Data Assimilation with Gaussian Mixture Models using the Dynamically Orthogonal Field Equations. Part I. Theory and Scheme

Sondergaard, T. and P.F.J. Lermusiaux (2012). Data Assimilation with Gaussian Mixture Models using the Dynamically Orthogonal Field Equations. Part I: Theory and Scheme, Monthly Weather Review, in press.

Data Assimilation with Gaussian Mixture Models using the Dynamically Orthogonal Field Equations. Part I: Theory and Scheme

Thomas Sondergaard and Pierre F. J. Lermusiaux *

Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts

ABSTRACT

This work introduces and derives an efficient, data-driven assimilation scheme, focused on a time-dependent stochastic subspace, that respects nonlinear dynamics and captures non-Gaussian statistics as it occurs. The motivation is to obtain a filter that is applicable to realistic geophysical applications but that also rigorously utilizes the governing dynamical equations with information theory and learning theory for efficient Bayesian data assimilation. Building on the foundations of classical filters, the underlying theory and algorithmic implementation of the new filter are developed and derived. The stochastic Dynamically Orthogonal (DO) field equations and their adaptive stochastic subspace are employed to predict prior probabilities for the full dynamical state, effectively approximating the Fokker-Planck equation. At assimilation times, the DO realizations are fit to semiparametric Gaussian mixture models (GMMs) using the Expectation-Maximization algorithm and the Bayesian Information Criterion. Bayes' Law is then efficiently carried out analytically within the evolving stochastic subspace. The resulting GMM-DO filter is illustrated in a very simple example. Variations of the GMM-DO filter are also provided along with comparisons with related schemes.

1. Introduction

Data assimilation (DA) is the process of quantitatively estimating dynamically evolving fields by melding information from observations with that predicted by computational models. DA has a long and interesting history; thorough expositions include Daley (1991), Ghil and Malanotte-Rizzoli (1991), Bennett (1992, 2002), Wunsch (1996), Malanotte-Rizzoli (1996), Robinson et al. (1998), Kalnay (2003) and Evensen (2007). Most schemes are derived from estimation theory (Jazwinski 1970; Gelb 1974), information theory (Sobczyk 2001; Cover and Thomas 2006), control theory (Lions 1971; Dimet and Talagrand 1986), and optimization theory and inverse problem theory (Tarantola 2005). While traditionally grounded in linear theory and the Gaussian approximation (Kalman 1960), recent years have seen the emergence of advanced DA schemes attempting to shed such limitations. One research thrust has been the development of efficient methods that respect nonlinear dynamics and capture non-Gaussian features. Most such methods are either challenging to employ with large realistic systems or still based on some ad hoc approximations. Our motivation here is to allow for realistic geophysical applications while rigorously utilizing the governing dynamical equations with information theory and learning theory for efficient Bayesian inference.

It is well known that geophysical dynamics can be very nonlinear and intermittent. The importance of accounting for nonlinearities in DA is also known for some time, e.g. (Miller et al. 1994). Nonlinearities not only affect prediction, but also the melding of measured and predicted information. As a result, oceanic and atmospheric fields can be characterized by complex, far-from-Gaussian statistics (CPSMA 1993; Lermusiaux et al. 2002a; Auclair et al. 2003; Dee and Silva 2003; Lermusiaux et al. 2006; Sura 2010). With the introduction of the Ensemble Kalman filter (Evensen 1994; Houtekamer et al. 1998), error subspace schemes (Lermusiaux and Robinson 1999) and square-root filters (Whitaker and Hamill 2002; Tippett et al. 2003) came the adoption of Monte Carlo methods (Doucet et al. 2001) within the DA community. In addition to utilizing the inherent nonlinearities of the governing equations, Monte Carlo methods allow exploration and exploitation of probabilistic structures beyond the simple Gaussian melding of information. One type of such methods are particle filters, e.g. (Pham 2001; van Leeuwen 2009), which evolve probability density functions (pdfs) using a discrete set of models states or particles and a corresponding mixture of "Dirac functions". Extensions include diffusion kernel filters (e.g. Krause and Restrepo 2009) and parametric filters (e.g. Kim et al. 2009). A related interest has been the approximation of distributions by Gaussian Mixture Models (GMMs) (Bocquet et al. 2010). Examples include Alspach and Sorenson (1972), Anderson and Anderson (1999), Chen and Liu (2000), Bengtsson et al. (2003),

Kotecha and Djuric (2003), Eyink and Kim (2006), Smith (2007), Hoteit et al. (2008) and Dovera and Rossa (2010), many of which will be examined later in this work. As will be shown, GMMs provide an attractive method for approximating distributions for the purposes of Bayesian inference. When fit to Monte Carlo data using the Expectation-Maximization algorithm (Dempster et al. 1977) and the Bayesian Information Criterion (Schwartz 1978), an accurate representation of the true pdf results. This is to be developed in this work.

A concern with present nonlinear DA schemes is their difficulty in handling the dimensionality of state vectors commonly encountered in oceanic and atmospheric applications, typically on the order of $n \sim 10^6 - 10^{10}$. A common useful remedy has been the adoption of various localization approximations (Bengtsson et al. 2003) and heuristic arguments (Anderson and Anderson 1999). A number of filters e.g. (Lermusiaux and Robinson 1999) have opted to focus on a *time-dependent* dominant subspace of the full state space, thereby allocating computational resources solely to the states that matter most. In a similar manner, we employ here the Dynamically Orthogonal (DO) field equations (Sapsis and Lermusiaux 2009; Sapsis 2010). The DO equations originate directly from the governing dynamical equations, i.e. the stochastic partial differential equations describing the evolution of the full geophysical system. By applying an orthogonality condition on the evolution of the stochastic subspace, the governing equations are reduced to evolution equations for (i) the mean field; (ii) the stochastic subspace; and (iii) the probabilistic variability contained within the subspace. These DO equations efficiently represent the true evolving pdf in between assimilation times and effectively approximate the Fokker-Planck equation.

In part I of this two-part paper, we develop and derive the underlying theory and algorithms of the proposed DA scheme: the GMM-DO filter. In section 2, we introduce and define the filter's core components. The derivation of the filter with a key proof are completed in section 3. Section 4 provides a simple example illustrating the filter's update step, while section 5 places the GMM-DO filter in the context of contemporary schemes based on related ideas. Conclusions are in section 6. In Appendices A and B, we present the EM algorithm and outline variations of the filter, respectively. In part II of this two-part paper (Sondergaard and Lermusiaux 2012), we apply the GMM-DO filter in a dynamical systems setting. Specifically, we evaluate its performance against contemporary filters when applied to (1) the Double Well Diffusion Experiment and (2) the Sudden Expansion fluid flow.

2. GMM-DO Filter Components

In this section, we introduce the core components that we ultimately combine into the GMM-DO filter, specifically:

- Gaussian mixture models;
- Expectation-Maximization algorithm;
- Bayesian Information Criterion; and
- Dynamically Orthogonal field equations.

In each case, we provide definitions and briefly justify the choices of these components in the context of oceanic and atmospheric DA. As a whole, the DO equations provide prior probabilities for a semiparametric assimilation framework based on Gaussian mixture models that are fit with an Expectation-Maximization algorithm and a Bayesian Information Criterion. Bayes' Law is then efficiently employed analytically to combine the predicted and observed information. The objective is to estimate the probabilistic properties of the dynamical state of the system under study, denoted as random state vector $\boldsymbol{X}$. For ease of notation, expositions in this section are completed in the corresponding dynamical state space. However, in computations, all Bayesian updates occur within the evolving subspace (see Sect. 3). Table 1 summarizes the notation specific to this manuscript.

a. Gaussian Mixture Models

The pdf for a random vector, $\boldsymbol{X} \in \mathbb{R}^n$, distributed according to a multivariate Gaussian mixture model (GMM) is given by

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{j=1}^{M} \pi_j \times \mathcal{N}\big(\boldsymbol{x}; \bar{\boldsymbol{x}}_j, \boldsymbol{P}_j\big), \tag{1}$$

subject to the constraint that

$$\sum_{j=1}^{M} \pi_j = 1. \tag{2}$$

We refer to $M \in \mathbb{N}$ as the mixture complexity; $\pi_j \in [0,1]$ as the mixture weights; $\bar{\boldsymbol{x}}_j \in \mathbb{R}^n$ as the mixture mean vectors; and $\boldsymbol{P}_j \in \mathbb{R}^{n \times n}$ as the mixture covariance matrices. The multivariate Gaussian density function takes the form:

$$\mathcal{N}\big(\boldsymbol{x}; \bar{\boldsymbol{x}}, \boldsymbol{P}\big) \equiv \frac{1}{(2\pi)^{n/2}\,|\boldsymbol{P}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}})^T \boldsymbol{P}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}})}. \tag{3}$$

GMMs provide an attractive *semiparametric* framework in which to approximate unknown distributions based on a set of ensemble realizations (McLachlan and Peel 2000). They are a flexible compromise between (a) a fully parametric (Gaussian) distribution for which $M = 1$ and (b) a (Gaussian) kernel density estimator (Silverman 1992) for which $M = N$, with N being the number of realizations. A single parametric distribution, while justified based on

maximum entropy arguments (Cover and Thomas 2006) often enforces too much structure onto the ensemble set and cannot model highly skewed or multimodal distributions. A kernel density estimator, on the other hand, usually requires one to retain all N realizations for the purposes of inference – a computationally burdensome task. Furthermore, due to the granularity associated with fitting a kernel to every realization, it often necessitates an heuristic choice of the kernel's shape parameter (see Sect. 5).

Mixture models efficiently summarize the ensemble set by a parameter vector, while retaining the ability to accurately model complex distributions (see figure 1). In fact, in the limit of large complexity and small covariance, a GMM converges uniformly to any sufficiently smooth distribution (Alspach and Sorenson 1972). Other mixtures and expansions have been used to approximate arbitrary probability distributions, among them the Gram-Charlier expansion, Edgeworth expansion and Pearson-type density functions (Alspach and Sorenson 1972). While the former two suffer from being invalid distributions when truncated (namely, that they must integrate to one and be positive everywhere), the latter does not lend itself well to Bayesian inference. In contrast, GMMs (1)-(3) are clearly valid.

An important property of GMMs is that they are conjugate priors to the commonly used Gaussian observation models: their Bayesian update then remains a Gaussian mixture (Casella and Berger 2001; Sondergaard 2011). Specifically, for a prior multivariate GMM,

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{j=1}^{M} \pi_j^f \times \mathcal{N}\big(\boldsymbol{x}; \bar{\boldsymbol{x}}_j^f, \boldsymbol{P}_j^f\big), \tag{4}$$

and a multivariate Gaussian observation model,

$$p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{y}; \boldsymbol{H}\boldsymbol{x}, \boldsymbol{R}\big), \tag{5}$$

the Bayesian update remains a multivariate GMM,

$$p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \sum_{j=1}^{M} \pi_j^a \times \mathcal{N}\big(\boldsymbol{x}; \bar{\boldsymbol{x}}_j^a, \boldsymbol{P}_j^a\big), \tag{6}$$

with posterior parameters:

$$\begin{aligned} \pi_j^a &= \frac{\pi_j^f \times \mathcal{N}\big(\boldsymbol{y}; \boldsymbol{H}\bar{\boldsymbol{x}}_j^f, \boldsymbol{H}\boldsymbol{P}_j^f\boldsymbol{H}^T + \boldsymbol{R}\big)}{\sum_{m=1}^{M} \pi_m^f \times \mathcal{N}\big(\boldsymbol{y}; \boldsymbol{H}\bar{\boldsymbol{x}}_m^f, \boldsymbol{H}\boldsymbol{P}_m^f\boldsymbol{H}^T + \boldsymbol{R}\big)} \\ \bar{\boldsymbol{x}}_j^a &= \bar{\boldsymbol{x}}_j^f + \boldsymbol{K}_j(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}_j^f) \\ \boldsymbol{P}_j^a &= (\boldsymbol{I} - \boldsymbol{K}_j\boldsymbol{H})\,\boldsymbol{P}_j^f, \end{aligned} \tag{7}$$

where

$$\boldsymbol{K}_j = \boldsymbol{P}_j^f\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}_j^f\boldsymbol{H}^T + \boldsymbol{R})^{-1} \tag{8}$$

is the Kalman gain matrix associated with mixture component j.

Consequently, for Gaussian observation models with GMMs as priors, the usually intractable Bayesian update reduces to an update of the elements of the parameter set, $\{\pi_1, \ldots, \pi_M, \bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_M, \boldsymbol{P}_1, \ldots, \boldsymbol{P}_M\}$, given by (7). Specifically, the individual mixture mean vectors and covariance matrices are found to be updated in accordance with familiar Kalman filter equations, the coupling occurring solely through the mixture weights.

Having introduced GMMs as an attractive method for approximating distributions for the purposes of Bayesian inference, its optimal parameter values,

$$\{\pi_1, \ldots, \pi_M, \bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_M, \boldsymbol{P}_1, \ldots, \boldsymbol{P}_M\}_{optimal},$$

need to be estimated based on a set of N ensemble realizations, $\{\boldsymbol{x}\} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. Here, we seek the value for the parameters that maximizes the probability of obtaining the given realizations; the Maximum Likelihood (ML) estimators. For this we make use of the Expectation-Maximization (EM) algorithm.

b. The Expectation-Maximization Algorithm

The EM algorithm is an iterative procedure for estimating the parameters θ_i of a target distribution that maximize the probability of obtaining a given set of realizations, $\{\boldsymbol{x}\} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. While resulting ML estimators can be justified based on intuition alone, they are also consistent and asymptotically efficient (Bertsekas and Tsitsiklis 2008). For most cases, differentiating the parametric probability distribution, $p_{\{\boldsymbol{X}\}}\big(\{\boldsymbol{x}\}; \theta_1, \ldots, \theta_M\big)$, with respect to θ_i, and equating the result to zero for maximization,

$$\frac{\partial p_{\{\boldsymbol{X}\}}\big(\{\boldsymbol{x}\}; \theta_1, \ldots, \theta_M\big)}{\partial \theta_i} = 0, \quad i = 1, \ldots, M, \tag{9}$$

results in nonlinear systems for θ_i's that lack closed form solutions. Such is also the case for GMMs. Hence, one resorts to numerical methods for obtaining the ML estimate. While various hill-climbing schemes exist, the EM algorithm takes advantage of properties of probability distributions.

Specifically, the EM algorithm (see App. A.a) is an iterative succession of expectation and maximization steps for obtaining the ML estimate. It successively estimates the weights with which a given realization is associated with each of the M mixture components. This is done based on present parameter estimates, followed by optimizing these parameters again using the newly calculated weights. Repeating this, it ultimately arrives at an estimate for the ML parameter vector based on the set of ensemble realizations, $\{\boldsymbol{x}\}$. In App. A.b, we present the EM algorithm for GMMs. The result is:
Given the set of ensemble realizations, $\{\boldsymbol{x}\} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, *and initial parameter estimate,*

$$\boldsymbol{\theta}^{(0)} = \{\pi_1^{(0)}, \ldots, \pi_M^{(0)}, \bar{\boldsymbol{x}}_1^{(0)}, \ldots, \bar{\boldsymbol{x}}_M^{(0)}, \boldsymbol{P}_1^{(0)}, \ldots, \boldsymbol{P}_M^{(0)}\},$$

repeat until convergence:

- *For all $i \in \{1,\dots,N\}$ and $j \in \{1,\dots,M\}$, use the present parameter estimate, $\boldsymbol{\theta}^{(k)}$, to form*

$$\tau_j(\boldsymbol{x}_i;\boldsymbol{\theta}^{(k)}) = \frac{\pi_j^{(k)} \times \mathcal{N}\big(\boldsymbol{x}_i;\bar{\boldsymbol{x}}_j^{(k)},\boldsymbol{P}_j^{(k)}\big)}{\sum_{m=1}^{M}\pi_m^{(k)} \times \mathcal{N}\big(\boldsymbol{x}_i;\bar{\boldsymbol{x}}_m^{(k)},\boldsymbol{P}_m^{(k)}\big)}. \quad (10)$$

- *For all $j \in \{1,\dots,M\}$, update the parameter estimate, $\boldsymbol{\theta}^{(k+1)}$, according to*

$$\pi_j^{(k+1)} = \frac{N_j^{(k)}}{N} \quad (11)$$

$$\bar{\boldsymbol{x}}_j^{(k+1)} = \frac{1}{N_j^{(k)}}\sum_{i=1}^{N}\tau_j(\boldsymbol{x}_i;\boldsymbol{\theta}^{(k)}) \times \boldsymbol{x}_i \quad (12)$$

$$\boldsymbol{P}_j^{(k+1)} = \frac{1}{N_j^{(k)}}\sum_{i=1}^{N}\tau_j(\boldsymbol{x}_i;\boldsymbol{\theta}^{(k)}) \times (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j^{(k+1)})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j^{(k+1)})^T \quad (13)$$

where

$$N_j^{(k)} \equiv \sum_{i=1}^{N}\tau_j(\boldsymbol{x}_i;\boldsymbol{\theta}^{(k)}). \quad (14)$$

Inspection of the above satisfis intuition. In the E-step of the EM algorithm, eqn. (10), we calculate the probability of mixture component j having generated realization $\boldsymbol{x}_i$ based on the present parameter estimates. We do so across all possible pairs of realizations and components. In the M-step of the EM algorithm, eqns. (11) - (13), the parameter values are updated in accordance with their weighted averages across all realizations (similar in form to eqns. (A2) - (A4) for the complete data set). As proved in App. A.b, repeated iterations of the above ensures that a local maximum for the ML parameter estimate is met. We thus arrive at an optimal fit of a GMM of complexity M to the set of N realizations, $\{\boldsymbol{x}\}$.

c. The Bayesian Information Criterion

Until now, we have assumed the mixture complexity, M, to be fixed and known. Such is rarely the case in practice, however. Determining the optimal complexity of a GMM can be a complicated task, particularly given limited *a priori* knowledge, and is often guided by empirical evidence, namely the set of ensemble realizations. Such a task is formally referred to as 'model selection'. While numerous schemes exist (e.g. Eisenberger 1964; McLachlan and Peel 2000; Duda et al. 2001), here we focus on the Bayesian Information Criterion (BIC).

