atmosphere

MDPI

Article

Cluster Sampling Filters for Non-Gaussian Data Assimilation

Ahmed Attia [1,†], Azam Moosavi [2,†] and Adrian Sandu [2,*,†]

[1] Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL60439, USA; attia@mcs.anl.gov or attia@vt.edu

[2] Computational Science Laboratory, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA; azmosavi@vt.edu

* Correspondence: sandu@cs.vt.edu; Tel.: +1-540-231-2193

† These authors contributed equally to this work.

Received: 20 March 2018; Accepted: 2 May 2018; Published: date

Abstract: This paper presents a fully non-Gaussian filter for sequential data assimilation. The filter is named the "*cluster sampling filter*", and works by directly sampling the posterior distribution following a Markov Chain Monte-Carlo (MCMC) approach, while the prior distribution is approximated using a Gaussian Mixture Model (GMM). Specifically, a clustering step is introduced after the forecast phase of the filter, and the prior density function is estimated by fitting a GMM to the prior ensemble. Using the data likelihood function, the posterior density is then formulated as a mixture density, and is sampled following an MCMC approach. Four versions of the proposed filter, namely $\mathcal{C}\ell$MCMC, $\mathcal{C}\ell$HMC, MC-$\mathcal{C}\ell$HMC, and MC-$\mathcal{C}\ell$HMC are presented. $\mathcal{C}\ell$MCMC uses a Gaussian proposal density to sample the posterior, and $\mathcal{C}\ell$HMC is an extension to the Hamiltonian Monte-Carlo (HMC) sampling filter. MC-$\mathcal{C}\ell$MCMC and MC-$\mathcal{C}\ell$HMC are multi-chain versions of the cluster sampling filters $\mathcal{C}\ell$MCMC and $\mathcal{C}\ell$HMC respectively. The multi-chain versions are proposed to guarantee that samples are taken from the vicinities of all probability modes of the formulated posterior. The new methodologies are tested using a simple one-dimensional example, and a quasi-geostrophic (QG) model with double-gyre wind forcing and bi-harmonic friction. Numerical results demonstrate the usefulness of using GMMs to relax the Gaussian prior assumption especially in the HMC filtering paradigm.

Keywords: data assimilation; ensemble filters; markov chain monte-carlo sampling; hamiltonian monte-carlo; gaussian mixture models

1. Introduction

Data assimilation (DA) is a complex process that involves combining information from different sources in order to produce accurate estimates of the true state of a physical system such as the atmosphere. Sources of information include computational models of the system, a background probability distribution, and observations collected at discrete time instances. With model state denoted by $\mathbf{x} \in \mathbb{R}^{\mathrm{N_{var}}}$, the prior probability density $\mathcal{P}^{\mathrm{b}}(\mathbf{x})$ encapsulates the knowledge about the system state before incorporating any other source of information such as the observations. Let $\mathbf{y} \in \mathbb{R}^{\mathrm{m}}$ be a measurement (observation) vector. The observation likelihood function $\mathcal{P}(\mathbf{y}|\mathbf{x})$ quantifies the mismatch between the model predictions (of observed quantities) and the collected measurements. A standard application of Bayes' theorem provides the posterior probability distribution $\mathcal{P}(\mathbf{x}|\mathbf{y})$ that provides an improved description of the unknown true state of the system of interest.

In the ideal case where the underlying probability distributions are Gaussian, the model dynamics is linear, and the observations are linearly related to the model state, the posterior can be obtained analytically for example by applying Kalman filter (KF) Equations [1,2]. For large dimensional problems the computational cost of the standard Kalman filter is prohibitive, and in practice the

probability distributions are approximated using small ensembles. The ensemble-based approximation has led to the ensemble Kalman filter (EnKF) family of methods [3–6]. Several modifications of EnKF, for example [7–12], have been introduced in the literature to solve practical DA problems of different complexities.

One of the drawbacks of the EnKF family is the reliance on an ensemble update formula that comes from the linear Gaussian theory. Several approaches have been proposed in the literature to alleviate the limitations of the Gaussian assumptions. The maximum likelihood ensemble filter (MLEF) [13–15] computes the maximum a posteriori estimate of the state in the ensemble space. The iterative EnKF [10,16] (IEnKF) extends MLEF to handle nonlinearity in models as well as in observations. IEnKF, however, assumes that the underlying probability distributions are Gaussian and the analysis state is best estimated by the posterior mode.

These families of filters can generally be tuned (e.g., using inflation and localization) for optimal performance on the problem at hand. However, if the posterior is a multimodal distribution, these filters are expected to diverge, or at best capture a single probability mode, especially in the case of long-term forecasts. Only a small number of filtering methodologies designed to work in the presence of highly non-Gaussian errors are available, and their efficiency with realistic models is yet to be established. These promising methods can be classified into two classes, particle filtering, and MCMC sampling. In this work, we focus on using MCMC to sample the posterior distribution.

The Hybrid/Hamiltonian Monte Carlo (HMC) sampling filter was proposed in [17] as a fully non-Gaussian filtering algorithm, and has been extended to the four-dimensional (smoothing) setting in [17–20]. The HMC sampling filter is a sequential DA filtering scheme that works by directly sampling the posterior probability distribution via an HMC approach [21,22]. The HMC filter is designed to handle cases where the underlying probability distributions are non-Gaussian. Nevertheless, the first HMC formulation presented in [17] assumes that the prior distribution can always be approximated by a Gaussian distribution. This assumption was introduced for simplicity of implementation; however, it can be too restrictive in many cases, and may lead to inaccurate conclusions. This strong assumption needs to be relaxed in order to accurately sample from the true posterior, while preserving computational efficiency.

In this work, we propose relaxing the Gaussian prior assumption, using a Gaussian Mixture Model (GMM) to approximate the prior distribution. Specifically, in the forecast phase of the filter, the prior is represented by a GMM that is fitted to the forecast ensemble via a clustering step. The posterior is formulated accordingly. In the analysis step the resulting mixture posterior is sampled following an MCMC approach. The resulting algorithm is named the "*cluster MCMC* ($\mathcal{C}\ell$MCMC) *sampling filter*", and the version in which HMC is used to sample the posterior is named *cluster HMC* ($\mathcal{C}\ell$HMC) *sampling filter*. In order to improve the sampling from the mixture posterior, more efficient versions namely "*multi-chain* (MC-$\mathcal{C}\ell$MCMC)", and "*multi-chain* $\mathcal{C}\ell$HMC (MC-$\mathcal{C}\ell$HMC)", filters are also discussed.

The proposed MCMC filtering algorithms are not suggested as replacements for EnKF in the linear-Gaussian settings. MCMC algorithms are generally expensive, compared to EnKF, and should be considered only when the linear-Gaussian assumption is highly violated, which can cause the conventional EnKF to fail. The numerical experiments presented herein, are carried out in both linear and nonlinear settings. Experiments with linear settings aim to compare the performance of the proposed algorithms, in the presence of benchmark results produced by EnKF. This has the benefit of demonstrating the advantage of replacing the Gaussian prior with a GMM, for MCMC sampling, even in the simplified linear settings. Numerical results with nonlinear settings, where EnKF fails, suggest that the proposed relaxation of the Gaussian-prior assumption is beneficial, especially for the sequential application of MCMC sampling filters in the presence of nonlinearities.

Using a GMM to approximate the prior density, given the forecast ensemble, was presented in [11,23] as a means to solve the nonlinear filtering problem. In [23], a continuous approximation of the prior density was built as a sum of Gaussian kernels, where the number of kernels is equal to the ensemble size. Assuming a Gaussian likelihood function, the posterior was formulated as a GMM

with updated mixture parameters. The updated means and covariance matrices of the GMM posterior were obtained by applying the convolution rule of Gaussians to the prior mixture components and the likelihood, and the analysis ensemble was generated by direct sampling from the GMM posterior. On the other hand, the approach presented in [11] works by fitting a GMM to the prior ensemble with the number of mixture components detected using Akaike information criterion. The EnKF equations are applied to each of the components in the mixture distribution to generate an analysis ensemble from the GMM posterior.

Unlike the existing approaches [11,23], the methodology proposed herein is fully non-Gaussian, and does not limit the posterior density to a Gaussian mixture distribution or Gaussian likelihood functions. Moreover, the posterior distribution is directly sampled, and is not approximated by a Gaussian mixture distribution.

The remaining part of the paper is organized as follows. Section 2 reviews the original formulation of the HMC sampling filter. Section 3 explains how GMM can be used to approximate the prior distribution, and presents the new cluster sampling filters. Numerical results and discussions are presented in Section 4. Conclusions are drawn in Section 5.

2. The HMC Sampling Filter

In this section we present a brief overview of the HMC sampling methodology, followed by the original formulation of the HMC sampling filter.

2.1. HMC Sampling

HMC sampling follows an auxiliary-variable approach [24,25] to accelerate the sampling process of Markov chain Monte-Carlo (MCMC) algorithms. In this approach, the MCMC sampler is devoted to sampling the joint probability density of the target variable, along with an auxiliary variable. The auxiliary variable is chosen carefully to allow for the construction of a Markov chain that mixes faster, and is easier to simulate than sampling the marginal density of the target variable [26].

The main component of the HMC sampling scheme is an auxiliary Hamiltonian system that plays the role of the proposal (jumping) distribution. The Hamiltonian dynamical system operates in a phase space of points $(\mathbf{p}, \mathbf{x}) \in \mathbb{R}^{2\mathrm{N_{var}}}$, where the individual variables are the position $\mathbf{x} \in \mathbb{R}^{\mathrm{N_{var}}}$, and the momentum $\mathbf{p} \in \mathbb{R}^{\mathrm{N_{var}}}$. The total energy of the system, given the position and the momentum, is described by the Hamiltonian function $H(\mathbf{p}, \mathbf{x})$. A general formulation of the Hamiltonian function (the Hamiltonian) of the system is given by:

$$H(\mathbf{p}, \mathbf{x}) = \frac{1}{2}\,\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} - \log(\phi(\mathbf{x})) = \frac{1}{2}\,\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} + \mathcal{J}(\mathbf{x})\,, \tag{1}$$

where $\mathbf{M} \in \mathbb{R}^{\mathrm{N_{var}}\times\mathrm{N_{var}}}$ is a symmetric positive definite matrix referred to as the *mass matrix*. The first term in the sum (1) quantifies the kinetic energy of the Hamiltonian system, while the second term is the associated potential energy.

The dynamics of the Hamiltonian system in time is described by the following ordinary differential equations (ODEs):

$$\frac{d\mathbf{x}}{dt} = \nabla_{\mathbf{p}}\, H\,, \qquad \frac{d\mathbf{p}}{dt} = -\nabla_{\mathbf{x}}\, H. \tag{2}$$

The time evolution of the system (2) in the phase space is described by the flow: [27,28]

$$\Phi_T : \mathbb{R}^{2\mathrm{N_{var}}} \to \mathbb{R}^{2\mathrm{N_{var}}}, \quad \Phi_T\big(\mathbf{p}(0), \mathbf{x}(0)\big) = \big(\mathbf{p}(T), \mathbf{x}(T)\big), \tag{3}$$

which maps the initial state of the system $(\mathbf{p}(0), \mathbf{x}(0))$ to $(\mathbf{p}(T), \mathbf{x}(T))$, the state of the system at time T. In practical applications, the analytic flow Φ_T is replaced by a numerical solution using a time reversible and symplectic numerical integration method [28,29]. The length of the Hamiltonian

trajectory T can generally be long, and may lead to instability of the time integrator if the step size is set to T. In order to accurately approximate Φ_T , the symplectic integrator typically takes m steps of size $h = T/m$ where h is chosen such as to maintain stability. We will use Φ_T hereafter to represent the numerical approximation of the Hamiltonian flow.

Given the formulation of the Hamiltonian (1), the dynamics of the Hamiltonian system is governed by the equations

$$\frac{d\mathbf{x}}{dt} = \mathbf{M}^{-1}\mathbf{p}\,, \qquad \frac{d\mathbf{p}}{dt} = -\nabla_{\mathbf{x}}\mathcal{J}(\mathbf{x})\,. \tag{4}$$

The canonical probability distribution of the state $(\mathbf{p}, \mathbf{x})$ of the Hamiltonian system in the phase space $\mathbb{R}^{2N_{\text{var}}}$, up to a scaling factor, is given by

$$\exp\left(-H(\mathbf{p},\mathbf{x})\right) = \exp\left(-\frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} - \mathcal{J}(\mathbf{x})\right) \propto \exp\left(-\frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}\right)\phi(\mathbf{x}). \tag{5}$$

The product form of this joint probability distribution shows that the two variables $\mathbf{p}$, and $\mathbf{x}$ are independent [29]. The marginal distribution of the momentum variable is Gaussian, $\mathbf{p} \sim \mathcal{N}(0, \mathbf{M})$, while the marginal distribution of the position variable is proportional to the negative-logarithm (negative-log) of the potential energy, that is $\mathbf{x} \sim f(\mathbf{x}) \propto \phi(\mathbf{x}) = \exp\left(-\mathcal{J}(\mathbf{x})\right)$. Here $f(\mathbf{x})$ is the normalized marginal density of the position variable, while $\phi(\mathbf{x})$ drops the scaling factor (e.g., the normalization constant) of the density function.

