COMMUN. MATH. SCI.
Vol. 15, No. 2, pp. 413–450

DERIVATION AND ANALYSIS OF SIMPLIFIED FILTERS*

WONJUNG LEE† AND ANDREW STUART‡

Abstract. Filtering is concerned with the sequential estimation of the state, and uncertainties, of a Markovian system, given noisy observations. It is particularly difficult to achieve accurate filtering in complex dynamical systems, such as those arising in turbulence, in which effective low-dimensional representation of the desired probability distribution is challenging. Nonetheless recent advances have shown considerable success in filtering based on certain carefully chosen simplifications of the underlying system, which allow closed form filters. This leads to filtering algorithms with significant, but judiciously chosen, model error. The purpose of this article is to analyze the effectiveness of these simplified filters, and to suggest modifications of them which lead to improved filtering in certain time-scale regimes. We employ a Markov switching process for the true signal underlying the data, rather than working with a fully resolved DNS PDE model. Such Markov switching models haven been demonstrated to provide an excellent surrogate test-bed for the turbulent bursting phenomena which make filtering of complex physical models, such as those arising in atmospheric sciences, so challenging.

Keywords. Bayesian statistics, sequential data assimilation, filtering with model error.

AMS subject classifications. 60G35, 93E11, 94A12.

1. Introduction

1.1. Overview.

Filtering is concerned with the sequential updating of Markovian systems, given noisy, partial observations of the system state [29,30,37]. Due to the increasing prevalence of data in all areas of science and engineering, and due to the inherent complexity of physical models developed for the description of many phenomena arising in science and engineering, the need for accurate and speedy filters is paramount. However in its full form filtering requires the description of a time-evolving probability distribution on the system state, conditioned on data, which for many systems can be hard to represent in a computationally tractable way. This is a particular challenge for the complex physical models arising in areas such as atmospheric sciences [26], oceanography [2] and oil reservoir simulation [35]. However a recent body of work by Majda and coworkers [6–8,21,22,27,30,31,38] has demonstrated the possibility of using drastic simplifications of the models for complex turbulent phenomena in order to construct effective filters which are computationally tractable in real-time. The underlying philosophy of this work is to replace the true underlying Markovian model (often deterministic, but chaotic) with a simplified stochastic model which captures the key physical phenomena at the statistical level yet is amenable to closed form expressions for the purpose of filtering. It is possible to interpret this work as providing an important step towards the adoption of *physically informed machine learning*, going beyond traditional machine learning methodologies which often attempt to build models from the data alone [4,34]. The purpose of our work is to shed further light on this body of work, through analysis, through the derivation of new methods in the same spirit, and through careful numerical experiments.

In order to carry out this program we do not work with a full complex model of turbulence for our true signal, but rather work with a simple switching stochastic model

*Received: December 15, 2015; accepted (in revised form): June 11, 2016. Communicated by Arnulf Jentzen.

†Mathematics Institute and Centre for Predictive Modelling, School of Engineering, University of Warwick, U.K., and Department of Mathematics, City University of Hong Kong, Hong Kong (lee.wonjung@cityu.edu.hk).

‡Mathematics Institute, University of Warwick, U.K. (A.M.Stuart@warwick.ac.uk).

(SSM), a stochastic differential equation driven by a sign-alternating two-state Markov process [32,41]. The system is either forced or dissipated depending on the sign of the driving signal, and as a consequence admits intermittent bursting phenomena, similar to what is seen in real turbulent signals [5,14,42]. The use of this model as a simplified model for turbulent bursting, and demonstration of its effectiveness in this context, may be seen from the papers [16,17]. This SSM, then, is viewed as the "true" Markov model whose signals generate the data. Our objective is to find simplified models, amenable to filtering, which capture the essential features of the SSM. We now define the filtering problem and outline the simplified models that we study.