Introducing a Bayesian framework, the parameter vector $\boldsymbol{\theta}$ is assumed random and M is considered constant but unknown. We denote $p_{\boldsymbol{\Theta}}\big(\boldsymbol{\theta};M\big)$ as the (arbitrary) prior distribution for $\boldsymbol{\theta}$ at a given M, and $p_{\{\boldsymbol{X}\}|\boldsymbol{\Theta}}\big(\{\boldsymbol{x}\}|\boldsymbol{\theta};M\big)$ as the distribution for the ensemble set conditioned on a $\boldsymbol{\theta}$ at a given M. In this work, the latter is a GMM.

The goal is to select the model complexity, M, that maximizes the likelihood of obtaining $\{\boldsymbol{x}\}$. In other words, by the assumed independence of the realizations, we seek M for which

$$p_{\{\boldsymbol{X}\}}\big(\{\boldsymbol{x}\};M\big) = \prod_{i=1}^{N} p_{\boldsymbol{X}_i}\big(\boldsymbol{x}_i;M\big) \quad (15)$$

is a maximum. A derivation of this optimum M is given in Sondergaard (2011). In summary, Laplace's approximation is applied to the left hand of side of Bayes' Law (MacKay 2003),

$$p_{\boldsymbol{\Theta}|\{\boldsymbol{X}\}}\big(\boldsymbol{\theta}|\{\boldsymbol{x}\};M\big) = \frac{p_{\{\boldsymbol{X}\}|\boldsymbol{\Theta}}\big(\{\boldsymbol{x}\}|\boldsymbol{\theta};M\big)p_{\boldsymbol{\Theta}}\big(\boldsymbol{\theta};M\big)}{p_{\{\boldsymbol{X}\}}\big(\{\boldsymbol{x}\};M\big)}, \quad (16)$$

evaluated at the ML estimate for the parameter vector, $\boldsymbol{\theta}$. Ultimately, we obtain:

$$\frac{1}{N}L_{\boldsymbol{x}}^{N}(M) = \frac{1}{N}L_{\boldsymbol{x}}^{N}(\hat{\boldsymbol{\theta}}_{ML},M) + \frac{1}{N}\log p_{\boldsymbol{\Theta}}\big(\hat{\boldsymbol{\theta}}_{ML};M\big) + \frac{K_M}{2N}\log 2\pi - \frac{K_M}{2N}\log N - \frac{1}{N}\log\big|\boldsymbol{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{ML})\big|, \quad (17)$$

where K_M denotes the length of the parameter vector, $\boldsymbol{J}_{\boldsymbol{x}}(\hat{\boldsymbol{\theta}}_{ML})$ defines the expected Fisher information (Bishop 2006) in any one realization, $\boldsymbol{x}_i$, evaluated at the ML estimate for the parameter vector, $\boldsymbol{\theta}$, and where we have defined the log-likelihoods:

$$L_{\boldsymbol{x}}^{N}(M) = \sum_{i=1}^{N}\log p_{\boldsymbol{X}_i}\big(\boldsymbol{x}_i;M\big) \quad (18)$$

$$L_{\boldsymbol{x}}^{N}(\hat{\boldsymbol{\theta}}_{ML},M) = \sum_{i=1}^{N}\log p_{\boldsymbol{X}_i|\boldsymbol{\Theta}}\big(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_{ML};M\big). \quad (19)$$

For large N, however, we keep only the order one terms of (17) to arrive at the BIC:

$$BIC = \min_{M}\big\{-2L_{\boldsymbol{x}}^{N}(M)\big\} \approx \min_{M}\Big\{K_M\log N - 2L_{\boldsymbol{x}}^{N}(\hat{\boldsymbol{\theta}}_{ML},M)\Big\}, \quad (20)$$

where N is the number of realizations; M is the mixture complexity; $L_{\boldsymbol{x}}^{N}(M)$ is the log-likelihood of the ensemble set integrated across all possible parameter values; $L_{\boldsymbol{x}}^{N}(\hat{\boldsymbol{\theta}}_{ML},M)$ is the log-likelihood of the ensemble set evaluated at the ML estimate for the parameter vector; and K_M is the number of parameters. M needs to be chosen to minimize the BIC.

The BIC is a quantitative equivalent of the 'Occam's Razor' (MacKay 2003; Duda et al. 2001), namely that one should favor the simplest hypothesis consistent with

the ensemble. Here, a balance is struck between underfitting - and thus imposing too much onto the data - and overfitting, for which we limit our predictive capacity beyond the ensemble. This is done by penalizing the fit of the realizations, quantified by twice the log-likelihood of the ensemble set evaluated at the ML parameter vector, $2L_{\boldsymbol{x}}^N(\hat{\boldsymbol{\theta}}_{ML}, M)$, with a term proportional to the mixture complexity, $K_M \log N$.

At this point, what remains for our DA scheme is an efficient method for evolving the probabilistic description of the state in time. For this, we employ the DO equations.

d. The Dynamically Orthogonal Field Equations

The DO equations (Sapsis and Lermusiaux 2009; Sapsis 2010), are a closed reduced set of evolution equations for general stochastic continuous fields, $X(\boldsymbol{r}, t; \omega)$, described by a stochastic partial differential equation (SPDE):

$$\frac{\partial X(\boldsymbol{r}, t; \omega)}{\partial t} = \mathcal{L}[X(\boldsymbol{r}, t; \omega); \omega], \tag{21}$$

with initial conditions

$$X(\boldsymbol{r}, t_0; \omega) = X_0(\boldsymbol{r}; \omega) \tag{22}$$

and boundary conditions

$$\mathcal{B}\left[X(\boldsymbol{r}, t; \omega)\right]\big|_{\boldsymbol{r}=\boldsymbol{\xi}} = h(\boldsymbol{\xi}, t; \omega), \tag{23}$$

where $\boldsymbol{r}$ denotes the position in space; t is time; ω a random event; $\mathcal{L}[\cdot]$ a general, potentially nonlinear, differential operator (presently, an ocean or fluid flow model); $\mathcal{B}$ a linear differential operator; and, $\boldsymbol{\xi}$ the spatial coordinate denoting the boundary. Two main assumptions are made in the derivation of the DO equations. First, a generalized, *time-dependent* Karhunen-Loeve decomposition of the fields (Lermusiaux 2006; Sapsis and Lermusiaux 2009) is used,

$$X(\boldsymbol{r}, t; \omega) = \bar{x}(\boldsymbol{r}, t) + \sum_{i=1}^{s(t)} \tilde{x}_i(\boldsymbol{r}, t)\Phi_i(t; \omega), \tag{24}$$

where $\bar{x}(\boldsymbol{r}, t) = E\left[X(\boldsymbol{r}, t; \omega)\right]$ are the mean fields with $E\left[\bullet\right]$ being the expectation operator over ω; $\tilde{x}_i(\boldsymbol{r}, t)$ are orthonormal modes spanning the time-dependent stochastic subspace; and $\Phi_i(t; \omega)$ are zero-mean, stochastic coefficients. The decomposition (24) defines generalized Empirical Orthogonal Functions. In addition to $\tilde{x}_i(\boldsymbol{r}, t)$, the dimension of the subspace s also varies with time, but in what follows, for ease of notation, we omit t next to s. Second, after insertion of (24) into (21), a DO condition is imposed, i.e. the rate of change of the stochastic subspace basis is orthogonal to itself over the physical domain,

$$\left\langle \frac{\partial \tilde{x}_i(\cdot, t)}{\partial t}, \tilde{x}_j(\cdot, t) \right\rangle = 0 \quad \forall i, j \in \{1, \ldots, s\}\,. \tag{25}$$

With these assumptions, the original SPDE is reduced to DO equations (see definition below):

i. a PDE (26) for the evolution of the mean field, $\bar{x}(\boldsymbol{r}, t)$;

ii. a family of s PDEs (27) for the evolution of the orthonormal modes $\tilde{x}_i(\boldsymbol{r}, t)$ describing a basis for the time-dependent dominant stochastic subspace; and,

iii. a system of s stochastic differential equations (28) for the coefficients, $\Phi_i(t; \omega)$, that define how the stochasticity evolves within the stochastic subspace.

Mathematically, for the governing dynamics (21)*, with initial and boundary conditions* (22) *and* (23)*, the coupled DO evolution equations are (using Einstein notation, $\sum_i a_i b_i \equiv a_i b_i$):*

$$\frac{\partial \bar{x}(\boldsymbol{r}, t)}{\partial t} = E\left[\mathcal{L}[X(\boldsymbol{r}, t; \omega); \omega]\right], \tag{26}$$

$$\frac{\partial \tilde{x}_i(\boldsymbol{r}, t)}{\partial t} = \boldsymbol{\Pi}^{\perp}\left(E\left[\mathcal{L}[X(\boldsymbol{r}, t; \omega); \omega]\phi_j(t; \omega)\right]\right) C^{-1}_{\Phi_i(t)\Phi_j(t)}, \tag{27}$$

$$\frac{d\Phi_i(t; \omega)}{dt} = \left\langle \mathcal{L}[X(\cdot, t; \omega); \omega] - E\left[\mathcal{L}[X(\cdot, t; \omega); \omega]\right], \tilde{x}_i(\cdot, t)\right\rangle, \tag{28}$$

where

$$\boldsymbol{\Pi}^{\perp}(F(\boldsymbol{r})) \equiv F(\boldsymbol{r}) - \langle F(\cdot), \tilde{x}_k(\cdot, t)\rangle\, \tilde{x}_k(\boldsymbol{r}, t) \tag{29}$$

is the projection of $F(\boldsymbol{r})$ onto the null space of the stochastic subspace; and,

$$C_{\Phi_i(t)\Phi_j(t)} \equiv E\left[\Phi_i(t; \omega)\Phi_j(t; \omega)\right] \tag{30}$$

is the correlation between random variables $\Phi_i(t; \omega)$ and $\Phi_j(t; \omega)$. The associated boundary conditions take the form

$$\mathcal{B}\left[\bar{x}(\boldsymbol{r}, t)\right]\big|_{\boldsymbol{r}=\boldsymbol{\xi}} = E\left[h(\boldsymbol{\xi}, t; \omega)\right] \tag{31}$$

$$\mathcal{B}\left[\tilde{x}_i(\boldsymbol{r}, t)\right]\big|_{\boldsymbol{r}=\boldsymbol{\xi}} = E\left[h(\boldsymbol{\xi}, t; \omega)\Phi_j(t; \omega)\right] C^{-1}_{\Phi_i(t)\Phi_j(t)} \tag{32}$$

and the initial conditions are given by

$$\bar{x}(\boldsymbol{r}, t_0) = \bar{x}_0(\boldsymbol{r}) = E\left[X_0(\boldsymbol{r}; \omega)\right] \tag{33}$$

$$\tilde{x}_i(\boldsymbol{r}, t_0) = \tilde{x}_{i0}(\boldsymbol{r}) \tag{34}$$

$$\Phi_i(t_0; \omega) = \langle X_0(\cdot; \omega) - \bar{x}_0(\cdot), \tilde{x}_{i0}(\cdot)\rangle \tag{35}$$

where $i = 1, \ldots, s$ and $\tilde{x}_{i0}(\boldsymbol{r})$ are the orthonormal modes for the stochastic subspace at t_0.

With the DO equations, both the stochastic subspace and the stochastic coefficients are dynamically evolved in time. They are initialized based on the initial pdf and thereafter evolved in accord with the SPDE governing $X(\boldsymbol{r}, t; \omega)$ and its boundary conditions. This evolution is an advantage when compared to the Proper Orthogonal Decomposition (Papoulis 1965; Holmes et al. 1996) and Polynomial

Chaos (Ghanem and Spanos 1991) which both fix in time parts of their truncated expansion, the former the stochastic subspace and the latter the form of the stochastic coefficients. We note that s can also be evolved based on the dynamics and external observations (Sapsis and Lermusiaux 2011), as done in Error Subspace Statistical Estimation (ESSE), (Lermusiaux 1999b).

3. The GMM-DO Filter

Combining the components described in Section 2, and building on the foundations of classical assimilation schemes, we now complete the derivation of the GMM-DO filter: data assimilation with GMMs using the DO equations. The result is an efficient, data-driven scheme that preserves non-Gaussian statistics and respects nonlinear dynamics.

The GMM-DO filter consists of a recursive succession of two distinct steps: a forecast step and an update step. The Bayesian assimilation is the update step. As will be proved, this update is efficiently computed within the evolving subspace and the result is equivalent to the Bayesian update in the dynamical state space. For today's ocean and atmosphere simulations, the subspace update is computationally feasible. We refer to table 1 for notation.

a. Initial Conditions

We initialize the state vector at discrete time $k = 0$ in a decomposed form,

$$\boldsymbol{X}_0 = \bar{\boldsymbol{x}}_0 + \sum_{i=1}^{s_0} \tilde{\boldsymbol{x}}_{i,0}\, \Phi_{i,0}(\omega)\,, \tag{36}$$

that accords with the DO equations. The initial state mean, $\bar{\boldsymbol{x}}_0$, orthonormal modes, $\tilde{\boldsymbol{x}}_{i,0}$, and stochastic coefficients, $\Phi_{i,0}(\omega)$, are chosen so as to best represent the initial probabilistic state. Various representations and discretizations for the coefficients, $\Phi_i(t;\omega)$, exist (Sapsis and Lermusiaux 2009; Ueckermann et al. 2012), several of which can be employed with our GMM-DO scheme. Here, we adopt a Monte Carlo approach: we draw N realizations of the multivariate random vector, $\{\Phi_{1,0}(\omega), \ldots, \Phi_{s_0,0}(\omega)\}$, to obtain the matrix,

$$\{\boldsymbol{\phi}_0\} = \{\boldsymbol{\phi}_{1,0}\,, \ldots,\, \boldsymbol{\phi}_{N,0}\}. \tag{37}$$

We emphasize that the $\boldsymbol{\phi}_{r,0} \in \mathbb{R}^{s_0}$ represent realizations residing in the initial stochastic subspace of dimension s_0. With this, we rewrite (36) in its Monte Carlo ensemble form,

$$\boldsymbol{x}_{r,0} = \bar{\boldsymbol{x}}_0 + \boldsymbol{\mathcal{X}}_0\, \boldsymbol{\phi}_{r,0}\,, \quad r = \{1, \ldots, N\}, \tag{38}$$

where $\boldsymbol{\mathcal{X}}_0 \in \mathbb{R}^{n \times s}$ (table 1) is the matrix of modes forming an orthonormal basis for the initial subspace. This $\boldsymbol{\mathcal{X}}_0$ is evolved in time by dynamics and random forcing in (27).

b. Forecast

Starting from either the initial DO conditions or the posterior state description following the assimilation of data at time $k-1$ (i.e. the Bayesian GMM update at $k-1$),

$$\boldsymbol{x}^a_{r,k-1} = \bar{\boldsymbol{x}}^a_{k-1} + \boldsymbol{\mathcal{X}}^a_{k-1}\, \boldsymbol{\phi}^a_{r,k-1}, \quad r = \{1, \ldots, N\}, \tag{39}$$

we use the stochastic DO equations, (26)–(28), to efficiently evolve the probabilistic description of the state vector in time, arriving at a forecast for observation time k:

$$\boldsymbol{x}^f_{r,k} = \bar{\boldsymbol{x}}^f_k + \boldsymbol{\mathcal{X}}^f_k\, \boldsymbol{\phi}^f_{r,k}, \quad r = \{1, \ldots, N\}. \tag{40}$$

This forecast is efficiently computed using the numerical schemes derived by Ueckermann et al. (2012). Specifically, for the mean and modes, we employ a second-order finite-volume spatial discretization and DO-specific projection method, and for the stochastic coefficients, a second or fourth order integration scheme in time.

