# Efficient Nonlinear Optimal Smoothing and Sampling Algorithms for Complex Turbulent Nonlinear Dynamical Systems with Partial Observations

Nan Chen

*Department of Mathematics, University of Wisconsin-Madison, Madison, Wisconsin, USA*

Andrew J. Majda

*Department of Mathematics and Center for Atmosphere Ocean Science, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA*
*Center for Prototype Climate Modeling, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, UAE*

**Abstract**

A nonlinear optimal smoother and an associated optimal strategy of sampling hidden model trajectories are developed for a rich class of complex nonlinear turbulent dynamical systems with partial and noisy observations. Despite the strong nonlinearity and the significant non-Gaussian characteristics in the underlying systems, both the optimal smoother estimates and the sampled trajectories can be solved via closed analytic formulae. Thus, they are computationally efficient and the methods are applicable to high-dimensional systems. The nonlinear optimal smoother is able to estimate the hidden model states associated with various non-Gaussian phenomena and is particularly skillful in capturing the onset, demise and amplitude of the observed and hidden extreme events. On the other hand, the sampled hidden trajectories succeed in recovering both the dynamical and statistical features of the underlying nonlinear systems, including the fat-tailed non-Gaussian probability density function and the temporal autocorrelation function. In the situations with only a short period

*Corresponding author: Nan Chen
*Email addresses:* chennan@math.wisc.edu (Nan Chen), jonjon@cims.nyu.edu (Andrew J. Majda)

of partially observed training time series, the optimal sampling strategy can be used to efficiently create a sufficient number of samples in an unbiased fashion that facilitates an accurate prediction of important non-Gaussian features in both the observed and hidden variables. In addition, the information provided by the sampled trajectories based on imperfect models allows an effective way of quantifying the model error. It also offers a systematic approach to improve approximate models and stochastic parameterizations in highly non-Gaussian systems and thus advances the real-time forecasts.

*Keywords:* Nonlinear optimal smoothing, Optimal backward sampling, Complex Nonlinear Turbulent Systems, Hidden variables, Extreme events, Model error

*2010 MSC:* 93E14, 62M20, 65C30, 34F05, 37M10

## 1. Introduction

Complex nonlinear turbulent dynamical systems [1, 2, 3, 4] are characterized by multiscale spatiotemporal structures, strong nonlinear interactions between different variables and across different scales, and significant non-Gaussian behavior such as the fat-tailed probability density function (PDF), intermittency and extreme events [5, 6, 7]. High dimensionality and model error are also the common issues in the study of these nonlinear dynamical systems. Due to the complexity in the turbulent systems, observations are often combined with dynamical models in reducing the model bias and model uncertainty. However, only partial and noisy observations are available in many practical applications. Despite the lack of the observational data, the unresolved variables nevertheless play a crucial role in transferring energy with the observed or resolved variables in a highly nonlinear way. These unobserved variables are also able to trigger various non-Gaussian phenomena including extreme events in both the resolved and unresolved scales. Therefore, estimating the states and recovering the nonlinear and non-Gaussian dynamical and statistical features of the hidden variables are central topics in studying complex nonlinear dynamical systems.

Filtering is one of the widely used methods for state estimation [8, 9, 10, 11, 12]. It utilizes the information up to the current time instant and thus has the advantage of providing an improved initial value for real-time prediction. However, the state estimation based on the information only in the past can be biased, especially for detecting the triggering phases and the nonlinear response of various non-Gaussian phenomena such as intermittency and extreme events. In addition, many important path-wise properties of intermittent trajectories, which involve crucial nonlinear dynamical features of the underlying system, are usually not well represented by the biased statistical description via the filter estimates. Therefore, for the purpose of an off-line optimal estimation of the hidden states, using the information of the entire observational period is a more appropriate choice. This is known as the smoothing technique [13, 14, 12]. In addition to improving the statistical state estimation, the optimal smoother estimates can be further applied to the development of effective methods for sampling the missing trajectories of the hidden variables, which involve both the path-wise and statistical characteristics of the underlying model. The smoother technique typically contains a forward pass using a certain filtering method followed by a backward pass to obtain the optimal smoothed state estimates. For linear models with Gaussian noise, Kalman filter [15] is often used as the forward pass and different smoothers have been developed, such as the Rauch-Tung-Striebel (RTS) and the Bryson-Frazier smoothers [13, 16]. Unfortunately, there is no general closed analytic form for optimal smoothers associated with complex nonlinear systems. Particle methods have to be used in order to obtain the nonlinear smoother estimates. However, these particle methods typically suffer from the curse of dimensionality [17] and are thus difficult to apply to high-dimensional complex turbulent dynamical systems. On the other hand, applying linear optimal smoothers as approximations to nonlinear turbulent systems often fails to capture crucial nonlinear and non-Gaussian features. Such linearizations may also lead to severe model divergence and bring about unstable dynamical behavior.

This article aims at developing a nonlinear optimal smoother and an associ-

ated optimal strategy of sampling the hidden model trajectories for a rich class of complex nonlinear turbulent dynamical models. In light of the model structure, the nonlinear optimal smoother and the optimal sampling strategy can be solved via closed analytic formulae and therefore they are computationally efficient and the methods are applicable to high-dimensional nonlinear turbulent dynamical systems. It is shown that the estimated states from the nonlinear optimal smoother provide a more accurate description of the non-Gaussian features than the nonlinear optimal filter estimates, including the timing, duration and amplitudes of the hidden extreme events. On the other hand, the optimal sampling technique is extremely useful in obtaining both the statistical and path-wise properties of the unobserved or unresolved variables. It is also able to provide extra dynamical information of the hidden processes which cannot be fully described by the filter and smoother estimates, such as the temporal correlation of the underlying systems. Note that although this optimal sampling strategy is not directly applicable in an online fashion, the dynamical and statistical information provided by the resulting sampled trajectories can be adopted to systematically improve the approximate models and the stochastic parameterizations of unresolved variables, which advance the reduction of the error and uncertainty in real-time forecasts. In addition, in many practical applications only a limited size of observations is available, which is not enough to accurately recover the fat-tailed PDFs and other non-Gaussian statistical quantities in complex nonlinear dynamical systems. This optimal sampling strategy can then be used to create a sufficient number of data in an unbiased fashion for both the observed and hidden variables, which facilitates the statistical description of various significant non-Gaussian features. Note that the sampled trajectories from this optimal sampling strategy is particularly useful in the presence of model error. In fact, due to the extra information provided by observations, the resulting statistics from the sampled trajectories are often much more accurate than those from a free run of the imperfect model.

The class of the complex nonlinear turbulent models used here is the so-called conditional Gaussian nonlinear models [18, 19]. These systems are highly nonlin-

ear and non-Gaussian, where both the joint and marginal PDFs can be skewed with fat tails. Extreme events, intermittency and highly nontrivial nonlinear interactions between different variables all appear in the conditional Gaussian systems. The name conditional Gaussian comes from the fact that once the trajectories of a subset of the variables are known, the statistics of the remaining variables conditioned on the trajectories of these known ones are Gaussian. The conditional Gaussian modeling framework includes a large class the physics-constrained nonlinear low-order stochastic models [20, 21], many stochastically coupled reaction-diffusion models in neuroscience and ecology [22, 23], and quite a few important large-scale dynamical models in turbulence, fluids and geophysical flows [24, 25]. A gallery of examples of conditional Gaussian systems can be found in [18].

The remaining of this article is organized as follows. Section 2 presents the general form of the conditional Gaussian nonlinear models and the corresponding nonlinear filter estimates. Section 3 focuses on the development of the nonlinear optimal smoother and the nonlinear optimal strategy of sampling hidden model trajectories, where the theories are built for both the continuous and discrete time dynamics. Section 4 illustrates several important applications of the nonlinear optimal filter, smoother and sampling strategy, including recovering the nonlinear dynamics and non-Gaussian statistics of complex nonlinear systems, state estimation of extreme events, effective sampling and predicting the fat-tailed PDFs with very short observational training data and improving the stochastic parameterizations using models with multiplicative noise. Both the perfect model setup and the tests in the presence of model error are studied here. The article is concluded in Section 5. All the proofs are shown in Appendix.

## 2. Conditional Gaussian Nonlinear Systems

The general form of the conditional Gaussian nonlinear systems is as follows [26, 19, 18],

$$\mathrm{d}\mathbf{X}(t) = \Big[\mathbf{A_0}(\mathbf{X},t) + \mathbf{A_1}(\mathbf{X},t)\mathbf{Y}(t)\Big]\,\mathrm{d}t + \mathbf{B_1}(\mathbf{X},t)\,\mathrm{d}\mathbf{W_1}(t) + \mathbf{B_2}(\mathbf{X},t)\,\mathrm{d}\mathbf{W_2}(t), \tag{1a}$$

$$\mathrm{d}\mathbf{Y}(t) = \Big[\mathbf{a_0}(\mathbf{X},t) + \mathbf{a_1}(\mathbf{X},t)\mathbf{Y}(t)\Big]\,\mathrm{d}t + \mathbf{b_1}(\mathbf{X},t)\,\mathrm{d}\mathbf{W_1}(t) + \mathbf{b_2}(\mathbf{X},t)\,\mathrm{d}\mathbf{W_2}(t), \tag{1b}$$

where the vector $\mathbf{X}$ stands for the observed variables while $\mathbf{Y}$ is the collection of the unobserved variables. In (1), $\mathbf{A}_0, \mathbf{A}_1, \mathbf{a}_0, \mathbf{a}_1, \mathbf{B}_1, \mathbf{B}_2, \mathbf{b}_1$ and $\mathbf{b}_2$ and are vectors and matrices that depend nonlinearly on the state variables $\mathbf{X}$ and time $t$ while $\mathbf{W_1}$ and $\mathbf{W_2}$ are independent white noise. With these two independent noise sources, the system in (1) is a generalized version of those in [19, 18]. For the notation simplicity, we remove the explicit dependence of $\mathbf{X}$ and $t$ in the matrices and vectors in (1). That is, we denote $\mathbf{A_0} := \mathbf{A_0}(\mathbf{X},t)$ and the same for other matrices and vectors.

Despite the conditional Gaussianity, the coupled system (1) remains highly nonlinear and is able to capture the non-Gaussian features as in nature. This conditional Gaussian nonlinear modeling framework includes many physics-constrained nonlinear stochastic models [20, 21], large-scale dynamical models in turbulence, fluids and geophysical flows [24, 25], as well as stochastically coupled reaction-diffusion models in neuroscience and ecology [22, 23]. See a recent work [18] for a gallery of examples of the conditional Gaussian systems. Applications of the conditional Gaussian systems to strongly nonlinear systems include developing low-order nonlinear stochastic models for predicting the intermittent time series of the Madden-Julian oscillation (MJO) and the monsoon intraseasonal variabilities [27, 28, 29, 30], filtering the stochastic skeleton model for the MJO [31], and recovering the turbulent ocean flows with noisy observations from Lagrangian tracers [32, 33, 34]. Other studies that also fit into the conditional Gaussian framework includes the cheap exactly solvable forecast models

in dynamic stochastic superresolution of sparsely observed turbulent systems [35, 36], stochastic superparameterization for geophysical turbulence [37] and blended particle filters for large-dimensional chaotic systems [38].

One important feature of the conditional Gaussian nonlinear system (1) is the following.

**Theorem 2.1.** *Given one realization of the time series* $\mathbf{X}(t)$ *for* $t \in [0, t]$*, the conditional distribution*

$$p(\mathbf{Y}(t)|\mathbf{X}(s), s \le t) \sim \mathcal{N}(\boldsymbol{\mu}(t), \mathbf{R}(t)) \tag{2}$$

*is Gaussian, where the conditional mean* $\boldsymbol{\mu}$ *and the conditional covariance* $\mathbf{R}$ *are given by the following explicit formulae*

$$\mathrm{d}\boldsymbol{\mu} = (\mathbf{a_0} + \mathbf{a_1}\boldsymbol{\mu})\,\mathrm{d}t + (\mathbf{b} \circ \mathbf{B} + \mathbf{R}\mathbf{A}_\mathbf{1}^*)(\mathbf{B} \circ \mathbf{B}^*)^{-1}(\,\mathrm{d}\mathbf{X} - (\mathbf{A_0} + \mathbf{A_1}\boldsymbol{\mu})\,\mathrm{d}t), \tag{3a}$$

$$\mathrm{d}\mathbf{R} = \big(\mathbf{a_1}\mathbf{R} + \mathbf{R}\mathbf{a}_\mathbf{1}^* + \mathbf{b} \circ \mathbf{b} - (\mathbf{b} \circ \mathbf{B} + \mathbf{R}\mathbf{A}_\mathbf{1}^*)(\mathbf{B} \circ \mathbf{B})^{-1}(\mathbf{b} \circ \mathbf{B} + \mathbf{A_1}\mathbf{R})\big)\,\mathrm{d}t, \tag{3b}$$

*with*

$$\mathbf{b} \circ \mathbf{b} = \mathbf{b_1}\mathbf{b}_\mathbf{1}^* + \mathbf{b_2}\mathbf{b}_\mathbf{2}^*,$$
$$\mathbf{b} \circ \mathbf{B} = \mathbf{b_1}\mathbf{B}_\mathbf{1}^* + \mathbf{b_2}\mathbf{B}_\mathbf{2}^*,$$
$$\mathbf{B} \circ \mathbf{B} = \mathbf{B_1}\mathbf{B}_\mathbf{1}^* + \mathbf{B_2}\mathbf{B}_\mathbf{2}^*.$$

See [26] for the proof of Theorem 2.1. The formulae in (3) are the nonlinear optimal filter estimates for the conditional Gaussian nonlinear systems, where the conditional covariance is driven by a random Riccati equation. The conditional mean $\boldsymbol{\mu}$ and the conditional covariance $\mathbf{R}$ in (3) are also named as posterior mean and posterior covariance or filter mean and filter covariance.

In addition to the applications in effective data assimilation and real-time forecast, these filter estimates in (3) play an important role in deriving the explicit formulae for the nonlinear optimal smoothing and the optimal strategy of sampling the unobserved model trajectories.

## 3. Conditional Gaussian Nonlinear Optimal Smoother and Backward Sampling

### 3.1. Continuous time dynamics

Consider the conditional Gaussian nonlinear systems in (1). Assume one realization of $\mathbf{X}(t)$ for $t \in [0, T]$ is available.

For the convenience of discussion, the statement below starts with a discrete approximation of the original nonlinear continuous system in time by adopting an Euler-Maruyama scheme [39]. Thus, the values of $\mathbf{X}$ and $\mathbf{Y}$ are taken at discrete points in time $\{\widetilde{\mathbf{X}}^0, \ldots, \widetilde{\mathbf{X}}^j, \ldots, \widetilde{\mathbf{X}}^J\}$ and $\{\widetilde{\mathbf{Y}}^0, \ldots, \widetilde{\mathbf{Y}}^j, \ldots, \widetilde{\mathbf{Y}}^J\}$, where $\widetilde{\mathbf{X}}^j := \mathbf{X}(t_j)$ and $\widetilde{\mathbf{Y}}^j = \mathbf{Y}(t_j)$. So do the resulting statistical estimates. Here, the variable with tilde and superscript $j$, namely $\widetilde{\cdot}^j$, denotes the discrete approximation of its continuous form at time $t_j$, where the entire time interval $[0, T]$ is divided into $J$ equipartition subintervals with $0 = t_0, t_1, t_2, \ldots, t_J = T$. Denote $\Delta t = t_{j+1} - t_j$ and therefore $J\Delta t = T$. In the analysis of the system with the discrete approximation, $J$ is assumed to be a large finite number (or equivalently $\Delta t$ is a small but finite quantity). Eventually, the limit $\Delta t \to 0$ is taken for the discrete approximation to retrieve the original continuous dynamics.

We start with the following Lemma, which is the basis for the development of both the nonlinear optimal smoother and the optimal backward sampling strategy.

**Lemma 3.1.** *The conditional distribution*

$$p(\widetilde{\mathbf{Y}}^j | \widetilde{\mathbf{Y}}^{j+1}, \widetilde{\mathbf{X}}^s, s \le j) \sim \mathcal{N}(\widetilde{\mathbf{m}}^j, \widetilde{\mathbf{P}}^j) \tag{4}$$

*is Gaussian, where the conditional mean* $\widetilde{\mathbf{m}}^j$ *and conditional covariance* $\widetilde{\mathbf{P}}^j$ *satisfy the following equations*

$$\widetilde{\mathbf{m}}^j = \widetilde{\boldsymbol{\mu}}^j + \widetilde{\mathbf{C}}^j\big(\widetilde{\mathbf{Y}}^{j+1} - \widetilde{\mathbf{a}}_{\mathbf{0}}^j \Delta t - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\boldsymbol{\mu}}^j\big), \tag{5a}$$

$$\widetilde{\mathbf{P}}^j = \widetilde{\mathbf{R}}^j - \widetilde{\mathbf{C}}^j\big(\widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j \Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\mathbf{R}}^j (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)^*\big)(\mathbf{C}^j)^*, \tag{5b}$$

*and the auxiliary matrix* $\mathbf{C}$ *is given by*

$$\widetilde{\mathbf{C}}^j = \widetilde{\mathbf{R}}^j (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)^* \big(\widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j \Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\mathbf{R}}^j (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j)^*\big)^{-1}. \tag{6}$$

With Lemma 3.1 in hand, the nonlinear optimal smoother estimate of the conditional Gaussian nonlinear system (1) is given as follows.

**Theorem 3.2** (Optimal Nonlinear Smoother)**.** *Given one realization of the observed variable* $\mathbf{X}(t)$ *for* $t \in [0,T]$*, the optimal smoother estimate* $p(\mathbf{Y}(t)|\mathbf{X}(s), s \in [0,T])$ *is conditional Gaussian,*

$$p(\mathbf{Y}(t)|\mathbf{X}(s), s \in [0,T]) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{s}}(t), \mathbf{R}_{\mathbf{s}}(t)), \tag{7}$$

*where the conditional mean* $\boldsymbol{\mu}_{\mathbf{s}}(t)$ *and conditional covariance* $\mathbf{R}_{\mathbf{s}}(t)$ *of the smoother at time* $t_j$ *satisfy the following equations*

$$\boldsymbol{\mu}_{\mathbf{s}}(t_j) = \lim_{\Delta t \to 0} \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^j = \widetilde{\boldsymbol{\mu}}^j + \widetilde{\mathbf{C}}^j \big( \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \widetilde{\mathbf{a}}_{\mathbf{0}}^j \Delta t - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t) \widetilde{\boldsymbol{\mu}}^j \big) \tag{8a}$$

$$\mathbf{R}_{\mathbf{s}}(t_j) = \lim_{\Delta t \to 0} \widetilde{\mathbf{R}}_{\mathbf{s}}^j = \widetilde{\mathbf{R}}^j + \widetilde{\mathbf{C}}^j \big( \widetilde{\mathbf{R}}_{\mathbf{s+1}}^j - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t) \widetilde{\mathbf{R}}^j (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)^* - \widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j \Delta t \big) (\widetilde{\mathbf{C}}^j)^*, \tag{8b}$$

*and* $\widetilde{\mathbf{C}}^j$ *is the same as that in* (6)*. Note that the optimal smoother estimate and the optimal filter estimate at the end point* $t = T$ *are the same, namely* $\boldsymbol{\mu}_{\mathbf{s}}(T) = \boldsymbol{\mu}(T)$ *and* $\mathbf{R}_{\mathbf{s}}(T) = \mathbf{R}(T)$*.*

Theorem 3.2 provides an optimal way of estimating both the state and the associated uncertainty at each time instant $t_j$ based on one realization of the entire observational time series $\mathbf{X}(t)$ with $t \in [0,T]$. Another important topic in studying the nonlinear turbulent dynamical system is to recover the dynamical features of the unobserved process, which, in addition to the estimation of these statistical quantities, also requires an efficient and accurate recovery of the path-wise information of the unobserved variable $\mathbf{Y}$ that contains the temporal dependence of $\mathbf{Y}$ at different time instants. The closed analytic formula in the following theorem provides an effective way of sampling the unobserved trajectories of $\mathbf{Y}$ conditioned on the given time series $\mathbf{X}(t)$ with $t \in [0,T]$, which facilitates the study of various hidden nonlinear dynamical features including the intermittency and extreme events in complex nonlinear dynamical systems.