1.2. Background on filtering. Consider an $\mathbb{R}^d$-valued Markov process $x(t)$ where $t \geq 0$. The process is hidden and we only have access to y_n, $n \in \mathbb{N}$, which is a (partial) noisy observation of $x_n \equiv x(nT)$ for some $T > 0$. The dependency diagram

$$\begin{tikzcd}
x_1 \arrow[r] \arrow[d] & x_2 \arrow[r] \arrow[d] & x_3 \arrow[r] \arrow[d] & x_4 \arrow[r] \arrow[d] & {} \\
y_1 & y_2 & y_3 & y_4 &
\end{tikzcd}$$

is an intuitive graphical way to illustrate the problem setting in which the hidden states form a Markov chain emitting a discrete time series of observations [4,34]. For $Y_n := \{y_1, \cdots, y_n\}$ the key objective in probabilistic filtering is the sequential updating of $\mathbb{P}(x_n|Y_n)$ [1,12,23,28,30].

Let the underlying process be governed by

$$x_1 \sim \mathbb{P}(x_1) \quad \text{and} \quad x_n|x_{n-1} \sim \mathbb{P}(x_n|x_{n-1}), \quad n \geq 2$$

where $\sim$ means distributed according to, and let the marginal distribution of y_n be given by

$$y_n|x_n \sim \mathbb{P}(y_n|x_n)$$

which is termed the likelihood and in what follows is viewed as function of x_n, parameterized by y_n. To perform filtering, the standard approach is to alternate the uncertainty propagation $\mathbb{P}(x_{n-1}|Y_{n-1}) \mapsto \mathbb{P}(x_n|Y_{n-1})$, and the data acquisition $\mathbb{P}(x_n|Y_{n-1}) \mapsto \mathbb{P}(x_n|Y_n)$ in a sequential manner. The former step corresponds to probabilistic solution of the governing equation for $x(t)$, obtained from

$$\mathbb{P}(x_n|Y_{n-1}) = \int \mathbb{P}(x_{n-1}|Y_{n-1}) \mathbb{P}(x_n|x_{n-1}) dx_{n-1},$$

while the latter step is accomplished by Bayes' rule $\mathbb{P}(x_n|Y_n) \propto \mathbb{P}(x_n|Y_{n-1}) \mathbb{P}(y_n|x_n)$, which asserts that the posterior distribution is proportional to the product of the prior distribution and the likelihood.

When the probability distribution of interest is Gaussian, the filtering problem can explicitly be solved by the Kalman filter which describes the evolution of the mean and covariance [24,25]. The extended Kalman filter [15] employs the first two moments to approximately represent the target probability distribution in non-Gaussian scenarios. A family of weighted Dirac delta masses (ensemble Kalman filter [13] and particle filter [20]) and mixture of Gaussian kernels (Gaussian mixture filter [9,39,40]) are also used to approximate the filtering distribution.

1.3. The true model and filtering with model error. In this paper the true model x_n underlying the data will be found from discrete time sampling of the following switching stochastic model, or SSM for short:

$$\textbf{(SSM)} \qquad \begin{cases} du & = -\gamma u dt + \sigma_u dB_u \\ \gamma & \in \{\gamma_+, \gamma_-\} \end{cases} \tag{1.1}$$

where $\gamma(t)$ is a Markov process, alternately taking constant values of $\{\gamma_+ > 0, \gamma_- < 0\}$. The distribution functions of the random variables

$$\tau^{\gamma_+} = \inf\{t : \gamma(t) = \gamma_- | \gamma(0) = \gamma_+\}$$
$$\tau^{\gamma_-} = \inf\{t : \gamma(t) = \gamma_+ | \gamma(0) = \gamma_-\}$$

are given by

$$\mathbb{P}(\tau^{\gamma_+} < t) = 1 - e^{-\frac{\lambda_+}{\epsilon} t}$$
$$\mathbb{P}(\tau^{\gamma_-} < t) = 1 - e^{-\frac{\lambda_-}{\epsilon} t}$$

respectively. The positive parameter ϵ determines the transition rates, accounting for the time-scale separation between input signal γ and output response u. In case of small ϵ, there is rapid switching between γ_+ and γ_-. On the other hand, switching is a rare event when ϵ is large. In the general notation above we have $x = (u, \gamma)$.