As (39) and (40) indicate, all of the mean, orthonormal modes and coefficients are evolved during the forecast from t_{k-1} to t_k. In particular, the span of the modes $\boldsymbol{\mathcal{X}}^f_k$ differs from that of $\boldsymbol{\mathcal{X}}^a_{k-1}$: the subspace evolves with time in between data assimilation.

c. Observation

Common to oceanic and atmospheric applications, we employ here a linear (or linearized) observation model,

$$\boldsymbol{Y}_k = \boldsymbol{H}\boldsymbol{X}_k + \boldsymbol{\Upsilon}_k, \quad \boldsymbol{\Upsilon}_k \sim \mathcal{N}(\boldsymbol{\upsilon}_k; \boldsymbol{0}, \boldsymbol{R})\,. \tag{41}$$

where $\boldsymbol{Y}_k \in \mathbb{R}^p$ is the observation random vector at time k, $\boldsymbol{H} \in \mathbb{R}^{p \times n}$ is the linear observation model and $\boldsymbol{\Upsilon}_k \in \mathbb{R}^p$ the corresponding random noise vector, assumed to be of a Gaussian distribution. We denote the realized observation vector by $\boldsymbol{y}_k \in \mathbb{R}^p$ and realized noise vector by $\boldsymbol{\upsilon}_k \in \mathbb{R}^p$. This observation model could be generalized to other forms, which would lead to variations in the following update scheme.

d. Update

The whole update occurs at fixed discrete time instant and, in what follows, we thus omit the subscript time index k. In the update, the subspace is for now assumed unchanged by the observations[1]: the notation $(\cdot)^f$ or $(\cdot)^a$ is thus not used on the modes $\boldsymbol{\mathcal{X}}$. Of course, observations affect the subspace evolution after each assimilation since the DO equations (26)–(28) are coupled. In conclusion, starting from the prior, here the DO forecast,

$$\boldsymbol{x}^f_r = \bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}}\boldsymbol{\phi}^f_r, \quad r = \{1, \ldots, N\}, \tag{42}$$

[1] As an aside, in ESSE (Lermusiaux 1999b), the update consists of two parts: data assimilation in a fixed-subspace followed by a correction of the subspace based on the innovation vector and posterior misfit. This results in prior and posterior subspaces that differ. We can generalize this subspace learning scheme to the present Bayesian GMM-DO framework, but this is not done here.

the goal is to update the mean state $\bar{\boldsymbol{x}}^f$ and set of realizations, $\{\boldsymbol{\phi}^f\} = \{\boldsymbol{\phi}_1^f, \ldots, \boldsymbol{\phi}_N^f\}$, in accordance with (41) and realized observations $\boldsymbol{y}$, to obtain the posterior GMM-DO estimate:

$$\boldsymbol{x}_r^a = \bar{\boldsymbol{x}}^a + \boldsymbol{\mathcal{X}}\boldsymbol{\phi}_r^a, \quad r = \{1, \ldots, N\}. \tag{43}$$

To do so, we first optimally fit a GMM (Section 2a) to the forecast set of realizations in the stochastic subspace. This prior GMM estimate is then updated within the subspace, in accordance with observations and Bayes' Law, ultimately leading to the posterior GMM-DO estimate (43). In what follows, we derive and describe this GMM-DO update algorithm.

1) GMM representation of prior set of ensemble realizations

At the time of a new set of measurements, $\boldsymbol{y}$, we use the EM algorithm and BIC to determine the GMM that best represents the set of ensemble realizations *within the stochastic subspace*, $\{\boldsymbol{\phi}^f\} = \{\boldsymbol{\phi}_1^f, \ldots, \boldsymbol{\phi}_N^f\}$. We denote the parameters of the GMM by

$$\pi_j^f, \boldsymbol{\mu}_j^f, \boldsymbol{\Sigma}_j^f, \quad j = 1, \ldots, M,$$

where $\pi_j^f \in [0,1]$, $\boldsymbol{\mu}_j^f \in \mathbb{R}^s$ and $\boldsymbol{\Sigma}_j^f \in \mathbb{R}^{s\times s}$. We again stress that the GMM efficiently resides in an s-dimensional subspace of the n-dimensional dynamical state space, with $s \ll n$, thus making the prior estimation procedure computationally feasible.

We determine the optimal mixture complexity by application of the BIC, (20), successively fitting GMMs of increasing complexity (i.e. $M = 1, 2, 3, \ldots$) with the EM algorithm, until a minimum of the BIC is met. The final result is a GMM optimally fit to the ensemble of realizations *in the stochastic subspace*. We write the resulting prior pdf of this GMM as:

$$p_{\boldsymbol{\Phi}^f}(\boldsymbol{\phi}^f) = \sum_{j=1}^{M} \pi_j^f \times \mathcal{N}(\boldsymbol{\phi}^f; \boldsymbol{\mu}_j^f, \boldsymbol{\Sigma}_j^f). \tag{44}$$

Due to the affine transformation (42) linking the stochastic subspace with the state space, we may expand the previously determined GMM into the state space according to:

$$\bar{\boldsymbol{x}}_j^f = \bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}}\boldsymbol{\mu}_j^f \tag{45}$$

$$\boldsymbol{P}_j^f = \boldsymbol{\mathcal{X}}\boldsymbol{\Sigma}_j^f\boldsymbol{\mathcal{X}}^T. \tag{46}$$

This is a key property of our GMM-DO filter. The mixture weights, π_j^f, naturally remain unchanged. We note that $\bar{\boldsymbol{x}}_j^f$ and $\boldsymbol{P}_j^f$ now refer to the mean vector and covariance matrix, respectively, for mixture component j *in the state space*. We thus arrive at the prior distribution for the state vector in state space, taking the form of the following GMM:

$$p_{\boldsymbol{X}^f}(\boldsymbol{x}^f) = \sum_{j=1}^{M} \pi_j^f \times \mathcal{N}(\boldsymbol{x}^f; \bar{\boldsymbol{x}}_j^f, \boldsymbol{P}_j^f). \tag{47}$$

We emphasize that, due to the affine transformation (42), this distribution would equally have been obtained had we performed the prior fitting of the GMM directly in the state space based on the set of realizations $\{\boldsymbol{x}^f\} = \{\boldsymbol{x}_1^f, \ldots, \boldsymbol{x}_N^f\}$.

2) Bayesian update

Since the uncertainty of the state is restricted to the stochastic subspace, we prove next that the Bayesian update can be performed therein. In doing so, we again make use of the affine transformations (42)–(43) linking the stochastic subspace with the state space. We re-emphasize that presently, this subspace, described by the matrix $\boldsymbol{\mathcal{X}}$, is assumed to remain unaffected by the assimilation. The result of the theorem, of course, provides an efficient implementation of the GMM-DO filter's update step, with significant computational savings due to the reduced dimensionality, $s \ll n$. For realistic modeling with large state vectors, only this update is computationally feasible.

Theorem 1
Given the GMM fit (47) *to the DO forecast as prior distribution and the realized observation vector* $\boldsymbol{y}$ *with observation model* (41) *of Gaussian distribution, the posterior distribution* $p_{\boldsymbol{X}^a}(\boldsymbol{x}^a)$ *of the state vector in the state space is obtained by Bayesian update of* (44) *carried out in the stochastic subspace. The result* $p_{\boldsymbol{\Phi}^a}(\boldsymbol{\phi}^a)$ *is equivalent to updating* $p_{\boldsymbol{X}^f}(\boldsymbol{x}^f)$ *directly. Specifically, the update equations for the mean* $\bar{\boldsymbol{x}}^f$ *and parameters* π_j^f, $\boldsymbol{\mu}_j^f$ *and* $\boldsymbol{\Sigma}_j^f$ *are:*

$$\bar{\boldsymbol{x}}^a = \bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}}\sum_{j=1}^{M} \pi_j^a \times \hat{\boldsymbol{\mu}}_j^a \tag{48}$$

$$= \bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}}\sum_{j=1}^{M} \pi_j^a \times (\boldsymbol{\mu}_j^f + \tilde{\boldsymbol{K}}_j(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{H}}\boldsymbol{\mu}_j^f)) \tag{49}$$

$$\pi_j^a = \frac{\pi_j^f \times \mathcal{N}(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{H}}\boldsymbol{\mu}_j^f, \tilde{\boldsymbol{H}}\boldsymbol{\Sigma}_j^f\tilde{\boldsymbol{H}}^T + \boldsymbol{R})}{\sum_{m=1}^{M} \pi_m^f \times \mathcal{N}(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{H}}\boldsymbol{\mu}_m^f, \tilde{\boldsymbol{H}}\boldsymbol{\Sigma}_m^f\tilde{\boldsymbol{H}}^T + \boldsymbol{R})} \tag{50}$$

$$\boldsymbol{\mu}_j^a = \hat{\boldsymbol{\mu}}_j^a - \sum_{j=1}^{M} \pi_j^a \times \hat{\boldsymbol{\mu}}_j^a \tag{51}$$

$$\boldsymbol{\Sigma}_j^a = (\boldsymbol{I} - \tilde{\boldsymbol{K}}_j\tilde{\boldsymbol{H}})\boldsymbol{\Sigma}_j^f. \tag{52}$$

with the definitions

$$\tilde{\boldsymbol{H}} \equiv \boldsymbol{H}\boldsymbol{\mathcal{X}} \tag{53}$$

$$\tilde{\boldsymbol{y}} \equiv \boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}^f \tag{54}$$

$$\tilde{\boldsymbol{K}}_j \equiv \boldsymbol{\Sigma}_j^f\tilde{\boldsymbol{H}}^T(\tilde{\boldsymbol{H}}\boldsymbol{\Sigma}_j^f\tilde{\boldsymbol{H}}^T + \boldsymbol{R})^{-1} \equiv \boldsymbol{\mathcal{X}}^T\boldsymbol{K}_j\,. \tag{55}$$

Proof

Bayesian update in the state space. Applying the Bayesian update equations (4)–(7) of Section 2a to the GMM prior (47) and observation model (41), we first obtain the posterior distribution for the state vector in the state space:

$$p_{\boldsymbol{X}^a}(\boldsymbol{x}^a) = \sum_{j=1}^{M} \pi_j^a \times \mathcal{N}\left(\boldsymbol{x}^a; \bar{\boldsymbol{x}}_j^a, \boldsymbol{P}_j^a\right) \tag{56}$$

with

$$\pi_j^a = \frac{\pi_j^f \times \mathcal{N}\big(\boldsymbol{y}; \boldsymbol{H}\bar{\boldsymbol{x}}_j^f, \boldsymbol{H}\boldsymbol{P}_j^f\boldsymbol{H}^T + \boldsymbol{R}\big)}{\sum_{m=1}^{M} \pi_m^f \times \mathcal{N}\big(\boldsymbol{y}; \boldsymbol{H}\bar{\boldsymbol{x}}_m^f, \boldsymbol{H}\boldsymbol{P}_m^f\boldsymbol{H}^T + \boldsymbol{R}\big)} \tag{57}$$

$$\bar{\boldsymbol{x}}_j^a = \bar{\boldsymbol{x}}_j^f + \boldsymbol{K}_j(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}_j^f) \tag{58}$$

$$\boldsymbol{P}_j^a = (\boldsymbol{I} - \boldsymbol{K}_j\boldsymbol{H})\boldsymbol{P}_j^f \tag{59}$$

where

$$\boldsymbol{K}_j = \boldsymbol{P}_j^f\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}_j^f\boldsymbol{H}^T + \boldsymbol{R})^{-1} \tag{60}$$

is the Kalman gain matrix associated with mixture component j.

With this, we can derive the expression for the posterior mean field in the state space,

$$\bar{\boldsymbol{x}}^a = \sum_{j=1}^{M} \pi_j^a \times \bar{\boldsymbol{x}}_j^a \tag{61}$$

$$= \sum_{j=1}^{M} \pi_j^a \times (\bar{\boldsymbol{x}}_j^f + \boldsymbol{K}_j(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}_j^f))\,, \tag{62}$$

as well as for other moments in the state space (see Remark hereafter). This completes the Bayesian update in the full state space, with the posterior mean vector $\bar{\boldsymbol{x}}^a$ and GMM parameters all expressed in terms of the state space quantities and realized observations $\boldsymbol{y}$.

Bayesian update in the stochastic space. Now, we show that the Bayesian update in the state space defined by (56)–(59) and (62) is equivalent to a Bayesian update in the stochastic DO subspace. We first remark that using (53) and (55) is computationally efficient. To derive (55), we use identity (46), orthonormality of the modes and definition (53),

$$\tilde{\boldsymbol{K}}_j \equiv \boldsymbol{\Sigma}_j^f\tilde{\boldsymbol{H}}^T(\tilde{\boldsymbol{H}}\boldsymbol{\Sigma}_j^f\tilde{\boldsymbol{H}}^T + \boldsymbol{R})^{-1} = \boldsymbol{\Sigma}_j^f\boldsymbol{\mathcal{X}}^T\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{\mathcal{X}}\boldsymbol{\Sigma}_j^f\boldsymbol{\mathcal{X}}^T\boldsymbol{H}^T + \boldsymbol{R})^{-1}$$
$$= \boldsymbol{\mathcal{X}}^T\boldsymbol{P}_j^f\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}_j^f\boldsymbol{H}^T + \boldsymbol{R})^{-1} = \boldsymbol{\mathcal{X}}^T\boldsymbol{K}_j\,.$$

Deriving next the update equation (50) for the mixture weights, we start from (57) and use (45) and (46), to obtain:

$$\pi_j^a = \frac{\pi_j^f \times \mathcal{N}\big(\boldsymbol{y}; \boldsymbol{H}\bar{\boldsymbol{x}}_j^f, \boldsymbol{H}\boldsymbol{P}_j^f\boldsymbol{H}^T + \boldsymbol{R}\big)}{\sum_{m=1}^{M} \pi_m^f \times \mathcal{N}\big(\boldsymbol{y}; \boldsymbol{H}\bar{\boldsymbol{x}}_m^f, \boldsymbol{H}\boldsymbol{P}_m^f\boldsymbol{H}^T + \boldsymbol{R}\big)} \tag{63}$$

$$= \frac{\pi_j^f \times \mathcal{N}\big(\boldsymbol{y}; \boldsymbol{H}(\bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}}\boldsymbol{\mu}_j^f), \boldsymbol{H}\boldsymbol{\mathcal{X}}\boldsymbol{\Sigma}_j^f\boldsymbol{\mathcal{X}}^T\boldsymbol{H}^T + \boldsymbol{R}\big)}{\sum_{m=1}^{M} \pi_m^f \times \mathcal{N}\big(\boldsymbol{y}; \boldsymbol{H}(\bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}}\boldsymbol{\mu}_m^f), \boldsymbol{H}\boldsymbol{\mathcal{X}}\boldsymbol{\Sigma}_m^f\boldsymbol{\mathcal{X}}^T\boldsymbol{H}^T + \boldsymbol{R}\big)}, \tag{64}$$

which becomes by simple rearranging of terms,

$$= \frac{\pi_j^f \times \mathcal{N}\big(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}^f; \boldsymbol{H}\boldsymbol{\mathcal{X}}\boldsymbol{\mu}_j^f, \boldsymbol{H}\boldsymbol{\mathcal{X}}\boldsymbol{\Sigma}_j^f\boldsymbol{\mathcal{X}}^T\boldsymbol{H}^T + \boldsymbol{R}\big)}{\sum_{m=1}^{M} \pi_m^f \times \mathcal{N}\big(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}^f; \boldsymbol{H}\boldsymbol{\mathcal{X}}\boldsymbol{\mu}_m^f, \boldsymbol{H}\boldsymbol{\mathcal{X}}\boldsymbol{\Sigma}_m^f\boldsymbol{\mathcal{X}}^T\boldsymbol{H}^T + \boldsymbol{R}\big)}. \tag{65}$$

Then, applying definitions (53) and (54) leads to:

$$\pi_j^a = \frac{\pi_j^f \times \mathcal{N}\big(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{H}}\boldsymbol{\mu}_j^f, \tilde{\boldsymbol{H}}\boldsymbol{\Sigma}_j^f\tilde{\boldsymbol{H}}^T + \boldsymbol{R}\big)}{\sum_{m=1}^{M} \pi_m^f \times \mathcal{N}\big(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{H}}\boldsymbol{\mu}_m^f, \tilde{\boldsymbol{H}}\boldsymbol{\Sigma}_m^f\tilde{\boldsymbol{H}}^T + \boldsymbol{R}\big)}, \tag{66}$$

With this, we obtain an efficient update equation for the mixture weights using vectors and matrices specific to the subspace, all the while retaining the familiar structure of (57).