In order to draw samples $\{\mathbf{x}(e)\}_{e=1,2,\ldots,N_{\text{ens}}}$ from a given probability distribution $f(\mathbf{x}) \propto \phi(\mathbf{x})$, HMC makes the following analogy with the Hamiltonian mechanical system (2). The state $\mathbf{x}$ is *viewed* as a position variable, and an auxiliary momentum variable $\mathbf{p} \sim \mathcal{N}(0,\, \mathbf{M})$ is included. The negative-log of the target probability density $\mathcal{J}(\mathbf{x}) = -\log(\phi(\mathbf{x}))$ is viewed as the potential energy of an auxiliary Hamiltonian system. The kinetic energy of the system is given by the negative-log of the Gaussian distribution of the auxiliary momentum variable. The mass matrix $\mathbf{M}$ is a user-defined parameter that is assumed to be symmetric positive definite. To achieve favorable performance of the HMC sampler, $\mathbf{M}$ is generally assumed to be diagonal, with values on the diagonal chosen to reflect the scale of the components of the target variable under the target density [17,27]. The HMC sampler proceeds by constructing a Markov chain whose stationary distribution is set to the canonical joint density (5). The chain is initialized to some position and momentum values, and at each step of the chain, a Hamiltonian trajectory starting at the current state is constructed to propose a new state. A Metropolis-Hastings-like acceptance rule is used to either accept or reject the proposed state. Since both position and momentum are statistically independent, the retained position samples are actually sampled from the target density $f(\mathbf{x})$. The collected momentum samples are discarded, and the position samples are returned as the samples from the target probability distribution $f(\mathbf{x})$.

The performance of the HMC sampling scheme is greatly influenced by the settings of the Hamiltonian trajectory, that is the choice of the two parameters m, h. The step size h should be small enough to maintain stability, while m should be generally large for the sampler to reach distant points in the state space. The parameters of the Hamiltonian trajectory can be set empirically [27] to achieve an acceptable rejection rate of at most 25% to 30% , or be automatically adapted using tuning schemes such as the No-U-Turn sampler(NUTS) [30], or the Riemannian Manifold HMC sampler (RMHMC) [31].

The ideas presented in this work can be easily extended to incorporate any of the HMC sampling algorithms with automatically tuned parameters. In this paper we tune the parameters of the Hamiltonian trajectory following the empirical approach, and focus on the sampler performance due to the choice of the prior distribution in the sequential filtering context.

2.2. HMC Sampling Filter

In the filtering framework, following a perfect-model approach, the posterior distribution $\mathcal{P}^{\mathrm{a}}(\mathbf{x}_k)$ at a time instance t_k follows from Bayes' theorem:

$$\mathcal{P}^{\mathrm{a}}(\mathbf{x}_k) = \mathcal{P}(\mathbf{x}_k|\mathbf{y}_k) = \frac{\mathcal{P}(\mathbf{y}_k|\mathbf{x}_k)\mathcal{P}^{\mathrm{b}}(\mathbf{x}_k)}{\mathcal{P}(\mathbf{y}_k)} \propto \mathcal{P}(\mathbf{y}_k|\mathbf{x}_k)\mathcal{P}^{\mathrm{b}}(\mathbf{x}_k)\,, \tag{6}$$

where $\mathcal{P}^{\mathrm{b}}(\mathbf{x}_k)$ is the prior distribution, $\mathcal{P}(\mathbf{y}_k|\mathbf{x}_k)$ is the likelihood function, all at time instance t_k. $\mathcal{P}(\mathbf{y}_k)$ acts as a scaling factor and is ignored in the HMC context.

As mentioned in Section 1, the formulation of the HMC sampling filter proposed in [17] assumes that the prior distribution $\mathcal{P}^{\mathrm{b}}(\mathbf{x}_k)$ can be represented by a Gaussian distribution $\mathcal{N}(\mathbf{x}_k^{\mathrm{b}}, \mathbf{B_k})$, that is

$$\mathcal{P}^{\mathrm{b}}(\mathbf{x}_k) = \frac{(2\pi)^{-\frac{\mathrm{N_{var}}}{2}}}{\sqrt{|\mathbf{B}_k|}} \exp\left(-\frac{1}{2}\|\mathbf{x}_k - \mathbf{x}_k^{\mathrm{b}}\|^2_{\mathbf{B}_k^{-1}}\right), \tag{7}$$

where $\mathbf{x}_k^{\mathrm{b}}$, is the background state, and $\mathbf{B}_k \in \mathbb{R}^{\mathrm{N_{var}}\times\mathrm{N_{var}}}$ is the background error covariance matrix. The background state $\mathbf{x}_k^{\mathrm{b}}$ is generally taken as the mean of an ensemble of forecasts $\{\mathbf{x}_k^{\mathrm{b}}(e)\}_{e=1,2,\ldots,\mathrm{N_{ens}}}$, obtained by forward model runs from a previous assimilation cycle. The associated weighted norm is defined as:

$$\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{C}} = (\mathbf{a}-\mathbf{b})^{\mathbf{T}}\mathbf{C}(\mathbf{a}-\mathbf{b}). \tag{8}$$

Under the traditional, yet non-restrictive assumption, that the observation errors are distributed according to a Gaussian distribution with zero mean, and observation error covariance matrix $\mathbf{R}_k \in \mathbb{R}^{\mathrm{m}\times\mathrm{m}}$, the likelihood function takes the form

$$\mathcal{P}(\mathbf{y}_k|\mathbf{x}_k) = \frac{(2\pi)^{-\frac{\mathrm{m}}{2}}}{\sqrt{|\mathbf{R}_k|}} \exp\left(-\frac{1}{2}\|\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}}\right), \tag{9}$$

where $\mathcal{H}_k : \mathbb{R}^{\mathrm{N_{var}}} \to \mathbb{R}^{\mathrm{m}}$ is the observation operator that maps a given state $\mathbf{x}_k$ to the observation space at time instance t_k. The dimension of the observation space m is generally much smaller than the state space dimension, that is $\mathrm{m} \ll \mathrm{N_{var}}$.

The posterior follows immediately from (6), (7), and (9)

$$\mathcal{P}^{\mathrm{a}}(\mathbf{x}_k) \propto \phi(\mathbf{x}_k) = \exp\Big(-\mathcal{J}(\mathbf{x}_k)\Big), \tag{10a}$$

$$\mathcal{J}(\mathbf{x}_k) = \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}_k^{\mathrm{b}}\|^2_{\mathbf{B}_k^{-1}} + \frac{1}{2}\|\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}}, \tag{10b}$$

where $\mathcal{J}(\mathbf{x}_k)$ is the negative-log of the posterior distribution (10b). The derivative of $\mathcal{J}(\mathbf{x}_k)$ with respect to the system state $\mathbf{x}_k$ is given by

$$\nabla_{\mathbf{x}}\mathcal{J}(\mathbf{x}_k) = \mathbf{B}_k^{-1}\left(\mathbf{x}_k - \mathbf{x}_k^{\mathrm{b}}\right) - \mathbf{H}_k^T\,\mathbf{R}_k^{-1}\left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x})\right), \tag{11}$$

where $\mathbf{H}_k = \mathcal{H}_k'(\mathbf{x})$ is the linearized observation operator (e.g., the Jacobian).

The HMC sampling filter [17] proceeds in two steps, namely a forecast step and an analysis step. Given an analysis ensemble of states $\{\mathbf{x}_{k-1}^{\mathrm{a}}(e)\}_{e=1,2,\ldots,\mathrm{N_{ens}}}$ at time t_{k-1}, an ensemble of forecasts at time t_k is generated using the forward model $\mathcal{M}$:

$$\mathbf{x}_k^{\mathrm{b}}(e) = \mathcal{M}_{t_{k-1}\to t_k}\left(\mathbf{x}_{k-1}^{\mathrm{a}}(e)\right), \quad e = 1,2,\ldots,\mathrm{N_{ens}}. \tag{12}$$

In the analysis step, the posterior (10) is sampled by running a HMC sampler with potential energy set to (10b), where $\mathbf{B}_k$ is approximated using the available ensemble of forecasts.

The formulation of the HMC filter presented in [17], and reviewed above, tends to be restrictive due to the assumption that the prior is always approximated by a Gaussian distribution. The prior distribution can be viewed as the result of propagating the posterior of the previous assimilation cycle using model dynamics. In the case of nonlinear model dynamics, the prior distribution is a nonlinear transformation of a non-Gaussian distribution which is generally expected to be non-Gaussian. Tracking the prior distribution exactly however is not possible, and a relaxation assumption must take place.

We propose conducting a more accurate density estimation of the prior, by fitting a GMM to the available prior ensemble, replacing the Gaussian prior with a Gaussian mixture prior.

3. Cluster Sampling Filters

Section 3.1 provides a brief overviews on mixture distributions, and review how a GMM can be used to represent the prior distribution given an ensemble of forecasts. Section 3.2 describes the posterior distribution, and presents the new cluster sampling filters.

3.1. Mixture Models

The probability distribution $\mathcal{P}(\mathbf{x})$ is said to be a mixture of $\mathrm{N_c}$ probability distributions $\{\mathcal{C}_i(\mathbf{x})\}_{i=1,2,\ldots,\mathrm{N_c}}$, if $\mathcal{P}(\mathbf{x})$ takes the form:

$$\mathcal{P}(\mathbf{x}) = \sum_{i=1}^{\mathrm{N_c}} \tau_i \, \mathcal{C}_i(\mathbf{x}) \quad \text{where} \quad \tau_i > 0, \; \forall i \;\; \text{and} \;\; \sum_{i=1}^{\mathrm{N_c}} \tau_i = 1 \,. \tag{13}$$

The weights τ_i are commonly referred to as the mixing weights, and $\mathcal{C}_i(\mathbf{x})$ are the densities of the mixing components.

3.1.1. Gaussian Mixture Models (GMM)

A GMM is a special case of (13) where the mixture components are Gaussian densities, that is $\mathcal{C}_i(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \Theta_i)$ with $\Theta_i = \{\mu_i, \, \mathbf{\Sigma}_i\}$ being the parameters of the ith Gaussian component.

Fitting a GMM to a given data set is one of the most popular approaches for density estimation [32–34]. Given a data set $\{\mathbf{x}(e)\}_{e=1,2,\ldots,\mathrm{N_{ens}}}$, sampled from an unknown probability distribution $\mathcal{P}(\mathbf{x})$, one can estimate the density function $\mathcal{P}(\mathbf{x})$ by a GMM; the parameters of the GMM, i.e., the mixing weights τ_i, the means μ_i, and the covariances $\mathbf{\Sigma}_i$ of the mixture components, can be inferred from the data.

The most popular approach to obtain the maximum likelihood estimate of the GMM parameters is the expectation-maximization (EM) algorithm [34]. EM is an iterative procedure that alternates between two steps, expectation (E) and maximization (M). At iteration $t+1$ the E-step computes the expectation of the complete log-likelihood based on the posterior probability of $\mathbf{x}$ belonging to the ith component, with the parameters $\Theta^{\{t\}}$ from the previous iteration. In particular, the following quantity $Q(\Theta|\Theta^{\{t\}})$ is evaluated:

$$\begin{aligned} Q(\Theta|\Theta^{\{t\}}) &= \sum_{e=1}^{\mathrm{N_{ens}}} \sum_{i=1}^{\mathrm{N_c}} r_{e,i} \, \log\left(\tau_i \, \mathcal{N}\left(\mathbf{x}(e); \Theta_i\right)\right), \\ r_{e,i} &= \frac{\tau_i^{\{t\}} \, \mathcal{N}\left(\mathbf{x}(e); \Theta_i^{\{t\}}\right)}{\sum_{\ell=1}^{\mathrm{N_c}} \tau_\ell^{\{t\}} \, \mathcal{N}\left(\mathbf{x}(e); \Theta_\ell^{\{t\}}\right)} \,. \end{aligned} \tag{14}$$

Here $\Theta = \{\tau_i, \Theta_i\}_{i=1\ldots\mathrm{N_c}}$ is the parameter set of all the mixture components, and $r_{e,i}$ is the probability that the eth ensemble member lies under the ith mixture component.

