A Gaussian Mixture Model Smoother for Continuous Nonlinear Stochastic Dynamical Systems: Theory and Scheme

TAPOVAN LOLLA AND PIERRE F. J. LERMUSIAUX

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

ABSTRACT

Retrospective inference through Bayesian smoothing is indispensable in geophysics, with crucial applications in ocean and numerical weather estimation, climate dynamics, and Earth system modeling. However, dealing with the high-dimensionality and nonlinearity of geophysical processes remains a major challenge in the development of Bayesian smoothers. Addressing this issue, a novel subspace smoothing methodology for high-dimensional stochastic fields governed by general nonlinear dynamics is obtained. Building on recent Bayesian filters and classic Kalman smoothers, the fundamental equations and forward–backward algorithms of new Gaussian Mixture Model (GMM) smoothers are derived, for both the full state space and dynamic subspace. For the latter, the stochastic Dynamically Orthogonal (DO) field equations and their time-evolving stochastic subspace are employed to predict the prior subspace probabilities. Bayesian inference, both forward and backward in time, is then analytically carried out in the dominant stochastic subspace, after fitting semiparametric GMMs to joint subspace realizations. The theoretical properties, varied forms, and computational costs of the new GMM smoother equations are presented and discussed.

1. Introduction

Data assimilation traditionally refers to the process of quantitatively estimating the state of a time-varying system using all appropriate modeled and measured information available. In geophysical applications, such as in meteorology and oceanography, the primary purpose of data assimilation has been to accurately estimate the flows in the atmosphere and the ocean (Ghil and Malanotte-Rizzoli 1991; Bennett 1992; Wunsch 1996; Robinson et al. 1998). In these systems, the available information essentially consists of the physical laws that govern the flows, and the indirect, noisy measurements gathered by the sensors observing the system (Talagrand 1997; Kalnay 2003; Daley 1993). In practice, the former is usually available through forecasts and predictions from computational models. Probabilistic frameworks for data assimilation (Van Leeuwen and Evensen 1996) allow us to naturally combine the information arising from noisy measurements with those given by model

predictions and obtain a statistically accurate estimate of the variables of interest. In a Bayesian setting, this combination amounts to accurately computing the posterior distribution of the state variables, conditioned on the observations (Särkkä 2013).

Bayesian filtering and smoothing are two classes of data assimilation problems that differ in their estimation timeline. While filters in their basic form only estimate the current state of the system given all the past measurements, smoothers are used to reconstruct the entire history of states prior to the current time using measurements distributed across time, both past and future (Gelb 1974; Jazwinski 2007). Albeit more computationally challenging than filtering, smoothing is applicable to a much broader range of problems. These include generalized inversions (Bennett 1992) for state estimation and the related variational assimilation schemes (Dimet and Talagrand 1986; Sasaki 1970), adaptive sampling for autonomous vehicles (Choi and How 2010, 2009), stochastic optimal control (Lee and Campbell 2015; Hsieh and Chirikjian 2014), target tracking (Crassidis and Junkins 2011; Thrun et al. 2005), multiresolution imaging (Willsky 2002) and robotic navigation (Kaess et al. 2012;

Corresponding author e-mail: Pierre F. J. Lermusiaux, pierrel@mit.edu

Särkkä 2013), to name a few. In all these applications, since a smoother utilizes more information through future observations, it is expected to yield better estimates than a filter. Smoothing is essential for several geophysical applications, including atmospheric sciences and meteorology (Cohn et al. 1994; Evensen and Van Leeuwen 2000; Khare et al. 2008), as well as ocean modeling (e.g., Lermusiaux and Robinson 1999). Reanalyses especially benefit from smoothing since observations that are subsequent to the estimated states are then also used (e.g., Lermusiaux et al. 2002; Stammer et al. 2002; Moore et al. 2004; Wunsch and Heimbach 2007; Di Lorenzo et al. 2007; Cosme et al. 2012). Other geophysical applications that benefit from smoothing include the estimation of atmospheric chemical sources (Bocquet 2005), adjustment of ocean forcings (Skandrani et al. 2009), and estimation of boundary conditions (Barth et al. 2010).

The landscape of smoothing for linear Gaussian systems is well established. The Kalman smoother then provides the optimal solution, in a Bayesian sense (Gelb 1974). Nonetheless, several optimal linear smoother algorithms exist. These include the fixed-point smoother, the fixed-lag smoother, the fixed-interval smoother (Kitagawa 1987), the Rauch–Tung–Striebel (RTS) smoother (Rauch et al. 1965; Raanes 2016), and the two-filter smoother (Kitagawa 1994). These smoothers are all based on Kalman's hypotheses (Kalman 1960) and the equations of the Kalman filter and, beyond their algorithmic differences, differ from the filter only by handling cross covariances in time to account for future observations (Cosme et al. 2012). They all yield strictly equivalent results when the linear Gaussian assumptions hold. However, this is rarely the case in geophysical systems, well known to be highly nonlinear and chaotic (Miller et al. 1999). As a result, ocean and atmospheric fields can develop complex, far-from-Gaussian statistics (e.g., Lermusiaux 2006). Nonlinearities thus not only affect forecasts, but also the melding of information from future observations with state variables in the past. Therefore, smoothing schemes should fully respect the nonlinearity of the known dynamics as they estimate the effect of observations through time.

The relative simplicity of the Kalman framework has prompted the development of similar types of smoothers, but applicable to high-dimensional non-linear problems in geophysics (Bocquet et al. 2010). Ensemble-based methods, in particular, stand out. These include the ensemble Kalman smoother (EnKS; Evensen and Van Leeuwen 2000) and the error

subspace smoother (ESSE; Lermusiaux and Robinson 1999), as well as fast ensemble smoothers (Ravela and McLaughlin 2007). These schemes represent the state variables in the form of Monte Carlo particles, and advance them in time using the nonlinear governing equations. This allows the exploration and exploitation of probabilistic structures beyond the basic Gaussian representation. However, these methods typically perform Gaussian updates, either for each particle in the full state space (EnKS) or for the mean in a reduced subspace (ESSE). Even though these Gaussian updates ignore the higher-order moments of the distribution, ensemble smoothers are popular because of their relative simplicity. For other linear smoother algorithms extended to geophysical applications, we refer to Cohn et al. (1994) and Cosme et al. (2010).

Sequential Monte Carlo (SMC) smoothers or particle smoothers are a class of Monte Carlo smoothing methods that sample successively from a sequence of target smoothed probability densities (Doucet and Johansen 2009). These schemes are not always related to Kalman-based approaches but they aim to overcome the limitations of a Gaussian update while retaining the ability to capture the non-Gaussian state features and also utilize the nonlinear dynamics. For example, Bresler (1986) extends the traditional two-filter smoother to a nonlinear, non-Gaussian setting. Similarly, Godsill et al. (2004) and Briers et al. (2010) develop RTS-style forward–backward smoothers for general state-space models. These schemes are asymptotically optimal, in the limit of infinite particles. For more on their implementations, we also refer to Klaas et al. (2006).

Even though particle smoothers are attractive because of their asymptotic optimality in nonlinear, non-Gaussian settings, several challenges remain for their use in geophysical systems. A major challenge is the high dimensionality (today, 10^6–10^{12}) of state vectors commonly encountered in oceanic and atmospheric applications. Resolving such high-dimensional state vectors requires a prohibitively large set of particles. Moreover, in many applications, particle methods can suffer from sample impoverishment, a phenomenon in which ensembles collapse into a handful of heavily weighted samples. Implicit particle smoothers (Weir et al. 2013; Atkins et al. 2013) are outcomes of recent efforts to address such issues. A related interest in smoothing has been the approximation of distributions by Gaussian Mixture Models (GMMs), as explored in Lee and Campbell (2015), Tagade et al. (2014), and Vo et al. (2012).

To address the challenges of high dimensionality, uncertainty quantification can focus on the *time-dependent*

dominant error subspace (Lermusiaux 1997), thereby allocating computational resources to the probabilistic states that matter most. The Dynamically Orthogonal (DO) field equations (Sapsis and Lermusiaux 2009, 2012) provide optimal reduced-order differential equations to evolve this dominant subspace forward in time, using the governing nonlinear dynamics (Feppon and Lermusiaux 2017). Building on the recent GMM–DO filter (Sondergaard and Lermusiaux 2013a), we first derive the fundamental equations of the full state-space, and subspace, GMM smoothers. We then develop an RTS-style implementation scheme for these non-Gaussian smoothers, where filtering is carried out in the forward pass using the state-space, or subspace, GMM filter. Starting at the final observation time, smoothing is then performed by propagating information backward in time *without linearizing* the dynamics, while also retaining the non-Gaussian GMM nature of the joint state densities across time. For the optimal reduced-order representation of high-dimensional stochastic fields governed by nonlinear dynamics, we finally obtain the equations for the GMM–DO smoother, a particular case of subspace-GMM smoothers. The GMM–DO smoother first uses the GMM–DO filter to quantify uncertainty prior to smoothing and then performs the GMM smoothing by carrying out Bayes's law analytically in the low-dimensional, time-evolving DO subspace. Critically, under the DO representation, the state-space GMM and subspace GMM–DO smoothers are shown to be equivalent.

This paper is structured as follows. In section 2, we introduce the notation and state the smoothing problem. Section 3 derives the core equations of the state-space and subspace GMM smoothers. For the GMM–DO smoother, we prove a key theorem that shows the equivalence between Bayesian smoothing performed in the full state space and that performed in the reduced stochastic DO subspace. We then discuss the theoretical properties, computational costs, and other forms of the smoothers. Conclusions and future work are provided in section 4. Tables 1 and 2 summarize the notation and main derived equations. The GMM–DO filter equations and schemes are summarized in appendixes A and B. In a companion paper (Lolla and Lermusiaux 2017), we illustrate and validate the GMM–DO smoother. There, we compare its performance to other smoothers, using three complementary dynamical system applications: a double-well diffusion experiment, reversible passive tracer advection, and a simulated ocean flow exiting a strait/estuary.

2. Notation and problem statement

a. Dynamical model

Let $X(\mathbf{r}, t; \omega)$: $\mathbb{R}^n \times [0, T] \to \mathbb{R}$ be a continuous stochastic field governed by a stochastic partial differential equation (SPDE) with stochastic initial conditions and boundary conditions:

$$\frac{\partial X(\mathbf{r}, t; \omega)}{\partial t} = \mathcal{L}[X(\mathbf{r}, t; \omega); \omega], \quad t \geq 0 \tag{1a}$$

$$X(\mathbf{r}, 0; \omega) = X_0(\mathbf{r}, \omega), \quad \text{and} \tag{1b}$$

$$\mathcal{B}[X(\mathbf{r}, t; \omega)]\big|_{\mathbf{r}=\boldsymbol{\xi}} = h(\boldsymbol{\xi}, t; \omega), \tag{1c}$$

where $\mathbf{r}$ and $\boldsymbol{\xi}$, respectively, denote the interior and boundary spatial coordinates; t is time; and ω is a random event. We use $\mathcal{L}[\cdot]$ to represent a general nonlinear differential operator for the dynamics, and $\mathcal{B}$ is a linear differential operator. The state variable X can, for example, represent atmospheric, oceanic, or fluid flow fields.

Let $\mathbf{X}(t; \omega)$ denote the spatially discretized state vector of the continuous field $X(\mathbf{r}, t; \omega)$, and $N_{\mathbf{X}}$ denote the dimensionality of the state space, that is, the size of $\mathbf{X}(t; \omega)$. We use a bold roman font to denote vectors and bold sans serif font for matrices. Uppercase letters and symbols parameterized by ω are random variables and their corresponding lowercase counterparts denote specific realizations. We omit ω in some cases where no confusion is expected. Finally, a multivariate Gaussian pdf with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is denoted by $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

b. Observation model

We are provided access to indirect, noisy observations of $\mathbf{X}(t; \omega)$ through the linear (or linearized) observation model:

$$\mathbf{Y}(t; \omega) = \mathbf{H}\mathbf{X}(t; \omega) + \mathbf{Y}(t; \omega), \quad \mathbf{Y}(t; \omega) \sim \mathcal{N}(\cdot; \mathbf{0}, \mathbf{R}), \tag{2}$$

where $\mathbf{H}$ is the observation matrix and $\mathbf{Y}$ is a zero-mean, uncorrelated Gaussian measurement noise with the covariance matrix $\mathbf{R}$. Observations are made at times t_k, for $k = 1, 2, \ldots, K$. For ease of notation, we denote $\mathbf{X}(t_k; \omega)$ and $\mathbf{Y}(t_k; \omega)$ by $\mathbf{X}_k$ and $\mathbf{Y}_k$, respectively. Posterior quantities (i.e., conditioned on the observations) are also indicated through subscripts; for example, $\mathbf{X}_k$ conditioned on observations $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_l$ is denoted by $\mathbf{X}_{k|1:l}$. With this notation, filtering and smoothing then amount to computing $\mathbf{X}_{k|1:k}$ and $\mathbf{X}_{k|1:l}$, respectively, for $1 \leq k \leq l \leq K$. Our goal here is to determine the smoothed quantities $\mathbf{X}_{k|1:K}$, for all $k = 1, 2, \ldots, K - 1$. Note that in the above setup, we have assumed without loss of generality, that smoothing times coincide with the

TABLE 1. Table of notation.

Symbol	Type	Description
Scalars		
t		Time
ω		Experiment number
i	$\in \mathbb{N}$	Stochastic subspace index
j	$\in \mathbb{N}$	Mixture component index
n	$\in \mathbb{N}$	Dimension of spatial coordinate
$N_{\mathbf{X}}$	$\in \mathbb{N}$	Dimension of discrete state vector $\mathbf{X}(t; \omega)$
N_{obs}	$\in \mathbb{N}$	Dimension of observation vector $\mathbf{Y}(t; \omega)$
s	$\in \mathbb{N}$	Dimension of stochastic space
N_r	$\in \mathbb{N}$	No. of (Monte Carlo) realizations
M	$\in \mathbb{N}$	No. of mixture components
r	$\in \{1, 2, \ldots, N_r\}$	Realization index
K	$\in \mathbb{N}$	No. of smoothing indices
k, l	$\in \{1, 2, \ldots, K\}$	Smoothing indices
t_k	$\in \mathbb{R}$	Time at smoothing index k
$\pi_k^{\ j}$	$\in \mathbb{R}^+$	jth component weight of the prior forecast GMM pdf, $p_{\mathbf{\Phi}_k \mid \mathbf{Y}_{1:k-1}}$
$\hat{\pi}_k^{\ j}$	$\in \mathbb{R}^+$	jth component weight of the posterior filtered GMM pdf, $p_{\mathbf{\Phi}_k \mid \mathbf{Y}_{1:k}}$
$X(\mathbf{r}, t; \omega)$	$\in \mathbb{R}$	Continuous stochastic field
$\bar{x}(\mathbf{r}, t)$	$\in \mathbb{R}$	Continuous mean field [mean of $X(\mathbf{r}, t; \omega)$]
$\tilde{x}_i(\mathbf{r}, t)$	$\in \mathbb{R}$	Continuous DO mode i: orthonormal basis for stochastic subspace
$\Phi_i(t; \omega)$	$\in \mathbb{R}$	Stochastic coefficient i
$\phi_i^{(r)}(t)$	$\in \mathbb{R}$	Realization No. r of stochastic coefficient i
Vectors		
$\mathbf{r}$	$\in \mathbb{R}^n$	Spatial coordinate
$\mathbf{\Phi}(t; \omega)$	$\in \mathbb{R}^s$	Vector of stochastic coefficients, $[\Phi_1(t; \omega), \Phi_2(t; \omega), \ldots, \Phi_s(t; \omega)]^{\mathrm{T}}$
$\mathbf{\Phi}_k(\omega)$	$\in \mathbb{R}^s$	Vector of stochastic coefficients at time t_k
$\boldsymbol{\phi}^{(r)}(t)$	$\in \mathbb{R}^s$	Realization r of the vector of stochastic coefficients, $[\phi_1^{(r)}(t), \phi_2^{(r)}(t), \ldots, \phi_s^{(r)}(t)]^{\mathrm{T}}$
$\boldsymbol{\phi}_k^{(r)}$	$\in \mathbb{R}^s$	Realization r of the random vector of stochastic coefficients at time t_k
$\mathbf{X}(t; \omega)$	$\in \mathbb{R}^{N_{\mathbf{X}}}$	State vector [spatially discretized $X(\mathbf{r}, t; \omega)$]
$\bar{\mathbf{x}}(t)$	$\in \mathbb{R}^{N_{\mathbf{X}}}$	Discrete mean field [mean of $\mathbf{X}(t; \omega)$]
$\tilde{\mathbf{x}}_i(t)$	$\in \mathbb{R}^{N_{\mathbf{X}}}$	Discrete DO mode i, forming the orthonormal basis for stochastic subspace
$\bar{\mathbf{x}}^j(t)$	$\in \mathbb{R}^{N_{\mathbf{X}}}$	Mean vector of mixture component j in state space
$\mathbf{x}^{(r)}(t)$	$\in \mathbb{R}^{N_{\mathbf{X}}}$	Realization number r in state space
$\mathbf{x}_k^{(r)}$	$\in \mathbb{R}^{N_{\mathbf{X}}}$	Realization number r in state space at time t_k
$\bar{\boldsymbol{\mu}}^j(t)$	$\in \mathbb{R}^s$	Mean vector of mixture component j in stochastic subspace
$\mathbf{Y}$	$\in \mathbb{R}^{N_{\text{obs}}}$	Observation vector
$\mathbf{y}$	$\in \mathbb{R}^{N_{\text{obs}}}$	Realization of observation vector $\mathbf{Y}$
$\mathbf{\Upsilon}$	$\in \mathbb{R}^{N_{\text{obs}}}$	Observation noise
υ	$\in \mathbb{R}^{N_{\text{obs}}}$	Realization of observation noise
$\mathbf{\Psi}_{k \mid 1:l}$		Any random vector $\mathbf{\Psi}(t_k; \omega)$ conditioned on observations $\mathbf{Y}_{1:l}$
$\mathbf{\Psi}_{k,k+1 \mid 1:l}$		Vector formed by augmenting $\mathbf{\Psi}_{k+1 \mid 1:l}$ to $\mathbf{\Psi}_{k \mid 1:l}$
Matrices		
$\mathcal{X}$	$\in \mathbb{R}^{N_{\mathbf{X}} \times s}$	Matrix of orthonormal DO basis vectors $[\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_s]$
$\mathbf{P}$	$\in \mathbb{R}^{N_{\mathbf{X}} \times N_{\mathbf{X}}}$	Covariance matrix in state space
$\mathbf{\Sigma}_j$	$\in \mathbb{R}^{s \times s}$	Covariance matrix of mixture component j in the stochastic subspace
$\mathbf{P}_j$	$\in \mathbb{R}^{N_{\mathbf{X}} \times N_{\mathbf{X}}}$	Covariance matrix of mixture component j in the state space
$\mathbf{R}$	$\in \mathbb{R}^{N_{\text{obs}} \times N_{\text{obs}}}$	Covariance matrix of observation noise
$\mathbf{H}$	$\in \mathbb{R}^{N_{\text{obs}} \times N_{\mathbf{X}}}$	Observation matrix
Densities		
$p_{\mathbf{\Psi}_k \mid \mathbf{Y}_{1:l}}$		The pdf of the vector $\mathbf{\Psi}(t_k; \omega)$ conditioned on observations $\mathbf{Y}_{1:l}$—filtering $(k = l)$, forecast $(k > l)$, and smoothing $(k < l)$
$p_{\mathbf{\Psi}_k, \mathbf{\Psi}_m \mid \mathbf{Y}_{1:l}}$		Joint pdf of vectors $\mathbf{\Psi}_k$ and $\mathbf{\Psi}_m$ conditioned on $\mathbf{Y}_{1:l}$
$p_{\mathbf{\Psi}_k \mid \mathbf{\Psi}_{k+1}, \mathbf{Y}_{1:k}}$		The pdf of $\mathbf{\Psi}_k$, conditioned on both $\mathbf{\Psi}_{k+1}$ and $\mathbf{Y}_{1:k}$