In a similar manner, to derive (48), (49) and (51) for the posterior mean $\bar{\boldsymbol{x}}^a$ and mixture means $\boldsymbol{\mu}_j^a$, we start with (62), use (45) and apply definition (55) to obtain:

$$\bar{\boldsymbol{x}}^a = \sum_{j=1}^{M} \pi_j^a \times (\bar{\boldsymbol{x}}_j^f + \boldsymbol{K}_j(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}_j^f)) \tag{67}$$

$$= \sum_{j=1}^{M} \pi_j^a \times (\bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}}\boldsymbol{\mu}_j^f + \boldsymbol{\mathcal{X}}\tilde{\boldsymbol{K}}_j(\boldsymbol{y} - \boldsymbol{H}(\bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}}\boldsymbol{\mu}_j^f))) \tag{68}$$

which becomes, using $\sum_{j=1}^{M} \pi_j^a \times \bar{\boldsymbol{x}}^f = \bar{\boldsymbol{x}}^f$ and applying definitions (53) and (54),

$$= \bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}} \sum_{j=1}^{M} \pi_j^a \times (\boldsymbol{\mu}_j^f + \tilde{\boldsymbol{K}}_j(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{H}}\boldsymbol{\mu}_j^f))\,. \tag{69}$$

As a result, we obtain,

$$\bar{\boldsymbol{x}}^a \equiv \bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}} \sum_{j=1}^{M} \pi_j^a \times \hat{\boldsymbol{\mu}}_j^a, \tag{70}$$

where we have defined "intermediate" mean vectors in the stochastic subspace,

$$\hat{\boldsymbol{\mu}}_j^a = \boldsymbol{\mu}_j^f + \tilde{\boldsymbol{K}}_j(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{H}}\boldsymbol{\mu}_j^f). \tag{71}$$

These "intermediate" vectors, when adequately combined and weighted, are the contribution of our Bayesian GMM-DO update to the conditional mean state $\bar{\boldsymbol{x}}^a$ from the forecast mean state $\bar{\boldsymbol{x}}^f$. We refer to these M vectors as "intermediate" means from the fact that our DO framework requires that the parametric distribution describing the stochastic subspace is of mean zero, i.e. $\sum_{j=1}^{M} \pi_j^a \times \boldsymbol{\mu}_j^a = 0$. This condition is obviously not satisfied by $\hat{\boldsymbol{\mu}}_j^a$. The actual

means of the posterior mixture components in the subspace can be obtained by a reset of these intermediate means:

$$\boldsymbol{\mu}_j^a \mapsto \hat{\boldsymbol{\mu}}_j^a - \sum_{j=1}^{M} \pi_j^a \times \hat{\boldsymbol{\mu}}_j^a. \tag{72}$$

Rather than merely stating this as a matter of fact, however, we now derive this result. Similarly to (45), we first write:

$$\bar{\boldsymbol{x}}_j^a = \bar{\boldsymbol{x}}^a + \boldsymbol{\mathcal{X}}\boldsymbol{\mu}_j^a. \tag{73}$$

By subtraction of $\bar{\boldsymbol{x}}^a$ and left multiplication by $\boldsymbol{\mathcal{X}}^T$, we then obtain:

$$\boldsymbol{\mu}_j^a = (\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}})^{-1}\boldsymbol{\mathcal{X}}^T\left(\bar{\boldsymbol{x}}_j^a - \bar{\boldsymbol{x}}^a\right) \tag{74}$$

$$= \boldsymbol{\mathcal{X}}^T\left(\bar{\boldsymbol{x}}_j^a - \bar{\boldsymbol{x}}^a\right), \tag{75}$$

where (75) results from the orthonormality of the modes, i.e. $\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}} = \boldsymbol{I}$. We now have, inserting (58) and (70) in (75),

$$\begin{aligned}\boldsymbol{\mu}_j^a &= \boldsymbol{\mathcal{X}}^T\left(\bar{\boldsymbol{x}}_j^a - \bar{\boldsymbol{x}}^a\right) \\ &= \boldsymbol{\mathcal{X}}^T(\bar{\boldsymbol{x}}_j^f + \boldsymbol{K}_j(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}_j^f) - \bar{\boldsymbol{x}}^f - \boldsymbol{\mathcal{X}}\sum_{j=1}^{M}\pi_j^a\times\hat{\boldsymbol{\mu}}_j^a),\end{aligned} \tag{76}$$

and then using (45), definition (55) and the orthonormality of the modes,

$$\begin{aligned}&= \boldsymbol{\mathcal{X}}^T(\boldsymbol{\mathcal{X}}\boldsymbol{\mu}_j^f + \boldsymbol{\mathcal{X}}\tilde{\boldsymbol{K}}_j(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}_j^f) - \boldsymbol{\mathcal{X}}\sum_{j=1}^{M}\pi_j^a\times\hat{\boldsymbol{\mu}}_j^a) \\ &= \boldsymbol{\mu}_j^f + \tilde{\boldsymbol{K}}_j(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{H}}\bar{\boldsymbol{x}}_j^f) - \sum_{j=1}^{M}\pi_j^a\times\hat{\boldsymbol{\mu}}_j^a\,.\end{aligned} \tag{77}$$

Hence, we derive (51),

$$\boldsymbol{\mu}_j^a = \hat{\boldsymbol{\mu}}_j^a - \sum_{j=1}^{M}\pi_j^a\times\hat{\boldsymbol{\mu}}_j^a\,. \tag{78}$$

Finally, to derive (52) that expresses the updated mixture covariance matrices, $\boldsymbol{\Sigma}_j^a$, in terms of DO subspace quantities, we proceed similarly. As in (46), we expend $\boldsymbol{P}_j^a$

$$\boldsymbol{P}_j^a = \boldsymbol{\mathcal{X}}\boldsymbol{\Sigma}_j^a\boldsymbol{\mathcal{X}}^T \tag{79}$$

and then equate (79) to (59), inserting (46), to obtain,

$$\begin{aligned}\boldsymbol{P}_j^a &= \boldsymbol{\mathcal{X}}\boldsymbol{\Sigma}_j^a\boldsymbol{\mathcal{X}}^T \\ &= (\boldsymbol{I} - \boldsymbol{K}_j\boldsymbol{H})\boldsymbol{P}_j^f = (\boldsymbol{I} - \boldsymbol{K}_j\boldsymbol{H})\boldsymbol{\mathcal{X}}\boldsymbol{\Sigma}_j^f\boldsymbol{\mathcal{X}}^T\,.\end{aligned}$$

We then left multiply by $\boldsymbol{\mathcal{X}}^T$ and right multiply by $\boldsymbol{\mathcal{X}}$, and use definition (55), to obtain:

$$\begin{aligned}\boldsymbol{\Sigma}_j^a &= \boldsymbol{\mathcal{X}}^T(\boldsymbol{I} - \boldsymbol{K}_j\boldsymbol{H})\boldsymbol{P}_j^f\boldsymbol{\mathcal{X}} \\ &= \boldsymbol{\mathcal{X}}^T(\boldsymbol{I} - \boldsymbol{\mathcal{X}}\tilde{\boldsymbol{K}}_j\boldsymbol{H})\boldsymbol{\mathcal{X}}\boldsymbol{\Sigma}_j^f\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}} \\ &= (\boldsymbol{I} - \tilde{\boldsymbol{K}}_j\tilde{\boldsymbol{H}})\boldsymbol{\Sigma}_j^f.\end{aligned} \tag{80}$$

where the orthonormality of the modes and definition (53) have been used. □

With the above theorem, we have derived efficient expressions (48)–(52) for the GMM-DO update in the time-dependent stochastic subspace. To conclude, we note the similarity of these GMM-DO filter equations for a Bayesian update with the corresponding ESSE equations for Gaussian update, both of which occur in the stochastic subspace.

Remark: *Although strictly unnecessary for the GMM-DO filter, we can also obtain all updated state space quantities. For example, the full posterior covariance matrix in the state space can be obtained using the Law of Total Variance (Bertsekas and Tsitsiklis 2008):*

$$\boldsymbol{P}^a = \sum_{j=1}^{M}\pi_j^a\times\boldsymbol{P}_j^a + \sum_{j=1}^{M}\pi_j^a\times\left(\bar{\boldsymbol{x}}_j^a - \bar{\boldsymbol{x}}^a\right)\left(\bar{\boldsymbol{x}}_j^a - \bar{\boldsymbol{x}}^a\right)^T. \tag{81}$$

3) Generation of posterior set of ensemble realizations

We complete the update step, as with ESSE scheme A (Lermusiaux and Robinson 1999), by generating a posterior set of realizations *within the stochastic subspace*, $\{\boldsymbol{\phi}^a\} = \{\boldsymbol{\phi}_1^a, \ldots, \boldsymbol{\phi}_N^a\}$, according to the posterior multivariate GMM, $p_{\boldsymbol{\Phi}^a}(\boldsymbol{\phi}^a)$, with parameters

$$\pi_j^a, \boldsymbol{\mu}_j^a, \boldsymbol{\Sigma}_j^a, \qquad j = 1, \ldots, M.$$

With this, we arrive at the posterior DO representation in Monte Carlo form for the state vector based on a Bayesian assimilation of the observations, $\boldsymbol{y}$, at time k:

$$\boldsymbol{x}_{r,k}^a = \bar{\boldsymbol{x}}_k^a + \boldsymbol{\mathcal{X}}_k\boldsymbol{\phi}_{r,k}^a\,, \quad r = \{1, \ldots, N\}. \tag{82}$$

We note that the size of the prior and posterior ensembles at time k in the stochastic subspace do not need to be the same: e.g. N can be evolved by a convergence criterion for the DO forecast from time k to the next observation time $k+1$ (Lermusiaux 2007; Ueckermann et al. 2012). This concludes the derivation of the GMM-DO filter. We summarize the algorithm using the flowchart displayed in figure 2. We note that extensions of this GMM-DO filter algorithm are provided in Appendix B: specifically, an algorithm for limiting the GMM fit to a dominant subspace in the full stochastic DO subspace as well as an algorithm for constraining the means of the GMM.

Next, we illustrate the GMM-DO filter procedure by way of a simple toy example. More realistic applications are provided in Part II (Sondergaard and Lermusiaux 2012).

4. Example

Assume we are provided with the following (arbitrarily chosen) forecast for the DO decomposed representation of the state:

$$\bar{\boldsymbol{x}}^f = \begin{bmatrix}1\\2\\3\end{bmatrix} \quad \text{and} \quad \boldsymbol{\mathcal{X}} = \begin{bmatrix}1 & 0\\0 & 1\\0 & 0\end{bmatrix},$$

with one hundred subspace realizations, $\{\boldsymbol{\phi}^f\} = \{\boldsymbol{\phi}^f_1, \ldots, \boldsymbol{\phi}^f_{100}\}$, generated from a Gaussian mixture model of complexity two:

$$p_{\boldsymbol{\Phi}^f}(\boldsymbol{\phi}^f) = \sum_{j=1}^{2} \pi_j \times \mathcal{N}(\boldsymbol{\phi}^f; \boldsymbol{\mu}^f_j, \boldsymbol{\Sigma}^f_j).$$

Let us further assume the following forecast parameters:

$$\pi^f_1 = 0.5, \qquad \boldsymbol{\mu}^f_1 = \begin{bmatrix}-10\\-1\end{bmatrix}, \qquad \boldsymbol{\Sigma}^f_1 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$$
$$\pi^f_2 = 0.5, \qquad \boldsymbol{\mu}^f_2 = \begin{bmatrix}10\\1\end{bmatrix}, \qquad \boldsymbol{\Sigma}^f_2 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}.$$

For simplicity, we will take the true field to coincide with one of the realizations, i.e.

$$\boldsymbol{x}^t = \bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}}\boldsymbol{\phi}^f_1.$$

We make noisy measurements of the first and third elements of the state vector, i.e.

$$\boldsymbol{H} = \begin{bmatrix}1 & 0 & 0\\0 & 0 & 1\end{bmatrix},$$

normally distributed with an error covariance matrix given by

$$\boldsymbol{R} = \sigma^2_{obs} \times \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix},$$

where $\sigma_{obs} = 5$. We illustrate all of the above in panel (a) of figure 3. With this, we proceed with the update step, using the GMM-DO flowchart, figure 2. We bypass illustrating the application of the BIC and rather present results directly for GMMs of complexity, M, one and two. The former is a single Gaussian parametric distribution while the latter would, with high probability, be obtained using the BIC criterion in the present example.

a. Fitting of GMM

i. Use the EM algorithm to obtain the prior mixture parameters

$$\pi^f_j, \boldsymbol{\mu}^f_j, \boldsymbol{\Sigma}^f_j, \qquad j = 1, \ldots, M$$

within the stochastic subspace based on the set of ensemble realizations, $\{\boldsymbol{\phi}^f\} = \{\boldsymbol{\phi}^f_1, \ldots, \boldsymbol{\phi}^f_{100}\}$. The identified mixtures (of complexities one and two), along with their marginal distributions, are displayed in panel b-(i) of figure 3.

b. Update

i. Calculate parameters:

$$\tilde{\boldsymbol{H}} \equiv \boldsymbol{H}\boldsymbol{\mathcal{X}}$$
$$\tilde{\boldsymbol{y}} \equiv \boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}^f$$

and determine the mixture Kalman gain matrices:

$$\tilde{\boldsymbol{K}}_j = \boldsymbol{\Sigma}^f_j \tilde{\boldsymbol{H}}^T (\tilde{\boldsymbol{H}} \boldsymbol{\Sigma}^f_j \tilde{\boldsymbol{H}}^T + \boldsymbol{R})^{-1}.$$

ii. Assimilate the measurements, $\boldsymbol{y}$, by calculating the 'intermediate' mixture means in the stochastic subspace,

$$\hat{\boldsymbol{\mu}}^a_j = \boldsymbol{\mu}^f_j + \tilde{\boldsymbol{K}}_j(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{H}}\boldsymbol{\mu}^f_j),$$

and further compute the posterior mixture weights:

$$\pi^a_j = \frac{\pi^f_j \times \mathcal{N}(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{H}}\boldsymbol{\mu}^f_j, \tilde{\boldsymbol{H}}\boldsymbol{\Sigma}^f_j\tilde{\boldsymbol{H}}^T + \boldsymbol{R})}{\sum_{m=1}^{M} \pi^f_m \times \mathcal{N}(\tilde{\boldsymbol{y}}; \tilde{\boldsymbol{H}}\boldsymbol{\mu}^f_m, \tilde{\boldsymbol{H}}\boldsymbol{\Sigma}^f_m\tilde{\boldsymbol{H}}^T + \boldsymbol{R})}.$$

iii. Update the DO mean field (displayed in panel c-(ii) of figure 3),

$$\bar{\boldsymbol{x}}^a = \bar{\boldsymbol{x}}^f + \boldsymbol{\mathcal{X}} \sum_{j=1}^{M} \pi^a_j \times \hat{\boldsymbol{\mu}}^a_j,$$

as well as the mixture parameters within the stochastic subspace:

$$\boldsymbol{\mu}^a_j = \hat{\boldsymbol{\mu}}^a_j - \sum_{j=1}^{M} \pi^a_j \times \hat{\boldsymbol{\mu}}^a_j$$
$$\boldsymbol{\Sigma}^a_j = (\boldsymbol{I} - \tilde{\boldsymbol{K}}_j\tilde{\boldsymbol{H}})\boldsymbol{\Sigma}^f_j.$$

iv. Generate the posterior set of ensemble realizations within the stochastic subspace, $\{\boldsymbol{\phi}^a\} = \{\boldsymbol{\phi}^a_1, \ldots, \boldsymbol{\phi}^a_{100}\}$, based on the multivariate GMM with posterior parameters

$$\pi^a_j, \boldsymbol{\mu}^a_j, \boldsymbol{\Sigma}^a_j, \qquad j = 1, \ldots, M.$$

We display the posterior set of realizations in panel c-(i) of figure 3.