In the M-step, the new parameters $\Theta^{\{t+1\}} = \arg\max_{\Theta} Q$ are obtained by maximizing the conditional probability Q in (14) with respect to the parameters Θ. The updated parameters $\Theta^{\{t+1\}}$ are given by the analytical formulas:

$$\begin{aligned} \tau_i^{\{t+1\}} &= \frac{\sum_{e=1}^{\mathrm{N_{ens}}} r_{e,i}}{\mathrm{N_{ens}}} = \frac{w_i}{\mathrm{N_{ens}}}\,, \\ \mu_i^{\{t+1\}} &= \sum_{e=1}^{\mathrm{N_{ens}}} \mathbf{x}(e)\,\frac{r_{e,i}}{w_i}\,, \\ \Sigma_i^{\{t+1\}} &= \sum_{e=1}^{\mathrm{N_{ens}}} \left(\mathbf{x}(e) - \mu_i^{\{t+1\}}\right)\left(\mathbf{x}(e) - \mu_i^{\{t+1\}}\right)^T \frac{r_{e,i}}{w_i} \\ \text{where} \quad w_i &= \sum_{e=1}^{\mathrm{N_{ens}}} r_{e,i}\,. \end{aligned} \tag{15}$$

To initialize the parameters for the EM iterations, the mixing weights are simply chosen to be equal $\tau_i = \mathrm{N_c}^{-1}$, the means μ_i can be randomly selected from the given ensemble, and the covariance matrices of the components can be all set to covariance matrix of the full ensemble. Regardless of the initialization, the convergence of the EM algorithm is ensured by the fact that it monotonically increases the observed data log-likelihood at each iteration [34], that is:

$$\sum_{e=1}^{\mathrm{N_{ens}}} \log\left(\sum_{i=1}^{\mathrm{N_c}} \tau_i^{\{t+1\}}\,\mathcal{N}\left(\mathbf{x}(e);\Theta_i^{\{t+1\}}\right)\right) \geq \sum_{e=1}^{\mathrm{N_{ens}}} \log\left(\sum_{i=1}^{\mathrm{N_c}} \tau_i^{\{t\}}\,\mathcal{N}\left(\mathbf{x}(e);\Theta_i^{\{t\}}\right)\right).$$

EM algorithm achieves the improvement of the data log-likelihood indirectly by improving the quantity $Q(\Theta|\Theta^{\{t\}})$ over consecutive iterations, i.e., $Q(\Theta|\Theta^{\{t+1\}}) \geq Q(\Theta|\Theta^{\{t\}})$.

3.1.2. Model Selection

Before EM iterations start, the number of mixture components $\mathrm{N_c}$ must be detected. To decide on the number of components in the prior mixture, model selection is employed. This process refers to the statistical decision of choosing a model, out of a set of candidate models, to give the best trade-off between model fit and complexity. Here, the best number of components $\mathrm{N_c}$ can be selected with common model selection methodologies such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC):

$$\begin{aligned} AIC &= -2\sum_{e=1}^{\mathrm{N_{ens}}} \log\left(\sum_{i=1}^{\mathrm{N_c}} \widehat{\tau}_i\,\mathcal{N}\left(\mathbf{x}(e);\widehat{\Theta}_i\right)\right) + 2\left(3\mathrm{N_c} - 1\right), \\ BIC &= -2\sum_{e=1}^{\mathrm{N_{ens}}} \log\left(\sum_{i=1}^{\mathrm{N_c}} \widehat{\tau}_i\,\mathcal{N}\left(\mathbf{x}(e);\widehat{\Theta}_i\right)\right) + \log(\mathrm{N_{ens}})\left(3\mathrm{N_c} - 1\right), \end{aligned} \tag{16}$$

where $\{\widehat{\tau}_i, \widehat{\Theta}_i\}_{i=1\ldots\mathrm{N_c}}$ is the set of optimal parameters for the candidate GMM model with $\mathrm{N_c}$ components.

The best number of components $\mathrm{N_c}$ minimizes the AIC or BIC criterion [35,36]. The main difference between the two criteria, as explained by the second terms in Equation (16), is that BIC imposes greater penalty on the number of parameters $(3\mathrm{N_c} - 1)$ of the candidate GMM model. For small or moderate numbers of samples BIC often chooses models that are too simple because of its heavy penalty on complexity.

3.2. Cluster Sampling Filters

The prior distribution is approximated by a GMM fitted to the forecast ensemble, e.g., using an EM clustering step. The prior PDF reads:

$$\mathcal{P}^{\mathrm{b}}(\mathbf{x}_k) = \sum_{i=1}^{\mathrm{N_c}} \tau_{k,i}\, \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{k,i}, \boldsymbol{\Sigma}_{k,i}) = \sum_{i=1}^{\mathrm{N_c}} \tau_{k,i} \frac{(2\pi)^{-\frac{\mathrm{Nvar}}{2}}}{\sqrt{|\boldsymbol{\Sigma}_{k,i}|}} \exp\left(-\frac{1}{2}\|\mathbf{x}-\boldsymbol{\mu}_{k,i}\|^2_{\boldsymbol{\Sigma}_{\mathbf{k,i}}{}^{-1}}\right), \tag{17}$$

where the weights $\tau_{k,i}$ quantify the probability that an ensemble member $\mathbf{x}_k(e)$ belongs to the ith component, and $(\boldsymbol{\mu}_{k,i}, \boldsymbol{\Sigma}_{k,i})$ are the mean and the covariance matrix associated with the ith component of the mixture model at time instance t_k.

Assuming Gaussian observation errors, the posterior can be formulated using Equations (6), (9), and (17) as follows:

$$\begin{aligned} f(\mathbf{x}_k) &= \mathcal{P}^{\mathrm{a}}(\mathbf{x}_k) \\ &= \frac{(2\pi)^{-\frac{\mathrm{m}}{2}}}{\sqrt{|\mathbf{R}_k|}} \exp\left(-\frac{1}{2}\|\mathcal{H}_k(\mathbf{x}_k)-\mathbf{y}_k\|^2_{\mathbf{R}_k^{-1}}\right) \sum_{i=1}^{\mathrm{N_c}} \tau_{k,i} \frac{(2\pi)^{-\frac{\mathrm{Nvar}}{2}}}{\sqrt{|\boldsymbol{\Sigma}_{k,i}|}} \exp\left(-\frac{1}{2}\|\mathbf{x}_k-\boldsymbol{\mu}_{k,i}\|^2_{\boldsymbol{\Sigma}_{\mathbf{k,i}}{}^{-1}}\right) \\ &\propto \phi(\mathbf{x}_k) = \sum_{i=1}^{\mathrm{N_c}} \frac{\tau_{k,i}}{\sqrt{|\boldsymbol{\Sigma}_{k,i}|}} \exp\left(-\frac{1}{2}\|\mathbf{x}_k-\boldsymbol{\mu}_{k,i}\|^2_{\boldsymbol{\Sigma}_{\mathbf{k,i}}{}^{-1}} - \frac{1}{2}\|\mathcal{H}_k(\mathbf{x}_k)-\mathbf{y}_k\|^2_{\mathbf{R}_k^{-1}}\right). \end{aligned} \tag{18}$$

In general the posterior PDF (18) will not correspond to a Gaussian mixture due to the nonlinearity of the observation operator. This makes analytical solutions not possible. Here we seek to sample directly from the posterior PDF (18) following a MCMC approach. A proposal distribution and acceptance/rejection criterion, are the backbones of any MCMC sampler.

The simplest proposal is a Gaussian centered around the current state of the Markov chain. Specifically, at the rth iteration of the chain, the proposal is $\mathcal{N}(\mathbf{x}_k^r, \mathbf{B}_k^{\mathrm{ens}})$, where $\mathbf{x}_k^r$ is the current state of the chain, and $\mathbf{B}_k^{\mathrm{ens}}$ is the ensemble-based covariance matrix

$$\mathbf{B}_k^{\mathrm{ens}} = \frac{1}{\mathrm{N_{ens}}-1} \sum_{e=1}^{\mathrm{N_{ens}}} \left(\mathbf{x}_k(e)-\bar{\mathbf{x}}_k\right)\left(\mathbf{x}_k(e)-\bar{\mathbf{x}}_k\right)^T, \tag{19}$$

where $\bar{\mathbf{x}}_k = \sum_{e=1}^{\mathrm{N_{ens}}} \mathbf{x}_k(e)$ is the forecast ensemble mean. The Metropolis-Hastings acceptance/rejection, in this case, can be evaluated using the acceptance probability $u^{(r)} = \phi(\mathbf{x}_k^*)/\phi(\mathbf{x}_k^r)$, where $\mathbf{x}_k^*$ is the proposed state, e.g., sampled from $\mathcal{N}(\mathbf{x}_k^r, \mathbf{B}_k^{\mathrm{ens}})$. We refer to this approach of filtering as the cluster MCMC ($\mathcal{C}\ell$MCMC) sampling filter.

Gradient-based MCMC sampling methods, such as the HMC sampler, use the gradient of the negative log-posterior to direct the sampler towards high-probability regions, and thus yielding low rejection rates. Specifically, the HMC sampler require setting the potential energy term in the Hamiltonian (1) to the negative-log of the posterior distribution (18). The potential energy term $\mathcal{J}(\mathbf{x}_k)$

$$\begin{aligned} \mathcal{J}(\mathbf{x}_k) &= -\log\left(\sum_{i=1}^{\mathrm{N_c}} \frac{\tau_{k,i}}{\sqrt{|\boldsymbol{\Sigma}_{k,i}|}} \exp\left(-\frac{1}{2}\|\mathbf{x}_k-\boldsymbol{\mu}_{k,i}\|^2_{\boldsymbol{\Sigma}_{\mathbf{k,i}}{}^{-1}} - \frac{1}{2}\left(\|\mathcal{H}_k(\mathbf{x}_k)-\mathbf{y}_k\|^2_{\mathbf{R}_k^{-1}}\right)\right)\right) \qquad (20a) \\ &= \frac{1}{2}\|\mathcal{H}_k(\mathbf{x}_k)-\mathbf{y}_k\|^2_{\mathbf{R}_k^{-1}} - \log\left(\sum_{i=1}^{\mathrm{N_c}} \frac{\tau_{k,i}}{\sqrt{|\boldsymbol{\Sigma}_{k,i}|}} \exp\left(-\frac{1}{2}\|\mathbf{x}_k-\boldsymbol{\mu}_{k,i}\|^2_{\boldsymbol{\Sigma}_{\mathbf{k,i}}{}^{-1}}\right)\right) \\ &= \frac{1}{2}\|\mathcal{H}_k(\mathbf{x}_k)-\mathbf{y}_k\|^2_{\mathbf{R}_k^{-1}} - \log\left(\sum_{i=1}^{\mathrm{N_c}} \frac{\tau_{k,i}}{\sqrt{|\boldsymbol{\Sigma}_{k,i}|}} \exp\left(-\mathcal{J}_{k,i}(\mathbf{x}_k)\right)\right), \\ \mathcal{J}_{k,i}(\mathbf{x}_k) &= \frac{1}{2}\|\mathbf{x}_k-\boldsymbol{\mu}_{k,i}\|^2_{\boldsymbol{\Sigma}_{\mathbf{k,i}}{}^{-1}}. \qquad (20b) \end{aligned}$$

