\begin{bmatrix} u & 0 & \phi/2 \\ 0 & u & 0 \\ \phi/2 & 0 & u \end{bmatrix}, \quad B = -\begin{bmatrix} v & 0 & 0 \\ 0 & v & \phi/2 \\ 0 & \phi/2 & v \end{bmatrix}, \quad C = \begin{bmatrix} 0 & f & 0 \\ -f & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{45}$$

where f is the Coriolis term

$$f = \hat{f} + \beta(y - D/2), \quad \beta = \frac{\partial f}{\partial y}, \quad \forall\, y \in [0, D], \tag{46}$$

with $\hat{f}$ and β constants. We assume periodic solutions in the x direction for all three state variables while in the y direction $v(x, 0, t) = v(x, D, t) = 0, \quad x \in [0, L], \quad t \in (0, t_{\mathrm{f}}]$, and Neumann boundary condition are considered for u and ϕ.

The numerical scheme is implemented in Fortran and uses a sparse matrix environment. For operations with sparse matrices we utilize SPARSEKIT library [48], and the sparse linear systems resulted from quasi-Newton iterations is solved using MGMRES library [9, 33, 49]. More details on the implementation can be found in [61].

6.2. Smoothing experimental settings

To test the HMC smoothers with SWE model in the context for data assimilation, we construct an assimilation window of length 91 units, with 10 observations distributed over the window. Here the observations are linearly related to model state with $\mathcal{H} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. 4D-Var is carried out in both high-fidelity space (Full 4D-Var) and reduced-order space (Reduced 4D-Var) against the HMC sampling smoother in the following settings

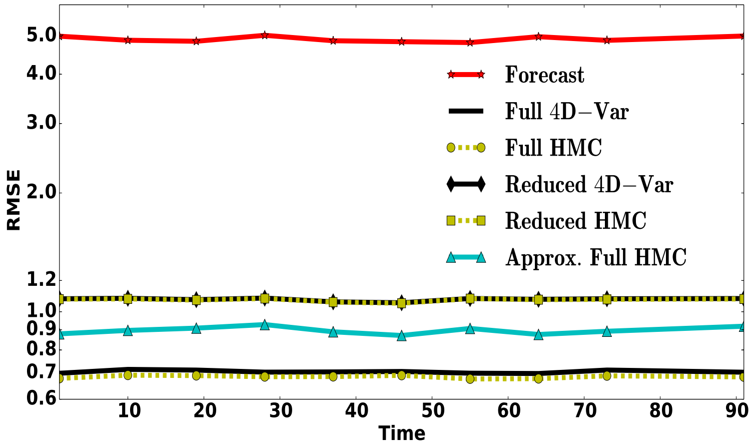
i) Sampling the high-fidelity space using the original HMC smoother [5] ("Full HMC"),

ii) Sampling the reduced space, i.e. sampling (15a) ("Reduced HMC"),

iii) Sampling the high-fidelity space with approximate gradients, i.e. sampling (17a) ("Approximate Full HMC"),

In the three cases, the symplectic integrator used is the position Verlet (11) with step size parameters tuned empirically through a preprocessing step. Higher order integrators [6, 50] and automatic tuning of parameters should be considered when theses algorithms are applied to more complicated, e.g., when $\mathcal{H}$ is nonlinear or when the Gaussian prior assumption is relaxed. The reduced basis $\mathbf{V}$ is constructed using initial trajectories of the high-fidelity forward and adjoint models as well as the associated gradient of the full cost function [61]. Later on this basis is updated using the current proposal and the corresponding trajectories.

Figure 1: Data assimilation results using 4D-Var schemes, and HMC smoother, in both high-fidelity space in reduced-order space. Errors for HMC smoother are obtained for 100 ensemble members with 25 burn-in steps, and 5 mixing steps. The steps size for the symplectic integrator is empirically tuned and unified to $T = 0.1$ with $h = 0.01$, and $m = 10$.



6.3. Numerical results

Due to the simple settings described above the posterior distribution is not expected to deviate notably from a Gaussian. This will enable us to easily test the quality of the ensemble by testing the first two moments generated from the ensemble.