times at which observations occur. All the key symbols are listed in Table 1.

3. Subspace-GMM smoothers and the GMM–DO smoother

a. Preliminaries

GMM smoothers are a general class of nonlinear, non-Gaussian smoothers that assume GMM distributions at the assimilation step. Their uncertainty prediction schemes to integrate the governing dynamics (1) and their filtering and smoothing schemes to assimilate observations (2) must respect the nonlinearities in the dynamics and capture the non-Gaussian statistics of the system in the GMM sense. Their subspace-GMM version employs reduced-order decompositions of the form (3) for both the forecast–prior ($\mathbf{X}_{k+1|1:k}$) and filtered–posterior ($\mathbf{X}_{k|1:k}$) state vectors:

$$\mathbf{X}_{l|1:k}(\omega) = \bar{\mathbf{x}}_{l|1:k} + \boldsymbol{\mathcal{X}}_l \boldsymbol{\Phi}_{l|1:k}(\omega), \quad l \in \{k, k+1\}, \quad l \le K,$$
$$k = 1, 2, \ldots, K, \tag{3}$$

where $\bar{\mathbf{x}}_{l|1:k}$ denotes the mean of $\mathbf{X}_{l|1:k}(\omega)$, $\boldsymbol{\mathcal{X}}_l$ is an $N_{\mathbf{X}} \times s$ matrix of orthonormal columns (modes) (i.e., $\boldsymbol{\mathcal{X}}_l^{\mathrm{T}} \boldsymbol{\mathcal{X}}_l = \mathbf{I}$), and $\boldsymbol{\Phi}_{l|1:k}(\omega)$ is a time-dependent $s \times 1$ vector of zero-mean stochastic coefficients (Table 1). The columns of $\boldsymbol{\mathcal{X}}_l$ form an orthonormal basis for a time-dependent s-dimensional stochastic subspace, whereas the vector $\boldsymbol{\Phi}_{l|1:k}(\omega)$ describes the randomness of $\mathbf{X}_{l|1:k}(\omega)$ within that subspace. The stochastic vector $\boldsymbol{\Phi}_{l|1:k}(\omega)$ is represented by its realizations, $\boldsymbol{\phi}_{l|1:k}^{(r)}$, $r = 1, 2, \ldots, N_r$. Since l has two values for each $k < K$, (3) denotes a total of $2K - 1$ reduced-order decompositions. Finally, we remark that if filtering at t_k leads to data-driven learning of the stochastic subspace and thus an increase of its dimension (e.g., as in Lermusiaux 1999, 2007), $\boldsymbol{\mathcal{X}}_k$ then corresponds to that larger learned filtered subspace (i.e., $= \boldsymbol{\mathcal{X}}_{k|1:k}$) and contains the forecast subspace $\boldsymbol{\mathcal{X}}_{k|1:k-1}$ by construction.

The requirement of the reduced-order decompositions (3) is not very restrictive. In fact, most ensemble-based schemes for data assimilation can be cast in this form. In such schemes, the columns of $\boldsymbol{\mathcal{X}}_l$ correspond to the (leading) singular vectors of the prior/posterior ensemble spread matrix, and the elements of $\boldsymbol{\phi}_{l|1:k}^{(r)}$ are the projections of the mean-removed state-space realizations $\mathbf{x}_{l|1:k}^{(r)} - \bar{\mathbf{x}}_{l|1:k}$ on to the columns of $\boldsymbol{\mathcal{X}}_l$. Similarly, some ensemble schemes directly provide a GMM representation for the state variable (Hoteit et al. 2008). In this paper and its companion (Lolla and Lermusiaux 2017), we emphasize the specific case of the GMM–DO smoother, which uses the GMM–DO filter

(Sondergaard and Lermusiaux 2013a) for uncertainty prediction and filtering. A summary of this filter, as well as its use of the Expectation-Maximization (EM) algorithm and Bayesian Information Criterion (BIC) to fit GMMs to DO subspace realizations, are provided in appendixes A and B, respectively. Nonetheless, the equations for subspace-GMM smoothing that we derive are applicable to the many assimilation schemes that satisfy (3).

In the RTS form, a subspace-GMM smoother starts at the final observation time t_K, uses the forecast–filtered decompositions in (3), and marches *backward* in time through each time t_k, for $k = K - 1, K - 2, \ldots, 1$. Its goal is to compute a smoothed reduced-order decomposition of the form

$$\mathbf{X}_{k|1:K}(\omega) = \bar{\mathbf{x}}_{k|1:K} + \boldsymbol{\mathcal{X}}_k \boldsymbol{\Phi}_{k|1:K}(\omega), \tag{4}$$

where $\bar{\mathbf{x}}_{k|1:K}$ is the mean of the smoothed state pdf $p_{\mathbf{X}_k|\mathbf{Y}_{1:K}}(\cdot \mid \mathbf{y}_{1:K})$ and $\boldsymbol{\Phi}_{k|1:K}(\omega)$ is the s-dimensional vector of zero-mean smoothed stochastic coefficients that describe the randomness of $\mathbf{X}_{k|1:K}(\omega)$ within the subspace spanned by the columns of $\boldsymbol{\mathcal{X}}_k$. For now, the matrices $\boldsymbol{\mathcal{X}}_k$ are assumed to be unchanged by the observations collected *after* t_k. In other words, the smoothing process does not change the filtered subspace at t_k; of course, it changes the stochastic coefficients $\boldsymbol{\Phi}_{k|1:K}(\omega)$ and thus the ensemble, but, beyond rotations, it does not change the discrete modes given by the columns of $\boldsymbol{\mathcal{X}}_k$. This point is further discussed in section 3f, along with smoothing schemes that adapt the filtered subspace. As with the filter, the smoothed stochastic coefficient vector $\boldsymbol{\Phi}_{k|1:K}(\omega)$ is represented by its realizations, denoted by $\boldsymbol{\phi}_{k|1:K}^{(r)}$, for $r = 1, 2, \ldots, N_r$. We also note that at the final observation time t_K (i.e., for $k = K$) the smoothed and filtered distributions [i.e., (4), with $k = K$] are identical, by definition.

In what follows, we first provide the recursive equation for smoothing. Then, we derive the fundamental GMM smoother updates in the full state space and in the stochastic subspace through a $k + 1 \leftrightarrow k$ *joint subspace* GMM fit and a backward-smoothing pass for $k = K - 1, K - 2, \ldots, 1$. Specifically, we derive the Bayesian smoothing equations for the mean $\bar{\mathbf{x}}_{k|1:K}$, the GMM updates in the stochastic subspace, and the corresponding realizations $\boldsymbol{\phi}_{k|1:K}^{(r)}$ representing $\boldsymbol{\Phi}_{k|1:K}(\omega)$. Within the GMM assumptions, these equations are exact.

b. The recursive equation for smoothing

We first derive an equation that relates $p_{\mathbf{X}_k|\mathbf{Y}_{1:K}}(\mathbf{x}_k \mid \mathbf{y}_{1:K})$, the smoothed pdf at time t_k, to that at time t_{k+1}. This recursion will later be used to develop the backward-smoothing

TABLE 2. GMM–DO smoother: summary of equations and algorithm.

GMM–DO smoother

a. *Forward GMM–DO filter pass*: Solve the DO equations (A3)–(A5) to predict the state pdf. At each observation time t_k, perform the analysis step of the GMM–DO filter (section 4). Save the following:

　1) mean vectors $\bar{\mathbf{x}}_{k|1:k}$ (filtered) and $\bar{\mathbf{x}}_{k+1|1:k}$ (forecast) for $k = 1, 2, \ldots, K - 1$;

　2) sets of stochastic coefficients $\{\boldsymbol{\phi}_{k|1:k}^{(r)}\}_{r=1}^{N_r}$ (filtered) and $\{\boldsymbol{\phi}_{k+1|1:k}^{(r)}\}_{r=1}^{N_r}$ (forecast) for $k = 1, 2, \ldots, K - 1$;

　3) matrices of modes $\boldsymbol{\mathcal{X}}_k$ for $k = 1, 2, \ldots, K$;

　4) the final-time filtered variables—the stochastic coefficients $\{\boldsymbol{\phi}_{K|1:K}^{(r)}\}_{r=1}^{N_r}$ and the mean vector $\bar{\mathbf{x}}_{K|1:K}$.

b. *Joint subspaces GMM-fitting pass*: Form the realizations of $\boldsymbol{\Phi}_{k,k+1|k}(\omega)$, as per (28): $\boldsymbol{\phi}_{k,k+1|1:k}^{(r)} = \begin{bmatrix} \boldsymbol{\phi}_{k|1:k}^{(r)} \\ \boldsymbol{\phi}_{k+1|1:k}^{(r)} \end{bmatrix}$.

Fit a GMM to each joint ensemble $\{\boldsymbol{\phi}_{k,k+1|1:k}^{(r)}\}_{r=1}^{N_r}$ using the EM–BIC procedure, to obtain the joint filtered GMMs (15) for $k = 1, 2, \ldots, K - 1$:

$$p_{\boldsymbol{\Phi}_k, \boldsymbol{\Phi}_{k+1}|\mathbf{Y}_{1:k}}(\boldsymbol{\phi}_k, \boldsymbol{\phi}_{k+1} | \mathbf{y}_{1:k}) = \sum_{j=1}^{M} \pi^j \times \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\phi}_k \\ \boldsymbol{\phi}_{k+1} \end{bmatrix}; \begin{bmatrix} \overline{\boldsymbol{\mu}}_{k|1:k}^j \\ \overline{\boldsymbol{\mu}}_{k+1|1:k}^j \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{k,k|1:k}^j & \boldsymbol{\Sigma}_{k,k+1|1:k}^j \\ \boldsymbol{\Sigma}_{k+1,k|1:k}^j & \boldsymbol{\Sigma}_{k+1,k+1|1:k}^j \end{bmatrix} \right).$$

c. *Backward-smoothing pass*: Execute the following steps sequentially, starting from $k = K - 1$ until $k = 1$:

　1) For each $r = 1, 2, \ldots, N_r$

　　(i) determine the subspace conditional pdf $p_{\boldsymbol{\Phi}_k|\boldsymbol{\Phi}_{k+1}, \mathbf{Y}_{1:k}}(\cdot | \boldsymbol{\phi}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k})$ from (20)–(21), where

$$p_{\boldsymbol{\Phi}_k|\boldsymbol{\Phi}_{k+1}, \mathbf{Y}_{1:k}}(\boldsymbol{\phi}_k | \boldsymbol{\phi}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k}) = \sum_{j=1}^{M} \hat{\pi}^{j,(r)} \times \mathcal{N}(\boldsymbol{\phi}_k; \tilde{\boldsymbol{\mu}}_k^{j,(r)}, \tilde{\boldsymbol{\Sigma}}_k^j);$$

　　(ii) draw the sample $\tilde{\boldsymbol{\phi}}_{k|1:K}^{(r)}$ from (27), where $\tilde{\boldsymbol{\phi}}_{k|1:K}^{(r)} \sim p_{\boldsymbol{\Phi}_k|\boldsymbol{\Phi}_{k+1}, \mathbf{Y}_{1:k}}(\cdot | \boldsymbol{\phi}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k})$;

　2) compute the smoothed mean $\bar{\mathbf{x}}_{k|1:K}$ from (29), where $\bar{\mathbf{x}}_{k|1:K} = \bar{\mathbf{x}}_{k|1:k} + \boldsymbol{\mathcal{X}}_k \times \left(\dfrac{1}{N_r} \displaystyle\sum_{r=1}^{N_r} \tilde{\boldsymbol{\phi}}_{k|1:K}^{(r)} \right)$;

　3) compute the zero-mean vectors of smoothed stochastic DO coefficients $\boldsymbol{\phi}_{k|1:K}^{(r)}$ from (30), where

$$\boldsymbol{\phi}_{k|1:K}^{(r)} = \tilde{\boldsymbol{\phi}}_{k|1:K}^{(r)} - \frac{1}{N_r} \sum_{r=1}^{N_r} \tilde{\boldsymbol{\phi}}_{k|1:K}^{(r)}, \quad r = 1, 2, \ldots, N_r;$$

　4) decrement k by 1 and go to step 1).

pass. For any $k \in \{1, 2, \ldots, K - 1\}$, the recursion is based on the $k + 1 \leftrightarrow k$ joint smoothed pdf $p_{\mathbf{X}_k, \mathbf{X}_{k+1}|\mathbf{Y}_{1:K}}$ and its marginalization to $\mathbf{X}_k$. Using the definition of conditional pdfs, $p_{\mathbf{X}_k, \mathbf{X}_{k+1}|\mathbf{Y}_{1:K}}$ is written as the product of the conditional pdf $p_{\mathbf{X}_k|\mathbf{X}_{k+1}, \mathbf{Y}_{1:K}}$ and the marginal smoothed pdf $p_{\mathbf{X}_{k+1}|\mathbf{Y}_{1:K}}$. This yields

$$p_{\mathbf{X}_k|\mathbf{Y}_{1:K}}(\mathbf{x}_k | \mathbf{y}_{1:K}) =$$
$$\int p_{\mathbf{X}_k|\mathbf{X}_{k+1}, \mathbf{Y}_{1:K}}(\mathbf{x}_k | \mathbf{x}_{k+1}, \mathbf{y}_{1:K}) p_{\mathbf{X}_{k+1}|\mathbf{Y}_{1:K}}(\mathbf{x}_{k+1} | \mathbf{y}_{1:K}) \, d\mathbf{x}_{k+1}.$$
$$(5)$$