Equation (20) is expected to suffer from numerical difficulties due to evaluating the logarithm of a sum of very small values. To address the accumulation of roundoff errors, and without loss of generality, we assume from now on that the terms in Equation (20) under the sum are sorted in decreasing order, i.e., $(\boldsymbol{\tau}_{k,i}/\sqrt{|\boldsymbol{\Sigma}_{k,i}|})\,\exp\left(-\mathcal{J}_{k,i}(\mathbf{x}_k)\right) > (\tau_{k,i+1}/\sqrt{|\Sigma_{k,i+1}|})\,\exp\left(-\mathcal{J}_{k,i+1}(\mathbf{x}_k)\right),\ \ \forall\, i = 1,\ldots,\mathrm{N_c}-1$. The potential energy function (20) is rewritten as:

$$\begin{aligned}\mathcal{J}(\mathbf{x}_k) &= \frac{1}{2}\|\mathcal{H}_k(\mathbf{x}_k)-\mathbf{y}_k\|^2_{\mathbf{R}_k^{-1}} \\ &\quad - \left[\log\left(\frac{\tau_{k,1}\,\exp\left(-\mathcal{J}_{k,1}(\mathbf{x}_k)\right)}{\sqrt{|\Sigma_{k,1}|}}\right) + \log\left(1+\sum_{i=2}^{\mathrm{N_c}}\frac{\frac{\tau_{k,i}}{\sqrt{|\boldsymbol{\Sigma}_{k,i}|}}\,\exp\left(-\mathcal{J}_{k,i}(\mathbf{x}_k)\right)}{\frac{\tau_{k,1}}{\sqrt{|\Sigma_{k,1}|}}\,\exp\left(-\mathcal{J}_{k,1}(\mathbf{x}_k)\right)}\right)\right] \\ &= \frac{1}{2}\|\mathcal{H}_k(\mathbf{x}_k)-\mathbf{y}_k\|^2_{\mathbf{R}_k^{-1}} + \mathcal{J}_{k,1}(\mathbf{x}_k) \\ &\quad - \log\left(\frac{\tau_{k,1}}{\sqrt{|\Sigma_{k,1}|}}\right) - \log\left(1+\sum_{i=2}^{\mathrm{N_c}}\frac{\tau_{k,i}\,\sqrt{|\Sigma_{k,1}|}}{\tau_{k,1}\,\sqrt{|\boldsymbol{\Sigma}_{k,i}|}}\,\exp\left(\mathcal{J}_{k,1}(\mathbf{x}_k)-\mathcal{J}_{k,i}(\mathbf{x}_k)\right)\right).\end{aligned} \tag{21}$$

The gradient of the potential energy (21) is:

$$\begin{aligned}\nabla_{\mathbf{x}_k}\mathcal{J}(\mathbf{x}_k) &= \mathbf{H}_k^T\mathbf{R}_k^{-1}(\mathcal{H}_k(\mathbf{x}_k)-\mathbf{y}_k) + \nabla_{\mathbf{x}_k}\mathcal{J}_{k,1}(\mathbf{x}_k) \\ &\quad - \mathrm{s}\sum_{i=2}^{\mathrm{N_c}}\left\{\frac{\tau_{k,i}\,\sqrt{|\Sigma_{k,1}|}}{\tau_{k,1}\,\sqrt{|\boldsymbol{\Sigma}_{k,i}|}}\,\exp\left(\mathcal{J}_{k,1}(\mathbf{x}_k)-\mathcal{J}_{k,i}(\mathbf{x}_k)\right)\left[\nabla_{\mathbf{x}_k}\mathcal{J}_{k,1}(\mathbf{x}_k)-\nabla_{\mathbf{x}_k}\mathcal{J}_{k,i}(\mathbf{x}_k)\right]\right\}, \\ \mathrm{s} &= \frac{1}{\left(1+\sum_{i=2}^{\mathrm{N_c}}\frac{\tau_{k,i}\,\sqrt{|\Sigma_{k,1}|}}{\tau_{k,1}\,\sqrt{|\boldsymbol{\Sigma}_{k,i}|}}\,\exp\left(\mathcal{J}_{k,1}(\mathbf{x}_k)-\mathcal{J}_{k,i}(\mathbf{x}_k)\right)\right)}, \\ \nabla_{\mathbf{x}_k}\mathcal{J}_{k,i}(\mathbf{x}_k) &= \boldsymbol{\Sigma}_{k,i}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}_{k,i}) \qquad \forall\, i = 1,2,\ldots,\mathrm{N_c}\,.\end{aligned} \tag{22}$$

The cluster HMC sampling filter ($\mathcal{C}\ell$HMC) results by replacing the potential energy function (10b) and its derivative (11) in the HMC sampling filter, with Equations (21) and (22) respectively. The steps of the $\mathcal{C}\ell$HMC sampling filter are explained in Algorithm 1.

Note that in the case where the mixture contains a single component (one Gaussian distribution), the potential energy function (21) and its gradient (22) reduce to the following, respectively:

$$\begin{aligned}\mathcal{J}(\mathbf{x}_k) &= \frac{1}{2}\|\mathbf{x}_k-\mathbf{x}_k^{\mathrm{b}}\|^2_{\mathbf{B}_k^{-1}} + \frac{1}{2}\|\mathcal{H}_k(\mathbf{x}_k)-\mathbf{y}_k\|^2_{\mathbf{R}_k^{-1}}\,, \\ \nabla_{\mathbf{x}_k}\mathcal{J}(\mathbf{x}_k) &= \mathbf{B}_k^{-1}\left(\mathbf{x}_k-\mathbf{x}_k^{\mathrm{b}}\right) + \mathbf{H}_k^T\mathbf{R}_k^{-1}\left(\mathcal{H}_k(\mathbf{x}_k)-\mathbf{y}_k\right)\,.\end{aligned} \tag{23}$$

This shows that the $\mathcal{C}\ell$HMC sampling filter proposed herein, reduces to the original HMC filter when the EM algorithm detects a single component during the prior density approximation phase.

The $\mathcal{C}\ell$HMC Sampling Algorithm

As in the HMC sampling filter, information about the analysis probability density at the previous time t_{k-1} is captured by the analysis ensemble of states $\{\mathbf{x}^{\mathrm{a}}_{k-1}(e)\}_{e=1,\ldots,\mathrm{N_{ens}}}$. The forecast step consists of two stages. First, the model (12) is used to integrate each analysis ensemble member forward to time t_k where observations are available. Next, a clustering scheme (e.g., EM) is used to generate the parameters of the GMM. The analysis step constructs a Markov chain starting from an initial state $\mathbf{x}_k^0$, and proceeds by sampling the posterior PDF (18) at stationarity. Here the superscript over $\mathbf{x}_k$ refers to the iteration number in the Markov chain.

It is worth mentioning that the Hamiltonian system used as a mechanism to generate proposals for the Markov Chain, along with its associated pseudo time-step settings, are independent from the physical model being simulated.

As discussed in [17], Algorithm 1 can be used either as a non-Gaussian filter, or as a replenishment tool for parallel implementations of the traditional filters such as EnKF.

Algorithm 1 Cluster HMC sampling filter ($\mathcal{C}\ell$HMC)

1: **Forecast step:** given an analysis ensemble $\{\mathbf{x}^{\mathrm{a}}_{k-1}(e)\}_{e=1,2,\ldots,\mathrm{N_{ens}}}$ at time t_{k-1};

i- generate the forecast ensemble using the model $\mathcal{M}$:

$$\mathbf{x}^{\mathrm{b}}_{k}(e) = \mathcal{M}_{t_{k-1}\to t_k}\left(\mathbf{x}^{\mathrm{a}}_{k-1}(e)\right), \quad e = 1, 2, \ldots, \mathrm{N_{ens}}.$$

ii- Use AIC/BIC criteria to detect the number of mixture components $\mathrm{N_c}$ in the GMM, then use EM to estimate the GMM parameters $\{(\boldsymbol{\tau}_{k,i};\, \boldsymbol{\mu}_{k,i},\, \boldsymbol{\Sigma}_{k,i})\}_{i=1,2,\ldots,\mathrm{N_c}}$.

2: **Analysis step:** given the observation $\mathbf{y}_k$ at time point t_k , follow the steps i to v:

i- Initialize the Markov Chain ($\mathbf{x}^0_k$) to be to the best estimate available, e.g., to the mean of the joint forecast ensemble, or the mixture component mean with maximum likelihood.
ii- Choose a positive definite mass matrix $\mathbf{M}$. A recommended choice is to set $\mathbf{M}$ to be a diagonal matrix whose diagonal is equal to the diagonal of the posterior precision matrix. The precisions calculated from the prior ensemble can be used as a proxy.
iii- Set the potential energy function to (21), and its derivative to (22).
iv- Initialize the chain with a state $\mathbf{x}^0_k$ and generate $\mathrm{N_{ens}}$ ensemble members from the posterior distribution (18) as follows:

(1) Draw a random vector $\mathbf{p}^r \sim \mathcal{N}(0, \mathbf{M})$.
(2) Use a symplectic numerical integrator (e.g., Verlet, 2-Stage, or 3-Stage [17,29]) to advance the current state $(\mathbf{p}^r,\, \mathbf{x}^r_k)$ by a pseudo-time increment T to obtain a *proposal* state $(\mathbf{p}^*,\, \mathbf{x}^*_k)$:

$$(\mathbf{p}^*,\, \mathbf{x}^*_k) = \Phi_T\big((\mathbf{p}^r,\, \mathbf{x}^r_k)\big). \tag{24}$$

(a) Evaluate the energy loss : $\Delta H = H(\mathbf{p}^*,\, \mathbf{x}^*_k) - H(\mathbf{p}^r,\, \mathbf{x}^r_k)$.
(b) Calculate the acceptance probability: $a^{(r)} = 1 \wedge e^{-\Delta H}$.
(c) Discard both $\mathbf{p}^*$, $\mathbf{p}^r$.
(d) **(Acceptance/Rejection)** Draw a uniform random variable $u^{(r)} \sim \mathcal{U}(0, 1)$:

i- If $a^{(r)} > u^{(r)}$ accept the proposal as the next sample: $\mathbf{x}^{r+1}_k := \mathbf{x}^*_k$;
ii- If $a^{(r)} \leq u^{(r)}$ reject the proposal and continue with the current state: $\mathbf{x}^{r+1}_k := \mathbf{x}^r_k$.

(e) Repeat steps 1 to 6 until $\mathrm{N_{ens}}$ distinct samples are drawn.

v- Use the generated samples $\{\mathbf{x}^{\mathrm{a}}_k(e)\}_{e=1,2,\ldots,\mathrm{N_{ens}}}$ as an analysis ensemble. The analysis ensemble can be used to infer the posterior moments, e.g., posterior mean and posterior covariance matrix.

3: Increase time $k := k + 1$ and repeat steps 1 and 2.

3.3. Computational Considerations

To initialize the Markov chain one seeks a state that is likely with respect to the analysis distribution. Therefore one can start with the background ensemble mean, or with the mean of the component that has the highest weight. Alternatively, one can apply a traditional EnKF step and use the mean analysis to initialize the chain.

The joint ensemble mean and covariance matrix can be evaluated using the forecast ensemble, or using the GMM parameters. Given the GMM parameters $(\boldsymbol{\tau}_{k,i};\, \boldsymbol{\mu}_{k,i},\, \boldsymbol{\Sigma}_{k,i})$, the joint background mean and covariance matrix are, respectively:

$$\bar{\mathbf{x}}_k^{\mathrm{b}} = \sum_{i=1}^{\mathrm{N_c}} \boldsymbol{\tau}_{k,i}\, \boldsymbol{\mu}_{k,i}\,, \tag{25a}$$

$$\mathbf{B}_k^{\mathrm{ens}} = \sum_{i=1}^{\mathrm{N_c}} \boldsymbol{\tau}_{k,i}\, \boldsymbol{\Sigma}_{k,i} + \sum_{i=1}^{\mathrm{N_c}} \boldsymbol{\tau}_{k,i}\, (\boldsymbol{\mu}_{k,i} - \bar{\mathbf{x}}_k^{\mathrm{b}})(\boldsymbol{\mu}_{k,i} - \bar{\mathbf{x}}_k^{\mathrm{b}})^T\,. \tag{25b}$$

Both the potential energy (21) and its gradient (22) require evaluating the determinants of the covariance matrices associated with the mixture components. This is a computationally expensive process that is best avoided for large-scale problems. A simple remedy is to force the covariance matrices $\Sigma_{k,i}$, $\forall i = 1, 2, \ldots, \mathrm{N_c}$ to be diagonal while constructing the GMM.

When the Algorithm 1 is applied sequentially, at some steps a single mixture component could be detected in the prior ensemble. In this case, forcing a diagonal covariance structure does not help, and we fall back to the standard HMC sampler, where the full ensemble covariance matrix is utilized.

*3.4. A Multi-Chain Version of the $\mathcal{C}\ell$*HMC *Filter (MC-$\mathcal{C}\ell$*HMC*)*

Given the special geometry of the posterior mixture distribution, one can construct separate Markov chains for different components of the posterior. These chains can run in parallel to independently sample different regions of the analysis distribution. By running a Markov chain starting at each component of the mixture distribution we ensure that the proposed algorithm navigates all modes of the posterior, and covers all regions of high probability. The parameters of the jumping distribution for each of the chains can be tuned locally based on the statistics of the ensemble points belonging to the corresponding component in the mixture.