The mean of the ensemble generated by HMC smoother is an MVUE of the posterior mean, and we are interested in comparing it against the 4D-Var solution. Figure 1 shows the Root mean squared (RMSE) errors associated with the 4D-Var and HMC estimates of the posterior mean. The size of the ensemble generated by the different HMC smoothers here is $\textsc{Nens} = 100$. We see clearly that the MVUE generated by HMC in both full and reduced space is at least as good as the 4D-Var minimizer. It is obvious that using Algorithm 2 to sample the full space, while approximating the gradient using reduced-space information, results in an analysis that is better than the case where the sampler is limited to the reduced space. In addition to testing the quality of the analysis (first-order moment here), we are interested in quantifying the quality of the analysis error covariance matrix generated by HMC. For reference we use HMC in full space to sample $\textsc{Nens} = 1000$ members to produce a good estimate $\mathbf{A}_0^{\text{ens}} \approx \mathbf{A}_0$. In the cases of reduced space sampling and approximate sampling in the full space we fix the ensemble size to $\textsc{Nens} = 100$. To compare analysis error covariances obtained in the different scenarios we perform a statistical test of the hypothesis $\mathbf{H}_0 : \Sigma_1 = \Sigma_2$ for the equality of two covariance matrices. Since the state space dimension can be much larger than ensemble size, we choose the test statistic [53] that works in high dimensional settings. Assume we have two probability distributions with covariance matrices Σ_1, Σ_2 respectively, and consider sample estimates $\mathbf{S}_1$, $\mathbf{S}_2$ obtained using ensembles of sizes n_1 and n_2, respectively. The test statistic t^*_{mn} defined in (47) asymptotically follows a standard normal distribution in the limit of large ensemble size and state space dimension. At a significance level α, the two sided test $\mathbf{H}_0 : \Sigma_1 = \Sigma_2$ is rejected only if $|t^*_{mn}| > z_{\alpha/2}$, where $Z = \mathcal{N}(0,)$, and $\mathcal{P}(Z \geq z_{\alpha/2}) = \alpha/2$.

$$
\begin{aligned}
t^*_{mn} &= \frac{t_{mn}}{\hat{\theta}} \,, \\
t_{mn} &= \left(1 - \frac{n_1 - 2}{\eta_1} \mathrm{Tr}\left(\mathbf{S}_1^2\right)\right) + \left(1 - \frac{n_2 - 2}{\eta_2} \mathrm{Tr}\left(\mathbf{S}_2^2\right)\right) - 2\mathrm{Tr}(\mathbf{S}_1 \mathbf{S}_2) - \frac{n_1}{\eta_1} \left(\mathrm{Tr}(\mathbf{S}_1)\right)^2 - \frac{n_2}{\eta_2} \left(\mathrm{Tr}(\mathbf{S}_2)\right)^2 \,, \\
\eta_1 &= (n_1 + 2)(n_1 - 1), \quad \eta_2 = (n_2 + 2)(n_2 - 1) \,, \qquad (47) \\
n &= n_1 + n_2, \quad \mathbf{S} = \frac{n_1}{n} \mathbf{S}_1 + \frac{n_2}{n} \mathbf{S}_2 \,, \\
\hat{\theta} &= \sqrt{4a^2 \left(\frac{n_1 + n_2}{n_1 n_2}\right)^2} \,, \quad a = \frac{n^2}{(n+2)(n-1)} \left(\mathrm{Tr}(\mathbf{S}^2) - \frac{(\mathrm{Tr}(\mathbf{S}))^2}{n}\right) \,.
\end{aligned}
$$

Table 1 shows the results of the tests conducted to compare the covariance matrices. In the case of sampling in the reduced space the null hypothesis is rejected due to strong evidence based on the samples' estimates. For the approximate full space sampling at a significance level $\alpha = 0.01$ there is no significant evidence to supports rejection. This gives a strong indication that the ensemble generated in the second case describes the uncertainty in the analysis much better than the first case. The test results at least don't oppose the conclusion that sampling (17a) using Algorithm 2 results in ensembles capable of estimating the posterior covariance matrix.