Due to the Markovian property of the dynamics in (1a), when conditioned on $\mathbf{X}_{k+1}$ (the smoothed present state), the future observations $\mathbf{Y}_{k+1:K}$ provide no additional information on the past state $\mathbf{X}_k$ (Cosme et al. 2012). Hence, the conditioning on $\mathbf{Y}_{k+1:K}$ may be dropped from $p_{\mathbf{X}_k|\mathbf{X}_{k+1}, \mathbf{Y}_{1:K}}$; that is,

$$p_{\mathbf{X}_k|\mathbf{X}_{k+1}, \mathbf{Y}_{1:K}}(\mathbf{x}_k | \mathbf{x}_{k+1}, \mathbf{y}_{1:K}) = p_{\mathbf{X}_k|\mathbf{X}_{k+1}, \mathbf{Y}_{1:k}}(\mathbf{x}_k | \mathbf{x}_{k+1}, \mathbf{y}_{1:k}).$$
$$(6)$$

Substituting (6) into (5) yields the final form of the *recursive smoothing equation*:

$$p_{\mathbf{X}_k|\mathbf{Y}_{1:K}}(\mathbf{x}_k | \mathbf{y}_{1:K}) =$$
$$\int p_{\mathbf{X}_k|\mathbf{X}_{k+1}, \mathbf{Y}_{1:k}}(\mathbf{x}_k | \mathbf{x}_{k+1}, \mathbf{y}_{1:k}) p_{\mathbf{X}_{k+1}|\mathbf{Y}_{1:K}}(\mathbf{x}_{k+1} | \mathbf{y}_{1:K}) \, d\mathbf{x}_{k+1}.$$
$$(7)$$

The recursive equation (7) may be interpreted as follows. Assuming that the $(k + 1)$th smoothed state pdf $p_{\mathbf{X}_{k+1}|\mathbf{Y}_{1:K}}(\cdot | \mathbf{y}_{1:K})$ and the $k \leftrightarrow k + 1$ conditional pdf $p_{\mathbf{X}_k|\mathbf{X}_{k+1}, \mathbf{Y}_{1:k}}(\cdot | \mathbf{x}_{k+1}, \mathbf{y}_{1:k})$ can be sampled from, (7) outlines a method for generating smoothed realizations from $p_{\mathbf{X}_k|\mathbf{Y}_{1:K}}(\cdot | \mathbf{y}_{1:K})$. Given any sample $\mathbf{x}_{k+1|1:K}^{(r)}$ drawn from the pdf $p_{\mathbf{X}_{k+1}|\mathbf{Y}_{1:K}}(\cdot | \mathbf{y}_{1:K})$, the corresponding smoothed sample $\mathbf{x}_{k|1:K}^{(r)}$ of $p_{\mathbf{X}_k|\mathbf{Y}_{1:K}}(\cdot | \mathbf{y}_{1:K})$ is obtained by drawing a sample from the conditional pdf $p_{\mathbf{X}_k|\mathbf{X}_{k+1}, \mathbf{Y}_{1:k}}(\cdot | \mathbf{x}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k})$; that is,

$$\mathbf{x}_{k|1:K}^{(r)} \sim p_{\mathbf{X}_k|\mathbf{X}_{k+1}, \mathbf{Y}_{1:k}}(\cdot | \mathbf{x}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k}). \quad (8)$$

In theory, the above process can be repeated for each realization $\mathbf{x}_{k+1|1:K}^{(r)}$, $r = 1, 2, \ldots, N_r$, in order to form the set of ensemble members $\{\mathbf{x}_{k|1:K}^{(1)}, \mathbf{x}_{k|1:K}^{(2)}, \ldots, \mathbf{x}_{k|1:K}^{(N_r)}\}$ representing the smoothed distribution $p_{\mathbf{X}_k|\mathbf{Y}_{1:K}}(\cdot | \mathbf{y}_{1:K})$. However, this simple approach suffers from two major issues, both of which are addressed in the next two

sections. First, the primary assumption in the above procedure is the availability of the conditional pdf $p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\cdot\,|\,\mathbf{x}_{k+1},\mathbf{y}_{1:k})$. For general nonlinear systems, the computation of this conditional pdf is nontrivial as it involves the $k + 1$ to k inversion of the k to $k + 1$ nonlinear operator that evolves $\mathbf{X}_k$ into $\mathbf{X}_{k+1}$. This direct model inversion is unstable for irreversible dynamical systems. Second, the pdfs $p_{\mathbf{X}_{k+1}|\mathbf{Y}_{1:K}}$ and $p_{\mathbf{X}_k|\mathbf{Y}_{1:K}}$ describe the smoothed variables in the full *state space* and the sampling operation (8) is also performed in the state space. This renders the above smoothing approach prohibitively expensive and impractical for nonlinear systems with high-dimensional state spaces, the main focus of this work.

In what follows, we address the two issues above and derive the GMM smoother equations. Section 3c discusses how to compute $p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}$, respecting and utilizing both the $k + 1$ to k nonlinearity in the dynamics and the non-Gaussian structures of the pdfs involved, in the full state space. Section 3d describes how the subspace smoother uses the reduced-order decompositions in (3) to solve (7) in the stochastic subspace and shows that, under these conditions (3), it is equivalent to the Bayes's update in the high-dimensional state space, as shown in section 3c.

c. Smoother updates in the state space

1) JOINT STATE-SPACE GMM

To perform the smoother updates given by the recursion (7) directly in the state space, we need to evaluate and draw samples from the conditional pdf $p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k\,|\,\mathbf{x}_{k+1|1:K}^{(r)},\mathbf{y}_{1:k})$ in (8). To this end, we start by representing the *joint filtered* pdf $p_{\mathbf{X}_k,\mathbf{X}_{k+1}|\mathbf{Y}_{1:k}}$ as a GMM, given by

$$p_{\mathbf{X}_k,\mathbf{X}_{k+1}|\mathbf{Y}_{1:k}}(\mathbf{x}_k,\mathbf{x}_{k+1}\,|\,\mathbf{y}_{1:k}) = \sum_{j=1}^{M} \pi^j$$
$$\times \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_k\\\mathbf{x}_{k+1}\end{bmatrix};\begin{bmatrix}\overline{\mathbf{x}}_{k|1:k}^j\\\overline{\mathbf{x}}_{k+1|1:k}^j\end{bmatrix},\begin{bmatrix}\mathbf{P}_{k,k|1:k}^j & \mathbf{P}_{k,k+1|1:k}^j\\\mathbf{P}_{k+1,k|1:k}^j & \mathbf{P}_{k+1,k+1|1:k}^j\end{bmatrix}\right).$$
$$(9)$$

This representation can be either exact, or a GMM best fit to the joint realizations of the filtered state vectors $\mathbf{X}_{k|1:k}(\omega)$ and $\mathbf{X}_{k+1|1:k}(\omega)$. The latter GMM best fit can be completed using the EM–BIC procedure (appendix B).

2) CONDITIONAL STATE-SPACE GMM

Since the joint distribution $p_{\mathbf{X}_k,\mathbf{X}_{k+1}|\mathbf{Y}_{1:k}}$ is a GMM [see (9)], the corresponding conditional distribution

$p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}$ is also a GMM and its pdf can be determined analytically. To see this, we start with the definition of the conditional distribution,

$$p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k\,|\,\mathbf{x}_{k+1},\mathbf{y}_{1:k}) = \frac{p_{\mathbf{X}_k,\mathbf{X}_{k+1}|\mathbf{Y}_{1:k}}(\mathbf{x}_k,\mathbf{x}_{k+1}\,|\,\mathbf{y}_{1:k})}{p_{\mathbf{X}_{k+1}|\mathbf{Y}_{1:k}}(\mathbf{x}_{k+1}\,|\,\mathbf{y}_{1:k})},$$

and substitute the expression for the numerator directly from (9) and the denominator by marginalizing $\mathbf{X}_k$ from (9) to obtain

$$p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k|\mathbf{x}_{k+1},\mathbf{y}_{1:k}) =$$
$$\frac{\sum_{j=1}^{M} \pi^j \times \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_k\\\mathbf{x}_{k+1}\end{bmatrix};\begin{bmatrix}\overline{\mathbf{x}}_{k|1:k}^j\\\overline{\mathbf{x}}_{k+1|1:k}^j\end{bmatrix},\begin{bmatrix}\mathbf{P}_{k,k|1:k}^j & \mathbf{P}_{k,k+1|1:k}^j\\\mathbf{P}_{k+1,k|1:k}^j & \mathbf{P}_{k+1,k+1|1:k}^j\end{bmatrix}\right)}{\sum_{j=1}^{M} \pi^j \times \mathcal{N}(\mathbf{x}_{k+1};\overline{\mathbf{x}}_{k+1|1:k}^j,\mathbf{P}_{k+1,k+1|1:k}^j)}.$$
$$(10)$$

The denominator in (10), containing no terms involving $\mathbf{x}_k$, simply normalizes the distribution. Using the expression for the multivariate normal pdf, we can expand each term in the numerator of (10) to obtain the expression for the conditional pdf $p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}$. The value of this conditional pdf for any smoothing state-space realization $\mathbf{x}_{k+1|1:K}^{(r)}$ (interpreted as a full-state observation) is

$$p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k\,|\,\mathbf{x}_{k+1|1:K}^{(r)},\mathbf{y}_{1:k}) = \sum_{j=1}^{M} \hat{\pi}^{j,(r)} \times \mathcal{N}(\mathbf{x}_k;\,\widehat{\overline{\mathbf{x}}}_k^{j,(r)},\hat{\mathbf{P}}_k^j),$$
$$(11)$$

where the conditional GMM components satisfy the following RTS-like equations:

$$\hat{\pi}^{j,(r)} \propto \pi^j \times \mathcal{N}(\mathbf{x}_{k+1|1:K}^{(r)};\overline{\mathbf{x}}_{k+1|1:k}^j,\mathbf{P}_{k+1,k+1|1:k}^j), \quad (12a)$$

$$\widehat{\overline{\mathbf{x}}}_k^{j,(r)} = \overline{\mathbf{x}}_{k|1:k}^j + \mathbf{K}^j[\mathbf{x}_{k+1|1:K}^{(r)} - \overline{\mathbf{x}}_{k+1|1:k}^j], \quad \text{and} \quad (12b)$$

$$\hat{\mathbf{P}}_k^j = \mathbf{P}_{k,k|1:k}^j - \mathbf{K}^j\mathbf{P}_{k+1,k|1:k}^j, \quad \text{with} \quad (12c)$$

$$\mathbf{K}^j = \mathbf{P}_{k,k+1|1:k}^j\mathbf{P}_{k+1,k+1|1:k}^{j+}, \quad (12d)$$

and $\cdot^+$ denotes the generalized inverse. These component update equations can also be derived using the property that GMM distributions are conjugate with respect to a linear Gaussian observation model (e.g., Ghanem and Spanos 2003). We note here that unlike the weights $\hat{\pi}^{j,(r)}$ and mean vectors $\widehat{\overline{\mathbf{x}}}_k^{j,(r)}$, the component covariance matrices $\hat{\mathbf{P}}_k^j$ do not depend on the specific realization $\mathbf{x}_{k+1|1:K}^{(r)}$ and can thus be computed independently. Hence, the superscript (r) is not used for them. A similar property holds

in the GMM–DO filter update (Sondergaard and Lermusiaux 2013a), wherein the posterior GMM covariances do not depend on the actual value of the observation.

The above RTS-like update equations (12) provide an analytical GMM representation of the conditional pdf $p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}$ and allow one to draw samples from it [i.e., (8)]. The result is the set of realizations $\mathbf{x}^{(r)}_{k|1:K}$ representing the smoothed pdf $p_{\mathbf{X}_k|\mathbf{Y}_{1:k}}$. This was our first objective, as required by the smoothing approach described in section 3b. Hence, the state-space GMM representation (9) of the *joint states* $\{\mathbf{X}_{k|1:k}(\omega), \mathbf{X}_{k+1|1:k}(\omega)\}$ and the above subsequent $k + 1$ to k inversion addresses the first of the two issues stated in section 3b.

d. The GMM–DO smoother with updates in the stochastic subspace

We now address the second issue discussed in section 3b and show how the subspace-GMM smoother uses the reduced-order decompositions (3) to solve the recursive smoothing in (7) directly in the dynamic low-dimensional stochastic subspace. To do so, the smoother exploits the joint reduced-order representations of the $k \leftrightarrow k + 1$ filtered variables $\{\mathbf{X}_{k|1:k}(\omega), \mathbf{X}_{k+1|1:k}(\omega)\}$, the smoothed variable $\mathbf{X}_{k+1|1:K}(\omega)$ with decomposition (4), and its realizations $\mathbf{x}^{(r)}_{k+1|1:K}$ given by

$$\mathbf{x}^{(r)}_{k+1|1:K} = \overline{\mathbf{x}}_{k+1|1:K} + \boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\phi}^{(r)}_{k+1|1:K}, \quad r = 1, 2, \ldots, N_r. \tag{13}$$

From (3), the joint filtered state-space variables $\{\mathbf{X}_{k|1:k}(\omega), \mathbf{X}_{k+1|1:k}(\omega)\}$ are related to the joint filtered stochastic coefficients $\{\boldsymbol{\Phi}_{k|1:k}(\omega), \boldsymbol{\Phi}_{k+1|1:k}(\omega)\}$ through the augmented affine transformation:

$$\mathbf{X}_{k,k+1|1:k}(\omega) := \begin{bmatrix} \mathbf{X}_{k|1:k}(\omega) \\ \mathbf{X}_{k+1|1:k}(\omega) \end{bmatrix} = \underbrace{\begin{bmatrix} \overline{\mathbf{x}}_{k|1:k} \\ \overline{\mathbf{x}}_{k+1|1:k} \end{bmatrix}}_{\overline{\mathbf{x}}_{k,k+1|1:k}}$$

$$+ \underbrace{\begin{bmatrix} \boldsymbol{\mathcal{X}}_k & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mathcal{X}}_{k+1} \end{bmatrix}}_{\boldsymbol{\mathcal{X}}_{k,k+1}} \times \underbrace{\begin{bmatrix} \boldsymbol{\Phi}_{k|1:k}(\omega) \\ \boldsymbol{\Phi}_{k+1|1:k}(\omega) \end{bmatrix}}_{\boldsymbol{\Phi}_{k,k+1|1:k}(\omega)}. \tag{14}$$

Equation (14) is the reduced-order decomposition of $\mathbf{X}_{k,k+1|1:k}(\omega)$, the $k \leftrightarrow k + 1$ *joint state-space vector* formed by augmenting $\mathbf{X}_{k|1:k}(\omega)$ with $\mathbf{X}_{k+1|1:k}(\omega)$. As noted in (14), we use the notation $\cdot_{k,k+1|1:l}$ to indicate vector augmentation and $\cdot_{k,k+1}$ for block matrix augmentation. The joint decomposition (14) exactly reproduces the individual representations of both $\mathbf{X}_{k|1:k}(\omega)$ and $\mathbf{X}_{k+1|1:k}(\omega)$. It is thus as accurate as each of these representations. In fact, it can be shown that when (3) is a DO decomposition (see appendix A), (14) is the dynamic Karhunen–Loéve (KL) expansion (Sapsis and Lermusiaux 2009; Ghanem and Spanos 2003) of $\mathbf{X}_{k,k+1|1:k}(\omega)$. The affine mapping (14) will be crucial in the next steps, where we first proceed as in section 3c but for the subspace and then prove the equivalence between smoothing updates in the full state space and in the stochastic subspace.

1) JOINT SUBSPACE GMM

Similar to the joint state-space pdf (9), the subspace-GMM smoother represents the augmented vector of filtered subspace coefficients $\boldsymbol{\Phi}_{k,k+1|1:k}(\omega)$ (and its joint filtered subspace pdf $p_{\boldsymbol{\Phi}_k,\boldsymbol{\Phi}_{k+1}|\mathbf{Y}_{1:k}}$) by a GMM. This can again be either exact (for truly GMM distributions) or in the sense of a EM–BIC best fit to joint realizations $\boldsymbol{\phi}^{(r)}_{k,k+1|1:k}$. The result is

$$p_{\boldsymbol{\Phi}_k,\boldsymbol{\Phi}_{k+1}|\mathbf{Y}_{1:k}}(\boldsymbol{\phi}_k, \boldsymbol{\phi}_{k+1} \,|\, \mathbf{y}_{1:k}) = \sum_{j=1}^{M} \pi^j \times \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\phi}_k \\ \boldsymbol{\phi}_{k+1} \end{bmatrix}; \begin{bmatrix} \overline{\boldsymbol{\mu}}^j_{k|1:k} \\ \overline{\boldsymbol{\mu}}^j_{k+1|1:k} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}^j_{k,k|1:k} & \boldsymbol{\Sigma}^j_{k,k+1|1:k} \\ \boldsymbol{\Sigma}^j_{k+1,k|1:k} & \boldsymbol{\Sigma}^j_{k+1,k+1|1:k} \end{bmatrix} \right), \tag{15}$$

where the component weights $\pi^j(\geq 0)$ sum to unity; $\overline{\boldsymbol{\mu}}^j_{k|1:k}$ and $\overline{\boldsymbol{\mu}}^j_{k+1|1:k} \in \mathbb{R}^s$ are the means of the jth mixture component of $\boldsymbol{\Phi}_{k|1:k}$ and $\boldsymbol{\Phi}_{k+1|1:k}$, respectively; $\boldsymbol{\Sigma}^j_{k,k|1:k}$ and $\boldsymbol{\Sigma}^j_{k+1,k+1|1:k}$ are the corresponding covariance matrices; and $\boldsymbol{\Sigma}^j_{k,k+1|1:k}$ and $\boldsymbol{\Sigma}^j_{k+1,k|1:k}$ are the corresponding cross-covariance matrices.