A multi-chain version of the $\mathcal{C}\ell$MCMC filter (MC-$\mathcal{C}\ell$MCMC) is developed by choosing the proposal of the ith chain to be $\mathcal{N}(\mathbf{x}^r_{k,i},\, \boldsymbol{\Sigma}_{k,i})$.

Running an HMC sampling chain under each of the posterior mixture components formulates the multi-chain $\mathcal{C}\ell$HMC (MC-$\mathcal{C}\ell$HMC) sampling filter. In this case, the diagonal of the mass matrix can be set globally for all components, for example using the diagonal of the precision matrix of the forecast ensemble, or can be chosen locally based on the second-order moments estimated from the prior ensemble under the corresponding component in the prior mixture. This local choice of the mass matrix does not change the marginal density of the target variable.

The local ensemble size (sample size per chain) can be specified based on the prior weight of the corresponding component multiplied by the likelihood of the mean of that component. Every chain is initialized to the mean of the corresponding component in the prior mixture.

The computational cost of the original HMC sampling filter depends greatly on the parameters of the Markov chain. A comprehensive discussion of the cost of the HMC sampling filter is given in Section 4.1 [17]. The $\mathcal{C}\ell$HMC sampling filter replaces the Gaussian prior with a GMM, which introduces an additional cost for utilizing the EM algorithm. This added cost, however, is negligible compared to cost of the sampling step. Moreover, MC-$\mathcal{C}\ell$HMC allows running the chains in parallel, which reduces the computational cost by a factor of up to the number of posterior probability modes $\mathrm{N_c}$. This approach is potentially very efficient, not only because it reduces the total running time of the sampler, but also because it favors an increased acceptance rate.

4. Numerical Results

We first apply the proposed algorithms to sample a simple one-dimensional mixture distribution. The cluster HMC sampling filters are then tested using a quasi-geostrophic (QG) model and compared against the original HMC sampling filter and against EnKF. We mainly use a nonlinear 1.5-layer reduced-gravity QG model with double-gyre wind forcing and bi-harmonic friction [37]. The data

assimilation testing suite (DATeS) [38,39] is used to carry out the numerical experiments presented in this work.

4.1. One-Dimensional Test Problem

We start with a prior ensemble generated from a GMM with $\mathrm{N_c} = 5$ and the following mixture parameters:

$$\begin{aligned}\{(\tau_i;\, \mu_i, \sigma_i^2)\}_{i=1,\dots,5} = \{&(0.2;\, -2.4, 0.05),\ (0.1;\, -1.0, 0.07),\\ &(0.1;\, 0, 0.02),\ (0.3;\, 1.0, 0.06),\ (0.3;\, 2.4, 0.1)\,\}\,.\end{aligned} \tag{26}$$

The EM algorithm is used to construct a GMM approximation of the true probability distribution from which the given prior ensemble is drawn. The model selection criterion used here is AIC. The generated GMM approximation of the prior has $\mathrm{N_c} = 4$ and the following parameters:

$$\begin{aligned}\{(\tau_i;\, \mu_i, \sigma_i^2)\}_{i=1,\dots,4} = \{&(0.169;\, -2.370, 0.052),\ (0.278;\, -0.727, 0.423),\\ &(0.229;\, 1.070, 0.065),\ (0.324;\, 2.436, 0.159)\,\}\,.\end{aligned} \tag{27}$$

The true prior, the prior ensemble, and the GMM approximation fitted to the prior ensemble, are shown in Figure 1.

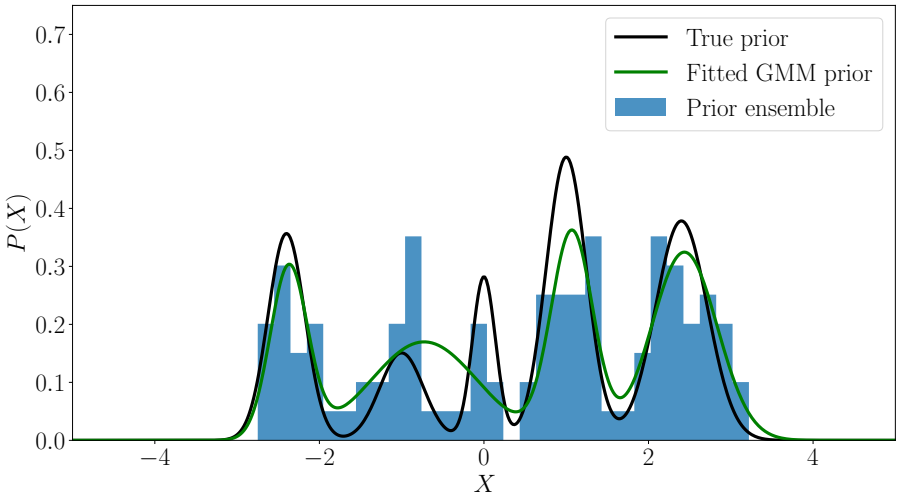


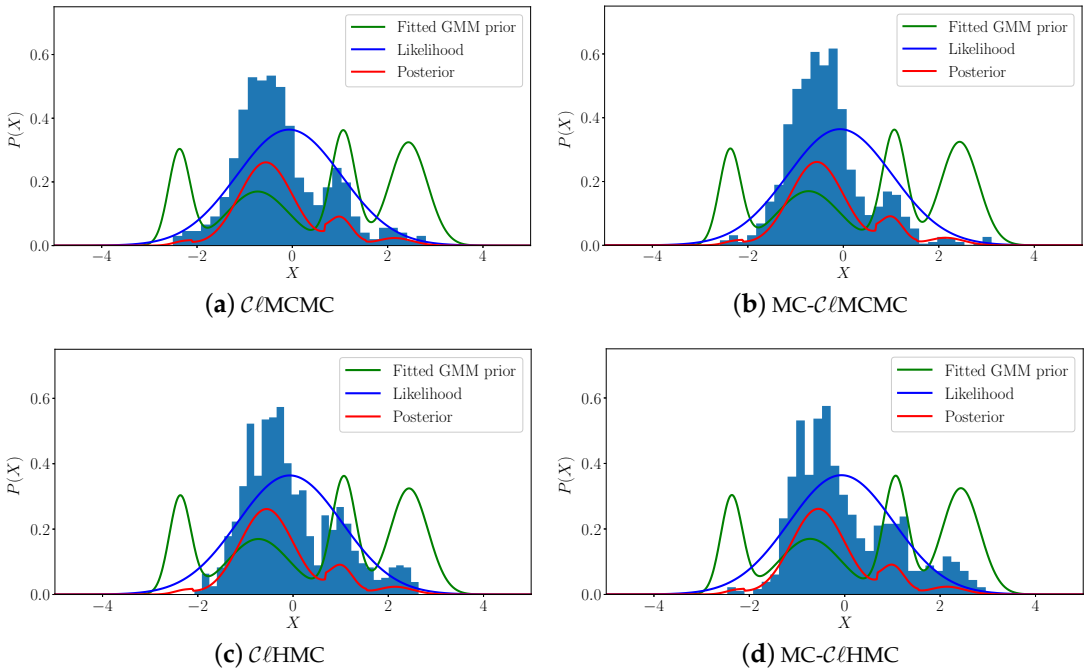
Figure 1. The one-dimensional example. A random sample of size $\mathrm{N_{ens}} = 100$ generated from the "true" GMM prior with parameters given by (26), and a GMM constructed by EM algorithm with AIC model selection criterion.

Assuming the observation likelihood function is given by:

$$\mathcal{P}(\mathbf{y}|\mathbf{x}) = \frac{1}{\sqrt{1.2}\,\sqrt{2\pi}}\ \exp\left(-\frac{1}{2}\,\frac{(\mathbf{x}-\mathbf{y})^2}{1.2}\right), \tag{28}$$

with an observation $\mathbf{y} = -0.06858$, the posterior and the histograms of $\mathrm{N_{ens}} = 1000$ sample points generated by $\mathcal{C}\ell$MCMC, MC-$\mathcal{C}\ell$MCMC, $\mathcal{C}\ell$HMC, and MC-$\mathcal{C}\ell$HMC algorithms, are shown in Figure 2. The acceptance rates of the cluster sampling filters are given in Table 1. In this example, the symplectic integrator used for the HMC-based filters, is Verlet with pseudo-time stepping parameters $T = mh$ with $m = 20$, and $h = 0.05$. Since the chains are initialized to the means of the prior mixture components, the burn-in stage is waived, i.e., the number of burn-in steps is set to zero. To reduce the correlation between the ensemble members of one chain we discard 20 states (mixing steps) between each two consecutive sampled points. In MC-$\mathcal{C}\ell$MCMC, and MC-$\mathcal{C}\ell$HMC filters, the ensemble size

per component (per chain) is set to $N_{ens} \times \ell_i \times \tau_i$, where ℓ_i is the likelihood of the mean of the i^{th} component in the prior mixture.



(**a**) $\mathcal{C}\ell$MCMC (**b**) MC-$\mathcal{C}\ell$MCMC

(**c**) $\mathcal{C}\ell$HMC (**d**) MC-$\mathcal{C}\ell$HMC

Figure 2. The one-dimensional example. A GMM prior, a Gaussian likelihood, and the resulting posterior, along with histograms of 1000 sample points generated by the $\mathcal{C}\ell$MCMC (**a**), MC-$\mathcal{C}\ell$MCMC (**b**), $\mathcal{C}\ell$HMC (**c**), and the MC-$\mathcal{C}\ell$HMC (**d**) sampling algorithms. The symplectic integrator used for HMC filters is Verlet with pseudo-time stepping parameters $T = mh$ with $m = 20$, and $h = 0.045$. The number of burn-in steps is zero, and the number of mixing steps is 20.

Table 1. Acceptance rates of the cluster sampling filters, for the one-dimensional example.

	Sampling Filter			
	$\mathcal{C}\ell$MCMC	MC-$\mathcal{C}\ell$MCMC	$\mathcal{C}\ell$HMC	MC-$\mathcal{C}\ell$HMC
Acceptance rate	44.32	77.39	99.21	99.23

The results reported in Figure 2 show that the four versions of the cluster sampling filter, $\mathcal{C}\ell$MCMC and MC-$\mathcal{C}\ell$MCMC, $\mathcal{C}\ell$HMC and MC-$\mathcal{C}\ell$HMC are capable of generating ensembles with mass distribution accurately representing the underlying target posterior. The serial version $\mathcal{C}\ell$HMC, however fails to sample one of the probability modes (the leftmost probability mode in Figure 2c), while the multi-chain version MC-$\mathcal{C}\ell$HMC generates samples from the vicinities of all posterior probability modes. Moreover, the acceptance rates shown in Table 1 explain that the HMC-based samplers yield much lower rejection rates, and thus are more favorable.

In large-scale settings, the $\mathcal{C}\ell$MCMC filter and the multi-chain version MC-$\mathcal{C}\ell$MCMC, are expected to suffer from random walk behavior, and would require large ensemble sizes to cover all probability modes properly. Unlike $\mathcal{C}\ell$MCMC, the $\mathcal{C}\ell$HMC sampler is suited for high-dimensional settings, and can explore the probability space quickly with small ensemble size. This is mainly because the HMC proposals target high-probability regions more frequently. The performance of the original HMC sampling filter was evaluated, and its usefulness in nonlinear settings was demonstrated, e.g., in [17]. The remainder of this Section is devoted to assessing the benefit of using a GMM, to relax the Gaussian-prior assumption posed by the original HMC filter, in sequential non-Gaussian filtering, and in relatively high-dimensional settings. Thus, in the numerical experiments shown in Section 4.2, we focus only on results of the $\mathcal{C}\ell$HMC, and MC-$\mathcal{C}\ell$HMC cluster sampling filters.

4.2. Quasi-Geostrophic Model

We employ the QG-1.5 model described by Sakov and Oke [37]. This model is a numerical approximation of the equations:

$$\begin{aligned} q_t &= \psi_x - \varepsilon J(\psi, q) - A\Delta^3\psi + 2\pi \sin(2\pi y)\,, \\ q &= \Delta\psi - F\psi\,, \\ J(\psi, q) &\equiv \psi_x q_x - \psi_y q_y\,, \end{aligned} \tag{29}$$

where $\Delta := \partial^2/\partial x^2 + \partial^2/\partial y^2$ and ψ is either the stream function or the surface elevation. We use the values of the model coefficients (29) from [37], as follows: $F = 1600$, $\varepsilon = 10^{-5}$, and $A = 2 \times 10^{-12}$. The domain of the model is a 1×1 [space units] square, with $0 \le x \le 1$, $0 \le y \le 1$, and is discretized by a grid of size 129×129 (including boundaries). Boundary conditions used are $\psi = \Delta\psi = \Delta^2\psi = 0$. The model state dimension is $\mathrm{N_{var}} = 16641$, while the model trajectories belong to affine subspaces with dimensions of the order of $10^2 - 10^3$ [37].