**Theorem 3.3** (Optimal Backward Sampling Formula)**.** *Conditioned on one realization of the observed variable* $\mathbf{X}(t)$ *for* $t \in [0,T]$*, the optimal strategy of*

*sampling the trajectories associated with the unobserved variable* $\mathbf{Y}$ *satisfies the following explicit formula,*

$$\mathrm{d}(-\mathbf{Y}) = \big(-\mathbf{a_0}-\mathbf{a_1}\mathbf{Y}\big)\,\mathrm{d}t+(\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1}\big(\boldsymbol{\mu}-\mathbf{Y}\big)\,\mathrm{d}t+\mathbf{b_1}\,\mathrm{d}\mathbf{W}_{\mathbf{Y},\mathbf{1}}+\mathbf{b_2}\,\mathrm{d}\mathbf{W}_{\mathbf{Y},\mathbf{2}}, \quad (9)$$

*where* $\boldsymbol{\mu}(t)$ *and* $\mathbf{R}(t)$ *are the conditional mean and conditional covariance from the filter estimates in* (3)*, and* $\mathbf{W}_{\mathbf{Y},\mathbf{1}}$ *and* $\mathbf{W}_{\mathbf{Y},\mathbf{2}}$ *are independent white noise sources. In* (9)*, the left hand side is understood as*

$$\mathrm{d}(-\mathbf{Y}) = \lim_{\Delta t\to 0} \mathbf{Y}(t) - \mathbf{Y}(t+\Delta t)$$

*while* $\mathbf{Y}$ *on the right hand side takes values at* $t+\Delta t$ *and the other coefficients are given at time* $t$*.*

*The formula* (9) *starts from* $t = T$ *and it is run backwards towards* $t = 0$*. Therefore, it is named as a backward sampling formula. The initial value of* $\mathbf{Y}$ *in* (9) *is drawn from the conditional Gaussian distribution* $\mathcal{N}(\boldsymbol{\mu}(T), \mathbf{R}(T))$*.*

*Remark.* Comparing with the true underlying dynamics of $\mathbf{Y}$ in (1b), the backward sampling equation (9) involves an extra term $(\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1}\big(\boldsymbol{\mu}-\mathbf{Y}\big)\,\mathrm{d}t$. This correction term plays an important role as a forcing and it drives the sampled trajectory to meander around the filter mean state $\boldsymbol{\mu}$. Yet, due to the memory of the process, the system response of the forcing has a delayed effect. The sampled trajectory $\mathbf{Y}$ actually fluctuates around the smoother mean state $\boldsymbol{\mu}_s$, which is a desirable feature since the optimal smoother estimate makes use of the entire observational information and is thus unbiased. The rigorous justification can be found in the first a few steps of the proof of Theorem 3.3. On the other hand, this correction term has a weight $(\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1}$. If the noise strength in the $\mathbf{Y}$ process is fixed, namely $\mathbf{b}\circ\mathbf{b}$ is a constant matrix, then the amplitude of $\mathbf{R}$ is positively correlated with the noise level of the observational process $\mathbf{X}$. A low noise level in $\mathbf{X}$ implies a small uncertainty in the filter covariance estimate $\mathbf{R}$, which leads to a large weight towards the correction term. In addition, the filter mean estimate $\boldsymbol{\mu}$ in such a situation is largely determined by the observations. As a consequence, the observations play a primary role in creating the sampled trajectories. Another important feature of the backward sampling equation (9)

is that it retains the dynamical structures of the true underlying dynamics of $\mathbf{Y}$ in (1b). Therefore the temporal autocorrelation function (ACF) and higher order temporal correlations associated with the underlying nonlinear systems can be accurately recovered using the sampled trajectories. Concrete examples are included in Section 4.

Next, in light of the backward sampling equation (9), there is an alternative way of calculating the nonlinear optimal smoother.

**Theorem 3.4** (An Alternative Way of Calculating the Optimal Nonlinear Smoother)**.** *An alternative way of calculating the optimal smoother is via the following equations,*

$$\mathrm{d}(-\boldsymbol{\mu}_{\mathbf{s}}) = \big( - \mathbf{a_0} - \mathbf{a_1}\boldsymbol{\mu}_{\mathbf{s}} + (\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{s}})\big)\,\mathrm{d}t, \tag{10a}$$

$$\mathrm{d}(-\mathbf{R}_{\mathbf{s}}) = -\big((\mathbf{a_1} + (\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1})\mathbf{R}_{\mathbf{s}} + \mathbf{R}_{\mathbf{s}}(\mathbf{a}_{\mathbf{1}}^{*} + (\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1}) - \mathbf{b}\circ\mathbf{b}\big)\,\mathrm{d}t. \tag{10b}$$

*In* (10)*, the terms of the left hand side are understood as*

$$\mathrm{d}(-\boldsymbol{\mu}_{\mathbf{s}}) = \lim_{\Delta t\to 0} \boldsymbol{\mu}_{\mathbf{s}}(t) - \boldsymbol{\mu}_{\mathbf{s}}(t+\Delta t)$$

$$\mathrm{d}(-\mathbf{R}_{\mathbf{s}}) = \lim_{\Delta t\to 0} \mathbf{R}_{\mathbf{s}}(t) - \mathbf{R}_{\mathbf{s}}(t+\Delta t)$$

*while* $\mathbf{Y}$ *on the right hand side of* (10) *takes values at* $t+\Delta t$ *and the other coefficients are given at* $t$*. The starting value of the nonlinear smoother* $(\boldsymbol{\mu}_{\mathbf{s}}(T), \mathbf{R}_{\mathbf{s}}(T))$ *is the same as the filter estimate at the endpoint* $(\boldsymbol{\mu}(T), \mathbf{R}(T))$*.*

The nonlinear optimal smoother formulae in Theorem 3.4 are more concise. It can be derived directly from the backward sampling equation (9) by using a mean-fluctuation decomposition [40] (see the proofs in Appendix for details). On the other hand, the optimal smoother formulae in Theorem 3.2 can be understood from a recursive point of view, where the procedure of the derivation is quite useful for finding the nonlinear optimal smoother of the discrete version of conditional Gaussian systems. The formulae in Theorem 3.2 can also be used to derive the Rauch-Tung-Striebel smoother [13] when the underlying system is linear with Gaussian noise, which will be discussed in Section 3.5.

*3.2. Discrete time systems*

As an analog to the continuous time conditional Gaussian systems, the general form of the discrete conditional Gaussian nonlinear models is as follows,

$$\begin{aligned}\mathbf{X}(t_{j+1}) &= \mathbf{A_0}(\mathbf{X}(t_j), t_j) + \mathbf{A_1}(\mathbf{X}(t_j), t_j)\mathbf{Y}(t_j) \\ &\quad + \mathbf{B_1}(\mathbf{X}(t_j), t_j)\boldsymbol{\varepsilon_1}(t_{j+1}) + \mathbf{B_2}(\mathbf{X}(t_j), t_j)\boldsymbol{\varepsilon_2}(t_{j+1}), \quad (11a)\\ \mathbf{Y}(t_{j+1}) &= \mathbf{a_0}(\mathbf{X}(t_j), t_j) + \mathbf{a_1}(\mathbf{X}(t_j), t_j)\mathbf{Y}(t_j) \\ &\quad + \mathbf{b_1}(\mathbf{X}(t_j), t_j)\boldsymbol{\varepsilon_1}(t_{j+1}) + \mathbf{b_2}(\mathbf{X}(t_j), t_j)\boldsymbol{\varepsilon_2}(t_{j+1}), \quad (11b)\end{aligned}$$

where $\boldsymbol{\varepsilon_1}$ and $\boldsymbol{\varepsilon_2}$ are independent white noise sources.

**Theorem 3.5** (Optimal Nonlinear Filter Estimate)**.** *For the discrete system* (11)*, assume a sequence of the observed variable* $\mathbf{X}$*, namely* $\{\mathbf{X}(t_0), \mathbf{X}(t_1), \ldots, \mathbf{X}(t_{j+1})\}$*, is available. Then the distribution of* $\mathbf{Y}(t_{j+1})$ *conditioned on this given observed sequence is conditional Gaussian,*

$$p(\mathbf{Y}(t_{j+1})|\mathbf{X}(s), s \le t_{j+1}) \sim \mathcal{N}(\boldsymbol{\mu}(t_{j+1}), \mathbf{R}(t_{j+1})). \quad (12)$$

*The time evolutions of the conditional mean* $\boldsymbol{\mu}(t_{j+1})$ *and conditional covariance* $\mathbf{R}(t_{j+1})$ *are given by the following explicit formulae,*

$$\begin{aligned}\boldsymbol{\mu}(t_{j+1}) &= \mathbf{a_0} + \mathbf{a_1}\boldsymbol{\mu}(t_j) + (\mathbf{b} \circ \mathbf{B} + \mathbf{a_1}\mathbf{R}(t_i)\mathbf{A_1^*}) \times \\ &\quad (\mathbf{B} \circ \mathbf{B} + \mathbf{A_1}\mathbf{R}(t_i)\mathbf{A_1^*})^{-1}(\mathbf{X}(t_{i+1}) - \mathbf{A_0} - \mathbf{A_1}\boldsymbol{\mu}(t_i)), \quad (13a)\\ \mathbf{R}(t_{j+1}) &= \mathbf{a_1}\mathbf{R}(t_i)\mathbf{a_1^*} + \mathbf{b} \circ \mathbf{b} - (\mathbf{b} \circ \mathbf{B} + \mathbf{a_1}\mathbf{R}(t_i)\mathbf{A_1^*}) \times \\ &\quad (\mathbf{B} \circ \mathbf{B} + \mathbf{A_1}\mathbf{R}(t_i)\mathbf{A_1^*})^{-1}(\mathbf{b} \circ \mathbf{B} + \mathbf{a_1}\mathbf{R}(t_i)\mathbf{A_1^*})^*, \quad (13b)\end{aligned}$$

*where*

$$\begin{aligned}\mathbf{b} \circ \mathbf{b} &= \mathbf{b_1}\mathbf{b_1^*} + \mathbf{b_2}\mathbf{b_2^*},\\ \mathbf{b} \circ \mathbf{B} &= \mathbf{b_1}\mathbf{B_1^*} + \mathbf{b_2}\mathbf{B_2^*},\\ \mathbf{B} \circ \mathbf{B} &= \mathbf{B_1}\mathbf{B_1^*} + \mathbf{B_2}\mathbf{B_2^*},\end{aligned}$$

*and all the matrices and vectors* $\mathbf{a_0}, \mathbf{a_1}, \mathbf{A_0}, \mathbf{A_1}, \mathbf{b_1}, \mathbf{b_2}, \mathbf{B_1}$ *and* $\mathbf{B_2}$ *are taking values at time* $t_j$*.*

Below, for the notation simplicity, we denote $\mathbf{Y}^j := \mathbf{Y}(t_j)$ and the same applies for other variables, matrices and vectors. Assume the total length of the observed sequence $\{\mathbf{X}^s\}$ is $n+1$.

**Theorem 3.6** (Optimal Nonlinear Smoother Estimate)**.** *Given one sequence of the observed variable* $\{\mathbf{X}^0, \ldots, \mathbf{X}^n\}$*, the nonlinear optimal smoother estimate* $p(\mathbf{Y}^j|\mathbf{X}^s, 0 \le s \le n)$ *is conditional Gaussian,*

$$p(\mathbf{Y}^j|\mathbf{X}^s, 0 \le s \le n) \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{s}^j, \mathbf{R}_\mathbf{s}^j),$$

*where the conditional mean* $\boldsymbol{\mu}_\mathbf{s}^j$ *and conditional covariance* $\mathbf{R}_\mathbf{s}^j$ *of the smoother are given by*

$$\boldsymbol{\mu}_\mathbf{s}^j = \boldsymbol{\mu}^j + \mathbf{C}^j(\boldsymbol{\mu}_\mathbf{s}^{j+1} - \mathbf{a}_\mathbf{0}^j - \mathbf{a}_\mathbf{1}^j\boldsymbol{\mu}^j), \tag{14a}$$

$$\mathbf{R}_\mathbf{s}^j = \mathbf{R}^j + \mathbf{C}^j(\mathbf{R}_\mathbf{s}^{j+1} - \mathbf{a}_\mathbf{1}^j\mathbf{R}^j(\mathbf{a}_\mathbf{1}^j)^* - \mathbf{b}\circ\mathbf{b})(\mathbf{C}^j)^*, \tag{14b}$$

*with*

$$\mathbf{C}^j = \mathbf{R}^j(\mathbf{a}_\mathbf{1}^j)^*(\mathbf{b}\circ\mathbf{b} + \mathbf{a}_\mathbf{1}^j\mathbf{R}^j(\mathbf{a}_\mathbf{1}^j)^*)^{-1} \tag{15}$$

*The optimal smoother is calculated backwards from* $s = n$ *to* $s = 0$*. The starting value of the smoother estimate* $(\boldsymbol{\mu}_\mathbf{s}^n, \mathbf{R}_\mathbf{s}^n)$ *is the same as the filter estimate at the endpoint* $(\boldsymbol{\mu}^n, \mathbf{R}^n)$*.*

**Theorem 3.7** (Optimal Backward Sampling Formula)**.** *Given one sequence of the observed variable* $\{\mathbf{X}^0, \ldots, \mathbf{X}^n\}$*, the optimal sample of a sequence of the unobserved variable* $\mathbf{Y}$ *can be drawn using the following explicit formula running backwards in time,*

$$p(\mathbf{Y}^j|\mathbf{Y}^{j+1}, \mathbf{X}^s, 0 \le s \le n) \sim \mathcal{N}(\mathbf{m}^j, \mathbf{P}^j), \tag{16}$$

*where*

$$\begin{aligned} \mathbf{m}^j &= \boldsymbol{\mu}^j + \mathbf{C}^j(\mathbf{Y}^{j+1} - \mathbf{a}_\mathbf{0}^j - \mathbf{a}_\mathbf{1}^j\boldsymbol{\mu}^j), \\ \mathbf{P}^j &= \mathbf{R}^j - \mathbf{C}^j(\mathbf{b}^j\circ\mathbf{b}^j + \mathbf{a}_\mathbf{1}^j\mathbf{R}^j(\mathbf{a}_\mathbf{1}^j)^*)(\mathbf{C}^j)^*, \end{aligned} \tag{17}$$

*and the auxiliary matrix* $\mathbf{C}$ *is given by*

$$\mathbf{C}^j = \mathbf{R}^j(\mathbf{a}_\mathbf{1}^j)^*(\mathbf{b}^j\circ\mathbf{b}^j + \mathbf{a}_\mathbf{1}^j\mathbf{R}^j(\mathbf{a}_\mathbf{1}^j)^*)^{-1}. \tag{18}$$

*3.3. Recovery of the transient and equilibrium PDFs of the unobserved variables*

One important application of the nonlinear optimal smoother and the backward sampling technique is to efficiently recover the PDFs of state variable $\mathbf{Y}$ at both the transient phases and the statistical equilibrium state. The focus here is on the continuous nonlinear conditional Gaussian system (1). It is straightforward to generalize the conclusions here to the discrete time systems.

**Theorem 3.8** (Transient PDF)**.** *Assume there are $L$ independent trajectories of the observed variable $\mathbf{X}(t)$ from $t = 0$ to $t = T$, denoted by $\mathbf{X}_l(t)$ with $l = 1, \ldots, L$. The PDF of $\mathbf{Y}$ at time instant $t$ is given by*

$$p(\mathbf{Y}(t)) = \lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} p\big(\mathbf{Y}(t)|\mathbf{X}_l(0 \le s \le T)\big), \tag{19}$$

*where for each $l$,*

$$p\big(\mathbf{Y}(t)|\mathbf{X}_l(0 \le s \le T)\big) \sim \mathcal{N}(\boldsymbol{\mu}_{l,\mathbf{s}}, \mathbf{R}_{l,\mathbf{s}}),$$

*with $\boldsymbol{\mu}_{l,\mathbf{s}}$ and $\mathbf{R}_{l,\mathbf{s}}$ being the mean and covariance computed from the nonlinear smoother in* (7)*.*

Note that the PDF of $\mathbf{Y}(t)$ can also be calculated using a combination of $L$ conditional Gaussian distributions from the filter estimates, which allows a real-time forecast of the PDF of $\mathbf{Y}$ [41]. In [42], it has been shown that only a small number of $L$ is needed for recovering $p(\mathbf{Y}(t))$ based on the filtered solutions regardless of the dimension of $\mathbf{Y}$. In addition, a hybrid strategy facilitates the recovery of the joint PDF $p(\mathbf{X}(t), \mathbf{Y}(t))$ in high-dimensional systems and overcomes the curse of dimensionality [43, 41]. Parallel theories can be built here using the smoother estimates.

If the coupled system (1) is ergodic, then an efficient way of computing the equilibrium PDF of $\mathbf{Y}$ is given as follows.

**Corollary 3.9** (Equilibrium PDF)**.** *Assume a long trajectory of $\mathbf{X}$ from $t = 0$ to $t = T$ is available. The equilibrium PDF of $\mathbf{Y}$, denoted by $p(\mathbf{Y}_\infty)$ is given by*

$$p(\mathbf{Y}_\infty) = \lim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} p\big(\mathbf{Y}(t_i)|\mathbf{X}(0 \le s \le T)\big), \tag{20}$$

*where all* $t_i$ *with* $0 \leq t_1 \leq \ldots \leq t_I \leq T$ *are distributed between* $t = 0$ *and* $t = T$ *with equal distance.*

The sampled trajectory from the backward sampling technique in (3.3) can be regarded as a path that fluctuates around the mean state of the smoother. The amplitude of the fluctuation is determined by the associated covariance at different time instants. Despite the temporal correlation between different points in the sampled trajectory, each point at time $t_i$ can be regarded as a sample from $p\big(\mathbf{Y}(t_i)|\mathbf{X}(0 \leq s \leq T)\big)$. Therefore, an even simpler way of recovering the equilibrium PDF of $\mathbf{Y}$ is given as follows.

**Corollary 3.10** (Equilibrium PDF; An Alternative Method)**.** *Under the same condition as Corollary* (3.9)*, an alternative way of solving the equilibrium PDF of* $\mathbf{Y}$ *is by collecting all the points in the trajectory calculated from the backward sampling equation* (9)*.*

*3.4. The temporal autocorrelation function (ACF).*

Autocorrelation is the correlation of a signal with a delayed copy of itself, as a function of delay [44]. For a zero mean and stationary random process $u$, the autocorrelation function (ACF) can be calculated as

$$\mathrm{ACF}(t) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{u(t+\tau)u^*(\tau)}{\mathrm{Var}(u)} d\tau, \tag{21}$$

where $\cdot^*$ denotes the complex conjugate. The ACF has been widely used to measure the system memory. It also plays an important role in improving the linear response via the fluctuation-dissipation theorem [45, 46]. If the perfect model and the approximate model share the similar ACFs, then the two systems usually have a similar dynamical behavior at least up to the second order statistics. However, for nonlinear and chaotic systems, high order statistics may play an important roles for extreme events. Therefore, the ACF can only be regarded as a crude indicator of the overall predictability of the underlying system. As a remark, the information theory is able to provide a rigorous and practical way to quantify the error in the two ACFs associated with the perfect

and approximate models by making use of their spectral representations. See [47, 48] for details.

In the study below, the ACFs associated with the sampled trajectories from the backward sampling strategy are compared with that of the truth. Such a comparison allows us to understand the skill of recovering the system memory and capturing important dynamical behavior using the sampled trajectories, which are not indicated by the equilibrium PDFs.

### *3.5. Special cases*

Recall that the general nonlinear conditional Gaussian models (1) allow the matrices and vectors $\mathbf{A_0}, \mathbf{A_1}, \mathbf{a_0}, \mathbf{a}_1, \mathbf{B_1}, \mathbf{B_2}, \mathbf{b_1}$ and $\mathbf{b_2}$ on the right hand side depend on the state variable $\mathbf{X}$ and such dependence can be highly nonlinear. In addition, the state variables $\mathbf{X}$ and $\mathbf{Y}$ have a mutual influence with each other in the coupled system, and the random noise are also coupled in the processes of $\mathbf{X}$ and $\mathbf{Y}$. In this subsection, several special and simplified cases of the general nonlinear conditional Gaussian models for filtering and smoothing are illustrated.

#### *3.5.1. The Kalman-Bucy model*

A special case of the conditional Gaussian nonlinear models is the Kalman-Bucy model [49, 50], which was originally proposed for continuous time filtering. It involves three simplifications of the general nonlinear conditional Gaussian framework. First, the variable $\mathbf{X}$ in the Kalman-Bucy model is regarded as the observation, which is a function of the variable $\mathbf{Y}$ that describes the underlying model. However, the observation does not influence the model itself. In other words, there is only an one-way interaction between the two variables, and such interaction is from $\mathbf{Y}$ to $\mathbf{X}$. Second, the Kalman-Bucy model is designed for linear system with linear observations. This means all the matrices and vectors have no dependence on $\mathbf{X}$, which is a significant simplification from the general nonlinear conditional Gaussian framework. Third, the noises in the $\mathbf{X}$ and $\mathbf{Y}$ processes of the Kalman-Bucy model are no longer coupled with each other.

In fact, the noise in the $\mathbf{X}$ process is the observational noise while that in $\mathbf{Y}$ represents the intrinsic small-scale variability of the underlying model. They are naturally independent and the noise coefficients are state-independent as well. Thus, the Kalman-Bucy model is given by

$$\mathrm{d}\mathbf{X}(t) = \Big[\mathbf{A_0}(t) + \mathbf{A_1}(t)\mathbf{Y}(t)\Big]\,\mathrm{d}t + \mathbf{B_2}(t)\,\mathrm{d}\mathbf{W_2}(t), \tag{22a}$$

$$\mathrm{d}\mathbf{Y}(t) = \Big[\mathbf{a_0}(t) + \mathbf{a_1}(t)\mathbf{Y}(t)\Big]\,\mathrm{d}t + \mathbf{b_1}(t)\,\mathrm{d}\mathbf{W_1}(t), \tag{22b}$$

Since the Kalman-Bucy model is a simple and special case of the general conditional Gaussian nonlinear framework (1), The filtering, smoothing and backward sampling formulae have the same forms as those appearing in Theorems 2.1, 3.3 and 3.4. The only difference in the formality is that $\mathbf{b} \circ \mathbf{b} = \mathbf{b_1}\mathbf{b_1^*}$, $\mathbf{B} \circ \mathbf{B} = \mathbf{B_2}\mathbf{B_2^*}$ and $\mathbf{b} \circ \mathbf{B} = 0$. However, in the general conditional Gaussian nonlinear framework, the filter estimate (3) involves solving a random Riccati equation for the covariance, which is not the case in the Kalman-Bucy model since all the coefficients of the filter estimate are state independent.