By way of this simple example, we draw two conclusions on the benefits of the GMM-DO filter. Due to the initial non-Gaussian statistics, the GMM was expectedly found to provide a posterior estimate superior to that of the Gaussian parametric distribution (PD), as evidenced for example by their posterior means, panel c-(ii) of figure 3. In particular, due to the PD's conservative estimate for the covariance matrix of the true pdf (panel b-(i) of figure 3), the noisy measurements were inherently favored during the update step, essentially resulting in an 'overshoot' of its

posterior estimate for the mean. Given the GMM's accurate representation of the non-Gaussian features, on the other hand, the prior information was properly balanced with that due to the measurements, resulting in a successful Bayesian update. While this was to be expected given the initial bimodal distribution, previous arguments suggest that this holds for arbitrary distributions as long as the fitting of GMMs based on the EM algorithm and BIC provides a good approximation of the true pdf.

The second conclusion refers to the posterior statistics, represented by the subspace realizations, $\{\boldsymbol{\phi}^a\} = \{\boldsymbol{\phi}^a_1, \ldots, \boldsymbol{\phi}^a_{100}\}$, in panel c-(i) of figure 3. In addition to the GMM's successful capture of the true solution, the compactness of its posterior set of realizations further emphasized an added belief in this estimate. The accuracy of the posterior representation of the true statistics clearly affects future assimilations (not shown here, however). We therefore hypothesize that the GMM-DO filter outperforms simpler schemes, e.g. the Gaussian parametric distribution, in this respect. In part II of this two-part paper, we support this hypothesis by applying the GMM-DO filter in truly dynamical systems.

5. Discussions and Comparisons with Related Schemes

In this section, we review a selection of past pioneering DA schemes that, as the GMM-DO filter, have adopted the use of GMMs for approximating the true pdf.

a. Alspach and Sorenson (1972)

GMMs were, to the best of our knowledge, first addressed in the context of filtering theory by Alspach and Sorenson (1972). Here, the authors were particularly motivated by the inappropriate use of the Gaussian parametric distribution, stating that *"the Gaussian [parametric] approximation greatly reduces the amount of information that is contained in the true density, particularly when it is multimodal"*. They emphasized the ability of GMMs to approximate arbitrary pdfs, all the while retaining the familiar computational tractability when placed in the context of Bayesian inference.

Based on an approximation of the known, initial (non-Gaussian) distribution by a GMM of complexity M, their scheme would essentially run M extended Kalman filters in parallel – one for each mixture component – coupled solely through the mixture weights. Their update would thus take a form structurally similar to that of the GMM-DO filter, set aside the latter's focus on a stochastic subspace nonlinearly evolving through fully coupled DO equations. While the authors freed themselves of the Gaussian parametric constraint, their scheme remained grounded in linear theory, however, having been inspired by the Extended Kalman filter. The authors also made no mention of the appropriate mixture complexity, nor the manner in which the initial mixture parameters were obtained. Moreover, while they alluded to the need for intermittently restarting the distribution – either due to a poor mismatch of forecast distribution with observations, or to the collapse of weights onto a single mixture component – no appropriate remedies were proposed.

b. Anderson and Anderson (1999)

Anderson and Anderson (1999), in part inspired by the recent advances of ensemble methods within the DA community (e.g. Evensen 1994; Lermusiaux 1997; Houtekamer et al. 1998), extended the work of Alspach and Sorenson by adopting a Monte Carlo approach for evolving the probabilistic description of the state in time. By arguing that *"one of the fundamental advantages of a Monte Carlo approach [is its] ability to represent non-Gaussian probability distributions"*, they chose to approximate the Monte Carlo realizations by use of a kernel density estimator,

$$p_{\boldsymbol{X}^f}(\boldsymbol{x}^f) = \sum_{i=1}^{N} \frac{1}{N} \times \mathcal{N}(\boldsymbol{x}^f; \boldsymbol{x}^f_i, \alpha \boldsymbol{\Sigma}^f), \tag{83}$$

with $\boldsymbol{x}_i$ representing realizations in state space; $\boldsymbol{\Sigma}$ the sample covariance matrix based on the set of ensemble realizations; and α an heuristically chosen scaling parameter.

Upon assimilating data from a Gaussian observation model, their posterior distribution for the state vector would thus take the familiar form

$$p_{\boldsymbol{X}^a}(\boldsymbol{x}^a) = \sum_{i=1}^{N} \pi^a_i \times \mathcal{N}(\boldsymbol{x}^a; \boldsymbol{x}^a_i, \alpha \boldsymbol{\Sigma}^a), \tag{84}$$

with parameters determined in accordance with (57)–(59), from which they would draw N new realizations.

The authors justifiably argued for the advantages over filters invoking the Gaussian parametric distribution, giving as example their respective performances when applied to the three-dimensional Lorenz-63 model (Lorenz 1963): while their kernel filter would represent states solely in accordance with model dynamics, simpler filters would potentially assign finite probability to regions of state space never visited.

One drawback of the filter lay in their arguments for choosing the scaling parameter, α. Specifically, the authors stated that while *"a number of methods for computing the constant covariance reduction factor, α, have been developed, ... the value of α is often subsumed into a tuning constant and so does not need to be calculated explicitly. ... Tuning a filter for a real system is complicated ... [and] must be chosen with care"*.

Hoteit et al. (2008) later extended the filter by allowing the realizations to carry uneven weights, drawing on the concepts of particle filters. Specifically, they retained the posterior form of equation (84) rather than drawing N new realizations following every assimilation step. To avoid

the collapse of weights onto only a few realizations, they proposed a number of interesting methods for resampling. While effective, these ideas are not discussed further.

c. Bengtsson et al. (2003)

Bengtsson et al. (2003) expressed a concern over Anderson and Anderson's use of kernel density methods for approximating distributions, arguing that the use of *"scaled versions of the full ensemble covariance around each center in the mixture ... cannot adapt as easily to local structure in the forecast distribution".* Instead, they proposed to approximate the set of realizations by a GMM (of complexity less than the number of realizations), estimating the mixture parameters using local knowledge of the ensemble distribution. They stated that such an approach would provide a more accurate approximation to the true pdf.

Their update step essentially proceeded as follows: M ensemble realizations would be *arbitrarily* chosen to act as means for the proposed Gaussian mixtures, from which N_n nearest neighbors to each of these realizations would be used to approximate their respective mixture covariance matrices. From here, one would proceed with the Bayesian update, conceptually inspired by the Ensemble Kalman filter (Evensen 1994).

As with Alspach and Sorensen, the authors left unanswered methods for determining both the mixture complexity, M, as well as the appropriate choice of N_n, the number of nearest neighbors. Furthermore, their choice of mixture means, based on the arbitrary sampling of ensemble realizations, would certainly invite for sampling noise.

The authors further expressed difficulties associated with manipulating pdfs in high dimensional spaces. They thus introduced a hierarchy of adaptations to the aforementioned filter in which they invoked varying degrees of localization approximations, all based on heuristic arguments. As a remedy, however, they concluded that *"a more sophisticated filter will likely rely on efficient, sequential identification of low-dimensional subspaces where non-Gaussian densities can be accurately represented and filtered using finite ensemble sizes".*

d. Smith (2007)

Indirectly extending the work by Bengtsson et al., Smith (2007) employed the EM algorithm to uncover the underlying structure represented by the set of ensemble realizations, thus alleviating former heuristic arguments. The author modified the Ensemble Kalman Filter to allow for a Gaussian mixture representation of the prior distribution, using Akaike's Information Criterion (AIC) as the method for selecting the appropriate mixture complexity. (As a side note, McLachlan and Peel (2000) found the BIC to outperform the AIC when fitting Gaussian mixtures to data; specifically, the latter would have the tendency to overestimate the mixture complexity.) Similar to the scheme of Bengtsson et al., Smith retained the concept of operating on individual ensemble realizations during the update step, imposing only – but somewhat surprisingly – that the posterior distribution be normally distributed.

For illustration, the author applied his Cluster Ensemble Kalman filter to a two-dimensional phytoplankton-zooplankton biological model. While successful for such simple models, he emphasized the difficulties of extending his scheme to test cases of larger dimensions, making, however, the useful comment that *"the state space could be projected onto a lower dimensional space depicting some relevant phenomenon, and the full covariance matrix in this state space could be used."*

e. Dovera and Rossa (2010)

Dovera and Rossa (2010) would later modify the approach by Smith, attempting to overcome the constraint that the posterior distribution be Gaussian. Their update step seemingly disagreed with the output of the EM algorithm, however – a point of view reflected in the recent work by Frei and Kunsch (2011).

The authors applied their scheme to both the Lorenz-63 model as well as a two-dimensional reservoir model, outperforming the regular Ensemble Kalman filter. As with previous schemes, however, they equally noted the problems caused by systems of high dimensionality, again using a number of localization arguments to overcome this burden. With the GMM-DO filter, all of these issues are addressed by: (i) adopting the generalized, time-dependent Karhunen-Loeve decomposition of the state dictated by the DO framework; and (ii) deriving the corresponding rigorous GMM-DO updates for fully Bayesian-based data assimilation.

6. Summary and Conclusions

A data assimilation framework that rigorously utilizes the governing dynamical equations with information theory and learning theory for efficient Bayesian geophysical data assimilation was presented. The theory and algorithm of the resulting filter, the GMM-DO filter, were developed and derived. The DO equations and their adaptive stochastic subspace are employed to provide prior probabilities, effectively approximating the Fokker-Planck equation. At assimilation times, the DO realizations are fit to semiparametric GMMs using the Expectation-Maximization (EM) algorithm and the Bayesian Information Criterion (BIC). Bayes' Law is then efficiently carried out analytically within the evolving stochastic subspace.

Past literature had identified the advantages of adopting GMMs in a filtering setting, allowing the update step to capture and retain potential non-Gaussian features. In

some cases, the EM algorithm and model selection criteria had been used to obtain optimal mixture parameter values, resulting in a more accurate approximation of the true pdf. However, existing schemes often reverted to heuristic approximations or surprising choices. A novelty of the GMM-DO filter lies in its rigorous coupling of – GMMs, the EM algorithm and the BIC – with the efficient DO equations. By focusing on the time-dependent dominant stochastic subspace of the state space, we address prior limitations caused by the dimensionality of geophysical applications. Particularly, we render obsolete ad hoc procedures. Contrary to the Ensemble Kalman filter, as well as several other methods, we presently refrain from operating directly on individual ensemble realizations during the update step. Rather, under the assumption that the fitted GMM accurately captures the true prior pdf, we analytically carry out Bayes' Law efficiently within the stochastic subspace.

The derived GMM-DO filter respects nonlinear dynamics and captures non-Gaussian statistics as it occurs, obviating the use of empirical arguments. Of course, variations of the present filter exist, two of which are derived in Appendix B. Additional areas for further research include the selection of the algorithms for fitting the GMMs to the DO realizations. Schemes based on the EM-BIC approach have the advantage of being generic, but there is a large body of literature on other estimators (McLachlan and Peel 2000), and some schemes could be tailored to specific oceanic or atmospheric applications. Constraints can also be added to this fitting procedure, leading to a supervised learning of the GMM properties. Other mixture models could be used, e.g. including Laplace mixtures for heavier tails, depending of the application and efficiency requirements. One advantage of the GMM is that if the number of Gaussians is one ($M = 1$), one recovers a classic Kalman update. Since our GMM-DO filter estimates the optimal M, if it is found to be one, a Kalman update in the subspace is used. The GMM-DO filter is thus a straightforward and efficient extension of the Kalman filter for nonlinear and non-Gaussian geophysical systems. The present GMM-DO update could also be augmented with a subspace learning scheme based on the innovation vector and posterior misfit, extending the ESSE learning to GMMs. Another variation of this update is to operate directly on individual realizations; such a variation exist in ESSE. Another research direction is the derivation of GMM-DO smoothers. A possibility is to employ a statistical linearization as in the ESSE smoother (Lermusiaux and Robinson 1999; Lermusiaux et al. 2002b), but other options are possible, including hybrid ones with variational schemes (e.g. Moore et al. 2004). Finally, for the case of white-noise stochastic forcing and for small enough stochastic subspace size, the Fokker-Planck equation that evolves the joint pdf for the stochastic coefficients of the DO expansion (Sapsis and Lermusiaux 2011) could be used instead of the stochastic differential equations for DO realizations. This approach would directly provide the prior joint pdf for the Bayesian update, but numerical schemes other than those employed here would then be needed (Ueckermann et al. 2012).

In part I of this two-part paper, we derived the GMM-DO filter, outlined its algorithmic implementation, and placed it in the context of current literature. In part II, we evaluate its performance when applied to the following test cases: (i) the Double-Well Diffusion Experiment and (ii) the Sudden Expansion fluid flow.

Acknowledgments.

We are very thankful to the MSEAS group members, in particular for helpful discussions with Mr. M.P. Ueckermann and Dr. T. Sapsis, as well as with Mr. T. Lolla and P. Lu. We especially thank Mr. M. Ueckermann for his help on the DO numerics. TS also acknowledges Prof. G. Wornell and his MIT course for a clear introduction to a number of the information theory concepts relevant to the present research. We are grateful to the Office of Naval Research for support under grants N00014-08-1-1097 (ONR6.1), N00014-09-1-0676 (Science of Autonomy – A-MISSION) and N00014-08-1-0586 (QPE). PFJL is also thankful to Sea Grant at MIT for the "2009 Doherty Professorship in Ocean Utilization" award.

APPENDIX A
EM algorithm (with Gaussian mixture models)

The EM algorithm is commonly introduced in the context of 'incomplete data' (Dempster et al. 1977), for which ML parameter estimation by partial differentiation, eqn. (9), fails to yield a closed form solution. To circumvent this, the main idea is to artificially 'complete' the data at hand with additional pseudo data (or knowledge about the data), thereby giving rise to closed form solutions for the ML parameters (McLachlan and Peel 2000). The data with which to complete the existing data set is chosen by the user and may have little physical relevance; its choice, however, ultimately dictates the efficiency of the algorithm. By conditioning the complete data on the available data, an improved estimate for the ML parameters is iteratively obtained. This procedure lies at the heart of the EM algorithm.

For the case of GMMs, we augment the available data set, represented by the set of ensemble realizations, $\{\boldsymbol{x}\} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, to form the *complete* data set,

$$\{\boldsymbol{z}\} = \{\boldsymbol{c}_1, \boldsymbol{x}_1, \ldots, \boldsymbol{c}_N, \boldsymbol{x}_N\}, \tag{A1}$$

where $\boldsymbol{c}_i$ represents an indicator vector of length M such that

$$(\boldsymbol{c}_i)_j = \begin{cases} 1 & \text{if realization } \boldsymbol{x}_i \text{ was generated by mixture component } j \\ 0 & \text{otherwise,} \end{cases}$$

with $(\boldsymbol{c}_i)_j$ referring to the j^{th} element of vector $\boldsymbol{c}_i$. (Here, these membership indicators have little physical relevance, and exist merely as a conceptual device within the EM framework.) Conditioned on the additional knowledge of the set $\{\boldsymbol{c}\} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N\}$, we assume known the origin of each realization, namely the mixture component that generated it. This knowledge gives rise to closed form solutions for the ML estimator of the parameter vector, specifically:

$$\pi_j = \frac{N_j}{N} \tag{A2}$$

$$\bar{\boldsymbol{x}}_j = \frac{1}{N_j}\sum_{i=1}^{N}(\boldsymbol{c}_i)_j \times \boldsymbol{x}_i \tag{A3}$$

$$\boldsymbol{P}_j = \frac{1}{N_j}\sum_{i=1}^{N}(\boldsymbol{c}_i)_j \times (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j)^T, \tag{A4}$$

where

$$N_j \equiv \sum_{i=1}^{N}(\boldsymbol{c}_i)_j. \tag{A5}$$

With the addition of the data set $\{\boldsymbol{c}\} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N\}$, we have thus *completed* the data vector (i.e. in some sense, we pretend that we know which mixture component generated each realization, so as to get the EM iterations started). In the real EM algorithm, however, a realization is not hard-wired to a particular mixture component, as done above. Rather, the algorithm iteratively estimates the weights with which a given realization is associated with each of the M mixture components.