2) EQUIVALENCE BETWEEN STATE-SPACE AND SUBSPACE GMMS

Now that the joint pdf $p_{\boldsymbol{\Phi}_k,\boldsymbol{\Phi}_{k+1}|\mathbf{Y}_{1:k}}$ is specified, the joint affine transformation (14) allows us to determine the

joint filtered state-space pdf $p_{\mathbf{X}_k,\mathbf{X}_{k+1}|\mathbf{Y}_{1:k}}$. Indeed, (14) implies that under (15), the distribution $p_{\mathbf{X}_k,\mathbf{X}_{k+1}|\mathbf{Y}_{1:k}}$ is also a GMM (Sondergaard and Lermusiaux 2013a) and is given by (9), where, for any $\alpha, \beta \in \{k, k + 1\}$, the component mean vectors and covariance matrices, respectively, satisfy

$$\overline{\mathbf{x}}^j_{\alpha|1:k} = \overline{\mathbf{x}}_{\alpha|1:k} + \boldsymbol{\mathcal{X}}_\alpha \overline{\boldsymbol{\mu}}^j_{\alpha|1:k} \quad \text{and} \tag{16a}$$

$$\mathbf{P}^j_{\alpha,\beta|1:k} = \boldsymbol{\mathcal{X}}_\alpha \boldsymbol{\Sigma}^j_{\alpha,\beta|1:k} \boldsymbol{\mathcal{X}}^{\mathrm{T}}_\beta. \tag{16b}$$

This relationship between the GMM parameters of the joint subspace and joint state space is what allows the

subspace-GMM smoother to solve the recursion (7) efficiently. It is used in the following key theorem to demonstrate the equivalence between the full-space and subspace smoothing updates, hence laying the foundation of the GMM–DO smoother. The first part of the theorem relates the components of the conditional state-space pdf (11) to those of the joint-subspace pdf (15). The second part outlines a procedure to implicitly draw samples from the pdf in (11), when $\mathbf{x}_{k+1|1:K}^{(r)}$ has a reduced-order decomposition. The various steps of the smoother are described in section 3e. All equations for the GMM–DO smoother are summarized in Table 2.

Theorem 1. *Let $\mathbf{X}_{k|1:k}(\omega)$ and $\mathbf{X}_{k+1|1:k}(\omega) \in \mathbb{R}^{N_\mathbf{x}}$, respectively, denote the filtered state vectors of the stochastic dynamical system (1a) at times t_k and t_{k+1}, conditioned on observations $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_k$. Let $\mathbf{X}_{k|1:k}(\omega)$ and $\mathbf{X}_{k+1|1:k}(\omega)$ also satisfy the augmented reduced-order decomposition (14), where $\overline{\mathbf{x}}_{k|1:k} = E_\omega[\mathbf{X}_{k|1:k}(\omega)]$ and $\overline{\mathbf{x}}_{k+1|1:k} = E_\omega[\mathbf{X}_{k+1|1:k}(\omega)]$; $\boldsymbol{\mathcal{X}}_k, \boldsymbol{\mathcal{X}}_{k+1} \in \mathbb{R}^{N_\mathbf{x} \times s}$ are the matrices of orthonormal modes ($\boldsymbol{\mathcal{X}}_k^\mathrm{T} \boldsymbol{\mathcal{X}}_k = \mathbf{I}$, $\boldsymbol{\mathcal{X}}_{k+1}^\mathrm{T} \boldsymbol{\mathcal{X}}_{k+1} = \mathbf{I}$); and $\boldsymbol{\Phi}_{k|1:k}(\omega)$ and $\boldsymbol{\Phi}_{k+1|1:k}(\omega) \in \mathbb{R}^s$ are zero-mean vectors of stochastic coefficients whose joint pdf is the multivariate GMM (15). Let $\mathbf{x}_{k+1|1:K}^{(r)} \in \mathbb{R}^{N_\mathbf{x}}$ be a smoothed state realization of the form*

$$\mathbf{x}_{k+1|1:K}^{(r)} = \overline{\mathbf{x}}_{k+1|1:K} + \boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\phi}_{k+1|1:K}^{(r)} \qquad (17)$$

for some realization index r, $\overline{\mathbf{x}}_{k+1|1:K} \in \mathbb{R}^{N_\mathbf{x}}$, and $\boldsymbol{\phi}_{k+1|1:K}^{(r)} \in \mathbb{R}^s$. Then:

1) *The $k \leftrightarrow k + 1$ conditional state pdf $p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k \mid \mathbf{x}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k})$ is a multivariate GMM whose samples are given by (11) and whose components satisfy*

$$\hat{\pi}^{j,(r)} \propto \pi^j \times \mathcal{N}(\boldsymbol{\phi}_{k+1|1:K}^{(r)}; \overline{\boldsymbol{\mu}}_{k+1|1:k}^j + \boldsymbol{\mathcal{X}}_{k+1}^\mathrm{T}(\overline{\mathbf{x}}_{k+1|1:k} - \overline{\mathbf{x}}_{k+1|1:K}), \boldsymbol{\Sigma}_{k+1,k+1|1:k}^j), \quad (18a)$$

$$\widehat{\overline{\mathbf{x}}}_k^{j,(r)} = \overline{\mathbf{x}}_{k|1:k} + \boldsymbol{\mathcal{X}}_k\{\overline{\boldsymbol{\mu}}_{k|1:k}^j + \tilde{\mathbf{K}}^j[\boldsymbol{\phi}_{k+1|1:K}^{(r)} - \overline{\boldsymbol{\mu}}_{k+1|1:k}^j - \boldsymbol{\mathcal{X}}_{k+1}^\mathrm{T}(\overline{\mathbf{x}}_{k+1|1:k} - \overline{\mathbf{x}}_{k+1|1:K})]\}, \quad (18b)$$

and

$$\hat{\mathbf{P}}_k^j = \boldsymbol{\mathcal{X}}_k(\boldsymbol{\Sigma}_{k,k|1:k}^j - \tilde{\mathbf{K}}^j \boldsymbol{\Sigma}_{k+1,k|1:k}^j)\boldsymbol{\mathcal{X}}_k^\mathrm{T}, \qquad (18c)$$

and the subspace component gain $\tilde{\mathbf{K}}^j$ is

$$\tilde{\mathbf{K}}^j = \boldsymbol{\Sigma}_{k,k+1|1:k}^j \boldsymbol{\Sigma}_{k+1,k+1|1:k}^{j\,+}, \qquad (19)$$

where $\cdot^+$ denotes the generalized inverse.

2) *Defining $\mathbf{x}_k := \overline{\mathbf{x}}_{k|1:k} + \boldsymbol{\mathcal{X}}_k\boldsymbol{\phi}_k$, the $k \leftrightarrow k + 1$ conditional full-state pdf is equivalent to the corresponding stochastic subspace pdf, i.e.,*

$$p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k \mid \mathbf{x}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k})$$
$$= p_{\boldsymbol{\Phi}_k|\boldsymbol{\Phi}_{k+1},\mathbf{Y}_{1:k}}(\boldsymbol{\phi}_k \mid \boldsymbol{\phi}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k}),$$

and this $k \leftrightarrow k + 1$ conditional subspace pdf $p_{\boldsymbol{\Phi}_k|\boldsymbol{\Phi}_{k+1},\mathbf{Y}_{1:k}}$ is given by the GMM

$$p_{\boldsymbol{\Phi}_k|\boldsymbol{\Phi}_{k+1},\mathbf{Y}_{1:k}}(\boldsymbol{\phi}_k \mid \boldsymbol{\phi}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k})$$
$$= \sum_{j=1}^M \hat{\pi}^{j,(r)} \times \mathcal{N}(\boldsymbol{\phi}_k; \tilde{\boldsymbol{\mu}}_k^{j,(r)}, \tilde{\boldsymbol{\Sigma}}_k^j), \qquad (20)$$

where

$$\tilde{\boldsymbol{\mu}}_k^{j,(r)} = \overline{\boldsymbol{\mu}}_{k|1:k}^j + \tilde{\mathbf{K}}^j[\boldsymbol{\phi}_{k+1|1:K}^{(r)} - \overline{\boldsymbol{\mu}}_{k+1|1:k}^j - \boldsymbol{\mathcal{X}}_{k+1}^\mathrm{T}(\overline{\mathbf{x}}_{k+1|1:k} - \overline{\mathbf{x}}_{k+1|1:K})] \quad \text{and} \qquad (21a)$$

$$\tilde{\boldsymbol{\Sigma}}_k^j = \boldsymbol{\Sigma}_{k,k|1:k}^j - \tilde{\mathbf{K}}^j \boldsymbol{\Sigma}_{k+1,k|1:k}^j. \qquad (21b)$$

Proof. 1) The augmented decomposition (14) is an affine mapping between the state-space vectors $\{\mathbf{X}_{k|1:k}(\omega), \mathbf{X}_{k+1|1:k}(\omega)\}$ and the subspace coefficients $\boldsymbol{\Phi}_{k|1:k}(\omega), \boldsymbol{\Phi}_{k+1|1:k}(\omega)$. Therefore, (15) implies that the joint filtered distribution $p_{\mathbf{X}_k,\mathbf{X}_{k+1}|\mathbf{Y}_{1:k}}$ is also a GMM with pdf (9), and its components given by (16). As a result, the conditional distribution $p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k \mid \mathbf{x}_{k+1|1:K}^{(r)}, \mathbf{Y}_{1:k})$ is also a GMM; it is given by (11), and its mixture components satisfy (12). Next, we start from these results (12), substitute (16) and (17), and simplify the expressions. This allows us to link the state-space component weights, means, covariances, and gains to their subspace counterparts and hence derive (18) and (19).

(i) Component weights—(18a)
For $j = 1, 2, \ldots, M$, using (12a), we obtain

$$\hat{\pi}^{j,(r)} \propto \pi^j \times \mathcal{N}(\mathbf{x}_{k+1|1:K}^{(r)}; \overline{\mathbf{x}}_{k+1|1:k}^j, \mathbf{P}_{k+1,k+1|1:k}^j)$$
$$= \pi^j \times \mathcal{N}(\overline{\mathbf{x}}_{k+1|1:K} + \boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\phi}_{k+1|1:K}^{(r)}; \overline{\mathbf{x}}_{k+1|1:k}^j, \boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\Sigma}_{k+1,k+1|1:k}^j\boldsymbol{\mathcal{X}}_{k+1}^\mathrm{T})$$
$$= \frac{\pi^j}{\sqrt{\det\!^*(2\pi\boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\Sigma}_{k+1,k+1|1:k}^j\boldsymbol{\mathcal{X}}_{k+1}^\mathrm{T})}} \times e^{\left\{-(1/2)[\overline{\mathbf{x}}_{k+1|1:K} + \boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\phi}_{k+1|1:K}^{(r)} - \overline{\mathbf{x}}_{k+1|1:k}^j]^\mathrm{T}(\boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\Sigma}_{k+1,k+1|1:k}^j\boldsymbol{\mathcal{X}}_{k+1}^\top)^+[\overline{\mathbf{x}}_{k+1|1:K} + \boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\phi}_{k+1|1:K}^{(r)} - \overline{\mathbf{x}}_{k+1|1:k}^j]\right\}}.$$

$$(22)$$

Here, det* denotes the pseudo-determinant. Observe that $\det^*(2\pi\boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\Sigma}^j_{k+1,k+1|1:k}\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}) = \det^*(2\pi\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}\boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\Sigma}^j_{k+1,k+1|1:k}) = \det^*(2\pi\boldsymbol{\Sigma}^j_{k+1,k+1|1:k})$ and

that $(\boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\Sigma}^j_{k+1,k+1|1:k}\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1})^+ = \boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\Sigma}^{j^+}_{k+1,k+1|1:k}\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}$. Substituting these expressions into (22), we obtain

$$
\begin{aligned}
\hat{\pi}^{j,(r)} &\propto \frac{\pi^j}{\sqrt{\det^*(2\pi\boldsymbol{\Sigma}^j_{k+1,k+1|1:k})}} \times e^{\left\{-(1/2)[\bar{\mathbf{x}}_{k+1|1:K}+\boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\phi}^{(r)}_{k+1|1:K}-\bar{\mathbf{x}}^j_{k+1|1:k}]^{\top}\boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\Sigma}^{j^+}_{k+1,k+1|1:k}\boldsymbol{\mathcal{X}}^{\top}_{k+1}[\bar{\mathbf{x}}_{k+1|1:K}+\boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\phi}^{(r)}_{k+1|1:K}-\bar{\mathbf{x}}^j_{k+1|1:k}]\right\}} \\
&= \frac{\pi^j}{\sqrt{\det^*(2\pi\boldsymbol{\Sigma}^j_{k+1,k+1|1:k})}} \times e^{\left\{-(1/2)(\boldsymbol{\phi}^{(r)}_{k+1|1:K}-\boldsymbol{\mathcal{X}}^{\top}_{k+1}(\bar{\mathbf{x}}^j_{k+1|1:k}-\bar{\mathbf{x}}_{k+1|1:K}))^{\top}\boldsymbol{\Sigma}^{j^+}_{k+1,k+1|1:k}[\boldsymbol{\phi}^{(r)}_{k+1|1:K}-\boldsymbol{\mathcal{X}}^{\top}_{k+1}(\bar{\mathbf{x}}^j_{k+1|1:k}-\bar{\mathbf{x}}_{k+1|1:K})]\right\}} \\
&= \pi^j \times \mathcal{N}(\boldsymbol{\phi}^{(r)}_{k+1|1:K}; \boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}(\bar{\mathbf{x}}^j_{k+1|1:k}-\bar{\mathbf{x}}_{k+1|1:K}), \boldsymbol{\Sigma}^j_{k+1,k+1|1:k}).
\end{aligned}
\tag{23}
$$

Finally, substituting the expression for $\bar{\mathbf{x}}^j_{k+1|1:k}$ from (16a) into (23) and using $\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}\boldsymbol{\mathcal{X}}_{k+1} = \mathbf{I}$ again gives us (18a).

(ii) Subspace component gain—(19)

To obtain the subspace component gain $\tilde{\mathbf{K}}^j$ defined by $\mathbf{K}^j = \boldsymbol{\mathcal{X}}_k\tilde{\mathbf{K}}^j\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}$, we start from the full-space component gain $\mathbf{K}^j$ in (12d) and substitute the expressions for $\mathbf{P}^j_{k,k+1|1:k}$ and $\mathbf{P}^j_{k+1,k+1|1:k}$ given in (16b):

$$
\begin{aligned}
\mathbf{K}^j &= \mathbf{P}^j_{k,k+1|1:k} \times \mathbf{P}^{j^+}_{k+1,k+1|1:k} \\
&= \boldsymbol{\mathcal{X}}_k\boldsymbol{\Sigma}^j_{k,k+1|1:k}\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1} \times \boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\Sigma}^{j^+}_{k+1,k+1|1:k}\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1} \\
&= \boldsymbol{\mathcal{X}}_k\boldsymbol{\Sigma}^j_{k,k+1|1:k}\boldsymbol{\Sigma}^{j^+}_{k+1,k+1|1:k}\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}.
\end{aligned}
\tag{24}
$$

This yields $\tilde{\mathbf{K}}^j = \boldsymbol{\Sigma}^j_{k,k+1|1:k}\boldsymbol{\Sigma}^{j^+}_{k+1,k+1|1:k}$ and thus (19).