*3.5.2. The Kalman filter*

The classical Kalman filter [15] is similar to the Kalman-Bucy model, which was designed for linear system with linear observations and it applies for discrete time sequence. The Kalman filter has the following general form,

$$\mathbf{X}(t_{j+1}) = \mathbf{G}(t_j)\mathbf{Y}(t_{j+1}) + \mathbf{B_2}(t_j)\boldsymbol{\varepsilon_2}(t_{j+1}), \tag{23a}$$

$$\mathbf{Y}(t_{j+1}) = \mathbf{a_0}(t_j) + \mathbf{a_1}(t_j)\mathbf{Y}(t_j) + \mathbf{b_1}(t_j)\boldsymbol{\varepsilon_1}(t_{j+1}), \tag{23b}$$

where $\mathbf{Y}$ is the state variable for the underlying dynamics and $\mathbf{X}$ is the observation. The linear function $\mathbf{G}(t_j)$ is the observational operator. The Kalman filter written in the classical form (23) does not belong to the general conditional Gaussian nonlinear framework, since the right hand side of (23a) involves the state variable $\mathbf{Y}$ at time $t_{j+1}$. Nevertheless, a slight modification can easily facilitate the Kalman filter to become a special case of the general conditional Gaussian nonlinear model. In fact, $\mathbf{Y}(t_{j+1})$ on the right hand side of (23a) can

be replaced by the equation (23b) and the resulting coupled system reads,

$$\mathbf{X}(t_{j+1}) = \mathbf{A_0}(t_j) + \mathbf{A_1}(t_j)\mathbf{Y}(t_j) + \mathbf{B_1}(t_j)\boldsymbol{\varepsilon_1}(t_{j+1}) + \mathbf{B_2}(t_j)\boldsymbol{\varepsilon_2}(t_{j+1}), \quad (24a)$$

$$\mathbf{Y}(t_{j+1}) = \mathbf{a_0}(t_j) + \mathbf{a_1}(t_j)\mathbf{Y}(t_j) + \mathbf{b_1}(t_j)\boldsymbol{\varepsilon_1}(t_{j+1}), \quad (24b)$$

where in (24a)

$$\mathbf{A_0}(t_j) = \mathbf{G}(t_j)\mathbf{a_0}(t_j), \qquad \mathbf{A_1}(t_j) = \mathbf{G}(t_j)\mathbf{a_1}(t_j) \qquad \text{and} \qquad \mathbf{B_1}(t_j) = \mathbf{G}(t_j)\mathbf{b_1}(t_j). \quad (25)$$

It is important to note that both the noise sources $\boldsymbol{\varepsilon_1}$ and $\boldsymbol{\varepsilon_2}$ enter into the observational process $\mathbf{X}$, which is different from the Kalman-Bucy model in (22). Comparing the Kalman filter (24) with the general framework of the conditional Gaussian nonlinear systems (1), it is easy to conclude that the Kalman filter is a special case of the latter with state-independent coefficients, linear model structure and additive noise.

### *3.5.3. The Rauch-Tung-Striebel (RTS) smoother*

The Rauch-Tung-Striebel (RTS) smoother [13] is an efficient two-pass algorithm for fixed interval smoothing of linear model with Gaussian noise. It is one of the most widely used smoothers in engineering, geophysics and turbulence.

The starting model for applying the RTS smoother is the same as that in (23) and the forward pass is simply the classical Kalman filter. These filtered prior and posterior state estimates $\boldsymbol{\mu}_-^{j+1}$, $\boldsymbol{\mu}^{j+1}$ and covariance $\mathbf{R}_-^{j+1}$, $\mathbf{R}^{j+1}$ are saved for use in the backwards pass. In the backwards pass, the smoothed state estimates $\boldsymbol{\mu}_\mathbf{s}^j$ and covariances $\mathbf{R}_\mathbf{s}^j$ are computed using the following recursive equations from $j = n$ back to $j = 0$,

$$\boldsymbol{\mu}_-^{j+1} = \mathbf{a}_\mathbf{0}^j + \mathbf{a}_\mathbf{1}^j\boldsymbol{\mu}^j, \quad (26a)$$

$$\mathbf{R}_-^{j+1} = \mathbf{a}_\mathbf{1}^j\mathbf{R}^j(\mathbf{a}_\mathbf{1}^j)^* + \mathbf{b}_\mathbf{1}^j(\mathbf{b}_\mathbf{1}^j)^*, \quad (26b)$$

$$\mathbf{C}^j = \mathbf{R}^j\mathbf{a}_\mathbf{1}^j(\mathbf{R}_-^{j+1})^{-1}, \quad (26c)$$

$$\boldsymbol{\mu}_\mathbf{s}^j = \boldsymbol{\mu}^j + \mathbf{C}^j(\boldsymbol{\mu}_\mathbf{s}^{j+1} - \boldsymbol{\mu}_-^{j+1}), \quad (26d)$$

$$\mathbf{R}_\mathbf{s}^j = \mathbf{R}^j + \mathbf{C}^j(\mathbf{R}_\mathbf{s}^{j+1} - \mathbf{R}_-^{j+1})(\mathbf{C}^j)^*. \quad (26e)$$

To build a connection between the RTS smoother in (26) and the general expression of the optimal smoother estimation associated with the conditional Gaussian systems in Theorem 3.6, plug (26a)–(26b) into (26c)–(26e) to eliminate the explicit dependence on the prior distribution,

$$\mathbf{C}^j = \mathbf{R}^j \mathbf{a}_1^j \big(\mathbf{a}_1^j \mathbf{R}^j (\mathbf{a}_1^j)^* + \mathbf{b}_1^j (\mathbf{b}_1^j)^*\big)^{-1}, \tag{27a}$$

$$\boldsymbol{\mu}_{\mathbf{s}}^j = \boldsymbol{\mu}^j + \mathbf{C}^j (\boldsymbol{\mu}_{\mathbf{s}}^{j+1} - \mathbf{a}_{\mathbf{0}}^j - \mathbf{a}_{\mathbf{1}}^j \boldsymbol{\mu}^j), \tag{27b}$$

$$\mathbf{R}_{\mathbf{s}}^j = \mathbf{R}^j + \mathbf{C}^j (\mathbf{R}_{\mathbf{s}}^{j+1} - \mathbf{a}_{\mathbf{1}}^j \mathbf{R}^j (\mathbf{a}_{\mathbf{1}}^j)^* - \mathbf{b}_{\mathbf{1}}^j (\mathbf{b}_{\mathbf{1}}^j)^*)(\mathbf{C}^j)^*. \tag{27c}$$

This is consistent with the general conclusions in Theorem 3.6 when the linear model (23) with $\mathbf{b_2} = 0$ is utilized. However, the filter estimate $(\boldsymbol{\mu}^j, \mathbf{R}^j)$ in the RTS smoother (27) is through the linear and Gaussian system while those in the general conditional Gaussian nonlinear systems involves nonlinearity and requires solving random Riccati equation for the filter covariance.

## 4. Applications

### *4.1. Recovering the path-wise and statistical information of hidden variables*

#### *4.1.1. A perfect model test*

We start with a perfect model test. The model here is a physics-constrained nonlinear dyad model [20, 21] with one observed variable $u$ and one unobserved variable $v$. The model reads,

$$\begin{aligned} \mathrm{d}u &= \big((-d_u + cv)u + F_u\big)\,\mathrm{d}t + \sigma_u\,\mathrm{d}W_u, \\ \mathrm{d}v &= (-d_v v - cu^2)\,\mathrm{d}t + \sigma_v\,\mathrm{d}W_v, \end{aligned} \tag{28}$$

In this dyad model, the energy in the nonlinear terms is conserved. It therefore satisfies the physics constraint. Note that the observed variable $v$ here serves as a stochastic damping in the dynamics of $u$. Once the variable $v$ goes beyond the threshold value $v^* = d_u/c$, the intermittency and extreme events appear in $u$. Such an intermittent behavior provides rich non-Gaussian features of $u$. The following parameters are used for the intermittent regime here.

$$F_u = 1, \qquad d_v = 0.8, \qquad d_u = 0.8, \qquad \sigma_v = 2, \qquad c = 1.2, \qquad \sigma_u = 0.5. \tag{29}$$

The blue curves in Panels (a)–(b) of Figure 1 show one realization of $u$ and $v$ respectively, and those in Panels (d)–(e) illustrate the corresponding equilibrium PDFs. It is clear that once $v$ exceeds the threshold value $d_u/c = 0.67$ , the associated $u$ becomes intermittently unstable with a fat-tailed distribution and the appearance of extreme events.

In many applications, the filter mean state (i.e., the posterior mean state from filtering) is simply treated as the estimated state. However, as is indicated by the red curves in Panels (e)–(f), both the PDF and the ACF of the time series of the filter mean are significantly biased from the truth. According to Panels (b)–(c), it is clear that despite the success in capturing the positive phases of $v$ to a large extent, most of the events with negative phases are missed. In fact, the positive phases of $v$ correspond to the intermittent events of $u$, at which the signal-to-noise ratio is large and therefore the state estimation is accurate. On the other hand, the negative phases of $v$ are associated with the quiescent events of $u$ and the resulting estate estimation has a large uncertainty, which implies a significant gap between the filter mean and the truth. As a comparison, the brown curves in Panel (b) shows the mean estimate from the nonlinear smoother. Since the smoother makes use of the entire observational period, its estimation is more accurate than the filter estimate and the uncertainty (Panel (c)) is smaller as well. As a consequence, the PDF and ACF associated with the smoother mean time series are overall closer to the truth. However, there is still an obvious disparity between the PDFs from these conditional mean estimates and the truth due to the non-negligible uncertainty in the smoother estimate.

The green curves of the PDF and ACF in Panels (e)–(f) are based on the time series resulting from the backward sampling strategy. The reason that the sampled trajectories are able to capture both the statistical and temporal information is that they make use of the mean state of the estimation and the uncertainty as well as the temporal dependence at different time instants. Panel (g) involves four different sampled trajectories. One of the major difference between these sampled trajectories and the filter mean estimates is that the former are able to capture the negative phases of $v$ and therefore they are more

skillful in recovering the associated PDF and ACF.

*4.1.2. A nonlinear model test in the presence of model error*

In most real applications, perfect model is never known. Approximate (or imperfect) models are used for state estimation and prediction. In particular, the unobserved processes often represent unresolved or small scale variables, the complete dynamics of which are hard to obtain. Simplified models are typically used to describe the unobserved variables. Therefore, it is important to understand the skill of recovering the dynamical and statistical features in the presence of model error. In this subsection, the following two-dimensional highly nonlinear model is adopted as the perfect model that generates the true signal,

$$\mathrm{d}u = \Big( -\gamma u + F_u \Big)\,\mathrm{d}t + \sigma_u\,\mathrm{d}W_u, \tag{30a}$$

$$\mathrm{d}\gamma = (a_\gamma \gamma + b_\gamma \gamma^2 + c_\gamma \gamma^3 + f_\gamma)\,\mathrm{d}t + (A_\gamma + B_\gamma \gamma)\,\mathrm{d}W_{\gamma,1} + \sigma_\gamma\,\mathrm{d}W_{\gamma,2}. \tag{30b}$$

In this model, $u$ and $\gamma$ are the observed and unobserved variables, respectively. The variable $\gamma$ acts as a stochastic damping in the equation of $u$ and the averaged value of $\gamma$ over time needs to be positive to guarantee the mean stability of $u$ [51]. Once the sign of $\gamma$ switches from positive values to negative values, $\gamma$ becomes anti-damping and it leads to the intermittent events in $u$. On the other hand, $\gamma$ is driven by a cubic nonlinear equation with correlated additive and multiplicative noise. This cubic model is a canonical model for low frequency atmospheric variability [52, 53]. This one-dimensional, normal form has been applied in a regression strategy for data from a prototype atmosphere and ocean model to build one-dimensional stochastic models for low-frequency patterns such as the North Atlantic Oscillation and the leading principal component that has features of the Arctic Oscillation. Given the non-Gaussian features and the potential physical explanations, the low-order model (30) becomes a useful testbed for developing suitable stochastic parameterization strategies of the hidden process that allows skillful prediction of the extreme events in the observed variable.

The following parameters are adopted in the coupled system (30),

$$F_u = 0.3, \qquad \sigma_u = 0.1, \qquad a_\gamma = -\frac{3}{8}, \qquad b_\gamma = 1, \qquad c_\gamma = -\frac{1}{2}, \\ A_\gamma = 0, \qquad B_\gamma = \frac{1}{2\sqrt{2}}, \qquad f_\gamma = 0.1, \qquad \sigma_\gamma = \frac{1}{2\sqrt{2}}. \tag{31}$$

One realization of the true signal of $u$ and $\gamma$ is shown in blue curves in Panels (a) and (b) of Figure 2, respectively. The associated PDFs are illustrated in Panels (d) and (e). It is clear that $u$ is highly intermittent with a strong one-sided fat tail in the PDF while $\gamma$ has a bimodal distribution with roughly two distinguished states. The dynamical switching between the two states of $\gamma$ corresponds to the interchange between the quiescent and active phases of $u$.

The approximate model here is developed using the stochastic parameterized equation technique [54, 55], the idea of which has been applied to the extended Kalman filters (known as the SPEKF-type model) and other prediction and data assimilation forecast models. The approximate model has the following form,

$$\mathrm{d}u = (-\gamma u + F_u)\,\mathrm{d}t + \sigma_u\,\mathrm{d}W_u, \tag{32a}$$

$$\mathrm{d}\gamma = -d_\gamma(\gamma - \hat{\gamma})\,\mathrm{d}t + \sigma_\gamma\,\mathrm{d}W_\gamma. \tag{32b}$$

In (32), the nonlinear process $\gamma$ with correlated additive and multiplicative noise in (30b) has been simplified to a linear process with only Gaussian additive noise. Nevertheless, the variable $\gamma$ remains switching between positive and negative phases, representing damping and anti-damping effects as a feedback to $u$. Therefore, the variable $\gamma$ is still able to trigger intermittent extreme events in $u$. Note that the approximate model (32) belongs to the conditional Gaussian framework while the perfect model (30) does not. The three parameters $\sigma_\gamma$, $d_\gamma$ and $\hat{\gamma}$ in (32) can be calibrated using a general model calibration method developed in [56]. But for the simplicity here, these parameters are calibrated by matching the mean, variance and decorrelation time of $\gamma$ in the perfect system, which provides the optimal Gaussian fit of $\gamma$ in the nonlinear model in (30).

The filter and smoother mean estimates are shown in Panel (b) of Figure 2 and the associated uncertainties are illustrated in Panel (c). The negative

phases of $\gamma$, corresponding to the intermittent phases of $u$, are recovered with high accuracy and small uncertainty. However, both the filter and smoother mean estimates fail to capture the positive phases of $\gamma$ due to a relatively small signal-to-noise ratio of the signal in the corresponding phases of the observed variable $u$. As a result, the PDFs formed by the time series from the filter and the smoother mean estimates contain a large error in capturing the right side of the true PDF, although they are able to recover the left peak of the truth. On the other hand, the $\gamma$ process in the imperfect approximate model (32) is Gaussian. Therefore, a free run of the approximate model leads to a Gaussian PDF of the unobserved variable. Despite capturing the Gaussian tail on the right side of the true PDF, such a Gaussian PDF completely misses the non-Gaussian features embodied by the second peak on the left side of the truth, which comes from the negative events of $\gamma$ that are associated with the intermittency in $u$.

The PDF associated with the trajectory from the backward sampling strategy combines the advantages of both the smoother mean estimate and the free run of the approximate model. The green curve in Panel (e) shows the PDF associated with the sampled trajectory. It is important to note that the PDF perfectly captures the peak of the left side of the truth and the Gaussian tail on the right is recovered more accurately than that from the filter and smoother mean estimates. In Panel (g), it is shown that the sampled trajectories succeed in capturing the negative phases of $\gamma$, which is obviously not the case for a free run of the model. The sampled trajectories also has a larger chance in capturing the positive phase of $\gamma$ compared with the conditional mean estimates from both the filter and the smoother. Finally, the ACF of $\gamma$ from the sampled trajectories also perfectly match that of the truth.

### *4.2. Nonlinear filtering and smoothing of physics-constrained nonlinear systems for detecting non-Gaussian features and predicting hidden extreme events*

Since both filtering and smoothing are designed for state estimation, it is important to study their difference in the resulting estimated states and explore

suitable situations for the applications.

*4.2.1. The nonlinear dyad model*

Here, the test model is the physics-constrained dyad model (28) with parameters given by (29). The reasons to apply this test model are the following. First, this model has energy-conserving nonlinear interactions between the observed and unobserved variables, satisfying the physics constraint, which mimics the dynamical behavior of many more realistic systems. Second, the unobserved variable $v$ serves as the triggering effect of the intermittent events of the observed variable $u$ while the decaying phase of the intermittent events in $u$ leads the relaxation of the unobserved variable $v$ back to its mean state. The energy is transferred nonlinearly through these non-Gaussian events. Third, if the observed variable $u$ and the unobserved variable $v$ are regarded as the large and small scale variables in turbulence, then the highly non-Gaussian PDF of $u$ and the nearly Gaussian statistics in $v$ are the typical feature as in many realistic systems.

Figure 3 shows a comparison between the state estimation using the nonlinear filtering and the nonlinear smoothing techniques. The true signals are given by the blue curves. The smoother and filter mean states are shown in the black dashed curves in Panels (b) and (c), respectively, and the associated uncertainties (represented by one standard deviation) are given by the red and green shading areas. One major difference in the recovered states is that the smoother estimate is able to capture both the timing and the duration of $v$ when it goes above the intermittent threshold $v^* = d_u/c = 0.67$ while the filter estimate always fails to detect the onset of such triggering phases of the intermittent events, e.g., at $t = 6$ and $t = 16.5$.

The fundamental reason of the failure of the filter in capturing the onset phases of the extreme events is that the filter estimate is calculated based on the observational data only in the past. In this dyad model, the intermittent events in $u$ are the response of the anti-damping of $v$. This response is always lagged behind the occurrence of the positive values of $v$. Therefore, before a

significant increase of the amplitude of $u$ that allows the filter to perceive such a triggering phase, the true signal of $v$ has already stayed in the positive phases for a certain period. This leads to the failure of the filter for timely predicting the triggering phases of the extreme events. In contrast, this is not the case for the nonlinear optimal smoother, since it is calculated based on the entire observational period, which is able to foresee the upcoming intermittent events and facilities an unbiased estimation of the timing of the onset phases. On the other hand, both the filter and smoother estimates are able to capture the period of $v$ that goes from anti-damping to damping phases, corresponding to the demise phases of $u$. This is because at the demise phases of the extreme events the signal of $v$ is mainly driven by $u$ though the feedback term $-cu^2$ and such a feedback is immediate.

*4.2.2. A four-dimensional stochastic climate model with multiscale features*

Now we consider a four-dimensional stochastic climate model with multiscale features. The model reads,

$$\begin{aligned}
\mathrm{d}x_1 &= \Big( - x_2(L_{12} + a_1 x_1 + a_2 x_2) + d_1 x_1 + F_1 \\
&\qquad + L_{13} y_1 + b_{123} x_2 y_1 \Big)\, \mathrm{d}t + \sigma_{x_1}\, \mathrm{d}W_{x_1}, \qquad (33a) \\
\mathrm{d}x_2 &= \Big( + x_1(L_{12} + a_1 x_1 + a_2 x_2) + d_2 x_2 + F_2 \\
&\qquad + L_{24} y_2 + b_{213} x_1 y_1 \Big)\, \mathrm{d}t + \sigma_{x_2}\, \mathrm{d}W_{x_2}, \qquad (33b) \\
\mathrm{d}y_1 &= \Big( - L_{13} x_1 + b_{312} x_1 x_2 + F_3 - \frac{\gamma_1}{\epsilon} y_1 \Big)\, \mathrm{d}t + \frac{\sigma_{y_1}}{\sqrt{\epsilon}}\, \mathrm{d}W_{y_1}. \qquad (33c) \\
\mathrm{d}y_2 &= \Big( - L_{24} x_2 + F_4 - \frac{\gamma_2}{\epsilon} y_2 \Big)\, \mathrm{d}t + \frac{\sigma_{y_2}}{\sqrt{\epsilon}}\, \mathrm{d}W_{y_2}, \qquad (33d)
\end{aligned}$$

where $b_{123} + b_{213} + b_{312} = 0$. This simple stochastic climate model [57, 58] features many of the important dynamical properties of comprehensive global circulation models (GCMs) but with many fewer degree of freedom. It contains a quadratic nonlinear part that conserves energy as well as a linear operator. The linear operator includes a skew-symmetric part that mimics the Coriolis effect and topographic Rossby wave propagation, and a negative definite symmetric part that is formally similar to the dissipation such as the surface drag

and radiative damping. The two variables $x_1$ and $x_2$ can be regarded as climate variables while the other two variables $y_1$ and $y_2$ become weather variables that occur in a much faster time scale when $\epsilon$ is small. The coupling in different variables is through both linear and nonlinear terms, where the nonlinear coupling through $b_{ijk}$ produces multiplicative noise. Note that when $\epsilon \to 0$, applying an explicit stochastic mode reduction results in a two-dimensional system for the climate variables [59, 60, 61].