For $x_n = (u_n, \gamma_n) = (u(nT), \gamma(nT))$ we assume the noisy observations are of the form

$$y_n = u_n + \eta_n, \qquad \eta_n \sim \mathcal{N}(0, R_n) \tag{1.2}$$

where $\{\eta_n\}_{n \geq 0}$ is an independent and identically distributed centred Gaussian. The filtering distribution $\mathbb{P}(x_n | Y_n)$, determined by Equations (1.1) and (1.2), does not allow for a closed-form representation. In the following, we address the problem through *filtering with model error*: that is, instead of a straightforward application to the genuine system, we replace the process x by a different Markov model which is more amenable to filtering explicitly than is the SSM. We tune the parameters of the new models to maximize their statistical resemblances with the SSM. It is important to note that in this paper, due to the low dimensionality of SSM, the introduction of reduced models used for filtering presumably does not lead to a significant saving of computational costs. However the aim is to understand the application of the methodology developed by Majda and coworkers, referred to above, which is targetted at situations where the true signal is very expensive to simulate, whilst the models used for filtering are orders of magnitude cheaper. Furthermore we investigate a new theory-based conceptual framework to illustrate this body of work, and to develop generalizations of it, working in a simple setting where the true signal of interest comes from the SSM.

Of course a typical real-world turbulent signal is governed by a highly complex dynamical system, not the SSM. Nonetheless it is our belief that our analysis of filtering with model error for the SSM sheds light on filtering with model error in the context of large scale geophysical fluid dynamics applications, because of the demonstrated ability of the SSM to represent turbulent bursting phenomena realistically.

There are four forms of filters with model error considered in this paper (acronyms explained later). The MSM and DSM are particularly relevant when ϵ is smaller, while the dMSM and dDSM are designed especially for larger ϵ. The MSM is found from the SSM by replacing the switching process γ by its mean (constant in time) value, giving rise to a process $\bar{u}$ instead of u. The DSM is found by replacing the switching process γ by the solution of an Ornstein-Uhlenbeck (OU) process, giving rise to a process $\widehat{u}$ instead of u. The dMSM is found by replacing u by a process with a constant γ in time, but choosing that constant randomly, according to carefully chosen weights. This leads to replacement of u by a process $\bar{u}'$. And finally the dDSM is found by replacing u by $\widehat{u}'$ in which γ is given by one of two OU processes for all time, but choosing the OU process randomly, according to carefully chosen weights. From now on, it will help to keep in mind that MSM and DSM are approximations of SSM for smaller ϵ, and dMSM and dDSM are approximations of SSM for larger ϵ.

1.4. Our contributions. Existing and extensive numerical studies naturally give rise to two fundamental questions about filtering with model error: (i) what are the precise conditions under which a given filter with model error is the best choice out of some class of filters; and (ii) how to choose the free parameters so as to maximize the consequent filtering accuracy. To address these questions we investigate the accuracy of the filters with model error via careful numerical experiments, and introduce a systematic approach for parameter determination. Specifically, our contributions in the present paper are as follows:

- in addition to studying the filters with model error MSM and DSM, introduced in [16, 17], we also introduce our own filters with model error: dMSM and dDSM;
- we build a Gaussian filter and a Gaussian mixture filter for SSM;
- we show the consistency of the reduced models in the extremely small (large) ϵ regime by proving limit theorems that connect the filter signal models MSM (dMSM) and DSM (dDSM) with the true signal model SSM;
- we use asymptotic analysis in the small (large) ϵ regime to obtain analytic formulae for the adaptive parameters of the simplified models MSM (dMSM) and DSM (dDSM);
- we employ optimization to solve minimization problem that yields suitable parameters for the simplifications when the scale-separation is not extreme but moderate or weak;
- we perform direct numerical simulations to show the accuracy and feasibility of the methods.

1.5. Organization of the paper. The paper is organized as follows. We precisely define the models used for filtering in Section 2. Our main results are in Section 3, where various tools, tuned to the relevant parameter regime for ϵ, are deployed to improve filtering accuracy. We perform numerical experiments in Section 4 and draw conclusions in Section 5. Lengthy calculations concerning the analysis of models are gathered in the appendices, in order to improve accessibility of the paper.

2. Filtering with model error: simplifications of SSM

Here we define four adaptive approximate models for SSM, based on the analysis of the qualitative behaviors of the switching process, and use them to build filters. Subsection 2.1 is concerned with the case when ϵ is small (scale-separation regime) and Subsection 2.2 is when ϵ is large (rare-event regime).