In what follows, to avoid lengthy expressions, we neglect random variable subscripts when describing pdfs with the understanding that their arguments are realizations of this random variable. For instance, for the pdfs

$$p\big(x;\boldsymbol{\theta}\big) \equiv p_X\big(x;\boldsymbol{\theta}\big), \tag{A6}$$

x is the realization of random variable X.

a. Derivation of EM algorithm

We let $\{x\} = \{x_1, \cdots, x_N\}$ denote the set of available data, $\{z\}$ the *complete* data vector and $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_M\}$ the set of parameters (to be determined) of the *chosen* distributional form, $p\big(\{z\};\boldsymbol{\theta}\big)$. We further assume, as is often the case, that the available data is a unique and deterministic function of the complete data, i.e. $\{x\} = g(\{z\})$. (For instance, this may simply be a subset of the complete data.) By the Total Probability Theorem (e.g. Bertsekas and Tsitsiklis (2008)), we may thus write:

$$p\big(\{z\};\boldsymbol{\theta}\big) = \sum_{\{x\}} p\big(\{z\}\big|\{x\};\boldsymbol{\theta}\big) \times p\big(\{x\};\boldsymbol{\theta}\big) \tag{A7}$$

$$= p\big(\{z\}\big|g(\{z\});\boldsymbol{\theta}\big) \times p\big(g(\{z\});\boldsymbol{\theta}\big). \tag{A8}$$

By taking logarithms, we consequently obtain for *any value of* $\{z\}$ *that satisfies* $\{x\} = g(\{z\})$:

$$\begin{aligned}\log\big(p\big(\{x\};\boldsymbol{\theta}\big)\big) = \log\big(p\big(\{z\};\boldsymbol{\theta}\big)\big) \\ - \log\big(p\big(\{z\}\big|\{x\};\boldsymbol{\theta}\big)\big).\end{aligned} \tag{A9}$$

By further taking expectations with respect to the complete data, conditioned on the available data and parametrized by an *arbitrary* vector $\tilde{\boldsymbol{\theta}}$ (to be optimized), i.e.

$$E\left[(\bullet) \mid \{x\};\tilde{\boldsymbol{\theta}}\right] = \int_{\{z\}} (\bullet) p\big(\{z\}\big|\{x\};\tilde{\boldsymbol{\theta}}\big) d\{z\}, \tag{A10}$$

the left hand side of equation (A9) remains unaffected,

$$E\left[\log\big(p\big(\{x\};\boldsymbol{\theta}\big)\big) \mid \{x\};\tilde{\boldsymbol{\theta}}\right] = \log\big(p\big(\{x\};\boldsymbol{\theta}\big)\big), \tag{A11}$$

and we thus obtain

$$\begin{aligned}\log\big(p\big(\{x\};\boldsymbol{\theta}\big)\big) = E\left[\log\big(p\big(\{z\};\boldsymbol{\theta}\big)\big) \mid \{x\};\tilde{\boldsymbol{\theta}}\right] \\ - E\left[\log\big(p\big(\{z\}\big|\{x\};\boldsymbol{\theta}\big)\big) \mid \{x\};\tilde{\boldsymbol{\theta}}\right].\end{aligned} \tag{A12}$$

For the sake of convenience, we define the notation

$$U(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) = E\left[\log\big(p\big(\{z\};\boldsymbol{\theta}\big)\big) \mid \{x\};\tilde{\boldsymbol{\theta}}\right] \tag{A13}$$

$$V(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) = -E\left[\log\big(p\big(\{z\}\big|\{x\};\boldsymbol{\theta}\big)\big) \mid \{x\};\tilde{\boldsymbol{\theta}}\right] \tag{A14}$$

to obtain the simplified expression

$$\log\big(p\big(\{x\};\boldsymbol{\theta}\big)\big) = U(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) + V(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}). \tag{A15}$$

By application of Gibbs' inequality (MacKay 2003), we see that

$$V(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) = -E\left[\log\big(p\big(\{z\}\big|\{x\};\boldsymbol{\theta}\big)\big) \mid \{x\};\tilde{\boldsymbol{\theta}}\right] \tag{A16}$$

$$\geq -E\left[\log\big(p\big(\{z\}\big|\{x\};\tilde{\boldsymbol{\theta}}\big)\big) \mid \{x\};\tilde{\boldsymbol{\theta}}\right] \tag{A17}$$

$$= V(\tilde{\boldsymbol{\theta}};\tilde{\boldsymbol{\theta}}). \tag{A18}$$

Therefore, if we denote $\tilde{\boldsymbol{\theta}}$ as our *present* estimate for the parameter vector, by *choosing* $\boldsymbol{\theta} \neq \tilde{\boldsymbol{\theta}}$ such that it further satisfies $U(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) \geq U(\tilde{\boldsymbol{\theta}};\tilde{\boldsymbol{\theta}})$, we guarantee that

$$\log\big(p\big(\{x\};\boldsymbol{\theta}\big)\big) = U(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) + V(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) \tag{A19}$$

$$\geq U(\tilde{\boldsymbol{\theta}};\tilde{\boldsymbol{\theta}}) + V(\tilde{\boldsymbol{\theta}};\tilde{\boldsymbol{\theta}}) \tag{A20}$$

$$= \log\big(p\big(\{x\};\tilde{\boldsymbol{\theta}}\big)\big). \tag{A21}$$

Consequently, upon repeated iterations, our estimate for the parameter vector monotonically increases the (log) likelihood of generating the data at hand, $\{x\} = \{x_1, \cdots, x_N\}$. Assuming further that the likelihood is bounded from above, we are thus guaranteed to converge to a stationary point and as such obtain an estimate for the ML parameter vector (Casella and Berger 2001). In summary, the EM algorithm proceeds as follows.

EM algorithm: *Given the available data,* $\{x\} = \{x_1, \ldots, x_N\}$, *initial parameter estimate,* $\boldsymbol{\theta}^{(0)}$, *proposed complete data vector* $\{z\}$ *with predetermined, user-specified distribution,* $p\big(\{z\};\boldsymbol{\theta}\big)$, *repeat until convergence:*

- *Using the present parameter estimate $\boldsymbol{\theta}^{(k)}$, form*

$$U\big(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\big) = E\left[\log\big(p\big(\{z\};\boldsymbol{\theta}\big)\big) \mid \{x\};\boldsymbol{\theta}^{(k)}\right]. \quad \text{(A22)}$$

- *Update the estimate for the parameter vector, $\boldsymbol{\theta}^{(k+1)}$, by maximizing $U\big(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\big)$:*

$$\boldsymbol{\theta}^{(k+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left(U(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})\right). \quad \text{(A23)}$$

Next, we apply the EM algorithm to multivariate GMMs. We provide the derivation in a condensed manner; we refer to Sondergaard (2011) for full details.

b. The EM algorithm with Gaussian mixture models (GMMs)

We augment the available data set, $\{\boldsymbol{x}\} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$, generated by a GMM of unknown parameters,

$$\boldsymbol{\theta} = \{\pi_1,\ldots,\pi_M,\bar{\boldsymbol{x}}_1,\ldots,\bar{\boldsymbol{x}}_M,\boldsymbol{P}_1,\ldots,\boldsymbol{P}_M\}, \quad \text{(A24)}$$

to form the *complete* data set

$$\{\boldsymbol{z}\} = \{\boldsymbol{c}_1,\boldsymbol{x}_1,\ldots,\boldsymbol{c}_N,\boldsymbol{x}_N\}, \quad \text{(A25)}$$

as described in equation (A1).

By the assumed independence of the data, the probability distribution for the complete data takes the form

$$p\big(\{\boldsymbol{z}\};\boldsymbol{\theta}\big) = \prod_{i=1}^{N} p\big(\boldsymbol{c}_i,\boldsymbol{x}_i;\boldsymbol{\theta}\big) \quad \text{(A26)}$$

$$= \prod_{i=1}^{N}\prod_{j=1}^{M}\left(\pi_j \times \mathcal{N}\left(\boldsymbol{x}_i;\bar{\boldsymbol{x}}_j,\boldsymbol{P}_j\right)\right)^{(\boldsymbol{c}_i)_j}. \quad \text{(A27)}$$

Upon taking logarithms we obtain

$$\log\left(p\big(\{\boldsymbol{z}\};\boldsymbol{\theta}\big)\right) = \sum_{i=1}^{N}\sum_{j=1}^{M}(\boldsymbol{c}_i)_j \times \big(\log\pi_j + \log\mathcal{N}\left(\boldsymbol{x}_i;\bar{\boldsymbol{x}}_j,\boldsymbol{P}_j\right)\big). \quad \text{(A28)}$$

By further taking the conditional expectation of equation (A28) with respect to the available data, arbitrarily parametrized by vector $\boldsymbol{\theta}^{(k)}$, we consequently obtain the expression to be maximized under the EM algorithm:

$$U(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) = E\left[\log\big(p\big(\{\boldsymbol{z}\};\boldsymbol{\theta}\big)\big) \mid \{\boldsymbol{x}\};\boldsymbol{\theta}^{(k)}\right] \quad \text{(A29)}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} E\left[(\boldsymbol{c}_i)_j \mid \{\boldsymbol{x}\};\boldsymbol{\theta}^{(k)}\right] \times \big(\log\pi_j + \log\mathcal{N}\big(\boldsymbol{x}_i;\bar{\boldsymbol{x}}_j,\boldsymbol{P}_j\big)\big). \quad \text{(A30)}$$

For convenience of notation, we define:

$$\tau_j\big(\boldsymbol{x}_i;\boldsymbol{\theta}^{(k)}\big) \equiv E\left[(\boldsymbol{c}_i)_j \mid \{\boldsymbol{x}\};\boldsymbol{\theta}^{(k)}\right] \quad \text{(A31)}$$

$$= \frac{\pi_j^{(k)} \times \mathcal{N}\big(\boldsymbol{x}_i;\bar{\boldsymbol{x}}_j^{(k)},\boldsymbol{P}_j^{(k)}\big)}{\sum_{m=1}^{M}\pi_m^{(k)} \times \mathcal{N}\big(\boldsymbol{x}_i;\bar{\boldsymbol{x}}_m^{(k)},\boldsymbol{P}_m^{(k)}\big)}. \quad \text{(A32)}$$

This completes the E-step of the EM algorithm, equation (A22)

We proceed with evaluating $\boldsymbol{\theta}^{(k+1)}$, the parameter vector, $\boldsymbol{\theta}$, which maximizes $U(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$. This forms the M-step of the EM algorithm, equation (A23). To determine the updated mixture weights, $\pi_j^{(k+1)}$, we augment $U(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ using Lagrange multipliers and so introduce the auxiliary function, Λ, with multiplier, λ:

$$\Lambda = \sum_{j=1}^{M}\sum_{i=1}^{N}\tau_j(\boldsymbol{x}_i;\boldsymbol{\theta}^{(k)}) \times \Big(\log\pi_j - \frac{n}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{P}_j| - \frac{1}{2}\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j\right)^T\boldsymbol{P}_j^{-1}\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j\right)\Big) + \lambda \times \Big(\sum_{k=1}^{M}\pi_k - 1\Big), \quad \text{(A33)}$$

By equating to zero the gradients of Λ with respect to π_p and λ, we obtain after manipulations the final expression:

$$\pi_p^{(k+1)} = \frac{\sum_{i=1}^{N}\tau_p\big(\boldsymbol{x}_i;\boldsymbol{\theta}^{(k)}\big)}{N} \equiv \frac{N_p^{(k)}}{N}, \quad \text{(A34)}$$

where $N_p^{(k)}$ is the sum total of particles associated with a given mixture component, p, under the present estimate for the parameter vector, $\boldsymbol{\theta}^{(k)}$. With this, we proceed to determine the unconstrained parameters, $\bar{\boldsymbol{x}}_p^{(k+1)}$ and $\boldsymbol{P}_p^{(k+1)}$. To obtain the updated mixture mean vectors, $\bar{\boldsymbol{x}}_p^{(k+1)}$, we equate the appropriate partial derivative of Λ with zero,

$$\frac{\partial\Lambda}{\partial\bar{\boldsymbol{x}}_p} = \boldsymbol{0} \quad \text{(A35)}$$

to obtain

$$\bar{\boldsymbol{x}}_p^{(k+1)} = \frac{1}{N_p^{(k)}}\sum_{i=1}^{N}\tau_p(\boldsymbol{x}_i;\boldsymbol{\theta}^{(k)}) \times \boldsymbol{x}_i. \quad \text{(A36)}$$

Similarly, to obtain the updated mixture covariance matrices, $\boldsymbol{P}_p^{(k+1)}$, we enforce (with knowledge of $\bar{\boldsymbol{x}}_p^{(k+1)}$)

$$\frac{\partial\Lambda}{\partial\boldsymbol{P}_p} = \boldsymbol{0} \quad \text{(A37)}$$

to ultimately arrive at

$$\boldsymbol{P}_p^{(k+1)} = \frac{1}{N_p^{(k)}}\sum_{i=1}^{N}\tau_p(\boldsymbol{x}_i;\boldsymbol{\theta}^{(k)}) \times (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_p^{(k+1)})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_p^{(k+1)})^T. \quad \text{(A38)}$$

This completes the condensed derivation of the EM algorithm as applied to GMMs. The algorithm is summarized in the main body of the text, equations (10)–(13). For additional remarks on the EM algorithm and its application to GMMs, including the choice of starting parameters and the issue of convergence, we refer to Sondergaard (2011).

APPENDIX B
Variations of the GMM-DO filter

a. EM algorithm in q-dominant space of stochastic subspace

Estimating and manipulating non-trivial pdfs in high-dimensional spaces can be a difficult task (Bengtsson et al. 2003). Heuristic arguments suggest that the number of realizations required to accurately represent multivariate pdfs grows exponentially with the dimension of the space (Silverman 1992). This is one of the reason why we investigate approximations to our main scheme that would allow efficient fitting of GMMs to realizations when the dimension of the stochastic subspace itself is large and may pose a difficulty. Another reason arises from oceanic and atmospheric applications. In such applications, the variance of the ESSE or DO modes is often found to decay rapidly with mode number, e.g. (Lermusiaux 1999a,b, 2001, 2007; Sapsis and Lermusiaux 2011). In addition, the accuracy of the low variance modes is not as good as that of the large variance modes: this is mainly because of their much smaller variance and of their proximity to the truncation index and thus un-modeled interactions with the truncated modes. As a result, trying to fit all structures of the marginal probabilities for these low variance modes is likely not needed and can in fact reduce the robustness in the Bayesian inversion. Finally, it reduces the computational cost.

As in the main text, we let the dimension of the stochastic subspace be s, i.e. $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n\times s}$. When deemed necessary on the grounds of tractability and mode variance decay, we can limit our estimation of mixtures to the stochastic coefficients associated with the space defined by the q most dominant modes, denoting this $\boldsymbol{\mathcal{X}}^q \in \mathbb{R}^{n\times q}$. We in turn approximate the stochastic coefficients of the remaining $s-q$ modes, $\{\Phi_{q+1},\ldots,\Phi_s\}$, as zero mean Gaussian with (co)variances based on the sample covariance matrix. For our purposes, an obvious and appropriate measure of dominance is the variance of each of the stochastic coefficients.

Next, we define this modified EM algorithm for GMMs in a q-dominant space.

EM algorithm in q-dominant space of stochastic subspace: *Given the set of realizations, $\{\boldsymbol{\phi}\} \in \mathbb{R}^{s\times N}$, associated with the stochastic subspace, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n\times s}$, we limit our attention to the ensemble set, $\{\boldsymbol{\phi}^q\} \in \mathbb{R}^{q\times N}$, associated with the q-dominant reduced space, $\boldsymbol{\mathcal{X}}^q \in \mathbb{R}^{n\times q}$, of the stochastic subspace (i.e. $q \le s$). We define q such that the following holds:*

$$1 \ge \frac{\sum_{i=1}^{q} var(\Phi_i)}{\sum_{j=1}^{s} var(\Phi_j)} \ge C \ge 0, \tag{B1}$$

where C denotes a user-specified constant chosen such that the majority of the energy in the stochastic subspace is captured. (Note, we assume that the stochastic coefficients, Φ_i, are ordered by decreasing variance, i.e. $var(\Phi_1) \ge var(\Phi_2) \ge \ldots \ge var(\Phi_s)$. Other ratios are also possible, e.g. (Lermusiaux 2007; Sapsis and Lermusiaux 2011).)