Assume one realization of the two climate variables $x_1$ and $x_2$ is given while the two weather variables $y_1$ and $y_2$ have no direct observations. The following parameters are used in the tests here,

$$\begin{gathered} L_{12} = 1, \qquad L_{13} = 0.5, \qquad L_{24} = 0.5, \qquad a_1 = 2, \qquad a_2 = 1, \\ d_1 = -1, \qquad d_2 = -0.4, \qquad \sigma_1 = 0.5, \qquad \sigma_2 = 2, \qquad \sigma_3 = 0.5, \qquad \sigma_4 = 1, \\ b_{123} = 1.5, \qquad b_{213} = 1.5, \qquad \gamma_1 = 0.5, \qquad \gamma_2 = 0.5, \\ F_1 = F_2 = F_3 = F_4 = 0. \end{gathered} \tag{34}$$

Depending on the scale separation parameter $\epsilon$, two dynamical regimes are considered:

$$\begin{aligned} \text{Regime I}: &\qquad \epsilon = 1, \\ \text{Regime II}: &\qquad \epsilon = 0.1, \end{aligned} \tag{35}$$

In the $\epsilon = 1$ regime, the weather and climate variables lie roughly on the same time scale. The blue curves in Figure 4 illustrate one realization of different model variables as well as the associated PDFs and ACFs. Here, one observed variable $x_1$ and one unobserved variable $y_1$ are significant non-Gaussian, where the associated PDFs are highly skewed and have an one-sided fat tail. Extreme events appear in both the trajectories of $x_1$ and $y_1$. The other two variables $x_2$ and $y_2$ are nearly Gaussian. Note that the ACF of $y_2$ releases more slowly than the other variables since it has the least influence from the nonlinearity.

On the other hand, due to the stronger linear damping and the larger noise strength, the two unobserved variables $y_1$ and $y_2$ become nearly Gaussian in the $\epsilon = 0.1$ regime. See the blue curves in Figure 5. Nevertheless, one of the

observed variables $x_1$ is still highly non-Gaussian. The ACFs of all the four variables release much faster than those in the $\epsilon = 1$ regime (note the difference in the a-axis for showing the ACFs in the two regimes).

Next, we compare the filter estimates, the smoother estimates and the statistics from the trajectories via the backward sampling strategy. First, as in many applications, the time series of the filter mean states can be regarded as an approximate estimates of the unobserved variables. In fact, the associated PDF of $y_1$ from the filter mean estimates is quite similar to the truth in the $\epsilon = 1$ regime and the filter mean succeeds in predicting the hidden extreme events in the path-wise sense as well. However, such an approximation leads to a large error for $y_2$ in recovering both the trajectory and the PDF. The failure of using the filter mean estimate to approximate $y_2$ is that the uncertainty in filtering $y_2$ is quite large (see Panel (a) of Figure 6) while that in filtering $y_1$ is tiny. Such a difference is due to the fact that $y_1$ has a strong interaction with the two observed variables via the energy-conserving nonlinear terms, which is not the case of $y_2$. Note that the time series of the filter mean estimate also fails to capture the temporal information of $y_2$ in the original dynamics in that the associated ACF is biased from the truth. With the decrease of $\epsilon$, the two unobserved variables $y_1$ and $y_2$ are more turbulent and therefore simply adopting the filter mean estimates provides much less information for recovering the statistics of the original model. In fact, in the $\epsilon = 0.1$ regime, the PDF of $y_1$ associated with the filter mean estimates already contains a large error. In addition, the filter mean estimation of $y_2$ indicates almost no information beyond the mean value of the statistical equilibrium state (see the last row of Figure 5 and Panel (c) of Figure 6).

Figure 6 illustrates the filter and smoother estimates. They are actually quite similar to each other in both the dynamical regimes. This indicates that the filter mean and the smoother mean estimates are both insufficient to recover the statistics of the original system when the system is strongly turbulent due to the ignorance of the large uncertainty. On the other hand, the sampled trajectory from the backward sampling strategy, which takes into account the information

in both the smoother mean and smoother uncertainty as well as the temporal dependence between different time instants, is able to perfectly capture both the statistical and dynamical information. In particular, the PDFs and ACFs associated with the sampled trajectories as shown in Figure 4 and 5 are identical to the truth. The sampled trajectories also succeed in predicting the timing and duration of the hidden extreme events of $y_1$ in the $\epsilon = 1$ regime.

### *4.3. Recovering non-Gaussian statistics of the observed variables using only short training period*

We have so far focused on the state estimation of the unobserved variables. Another key issue in practice is to predict the non-Gaussian statistics of the observed or resolved variables. It is important to note that in many applications in climate, atmosphere and ocean science the training data of variables lying in interannual or longer time scale is very limited since the satellite has only been used in the recent a few decades. Therefore, simply using the available observations may not be sufficient to describe many key non-Gaussian features in an accurate way, especially for recovering fat tails and extreme events which typically require a large number of samples.

The backward sampling strategy developed here can be used to generate a sufficient number of trajectories of the unobserved variables that are associated with the observations. Then plugging these sampled trajectories into the process of the observed variables facilitates the recovery of the non-Gaussian statistics of the observed variables.

#### *4.3.1. A perfect model test*

Let us start with a perfect model test, where the model that generates the true signal and the one for sampling the unobserved trajectories are both the dyad model (28) with parameters listed in (29). As was shown in Figure 1, the PDF of $u$ is highly non-Gaussian with an one-sided fat tail.

Assume a short observational period of $u$ with only 50 units is available, as shown in Panel (a) of Figure 7. This period contains three strong extreme

events ($t$ = 6,17 and 46) and several moderate strong events. It mimics the observed El Niño-Southern Oscillation (ENSO) [62]. In fact, since late 1970s when satellites became available for collecting the ENSO data three super El Niño events and a few moderate events were observed.

The blue curves in Panel (d) and Panel (e) show the PDF of this observed time series in linear and logarithm scales, respectively. As a comparison, the true PDF by running the model forward for 1000 units is shown in black color. It is clear that the fat tail of the PDF based only on the observed period is estimated with large errors, which is due to the insufficient number of samples.

Since filtering is widely used for state estimation and the filtered mean estimate is often regarded as the best estimation of the unobserved state, one natural and simple way of generating more samples is as follows. Plug the time series of the filter mean estimate into the process of $u$ to replace the $v$ variable there and repeat running the model of $u$ for $L$ times, which provides $L$ time series of $u$ with 50 units of each. These $L$ time series are different from each other because of the noise in the $u$ process. Then the PDF of $u$ is formed by collecting all these time series, which effectively gives a length of in total $50L$ units. However, such a method fails to recover the fat tail and extreme events. See the PDFs in red curves in Panels (d) and (e), where $L = 20$ is used here. There are at least two fundamental reasons that lead to the failure of such an approach. First, in addition to the filter mean state, the uncertainty in the filter estimate also plays an important role in the nonlinear interaction between the observed and unobserved variables, especially for the intermittent phases. Second, the time series of the filter mean state does not capture the exact dynamical information of the truth. For example, the onset of the intermittent phases is always delayed in this dyad model (See Section 4.2), which means the duration time of the intermittent phase is always underestimated and thus the method is unskillful in recovering the fat tail.

The backward sampling strategy developed here can actually resolve both the fundamental difficulties discussed above and therefore provides an unbiased way of recovering the non-Gaussian statistics of the observed variable. In fact,

the backward sampling is based on the smoothing estimates, which are able to capture both the timing and duration of the intermittent phases, as was shown in Section 4.2. In addition, the sampled trajectories of $v$ take into account the information in the smoother mean and smoother uncertainty as well as the dynamical information of the truth. Run the backward sampling $L = 20$ times and collect all the resulting data to form the PDF of $u$, which is shown by the green curves in Panels (d) and (e). It is clear that the resulting PDF perfectly captures the non-Gaussian fat tail of the truth. In addition, as is shown in Panel (f), the trajectories of $u$ based on the sampled $v$ from the backward sampling (green) are intermittent while those based on the filter mean time series of $v$ (red) are more quiescent.

One final remark here is that if the perfect model is known, then a more straightforward way of forming the PDF of $u$ can be done by running the perfect model forward. There is in fact no need to run the backward sampling for obtaining the sampled trajectories of $v$. Nevertheless, model error appears in many applications. Then the proposed method here becomes more powerful in recovering the non-Gaussian PDF than simply running the imperfect model forward. See the next subsection for a more realistic case with model error.

*4.3.2. A nonlinear model test in the presence of model error*

Consider a more realistic situation now. The perfect model that generates the true signal is still given by (28). However, it is assumed to be unknown. Therefore, an approximate model is used to recover the statistics and the approximate model contains model error. Here, the imperfect approximate model is given as follows

$$\begin{aligned} \mathrm{d}u &= \big((-d_u + cv)u + F_u\big)\,\mathrm{d}t + \sigma_u\,\mathrm{d}W_u, \\ \mathrm{d}v &= (-d_v v - cu^2 + F_v)\,\mathrm{d}t + \sigma_v\,\mathrm{d}W_v, \end{aligned} \tag{36}$$

where an extra term $F_v$ is added to the perfect model. Assume $F_v = -2$ while all the other parameters are taken as the same values as those in (29). Due to this negative forcing $F_v$, the trajectory of $u$ from a free run of the imperfect model

(36) is less intermittent and the associated PDF is more towards Gaussian. See the magenta curves in Panels (d)–(f) in Figure 8 for the PDF and the trajectories from a free run of the imperfect model.

On the other hand, with the help of the observations, the smoother estimates based on such an imperfect model actually do not differ too much from those using the perfect model. See Panels (b) and (c) in Figure 8. In fact, there is only a slight shift towards the negative value of $v$ in the smoother mean estimate. This error is much smaller than the model bias introduced by the extra forcing $F_v$. As a result, even in the presence of a significant model error in the imperfect model, the recovered PDF by plugging the sampled trajectories of $v$ into the $u$ process (green curves in Panels (b) and (c)) remains similar to the truth (black curves), especially the non-Gaussian fat tail.

### *4.4. Improving the stochastic parameterizations*

Stochastic parameterizations are widely used in practice to model the unresolved or unobserved variables, which aim at capturing the nonlinear interactions across different scales and recovering the statistical feedback from the unresolved to the resolved variables. However, due to the lack of observations, it is often not an easy task to design a skillful stochastic parameterization.

The backward sampling technique developed here can be used as a systematical framework for quantifying the bias in the given stochastic parameterization and providing guidelines for improving it.

Consider the following model as the perfect model that generates the true signal,

$$\mathrm{d}u = (-d_u u + \gamma v)\,\mathrm{d}t + \sigma_u\,\mathrm{d}W_u, \tag{37a}$$

$$\mathrm{d}v = -d_v(v - \hat{v})\,\mathrm{d}t + \sigma_v(v)\,\mathrm{d}W_v, \tag{37b}$$

where $u$ is the observed variable while $v$ is unobserved. The noise coefficient in the process $v$ here is not a constant. Instead, it is state dependent

$$\sigma_v(v) = \exp\left(-\frac{|v| - m_c}{v_c}\right). \tag{38}$$

The parameters in this model are given by

$$d_u = 0.5, \qquad d_v = 0.5, \qquad \hat{v} = 0, \qquad \gamma = 1, \qquad \sigma_u = 0.2, \qquad m_c = 1, \qquad v_c = 1. \tag{39}$$

One realization of the true signals of $u$ and $v$ is shown in Panels (a)–(b) of Figure 9. The blue curves in Panels (e)–(h) show the PDFs and the ACFs of the model. In particular, the PDFs of $u$ and $v$ are both highly non-Gaussian with bimodal distributions.

Since the true dynamics of $v$ in the perfect model (37) is unknown, a typical way of parameterizing $v$ is to adopt a linear and Gaussian process. This leads to the following approximate model,

$$\mathrm{d}u = (-d_u u + \gamma v)\,\mathrm{d}t + \sigma_u\,\mathrm{d}W_u, \tag{40a}$$

$$\mathrm{d}v = -d_v^M (v - \hat{v}^M)\,\mathrm{d}t + \sigma_v^M\,\mathrm{d}W_v, \tag{40b}$$

where $\sigma_v^M$ is a constant. Assume the mean and variance of $v$ in the perfect model are available. But we assume the decorrelation time of $v$ is not accurately measured, where the measured value is twice as large as the truth. The three parameters in (40b) are then estimated by matching the mean, variance and the decorrelation time from the measurement.

Because of the additive noise coefficient $\sigma_v^M$ in the approximate model (40), the model becomes linear and Gaussian. Therefore, the equilibrium PDFs of both $u$ and $v$, as shown in the magenta curves in Panels (e) and (f), are Gaussian. In addition, due to the overestimation of the decorrelation time, the ACFs, as shown in Panels (g) and (h), are also different from the truth. These results indicate that the stochastic parameterization in (40) is not a suitable one. Next, the backward sampling technique is used to improve the stochastic parameterization.

Despite that the approximate model (40) with the simplest stochastic parameterization contains a large error, it is nevertheless a useful starting model for improving the stochastic parameterization. In fact, combining the observations of $u$ with this linear and Gaussian approximate model, the backward sampling

strategy can be applied to obtain trajectories of the unobserved variable $v$. Notably, the sampled trajectories are no longer Gaussian since the observations play an important role in such a sampling procedure. Then the decorrelation time computed from the ACF, the mean state and the non-Gaussian PDF associated with the the sampled trajectories can be used to build an improved model of $v$ based on the following result.

**Theorem 4.1.** *Given the decorrelation time $\tau$, the mean value $m$ and non-Gaussian PDF $p(x)$, there is an unique stochastic differential equation that satisfies these conditions,*

$$\mathrm{d}x(t) = -\lambda(x(t) - m)\,\mathrm{d}t + \sigma(x)\,\mathrm{d}W(t), \tag{41}$$

*where $\lambda = 1/\tau$ and the multiplicative noise coefficient $\sigma(x)$ is given by*

$$\sigma^2(x) = \frac{2}{p(x)}\{-\lambda\Phi(x)\}, \qquad \textit{with} \qquad \Phi(x) = \int_b^x (y-m)p(y)dy.$$

Following Theorem 4.1, the improved model is as follows,

$$du = (-d_u u + \gamma v)\,\mathrm{d}t + \sigma_u\,\mathrm{d}W_u, \tag{42a}$$

$$dv = -\lambda^M(v - m^M)\,\mathrm{d}t + \sigma_v^M(v)\,\mathrm{d}W_v, \tag{42b}$$

which has essentially the same form as the perfect model but the parameters in (42b) are determined by the results in Theorem 4.1 using the statistics from the trajectories sampled by applying the backward sampling strategy to the starting linear Gaussian model (40).

In Panels (c) and (d), one realization of the trajectories from (42) is shown. For an unbiased comparison with the perfect model simulation, we adopt the same random number seeds in generating these trajectories. It is clear from these panels that the dynamics of the improved model is far from linear and Gaussian as in the starting imperfect model (40). In fact, as shown in Panels (e) and (f), the PDFs of both $u$ and $v$ from the improved model (42) are bimodal, and they are quite similar to the truth. Another important finding is that despite the overestimation of the decorrelation time in the starting imperfect model (40),

the observations play a significant role in the backward sampling procedure such that the decorrelation time in the sampled trajectory is nearly the same as the truth. As a consequence, the ACFs associated with the improved model almost overlap with the truth. See Panels (g) and (h). Finally, the multiplicative noise coefficient $\sigma_v^M(v)$ in (42b) is shown in the green curve in Panel (i). Calibrating the improved model (42) based on the sampled trajectories leads to an apparent multiplicative noise in the process of $v$, which is completely different from the additive noise in the starting imperfect model (40).

To summarize, despite the large model error in the starting imperfect model (40), the sampled trajectories by combining observations with this model are nevertheless able to provide extra useful information. The resulting improved model (42) based on the above sampled trajectories is therefore much more accurate in reproducing the non-Gaussian statistics. The improved model (42) is also a suitable approximate model for real-time forecast, which is expected to be more skillful in predicting extreme events and other non-Gaussian features than the starting imperfect model (40). Note that the difference in the statistics between the sampled hidden trajectory and that from a free run of the approximate model provides an effective way of quantifying the model error without knowing the perfect model. This is an extremely useful tool since the perfect model is unknown in practice.

## 5. Conclusion

In this article, a nonlinear optimal smoother and an associated nonlinear optimal sampling technique of the hidden trajectories are developed for a rich class of nonlinear complex turbulent dynamical systems with partial observations. The models considered here are the so-called conditional Gaussian nonlinear systems, which are highly nonlinear and highly non-Gaussian despite the conditional Gaussian structures. These models have wide applications in geophysics, engineer, neural science and other areas. Both the optimal smoother and the optimal sampling strategy have closed analytic form and therefore they

can be applied to high-dimensional nonlinear turbulent systems with high efficiency. Several important applications of the nonlinear optimal filter, smoother and sampling strategy are addressed in this article. They include recovering the nonlinear dynamics and non-Gaussian statistics of complex nonlinear systems, state estimation of extreme events, effective sampling and predicting the fat-tailed PDFs with very short observational training data and improving the stochastic parameterizations using models with multiplicative noise. Both the perfect model setup and the tests in the presence of model error are studied here.

One important future work is to study the online forecast skill of the improved approximate model developed by making use of the information in the sampled trajectories sampled by some starting imperfect models, especially for predicting the rare and extreme events. Recovering the statistical and dynamical information from the backward sampling can also be used for model selection and model identification. In addition, the proposed nonlinear framework can be applied to study the stochastic control for nonlinear turbulent systems.

**Acknowledgments**

The research of N.C. is partially supported by the Office of Vice Chancellor for Research and Graduate Education (VCRGE) at University of Wisconsin-Madison. The research of A.J.M. is partially supported by the Office of Naval Research (ONR) N00014-19-1-2680 and the Center for Prototype Climate Modeling (CPCM) at New York University Abu Dhabi Research Institute. The research of both N.C. and A.J.M is funded by the ONR MURI N00014-19-1-2421.

## Appendix A. Some useful properties of multivariate Gaussian distributions

Let us denote a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{R}$ by $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{R})$, where $\mathbf{x}$ is the random variable.

**Lemma Appendix A.1.** *For two Gaussian distributions,*

$$\int \mathcal{N}(\mathbf{x_2}|\mathbf{F}\mathbf{x_1}+\mathbf{b},\mathbf{R_2})\mathcal{N}(\mathbf{x_1}|\boldsymbol{\mu_1},\mathbf{R_1})\,\mathrm{d}\mathbf{x_1} = \mathcal{N}(\mathbf{x_2}|\mathbf{F}\boldsymbol{\mu_1}+\mathbf{b},\mathbf{F}\mathbf{R_1}\mathbf{F}^*+\mathbf{R_2}). \tag{A.1}$$

**Lemma Appendix A.2.** *Let the Gaussian random variables be*

$$\mathbf{x} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \end{pmatrix},$$

*with mean* $\boldsymbol{\mu}$ *and covariance* $\mathbf{R}$*,*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} \mathbf{R_{11}} & \mathbf{R_{12}} \\ \mathbf{R_{21}} & \mathbf{R_{22}} \end{pmatrix}.$$

*The conditional distribution*

$$p(\mathbf{x_1}|\mathbf{x_2}) \sim \mathcal{N}(\overline{\boldsymbol{\mu}},\overline{\mathbf{R}}),$$

*where*

$$\begin{aligned} \overline{\boldsymbol{\mu}} &= \boldsymbol{\mu_1} + \mathbf{R_{12}}\mathbf{R_{22}^{-1}}(\mathbf{x_2}-\boldsymbol{\mu}_2), \\ \overline{\mathbf{R}} &= \mathbf{R_{11}} - \mathbf{R_{12}}\mathbf{R_{22}^{-1}}\mathbf{R_{21}}. \end{aligned} \tag{A.2}$$

## Appendix B. Proof of the theorems related to the continuous time conditional Gaussian systems

The proofs will be based on applying a Euler-Maruyama temporal discretization for the coupled system (1) with a small $\Delta t$. Eventually the limit $\Delta t \to 0$ will be taken to recover the continuous time dynamics. The Euler-Maruyama temporal discretization of (1) yields,

$$\begin{aligned} \widetilde{\mathbf{X}}^{j+1} &= \widetilde{\mathbf{X}}^j + \Big(\widetilde{\mathbf{A}}_{\mathbf{0}}^j + \widetilde{\mathbf{A}}_{\mathbf{1}}^j\widetilde{\mathbf{Y}}^j\Big)\Delta t + \widetilde{\mathbf{B}}_{\mathbf{1}}^j\Delta\widetilde{\mathbf{W}}_{\mathbf{1}}^j + \widetilde{\mathbf{B}}_{\mathbf{2}}^j\Delta\widetilde{\mathbf{W}}_{\mathbf{2}}^j, \\ \widetilde{\mathbf{Y}}^{j+1} &= \widetilde{\mathbf{Y}}^j + \Big(\widetilde{\mathbf{a}}_{\mathbf{0}}^j + \widetilde{\mathbf{a}}_{\mathbf{1}}^j\widetilde{\mathbf{Y}}^j\Big)\Delta t + \widetilde{\mathbf{b}}_{\mathbf{1}}^j\Delta\widetilde{\mathbf{W}}_{\mathbf{1}}^j + \widetilde{\mathbf{b}}_{\mathbf{2}}^j\Delta\widetilde{\mathbf{W}}_{\mathbf{2}}^j, \end{aligned} \tag{B.1}$$

where $\widetilde{\mathbf{X}} = \mathbf{X}(t_{j+1})$ and $\widetilde{\mathbf{Y}} = \mathbf{Y}(t_{j+1})$. In the proofs below, the variables or functions with tilde $\widetilde{\cdot}$ always represent the time discrete form of the continuous equation.