2.1. Scale-separation regime.

2.1.1. Mean stochastic model (MSM). In many multi-scale problems, the governing equation in which the driving signal is significantly faster is replaced by an equation with non-oscillatory coefficient found as a limit (usually in a weak sense) of scale-separation [3, 10, 36]. This work suggests that, when ϵ is sufficiently small, the mean stochastic model (MSM)

$$\textbf{(MSM)} \qquad \begin{cases} d\bar{u} = -\bar{\gamma}\bar{u}\,dt + \sigma_u dB_u \\ \bar{\gamma} \;\; = \text{const} \end{cases} \tag{2.1}$$

can be a good approximation of SSM. Using MSM for filtering we note that, provided $\bar{u}_0$ is Gaussian, all distributions $\mathbb{P}(\bar{u}_{n-1}|Y_{n-1}) \mapsto \mathbb{P}(\bar{u}_n|Y_{n-1}) \mapsto \mathbb{P}(\bar{u}_n|Y_n)$ are Gaussians and may be updated by the Kalman filter [24].

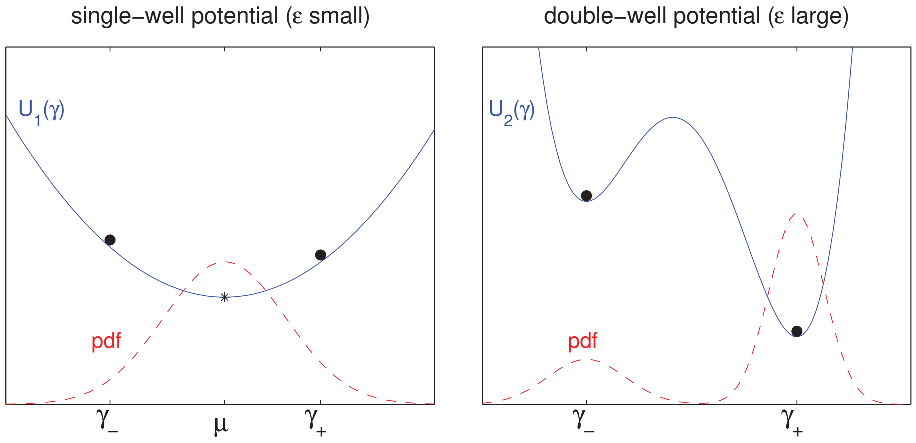

FIG. 2.1. *Regularized modeling of the qualitative behaviors of switching process.*

2.1.2. Diffusive stochastic model (DSM). The diffusive stochastic model (DSM) is given by

$$\textbf{(DSM)} \qquad \begin{cases} d\widehat{u} = -\widehat{\gamma}\widehat{u}\,dt + \sigma_u dB_u \\ d\widehat{\gamma} = -\frac{\nu}{\epsilon}(\widehat{\gamma} - \mu)\,dt + \frac{\sigma}{\sqrt{\epsilon}}dB_\gamma \end{cases}$$

Note that $\widehat{\gamma}$ is in an Ornstein-Uhlenbeck process: solution of the Langevin equation

$$d\widehat{\gamma} = -\frac{1}{\epsilon}\nabla U(\widehat{\gamma})dt + \frac{\sigma}{\sqrt{\epsilon}}dB_\gamma \tag{2.2}$$

with the potential

$$U(x) = U_1(x) := \frac{\nu}{2}(x-\mu)^2. \tag{2.3}$$

We aim to tune this process to match the response of the system, in the observed variable u. The reason for interest in this model is that, although exact filtering is not possible, it is possible to compute an approximate Gaussian filter, based on exact propagation of the first two moments. Indeed provided $(\widehat{u}(0), \widehat{\gamma}(0))$ is joint Gaussian, the mean and covariance of $(\widehat{u}(T), \widehat{\gamma}(T))$ are exactly solvable. Denoting $\widehat{\gamma}_n \equiv \widehat{\gamma}(nT)$, the resultant moment mapping can be used for uncertainty propagation: $\mathbb{P}(\widehat{u}_{n-1}, \widehat{\gamma}_{n-1}|Y_{n-1}) \mapsto \mathbb{P}(\widehat{u}_n, \widehat{\gamma}_n|Y_{n-1})$. Under this Gaussian approximation, the Kalman filter may be applied to obtain $\mathbb{P}(\widehat{u}_n,\widehat{\gamma}_n|Y_{n-1})\mapsto\mathbb{P}(\widehat{u}_n,\widehat{\gamma}_n|Y_n)$. The resulting filter is named the stochastic parametrization extended Kalman filter (SPEKF) in [18] where it was introduced. Finally, a proper marginalization at every step yields the object of interest: $\mathbb{P}(\widehat{u}_n|Y_n)$.