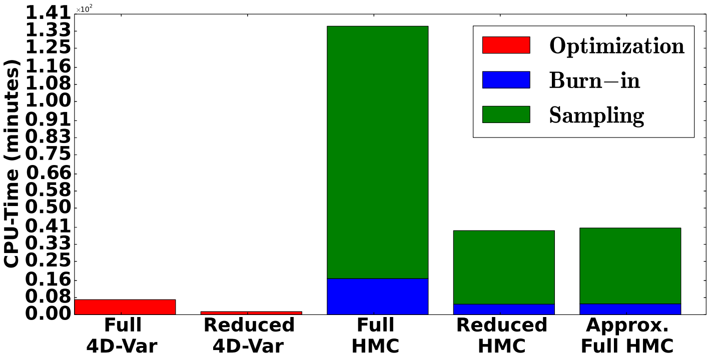
Table 1: Results of statistical tests conducted to compare covariance matrices obtained by HMC smoother in the three scenarios. $\mathbf{A}_0$ is the true posterior covariance of the distribution (2). $\widetilde{\mathbf{A}}_0$ is the true posterior covariance of the distribution with negative-log given by (15), while $\widetilde{\mathbf{A}}_0^{\mathrm{ens}}$ is the ensemble-based approximation obtained by Algorithm 2. $\widehat{\mathbf{A}}_0$ is the true posterior covariance of the distribution with negative-log given by (17), while $\widehat{\mathbf{A}}_0^{\mathrm{ens}}$ is the ensemble-based approximation obtained by Algorithm 2.

Test			Ensemble statistics	Test-statistic
1	Sampling the reduced space	$\mathbf{H_0} : \mathbf{A}_0 = \widetilde{\mathbf{A}}_0$ $\mathbf{H_a} : \mathbf{A}_0 \neq \widetilde{\mathbf{A}}_0$	$n_1 = 1000$, $n_2 = 100$ $\mathbf{S}_1 = \mathbf{A}_0^{\mathrm{ens}}$, $\mathbf{S}_2 = \widetilde{\mathbf{A}}_0^{\mathrm{ens}}$	$t^*_{nm} = 61.0258$
2	Sampling the full space with approximate gradient	$\mathbf{H_0} : \mathbf{A}_0 = \widehat{\mathbf{A}}_0$ $\mathbf{H_a} : \mathbf{A}_0 \neq \widehat{\mathbf{A}}_0$	$n_1 = 1000$, $n_2 = 100$ $\mathbf{S}_1 = \mathbf{A}_0^{\mathrm{ens}}$, $\mathbf{S}_2 = \widehat{\mathbf{A}}_0^{\mathrm{ens}}$	$t^*_{nm} = 2.4514$

6.4. Computational costs

The computational cost for HMC smoother in full space is much higher than the cost of 4D-Var, however it comes with the advantage of generating a consistent estimate of the analysis error covariance matrix. The bottleneck of HMC smoother is the propagation of the forward and backward model to evaluate the gradient of the potential energy. Using surrogate models radically reduces the computational cost. A detailed discussion of the computational cost in terms of model propagation can be found in [5]. Here we report the CPU time of the different scenarios as shown in Figure 2 and Table 2. The HMC CPU-time also depends on the settings of the parameters and the size of the ensemble. Following [5] we compare the CPU-times to generate 30 ensemble members. The

Figure 2: Data assimilation results using 4D-Var schems, and HMC smoother, in both high-fidelity space in reduced-order space. CPU-times for HMC smoother are obtained for 30 ensemble members with 25 burn-in steps, and 5 mixing steps. The steps size for the symplectic integrator is empirically tuned and unified to $T = 0.1$ with $h = 0.01$, and $m = 10$. The red color represents the CPU-time spent during optimization steps only. Blue and Green colors, respectively, represent CPU-time spent during the burn-in and the sampling(and mixing) steps.



CPU-times are almost similar when the two strategies in Algorithm 2 are applied, and both are approximately four times faster than the original HMC smoother. The online cost of the approximate smoother is still higher than the cost of 4D-Var in full space, however it is notably reduced by using information coming from a reduced space. The cost can be further reduced by cleverly tuning the sampler parameters or projecting the observation operator and observation error statistics in the reduced space. These ideas will be considered in the future to further reduce the cost of the HMC sampling smoother. It is very important to highlight the fact that the goal is not just to find an anlysis state but to approximate the whole posterior distribution. Despite the high cost of the HMC smoother, we obtain a consistent description of the uncertainty of the analysis state, e.g. an estimate of the posterior covariances.

Table 2: Data assimilation results using 4D-Var schemes, and HMC smoother, in both high-fidelity space in reduced-order space. CPU-times for HMC smoother are obtained for 30 ensemble members with 25 burn-in steps, and 5 mixing steps. The steps size for the symplectic integrator is empirically tuned and unified to $T = 0.1$ with $h = 0.01$, and $m = 10$.