(iii) Component mean vectors—(18b)

Substituting $\bar{\mathbf{x}}^j_{k|1:k}$ and $\bar{\mathbf{x}}^j_{k+1|1:k}$ from (16a), $\mathbf{x}^{(r)}_{k+1|1:K}$ from (17), and $\mathbf{K}^j$ from (24) into (12b), we obtain

$$
\begin{aligned}
\widehat{\bar{\mathbf{x}}}^{j,(r)}_k &= \bar{\mathbf{x}}^j_{k|1:k} + \mathbf{K}^j[\mathbf{x}^{(r)}_{k+1|1:K}-\bar{\mathbf{x}}^j_{k+1|1:k}] \\
&= \bar{\mathbf{x}}_{k|1:k} + \boldsymbol{\mathcal{X}}_k\bar{\boldsymbol{\mu}}^j_{k|1:k} + \boldsymbol{\mathcal{X}}_k\tilde{\mathbf{K}}^j\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}[\bar{\mathbf{x}}_{k+1|1:K} \\
&\quad + \boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\phi}^{(r)}_{k+1|1:K}-\bar{\mathbf{x}}_{k+1|1:k}-\boldsymbol{\mathcal{X}}_{k+1}\bar{\boldsymbol{\mu}}^j_{k+1|1:k}] \\
&= \bar{\mathbf{x}}_{k|1:k} + \boldsymbol{\mathcal{X}}_k\{\bar{\boldsymbol{\mu}}^j_{k|1:k} + \tilde{\mathbf{K}}^j[\boldsymbol{\phi}^{(r)}_{k+1|1:K} \\
&\quad - \bar{\boldsymbol{\mu}}^j_{k+1|1:k}-\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}(\bar{\mathbf{x}}_{k+1|1:k}-\bar{\mathbf{x}}_{k+1|1:K})]\},
\end{aligned}
$$

which is the component mean vector given by (18b).

(iv) Component covariance matrices—(18c)

Substituting $\mathbf{P}^j_{k,k|1:k}$ and $\mathbf{P}^j_{k+1,k|1:k}$ from (16b) into the full-state component covariances (12c), we obtain

$$
\begin{aligned}
\hat{\mathbf{P}}^j_k &= \mathbf{P}^j_{k,k|1:k} - \mathbf{K}^j \times \mathbf{P}^j_{k+1,k|1:k} \\
&= \boldsymbol{\mathcal{X}}_k\boldsymbol{\Sigma}^j_{k,k|1:k}\boldsymbol{\mathcal{X}}^{\mathrm{T}}_k - \boldsymbol{\mathcal{X}}_k\tilde{\mathbf{K}}^j\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1} \times \boldsymbol{\mathcal{X}}_{k+1}\boldsymbol{\Sigma}^j_{k+1,k|1:k}\boldsymbol{\mathcal{X}}^{\mathrm{T}}_k \\
&= \boldsymbol{\mathcal{X}}_k(\boldsymbol{\Sigma}^j_{k,k|1:k} - \tilde{\mathbf{K}}^j\boldsymbol{\Sigma}^j_{k+1,k|1:k})\boldsymbol{\mathcal{X}}^{\mathrm{T}}_k.
\end{aligned}
$$

This yields (18c) and completes the proof of part 1.

2) To prove this part of the theorem, we start from $\mathbf{x}_k = \bar{\mathbf{x}}_{k|1:k} + \boldsymbol{\mathcal{X}}_k\boldsymbol{\phi}_k$ and simplify the expression for the $k \leftrightarrow k+1$ conditional full-state distribution $p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}[\mathbf{x}_k \,|\, \mathbf{x}^{(r)}_{k+1|1:K}, \mathbf{y}_{1:k}]$ obtained in (11) until we arrive at the $k \leftrightarrow k+1$ conditional stochastic subspace pdf $p_{\boldsymbol{\Phi}_k|\boldsymbol{\Phi}_{k+1},\mathbf{Y}_{1:k}}[\boldsymbol{\phi}_k \,|\, \boldsymbol{\phi}^{(r)}_{k+1|1:K}, \mathbf{y}_{1:k}]$ given by (20). From (11), we have

$$
\begin{aligned}
&p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k \,|\, \mathbf{x}^{(r)}_{k+1|1:K}, \mathbf{y}_{1:k}) \\
&\quad = \sum_{j=1}^{M} \hat{\pi}^{j,(r)} \times \mathcal{N}(\mathbf{x}_k; \widehat{\bar{\mathbf{x}}}^{j,(r)}_k, \hat{\mathbf{P}}^j_k).
\end{aligned}
$$

Substituting $\widehat{\bar{\mathbf{x}}}^{j,(r)}_k$ from (18b), we obtain

$$
\begin{aligned}
&p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k \,|\, \mathbf{x}^{(r)}_{k+1|1:K}, \mathbf{y}_{1:k}) \\
&= \sum_{j=1}^{M} \hat{\pi}^{j,(r)} \times \mathcal{N}(\bar{\mathbf{x}}_{k|1:k}+\boldsymbol{\mathcal{X}}_k\boldsymbol{\phi}_k; \bar{\mathbf{x}}_{k|1:k}+\boldsymbol{\mathcal{X}}_k\{\bar{\boldsymbol{\mu}}^j_{k|1:k}+\tilde{\mathbf{K}}^j[\boldsymbol{\phi}^{(r)}_{k+1|1:K}-\bar{\boldsymbol{\mu}}^j_{k+1|1:k}-\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}(\bar{\mathbf{x}}_{k+1|1:k}-\bar{\mathbf{x}}_{k+1|1:K})]\}, \hat{\mathbf{P}}^j_k) \\
&= \sum_{j=1}^{M} \hat{\pi}^{j,(r)} \times \mathcal{N}(\boldsymbol{\mathcal{X}}_k\boldsymbol{\phi}_k; \boldsymbol{\mathcal{X}}_k\{\bar{\boldsymbol{\mu}}^j_{k|1:k}+\tilde{\mathbf{K}}^j[\boldsymbol{\phi}^{(r)}_{k+1|1:K}-\bar{\boldsymbol{\mu}}^j_{k+1|1:k}-\boldsymbol{\mathcal{X}}^{\mathrm{T}}_{k+1}(\bar{\mathbf{x}}_{k+1|1:k}-\bar{\mathbf{x}}_{k+1|1:K})]\}, \hat{\mathbf{P}}^j_k).
\end{aligned}
\tag{25}
$$

Using (21a), (25) reduces to

$$p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}[\mathbf{x}_k \,|\, \mathbf{x}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k}] = \sum_{j=1}^{M} \hat{\pi}^{j,(r)} \times \mathcal{N}(\mathcal{X}_k \boldsymbol{\phi}_k; \mathcal{X}_k \tilde{\boldsymbol{\mu}}_k^{j,(r)}, \hat{\mathbf{P}}_k^j).$$

Substituting $\hat{\mathbf{P}}_k^j$ from (18c), and setting $\tilde{\boldsymbol{\Sigma}}_k^j = \boldsymbol{\Sigma}_{k,k|1:k}^j - \tilde{\mathbf{K}}^j \boldsymbol{\Sigma}_{k+1,k|1:k}^j$, we obtain

$$p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k \,|\, \mathbf{x}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k}) = \sum_{j=1}^{M} \hat{\pi}^{j,(r)} \times \mathcal{N}(\mathcal{X}_k \boldsymbol{\phi}_k; \mathcal{X}_k \tilde{\boldsymbol{\mu}}_k^{j,(r)}, \mathcal{X}_k \tilde{\boldsymbol{\Sigma}}_k^j \mathcal{X}_k^{\mathrm{T}})$$

$$= \sum_{j=1}^{M} \frac{\hat{\pi}^{j,(r)}}{\sqrt{\det^*(2\pi \mathcal{X}_k \tilde{\boldsymbol{\Sigma}}_k^j \mathcal{X}_k^{\mathrm{T}})}} e^{\left\{-(1/2)\left[\mathcal{X}_k \boldsymbol{\phi}_k - \mathcal{X}_k \tilde{\boldsymbol{\mu}}_k^{j,(r)}\right]^{\mathrm{T}} \left(\mathcal{X}_k \tilde{\boldsymbol{\Sigma}}_k^j \mathcal{X}_k^{\top}\right)^{+} \left[\mathcal{X}_k \boldsymbol{\phi}_k - \mathcal{X}_k \tilde{\boldsymbol{\mu}}_k^{j,(r)}\right]\right\}}. \quad (26)$$

As seen in the proof of 1(i), $\det^*(2\pi \mathcal{X}_k \tilde{\boldsymbol{\Sigma}}_k^j \mathcal{X}_k^{\mathrm{T}}) = \det^*(2\pi \tilde{\boldsymbol{\Sigma}}_k^j)$, and $(\mathcal{X}_k \tilde{\boldsymbol{\Sigma}}_k^j \mathcal{X}_k^{\mathrm{T}})^{+} = \mathcal{X}_k \tilde{\boldsymbol{\Sigma}}_k^{j^+} \mathcal{X}_k^{\mathrm{T}}$. Substituting these expressions into (26), and using $\mathcal{X}_k^{\mathrm{T}} \mathcal{X}_k = \mathbf{I}$, we obtain

$$p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\mathbf{x}_k \,|\, \mathbf{x}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k}) = \sum_{j=1}^{M} \frac{\hat{\pi}^{j,(r)}}{\sqrt{\det^*(2\pi \tilde{\boldsymbol{\Sigma}}_k^j)}} e^{\left\{-(1/2)\left[\boldsymbol{\phi}_k - \tilde{\boldsymbol{\mu}}_k^{j,(r)}\right]^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}}_k^{j^+} \left[\boldsymbol{\phi}_k - \tilde{\boldsymbol{\mu}}_k^{j,(r)}\right]\right\}}$$

$$= \sum_{j=1}^{M} \hat{\pi}^{j,(r)} \times \mathcal{N}(\boldsymbol{\phi}_k; \tilde{\boldsymbol{\mu}}_k^{j,(r)}, \tilde{\boldsymbol{\Sigma}}_k^j) = p_{\boldsymbol{\Phi}_k|\boldsymbol{\Phi}_{k+1},\mathbf{Y}_{1:k}}(\boldsymbol{\phi}_k \,|\, \boldsymbol{\phi}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k}).$$

This completes the proof of part 2.∎

Theorem 1 outlines a procedure to efficiently draw a sample from the conditional pdf $p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\cdot \,|\, \mathbf{x}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k})$ [and therefore from the smoothed pdf $p_{\mathbf{X}_k|\mathbf{Y}_{1:K}}(\mathbf{x}_k \,|\, \mathbf{y}_{1:K})$], when $\mathbf{x}_{k+1|1:K}^{(r)}$ has a reduced-order decomposition of the form (17). *Although strictly unnecessary for the subspace-GMM smoother*, a sample from $p_{\mathbf{X}_k|\mathbf{X}_{k+1},\mathbf{Y}_{1:k}}(\cdot \,|\, \mathbf{x}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k})$ can be generated by drawing $\tilde{\boldsymbol{\phi}}_{k|1:K}^{(r)}$ from the *subspace* conditional pdf $p_{\boldsymbol{\Phi}_k|\boldsymbol{\Phi}_{k+1},\mathbf{Y}_{1:k}}(\boldsymbol{\phi}_k \,|\, \boldsymbol{\phi}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k})$ given by (20),

$$\tilde{\boldsymbol{\phi}}_{k|1:K}^{(r)} \sim p_{\boldsymbol{\Phi}_k|\boldsymbol{\Phi}_{k+1},\mathbf{Y}_{1:k}}(\cdot \,|\, \boldsymbol{\phi}_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k}), \quad (27)$$

and transforming these $\tilde{\boldsymbol{\phi}}_{k|1:K}^{(r)}$ into $\bar{\mathbf{x}}_{k|1:k} + \mathcal{X}_k \tilde{\boldsymbol{\phi}}_{k|1:K}^{(r)} = \mathbf{x}_{k|1:K}^{(r)}$. Instead, the subspace-GMM smoother performs all the computations for evaluating and drawing a sample from the conditional GMM distribution (20) strictly in the evolving stochastic subspace. The component state-space mean vectors $\widehat{\bar{\mathbf{x}}}_k^{j,(r)}$, covariance matrices $\hat{\mathbf{P}}_k^j$, and smoothed realizations $\mathbf{x}_{k|1:K}^{(r)}$ are never explicitly used nor calculated by the smoother. We now have all the elements necessary to present the smoother update equations. These are provided next.

e. *The subspace-GMM smoother: Summary of equations, algorithm, and computational cost*

As the smoothed reduced-order decomposition in (4) indicates, the goal of the subspace-GMM smoother is to recursively determine $\bar{\mathbf{x}}_{k|1:K}$ and the subspace realizations $\{\boldsymbol{\phi}_{k|1:K}^{(r)}\}_{r=1}^{N_r}$ representing $\boldsymbol{\Phi}_{k|1:K}(\omega)$, using their respective counterparts at time t_{k+1}. To obtain these quantities, the overall smoothing procedure consists of three steps or passes: forward filtering, joint subspaces GMM fitting, and backward smoothing.

1) FORWARD FILTERING PASS

A nonlinear, non-Gaussian filter is first used between times $t = 0$ and $t = t_K$ to sequentially assimilate the observations $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_K$ as they arrive. For $l = k, k + 1$, the filter provides the quantities $\bar{\mathbf{x}}_{l|1:k}$ and the realizations $\{\boldsymbol{\phi}_{l|1:k}^{(r)}\}_{r=1}^{N_r}$ of $\boldsymbol{\Phi}_{l|1:k}(\omega)$ that form the reduced-order decomposition in (3) of the state vectors $\mathbf{X}_{l|1:k}(\omega)$. The sets of subspace ensemble members $\{\boldsymbol{\phi}_{k|1:k}^{(r)}\}_{r=1}^{N_r}$ and $\{\boldsymbol{\phi}_{k+1|1:k}^{(r)}\}_{r=1}^{N_r}$ computed during this filtering run are stored, as needed for the second step of the algorithm, the GMM-fitting pass. The matrices $\mathcal{X}_k$, and the mean vectors $\bar{\mathbf{x}}_{k|1:k}, \bar{\mathbf{x}}_{k+1|1:k}$ are also stored. In this paper, the GMM–DO filter (Sondergaard and Lermusiaux 2013a) is utilized for the filtering pass (see appendix A).

2) JOINT SUBSPACES GMM-FITTING PASS

A crucial component of the new subspace-GMM smoother is the joint pdf in (15) of the $k \leftrightarrow k + 1$ pair of filtered stochastic coefficients, $\boldsymbol{\Phi}_{k|1:k}(\omega)$ and $\boldsymbol{\Phi}_{k+1|1:k}(\omega)$. To determine this pdf, the smoother optimally fits a

GMM distribution to the realizations of $\Phi_{k,k+1|1:k}(\omega)$, defined by (14). These realizations, denoted by $\phi_{k,k+1|1:k}^{(r)}$, $r = 1, 2, \ldots, N_r$ are assembled by augmenting the *corresponding* realizations of $\Phi_{k|1:k}(\omega)$ and $\Phi_{k+1|1:k}(\omega)$; i.e.,

$$\phi_{k,k+1|1:k}^{(r)} = \begin{bmatrix} \phi_{k|1:k}^{(r)} \\ \phi_{k+1|1:k}^{(r)} \end{bmatrix}. \tag{28}$$

Observe that the quantities $\Phi_{k|1:k}(\omega)$ and $\Phi_{k+1|1:k}(\omega)$ are conditioned only on the past observations $\mathbf{Y}_{1:k}$, and do not depend on the future observations $\mathbf{Y}_{k+1:K}$. Consequently, they can be determined by a filtering run; that is, no smoothing is required. This is why the above filter run forms the first pass of the subspace-GMM smoother.

Following the filtering run, for each k, a GMM is fit to the set of ensemble realizations $\{\phi_{k,k+1|1:k}^{(r)}\}_{r=1}^{N_r}$. The resulting GMM best represents the set of ensemble realizations in the 2s-dimensional *joint filtered subspace*. The total number of GMM-fitting operations in this step is $K - 1$. The EM–BIC scheme (appendix B) is the GMM-fitting procedure used in this paper.

3) BACKWARD SMOOTHING PASS

For each realization index r, the subspace-GMM smoother draws a sample $\tilde{\phi}_{k|1:K}^{(r)}$ from the conditional subspace distribution $p_{\Phi_k|\Phi_{k+1}, \mathbf{Y}_{1:k}}(\cdot \mid \phi_{k+1|1:K}^{(r)}, \mathbf{y}_{1:k})$, given by (20). To sample from this pdf, a two-step approach is followed. First, M independent samples are drawn, one from each Gaussian component of (20). Next, exactly one of these M samples is accepted, where the probability of accepting any given sample equals the weight $\hat{\pi}_k^j$ of the Gaussian component that generated it. Then, the smoothed mean state $\bar{\mathbf{x}}_{k|1:K}$ is computed as

$$\bar{\mathbf{x}}_{k|1:K} = \bar{\mathbf{x}}_{k|1:k} + \mathcal{X}_k \times \left[\frac{1}{N_r} \sum_{r=1}^{N_r} \tilde{\phi}_{k|1:K}^{(r)} \right], \tag{29}$$

and the ensemble of the zero-mean vectors of the stochastic coefficients $\phi_{k|1:K}^{(r)}$ is given by

$$\phi_{k|1:K}^{(r)} = \tilde{\phi}_{k|1:K}^{(r)} - \frac{1}{N_r} \sum_{r=1}^{N_r} \tilde{\phi}_{k|1:K}^{(r)}, \quad r = 1, 2, \ldots, N_r. \tag{30}$$

The smoothed $\bar{\mathbf{x}}_{k|1:K}$ and $\{\phi_{k|1:K}^{(r)}\}_{r=1}^{N_r}$ with the modes $\mathcal{X}_k$ together provide the decomposition of the smoothed state vector $\mathbf{X}_{k|1:K}(\omega)$ as per (4). This process is repeated for each time index k, starting from the final index K (where the smoothed and filtered distributions coincide) and marching backward in time with successive index decrements of 1, until we reach $k = 1$. Along the way, we determine the smoothed variables $\bar{\mathbf{x}}_{k|1:K}$ and $\{\phi_{k|1:K}^{(r)}\}_{r=1}^{N_r}$ for all $k = 1, 2, \ldots, K - 1$. This backward-smoothing pass constitutes the final step of the subspace-GMM smoother.