The DSM incorporates non-Gaussian features of SSM (and hence, potentially, of realistic models of turbulent dynamical systems) yet maintains the advantage of having low computational cost. One comment with respect to filtering of DSM is that SPEKF outperforms traditional methods due to at least two reasons. First, the new filter is characterized by long time stability, which is in contrast to the extended Kalman filter, where linearization error can accumulate to induce a filter instability, and second, that SPEKF makes use of judicious model errors which retain high filtering skill for complex turbulent signals, and as a consequence it is demonstrably more efficient, in terms of cost per unit error, than Monte Carlo based methods such as the ensemble Kalman filter and the particle filter. One might believe that SPEKF is only efficient in low dimensions, or in situations where sparse correlations may be exploited. However, the work of Majda [16–18, 31] demonstrates how sparse correlation structure can be exploited by allowing correlations which reflect aliasing and key nonlinear interactions arising in forced-dissipative systems with quadratic energy-conserving nonlinearities. For example, aliasing is usually viewed as a bad feature of numerical algorithms; however in the present context, judicious use of aliasing yields stochastic superresolution [6, 27, 30, 31].

2.2. Rare-event regime.

2.2.1. Dual-mode mean stochastic model (dMSM). When ϵ is large enough, transitions in γ are rare. To study this case, we build the following dual-mode mean stochastic model (dMSM):

$$\textbf{(dMSM)}\qquad \begin{cases} d\bar{u}' = -\bar{\gamma}'\bar{u}'\,dt+\sigma_u dB_u \\ \bar{\gamma}' \;= \begin{cases} \gamma_+ \text{ with probability } \bar{\rho}_+ \text{ for } t\geq 0 \\ \gamma_- \text{ with probability } \bar{\rho}_- = 1-\bar{\rho}_+ \text{ for } t\geq 0 \end{cases} \end{cases} \tag{2.4}$$

as the reduced modeling of SSM. This can be viewed as an example of the more general switching linear dynamical system model [19].

If the probability distribution of $\bar{u}'(0)$ is the sum of weighted Gaussian kernels, then note that

$$\mathbb{P}(\bar{u}'(T))=\mathbb{P}(\bar{\gamma}'=\gamma_+)\mathbb{P}(\bar{u}'(T)|\bar{\gamma}'=\gamma_+)+\mathbb{P}(\bar{\gamma}'=\gamma_-)\mathbb{P}(\bar{u}'(T)|\bar{\gamma}'=\gamma_-). \tag{2.5}$$

Under this assumption on $\bar{u}'(0)$, then, we may use the Gaussian mixture filter to obtain the exact filtering solution of dMSM. The procedure $\mathbb{P}(\bar{u}'_{n-1}|Y_{n-1})\mapsto\mathbb{P}(\bar{u}'_n|Y_{n-1})$ is performed using Equation (2.5), and a parallel application of the Kalman filter to each Gaussian kernel, along with updating of the weights of each kernel, completes the update $\mathbb{P}(\bar{u}'_n|Y_{n-1})\mapsto\mathbb{P}(\bar{u}'_n|Y_n)$.

In practice, the geometric growth in the number of kernels in the number of prediction steps prevents tractable exact inference as data is accumulated sequentially. One resolution that we adopt here is through the projection of the filtering solution onto the space of tractable distributions. Following the idea of assumed density filtering [33], a large mixture of Gaussians is replaced by a smaller mixture of Gaussians at regular time-intervals, while filtering progresses [11].

2.2.2. Dual-mode diffusive stochastic model (dDSM). As in the DSM we now try to use a diffusion process to model the switching process γ, in order to benefit from the possibility of propagating second moments exactly, as is done in the DSM. When ϵ is large, however, the process (2.2) with the single-well potential (2.3) is not suitable for mimicking rare transitions. We instead consider a double-well potential $U_2(x)$ for $U(x)$ (illustrated in the right-panel of Figure 2.1). In this scenario, the motion of $\widetilde{\gamma}$ is captured within either of the potential wells for significant time periods, but random perturbations allow it to effectively jump over the potential barrier and enter the parallel metastable state.