Based on the reduced ensemble set, $\{\boldsymbol{\phi}^q\} = \{\boldsymbol{\phi}_1^q,\ldots,\boldsymbol{\phi}_N^q\}$, and initial parameter estimate,

$$\boldsymbol{\theta}^{q,(0)} = \{\pi_1^{q,(0)},\ldots,\pi_M^{q,(0)},\bar{\boldsymbol{x}}_1^{q,(0)},\ldots,\bar{\boldsymbol{x}}_M^{q,(0)},\boldsymbol{P}_1^{q,(0)},\ldots,\boldsymbol{P}_M^{q,(0)}\},$$

appropriately sized for the reduced EM estimation procedure, we repeat until convergence:

- *For all $i \in \{1,\ldots,N\}$ and $j \in \{1,\ldots,M\}$, use the present parameter estimate, $\boldsymbol{\theta}^{q,(k)}$, to form*

$$\tau_j\big(\boldsymbol{\phi}_i^q;\boldsymbol{\theta}^{q,(k)}\big) = \frac{\pi_j^{q,(k)} \times \mathcal{N}\big(\boldsymbol{\phi}_i^q;\boldsymbol{\mu}_j^{q,(k)},\boldsymbol{\Sigma}_j^{q,(k)}\big)}{\sum_{m=1}^{M} \pi_m^{q,(k)} \times \mathcal{N}\big(\boldsymbol{\phi}_i^q;\boldsymbol{\mu}_m^{q,(k)},\boldsymbol{\Sigma}_m^{q,(k)}\big)}. \tag{B2}$$

- *For all $j \in \{1,\ldots,M\}$, update the parameter estimate, $\boldsymbol{\theta}^{q,(k+1)}$, according to*

$$\pi_j^{q,(k+1)} = \frac{N_j^{q,(k)}}{N} \tag{B3}$$

$$\boldsymbol{\mu}_j^{q,(k+1)} = \frac{1}{N_j^{q,(k)}} \sum_{i=1}^{N} \tau_j\big(\boldsymbol{\phi}_i^q;\boldsymbol{\theta}^{q,(k)}\big) \times \boldsymbol{\phi}_i^q \tag{B4}$$

$$\boldsymbol{\Sigma}_j^{q,(k+1)} = \frac{1}{N_j^{q,(k)}} \sum_{i=1}^{N} \tau_j\big(\boldsymbol{\phi}_i^q;\boldsymbol{\theta}^{q,(k)}\big) \times \big(\boldsymbol{\phi}_i^q - \boldsymbol{\mu}_j^{q,(k+1)}\big)\big(\boldsymbol{\phi}_i^q - \boldsymbol{\mu}_j^q \tag{B5}$$

where

$$N_j^{q,(k)} = \sum_{i=1}^{N} \tau_j\big(\boldsymbol{\phi}_i^q;\boldsymbol{\theta}^{q,(k)}\big). \tag{B6}$$

Once converged, we obtain the GMM associated with the stochastic subspace, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n\times s}$, by embedding the above q-dominant vectors and matrices into their adequately sized equivalent:

$$\boldsymbol{\mu}_j = \begin{bmatrix} \boldsymbol{\mu}_j^q \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{0} \in \mathbb{R}^{s-q} \tag{B7}$$

and

$$\boldsymbol{\Sigma}_j = \begin{bmatrix} \boldsymbol{\Sigma}_j^q & \boldsymbol{\Sigma}_{1:q,(q+1):s} \\ \boldsymbol{\Sigma}_{(q+1):s,1:q} & \boldsymbol{\Sigma}_{(q+1):s,(q+1):s} \end{bmatrix}, \tag{B8}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{s\times s}$ is the sample covariance matrix,

$$\boldsymbol{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} \boldsymbol{\phi}\boldsymbol{\phi}^T, \tag{B9}$$

and $\boldsymbol{\Sigma}_{a:b,c:d}$ denotes the sub-matrix of $\boldsymbol{\Sigma}$ defined by rows a-b and columns c-d.

In the above, we arrive at equations (B7) and (B8) by application of the Law of Iterated Expectations and the

Law of Total Variance, respectively (e.g. Bertsekas and Tsitsiklis 2008), ensuring that the stochastic coefficients, $\{\Phi_{q+1}, \ldots, \Phi_s\}$, are approximated as zero mean Gaussian distributions with variances based on the sample covariance matrix.

b. EM algorithm with a constrained mean for the Gaussian mixture model

In the DO decomposition (24), we impose a zero-mean constraint on the random vector, $\boldsymbol{\Phi}(\omega)$, represented by the ensemble set, $\{\boldsymbol{\phi}\} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N\}$. Since the EM algorithm is an unconstrained optimization procedure in this regard, however, the EM fit of the GMM may not necessarily itself be of zero mean, i.e.

$$\sum_{j=1}^{M} \pi_j \times \boldsymbol{\mu}_j \neq 0. \tag{B10}$$

While the test cases presented in part II of this two-part paper give evidence to suggest that this is little cause for concern (namely that this mean offset is negligible and tends to zero as N increases), we nonetheless propose two possible remedies:

i. When forming the auxiliary function in equation (A33), one may add the constraint that the GMM be of zero mean, i.e.

$$\sum_{j=1}^{M} \pi_j \times \boldsymbol{\mu}_j = 0, \tag{B11}$$

thus updating the auxiliary function (in the stochastic subspace) to:

$$\begin{aligned}\Lambda = &\sum_{j=1}^{M}\sum_{i=1}^{N} \tau_j(\boldsymbol{\phi}_i;\boldsymbol{\theta}^{(k)}) \times \big(\log \pi_j - \frac{s}{2}\log 2\pi \\ &- \frac{1}{2}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}\left(\boldsymbol{\phi}_i - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}_j^{-1}\left(\boldsymbol{\phi}_i - \boldsymbol{\mu}_j\right)\big) \\ &+ \lambda_1 \times (\sum_{k=1}^{M} \pi_k - 1) + \lambda_2 \times \sum_{l=1}^{M} \pi_l \times \boldsymbol{\mu}_l.\end{aligned} \tag{B12}$$

While this clearly provides a viable solution, a closer inspection reveals that such a constraint destroys the simplicity of the EM algorithm. Particularly, the closed form equations (11)–(13) for the updated mixture parameters then no longer arise. Rather, the GMM parameters to be optimized become intimately coupled.

ii. A complementary approach first estimates the parameter vector by means of our regular EM algorithm for GMMs. This estimate is then in turn fed as a first guess to the coupled set of equations obtained in i) above, for which an iteration procedure of choice may be utilized. Since based on experience we know that the first guess is good for N large enough, we expect that only a few iterations are needed to converge to an optimal set of parameter values satisfying the additional zero mean constraint.

REFERENCES

Alspach, D. L. and H. W. Sorenson, 1972: Nonlinear bayesian estimation using gaussian sum approximations. *IEEE Transactions on Automatic Control*, **AC-17 (4)**, 438–448.

Anderson, J. L. and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, **127 (12)**, 2741–2758.

Auclair, F., P. Marsaleix, and P. D. Mey, 2003: Space-time structure and dynamics of the forecast error in a coastal circulation model of the Gulf of Lions. *Dynamics of Atmospheres and Oceans*, **36**, 309–346.

Bengtsson, T., C. Snyder, and D. Nychka, 2003: Toward a nonlinear ensemble filter for high-dimensional systems. *Journal of Geophysical Research-Atmospheres*, **108 (D24)**.

Bennett, A., 1992: *Inverse Methods in Physical Oceanography*. Cambridge University Press.

Bennett, A., 2002: *Inverse Modeling of the Ocean and Atmosphere*. Cambridge University Press.

Bertsekas, D. P. and J. N. Tsitsiklis, 2008: *Introduction to Probability*. 2d ed., Athena Scientific.

Bishop, C. M., 2006: *Pattern Recognition and Machine Learning*. Springer, 738 pp.

Bocquet, M., C. A. Pires, and L. Wu, 2010: Beyond Gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review*, **138**, 2997–3023.

Casella, G. and R. L. Berger, 2001: *Statistical Inference*. Duxbury.

Chen, R. and J. S. Liu, 2000: Mixture Kalman filters. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **62**, 493–508.

Cover, T. M. and J. A. Thomas, 2006: *Elements of information theory*. Wiley-Interscience, New York, NY, USA.

CPSMA, 1993: *Statistics and Physical Oceanography*. The National Academies Press, 62 pp.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press.

Dee, D. P. and A. M. D. Silva, 2003: The choice of variable for atmospheric moisture analysis. *Monthly Weather Review*, **131**, 155–171.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39 (1)**, 1–38.

Dimet, F. X. L. and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations. *Tellus*, **38A**, 97–110.

Doucet, A., N. de Freitas, and N. Gordon, 2001: *Sequential Monte-Carlo Methods in Practice*. Springer-Verlag.

Dovera, L. and E. D. Rossa, 2010: Multimodal ensemble Kalman filtering using Gaussian mixture models. *Computational Geosciences*, 1–17.

Duda, R. O., P. E. Hart, and D. G. Stork, 2001: *Pattern Classification*. 2d ed., Wiley-Interscience.

Eisenberger, I., 1964: Genesis of bimodal distributions. *Technometrics*, **6**, 357–363.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *Journal of Geophysical Research-Oceans*, **99 (C5)**, 10 143–10 162.

Evensen, G., 2007: *Data Assimilation, the Ensemble Kalman Filter*. Springer.

Eyink, G. L. and S. Kim, 2006: A maximum entropy method for particle filtering. *Journal of Statistical Physics*, **123 (5)**, 1071–1128.

Frei, M. and H. R. Kunsch, 2011: Mixture ensemble Kalman filters. *Computational Statistics and Data Analysis*.

Gelb, A., 1974: *Applied optimal estimation*. MIT Press.

Ghanem, R. and P. Spanos, 1991: *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag.

Ghil, M. and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Adv. Geophys.*, **3**, 141–266.

Holmes, P., J. Lumley, and G. Berkooz, 1996: *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press.

Hoteit, I., D. T. Pham, G. Triantafyllou, and G. Korres, 2008: A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography. *Monthly Weather Review*, **136 (1)**, 317–334.

Houtekamer, P. L., H. L. Mitchell, and L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, **126**, 796–811.

Ide, K., P. Courtier, M. Ghil, and A. Lorenc, 1997: Unified notation for data assimilation: Operational, sequential and variational. . *Meteor. Soc. Japan*, **75**, 181–189.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory*. Academic Press.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, **82 (Series D)**, 35–45.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press.

Kim, S., G. L. Eyink, J. M. Restrepo, F. J. Alexander, and G. Johnson, 2009: Ensemble filtering for nonlinear dynamics. *Monthly Weather Review*, **131**, 2586–2594, doi:10.1175/1520-0493(2003)131⟨2586:EFFND⟩2.0.CO;2.

Kotecha, J. H. and P. A. Djuric, 2003: Gaussian particle filtering. *IEEE Transactions on Signal Processing*, **51 (10)**, 2592–2601.

Krause, P. and J. M. Restrepo, 2009: The diffusion kernel filter applied to lagrangian data assimilation. *Monthly Weather Review*, **137**, 4386–4400, doi:10.1175/2009MWR2889.1.

Lermusiaux, P. F. J., 1997: Data assimilation via error subspace statistical estimation. Ph.D. thesis, Harvard University, Division of Engineering and Applied Sciences.

Lermusiaux, P. F. J., 1999a: Data assimilation via error subspace statistical estimation, Part II: Middle Atlantic Bight shelfbreak front simulations. *Monthly Weather Review*, **127 (8)**, 1408–1432.

Lermusiaux, P. F. J., 1999b: Estimation and study of mesoscale variability in the Strait of Sicily. *Dynamics of Atmospheres and Oceans*, **29**, 255–303.

Lermusiaux, P. F. J., 2001: Evolving the subspace of the three-dimensional multiscale ocean variability: Massachusetts Bay. *Special issue on "Three-dimensional ocean circulation: Lagrangian measurements and diagnostic analyses"*, J. Marine Systems, Vol. 29, 385–422.

Lermusiaux, P. F. J., 2006: Uncertainty estimation and prediction for interdisciplinary ocean dynamics. *Journal of Computational Physics*, **29**, 176–199, doi:10.1016/j.jcp.2006.02.010, special issue on "Uncertainty Quantification".

Lermusiaux, P. F. J., 2007: Adaptive modeling, adaptive data assimilation and adaptive sampling. *Special issue on 'Mathematical Issues and Challenges in Data Assimilation for Geophysical Systems: Interdisciplinary Perspectives'*, C. K. R. T. Jones and K. Ide, Eds., Physica D, 172–196.

Lermusiaux, P. F. J., C.-S. Chiu, and A. R. Robinson, 2002a: Modeling uncertainties in the prediction of the acoustic wavefield in a shelfbreak environment. *Proceedings of the 5th ICTCA, May 21-25, 2001, in Theoretical and Computational Acoustics*, E.-C. Shang, Q. Li, and T. Gao, Eds., World Scientific Publishing Co., 191–200.

Lermusiaux, P. F. J. and A. Robinson, 1999: Data assimilation via error subspace statistical estimation, Part I: Theory and scheme. *Monthly Weather Review*, **127 (8)**, 1385–1407.

Lermusiaux, P. F. J., A. R. Robinson, P. J. Haley, and W. G. Leslie, 2002b: Advanced interdisciplinary data assimilation: Filtering and smoothing via error subspace statistical estimation. *Proceedings of "The OCEANS 2002 MTS/IEEE" conference*, Holland Publications, 795–802.

Lermusiaux, P. F. J., et al., 2006: Quantifying uncertainties in ocean predictions. *Oceanography*, **19**, 92–105, special issue on "Advances in Computational Oceanography".

Lions, J. L., 1971: *Optimal Control of Systems Governed by Partial Differential Equations*. Springer Verlag, Berlin.

Lorenz, E., 1963: Deterministic nonperiodic flow. *Journal of Atmospheric Science*, **20**, 130–141.

MacKay, D. J. C., 2003: *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 628 pp.

Malanotte-Rizzoli, P., 1996: *Modern Approaches to Data Assimilation in Ocean Modeling*. Amsterdam: Elsevier Oceanography Series, 455 pp.

McLachlan, G. and D. Peel, 2000: *Finite Mixture Models*. John Wiley & Sons, Inc.

Miller, R. N., M. Ghil, and F. Gauthiez, 1994: Advanced data assimilation in strongly nonlinear dynamical-systems. *Journal of the Atmospheric Sciences*, **51 (8)**, 1037–1056.

Moore, A. M., H. G. Arango, E. D. Lorenzo, B. D. Cornuelle, A. J. Miller, and D. J. Neilson, 2004: A comprehensive ocean prediction and analysis system based on the tangent linear and adjoint of a regional ocean model. *Ocean Modelling*, **7 (1-2)**, 227 – 258, doi:10.1016/j.ocemod.2003.11.001, URL `http://www.sciencedirect.com/science/article/pii/S146350030300057X`.

Papoulis, A., 1965: *Probability, Random Variables and Stochastic Processes*. McGraw-Hill.

Pham, D. T., 2001: Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Monthly Weather Review*, **129**, 1194–1207.

Robinson, A. R., P. F. J. Lermusiaux, and N. Q. Sloan, 1998: Data assimilation. *In THE SEA: The Global Coastal Ocean, Volume 10: Processes and Methods*, K. H. Brink and A. R. Robinson, Eds., John Wiley and Sons, NY, 541–594.

Sapsis, T., 2010: Dynamically orthogonal field equations. Ph.D. thesis, Massachusetts Institute of Technology, Department of Mechanical Engineering.

Sapsis, T. and P. F. J. Lermusiaux, 2009: Dynamically orthogonal field equations for continuous stochastic dynamical systems. *Physica D*, **238**, 2347–2360, doi:10.1016/j.physd.2009.09.017.

Sapsis, T. and P. F. J. Lermusiaux, 2011: Dynamical criteria for the evolution of the stochastic dimensionality in flows with uncertainty. *Physica D*, **241**, 60–76, doi:10.1016/j.physd.2011.10.001.

Schwartz, G. E., 1978: Estimating the dimension of a model. *Annals of Statistics*, **6 (11-12)**, 461–464.

Silverman, B., 1992: *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

Smith, K. W., 2007: Cluster ensemble Kalman filter. *Tellus Series a-Dynamic Meteorology and Oceanography*, **59**, 749–757.

Sobczyk, K., 2001: Information dynamics: Premises, challenges and results. *Mechanical Systems and Signal Processing*, **15**, 475–498.

Sondergaard, T., 2011: Data assimilation with Gaussian mixture models using the dynamically orthogonal field equations. M.S. thesis, Massachusetts Institute of Technology, Department of Mechanical Engineering.

Sondergaard, T. and P. F. J. Lermusiaux, 2012: Data assimilation with Gaussian mixture models using the dynamically orthogonal field equations. Part II: Applications. *Monthly Weather Review*, sub-judice.