Table 2 summarizes the equations corresponding to the above three steps for the GMM–DO smoother, a particular case of the subspace-GMM smoother, which uses the GMM–DO filter and its optimal DO reduced-order SPDEs during the forward-filtering pass.

4) COMPUTATIONAL AND STORAGE COSTS

We now describe the computational and storage costs of the subspace-GMM smoother. In particular, we compare and contrast the costs of backward smoothing to these of the GMM-fitting passes. We also provide potential strategies to accelerate GMM fitting.

At each step of the backward smoothing pass, the conditional pdf (20) must be evaluated for all coefficient realizations $\phi_{k+1|1:K}^{(r)}$. The computation of the reduced component gain $\tilde{\mathbf{K}}^j$ using (19) is an $O(s^3)$ process (recall that s is the size of the stochastic subspace). Repeating this calculation for all M components incurs a total cost of $O(s^3M)$. As mentioned in section 3c, the component covariances $\tilde{\boldsymbol{\Sigma}}_k^j$ are uniform across all r. From (21b), the total cost of evaluating $\tilde{\boldsymbol{\Sigma}}_k^j$ for all j equals $O(s^3M)$. Next, the quantity $\mathcal{X}_{k+1}^T(\bar{\mathbf{x}}_{k+1|1:k} - \bar{\mathbf{x}}_{k+1|1:K})$ is $O(sN_{\mathbf{X}})$ to calculate and is used to determine all component mean vectors $\tilde{\boldsymbol{\mu}}_k^{j,(r)}$ in (21a) and component weights $\hat{\pi}^{j,(r)}$ in (18a). Hence, it is evaluated and stored in memory. The cost of evaluating $\tilde{\boldsymbol{\mu}}_k^{j,(r)}$ from (21a) for all j and r is $O(s^2MN_r)$ For the component weights $\hat{\pi}^{j,(r)}$, evaluating the Gaussian pdf in (18a) for all j and r incurs a cost of $O(s^2MN_r)$, after the inverses of all the component covariances $\boldsymbol{\Sigma}_{k+1,k+1|1:k}^j$ have been determined [which is an $O(s^3M)$ process]. Finally, computing the smoothed mean vector $\bar{\mathbf{x}}_{k|1:K}$ and the smoothed coefficients $\phi_{k|1:K}^{(r)}$ from (29) and (30), respectively, incurs a cost of $O(sN_r + sN_{\mathbf{X}})$. Adding all of the above at each k, the total cost of a single step of the backward-smoothing pass is only $O(s^3M + s^2MN_r + sN_{\mathbf{X}})$. Thus, the overall cost of the backward-smoothing pass for the K assimilations is $O(s^3MK + s^2MN_rK + sN_{\mathbf{X}}K)$.

We now consider the computational cost of the joint subspaces GMM-fitting pass. Since the EM algorithm is an iterative optimizer, the actual cost of GMM fitting depends on the nature of the joint subspace distributions themselves, for example, how far from Gaussian they are (M), in addition to the subspace size s and number of realizations N_r. In particular, the total number of floating point operations required for the E step in (B2) is $O(s^3M + s^2MN_r)$, which includes the cost of inverting all the component covariances and the subsequent evaluation of the Gaussian pdfs in (B2). The individual costs of parts a–c in the M step of (B3) are $O(N_rM)$, $O(sN_rM)$, and $O(s^2N_rM)$, respectively, adding up to $O(s^2N_rM)$.

Thus, each iteration of the EM procedure incurs a cost of $O(s^3 M + s^2 M N_r)$. Since $s \ll N_r$ in practice, the dominating term in the expression is $O(s^2 N_r M)$. If N_i iterations are required for convergence, the total cost of GMM fitting using the EM scheme grows to $O(s^2 N_r M N_i)$. Assuming that each of the $K - 1$ GMM fits in the GMM-fitting pass has a mixture complexity M and requires $O(N_i)$ iterations, the overall dominating cost of the GMM-fitting pass is $O(s^2 N_r M N_i K)$. Hence, comparing this cost of the EM–BIC scheme for GMM fitting to that of backward smoothing, we find that for large N_i, the former dominates.

One may speed up the EM–BIC scheme by placing reasonable bounds on M or by using other efficient fitting schemes (e.g., Sondergaard and Lermusiaux 2013a; Bouveyron and Brunet-Saumard 2014). Convergence can also be accelerated by choosing a suitable initial guess for the unknown mixture components in the EM algorithm. For example, one can set the initial guess to be a random perturbation around the converged GMM parameters of a lower mixture complexity or of a previous time. We also note that the present joint GMM fitting is decoupled from the filtering run. Hence, the joint subspace GMM fits can be performed offline and in parallel, either after the filtering run ends or, ideally, as it progresses. The ensemble-based RTS ESSE smoother (Lermusiaux and Robinson 1999) and Kalman smoothers with their recent low-rank implementations (Cosme et al. 2010, 2012) are limited to single-component ($M = 1$) Gaussian updates. Thus, their analysis steps are cheaper than that of the subspace-GMM smoother. Of course, a major advantage of this particular GMM–DO smoother (Table 2) is the statistical resolution in its dynamic stochastic subspace, that is, the rich number of realizations $N_r \gg s$ in the subspace.

The subspace-GMM smoother stores the mean vectors $\bar{\mathbf{x}}_{k|1:k-1}$, $\bar{\mathbf{x}}_{k|1:k}$, $\bar{\mathbf{x}}_{k|1:K}$, the matrices $\boldsymbol{\mathcal{X}}_k$, and the coefficients $\{\boldsymbol{\phi}_{k|1:k-1}^{(r)}\}_{r=1}^{N_r}$, $\{\boldsymbol{\phi}_{k|1:k}^{(r)}\}_{r=1}^{N_r}$, $\{\boldsymbol{\phi}_{k|1:K}^{(r)}\}_{r=1}^{N_r}$, thereby incurring a storage cost of $O[(N_{\mathbf{X}} + N_r)sK]$. Furthermore, the total cost of storing the GMM components from the GMM-fitting pass is $O(s^2 M K)$. Thus, the total storage cost of the smoother is $O[(N_{\mathbf{X}} + N_r)sK + s^2 M K]$. In contrast, the EnKS operates in the full state space and incurs a significantly larger storage cost, even with the same number of realizations as the subspace-GMM smoother. The cost of storing N_r EnKS ensemble members in the state space is $O(N_{\mathbf{X}} N_r K)$, which is much larger than that of the subspace-GMM smoother as $N_{\mathbf{X}} \gg s$. The subspace-GMM smoother has a similar storage cost as the ensemble-based ESSE smoother and low-rank Kalman smoothers, since their smoother updates are also carried out in a dominant stochastic subspace.

f. Other remarks

1) USE OF GMMS

Our use of GMMs to represent the pdfs of the stochastic coefficients is motivated by two factors. First, GMMs can represent the pdfs of continuous random variables to any desired level of accuracy. In particular, they are superior in capturing multimodal distributions that are often encountered in weather and ocean-based systems. Second, many elegant properties of Gaussian pdfs also extend to GMMs. For example, GMMs are conjugate with respect to linear Gaussian measurement models. This key property allowed us to derive the equations of the backward-smoothing pass of the subspace-GMM smoother. Finally, the GMM smoother is a fundamental direct extension to the Gaussian smoother, which is retrieved for $M = 1$.

2) NONLINEARITY PRESERVATION

The use of GMMs for smoothing in linear (or linearized) systems has been recently studied. In such systems, an explicit algebraic relationship holds between the GMM state vectors at any two time instances. For example, Vo et al. (2012) derived smoother equations for linear systems with GMM state distributions. In an earlier work (Lolla 2016), we extended their approach to derive analytical expressions for the smoothed pdfs in weakly nonlinear systems with linear Gaussian measurement models. The idea is to compute the smoothed pdf at any time by performing a multiplicative correction to the corresponding filtered pdf. Similar to the subspace-GMM smoother, the filtered pdfs are computed and saved during a forward-filtering pass. The correction terms are computed analytically through a backward pass, *linearizing* the dynamics operator at each time step. We implemented this method for the chaotic Lorenz-63 model, but the results are not shown here. This approach suffers from the drawback of linearizing the dynamics during the backward pass. The present subspace-GMM smoother however performs no linearization during the backward-smoothing pass as the joint subspace GMMs directly represent the joint pdfs $p_{\boldsymbol{\Phi}_k, \boldsymbol{\Phi}_{k+1} | \mathbf{Y}_{1:k}}$. The nonlinear transformations of the filtered stochastic coefficients across time [e.g., $\boldsymbol{\Phi}_{k|1:k}(\omega) \rightarrow \boldsymbol{\Phi}_{k+1|1:k}(\omega)$] are fully captured by allowing the GMM-fitting process to adjust the mixture parameters and complexity in accordance with their joint realizations. Hence, the use of GMMs can potentially circumvent the issue of spurious correlations introduced by a Gaussian treatment of joint state variables (Nerger et al. 2014).

3) EFFECT OF SMOOTHING ON THE STOCHASTIC SUBSPACE

The present subspace-GMM smoother assumes that the stochastic subspaces (spanned by the columns of $\mathcal{X}_k$) remain unaltered by future observations. Smoothing then only updates the distribution of uncertainty within the subspaces. This is a logical assumption for practical applications, since for a properly chosen s, the errors outside the subspace are small. Thus, the smoother corrections outside the subspace are expected to be small. If not, one can successively increase the size of the subspace (s) to capture this missing fraction of the uncertainty (Lermusiaux 2007; Sapsis and Lermusiaux 2012), until the smoothed subspace coefficients converge statistically. Of course, the filtered subspace at t_k (i.e., $\mathcal{X}_k$) has been influenced by all observations taken prior to t_k. Nonetheless, as done in ESSE (Lermusiaux 1999), one can extend the present smoothing and expand–learn the subspace backward in time, based on the smoothed observation residuals. When these residuals lie outside the filtered subspace and not as expected by the model in (2), they are used to update filtered subspaces into smoothed ones and even to further correct the smoothed realizations. However, these schemes are not discussed here.

4. Conclusions and future work

Retrospective inference, or smoothing, is indispensable in geosciences. It can be used to perform reanalysis of ocean fields, detect sources of pollutants, initialize numerical weather predictions, and infer past climate states. Going beyond the realm of quadratic cost or Gaussian methods, recent years have witnessed a growth of Bayesian smoothing approaches. However, most existing methods perform poorly in high-dimensional systems, such as those arising in geosciences. Addressing these issues, we derived the fundamental equations of the full state-space, and subspace GMM smoothers, focusing on their RTS-style forward–backward form. We also obtained and discussed the theoretical properties and computational costs of the new GMM smoother equations.

The GMM–DO smoother, a particular case of the subspace-GMM smoother, uses the Bayesian GMM–DO filter (Sondergaard and Lermusiaux 2013a) to accurately assimilate observations sequentially over time. Uncertainties are forecast using DO equations (Sapsis and Lermusiaux 2009), efficiently reducing the dimensionality to the time-evolving dominant stochastic subspace. After filtering, a key step in the smoother is the joint GMM fitting performed within the joint stochastic subspaces across pairs of successive observation times. This joint fitting allows for the analytical, nonlinear back propagation of future information. In the backward-smoother pass, the filter estimates are updated by solving the smoothing recursion equation within the stochastic subspaces, rendering the algorithm practical and computationally efficient. Since all smoother operations are performed in the subspaces, ensemble members of the high-dimensional state space are never computed. This alleviates the storage costs associated with smoothing. The overall results are the equations and RTS-style algorithm of the GMM–DO smoother, tailored for high-dimensional problems. The GMM–DO smoother preserves the dominant non-Gaussian structure of the stochastic dynamical fields, accurately evolves them using the governing nonlinear PDEs, and propagates the observed information backward in a nonlinear GMM fashion. In a companion paper (Lolla and Lermusiaux 2017), we employ a double-well diffusion experiment, a reversible passive tracer advection, and a simulated ocean flow exiting a strait–estuary to validate the smoother and compare its performance to that of other smoothers.

It is straightforward to extend the RTS-style subspace GMM smoother to other forms, such as the fixed interval or the fixed lag. The procedure is similar to that of other classic fixed-interval or fixed-lag smoothers. Presently, all GMM fitting operations are completed using the EM–BIC scheme. A future direction is to investigate alternate techniques for GMM fitting using other schemes from information theory and machine learning (McLachlan and Peel 2000). A related idea introduced in Sondergaard and Lermusiaux (2013a) is to complete the GMM fitting only in a dominant subspace *within* the stochastic DO subspace. One could also directly work with GMMs in the subspace, and derive evolution equations of the GMM components, rather than employing a Monte Carlo method to forecast the stochastic coefficients. Finally, the backward-smoothing pass offers interesting fully Bayesian extensions of adjoint and variational methods, which could be further investigated.

anonymous reviewers and the editor, Dr. Jeff Anderson, for their comments.

APPENDIX A

Overview of the GMM–DO Filter

In this section, we review the GMM–DO filter (Sondergaard and Lermusiaux 2013a,b), which is the methodology for uncertainty prediction and filtering used in this two-part paper. Specifically, we outline (i) the Dynamically Orthogonal (DO) field equations for uncertainty prediction and (ii) the analysis step of the GMM–DO filter. In each case, we briefly justify the choice of these components within the context of oceanic and atmospheric data assimilation.

The GMM–DO filter preserves the non-Gaussian structure of the state variables and respects their non-linear dynamics. It employs the DO methodology to solve the governing SPDE (1a) and to forecast the prior pdfs of the state vector $\mathbf{X}(t; \omega)$. At each assimilation time t_k, $k = 1, 2, \ldots, K$, the filter uses the Expectation-Maximization (EM) algorithm, coupled with the Bayesian Information Criterion (BIC) to perform a semiparametric Gaussian Mixture Model fit (see appendix B) of the prior pdf given by the DO methodology. During the analysis step, the filtered posterior pdf is computed by analytically carrying out Bayes's law in the DO subspace in accordance with the measurement model (2). The DO methodology then advances this filtered posterior state vector forward in time to yield the forecast (prior) pdf at the next assimilation time, t_{k+1}. This process is then repeated until the final assimilation time t_K.

a. The Dynamically Orthogonal field equations for uncertainty prediction

The DO methodology (Sapsis and Lermusiaux 2009) is a reduced-order technique to solve (1a)–(1c) for continuous stochastic fields $X(\mathbf{r}, t; \omega)$. It decomposes the stochastic field $X(\mathbf{r}, t; \omega)$ using a generalized, time-dependent Karhunen–Loéve (KL) expansion:

$$X(\mathbf{r}, t; \omega) = \overline{x}(\mathbf{r}, t) + \sum_{i=1}^{s} \tilde{x}_i(\mathbf{r}, t)\Phi_i(t; \omega). \quad \text{(A1)}$$

Here, $\overline{x}(\mathbf{r}, t)$ represents the mean of $X(\mathbf{r}, t; \omega)$; that is, $\overline{x}(\mathbf{r}, t) = E_{\omega}[X(\mathbf{r}, t; \omega)]$, where E_{ω} denotes the expectation operator. The (possibly time dependent) scalar s is the total number of modes retained in the KL expansion, and is also the size of the stochastic subspace. The family of deterministic modes $\tilde{x}_i(\mathbf{r}, t)$, for $i = 1, 2, \ldots, s$, constitutes an orthonormal basis for the time-dependent stochastic subspace. The randomness within this subspace is captured by the zero-mean stochastic

coefficients $\Phi_i(t; \omega)$. Therefore, the DO methodology describes the randomness in the field $X(\mathbf{r}, t; \omega)$ entirely through the coefficients $\Phi_i(t; \omega)$, based on the affine mapping in (A1). Hereafter, the s-dimensional random vector of stochastic coefficients $[\Phi_1(t; \omega), \Phi_2(t; \omega), \ldots, \Phi_s(t; \omega)]^{\mathrm{T}}$ will be denoted by $\mathbf{\Phi}(t; \omega)$.