*Appendix B.1. Proof of Lemma 3.1*

*Proof.* Let us start with the joint distribution $p(\widetilde{\mathbf{Y}}^j, \widetilde{\mathbf{Y}}^{j+1}|\widetilde{\mathbf{X}}^s, s \le j)$. Making use of the roles of the conditional distribution yields

$$p(\widetilde{\mathbf{Y}}^j, \widetilde{\mathbf{Y}}^{j+1}|\widetilde{\mathbf{X}}^s, s \le j) = p(\widetilde{\mathbf{Y}}^{j+1}|\widetilde{\mathbf{Y}}^j, \widetilde{\mathbf{X}}^s, s \le j)\, p(\widetilde{\mathbf{Y}}^j|\widetilde{\mathbf{X}}^s, s \le j) \tag{B.2}$$

In light of the second equation in (B.1), the first term on the right hand side of (B.2) is given by

$$\begin{aligned} &p(\widetilde{\mathbf{Y}}^{j+1}|\widetilde{\mathbf{Y}}^j, \widetilde{\mathbf{X}}^s, s \le j) \\ &\qquad \sim \mathcal{N}\Big(\widetilde{\mathbf{a}}_{\mathbf{0}}^j \Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\boldsymbol{\mu}},\, \widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j \Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\mathbf{R}}^j (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)^*\Big). \end{aligned} \tag{B.3}$$

On the other hand, the second term on the right hand side of (B.2) is simply given by the filtering formula,

$$p(\widetilde{\mathbf{Y}}^j|\widetilde{\mathbf{X}}^s, s \le j) \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}^j, \widetilde{\mathbf{R}}^j). \tag{B.4}$$

The cross covariance term is given by

$$\langle \widetilde{\mathbf{Y}}'^{,j+1} (\widetilde{\mathbf{Y}}'^{,j})^* \rangle = (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\mathbf{R}}^j, \tag{B.5}$$

where $\widetilde{\mathbf{Y}}'^{,j+1}$ and $\widetilde{\mathbf{Y}}'^{,j}$ are $\widetilde{\mathbf{Y}}^{j+1}$ and $\widetilde{\mathbf{Y}}^j$ by removing their mean values. Therefore, collecting (B.2)–(B.5) leads to

$$\begin{aligned} &p(\widetilde{\mathbf{Y}}^j, \widetilde{\mathbf{Y}}^{j+1}|\widetilde{\mathbf{X}}^s, s \le j) \\ &\sim \mathcal{N}\left( \begin{pmatrix} \widetilde{\boldsymbol{\mu}}^j \\ \widetilde{\mathbf{a}}_{\mathbf{0}}^j \Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\boldsymbol{\mu}}^j \end{pmatrix}, \begin{pmatrix} \widetilde{\mathbf{R}}^j & \widetilde{\mathbf{R}}^j (1 + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)^* \\ (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\mathbf{R}}^j & \widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j \Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\mathbf{R}}^j (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)^* \end{pmatrix} \right). \end{aligned} \tag{B.6}$$

In light of Lemma Appendix A.2, the result in (B.6) yields the conditional distribution,

$$p(\widetilde{\mathbf{Y}}^j|\widetilde{\mathbf{Y}}^{j+1}, \widetilde{\mathbf{X}}^s, s \le J) = p(\widetilde{\mathbf{Y}}^j|\widetilde{\mathbf{Y}}^{j+1}, \widetilde{\mathbf{X}}^s, s \le j) = \mathcal{N}(\widetilde{\mathbf{m}}^j, \widetilde{\mathbf{P}}^j), \tag{B.7}$$

where

$$\widetilde{\mathbf{m}}^j = \widetilde{\boldsymbol{\mu}}^j + \widetilde{\mathbf{C}}^j \big( \widetilde{\mathbf{Y}}^{j+1} - \widetilde{\mathbf{a}}_{\mathbf{0}}^j \Delta t - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\boldsymbol{\mu}}^j \big), \tag{B.8a}$$

$$\widetilde{\mathbf{P}}^j = \widetilde{\mathbf{R}}^j - \widetilde{\mathbf{C}}^j \big( \widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j \Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\mathbf{R}}^j (\mathbf{I} + \mathbf{a}_{\mathbf{1}}^j \Delta t)^* \big) (\mathbf{C}^j)^*, \tag{B.8b}$$

and the auxiliary matrix $\mathbf{C}$ is given by

$$\widetilde{\mathbf{C}}^j = \widetilde{\mathbf{R}}^j(\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)^* \big(\widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j \Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\mathbf{R}}^j (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j)^*\big)^{-1}. \tag{B.9}$$

Note that the first equality in (B.7) is due to the Markovian property of the underlying system. In fact, if $\widetilde{\mathbf{Y}}^{j+1}$ is known, then the conditional distribution of $\widetilde{\mathbf{Y}}^j$ has no dependence on $\widetilde{\mathbf{X}}^s, s \geq j+1$. This finishes the proof of Lemma 3.1. □

*Appendix B.2. Proof of Theorem 3.2*

*Proof.* Making use of the rules of the conditional distribution leads to

$$\begin{aligned} p(\widetilde{\mathbf{Y}}^{j+1}, \widetilde{\mathbf{Y}}^j | \widetilde{\mathbf{X}}^s, s \leq J) &= p(\widetilde{\mathbf{Y}}^j | \widetilde{\mathbf{Y}}^{j+1}, \widetilde{\mathbf{X}}^s, s \leq J) p(\widetilde{\mathbf{Y}}^{j+1} | \widetilde{\mathbf{X}}^s, s \leq J) \\ &= p(\widetilde{\mathbf{Y}}^j | \widetilde{\mathbf{Y}}^{j+1}, \widetilde{\mathbf{X}}^s, s \leq j) p(\widetilde{\mathbf{Y}}^{j+1} | \widetilde{\mathbf{X}}^s, s \leq J) \\ &\sim \mathcal{N}(\widetilde{\mathbf{Y}}^j | \widetilde{\mathbf{m}}^j, \widetilde{\mathbf{P}}^j) \mathcal{N}(\widetilde{\mathbf{Y}}^{j+1} | \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1}, \widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1}), \end{aligned} \tag{B.10}$$

where the first conditional Gaussian distribution has been given by (B.7) in Lemma 3.1 and $\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1}$ and $\widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1}$ stand for the smoother mean and the smoother covariance at $t_{j+1}$. Next, in light of Lemma Appendix A.1, we arrive at the following result

$$\begin{aligned} p(\widetilde{\mathbf{Y}}^j | \widetilde{\mathbf{X}}^s, s \leq J) &\sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^j, \widetilde{\mathbf{R}}_{\mathbf{s}}^j) \\ &\sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}^j + \widetilde{\mathbf{C}}(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \widetilde{\mathbf{a}}_{\mathbf{0}}^j \Delta t - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\boldsymbol{\mu}}^j),\ \widetilde{\mathbf{P}}^j + \widetilde{\mathbf{C}}^j \widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} (\widetilde{\mathbf{C}}^j)^*). \end{aligned} \tag{B.11}$$

Finally, with the help of (B.8b), the mean and covariance from the smoother in (B.11) become

$$\begin{aligned} \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^j &= \widetilde{\boldsymbol{\mu}}^j + \widetilde{\mathbf{C}}\big(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \widetilde{\mathbf{a}}_{\mathbf{0}}^j \Delta t - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\boldsymbol{\mu}}^j\big) \\ \widetilde{\mathbf{R}}_{\mathbf{s}}^j &= \widetilde{\mathbf{P}}^j + \widetilde{\mathbf{C}}^j \widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} (\widetilde{\mathbf{C}}^j)^* \\ &= \widetilde{\mathbf{R}}^j + \widetilde{\mathbf{C}}^j \big(\widetilde{\mathbf{R}}_{\mathbf{s+1}}^j - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)\widetilde{\mathbf{R}}^j (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^j \Delta t)^* - \widetilde{\mathbf{b}}^j \circ \widetilde{\mathbf{b}}^j \Delta t\big)(\widetilde{\mathbf{C}}^j)^*. \end{aligned} \tag{B.12}$$

This finishes the proof of Theorem 3.2. □

*Appendix B.3. Proof of Theorem 3.3*

*Proof.* Recall from (B.7) that $\widetilde{\mathbf{Y}}^{j+1}$ is calculated by generate a multivariate Gaussian random variable with mean $\widetilde{\mathbf{m}}^j$ and covariance $\widetilde{\mathbf{P}}^j$. The backward equation of sampling $\mathbf{Y}$ can be derived by making use of (B.8) and (B.9).

Plugging (B.9) into (B.8b) yields,

$$\begin{aligned}
\widetilde{\mathbf{P}}^j &= \widetilde{\mathbf{R}}^j - \widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)^*\big(\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\Delta t + (\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)^*\big)^{-1}(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)\widetilde{\mathbf{R}}^j \\
&= \widetilde{\mathbf{R}}^j - \widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)^*\big(\widetilde{\mathbf{R}}^j + (\widetilde{\mathbf{a}}_1^j\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_1^j + \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t\big)^{-1}(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)\widetilde{\mathbf{R}}^j + O(\Delta t^2) \\
&= \widetilde{\mathbf{R}}^j - \widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)^*\big(\widetilde{\mathbf{R}}^j(\mathbf{I}+(\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{a}}_1^j\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_1^j + \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t)\big)^{-1} \\
&\qquad\qquad \times(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)\widetilde{\mathbf{R}}^j + O(\Delta t^2) \\
&= \widetilde{\mathbf{R}}^j - \widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)^*\big(\mathbf{I}-(\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{a}}_1^j\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_1^j + \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t\big)(\widetilde{\mathbf{R}}^j)^{-1} \\
&\qquad\qquad \times(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)\widetilde{\mathbf{R}}^j + O(\Delta t^2)
\end{aligned} \tag{B.13}$$

For notation simplicity, we define

$$\mathbf{F} = (\widetilde{\mathbf{R}}^j)^{-1}\big(\widetilde{\mathbf{a}}_1^j\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_1^j + \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\big), \tag{B.14}$$

and thus

$$\begin{aligned}
&\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)^*\big(\mathbf{I}-(\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{a}}_1^j\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_1^j + \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t\big)(\widetilde{\mathbf{R}}^j)^{-1}(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)\widetilde{\mathbf{R}}^j \\
=&\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)^*\big(\mathbf{I}-\mathbf{F}\Delta t\big)(\widetilde{\mathbf{R}}^j)^{-1}(\mathbf{I}+\widetilde{\mathbf{a}}_1^j\Delta t)\widetilde{\mathbf{R}}^j \\
=&(\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_1^j\Delta t)((\widetilde{\mathbf{R}}^j)^{-1} - \mathbf{F}(\widetilde{\mathbf{R}}^j)^{-1}\Delta t)(\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{a}}_\mathbf{j}^j\widetilde{\mathbf{R}}^j\Delta t) \\
=&(\mathbf{I} - \widetilde{\mathbf{R}}^j\mathbf{F}(\widetilde{\mathbf{R}}^j)^{-1} + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_1^j(\widetilde{\mathbf{R}}^j)^{-1}\Delta t)(\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{a}}_1^j\widetilde{\mathbf{R}}^j\Delta t) \\
=&\widetilde{\mathbf{R}}^j - \widetilde{\mathbf{R}}^j\mathbf{F}\Delta t + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_1^j\Delta t + \widetilde{\mathbf{a}}_1^j\widetilde{\mathbf{R}}^j\Delta t + O(\Delta t^2).
\end{aligned} \tag{B.15}$$

Plugging B.16 back to (B.13) yields

$$\begin{aligned}
\widetilde{\mathbf{P}}^j &= \widetilde{\mathbf{R}}^j - \Big(\widetilde{\mathbf{R}}^j - \widetilde{\mathbf{R}}^j\mathbf{F}\Delta t + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_1^j\Delta t + \widetilde{\mathbf{a}}_1^j\widetilde{\mathbf{R}}^j\Delta t\Big) + O(\Delta t^2) \\
&= \widetilde{\mathbf{R}}^j\mathbf{F}\Delta t - \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_1^j\Delta t - \widetilde{\mathbf{a}}_1^j\widetilde{\mathbf{R}}^j\Delta t + O(\Delta t^2) \\
&= \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\Delta t + O(\Delta t^2)
\end{aligned} \tag{B.16}$$

Applying a similar argument, plugging (B.9) into (B.8a) yields and subtracting $\widetilde{\mathbf{Y}}^{j+1}$ on both sides of (B.8a) yields,

$$\begin{aligned}
\widetilde{\mathbf{m}}^j - \widetilde{\mathbf{Y}}^{j+1} &= \widetilde{\boldsymbol{\mu}}^j + \widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_\mathbf{1}^j\Delta t)^*\big(\widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j\Delta t + (\mathbf{I}+\widetilde{\mathbf{a}}_\mathbf{1}^j\Delta t)\widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_\mathbf{1}^j)^*\big)^{-1} \\
&\qquad\qquad \times\big(\widetilde{\mathbf{Y}}^{j+1} - \widetilde{\mathbf{a}}_\mathbf{0}^j\Delta t - (\mathbf{I}+\widetilde{\mathbf{a}}_\mathbf{1}^j\Delta t)\widetilde{\boldsymbol{\mu}}^j\big) - \widetilde{\mathbf{Y}}^{j+1} \\
&= \widetilde{\boldsymbol{\mu}}^j + \widetilde{\mathbf{R}}^j(\mathbf{I}+\widetilde{\mathbf{a}}_\mathbf{1}^j\Delta t)^*\big(\mathbf{I} - (\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{a}}_\mathbf{1}^j\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_\mathbf{1}^j + \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t\big)(\widetilde{\mathbf{R}}^j)^{-1} \\
&\qquad\qquad \times\big(\widetilde{\mathbf{Y}}^{j+1} - \widetilde{\mathbf{a}}_\mathbf{0}^j\Delta t - (\mathbf{I}+\widetilde{\mathbf{a}}_\mathbf{1}^j\Delta t)\widetilde{\boldsymbol{\mu}}^j\big) - \widetilde{\mathbf{Y}}^{j+1} + O(\Delta t^2) \\
&= \widetilde{\boldsymbol{\mu}}^j + (\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_\mathbf{1}^j\Delta t)\big((\widetilde{\mathbf{R}}^j)^{-1} - (\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{a}}_\mathbf{1}^j\widetilde{\mathbf{R}}^j + \widetilde{\mathbf{R}}^j\widetilde{\mathbf{a}}_\mathbf{1}^j + \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)(\widetilde{\mathbf{R}}^j)^{-1}\Delta t\big) \\
&\qquad\qquad \times\big(\widetilde{\mathbf{Y}}^{j+1} - \widetilde{\mathbf{a}}_\mathbf{0}^j\Delta t - (\mathbf{I}+\widetilde{\mathbf{a}}_\mathbf{1}^j\Delta t)\widetilde{\boldsymbol{\mu}}^j\big) - \widetilde{\mathbf{Y}}^{j+1} + O(\Delta t^2) \\
&= \widetilde{\boldsymbol{\mu}}^j + \big(\mathbf{I} - (\widetilde{\mathbf{a}}_\mathbf{1}^j + \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j)\Delta t\big)\big(\widetilde{\mathbf{Y}}^{j+1} - \widetilde{\boldsymbol{\mu}}^j - (\widetilde{\mathbf{a}}_\mathbf{0}^j + \widetilde{\mathbf{a}}_\mathbf{1}^j\widetilde{\boldsymbol{\mu}}^j)\Delta t\big) - \widetilde{\mathbf{Y}}^{j+1} + O(\Delta t^2) \\
&= -(\widetilde{\mathbf{a}}_\mathbf{0}^j + \widetilde{\mathbf{a}}_\mathbf{1}^j\widetilde{\boldsymbol{\mu}}^j)\Delta t - (\widetilde{\mathbf{a}}_\mathbf{1}^j + \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j(\widetilde{\mathbf{R}}^j)^{-1})(\widetilde{\mathbf{Y}}^{j+1} - \widetilde{\boldsymbol{\mu}}^j)\Delta t + O(\Delta t^2) \\
&= -(\widetilde{\mathbf{a}}_\mathbf{0}^j + \widetilde{\mathbf{a}}_\mathbf{1}^j\widetilde{\mathbf{Y}}^{j+1})\Delta t - \widetilde{\mathbf{b}}^j\circ\widetilde{\mathbf{b}}^j(\widetilde{\mathbf{R}}^j)^{-1}(\widetilde{\mathbf{Y}}^{j+1} - \widetilde{\boldsymbol{\mu}}^j)\Delta t + O(\Delta t^2)
\end{aligned} \tag{B.17}$$

Combining (B.16) and (B.17) and taking the limit $\Delta t \to 0$ yields an explicit formula of sampling the unobserved processes $Z(t)$,

$$\mathrm{d}(-\mathbf{Y}) = \big(-\mathbf{a_0} - \mathbf{a_1}\mathbf{Y}\big)\,\mathrm{d}t + (\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1}\big(\boldsymbol{\mu} - \mathbf{Y}\big)dt + \mathbf{b_1}\,\mathrm{d}\mathbf{W_{Y,1}} + \mathbf{b_2}\,\mathrm{d}\mathbf{W_{Y,2}}, \tag{B.18}$$

This finishes the proof.

730 □

*Appendix B.4. Proof of Theorem 3.4*

*Proof.* At each fixed time $t$, a mean-fluctuation decomposition of $\mathbf{Y}$ yields

$$\mathbf{Y} = \langle\mathbf{Y}\rangle + \mathbf{Y}', \tag{B.19}$$

where $\langle\cdot\rangle$ denotes the ensemble mean and $\cdot'$ is the fluctuation with $\langle\cdot'\rangle = 0$. Therefore,

$$\boldsymbol{\mu}_\mathbf{s} = \langle\mathbf{Y}\rangle, \qquad \text{and} \qquad \mathbf{R}_\mathbf{s} = \langle\mathbf{Y}'(\mathbf{Y}')^*\rangle \tag{B.20}$$

Now taking the ensemble average of both the left and right hand sides of (9) yields

$$\mathrm{d}(-\boldsymbol{\mu}_\mathbf{s}) = \big(-\mathbf{a_0} - \mathbf{a_1}\boldsymbol{\mu}_\mathbf{s} + (\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_\mathbf{s})\big)\,\mathrm{d}t, \tag{B.21}$$

which is (10a).