The time integration scheme used is the fourth-order Runge-Kutta scheme with a time step 1.25 [time units].

For all experiments in this work, the model is run over 1000 model time steps, with observations made available every 10 time steps. In this synthetic model the scales are not relevant, and we use generic space, time, and solution amplitude units.

4.2.1. Observations and Observation Operators

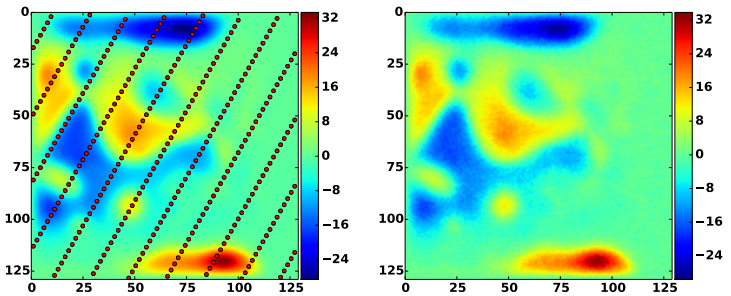
Two observation operators are used with this model.

- First we use a standard linear operator to observe 300 components of ψ. The observation error variance is 4.0 [units squared]. Synthetic the observations are obtained by adding white noise to measurements of the sea surface height (SSH) extracted from a model run with lower viscosity.
- The second observation operator measures the magnitude of the flow velocity $\sqrt{u^2 + v^2}$. The flow velocity components u, v are obtained using a finite difference approximation of the following relations to the stream function:

$$u = +\frac{\partial\psi}{\partial y}\,, \quad v = -\frac{\partial\psi}{\partial x}. \tag{30}$$

In both cases, the observed components are uniformly distributed over the state vector length, with a random offset, that is updated at each assimilation cycle. As mentioned in [37], the observational settings used here are motivated by typical distribution of satellite altimetry for oceanic applications.

The reference initial state along with an example of the observational grid used, and the initial forecast state are shown in Figure 3.



(**a**) Reference initial state and observational grid (**b**) Mean of the initial forecast ensemble

Figure 3. The QG-1.5 model. The red dots in (**a**) indicate the location of observations for one of the test cases employed.

4.2.2. Filter Tuning

We used a deterministic implementation of EnKF (DEnKF) with parameters tuned as suggested in [37]. Specifically, we apply a covariance localization by means of a Hadamard product as explained in [8]. The localization function used is Gaspari-Cohn [40] with localization radius set to 12 grid cells. Inflation is applied with factor $\delta = 1.06$ to the analysis ensemble of anomalies at the end of each assimilation cycle of DEnKF.

The parameters of the HMC and $\mathcal{C}\ell$HMC sampling filters are tuned empirically in a preprocessing step in the HMC filter to guarantee a rejection rate at most between 25% to 30%. Here we tune the parameters of the Hamiltonian trajectory only once at the beginning of the assimilation experiment. Specifically, the step size parameters of the symplectic integrator are set to $h = 0.075$, $m = 25$ in the presence of the linear observation operator, and are set to $h = 0.015$, $m = 25$ when the nonlinear observation operator (30) is used. The integrator used for the Hamiltonian system in all experiments is the three-stage symplectic integrator [17,29]. The mass matrix $\mathbf{M}$ is chosen to be a diagonal matrix whose nonzero entries are equal to the precisions, i.e., reciprocal of the variances, of the forecast ensemble. In the current experiments, the first 50 steps of the Markov chains are discarded as a burn-in stage. Alternatively, one can run a suboptimal minimization of the negative-log of the posterior to achieve convergence to the posterior.

The parameters of the MC-$\mathcal{C}\ell$HMC filter are set as follows. The step size parameters of the symplectic integrator are set to $h = 0.05/\mathrm{N_c}$, $m = 15$ in the experiments with linear observation operator, and $h = 0.0075/\mathrm{N_c}$, $m = 15$ in the case of the nonlinear observation operator (30). The mass matrix is a diagonal matrix whose diagonal is set to the diagonal of the precision matrix of the forecast ensemble labeled under the corresponding mixture component. To avoid numerical problems related to very small ensemble variances, for example in the case of outliers, the variances are averaged with the modeled forecast variances of 5 units squared.

The prior GMM is built with number of components determined using AIC model selection criteria, with a lower bound of 5 of the number of ensemble members belonging to each component of the mixture. This lower bound is enforced as a means to ameliorate the effect of outliers on the GMM construction. In all experiments involving $\mathcal{C}\ell$HMC, and MC-$\mathcal{C}\ell$HMC, the diagonal covariances relaxation assumption is imposed. However, this structure is not imposed if only one mixture component is detected, and $\mathcal{C}\ell$HMC and MC-$\mathcal{C}\ell$HMC filters fall back to the original HMC filter. For cases where a component contains a very small number of ensemble members covariance tapering [41] can prove useful.

The ensemble size for all filters used here is set to $\mathrm{N_{ens}} = 25$.

4.2.3. Assessment Metrics

To assess the accuracy of the tested filters we use the root mean squared error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N_{var}}} \sum_{i=1}^{\mathrm{N_{var}}} (x_i - x_i^{\mathrm{true}})^2}\,, \tag{31}$$

where $\mathbf{x}^{\mathrm{true}} = \psi^{\mathrm{true}}$ is the reference state of the system and $\mathbf{x}$ is the analysis state, e.g., the average of the analysis ensemble. Here $\mathrm{N_{var}} = 129 \times 129 = 16641$ is the dimension of the model state. We also use Talagrand (rank) histogram [42,43] to assess the quality of the ensemble spread around the true state.

4.2.4. Results with Linear Observation Operator

In the linear settings of this experiment, EnKF produces the best possible results, under both RMSE and Rank histogram uniformity metrics. Our main goal in this section, is to test the performance of the proposed algorithms, $\mathcal{C}\ell$HMC and MC-$\mathcal{C}\ell$HMC against the original HMC sampling filter, with EnKF results as a benchmark. The results below suggest that relaxing the Gaussian-prior using

an ensemble-based GMM estimate could be dangerous, unless the sampler is guaranteed to cover all posterior probability modes.

Figure 4 presents the RMSE (31) results of the analyses obtained using EnKF, HMC, $\mathcal{C}\ell$HMC, and MC-$\mathcal{C}\ell$HMC filters in the presence of a linear observation operator. Figure 4 shows that the results of all HMC filter versions improve quickly at the first few assimilation windows. While the results of the original HMC filter improve quickly at the first few assimilation windows, the performance of the original HMC filter degrades compared to the DEnKF filter performance especially in the long run. We believe that the two main factors contribute to the HMC filter degradation are the parameter tuning, and the development of non-Gaussianity in the prior distribution. The $\mathcal{C}\ell$HMC analysis drifts away quickly from the true trajectory. This is mainly because the HMC sampling strategy is unable to cover all probability modes in the posterior distribution. To guarantee that the sampling filter covers the truth well, the sampler has to be able to sample properly from all posterior probability modes. This is achieved by design by the MC-$\mathcal{C}\ell$HMC filter. The MC-$\mathcal{C}\ell$HMC version produces RMSE results comparable to the RMSE obtained by DEnKF.

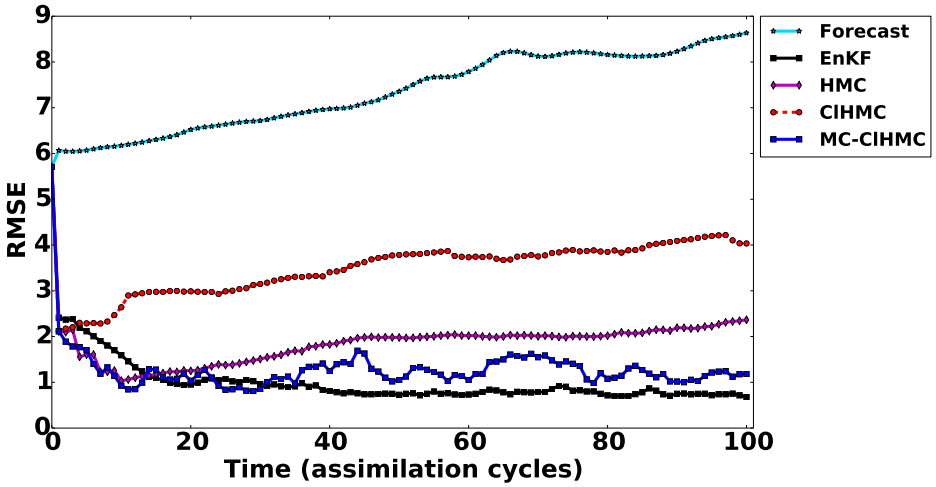


Figure 4. Data assimilation results with the linear observation operator. RMSE of the (31) analyses obtained by EnKF, HMC, $\mathcal{C}\ell$HMC, and MC-$\mathcal{C}\ell$HMC filters. Forecast results here refer to the RMSE obtained from a free run of the dynamical model, with initial condition set to the forecast state at the initial time.

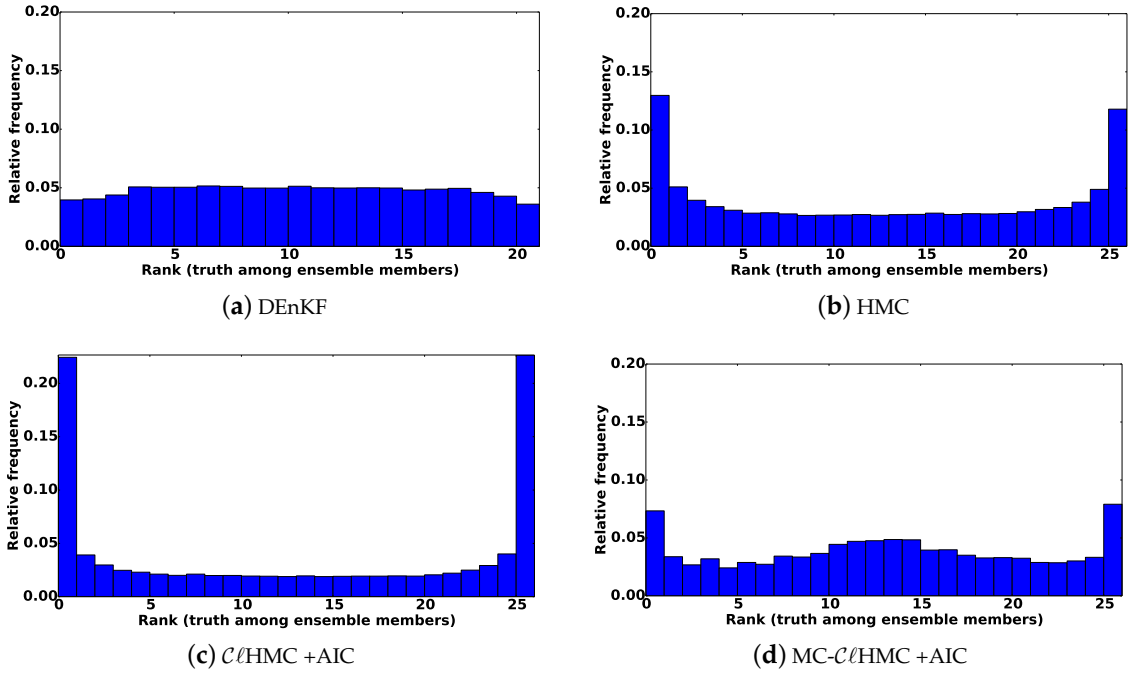
As discussed in [17] the performance of HMC filter can be further enhanced by automatically tuning the parameters of the symplectic integrator at the beginning of each assimilation cycle. Here however we are mainly interested in assessing the performance of the new methodologies compared to the original HMC filter using equivalent settings.

It is important to note that the MC-$\mathcal{C}\ell$HMC filter requires shorter Hamiltonian trajectories to explore the space under each local mixture component, which results in computational savings. Additional savings can be obtains by running the chains in parallel to sample different regions of the posterior.