A closed set of evolution equations for $\overline{x}(\mathbf{r}, t)$, $\{\tilde{x}_i(\mathbf{r}, t)\}_{i=1}^{s}$, and $\{\Phi_i(t; \omega)\}_{i=1}^{s}$ are derived by substituting (A1) into (1a) and imposing the "DO condition":

$$\left\langle \frac{\partial \tilde{x}_i(\cdot, t)}{\partial t}, \tilde{x}_j(\cdot, t) \right\rangle = 0, \quad \forall \quad i, j \in \{1, 2, \ldots, s\}. \quad \text{(A2)}$$

The DO condition (A2) dictates that, with respect to the chosen inner product, the transformation of the stochastic subspace is strictly orthogonal to its current configuration. This condition is imposed without loss of generality as it removes the redundancy of having both the basis and coefficients to represent the evolution of uncertainty within the stochastic subspace. Substituting (A2) and (A1) into (1a) and performing a Galerkin projection of the resulting equation onto each of the modes $\tilde{x}_i$, and a statistical average after multiplication with the coefficients, yields the following system of equations:

$$\frac{\partial \overline{x}(\mathbf{r}, t)}{\partial t} = E_{\omega}[\mathcal{L}[X(\mathbf{r}, t; \omega); \omega]], \quad \text{(A3a)}$$

$$\frac{\partial \tilde{x}_i(\mathbf{r}, t)}{\partial t} = \sum_{j=1}^{s} \mathbf{\Pi}^{\perp}(E_{\omega}[\mathcal{L}[X(\mathbf{r}, t; \omega); \omega]\Phi_j(t; \omega)])C_{\Phi_i(t)\Phi_j(t)}^{-1}, \quad \text{(A3b)}$$

and

$$\frac{d\Phi_i(t; \omega)}{dt} = \langle \mathcal{L}[X(\mathbf{r}, t; \omega); \omega] - E_{\omega}[\mathcal{L}[X(\mathbf{r}, t; \omega); \omega]], \tilde{x}_i(\mathbf{r}, t)\rangle, \quad \text{(A3c)}$$

where

$$\mathbf{\Pi}^{\perp}[F(\mathbf{r}, t)] = F(\mathbf{r}, t) - \sum_{j=1}^{s} \langle F(\cdot, t), \tilde{x}_j(\cdot, t)\rangle \tilde{x}_j(\mathbf{r}, t)$$

is the projection of a given field F onto the orthogonal complement of the stochastic subspace and C is the covariance operator; that is, $C_{\Phi_i(t)\Phi_j(t)} = E_{\omega}[\Phi_i(t; \omega)\Phi_j(t; \omega)]$. Using (1c), the boundary conditions on $\overline{x}$ and $\{\tilde{x}_i\}_{i=1}^{s}$ take the form

$$\mathcal{B}[\overline{x}(\mathbf{r}, t)]|_{\mathbf{r}=\xi} = E_{\omega}[h(\xi, t; \omega)] \quad \text{and} \quad \text{(A4a)}$$

$$\mathcal{B}[\tilde{x}_i(\mathbf{r}, t)]|_{\mathbf{r}=\xi} = \sum_{j=1}^{s} E_{\omega}[h(\xi, t; \omega)\Phi_j(t; \omega)]C_{\Phi_i(t)\Phi_j(t)}^{-1}. \quad \text{(A4b)}$$

Similarly, the initial conditions on $\overline{x}$, $\{\tilde{x}_i\}_{i=1}^{s}$, and $\{\Phi_i\}_{i=1}^{s}$ are

$$\overline{x}(\mathbf{r}, 0) := \overline{x}_0(\mathbf{r}) = E_\omega[X_0(\mathbf{r}; \omega)], \qquad \text{(A5a)}$$

$$\tilde{x}_i(\mathbf{r}, 0) = \tilde{x}_{i,0}(\mathbf{r}), \quad \text{and} \qquad \text{(A5b)}$$

$$\Phi_i(0; \omega) = \langle X_0(\cdot; \omega) - \overline{x}_0(\cdot), \tilde{x}_{i,0}(\cdot) \rangle. \qquad \text{(A5c)}$$

In addition to the stochastic coefficients $\Phi_i(t; \omega)$, the DO equations in (A3a)–(A3c) evolve the modes $\tilde{x}_i(\mathbf{r}, t)$ in accordance with the dynamics of $X(\mathbf{r}, t; \omega)$ and the boundary conditions. This dynamic evolution of the modes offers a significant advantage over schemes such as proper orthogonal decomposition, which fix the subspace in time (Sapsis and Lermusiaux 2009). We note that s can also be evolved based on the dynamics and external observations (Lermusiaux 1999; Sapsis and Lermusiaux 2012). However, without loss of generality for our purpose of smoothing, we will assume here that s is fixed. We denote the spatially discretized fields of $\overline{x}(\mathbf{r}, t)$ and $\tilde{x}_i(\mathbf{r}, t)$ by $\overline{\mathbf{x}}(t)$ and $\tilde{\mathbf{x}}_i(t)$, respectively. In this case, the matrix $\boldsymbol{\mathcal{X}}(t) = [\tilde{\mathbf{x}}_1(t) \,|\, \tilde{\mathbf{x}}_2(t) \,|\, \ldots \,|\, \tilde{\mathbf{x}}_s(t)]$ is formed by arranging the discretized modes $\tilde{\mathbf{x}}_i(t)$ as column vectors.

The final step of the GMM–DO filter at each observation time is the analysis step, which computes the posterior variables conditioned on the observation value. We summarize this step next.

b. The GMM–DO filter: Analysis step

One of the main features of the GMM–DO filter is its ability to retain the non-Gaussian character of the state variable $X(\mathbf{r}, t; \omega)$ during the analysis step. To do so, at each observation time t_k, the filter fits a semiparametric GMM to the forecast vector of the stochastic coefficients $\boldsymbol{\Phi}_{k|1:k-1}(\omega)$ using the EM algorithm (see appendix B). The mixture complexity (characterized by M, the number of mixture components) in the EM algorithm is typically not known a priori. The choice of M must be guided by the empirical evidence available, namely the set of ensemble realizations $\{\boldsymbol{\phi}_{k|1:k-1}^{(r)}\}_{r=1}^{N_r}$. To select a suitable value for M, the GMM–DO filter uses the BIC (appendix B).

At each assimilation time t_k, the DO differential equations forecast the reduced-order representation (3) for the prior state vector $\mathbf{X}_{k|1:k-1}(\omega)$. The present GMM–DO filter uses a Monte Carlo approach to integrate the stochastic coefficients $\boldsymbol{\Phi}_{k|1:k-1}(\omega)$ [we note that other schemes are possible; see Sapsis and Lermusiaux (2009) and Ueckermann et al. (2013)]. The stochastic ODEs in (A3c) are solved in a particle-wise manner. Immediately preceding the analysis step, the GMM–DO filter performs a GMM fit of the realizations of $\boldsymbol{\Phi}_{k|1:k-1}(\omega)$, using the EM–BIC criteria to yield the parameter set $\{\pi_k^j, \overline{\boldsymbol{\mu}}_{k|1:k-1}^j, \boldsymbol{\Sigma}_{k|1:k-1}^j\}_{j=1}^M$. Consequently, $\boldsymbol{\Phi}_{k|1:k-1}(\omega)$. has the following pdf:

$$p_{\boldsymbol{\Phi}_k|\mathbf{Y}_{1:k-1}}(\boldsymbol{\phi}_k \,|\, \mathbf{y}_{1:k-1}) = \sum_{j=1}^M \pi_k^j \times \mathcal{N}(\boldsymbol{\phi}_k; \overline{\boldsymbol{\mu}}_{k|1:k-1}^j, \boldsymbol{\Sigma}_{k|1:k-1}^j). \qquad \text{(A6)}$$

Upon the arrival of the noisy observation $\mathbf{y}_k$, the posterior values of the state mean vector $\overline{\mathbf{x}}_{k|1:k-1}$ and all GMM components $\{\pi_k^j, \overline{\boldsymbol{\mu}}_{k|1:k-1}^j, \boldsymbol{\Sigma}_{k|1:k-1}^j\}_{j=1}^M$ are *analytically* computed, using the property that GMMs are conjugate priors with respect to a linear Gaussian observation model. The posterior mean vector $\overline{\mathbf{x}}_{k|1:k}$ is given by

$$\overline{\mathbf{x}}_{k|1:k} = \overline{\mathbf{x}}_{k|1:k-1} + \boldsymbol{\mathcal{X}}_k \sum_{j=1}^M \hat{\pi}_k^j \times \hat{\boldsymbol{\mu}}_k^j, \qquad \text{(A7)}$$

and the filtered vector of stochastic coefficients $\boldsymbol{\Phi}_{k|1:k}(\omega)$ has the following pdf:

$$p_{\boldsymbol{\Phi}_k|\mathbf{Y}_{1:k}}(\boldsymbol{\phi}_k \,|\, \mathbf{y}_{1:k}) = \sum_{j=1}^M \hat{\pi}_k^j \times \mathcal{N}(\boldsymbol{\phi}_k; \overline{\boldsymbol{\mu}}_{k|1:k}^j, \boldsymbol{\Sigma}_{k|1:k}^j). \qquad \text{(A8)}$$

The components of this posterior GMM distribution have the following exact expressions:

$$\hat{\pi}_k^j \propto \pi_k^j \times \mathcal{N}(\tilde{\mathbf{y}}_k; \tilde{\mathbf{H}}_k \overline{\boldsymbol{\mu}}_{k|1:k-1}^j, \tilde{\mathbf{H}}_k \boldsymbol{\Sigma}_{k|1:k-1}^j \tilde{\mathbf{H}}_k^T + \mathbf{R}),$$

$$\text{with} \quad \sum_{j=1}^M \hat{\pi}_k^j = 1, \qquad \text{(A9)}$$

$$\overline{\boldsymbol{\mu}}_{k|1:k}^j = \hat{\boldsymbol{\mu}}_k^j - \sum_{l=1}^M \hat{\pi}_k^l \times \hat{\boldsymbol{\mu}}_k^l, \quad \text{and} \qquad \text{(A10)}$$

$$\boldsymbol{\Sigma}_{k|1:k}^j = (\mathbf{I} - \tilde{\mathbf{K}}_k^j \tilde{\mathbf{H}}_k) \boldsymbol{\Sigma}_{k|1:k-1}^j. \qquad \text{(A11)}$$

In the above expressions, the following definitions are used:

$$\hat{\boldsymbol{\mu}}_k^j = \overline{\boldsymbol{\mu}}_{k|1:k-1}^j + \tilde{\mathbf{K}}_k^j(\tilde{\mathbf{y}}_k - \tilde{\mathbf{H}}_k \overline{\boldsymbol{\mu}}_{k|1:k-1}^j), \qquad \text{(A12)}$$

$$\tilde{\mathbf{H}}_k = \mathbf{H} \boldsymbol{\mathcal{X}}_k, \qquad \text{(A13)}$$

$$\tilde{\mathbf{y}}_k = \mathbf{y}_k - \mathbf{H} \overline{\mathbf{x}}_{k|1:k-1}, \quad \text{and} \qquad \text{(A14)}$$

$$\tilde{\mathbf{K}}_k^j = \boldsymbol{\Sigma}_{k|1:k-1}^j \tilde{\mathbf{H}}_k^T(\tilde{\mathbf{H}}_k \boldsymbol{\Sigma}_{k|1:k-1}^j \tilde{\mathbf{H}}_k^T + \mathbf{R})^{-1}. \qquad \text{(A15)}$$

The GMM–DO analysis step is completed by drawing ensemble realizations of the posterior stochastic coefficient vector $\boldsymbol{\Phi}_{k|1:k}(\omega)$ through its pdf given by (A8). The DO representation of the posterior filtered state vector is advanced until the next assimilation (t_{k+1}) using the DO equations in (A3). This completes the outline of the GMM–DO filter (Sondergaard and Lermusiaux 2013a).

APPENDIX B

The Expectation-Maximization (EM) Algorithm and Bayesian Information Criterion (BIC) for Gaussian Mixture Models

a. The EM algorithm for Gaussian Mixture Models

In this section, we briefly summarize the EM algorithm specific to the case of a multivariate GMM fit of a random vector $\mathbf{\Phi}$, whose realizations are denoted by $\boldsymbol{\phi}^{(r)}$, $r = 1, 2, \ldots N_r$. Further details are available in Bilmes (1998) and Sondergaard and Lermusiaux (2013a) and the references therein.

The EM algorithm is an iterative procedure for estimating the parameters of a target distribution that maximize the probability of obtaining a given set of realizations. For a chosen mixture complexity M, the EM algorithm estimates the quantities π^j, $\overline{\boldsymbol{\mu}}^j$, and $\mathbf{\Sigma}^j$ for $j = 1, 2, \ldots, M$ in the GMM representation of $p_{\mathbf{\Phi}}(\boldsymbol{\phi})$:

$$p_{\mathbf{\Phi}}(\boldsymbol{\phi}) = \sum_{j=1}^{M} \pi^j \times \mathcal{N}(\boldsymbol{\phi}; \overline{\boldsymbol{\mu}}^j, \mathbf{\Sigma}^j). \quad (B1)$$

Here, the parameters π^j, $\overline{\boldsymbol{\mu}}^j$, and $\mathbf{\Sigma}^j$ denote the weight, mean vector, and the covariance matrix of the jth mixture component of the GMM, respectively. The EM algorithm is composed of a succession of expectation and maximization steps to obtain the maximum-likelihood (ML) estimate of these parameters. It successively estimates the weight with which the realizations $\boldsymbol{\phi}^{(r)}$ of $\mathbf{\Phi}$ are associated with each of the M mixture components. This is done based on the present parameter estimates, followed by an optimization of these parameters using the newly calculated weights. Repeating this process until convergence ultimately yields the ML estimate of the parameters based on the ensemble realizations $\boldsymbol{\phi}^{(r)}$. The result is as follows.

Given the initial parameter estimate,

$$\boldsymbol{\theta}_{(0)} = \{\pi^1_{(0)}, \ldots, \pi^M_{(0)}, \overline{\boldsymbol{\mu}}^1_{(0)}, \ldots, \overline{\boldsymbol{\mu}}^M_{(0)}, \mathbf{\Sigma}^1_{(0)}, \ldots, \mathbf{\Sigma}^M_{(0)}\};$$

repeat until convergence:

- *E step*—For all $r \in \{1, 2, \ldots, N_r\}, j \in \{1, 2, \ldots, M\}$, use the present parameter estimate $\theta_{(l)}$ to form

$$\tau^j(\boldsymbol{\phi}^{(r)}; \boldsymbol{\theta}_{(l)}) = \frac{\pi^j_{(l)} \times \mathcal{N}(\boldsymbol{\phi}^{(r)}; \overline{\boldsymbol{\mu}}^j_{(l)}, \mathbf{\Sigma}^j_{(l)})}{\displaystyle\sum_{m=1}^{M} \pi^m_{(l)} \times \mathcal{N}(\boldsymbol{\phi}^{(r)}; \overline{\boldsymbol{\mu}}^m_{(l)}, \mathbf{\Sigma}^m_{(l)})}. \quad (B2)$$

- *M step*—For all $j \in \{1, 2, \ldots, M\}$, update the parameter estimate to $\theta_{(l+1)}$ as follows:

$$\pi^j_{(l+1)} = \frac{N^j_{(l)}}{N_r}, \quad (B3a)$$

$$\overline{\boldsymbol{\mu}}^j_{(l+1)} = \frac{1}{N^j_{(l)}} \times \sum_{r=1}^{N_r} \tau^j(\boldsymbol{\phi}^{(r)}; \boldsymbol{\theta}_{(l)}) \times \boldsymbol{\phi}^{(r)}, \quad \text{and} \quad (B3b)$$

$$\mathbf{\Sigma}^j_{(l+1)} = \frac{1}{N^j_{(l)}} \times \sum_{r=1}^{N_r} \tau^j(\boldsymbol{\phi}^{(r)}; \boldsymbol{\theta}_{(l)}) \times [\boldsymbol{\phi}^{(r)} - \overline{\boldsymbol{\mu}}^j_{(l+1)}]$$
$$\times [\boldsymbol{\phi}^{(r)} - \overline{\boldsymbol{\mu}}^j_{(l+1)}]^{\mathrm{T}}, \quad (B3c)$$

where

$$N^j_{(l)} = \sum_{r=1}^{N_r} \tau^j(\boldsymbol{\phi}^{(r)}; \boldsymbol{\theta}_{(l)}). \quad (B4)$$

In the E step, we calculate the probability of mixture component j having generated the realization $\boldsymbol{\phi}^{(r)}$. based on the present parameter estimates. We do so across all pairs of realizations and components. In the M step, we update the parameter values in accordance with their weighted averages across all realizations. Upon repeating the above steps until convergence, we arrive at a maxima of the ML estimate of the parameters for a given mixture complexity M.

b. The Bayesian Information Criterion

The BIC is a quantitative equivalent of the Occam's razor principle, which states that one should favor the simplest model consistent with the ensemble. It tries to strike a balance between underfitting, which fails to capture the trend in the data, and overfitting, which limits predictive capability beyond the ensemble. We now briefly summarize this criterion, and refer the reader to Sondergaard and Lermusiaux (2013a) for more details.