Next, taking the difference between (9) and (B.21) yields,

$$\mathrm{d}(-\mathbf{Y}') = \big( -\mathbf{a_1} - (\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1}\big)\mathbf{Y}'\,\mathrm{d}t + \mathbf{b_1}\,\mathrm{d}\mathbf{W}_{\mathbf{Y},\mathbf{1}} + \mathbf{b_2}\,\mathrm{d}\mathbf{W}_{\mathbf{Y},\mathbf{2}}. \tag{B.22}$$

The covariance can be solved by making use of the Ito's formula [44],

$$\mathrm{d}(-\mathbf{R_s}) := \mathrm{d}(-\mathbf{Y}'(\mathbf{Y}')^*) = (\mathbf{Y}')^*\,\mathrm{d}(-\mathbf{Y}') + \mathbf{Y}'\,\mathrm{d}(-(\mathbf{Y}')^*) + \mathrm{d}(-\mathbf{Y}')\,\mathrm{d}(-(\mathbf{Y}')^*), \tag{B.23}$$

which combining with (B.22) leads to

$$\mathrm{d}(-\mathbf{R_s}) = -\big((\mathbf{a_1} + (\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1})\mathbf{R_s} + \mathbf{R_s}(\mathbf{a_1}^* + (\mathbf{b}\circ\mathbf{b})\mathbf{R}^{-1}) - \mathbf{b}\circ\mathbf{b}\big)\,\mathrm{d}t, \tag{B.24}$$

which is (10b). □

*Appendix B.5. Proof the equivalency of the results in Theorem 3.4 and Theorem 3.2*

*Proof.* Let us start with the result in Theorem 3.2

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j} &= \widetilde{\boldsymbol{\mu}}^{j} + \widetilde{\mathbf{C}}^{j}\big(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \widetilde{\mathbf{a}}_{\mathbf{0}}^{j}\Delta t - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)\widetilde{\boldsymbol{\mu}}^{j}\big) \\
\widetilde{\mathbf{R}}_{\mathbf{s}}^{j} &= \widetilde{\mathbf{R}}^{j} + \widetilde{\mathbf{C}}^{j}\big(\widetilde{\mathbf{R}}_{\mathbf{s+1}}^{j} - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)\widetilde{\mathbf{R}}^{j}(\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)^* - \widetilde{\mathbf{b}}^{j}\circ\widetilde{\mathbf{b}}^{j}\Delta t\big)(\widetilde{\mathbf{C}}^{j})^*,
\end{aligned} \tag{B.25}$$

with

$$\widetilde{\mathbf{C}}^{j} = \widetilde{\mathbf{R}}^{j}(\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)^*\big(\widetilde{\mathbf{b}}^{j}\circ\widetilde{\mathbf{b}}^{j}\Delta t + (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)\widetilde{\mathbf{R}}^{j}(\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)^*\big)^{-1}. \tag{B.26}$$

Subtracting $\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1}$ on both sides of the first equation in (B.25) yields,

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j} - \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} &= \widetilde{\boldsymbol{\mu}}^{j} - \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} + \widetilde{\mathbf{C}}^{j}\big(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \tilde{\mathbf{a}}_{\mathbf{0}}^{j}\Delta t - (\mathbf{I} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)\widetilde{\boldsymbol{\mu}}^{j}\big) \\
&= \widetilde{\boldsymbol{\mu}}^{j} - \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} + \widetilde{\mathbf{R}}^{j}(\mathbf{I} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)^{*}\big(\widetilde{\mathbf{R}}^{j} + (\widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\widetilde{\mathbf{R}}^{j} + \widetilde{\mathbf{R}}^{j}(\tilde{\mathbf{a}}_{\mathbf{1}}^{j})^{*})\Delta t\big)^{-1} \\
&\qquad \times \big(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \tilde{\mathbf{a}}_{\mathbf{0}}\Delta t - (\mathbf{I} + \tilde{\mathbf{a}}_{\mathbf{1}}\Delta t)\widetilde{\boldsymbol{\mu}}^{j}\big) + O(\Delta t^{2}) \\
&= \widetilde{\boldsymbol{\mu}}^{j} - \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} + \widetilde{\mathbf{R}}^{j}(\mathbf{I} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)^{*}\big(\mathbf{I} + (\widetilde{\mathbf{R}}^{j})^{-1}(\widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\widetilde{\mathbf{R}}^{j} + \widetilde{\mathbf{R}}^{j}(\tilde{\mathbf{a}}_{\mathbf{1}}^{j})^{*})\Delta t\big)^{-1}(\widetilde{\mathbf{R}}^{j})^{-1} \\
&\qquad \times \big(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \tilde{\mathbf{a}}_{\mathbf{0}}\Delta t - (\mathbf{I} + \tilde{\mathbf{a}}_{\mathbf{1}}\Delta t)\widetilde{\boldsymbol{\mu}}^{j}\big) + O(\Delta t^{2}) \\
&= \widetilde{\boldsymbol{\mu}}^{j} - \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} + \widetilde{\mathbf{R}}^{j}(\mathbf{I} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)^{*}\big(\mathbf{I} - (\widetilde{\mathbf{R}}^{j})^{-1}(\widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\widetilde{\mathbf{R}}^{j} + \widetilde{\mathbf{R}}^{j}(\tilde{\mathbf{a}}_{\mathbf{1}}^{j})^{*})\Delta t\big)(\widetilde{\mathbf{R}}^{j})^{-1} \\
&\qquad \times \big(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \tilde{\mathbf{a}}_{\mathbf{0}}^{j}\Delta t - (\mathbf{I} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)\widetilde{\boldsymbol{\mu}}^{j}\big) + O(\Delta t^{2}) \\
&= \widetilde{\boldsymbol{\mu}}^{j} - \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} + \big(\widetilde{\mathbf{R}}^{j} - (\widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\widetilde{\mathbf{R}}^{j} + \widetilde{\mathbf{R}}^{j}(\tilde{\mathbf{a}}_{\mathbf{1}}^{j})^{*})\Delta t + \widetilde{\mathbf{R}}^{j}(\tilde{\mathbf{a}}_{\mathbf{1}}^{j})^{*}\Delta t\big)(\widetilde{\mathbf{R}}^{j})^{-1} \\
&\qquad \times \big(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \tilde{\mathbf{a}}_{\mathbf{0}}^{j}\Delta t - (\mathbf{I} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)\widetilde{\boldsymbol{\mu}}^{j}\big) + O(\Delta t^{2}) \\
&= \widetilde{\boldsymbol{\mu}}^{j} - \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} + \big(\mathbf{I} - (\tilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}\widetilde{\mathbf{R}}^{j})^{-1}\Delta t\big) \\
&\qquad \times \big(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \tilde{\mathbf{a}}_{\mathbf{0}}^{j}\Delta t - (\mathbf{I} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)\widetilde{\boldsymbol{\mu}}^{j}\big) + O(\Delta t^{2}) \\
&= \widetilde{\boldsymbol{\mu}}^{j} - \widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} + (\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \tilde{\mathbf{a}}_{\mathbf{0}}^{j}\Delta t - (\mathbf{I} + \tilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)\widetilde{\boldsymbol{\mu}}^{j}) - (\tilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}(\widetilde{\mathbf{R}}^{j})^{-1}) \\
&\qquad \times (\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \widetilde{\boldsymbol{\mu}}^{j})\Delta t + O(\Delta t^{2}) \\
&= (\tilde{\mathbf{a}}_{\mathbf{0}}^{j} - \tilde{\mathbf{a}}_{\mathbf{0}}^{j}\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}(\widetilde{\mathbf{R}}^{j})^{-1}(\widetilde{\boldsymbol{\mu}}_{\mathbf{s}}^{j+1} - \widetilde{\boldsymbol{\mu}}^{j}))\Delta t + O(\Delta t^{2})
\end{aligned} \tag{B.27}$$

Taking the limit $\Delta t \to 0$, the above equation becomes

$$\mathrm{d}(-\boldsymbol{\mu}_{\mathbf{s}}) = \big(-\mathbf{a}_{\mathbf{0}} - \mathbf{a}_{\mathbf{1}}\boldsymbol{\mu}_{\mathbf{s}} + (\mathbf{b} \circ \mathbf{b})\mathbf{R}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathbf{s}})\big)\,\mathrm{d}t \tag{B.28}$$

which is exactly (10a) as in Theorem 3.4.

Similarly, subtracting $\widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1}$ on both sides of the second equation in (B.25)

yields,

$$\begin{aligned}
\widetilde{\mathbf{R}}_{\mathbf{s}}^{j} - \widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} &= \widetilde{\mathbf{R}}^{j} - \widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} + \widetilde{\mathbf{C}}^{j}\big(\widetilde{\mathbf{R}}_{\mathbf{s+1}}^{j} - (\mathbf{I} + \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)\widetilde{\mathbf{R}}^{j}(1 + \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\Delta t)^{*} - \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}\Delta t\big)(\widetilde{\mathbf{C}}^{j})^{*}, \\
&= \widetilde{\mathbf{R}}^{j} - \widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} + \big(\mathbf{I} - (\widetilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}\widetilde{\mathbf{R}}^{j})^{-1}\Delta t\big) \\
&\quad \times \big(\widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} - \widetilde{\mathbf{R}}^{j} - (\widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\widetilde{\mathbf{R}}^{j} + \widetilde{\mathbf{R}}^{j}\widetilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j})\Delta t\big)\big(\mathbf{I} - (\widetilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}\widetilde{\mathbf{R}}^{j})^{-1}\Delta t\big)^{*} \\
&= \widetilde{\mathbf{R}}^{j} - \widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} + \widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} - \widetilde{\mathbf{R}}^{j} - (\widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\widetilde{\mathbf{R}}^{j} + \widetilde{\mathbf{R}}^{j}\widetilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j})\Delta t \\
&\quad - (\widetilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}(\widetilde{\mathbf{R}}^{j})^{-1})(\widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} - \widetilde{\mathbf{R}}^{j})\Delta t - (\widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} - \widetilde{\mathbf{R}}^{j})(\widetilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}(\widetilde{\mathbf{R}}^{j})^{-1})^{*}\Delta t \\
&= \Big( - \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\widetilde{\mathbf{R}}^{j} - \widetilde{\mathbf{R}}^{j}\widetilde{\mathbf{a}}_{\mathbf{1}}^{j} - \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j} - (\widetilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}(\widetilde{\mathbf{R}}^{j})^{-1})\widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} \\
&\quad - \widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1}(\widetilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}(\widetilde{\mathbf{R}}^{j})^{-1})^{*} + \widetilde{\mathbf{a}}_{\mathbf{1}}^{j}\widetilde{\mathbf{R}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j} + \widetilde{\mathbf{R}}^{j}(\widetilde{\mathbf{a}}_{\mathbf{1}}^{j})^{*} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}\Big)\Delta t \\
&= -\Big((\widetilde{\mathbf{a}}_{\mathbf{1}}^{j} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}(\widetilde{\mathbf{R}}^{j})^{-1})\widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1} + \widetilde{\mathbf{R}}_{\mathbf{s}}^{j+1}((\widetilde{\mathbf{a}}_{\mathbf{1}}^{j})^{*} + \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}(\widetilde{\mathbf{R}}^{j})^{-1}) - \widetilde{\mathbf{b}}^{j} \circ \widetilde{\mathbf{b}}^{j}\Big)\Delta t
\end{aligned} \tag{B.29}$$

Taking the limit $\Delta t \to 0$, the above equation becomes

$$\mathrm{d}(-\mathbf{R}_{\mathbf{s}}) = -\big((\mathbf{a_1} + (\mathbf{b} \circ \mathbf{b})\mathbf{R}^{-1})\mathbf{R}_{\mathbf{s}} + \mathbf{R}_{\mathbf{s}}(\mathbf{a}_{\mathbf{1}}^{*} + (\mathbf{b} \circ \mathbf{b})\mathbf{R}^{-1}) - \mathbf{b} \circ \mathbf{b}\big)\,\mathrm{d}t, \tag{B.30}$$

which is exactly (10b) as in Theorem 3.4.

Therefore, the equivalency of the results in Theorem 3.4 and Theorem 3.2 has been proved. □

## Appendix C. Proof of the theorems related to the discrete time conditional Gaussian systems

*Appendix C.1. Proof of Theorem 3.5*

*Proof.* Consider the joint distribution $p(\mathbf{X}^{j+1}, \mathbf{Y}^{j+1} | \mathbf{X}^{s}, s \le j)$.

In light of (11a), it is easy to derive, by evolving the model forward, that

$$p(\mathbf{X}^{j+1} | \mathbf{X}^{s}, s \le j) \sim \mathcal{N}(\mathbf{A}_{\mathbf{0}}^{j} + \mathbf{A}_{\mathbf{1}}^{j}\boldsymbol{\mu}^{j}, \mathbf{A}_{\mathbf{1}}^{j}\mathbf{R}^{j}(\mathbf{A}_{\mathbf{1}}^{j})^{*} + \mathbf{B}^{j} \circ \mathbf{B}^{j}). \tag{C.1}$$

Using the same argument, running the model (11b) forward yields

$$p(\mathbf{Y}^{j+1} | \mathbf{X}^{s}, s \le j) \sim \mathcal{N}(\mathbf{a}_{\mathbf{0}}^{j} + \mathbf{a}_{\mathbf{1}}^{j}\boldsymbol{\mu}^{j}, \mathbf{a}_{\mathbf{1}}^{j}\mathbf{R}^{j}(\mathbf{a}_{\mathbf{1}}^{j})^{*} + \mathbf{b}^{j} \circ \mathbf{b}^{j}). \tag{C.2}$$

The cross-covariance term can be derived by first removing the mean in (11a) and then multiplying the resulting equation by $(\mathbf{Y}'^{,j+1})^*$, where $(\mathbf{Y}'^{,j+1})^*$ is $(\mathbf{Y}^{j+1})^*$ subtracting its mean. The result is

$$\langle \mathbf{X}'^{,j+1}(\mathbf{Y}'^{,j+1})^* \rangle = \mathbf{A}_\mathbf{1}^j \mathbf{R}^j (\mathbf{a}_\mathbf{1}^j)^* + (\mathbf{b}^j \circ \mathbf{B}^j)^*. \tag{C.3}$$

Collecting (C.1), (C.2) and (C.3) leads to

$$p(\mathbf{X}^{j+1}, \mathbf{Y}^{j+1} | \mathbf{X}^s, s \leq j)$$
$$\sim \mathcal{N}\left( \begin{pmatrix} \mathbf{A}_\mathbf{0}^j + \mathbf{A}_\mathbf{1}^j \boldsymbol{\mu}^j \\ \mathbf{a}_\mathbf{0}^j + \mathbf{a}_\mathbf{1}^j \boldsymbol{\mu}^j \end{pmatrix}, \begin{pmatrix} \mathbf{A}_\mathbf{1}^j \mathbf{R}^j (\mathbf{A}_\mathbf{1}^j)^* + \mathbf{B}^j \circ \mathbf{B}^j & \mathbf{A}_\mathbf{1}^j \mathbf{R}^j (\mathbf{a}_\mathbf{1}^j)^* + (\mathbf{b}^j \circ \mathbf{B}^j)^* \\ \mathbf{a}_\mathbf{1}^j \mathbf{R}^j (\mathbf{A}_\mathbf{1}^j)^* + \mathbf{b}^j \circ \mathbf{B}^j & \mathbf{a}_\mathbf{1}^j \mathbf{R}^j (\mathbf{a}_\mathbf{1}^j)^* + \mathbf{b}^j \circ \mathbf{b}^j \end{pmatrix} \right) \tag{C.4}$$

Then making use of (A.2) in Lemma Appendix A.2 finishes the proof. □

*Appendix C.2. Proof of Theorem 3.6*

*Proof.* Let us start with the joint distribution $p(\mathbf{Y}^j, \mathbf{Y}^{j+1} | \mathbf{X}^s, s \leq j)$. Applying the conditional distribution rules yields

$$p(\mathbf{Y}^j, \mathbf{Y}^{j+1} | \mathbf{X}^s, s \leq j) = p(\mathbf{Y}^{j+1} | \mathbf{Y}^j, \mathbf{X}^s, s \leq j)\, p(\mathbf{Y}^j | \mathbf{X}^s, s \leq j). \tag{C.5}$$

The first distribution on the right hand side of (C.5) can be calculated in light of (11b),

$$p(\mathbf{Y}^{j+1} | \mathbf{Y}^j, \mathbf{X}^s, s \leq j) \sim \mathcal{N}\big(\mathbf{a}_\mathbf{0}^j + \mathbf{a}_\mathbf{1}^j \boldsymbol{\mu}^j, \mathbf{b}^j \circ \mathbf{b}^j + \mathbf{a}_\mathbf{1}^j \mathbf{R}^j (\mathbf{a}_\mathbf{1}^j)^*\big). \tag{C.6}$$

The second distribution on the right hand side of (C.5) is simply the filter estimate,

$$p(\mathbf{Y}^j | \mathbf{X}^s, s \leq j) \sim \mathcal{N}(\boldsymbol{\mu}^j, \mathbf{R}^j). \tag{C.7}$$

The cross-covariance between $\mathbf{Y}^{j+1}$ and $\mathbf{Y}^j$ can be calculated by making use of (11b), which gives

$$\langle \mathbf{Y}'^{,j+1}(\mathbf{Y}'^{,j})^* \rangle = \mathbf{a}_\mathbf{1}^j \mathbf{R}^j, \tag{C.8}$$

where $\mathbf{Y}'^{,j+1}$ and $\mathbf{Y}'^{,j}$ are $\mathbf{Y}^{j+1}$ and $\mathbf{Y}^j$ subtracting their means. Therefore, collecting (C.6), (C.7) and (C.8) leads to

$$p(\mathbf{Y}^j, \mathbf{Y}^{j+1} | \mathbf{X}^s, s \leq j) \sim \mathcal{N}\left( \begin{pmatrix} \boldsymbol{\mu}^j \\ \mathbf{a}_\mathbf{0}^j + \mathbf{a}_\mathbf{1}^j \boldsymbol{\mu}^j \end{pmatrix}, \begin{pmatrix} \mathbf{R}^j & \mathbf{R}^j (\mathbf{a}_\mathbf{1}^j)^* \\ \mathbf{a}_\mathbf{1}^j \mathbf{R}^j & \mathbf{b}^j \circ \mathbf{b}^j + \mathbf{a}_\mathbf{1}^j \mathbf{R}^j (\mathbf{a}_\mathbf{1}^j)^* \end{pmatrix} \right). \tag{C.9}$$

With the result (C.9) in hand, it is easy to see that

$$p(\mathbf{Y}^j|\mathbf{Y}^{j+1}, \mathbf{X}^s, s \leq n) = p(\mathbf{Y}^j|\mathbf{Y}^{j+1}, \mathbf{X}^s, s \leq j) = p(\mathbf{m}^j, \mathbf{P}^j), \tag{C.10}$$

where

$$\begin{aligned} \mathbf{m}^j &= \boldsymbol{\mu}^j + \mathbf{C}^j(\mathbf{Y}^{j+1} - \mathbf{a_0^j} - \mathbf{a_1^j}\boldsymbol{\mu}^j), \\ \mathbf{P}^j &= \mathbf{R}^j - \mathbf{C}^j(\mathbf{b}^j \circ \mathbf{b}^j + \mathbf{a_1^j}\mathbf{R}^j(\mathbf{a_1^j})^*)(\mathbf{C}^j)^*, \end{aligned} \tag{C.11}$$

and the auxiliary matrix $\mathbf{C}$ is given by

$$\mathbf{C}^j = \mathbf{R}^j(\mathbf{a_1^j})^*(\mathbf{b}^j \circ \mathbf{b}^j + \mathbf{a_1^j}\mathbf{R}^j(\mathbf{a_1^j})^*)^{-1}. \tag{C.12}$$

Next, using a similar technique as (C.5) yields

$$\begin{aligned} p(\mathbf{Y}^j, \mathbf{Y}^{j+1}|\mathbf{X}^s, s \leq n) &= p(\mathbf{Y}^j|\mathbf{Y}^{j+1}, \mathbf{X}^s, s \leq n)\, p(\mathbf{Y}^{j+1}|\mathbf{X}^s, s \leq n) \\ &\sim \mathcal{N}(\mathbf{Y}^j|\mathbf{m}^j, \mathbf{P}^j)\mathcal{N}(\mathbf{Y}^{j+1}|\boldsymbol{\mu}_\mathbf{s}^{j+1}, \mathbf{R}_\mathbf{s}^{j+1}). \end{aligned} \tag{C.13}$$

Finally, applying (A.1) in Lemma Appendix A.1 to (C.13) gives

$$\begin{aligned} p(\mathbf{Y}^j|\mathbf{X}^s, s \leq n) &\sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{s}^j, \mathbf{R}_\mathbf{s}^j) \\ &\sim \mathcal{N}(\boldsymbol{\mu}^j + \mathbf{C}^j(\boldsymbol{\mu}_\mathbf{s}^{j+1} - \mathbf{a_0^j} - \mathbf{a_1^j}\boldsymbol{\mu}^j), \mathbf{P} + \mathbf{C}^j\mathbf{R}_\mathbf{s}^{j+1}(\mathbf{C}^j)^{-1}). \end{aligned} \tag{C.14}$$

Plugging (C.11) into (C.14) gives the recursive backward smoothing formulae in (14)

$$\boldsymbol{\mu}_\mathbf{s}^j = \boldsymbol{\mu}^j + \mathbf{C}^j(\boldsymbol{\mu}_\mathbf{s}^{j+1} - \mathbf{a_0^j} - \mathbf{a_1^j}\boldsymbol{\mu}^j), \tag{C.15a}$$

$$\mathbf{R}_\mathbf{s}^j = \mathbf{R}^j + \mathbf{C}^j(\mathbf{R}_\mathbf{s}^{j+1} - \mathbf{a_1^j}\mathbf{R}^j(\mathbf{a_1^j})^* - \mathbf{b}^j \circ \mathbf{b}^j)(\mathbf{C}^j)^*. \tag{C.15b}$$

This finishes the proof □

*Appendix C.3. Proof of Theorem 3.7*

*Proof.* The proof of Theorem 3.7 is finished by making use of (C.10) and (C.11).