Cost	**Experiment**							
	4D-Var		**HMC Smoother**					
	high-fidelity space	**reduced-order space**	**high-fidelity space**		**reduced-fidelity space**		**high-fidelity space with approximate gradient**	
			average per ensemble member	*total*	*average per ensemble member*	*total*	*average per ensemble member*	*total*
CPU-time (minutes)	7.04	1.44	0.68	118.42	0.20	34.50	0.20	35.58

7. Conclusions and Future Work

The HMC sampling smoother is developed as a general ensemble-based data assimilation framework to solve the non-Gaussian four-dimensional data assimilation problem. The original formulation of the HMC smoother works with the full dimensional model. It provides a consistent description of the posterior distribution, however it is very expensive due to the necessary large number of full model runs. The HMC sampling smoother employs reduced-order approximations of the model dynamics. It achieves computational efficiency while retaining most of the accuracy of the full space HMC smoother. The formulations discussed here still assume a Gaussian prior at the initial time, which is a weak assumption since the forward propagation through nonlinear model dynamics will result in a non-Gaussian likelihood. This assumption, however, can be easily relaxed using a mixture of Gaussians to represent the background at the initial time; this will be considered in future work. We plan to explore the possibility of using the KL-Divergence measure between the high fideltity distribution and both the projected and the approximate posterior distribution, to guide the optimal choice of the size of reduced-order basis. In future work we will also consider incorporating an HMC sampler capable of automatically tuning the parameters of the symplectic integrator, such as NUTS [29], in order to further enhance the smoother performance.

Acknowledgments

The work of Dr. Răzvan Stefanescu and Prof. Adrian Sandu was supported by awards NSF CCF–1218454, NSF DMS–1419003, AFOSR FA9550–12–1–0293–DEF, AFOSR 12-2640-06, and by the Computational Science Laboratory at Virginia Tech.

[1] K. Afanasiev and M. Hinze. Adaptive Control of a Wake Flow Using Proper Orthogonal Decomposition. Lecture Notes in Pure and Applied Mathematics, 216:317–332, 2001.

[2] A. C. Antoulas. Approximation of large-scale dynamical systems. SIAM, Philadelphia, 2005.

[3] P. Astrid, S. Weiland, K. Willcox, and T. Backx. Missing point estimation in models described by Proper Orthogonal Decomposition. IEEE Transactions on Automatic Control, 53(10):2237–2251, 2008.

[4] A. Attia, V. Rao, and A. Sandu. A sampling approach for four dimensional data assimilation. In Proceedings of the Dynamic Data Driven environmental System Science Conference, 2014.

[5] A. Attia, V. Rao, and A. Sandu. A Hybrid Monte-Carlo sampling smoother for four dimensional data assimilation. Arxiv preprint: `http://arxiv.org/abs/1505.04724`, 2015.

[6] A. Attia and A. Sandu. A hybrid Monte Carlo sampling filter for non-gaussian data assimilation. AIMS Geosciences, 1(geosci-01-00041):41–78, 2015.

[7] M. Barrault, Y. Maday, N. D. Nguyen, and A. T. Patera. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. Comptes Rendus Mathematique. Académie des Sciences, 339(9):667–672, 2004.

[8] M. Barrault, Y. Maday, N.C. Nguyen, and A.T. Patera. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. Comptes Rendus Mathematique, 339(9):667–672, 2004.

[9] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst. Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition. SIAM, Philadelphia, PA, 1994.

[10] D.A. Bistrian and I.M. Navon. Comparison of optimized Dynamic Mode Decomposition vs POD for the shallow water equations model reduction with large-time-step observations . Technical report, Submited to International Journal for Numerical Methods in Fluids, 2014.

[11] Y. Cao, J. Zhu, I.M. Navon, and Z. Luo. A reduced-order approach to four-dimensional variational data assimilation using Proper Orthogonal Decomposition. International Journal for Numerical Methods in Fluids, 53(10):1571–1584, 2007.

[12] K. Carlberg, C. Bou-Mosleh, and C. Farhat. Efficient non-linear model reduction via a least-squares Petrov-Galerkin projection and compressive tensor approximations. International Journal for Numerical Methods in Engineering, 86(2):155–181, 2011.