The goal of the BIC is to choose the model complexity M that maximizes the likelihood of obtaining the ensemble set $\{\boldsymbol{\phi}^{(r)}\}^{N_r}_{r=1}$. The parameter vector $\boldsymbol{\theta}$ is assumed to be random with an arbitrary prior distribution of $p_{\Theta}(\boldsymbol{\theta}; M)$, and M is considered constant but unknown. The distribution of the ensemble set, assuming independence of the realizations, is given by

$$p_{\{\mathbf{\Phi}\}}(\{\boldsymbol{\phi}\}; M) = \prod_{r=1}^{N_r} p_{\mathbf{\Phi}}(\boldsymbol{\phi}^{(r)}; M). \quad (B5)$$

The BIC seeks the value of M at which $p_{\{\mathbf{\Phi}\}}(\{\boldsymbol{\phi}\}; M)$ is maximized. Using the definition of conditional probability, we obtain

$$p_{\{\Phi\}}(\{\phi\}; M) = \frac{p_{\{\Phi\}|\Theta}(\{\phi\} \mid \theta; M) \, p_{\Theta}(\theta; M)}{p_{\Theta|\{\Phi\}}(\theta \mid \{\phi\}; M)}. \quad \text{(B6)}$$

The pdf $p_{\{\Phi\}|\Theta}(\{\phi\}; \theta; M)$ in the numerator is a GMM in our case. In the denominator $p_{\Theta|\{\Phi\}}(\theta \mid \{\phi\}; M)$ is evaluated using the Laplace approximation at the ML estimate of the parameter vector θ, denoted by $\hat{\theta}_{\text{ML}}$. The BIC is formally defined as

$$\text{BIC}(M) = K_M \log N_r - 2 \log p_{\{\Phi\}|\Theta}(\{\phi\} \mid \hat{\theta}_{\text{ML}}; M), \quad \text{(B7)}$$

where K_M denotes the length of the parameter vector θ. The above expression for the BIC is obtained through a sequence of approximations to $-2 \log p_{\{\Phi\}}(\{\phi\}; M)$, starting from (B6). The optimum value of M, therefore, minimizes the BIC defined in (B7).

REFERENCES

Atkins, E., M. Morzfeld, and A. J. Chorin, 2013: Implicit particle methods and their connection with variational data assimilation. *Mon. Wea. Rev.*, **141**, 1786–1803, doi:10.1175/MWR-D-12-00145.1.

Barth, A., A. Alvera-Azcárate, K.-W. Gurgel, J. Staneva, A. Port, J.-M. Beckers, and E. V. Stanev, 2010: Ensemble perturbation smoother for optimizing tidal boundary conditions by assimilation of high-frequency radar surface currents application to the German Bight. *Ocean Sci.*, **6**, 161–178, doi:10.5194/os-6-161-2010.

Bennett, A. F., 1992: *Inverse Methods in Physical Oceanography.* Cambridge University Press, 346 pp.

Bilmes, J. A., 1998: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute Tech. Rep. TR-97-021, 18 pp. [Available online at http://melodi.ee.washington.edu/people/bilmes/mypapers/em.pdf.]

Bocquet, M., 2005: Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: Theory. *Quart. J. Roy. Meteor. Soc.*, **131**, 2191–2208, doi:10.1256/qj.04.67.

——, C. A. Pires, and L. Wu, 2010: Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Wea. Rev.*, **138**, 2997–3023, doi:10.1175/2010MWR3164.1.

Bouveyron, C., and C. Brunet-Saumard, 2014: Model-based clustering of high-dimensional data: A review. *Comput. Stat. Data Anal.*, **71**, 52–78, doi:10.1016/j.csda.2012.12.008.

Bresler, Y., 1986: Two-filter formulae for discrete-time non-linear Bayesian smoothing. *Int. J. Control*, **43**, 629–641, doi:10.1080/00207178608933489.

Briers, M., A. Doucet, and S. Maskell, 2010: Smoothing algorithms for state–space models. *Ann. Inst. Stat. Math.*, **62**, 61–89, doi:10.1007/s10463-009-0236-2.

Choi, H.-L., and J. P. How, 2009: On the roles of smoothing in planning of informative paths. *American Control Conf. 2009*, St. Louis, MO, American Automatic Control Council, 2154–2159. [Available online at http://web.mit.edu/hanlimc/www/hl.docs/ChoiHow_ACC09.pdf.]

——, and ——, 2010: Continuous trajectory planning of mobile sensors for informative forecasting. *Automatica*, **46**, 1266–1275, doi:10.1016/j.automatica.2010.05.004.

Cohn, S. E., N. Sivakumaran, and R. Todling, 1994: A fixed-lag Kalman smoother for retrospective data assimilation. *Mon.*

Wea. Rev., **122**, 2838–2867, doi:10.1175/1520-0493(1994)122<2838:AFLKSF>2.0.CO;2.

Cosme, E., J.-M. Brankart, J. Verron, P. Brasseur, and M. Krysta, 2010: Implementation of a reduced rank square-root smoother for high resolution ocean data assimilation. *Ocean Modell.*, **33**, 87–100, doi:10.1016/j.ocemod.2009.12.004.

——, J. Verron, P. Brasseur, J. Blum, and D. Auroux, 2012: Smoothing problems in a Bayesian framework and their linear Gaussian solutions. *Mon. Wea. Rev.*, **140**, 683–695, doi:10.1175/MWR-D-10-05025.1.

Crassidis, J. L., and J. L. Junkins, 2011: *Optimal Estimation of Dynamic Systems.* 2nd ed. CRC Press, 733 pp.

Daley, R., 1993: *Atmospheric Data Analysis.* 2nd ed. Cambridge Atmospheric and Space Science Series, Cambridge University Press, 472 pp.

Di Lorenzo, E., A. M. Moore, H. G. Arango, B. D. Cornuelle, A. J. Miller, B. Powell, B. S. Chua, and A. F. Bennett, 2007: Weak and strong constraint data assimilation in the inverse Regional Ocean Modeling System (ROMS): Development and application for a baroclinic coastal upwelling system. *Ocean Modell.*, **16**, 160–187, doi:10.1016/j.ocemod.2006.08.002.

Dimet, F.-X. L., and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus*, **38A**, 97–110, doi:10.1111/j.1600-0870.1986.tb00459.x.

Doucet, A., and A. M. Johansen, 2009: A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, D. Crisan and B. Rozovskiĭ, Eds., Oxford University Press, 656–704.

Evensen, G., and P. J. Van Leeuwen, 2000: An ensemble Kalman smoother for nonlinear dynamics. *Mon. Wea. Rev.*, **128**, 1852–1867, doi:10.1175/1520-0493(2000)128<1852:AEKSFN>2.0.CO;2.

Feppon, F., and P. F. J. Lermusiaux, 2017: A geometric approach to dynamical model-order reduction. *SIAM J. Sci. Comput.*, in press.

Gelb, A., 1974: *Applied Optimal Estimation.* The MIT Press, 374 pp.

Ghanem, R. G., and P. D. Spanos, 2003: *Stochastic Finite Elements: A Spectral Approach.* Courier Corporation, 222 pp.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics*, Vol. 33, Academic Press, 141–266, doi:10.1016/S0065-2687(08)60442-2.

Godsill, S. J., A. Doucet, and M. West, 2004: Monte Carlo smoothing for nonlinear time series. *J. Amer. Stat. Assoc.*, **99**, 156–168, doi:10.1198/016214504000000151.

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. *Mon. Wea. Rev.*, **136**, 317–334, doi:10.1175/2007MWR1927.1.

Hsieh, M. A., and G. Chirikjian, 2014: *Distributed Autonomous Robotic Systems: The 11th International Symposium.* Springer Tracts in Advanced Robotics, Vol. 104, Springer, 450 pp.

Jazwinski, A. H., 2007: *Stochastic Processes and Filtering Theory.* Dover Publications, 400 pp.

Kaess, M., S. Williams, V. Indelman, R. Roberts, J. J. Leonard, and F. Dellaert, 2012: Concurrent filtering and smoothing. *15th Int. Conf. on Information Fusion*, Singapore, IEEE, 1300–1307.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *J. Fluids Eng.*, **82**, 35–45.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation, and Predictability.* Cambridge University Press, 341 pp.

Khare, S. P., J. L. Anderson, T. J. Hoar, and D. Nychka, 2008: An investigation into the application of an ensemble Kalman smoother to high-dimensional geophysical systems. *Tellus*, **60A**, 97–112, doi:10.1111/j.1600-0870.2007.00281.x.

Kitagawa, G., 1987: Non-Gaussian state–space modeling of non-stationary time series. *J. Amer. Stat. Assoc.*, **82**, 1032–1041.

——, 1994: The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother. *Ann. Inst. Stat. Math.*, **46**, 605–623, doi:10.1007/BF00773470.

Klaas, M., M. Briers, N. De Freitas, A. Doucet, S. Maskell, and D. Lang, 2006: Fast particle smoothing: If I had a million particles. *Proc. 23rd Int. Conf. on Machine Learning*, Pittsburgh, PA, ACM, 481–488.

Lee, D. J., and M. E. Campbell, 2015: Smoothing algorithm for nonlinear systems using Gaussian mixture models. *J. Guid. Control Dyn.*, **38**, 1438–1451, doi:10.2514/1.G000603.

Lermusiaux, P. F. J., 1997: Error subspace data assimilation methods for ocean field estimation: Theory, validation and applications. Ph.D. thesis, Harvard University, 402 pp.

——, 1999: Estimation and study of mesoscale variability in the Strait of Sicily. *Dyn. Atmos. Oceans*, **29**, 255–303, doi:10.1016/S0377-0265(99)00008-1.

——, 2006: Uncertainty estimation and prediction for interdisciplinary ocean dynamics. *J. Comput. Phys.*, **217**, 176–199, doi:10.1016/j.jcp.2006.02.010.

——, 2007: Adaptive modeling, adaptive data assimilation and adaptive sampling. *Physica D*, **230**, 172–196, doi:10.1016/j.physd.2007.02.014.

——, and A. Robinson, 1999: Data assimilation via error subspace statistical estimation. Part I: Theory and schemes. *Mon. Wea. Rev.*, **127**, 1385–1407, doi:10.1175/1520-0493(1999)127<1385:DAVESS>2.0.CO;2.

——, A. R. Robinson, P. J. Haley, and W. G. Leslie, 2002: Advanced interdisciplinary data assimilation: Filtering and smoothing via error subspace statistical estimation. *Proc. OCEANS 2002 MTS/IEEE Conf.*, Biloxi, MS, IEEE, 795–802.

Lolla, T., 2016: Path planning and adaptive sampling in the coastal ocean. Ph.D. thesis, Massachusetts Institute of Technology, 315 pp. [Available online at https://dspace.mit.edu/handle/1721.1/103438.]

——, and P. F. J. Lermusiaux, 2017: A Gaussian mixture model smoother for continuous nonlinear stochastic dynamical systems: Applications. *Mon. Wea. Rev.*, **145**, 2763–2790, doi:10.1175/MWR-D-16-0065.1.

McLachlan, G., and D. Peel, 2000: *Finite Mixture Models*. Wiley Series in Probability and Statistics, John Wiley and Sons, 419 pp.

Miller, R. N., E. F. Carter, and S. T. Blue, 1999: Data assimilation into nonlinear stochastic models. *Tellus*, **51A**, 167–194, doi:10.1034/j.1600-0870.1999.t01-2-00002.x.

Moore, A. M., H. G. Arango, E. Di 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 Modell.*, **7**, 227–258, doi:10.1016/j.ocemod.2003.11.001.

Nerger, L., S. Schulte, and A. Bunse-Gerstner, 2014: On the influence of model nonlinearity and localization on ensemble Kalman smoothing. *Quart. J. Roy. Meteor. Soc.*, **140**, 2249–2259, doi:10.1002/qj.2293.

Raanes, P. N., 2016: On the ensemble Rauch–Tung–Striebel smoother and its equivalence to the ensemble Kalman smoother. *Quart. J. Roy. Meteor. Soc.*, **142**, 1259–1264, doi:10.1002/qj.2728.

Rauch, H. E., C. Striebel, and F. Tung, 1965: Maximum likelihood estimates of linear dynamic systems. *AIAA J.*, **3**, 1445–1450, doi:10.2514/3.3166.

Ravela, S., and D. McLaughlin, 2007: Fast ensemble smoothing. *Ocean Dyn.*, **57**, 123–134, doi:10.1007/s10236-006-0098-6.

Robinson, A. R., P. F. J. Lermusiaux, and N. Q. Sloan III, 1998: Data assimilation. *The Global Coastal Ocean: Processes and Methods*, K. H. Brink and A. R. Robinson, Eds., The Sea—Ideas and Observations on Progress in the Study of the Seas, Vol. 10, John Wiley and Sons, 541–594.

Sapsis, T. P., 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.

——, and ——, 2012: 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.

Särkkä, S., 2013: *Bayesian Filtering and Smoothing*. Institute of Mathematical Statistics Textbooks, Vol. 3, Cambridge University Press, 232 pp.

Sasaki, Y., 1970: Some basic formalisms in numerical variational analysis. *Mon. Wea. Rev.*, **98**, 875–883, doi:10.1175/1520-0493(1970)098<0875:SBFINV>2.3.CO;2.

Skandrani, C., J.-M. Brankart, N. Ferry, J. Verron, P. Brasseur, and B. Barnier, 2009: Controlling atmospheric forcing parameters of global ocean models: Sequential assimilation of sea surface Mercator-Ocean reanalysis data. *Ocean Sci.*, **5**, 403–419, doi:10.5194/os-5-403-2009.

Sondergaard, T., and P. F. J. Lermusiaux, 2013a: Data assimilation with Gaussian mixture models using the dynamically orthogonal field equations. Part I: Theory and scheme. *Mon. Wea. Rev.*, **141**, 1737–1760, doi:10.1175/MWR-D-11-00295.1.

——, and ——, 2013b: Data assimilation with Gaussian mixture models using the dynamically orthogonal field equations. Part II: Applications. *Mon. Wea. Rev.*, **141**, 1761–1785, doi:10.1175/MWR-D-11-00296.1.

Stammer, D., and Coauthors, 2002: Global ocean circulation during 1992–1997, estimated from ocean observations and a general circulation model. *J. Geophys. Res.*, **107**, 3118, doi:10.1029/2001JC000888.

Tagade, P., H. Seybold, and S. Ravela, 2014: Mixture ensembles for data assimilation in dynamic data-driven environmental systems. *Proc. Comput. Sci.*, **29**, 1266–1276, doi:10.1016/j.procs.2014.05.114.

Talagrand, O., 1997: Assimilation of observations, an introduction. *J. Meteor. Soc. Japan*, **75** (1B), 191–209.

Thrun, S., W. Burgard, and D. Fox, 2005: *Probabilistic Robotics*. The MIT Press, 647 pp.

Ueckermann, M. P., P. F. J. Lermusiaux, and T. P. Sapsis, 2013: Numerical schemes for dynamically orthogonal equations of stochastic fluid and ocean flows. *J. Comput. Phys.*, **233**, 272–294, doi:10.1016/j.jcp.2012.08.041.

Van Leeuwen, P. J., and G. Evensen, 1996: Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Wea. Rev.*, **124**, 2898–2913, doi:10.1175/1520-0493(1996)124<2898:DAAIMI>2.0.CO;2.

Vo, B.-N., B.-T. Vo, and R. P. Mahler, 2012: Closed-form solutions to forward–backward smoothing. *IEEE Trans. Signal Process.*, **60**, 2–17, doi:10.1109/TSP.2011.2168519.

Weir, B., R. Miller, and Y. Spitz, 2013: A potential implicit particle method for high-dimensional systems. *Nonlinear Processes Geophys.*, **20**, 1047–1060, doi:10.5194/npg-20-1047-2013.

Willsky, A. S., 2002: Multiresolution Markov models for signal and image processing. *Proc. IEEE*, **90**, 1396–1458, doi:10.1109/JPROC.2002.800717.

Wunsch, C., 1996: *The Ocean Circulation Inverse Problem*. Cambridge University Press, 442 pp.

——, and P. Heimbach, 2007: Practical global oceanic state estimation. *Physica D*, **230**, 197–208, doi:10.1016/j.physd.2006.09.040.