□

## Appendix D. Proof of Theorem 3.8

*Proof.* The marginal distribution of $\mathbf{X}$ at any fixed time $s$ is given by

$$p(\mathbf{X}(s)) = \lim_{L\to\infty} \sum_{l=1}^{L} \delta(\mathbf{X}(s) - \mathbf{X}_l(s)), \tag{D.1}$$

where $\delta(\cdot)$ is the Dirac delta function with the point mass being at zero. Thus, the same conclusion applies for the joint distribution

$$p(\mathbf{X}(0 \le s \le T)) = \lim_{L\to\infty} \sum_{l=1}^{L} \delta(\mathbf{X}(0 \le s \le T) - \mathbf{X}_l(0 \le s \le T)), \tag{D.2}$$

which is understood in the sense of applying a temporal discretization of the continuous path $\mathbf{X}$. According to the fundamental relationship between joint, marginal and conditional distributions, the marginal distribution of $\mathbf{Y}$ at time $t$ is given by

$$\begin{aligned} p(\mathbf{Y}(t)) &= \int p(\mathbf{X}(0 \le s \le T), \mathbf{Y}(t)) \, \mathrm{d}\mathbf{X}(0 \le s \le T) \\ &= \int p(\mathbf{X}(0 \le s \le T)) p(\mathbf{Y}(t) | \mathbf{X}(0 \le s \le T)) \, \mathrm{d}\mathbf{X}(0 \le s \le T) \end{aligned} \tag{D.3}$$

Inserting (D.2) into (D.3) yields,

$$p(\mathbf{Y}(t)) = \lim_{L\to\infty} \sum_{l=1}^{L} p(\mathbf{Y}(t) | \mathbf{X}_l(0 \le s \le T)), \tag{D.4}$$

where for each $l$,

$$p\big(\mathbf{Y}(t) | \mathbf{X}_l(0 \le s \le T)\big) \sim \mathcal{N}(\boldsymbol{\mu}_{l,\mathbf{s}}, \mathbf{R}_{l,\mathbf{s}}),$$

This finishes the proof. □

## Appendix E. Proof of Theorem 4.1

*Proof.* Let us start writing down the general from of a 1-D SDE with multiplicative noise,

$$\mathrm{d}x(t) = -\lambda(x(t) - m) \, \mathrm{d}t + \sigma(x) \, \mathrm{d}W(t), \tag{E.1}$$

where $m$ is an arbitrary real constant, and $\lambda$ is a real positive constant. Assume $\sigma(x)$ is twice continuously differentiable. Then the Fokker-Planck equation associated with (E.1) is given by

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}[\lambda(x-m)p(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[\sigma^2(x)p(x,t)]. \tag{E.2}$$

Taking the integration of (E.2) with respect to $x$ once, the PDF $p(x)$ of the stationary solution to (E.1) satisfies the equation

$$\lambda(x-m)p(x) + \frac{1}{2}\frac{\partial}{\partial x}[\sigma^2(x)p(x)] = \text{const.} \tag{E.3}$$

If $p(x)$ does not vanish on $\mathbb{R}$ and $p(x)$ is twice continuously differentiable, then we can set

$$m = \langle x(t)\rangle, \qquad \sigma^2(x) = \frac{2}{p(x)}\{-\lambda\Phi(x)\} \tag{E.4}$$

where

$$\Phi(x) = \int_b^x (y-m)p(y)dy. \tag{E.5}$$

□

## References


[1] A. J. Majda, Introduction to turbulent dynamical systems in complex systems, Springer, 2016.

[2] S. H. Strogatz, Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering, CRC Press, 2018.

[3] D. Baleanu, J. A. T. Machado, A. C. Luo, Fractional dynamics and control, Springer Science & Business Media, 2011.

[4] T. Deisboeck, J. Y. Kresh, Complex systems science in biomedicine, Springer Science & Business Media, 2007.

[5] M. Farazmand, T. Sapsis, Extreme events: Mechanisms and prediction, Applied Mechanics Reviews.

[6] M. W. Denny, L. J. Hunt, L. P. Miller, C. D. Harley, On the prediction of extreme ecological events, Ecological Monographs 79 (3) (2009) 397–421.

[7] M. A. Mohamad, T. P. Sapsis, Probabilistic description of extreme events in intermittently unstable dynamical systems excited by correlated stochastic processes, SIAM/ASA Journal on Uncertainty Quantification 3 (1) (2015) 709–736.

[8] E. Kalnay, Atmospheric modeling, data assimilation and predictability, Cambridge university press, 2003.

[9] W. Lahoz, B. Khattatov, R. Ménard, Data assimilation and information, in: Data Assimilation, Springer, 2010, pp. 3–12.

[10] A. J. Majda, J. Harlim, Filtering complex turbulent systems, Cambridge University Press, 2012.

[11] G. Evensen, Data assimilation: the ensemble Kalman filter, Springer Science & Business Media, 2009.

[12] K. Law, A. Stuart, K. Zygalakis, Data assimilation: a mathematical introduction, Vol. 62, Springer, 2015.

[13] H. E. Rauch, C. Striebel, F. Tung, Maximum likelihood estimates of linear dynamic systems, AIAA journal 3 (8) (1965) 1445–1450.

[14] J. S. Simonoff, Smoothing methods in statistics, Springer Science & Business Media, 2012.

[15] R. E. Kalman, A new approach to linear filtering and prediction problems, Journal of basic Engineering 82 (1) (1960) 35–45.

[16] G. J. Bierman, Factorization methods for discrete sequential estimation, Courier Corporation, 2006.

[17] R. Bellman, R. E. Kalaba, Dynamic programming and modern control theory, Vol. 81, Citeseer, 1965.

[18] N. Chen, A. Majda, Conditional gaussian systems for multiscale nonlinear stochastic systems: Prediction, state estimation and uncertainty quantification, Entropy 20 (7) (2018) 509.

[19] N. Chen, A. J. Majda, Filtering nonlinear turbulent dynamical systems through conditional Gaussian statistics, Monthly Weather Review 144 (12) (2016) 4885–4917.

[20] A. J. Majda, J. Harlim, Physics constrained nonlinear regression models for time series, Nonlinearity 26 (1) (2012) 201.

[21] J. Harlim, A. Mahdi, A. J. Majda, An ensemble kalman filter for statistical estimation of physics constrained nonlinear regression models, Journal of Computational Physics 257 (2014) 782–812.

[22] B. Lindner, J. Garcıa-Ojalvo, A. Neiman, L. Schimansky-Geier, Effects of noise in excitable systems, Physics reports 392 (6) (2004) 321–424.

[23] A. B. Medvinsky, S. V. Petrovskii, I. A. Tikhonova, H. Malchow, B.-L. Li, Spatiotemporal complexity of plankton and fish dynamics, SIAM review 44 (3) (2002) 311–370.

[24] R. Salmon, Lectures on geophysical fluid dynamics, Oxford University Press, 1998.

[25] G. K. Vallis, Atmospheric and oceanic fluid dynamics, Cambridge University Press, 2017.

[26] R. S. Liptser, A. N. Shiryaev, Statistics of random processes II: Applications, Vol. 6, Springer Science & Business Media, 2013.

[27] N. Chen, A. J. Majda, D. Giannakis, Predicting the cloud patterns of the Madden-Julian oscillation through a low-order nonlinear stochastic model, Geophysical Research Letters 41 (15) (2014) 5612–5619.

[28] N. Chen, A. J. Majda, Predicting the real-time multivariate Madden–Julian oscillation index through a low-order nonlinear stochastic model, Monthly Weather Review 143 (6) (2015) 2148–2169.

[29] N. Chen, A. J. Majda, Predicting the cloud patterns for the boreal summer intraseasonal oscillation through a low-order stochastic model, Mathematics of Climate and Weather Forecasting 1 (1) (2015) 1–20.

[30] N. Chen, A. J. Majda, C. Sabeerali, R. Ajayamohan, Predicting monsoon intraseasonal precipitation using a low-order nonlinear stochastic model, Journal of Climate 31 (11) (2018) 4403–4427.

[31] N. Chen, A. J. Majda, Filtering the stochastic skeleton model for the Madden–Julian oscillation, Monthly Weather Review 144 (2) (2016) 501–527.

[32] N. Chen, A. J. Majda, X. T. Tong, Information barriers for noisy Lagrangian tracers in filtering random incompressible flows, Nonlinearity 27 (9) (2014) 2133.

[33] N. Chen, A. J. Majda, X. T. Tong, Noisy Lagrangian tracers for filtering random rotating compressible flows, Journal of Nonlinear Science 25 (3) (2015) 451–488.

[34] N. Chen, A. J. Majda, Model error in filtering random compressible flows utilizing noisy Lagrangian tracers, Monthly Weather Review 144 (11) (2016) 4037–4061.

[35] M. Branicki, A. J. Majda, Dynamic stochastic superresolution of sparsely observed turbulent systems, Journal of Computational Physics 241 (2013) 333–363.

[36] S. R. Keating, A. J. Majda, K. S. Smith, New methods for estimating ocean eddy heat transport using satellite altimetry, Monthly Weather Review 140 (5) (2012) 1703–1722.

[37] A. J. Majda, I. Grooms, New perspectives on superparameterization for geophysical turbulence, Journal of Computational Physics 271 (2014) 60–77.

[38] A. J. Majda, D. Qi, T. P. Sapsis, Blended particle filters for large-dimensional chaotic dynamical systems, Proceedings of the National Academy of Sciences (2014) 201405675.

[39] P. E. Kloeden, E. Platen, Higher-order implicit strong numerical schemes for stochastic differential equations, Journal of statistical physics 66 (1-2) (1992) 283–314.

[40] R. Adrian, K. Christensen, Z.-C. Liu, Analysis and interpretation of instantaneous turbulent velocity fields, Experiments in fluids 29 (3) (2000) 275–290.

[41] N. Chen, A. J. Majda, Efficient statistically accurate algorithms for the fokker–planck equation in large dimensions, Journal of Computational Physics 354 (2018) 242–268.

[42] N. Chen, A. J. Majda, X. T. Tong, Rigorous analysis for efficient statistically accurate algorithms for solving fokker–planck equations in large dimensions, SIAM/ASA Journal on Uncertainty Quantification 6 (3) (2018) 1198–1223.

[43] N. Chen, A. J. Majda, Beating the curse of dimension with accurate statistics for the fokker–planck equation in complex turbulent systems, Proceedings of the National Academy of Sciences 114 (49) (2017) 12864–12869.

[44] C. W. Gardiner, Handbook of stochastic methods for physics, chemistry and the natural sciences, vol. 13 of springer series in synergetics (2004).

[45] A. J. Majda, D. Qi, Strategies for reduced-order models for predicting the statistical responses and uncertainty quantification in complex turbulent dynamical systems, SIAM Review 60 (3) (2018) 491–549.

[46] A. J. Majda, D. Qi, Linear and nonlinear statistical response theories with prototype applications to sensitivity analysis and statistical control of complex turbulent dynamical systems, Chaos: An Interdisciplinary Journal of Nonlinear ScienceSubmitted.

[47] D. Qi, A. J. Majda, Predicting fat-tailed intermittent probability distributions in passive scalar turbulence with imperfect models through empirical information theory, Communications in Mathematical Sciences 14 (6) (2016) 1687–1722.

[48] D. Qi, A. J. Majda, Low-dimensional reduced-order models for statistical response and uncertainty quantification: Two-layer baroclinic turbulence, Journal of the Atmospheric Sciences 73 (12) (2016) 4609–4639.

[49] R. E. Kalman, R. S. Bucy, New results in linear filtering and prediction theory, Journal of basic engineering 83 (1) (1961) 95–108.

[50] R. S. Bucy, P. D. Joseph, Filtering for stochastic processes with applications to guidance, Vol. 326, American Mathematical Soc., 2005.

[51] M. Branicki, A. J. Majda, Quantifying uncertainty for predictions with model error in non-Gaussian systems with intermittency, Nonlinearity 25 (9) (2012) 2543.

[52] A. J. Majda, R. Abramov, B. Gershgorin, High skill in low-frequency climate response through fluctuation dissipation theorems despite structural instability, Proceedings of the National Academy of Sciences 107 (2) (2010) 581–586.

[53] A. J. Majda, C. Franzke, D. Crommelin, Normal forms for reduced stochastic climate models, Proceedings of the National Academy of Sciences 106 (10) (2009) 3649–3653.

[54] B. Gershgorin, J. Harlim, A. J. Majda, Test models for improving filtering with model errors through stochastic parameter estimation, Journal of Computational Physics 229 (1) (2010) 1–31.

[55] B. Gershgorin, J. Harlim, A. J. Majda, Improving filtering and prediction of spatially extended turbulent systems with model errors through stochastic parameter estimation, Journal of Computational Physics 229 (1) (2010) 32–57.

[56] N. Chen, A. J. Majda, A new efficient parameter estimation algorithm for high-dimensional complex nonlinear turbulent dynamical systems with partial observations, Journal of Computational Physics 397 (2019) 108836.

[57] A. J. Majda, C. Franzke, B. Khouider, An applied mathematics perspective on stochastic modelling for climate, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 366 (1875) (2008) 2427–2453.

[58] A. Majda, R. V. Abramov, M. J. Grote, Information theory and stochastics for multiscale nonlinear systems, Vol. 25, American Mathematical Soc., 2005.

[59] A. J. Majda, I. Timofeyev, E. V. Eijnden, Models for stochastic climate prediction, Proceedings of the National Academy of Sciences 96 (26) (1999) 14687–14691.

[60] A. J. Majda, I. Timofeyev, E. Vanden Eijnden, A mathematical framework for stochastic climate models, Communications on Pure and Applied Mathematics 54 (8) (2001) 891–974.

[61] A. Majda, I. Timofeyev, E. Vanden-Eijnden, Stochastic models for selected slow variables in large deterministic systems, Nonlinearity 19 (4) (2006) 769.

[62] S. G. H. Philander, El nino southern oscillation phenomena, Nature 302 (5906) (1983) 295.

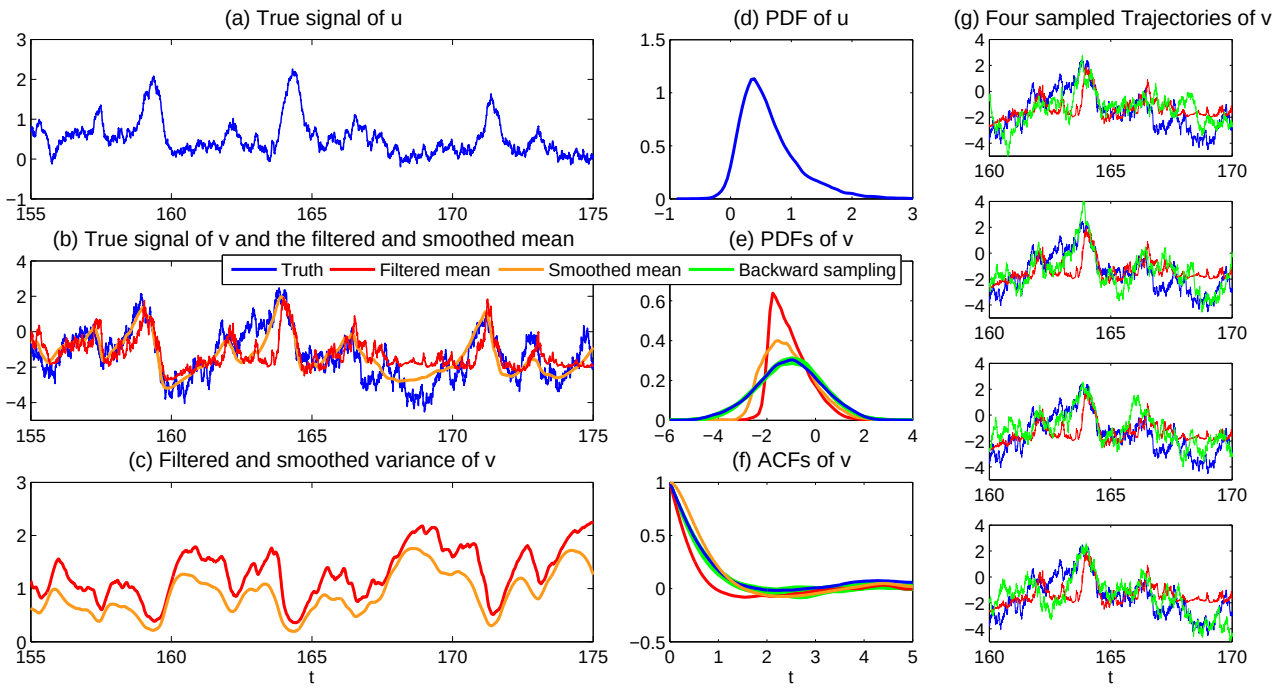


Figure 1: Perfect model test with the dyad model in (28). Panel (a): the true signal of $u$. Panel (b): the true signal (blue), the filter mean (red) and the smoother mean (brown) estimates. Panel (c): the filter covariance (red) and the smoother covariance (brown) estimates. Panel (d): the equilibrium PDF of the true signal of $u$, based on a trajectory with 500 time units. Panel (e): the equilibrium PDF of the true signal of $v$ (blue), the PDF associated with the time series of the filter mean (red) and smoother mean (brown) estimates, and the PDF associated with the sampled trajectory of $v$ using the backward sampling strategy (green). Panel (f): the temporal autocorrelation functions (ACFs) formed from different time series. Panel (g): comparison of the true signal (blue), the time series of the filter mean estimate (red) and four different sampled trajectories from the backward sampling strategy (green).

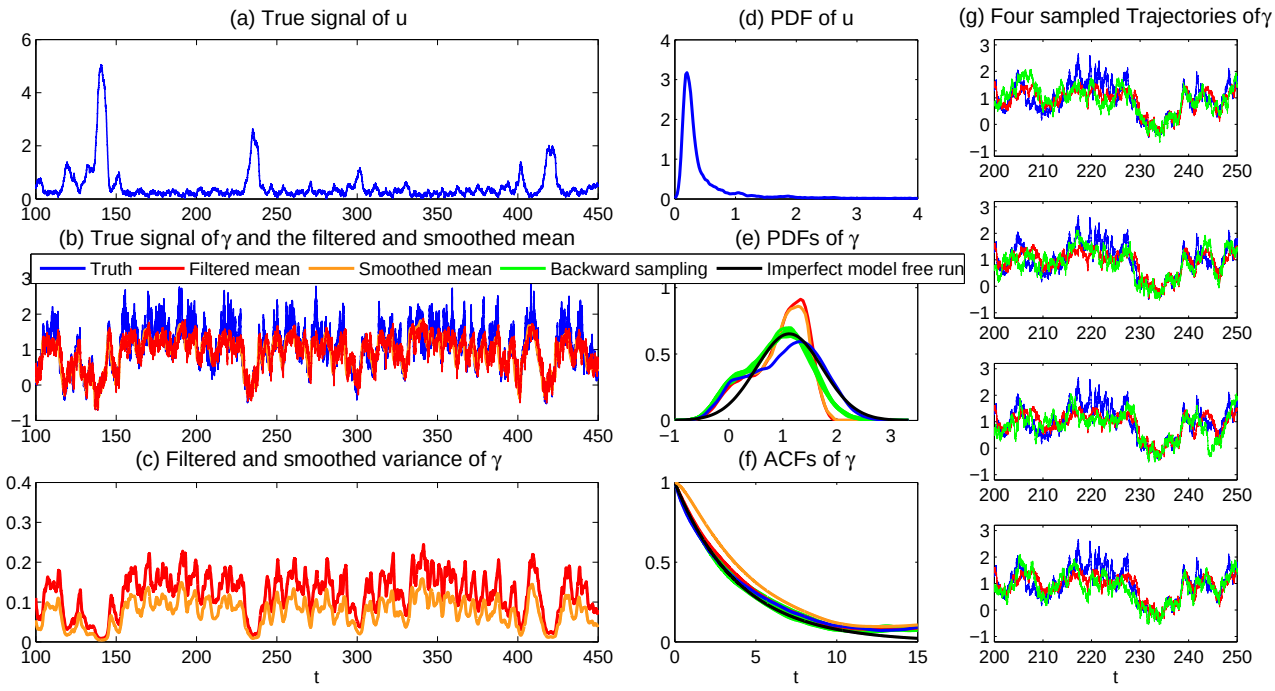


Figure 2: Similar illustration as Figure 1. Here the true signal is generated from the perfect model but the filter and smoother estimates as well as the backward sampling are all based on the approximate model.

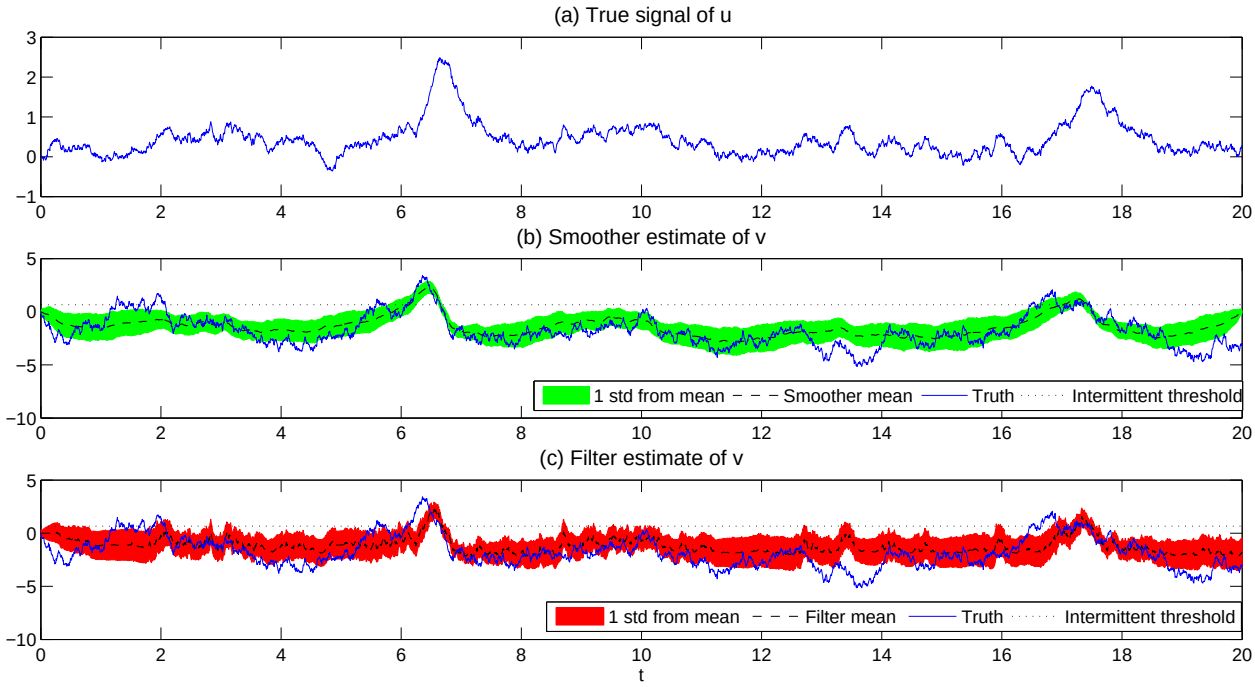


Figure 3: The physics-constrained dyad model (28) with parameters given by (29). Panel (a): one realization of the true signal $u$. Panel (b): the corresponding true signal of $v$ (blue) and the estimate from the nonlinear optimal smoothing, where the smoother mean is given by black dashed curve and one standard deviation (std) is shown in the green shading area. The horizontal dotted line shows the intermittent threshold $v^* = d_u/c = 0.67$. Panel (c): similar to (b) but for the filter estimates.