[13] K. Carlberg, R. Tuminaro, and P. Boggsz. Efficient structure-preserving model reduction for nonlinear mechanical systems with application to structural dynamics. preprint, Sandia National Laboratories, Livermore, CA 94551, USA, 2012.

[14] S. Chaturantabut. Dimension Reduction for Unsteady Nonlinear Partial Differential Equations via Empirical Interpolation Methods. Technical Report TR09-38,CAAM, Rice University, 2008.

[15] S. Chaturantabut and D.C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. SIAM Journal on Scientific Computing, 32(5):2737–2764, 2010.

[16] S. Chaturantabut and D.C. Sorensen. A state space error estimate for POD-DEIM nonlinear model reduction. SIAM Journal on Numerical Analysis, 50(1):46–63, 2012.

[17] T.M. Cover and J.A. Thomas. Elements of information theory. John Wiley & Sons, 2012.

[18] M. Dihlmann and B. Haasdonk. Certified PDE-constrained parameter optimization using reduced basis surrogate models for evolution problems. Submitted to the Journal of Computational Optimization and Applications, 2013.

[19] G. Dimitriu, I.M. Navon, and R. Ştefănescu. Application of POD-DEIM Approach for Dimension Reduction of a Diffusive Predator-Prey System with Allee Effect. In Large-Scale Scientific Computing, pages 373–381. Springer, 2014.

[20] Z. Drmac and S. Gugercin. A New Selection Operator for the Discrete Empirical Interpolation Method – improved a priori error bound and extensions, 2015.

[21] S. Duane, A.D. Kennedy, J. Brian B.J. Pendleton, and D. Roweth. Hybrid Monte Carlo. Physics Letters B, 195(2):216–222, 1987.

[22] G. Evensen and P.J. van Leeuwen. An ensemble Kalman smoother for nonlinear dynamics . Monthly Weather Review, 128:1852–1867, 2000.

[23] R. Everson and L. Sirovich. Karhunen-Loeve procedure for gappy data. Journal of the Optical Society of America A, 12:1657–64, 1995.

[24] F Fang, CC Pain, IM Navon, MD Piggott, GJ Gorman, PE Farrell, PA Allison, and AJH Goddard. A POD reduced-order 4D-Var adaptive mesh ocean modelling approach. International Journal for Numerical Methods in Fluids, 60(7):709–732, 2009.

[25] M. Girolami and B. Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(2):123–214, 2011.

[26] M.A. Grepl and A.T. Patera. A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations. ESAIM: Mathematical Modelling and Numerical Analysis, 39(01):157–181, 2005.

[27] B. Gustafsson. An alternating direction implicit method for solving the shallow water equations. Journal of Computational Physics, 7:239–254, 1971.

[28] R. Heikes and D.A. Randall. Numerical integration of the shallow-water equations on a twisted icosahedral grid. part i: Basic design and results of tests. Monthly Weather Review, 123(6):1862–1880, 1995.

[29] M.D. Homan and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. The Journal of Machine Learning Research, 15(1):1593–1623, 2014.

[30] H. Hotelling. Analysis of a complex of statistical variables with principal components. Journal of Educational Psychology, 24:417–441, 1933.

[31] R.E. Kalman. A new approach to linear filtering and prediction problems . Transaction of the ASME- Journal of Basic Engineering, 82:35–45, 1960.

[32] K. Karhunen. Zur spektraltheorie stochastischer prozesse. Annales Academiae Scientarum Fennicae, 37, 1946.

[33] C. T. Kelley. Iterative Methods for Linear and Nonlinear Equations. Number 16 in Frontiers in Applied Mathematics. SIAM, 1995.

[34] CG Khatri. Some results for the singular normal multivariate regression models. Sankhyā: The Indian Journal of Statistics, Series A, pages 267–280, 1968.

[35] K. Kunisch and S. Volkwein. Proper Orthogonal Decomposition for Optimality Systems. Math. Modelling and Num. Analysis, 42:1–23, 2008.

[36] C. Lieberman, K. Willcox, and O. Ghattas. Parameter and state model reduction for large-scale statistical inverse problems. SIAM Journal on Scientific Computing, 32(5):2523–2542, 2010.

[37] M.M. Loève. Probability Theory. Van Nostrand, Princeton, NJ, 1955.

[38] E.N. Lorenz. Empirical Orthogonal Functions and Statistical Weather Prediction. Technical report, Massachusetts Institute of Technology, Dept. of Meteorology, 1956.