Since we are not interested in only a *single* best estimate of the true state of the system, RMSE alone is not sufficient to judge the quality of the filtering system. The analysis ensemble sampled from the posterior should be spread widely enough to cover the truth and avoid filter collapse. The rank histograms of the analysis ensembles are shown in Figure 5. Generally speaking, Talagrand diagram is a histogram constructed by calculating the rank of the true state, compared to the ensemble members ordered increasingly in magnitude. The ranks are calculated over several assimilation cycles, for individual entries of the true state with respect to the corresponding entries of the ensemble members. Observations are used instead of model states in real experiments where the truth is unknown. A U-shaped rank histogram indicates an under-dispersed ensemble, while a mound rank

histogram indicates over-dispersion. A nearly-uniform rank histogram is desirable, and suggests that the truth is indistinguishable from the ensemble members.

The two small spikes in Figure 5b suggest that the performance of the original HMC filter could be enhanced by increasing the length of the Hamiltonian trajectories in some assimilation cycles.



(**a**) DEnKF (**b**) HMC

(**c**) $\mathcal{C}\ell$HMC +AIC (**d**) MC-$\mathcal{C}\ell$HMC +AIC

Figure 5. Data assimilation results with the linear observation operator. The rank histograms of where the truth ranks among posterior ensemble members. The ranks are evaluated for every 16th variable in the state vector (past the correlation bound) at 100 assimilation times.

The rank histogram shown in Figure 5c shows that the analysis ensembles produced by the $\mathcal{C}\ell$HMC filter tend to be under-dispersed. Since the ensemble size is relatively small and the prior GMM is multimodal, with regions of low-probability between the different mixture components, a multimodal mixture posterior with isolated components is obtained. As explained in [44], this is a case where HMC sampling in general can suffer from being entrapped in a local minimum (and fails to jump between different high probability modes). This behavior is expected to result in ensemble collapse, as seen in Figure 5c, leading to filter degradation in the long run as illustrated by the RMS errors shown in Figure 4.

The results shown in Figure 5c suggest that the analysis ensemble collected by $\mathcal{C}\ell$HMC fails to cover all mixture components, thereby losing its dispersion when it is applied repetitively. This is supported by the results in Figure 6, where the rank histograms are plotted using results from the first two, five, and 10 cycles, respectively.

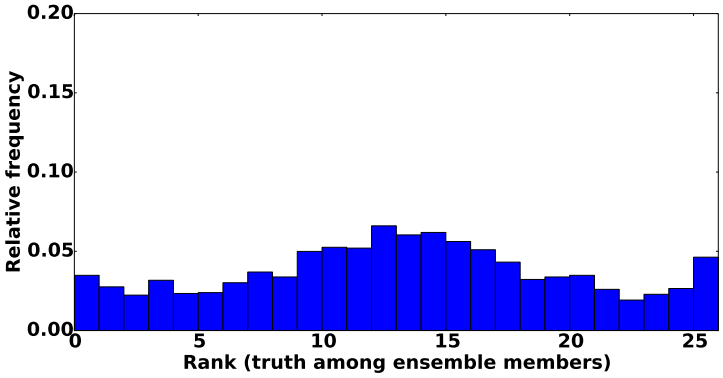
The ensemble collapse can be avoided if we force the sampler to collect ensemble members from all the probability modes. This is illustrated by the rank histograms of results obtained using the MC-$\mathcal{C}\ell$HMC filter with AIC criteria as shown in Figure 5d.

We believe that having isolated regions of high probability, e.g., with very small number of ensemble members in each component, can be the critical factor leading the poor long-term performance of $\mathcal{C}\ell$HMC. This is alleviated here by imposing a minimum number of 3 ensemble points in each component, e.g., via hard assignment, of the mixture while constructing the GMM approximation of the prior.

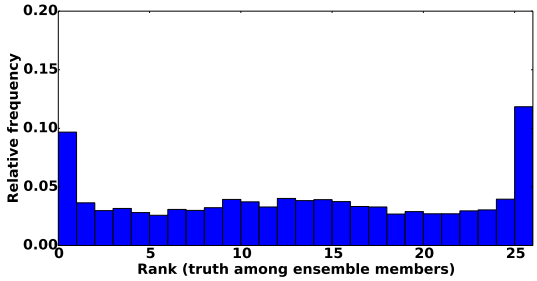
With automatic tuning of the Hamiltonian parameters the performance of both HMC and MC-$\mathcal{C}\ell$HMC filters is expected to be greatly enhanced. We have only shown the results of $\mathcal{C}\ell$HMC,

and MC-$\mathcal{C}\ell$HMC with AIC information criterion; experiments carried out using other model selection criteria such as BIC have proven to be very similar.

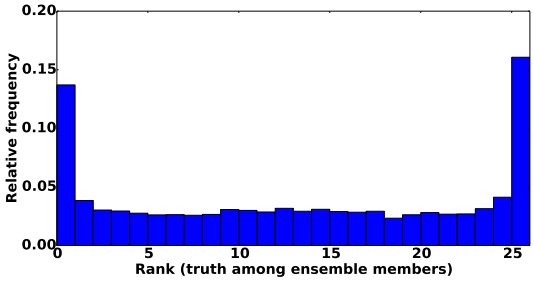
To help decide whether to apply the original formulation of the HMC filter, or the proposed methodology, one can run tests of non-Gaussianity on the forecast ensemble. To asses non-Gaussianity of the forecast several numeric or visualization normality tests are available, e.g., the Mardia test [45] based on multivariate extensions of skewness and kurtosis measures. Indication of non-Gaussianity can be found by visually inspecting several bivariate contour plots of the joint distribution of selected components in the state vector. Visualization methods for multivariate normality assessment such as chi-square QQ-plots can be very useful as well. Given a multivariate normal random variable $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, the squared Mahalanobis distance $d^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ follow a Chi-Squared distribution. Multivariate QQ-plot assesses the normality of a random variable $\mathbf{x}$, by comparing the quantiles of a sample of Mahalanobis distances to the quantiles of the correct Chi-Squared distribution, under the normality assumption. Deviations from the true distribution, e.g., the solid line in Figure 7, indicate possible departure of the sampled variable from multivariate normality.



(**a**) $\mathcal{C}\ell$HMC +AIC; first two observation windows



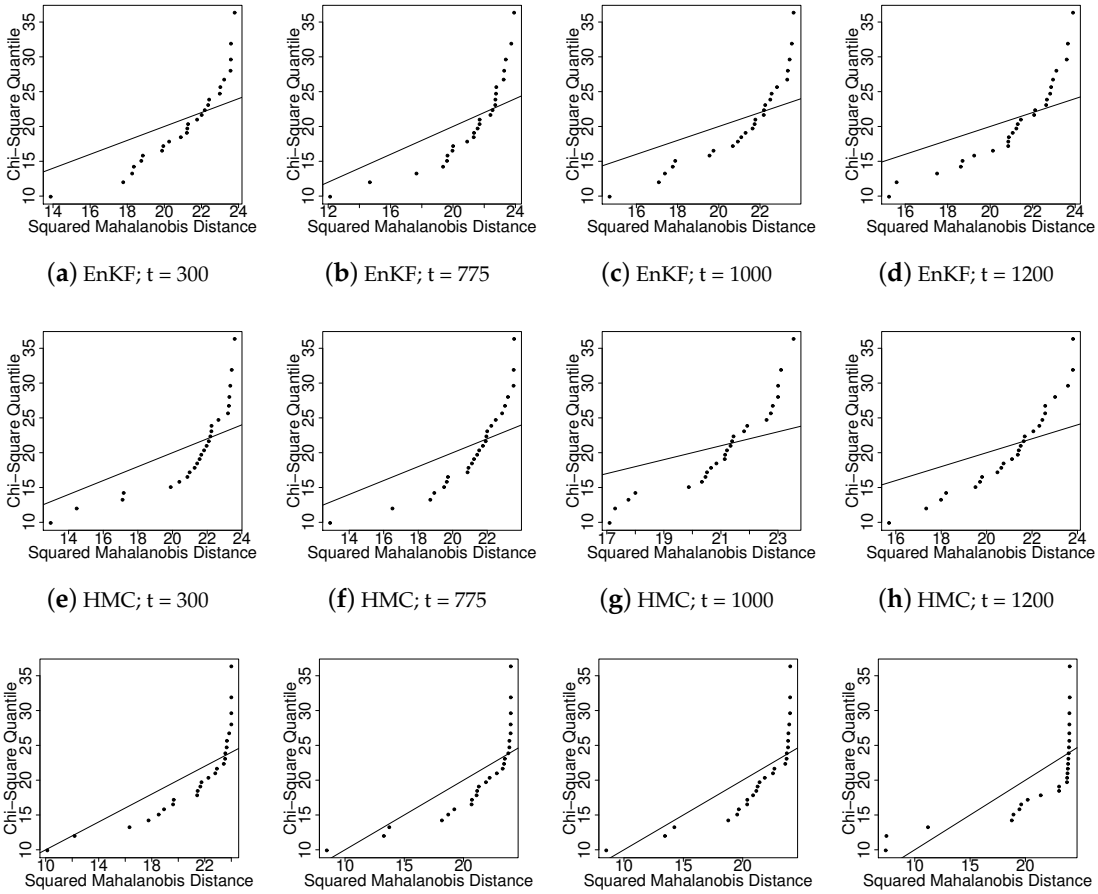
(**b**) $\mathcal{C}\ell$HMC +AIC; first five observation windows



(**c**) $\mathcal{C}\ell$HMC +AIC; first ten observation windows

Figure 6. Data assimilation results using a linear observation operator. Rank histograms of where the truth ranks among posterior ensemble members. The ranks are evaluated for every 16th variable in the state vector (past the correlation bound). Rank histograms of $\mathcal{C}\ell$HMC results obtained at the first two, five, and 10 assimilation cycles, respectively, are shown. The model selection criterion used is AIC.

Figure 7 shows several chi-square Q-Q plots of the forecast ensembles generated from the result of EnKF, HMC, and MC-$\mathcal{C}\ell$HMC filters at different time instances. These plots show strong signs of non-Gaussianity in the forecast ensemble, and suggest that the Gaussian-prior assumption may in general lead to inaccurate conclusions.



(**a**) EnKF; t = 300 (**b**) EnKF; t = 775 (**c**) EnKF; t = 1000 (**d**) EnKF; t = 1200

(**e**) HMC; t = 300 (**f**) HMC; t = 775 (**g**) HMC; t = 1000 (**h**) HMC; t = 1200

(**i**) MC-$\mathcal{C}\ell$HMC; t = 300 (**j**) MC-$\mathcal{C}\ell$HMC; t = 775 (**k**) MC-$\mathcal{C}\ell$HMC; t = 1000 (**l**) MC-$\mathcal{C}\ell$HMC; t = 1200

Figure 7. Data assimilation with a linear observation operator. Chi-square Q-Q plots for the forecast ensembles obtained from propagating analyses of EnKF, HMC, and MC-$\mathcal{C}\ell$HMC filtering systems to times t = 300, 775, 1000, and 1200 provide a strong indication of non-Gaussianity. The filtering methodology, and the assimilation time are given under each panel. Localization is applied to the ensemble covariance matrix to avoid singularity while evaluating the Mahalanobis distances of the ensemble members.

4.2.5. Results with Nonlinear Wind-Magnitude Observations

In the presence of a nonlinear observation operator the distribution is expected to show even stronger signs of non-Gaussianity. With stronger non-Gaussianity, the cluster methodology is expected to outperform the original formulation of the HMC sampling filter. In the settings used in this Section, EnKF diverges after the third cycle, and its results are omitted for clarity.

Figure 8 shows RMSE results, with the nonlinear observation operator (30), for the analyses obtained by HMC, $\mathcal{C}\ell$HMC, MC-$\mathcal{C}\ell$HMC filtering systems. While EnKF diverges quickly under these settings, the HMC algorithms, i.e., HMC, $\mathcal{C}\ell$HMC, and MC-$\mathcal{C}\ell$HMC sampling filters, continue to show behavior similar to the case where the linear observation operator is used (Section 4.2.4).

Figure 9 shows rank histograms of HMC, $\mathcal{C}\ell$HMC, and MC-$\mathcal{C}\ell$HMC, with a nonlinear observation operator. We can see that $\mathcal{C}\ell$HMC performance is similar to the case when the linear observation operator is used. It seems to be entrapped into a local minimum losing its dispersion quickly. The results of the MC-$\mathcal{C}\ell$HMC filter avoid this effect and show a reasonable spread.