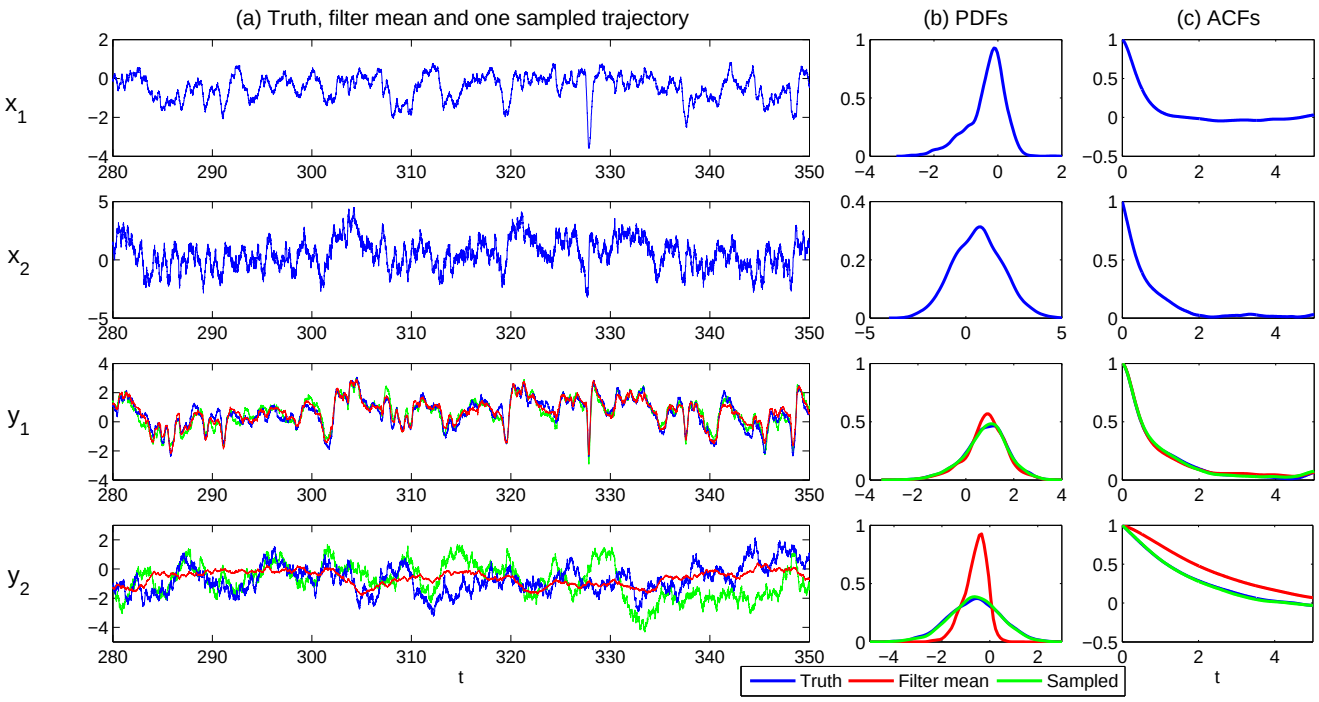


Figure 4: The four-dimensional stochastic climate model (33) with parameters (34). Blue curves show the true trajectories, the associated PDFs and ACFs. The red curves show those of the filtered mean state. The green curves show those of one sampled trajectories of the unobserved variables $y_1$ and $y_2$ from the backward sampling strategy.

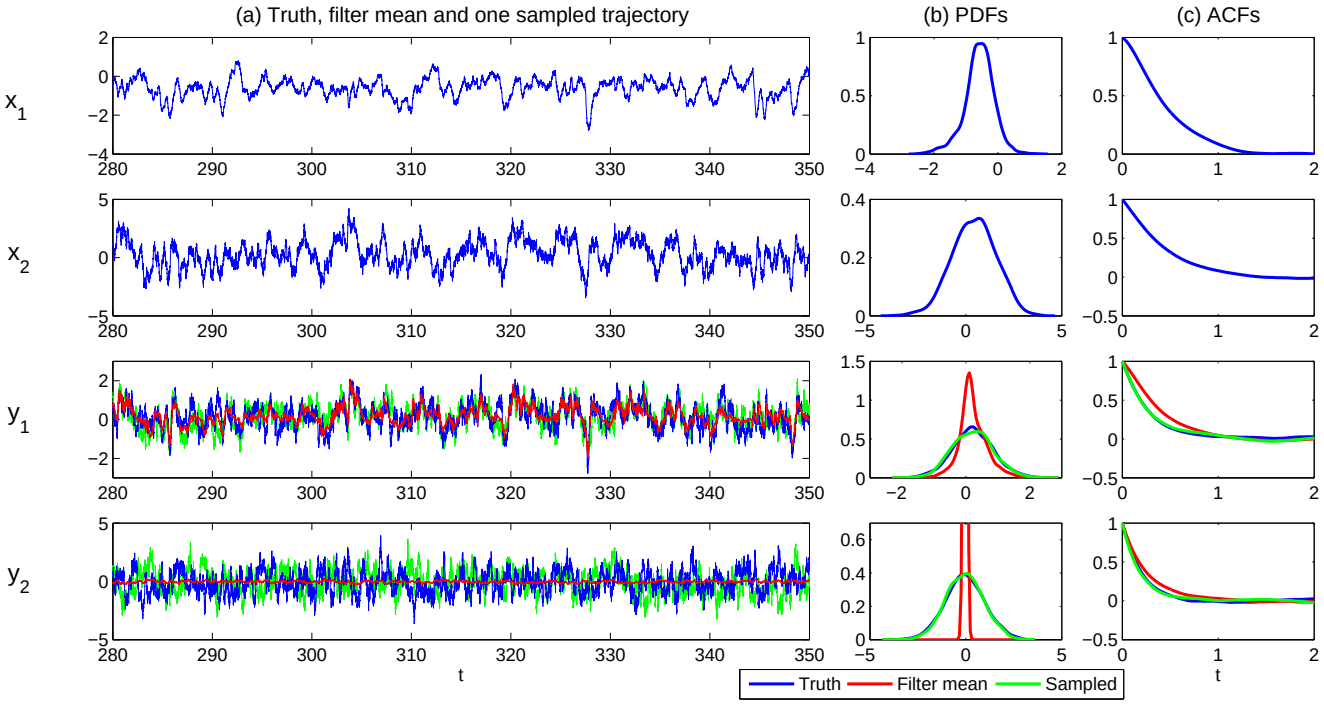


Figure 5: Similar to Figure 5 but for $\epsilon = 0.1$ regime.

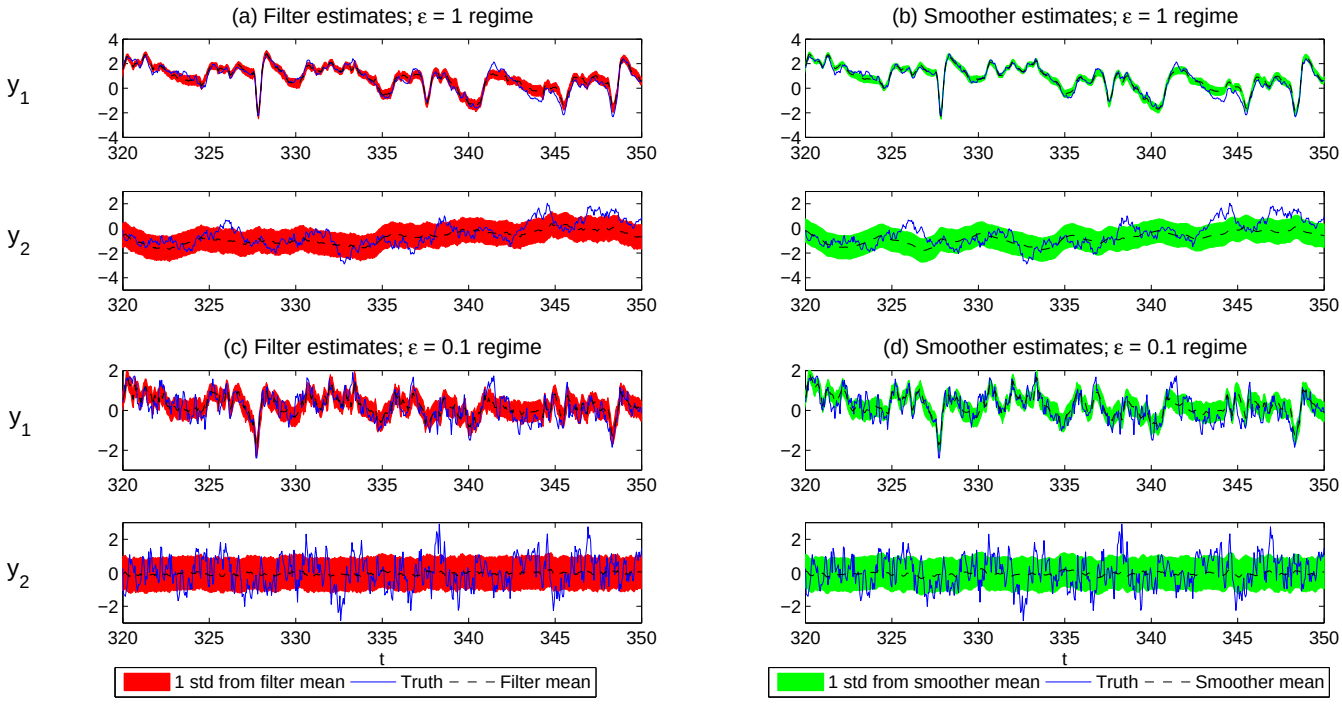


Figure 6: The four-dimensional stochastic climate model (33) with parameters (34). Comparison of the filter and smoother estimates. The true signals of $y_1$ and $y_2$ are shown in blue curves. The filter and smoother mean estimates are given by the black dashed curves. The uncertainty of the filter and smoother estimates, represented by the one standard deviation, are shown by the red and green shading areas, respectively.

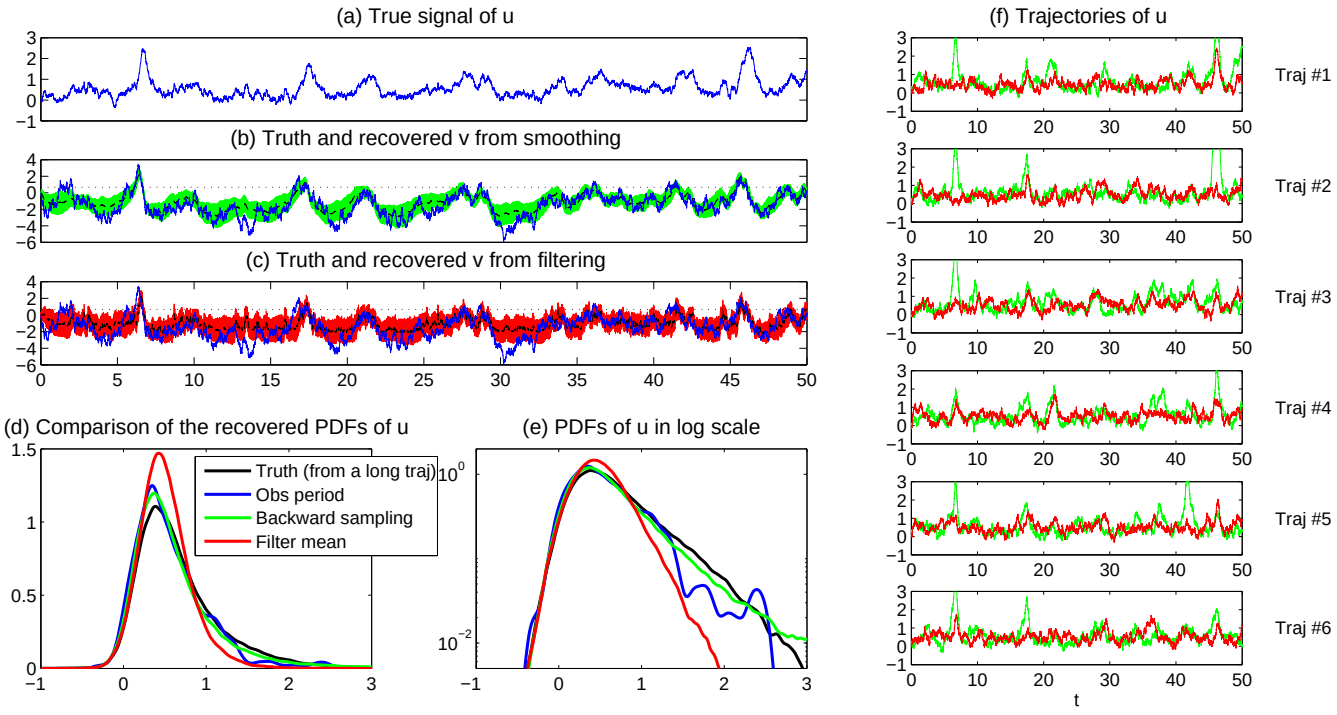


Figure 7: Dyad model (28) with parameters (29). Panel (a): a short observational period of $u$ with only 50 units. Panel (b): the true signal of $v$ (blue) and the smoother estimates, where the smoother mean is given by black dashed curve and the one standard deviation from the smoother mean is shown by the green shading area. Panel (c) is similar to Panel (b) but for the filter mean and filter uncertainty. Panel (d): the PDF of $u$. The black curve is formed by a long trajectory with 1000 time units from the perfect model free run. The blue one is formed by the observed time series shown in Panel (a). The green curve is formed by first applying the backward sampling to the variable $v$ and then plugging the sampled trajectory of $v$ into the $u$ process. This is repeated 20 times with each trajectory having length 50 units and thus the effective length of $u$ is 1000. The red curve is formed in a similar way but the associated $v$ trajectory is given by the filtered mean time series. Panel (e) shows the PDFs in the logarithm scale, which is a good representation of the fat tails. Panel (f) shows a few realizations of $u$, where the associated $v$ trajectories are either from the filter mean time series (red) or from the backward sampling strategy (green).

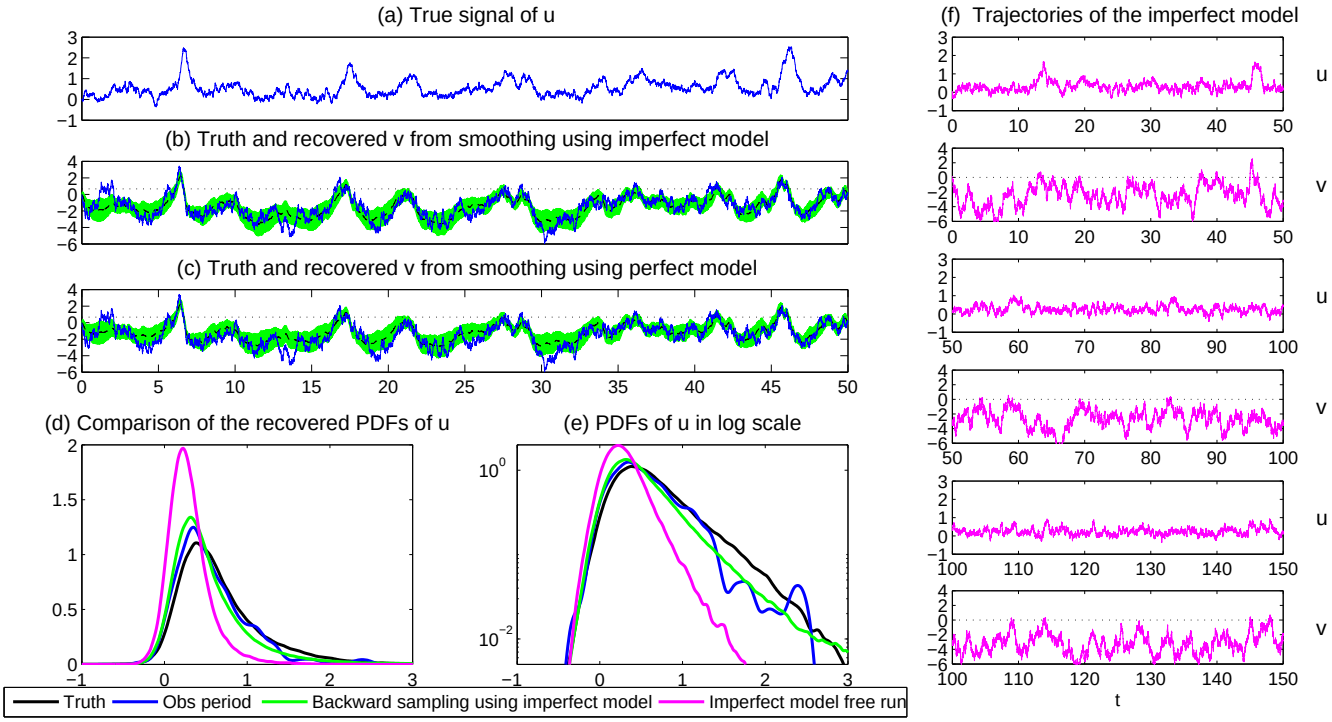


Figure 8: Similar to Figure 7 except that the model for recovering the statistics (36) is different from the perfect model (28). Panel (a) shows a short period of the observed variable $u$ from the perfect model run. Panels (b)–(c) show the smoother estimates using the imperfect and perfect models, respectively. Panels (d)–(e) show the PDFs of $u$ using different methods. Panel (f) shows the trajectories from a free run of the imperfect model (36).

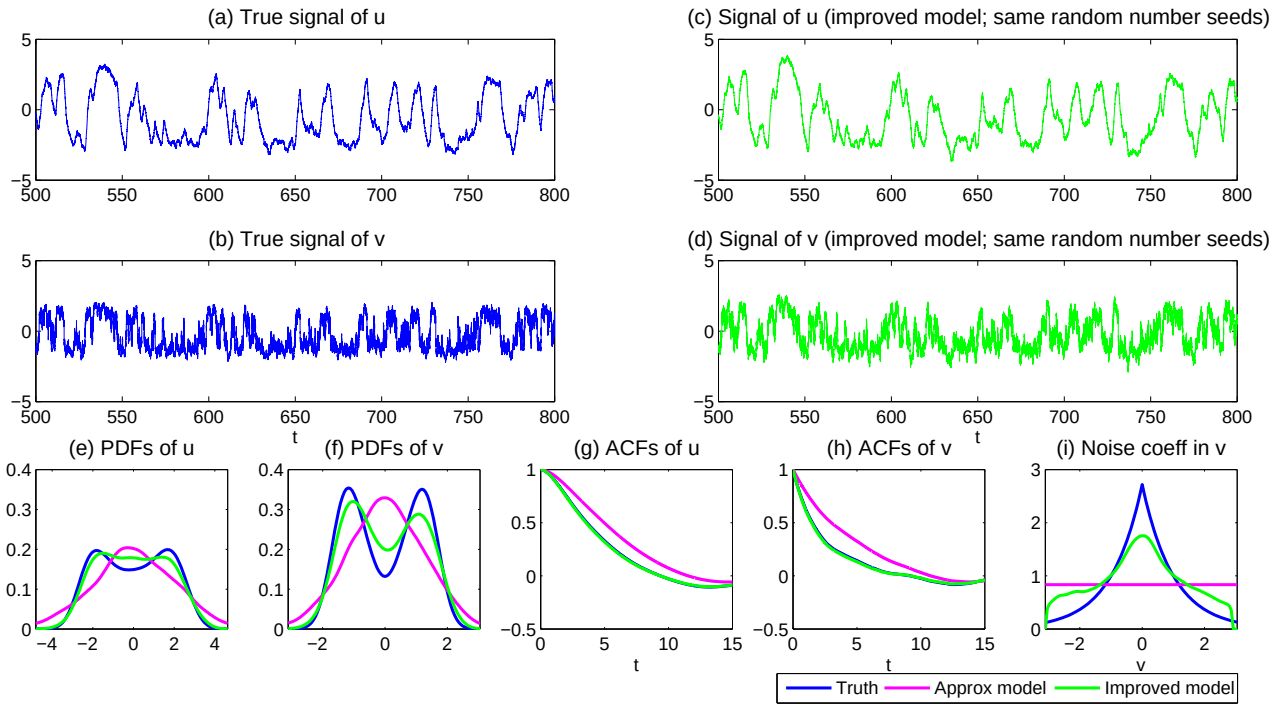


Figure 9: Improving the stochastic parameterizations. Panels (a)–(b): One realization of the true signals of $u$ and $v$ from (37). Panels (c)–(d): One realization of $u$ and $v$ from the improved model (using the same random number seeds as generating the true signal). Panels (e)–(h): comparison of the PDFs and ACFs. Panel (i): comparison of the noise coefficient in the $v$ process.