[39] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. The Journal of Chemical Physics, 21(6):1087–1092, 1953.

[40] R.M. Neal. MCMC using Hamiltonian dynamics. Handbook of Markov chain Monte Carlo, 2011.

[41] N.C. Nguyen, A.T. Patera, and J. Peraire. A 'best points' interpolation method for efficient approximation of parametrized function. International Journal for Numerical Methods in Engineering, 73:521–543, 2008.

[42] B.R. Noack, K. Afanasiev, M. Morzynski, G. Tadmor, and F. Thiele. A hierarchy of low-dimensional models for the transient and post-transient cylinder wake. Journal of Fluid Mechanics, 497:335–363, 2003.

[43] A.T. Patera and G. Rozza. Reduced basis approximation and a posteriori error estimation for parametrized partial differential equations, 2007.

[44] C.R. Rao. Linear statistical inference and its applications, volume 22. John Wiley & Sons, 2009.

[45] S.S. Ravindran. Adaptive Reduced-Order Controllers for a Thermal Flow System Using Proper Orthogonal Decomposition. Journal of Scientific Computing, 23:1924–1942, 2002.

[46] C.W. Rowley, I. Mezic, S. Bagheri, P.Schlatter, and D.S. Henningson. Spectral analysis of nonlinear flows. Journal of Fluid Mechanics, 641:115–127, 2009.

[47] G. Rozza, D.B.P. Huynh, and A.T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. Archives of Computational Methods in Engineering, 15(3):229–275, 2008.

[48] Y. Saad. Sparsekit: a basic tool kit for sparse matrix computations. Technical Report, Computer Science Department, University of Minnesota, 1994.

[49] Y. Saad. Iterative Methods for Sparse Linear Systems. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 2003.

[50] J.M. Sanz-Serna. Markov chain Monte Carlo and numerical differential equations. In Current Challenges in Stability Issues for Numerical Differential Equations, pages 39–88. Springer, 2014.

[51] J.M. Sanz-Serna and M-P.Calvo. Numerical Hamiltonian problems, volume 7. Chapman & Hall London, 1994.

[52] P.J. Schmid. Dynamic mode decomposition of numerical and experimental data. Journal of Fluid Mechanics, 656:5–28, 2010.

[53] J.R. Schott. A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. Computational Statistics & Data Analysis, 51(12):6535–6542, 2007.

[54] L. Sirovich. Turbulence and the dynamics of coherent structures. I. Coherent structures. Quarterly of Applied Mathematics, 45(3):561–571, 1987.

[55] L. Sirovich. Turbulence and the dynamics of coherent structures. II. Symmetries and transformations. Quarterly of Applied Mathematics, 45(3):573–582, 1987.

[56] L. Sirovich. Turbulence and the dynamics of coherent structures. III. Dynamics and scaling. Quarterly of Applied Mathematics, 45(3):583–590, 1987.

[57] R. Stefanescu and I.M. Navon. POD/DEIM Nonlinear model order reduction of an ADI implicit shallow water equations model. Journal of Computational Physics, 237:95–114, 2013.

[58] R. Stefanescu, I.M. Navon, H. Fuelberg, and M. Marchand. 1D+4D–VAR Data Assimilation of lightning with WRFDA model using nonlinear observation operators. Technical report, Arxiv preprint: http://arxiv.org/abs/1211.2521, 2013.

[59] R. Stefanescu, A. Sandu, and I.M. Navon. Comparison of POD Reduced Order Strategies for the Nonlinear 2D Shallow Water Equations. International Journal for Numerical Methods in Fluids, 76(8):497–521, 2014.

[60] R. Ştefănescu, A. Sandu, and I.M. Navon. Comparison of POD reduced order strategies for the nonlinear 2D shallow water equations. International Journal for Numerical Methods in Fluids, 76(8):497–521, 2014.

[61] R. Ştefănescu, A. Sandu, and I.M. Navon. POD/DEIM reduced-order strategies for efficient four dimensional variational data assimilation. Journal of Computational Physics, 295:569–595, 2015.

[62] G. Tissot, L. Cordir, N. Benard, and B. Noack. Model reduction using Dynamic Mode Decomposition. Comptes Rendus Mécanique, 342(6-7):410–416, 2014.

[63] PTM Vermeulen and AW Heemink. Model-reduced variational data assimilation. Monthly Weather Review, 134(10):2888–2899, 2006.