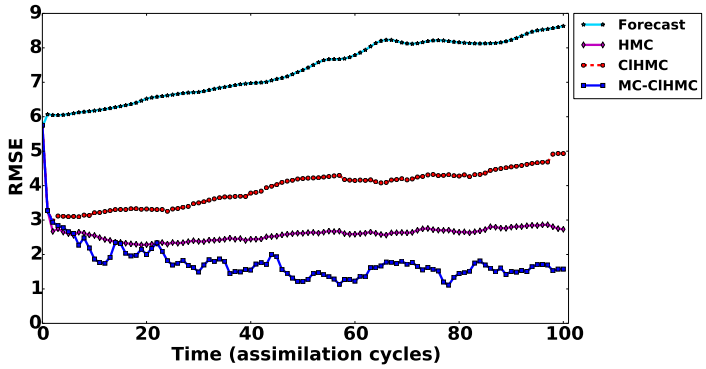
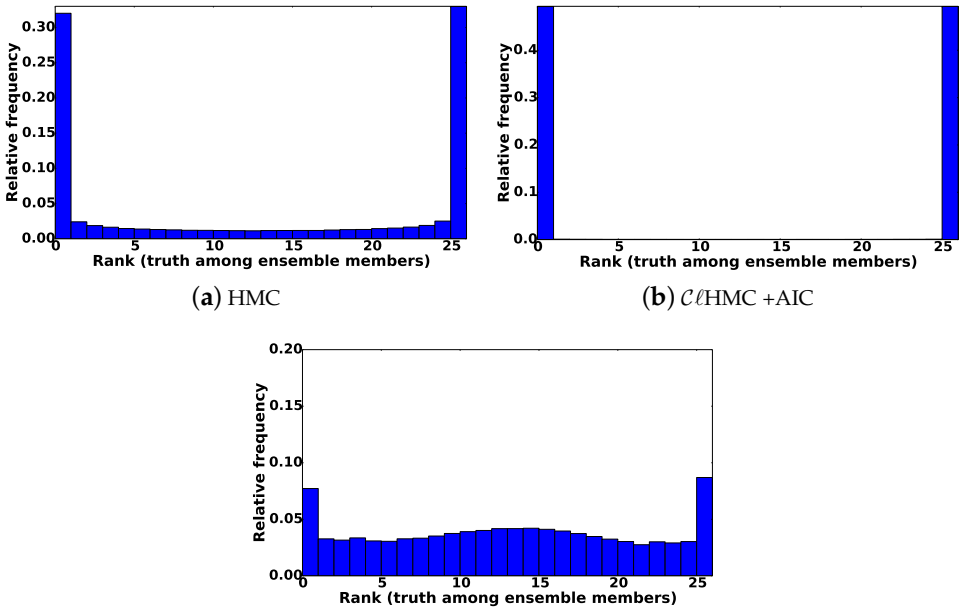


Figure 8. Data assimilation results with the nonlinear observation operator (30). RMSE of the analyses obtained by HMC, $\mathcal{C}\ell$HMC, and MC-$\mathcal{C}\ell$HMC filtering schemes. The Forecast RMSE results are obtained from a free run of the dynamical model.



(**a**) HMC (**b**) $\mathcal{C}\ell$HMC +AIC

(**c**) MC-$\mathcal{C}\ell$HMC +AIC

Figure 9. Data assimilation results using the nonlinear observation operator (30). The rank histograms of where the truth ranks among posterior ensemble members. The ranks are evaluated for every 16th variable in the state vector (past the correlation bound) at 100 assimilation times. The filtering scheme used is indicated under each panel.

The results presented here suggest that the cluster formulation of the HMC sampling filter is advantageous, especially in the presence of highly nonlinear observation operator, or strong indication of non-Gaussianity.

5. Conclusions and Future Work

This work presents a set of fully non-Gaussian sampling filters for sequential data assimilation. The filters use a GMM to approximate the prior distribution, given the forecast ensemble. The posterior mixture distribution is directly sampled following a MCMC approach. Several proposals are suggested for the MCMC sampling step. The $\mathcal{C}\ell$MCMC filter uses a Gaussian proposal, and $\mathcal{C}\ell$HMC follows a HMC approach for sampling the posterior, using a single Markov chain. These two versions may suffer from high-rejection rates, and may fail to sample all posterior probability modes, especially

in highly non-linear settings. More efficient multi-chain versions of $\mathcal{C}\ell$MCMC and $\mathcal{C}\ell$HMC, namely MC-$\mathcal{C}\ell$MCMC and MC-$\mathcal{C}\ell$HMC respectively, are developed in order to alleviate such difficulties.

Numerical experiments are carried out using a simple one-dimensional example, and a nonlinear 1.5-layer reduced-gravity quasi geostrophic model in the presence of observation operators of different levels of nonlinearity. The results show that the new methodologies are much more efficient that the original HMC sampling filter especially in the presence of a highly nonlinear observation operator.

The multi-chain cluster sampling filters, MC-$\mathcal{C}\ell$HMC and MC-$\mathcal{C}\ell$HMC, deserve further investigation. For example the local sample sizes here are selected based on the prior weight multiplied by the likelihood of the corresponding component mean. An optimal selection of the local ensemble size is required to guarantee efficient sampling from the target distribution.

Instead of using MC-$\mathcal{C}\ell$HMC filter, one can use $\mathcal{C}\ell$HMC with geometrically tempered Hamiltonian sampler as recently proposed in [44], such as to guarantee navigation between separate modes of the posterior. Alternatively, the posterior distribution can be split into N_c target distributions with different potential energy functions and associated gradients. This is equivalent to running independent HMC sampling filters in different regions of the state space under the target posterior.

The authors have started to investigate the ideas discussed here, in addition to testing the proposed methodologies with automatically tuned HMC samplers.

Author Contributions: Ahmed Attia and Adrian Sandu conceived and designed the experiments; Ahmed Attia and Azam Moosavi performed the experiments and analyzed the data; Ahmed Attia, Azam Moosavi, and Adrian Sandu wrote the paper.

Acknowledgments: This work was supported in part by the Air Force Office of Scientific Research (AFOSR) Dynamic Data Driven Application Systems program, by the National Science Foundation award NSF CCF (Algorithmic foundations)–1218454, and by the Computational Science Laboratory (CSL) in the Department of Computer Science at Virginia Tech. The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Kalman, R.E. A new approach to linear filtering and prediction problems. *Trans. ASME* **1960**, *82*, 35–45.
2. Kalman, R.E.; Bucy, R.S. New results in linear filtering and prediction theory. *J. Basic Eng.* **1961**, *83*, 95–108.
3. Burgers, G.; Jan van Leeuwen, P.; Evensen, G. Analysis scheme in the Ensemble Kalman Filter. *Mon. Weather Rev.* **1998**, *126*, 1719–1724.
4. Evensen, G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forcast error statistics. *J. Geophys. Res.* **1994**, *99*, 10143–10162.
5. Evensen, G. The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dyn.* **2003**, *53*, 343–367.
6. Houtekamer, P.L.; Mitchell, H.L. Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.* **1998**, *126*, 796–811.
7. Hamill, T.M.; Whitaker, J.S.; Snyder, C. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Weather Rev.* **2001**, *129*, 2776–2790.
8. Houtekamer, P.L.; Mitchell, H.L. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **2001**, *129*, 123–137.
9. Whitaker, J.S.; Hamill, T.M. Ensemble data assimilation without perturbed observations. *Mon. Weather Rev.* **2002**, *130*, 1913–1924.
10. Sakov, P.; Oliver, D.S.; Bertino, L. An iterative EnKF for strongly nonlinear systems. *Mon. Weather Rev.* **2012**, *140*, 1988–2004.
11. Smith, K.W. Cluster ensemble Kalman filter. *Tellus A* **2007**, *59*, 749–757.

12. Tippett, M.K.; Anderson, J.L.; Bishop, C.H.; Hamill, T.M.; Whitaker, J.S. Ensemble square root filters. *Mon. Weather Rev.* **2003**, *131*, 1485–1490.
13. Lorenc, A.C. Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **1986**, *112*, 1177–1194.
14. Zupanski, M. Maximum likelihood ensemble filter: Theoretical aspects. *Mon. Weather Rev.* **2005**, *133*, 1710–1726.
15. Zupanski, M.; Navon, I.M.; Zupanski, D. The Maximum Likelihood Ensemble Filter as a non-differentiable minimization algorithm. *Q. J. R. Meteorol. Soc.* **2008**, *134*, 1039–1050.
16. Gu, Y.; Oliver, D.S. An iterative ensemble Kalman filter for multiphase fluid flow data assimilation. *SPE J.* **2007**, *12*, 438–446.
17. Attia, A.; Sandu, A. A Hybrid Monte Carlo sampling filter for non-Gaussian data assimilation. *AIMS Geosci.* **2015**, *1*, 41–78.
18. Attia, A.; Rao, V.; Sandu, A. A sampling approach for four dimensional data assimilation. In *Dynamic Data-Driven Environmental Systems Science*; Springer: Heidelberg/Berlin, Germany, 2015; pp. 215–226.
19. Attia, A.; Rao, V.; Sandu, A. A Hybrid Monte-Carlo sampling smoother for four dimensional data assimilation. *Int. J. Nume. Methods Fluids* **2016**, *83*, 90–112, doi:10.1002/fld.4259.
20. Attia, A.; Stefanescu, R.; Sandu, A. The Reduced-Order Hybrid Monte Carlo Sampling Smoother. *Int. J. Nume. Methods Fluids* **2016**, *83*, 28–51, doi:10.1002/fld.4255.
21. Duane, S.; Kennedy, A.; B.J. Pendleton, J.B.; Roweth, D. Hybrid Monte Carlo. *Phys. Lett. B* **1987**, *195*, 216–222.
22. Toral, R.; Ferreira, A. A general class of hybrid Monte Carlo methods. *Proc. Phys. Comput.* **1994**, *94*, 265–268.
23. Anderson, J.L.; Anderson, S.L. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev.* **1999**, *127*, 2741–2758.
24. Besag, J.; Green, P.J. *Spatial Statistics and Bayesian Computation*; Wiley: Hoboken, NJ, USA, 1993; pp. 25–37.
25. Sokal, A. *Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms*; Springer: Heidelberg/Berlin, Germany, 1997.
26. Higdon, D.M. Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Am. Stat. Assoc.* **1998**, *93*, 585–595.
27. Neal, R. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*; CRC Press: Boca Raton, FL, USA, 2011.
28. Sanz-Serna, J.; M-P.Calvo. *Numerical Hamiltonian Problems*; Chapman & Hall: London, UK, 1994; Volume 7.
29. Sanz-Serna, J. Markov chain Monte Carlo and numerical differential equations. In *Current Challenges in Stability Issues for Numerical Differential Equations*; Springer: Heidelberg/Berlin, Germany 2014; pp. 39–88.
30. Homan, M.; Gelman, A. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623.
31. Girolami, M.; Calderhead, B. Riemann manifold langevin and hamiltonian monte carlo methods. *J. R. Stat. Soc. Ser. B* **2011**, *73*, 123–214.
32. Bilmes, J.A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *Int. Comput. Sci. Institute* **1998**, *4*, 126.
33. Shalizi, C.R. *Advanced Data Analysis from an Elementary Point of View*; Cambridge University Press: Cambridge, UK, 2013.
34. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 1–38.
35. Hu, X.; Xu, L. Investigation on several model selection criteria for determining the number of cluster. *Neur. Inf. Proc. Lett. Rev.* **2004**, *4*, 1–10.
36. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
37. Sakov, P.; Oke, P.R. A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. *Tellus A* **2008**, *60*, 361–371.
38. Attia, A.; Sandu, A. DATeS: A Highly-Extensible Data Assimilation Testing Suite. *arXiv* **2018**, arXiv:1704.05594.
39. Attia, A.; Glandon, R.; Tranquilli, P.; Narayanamurthi, M.; Sarshar, A.; Sandu, A. DATeS: A Highly-Extensible Data Assimilation Testing Suite. 2016. Available online: people.cs.vt.edu/~attia/DATeS (accessed on 28 May 2012).
40. Gaspari, G.; Cohn, S.E. Construction of correlation functions in two and three dimensions. *Q. J. R. Meteorol. Soc.* **1999**, *125*, 723–757.

41. Furrer, R.; Genton, M.G.; Nychka, D. Covariance tapering for interpolation of large spatial datasets. *J. Comput. Gr. Stat.* **2006**, *15*, 502–523.
42. Anderson, J.L. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Clim.* **1996**, *9*, 1518–1530.
43. Candille, G.; Talagrand, O. Evaluation of probabilistic prediction systems for a scalar variable. *Q. J. R. Meteorol. Soc.* **2005**, *131*, 2131–2150.
44. Nishimura, A.; Dunson, D. Geometrically Tempered Hamiltonian Monte Carlo. *arXiv* **2016**, arXiv:1604.00872.
45. Mardia, K.V. Measures of multivariate skewness and kurtosis with applications. *Biometrika* **1970**, *57*, 519–530.