Sura, P., 2010: On non-Gaussian SST variability in the Gulf Stream and other strong currents. *Ocean Dynamics*, **60**, 155–170.

Tarantola, A., 2005: *Inverse Problem Theory and Model Parameter Estimation*. SIAM.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square root filters. *Monthly Weather Review*, **131**, 1485–1490.

Ueckermann, M. P., P. F. J. Lermusiaux, and T. P. Sapsis, 2012: Numerical schemes for dynamically orthogonal equations of stochastic fluid and ocean flows. *Journal of Computational Physics*, doi:10.1016/j.jcp.2012.08.041, in press.

van Leeuwen, P. J., 2009: Particle filtering in geophysical systems. *Monthly Weather Review*, **137 (12)**, 4089–4114.

Whitaker, J. S. and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, **130**, 1913–1924.

Wunsch, C., 1996: *The Ocean Circulation Inverse Problem*. Cambridge University Press.

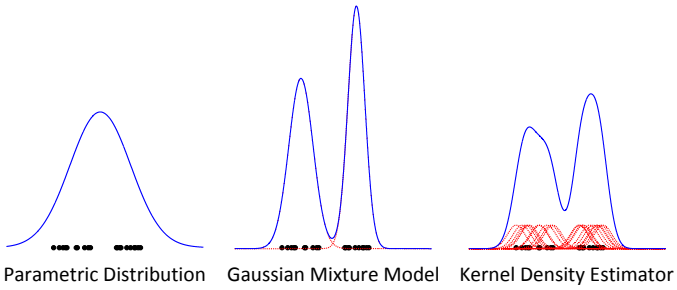

FIG. 1. Gaussian (parametric) distribution, Gaussian mixture model and Gaussian (kernel) density estimator based on 20 samples generated from the mixture of uniform distributions: $p_X\big(x\big) = \frac{1}{2} \times \mathcal{U}(x; -8, -1) + \frac{1}{2} \times \mathcal{U}(x; 1, 8)$, where $\mathcal{U}(x; a, b) = \frac{1}{b-a}$ denotes the continuous uniform pdf for random variable X.

Initial Conditions: *Initialize the state vector in a decomposed form that accords with the Dynamically Orthogonal field equations:*

$$\boldsymbol{x}_{r,0} = \bar{\boldsymbol{x}}_0 + \boldsymbol{\mathcal{X}}_0 \boldsymbol{\phi}_{r,0}, \quad r = \{1, \ldots, N\}.$$

$\bar{\boldsymbol{x}} \in \mathbb{R}^n$ is the mean vector, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n \times s}$ defines the matrix of modes describing an orthonormal basis for the stochastic subspace, and the $\boldsymbol{\phi}_r \in \mathbb{R}^s$ represent N realizations drawn from the multivariate random vector described by $\{\Phi_{1,0}(\omega), \ldots, \Phi_{s,0}(\omega)\}$ that reside in the stochastic subspace of dimension s.

Forecast: *Using either the initial DO conditions or the posterior state description following the assimilation of data at time $k-1$,*

$$\boldsymbol{x}^a_{r,k-1} = \bar{\boldsymbol{x}}^a_{k-1} + \boldsymbol{\mathcal{X}}_{k-1} \boldsymbol{\phi}^a_{r,k-1}, \quad r = \{1, \ldots, N\},$$

apply the DO equations to efficiently evolve the probabilistic description of the state vector in time, arriving at a forecast for observation time k:

$$\boldsymbol{x}^f_{r,k} = \bar{\boldsymbol{x}}^f_k + \boldsymbol{\mathcal{X}}_k \boldsymbol{\phi}^f_{r,k}, \quad r = \{1, \ldots, N\}.$$

Observation: *Employ a linear (or linearized) observation model,*

$$\boldsymbol{Y}_k = \boldsymbol{H}\boldsymbol{X}_k + \boldsymbol{\Upsilon}_k, \quad \boldsymbol{\Upsilon}_k \sim \mathcal{N}(\boldsymbol{\upsilon}_k; \boldsymbol{0}, \boldsymbol{R}).$$

where $\boldsymbol{Y}_k \in \mathbb{R}^p$ is the observation random vector at discrete time k; $\boldsymbol{H} \in \mathbb{R}^{p \times n}$ is the linear observation model; and $\boldsymbol{\Upsilon}_k \in \mathbb{R}^p$ the corresponding random noise vector, assumed to be of a Gaussian distribution. The realized observation vector is denoted by $\boldsymbol{y}_k \in \mathbb{R}^p$.

Fitting of GMM: *For Gaussian mixture models of increasing complexity (i.e. $M = 1, 2, 3, \ldots$), repeat until a minimum of the BIC is met:*

i. Use the EM algorithm to obtain the prior mixture parameters

$$\pi^f_{j,k}, \boldsymbol{\mu}^f_{j,k}, \boldsymbol{\Sigma}^f_{j,k}, \quad j = 1, \ldots, M$$

within the stochastic subspace based on the set of ensemble realizations, $\{\boldsymbol{\phi}^f_k\} = \{\boldsymbol{\phi}^f_{1,k}, \ldots, \boldsymbol{\phi}^f_{N,k}\}$.

ii. Use the Bayesian Information Criterion to evaluate the fit of the Gaussian mixture model for the given M.

Update:

i. Compute:

$$\tilde{\boldsymbol{H}}_k \equiv \boldsymbol{H}\boldsymbol{\mathcal{X}}_k$$
$$\tilde{\boldsymbol{y}}_k \equiv \boldsymbol{y}_k - \boldsymbol{H}\bar{\boldsymbol{x}}^f_k$$

and determine the individual Kalman gain matrices:

$$\tilde{\boldsymbol{K}}_{j,k} = \boldsymbol{\Sigma}^f_{j,k}\tilde{\boldsymbol{H}}^T_k(\tilde{\boldsymbol{H}}_k\boldsymbol{\Sigma}^f_{j,k}\tilde{\boldsymbol{H}}^T_k + \boldsymbol{R})^{-1}, \quad j = 1, \ldots, M.$$

ii. Assimilate the measurement, $\boldsymbol{y}_k$, by calculating the 'intermediate' mixture means in the stochastic subspace,

$$\hat{\boldsymbol{\mu}}^a_{j,k} = \boldsymbol{\mu}^f_{j,k} + \tilde{\boldsymbol{K}}_{j,k}(\tilde{\boldsymbol{y}}_k - \tilde{\boldsymbol{H}}_k\boldsymbol{\mu}^f_{j,k}),$$

and further compute the posterior mixture weights:

$$\pi^a_{j,k} = \frac{\pi^f_{j,k} \times \mathcal{N}(\tilde{\boldsymbol{y}}_k; \tilde{\boldsymbol{H}}_k\boldsymbol{\mu}^f_{j,k}, \tilde{\boldsymbol{H}}_k\boldsymbol{\Sigma}^f_{j,k}\tilde{\boldsymbol{H}}^T_k + \boldsymbol{R})}{\sum_{l=1}^M \pi^f_{l,k} \times \mathcal{N}(\tilde{\boldsymbol{y}}_k; \tilde{\boldsymbol{H}}_k\boldsymbol{\mu}^f_{l,k}, \tilde{\boldsymbol{H}}_k\boldsymbol{\Sigma}^f_{l,k}\tilde{\boldsymbol{H}}^T_k + \boldsymbol{R})}.$$

iii. Update the DO mean field,

$$\bar{\boldsymbol{x}}^a_k = \bar{\boldsymbol{x}}^f_k + \boldsymbol{\mathcal{X}}_k \sum_{j=1}^M \pi^a_{j,k} \times \hat{\boldsymbol{\mu}}^a_{j,k},$$

as well as the mixture parameters within the stochastic subspace:

$$\boldsymbol{\mu}^a_{j,k} = \hat{\boldsymbol{\mu}}^a_{j,k} - \sum_{j=1}^M \pi^a_{j,k} \times \hat{\boldsymbol{\mu}}^a_{j,k}$$
$$\boldsymbol{\Sigma}^a_{j,k} = (\boldsymbol{I} - \tilde{\boldsymbol{K}}_{j,k}\tilde{\boldsymbol{H}}_k)\boldsymbol{\Sigma}^f_{j,k}.$$

iv. Generate the posterior set of subspace ensemble realizations, $\{\boldsymbol{\phi}^a_k\} = \{\boldsymbol{\phi}^a_{1,k}, \ldots, \boldsymbol{\phi}^a_{N,k}\}$, according to the multivariate Gaussian mixture model with posterior parameter values:

$$\pi^a_{j,k}, \boldsymbol{\mu}^a_{j,k}, \boldsymbol{\Sigma}^a_{j,k}, \quad j = 1, \ldots, M.$$

FIG. 2. GMM-DO filter flowchart.

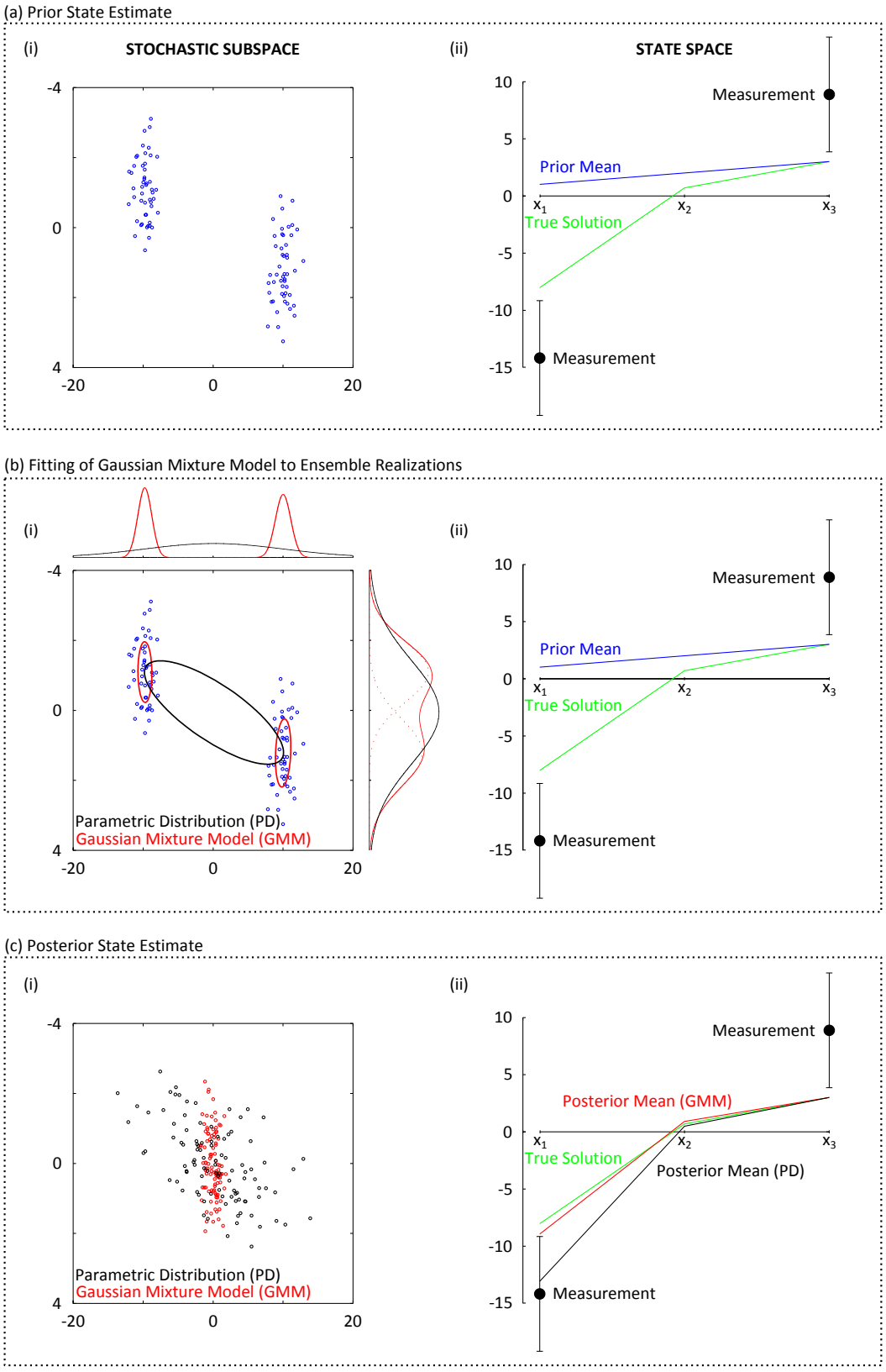

FIG. 3. GMM-DO filter update. In column (i), we plot the set of ensemble realizations within the stochastic subspace, $\{\phi\} = \{\phi_1, \ldots, \phi_{100}\}$; in column (ii), we display the vectors and information residing in the state space. Panel (a) shows the prior state estimate; in panel (b), we show the fitting of Gaussian mixture models of complexity $M = 1$ (PD) and $M = 2$ (GMM), and plot their marginal distributions for each of the stochastic coefficients, Φ_1 and Φ_2; in panel (c), we provide the posterior state estimate again in the decomposed form that accords with the DO equations

TABLE 1. Notation relevant to the GMM-DO filter. (While we have primarily adopted notation specific to probability theory, information theory and estimation theory, where possible we also utilize the notation advocated by Ide et al. (1997).)

Descriptors		
$(\cdot)^f$		forecast
$(\cdot)^a$		analysis

Scalars		
i	$\in \mathbb{N}$	stochastic subspace index
j	$\in \mathbb{N}$	mixture component index
k	$\in \mathbb{N}$	discrete time index
n	$\in \mathbb{N}$	dimension of state vector
p	$\in \mathbb{N}$	dimension of observation vector
q	$\in \mathbb{N}$	dimension of dominant stochastic subspace
r	$\in \mathbb{N}$	realization index
s	$\in \mathbb{N}$	dimension of stochastic subspace
M	$\in \mathbb{N}$	complexity of Gaussian Mixture Model
N	$\in \mathbb{N}$	number of Monte Carlo members
Φ_i	$\in \mathbb{R}$	random variable describing the pdf for orthonormal mode $\tilde{\boldsymbol{x}}_i$

Vectors		
$\boldsymbol{X}$	$\in \mathbb{R}^n$	state (random) vector
$\boldsymbol{x}$	$\in \mathbb{R}^n$	state realization
$\tilde{\boldsymbol{x}}_i$	$\in \mathbb{R}^n$	DO mode i: dynamically orthonormal basis for stochastic subspace
$\bar{\boldsymbol{x}}$	$\in \mathbb{R}^n$	mean state vector
$\boldsymbol{Y}$	$\in \mathbb{R}^p$	observation (random) vector
$\boldsymbol{y}$	$\in \mathbb{R}^p$	observation realization
$\bar{\boldsymbol{x}}_j$	$\in \mathbb{R}^n$	mean vector of mixture component j in state space
$\boldsymbol{\mu}_j$	$\in \mathbb{R}^s$	mean vector of mixture component j in stochastic subspace
$\boldsymbol{\Phi}$	$\in \mathbb{R}^s$	multivariate random vector, $[\Phi_1 \ \dots \ \Phi_s]$
$\boldsymbol{\phi}$	$\in \mathbb{R}^s$	realization residing in stochastic subspace
$\boldsymbol{\Upsilon}$	$\in \mathbb{R}^p$	observation noise (random) vector
$\boldsymbol{\upsilon}$	$\in \mathbb{R}^p$	observation noise realization

Matrices		
$\boldsymbol{P}$	$\in \mathbb{R}^{n \times n}$	covariance matrix in state space
$\boldsymbol{\Sigma}_j$	$\in \mathbb{R}^{s \times s}$	covariance matrix of mixture component j in stochastic subspace
$\boldsymbol{P}_j$	$\in \mathbb{R}^{n \times n}$	covariance matrix of mixture component j in state space
$\boldsymbol{R}$	$\in \mathbb{R}^{p \times p}$	observation covariance matrix
$\boldsymbol{H}$	$\in \mathbb{R}^{m \times n}$	(linear) observation model
$\boldsymbol{\mathcal{X}}$	$\in \mathbb{R}^{n \times s}$	matrix of s DO modes, $[\tilde{\boldsymbol{x}}_1 \ \dots \ \tilde{\boldsymbol{x}}_s]$
$\{\boldsymbol{\phi}\}$	$\in \mathbb{R}^{s \times N}$	set of subspace ensemble realizations, $\{\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_N\}$
$\{\boldsymbol{x}\}$	$\in \mathbb{R}^{n \times N}$	set of state space ensemble realizations, $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$