Based upon the quadratic expansions

$$U_2(x) \simeq \begin{cases} U_2(\mu_+) + \frac{\nu_+}{2}(x-\mu_+)^2 & \text{when } |x-\mu_+| \text{ is small} \\ U_2(\mu_-) + \frac{\nu_-}{2}(x-\mu_-)^2 & \text{when } |x-\mu_-| \text{ is small} \end{cases}$$

we build a new model

$$\textbf{(dDSM)} \qquad \begin{cases} d\widehat{u}' = -\widehat{\gamma}'\widehat{u}'\,dt + \sigma_u dB_u \\ d\widehat{\gamma}' = \begin{cases} -\frac{\nu_+}{\epsilon}(\widehat{\gamma}'-\mu_+)\,dt + \frac{\sigma_+}{\sqrt{\epsilon}}dB_\gamma \text{ with probability } \widehat{\rho}_+ \\ -\frac{\nu_-}{\epsilon}(\widehat{\gamma}'-\mu_-)\,dt + \frac{\sigma_-}{\sqrt{\epsilon}}dB_\gamma \text{ with probability } \widehat{\rho}_- = 1-\widehat{\rho}_+ \end{cases} \end{cases} \tag{2.6}$$

where the uncertainty is separately delivered by two independent sets of SDEs. Equation (2.6) is named by the dual-mode diffusive stochastic model (dDSM).

When $(\widehat{u}'(0),\widehat{\gamma}'(0))$ is a Gaussian mixture, utilizing the exact solvability of the first two moments of the propagated distributions (as for DSM), the probability of $(\widehat{u}'(T),\widehat{\gamma}'(T))$ can be approximated as Gaussian mixture with the number of kernels doubled, similarly to dMSM. As for dMSM we may perform a reduction of the number of mixtures to retain computational tractability. In this way, the approximate Gaussian mixture filter $\mathbb{P}(\widehat{u}'_{n-1},\widehat{\gamma}'_{n-1}|Y_{n-1}) \mapsto \mathbb{P}(\widehat{u}'_n,\widehat{\gamma}'_n|Y_{n-1}) \mapsto \mathbb{P}(\widehat{u}'_n,\widehat{\gamma}'_n|Y_n)$ is established.

3. Model validations

In this section, we proceed (i) to validate the proposed models, and (ii) to determine the adaptive parameters. We classify the ϵ parameter regime into the six regions; the scale-separation limit $\{\epsilon \to 0\}$, the sharp scale-separation regime $\{\epsilon \ll 1\}$, the imprecise scale-separation regime $\{\epsilon < 1\}$, the moderately rare-event regime $\{\epsilon > 1\}$, the extremely rare-event regime $\{\epsilon \gg 1\}$, the rare-event limit $\{\epsilon \to \infty\}$. Subsection 3.1 is devoted to the study of the case $\{\epsilon \to 0, \epsilon \to \infty\}$, and Subsection 3.2 to $\{\epsilon \ll 1, \epsilon \gg 1\}$, and Subsection 3.3 to $\{\epsilon < 1, \epsilon > 1\}$.

3.1. Convergence results.

Here we demonstrate the consistency of the simplified models by showing that $u_T, \widehat{u}_T \to \bar{u}_T$ (subscript notation will be abused and $u_T = u(T)$) as $\epsilon \to 0$ and that $u_T, \widehat{u}'_T \to \bar{u}'_T$ as $\epsilon \to \infty$ in senses elucidated in what follows. All proofs are deferred to the Appendix C.

3.1.1. Scale-separation limit. The main results here are the following theorem and corollary; the constants are defined in the developments following their statements.

THEOREM 3.1. *Assume that u_0, $\bar{u}_0$ and $\widehat{u}_0$ are identically distributed Gaussian random variables, and assume that (u_0,γ_0) and $(\widehat{u}_0,\widehat{\gamma}_0)$ are independent pairs of random variables. If $\bar{\gamma}$ and μ are equal to $\bar{\gamma}_\infty \equiv \frac{\lambda_-\gamma_+ + \lambda_+\gamma_-}{\lambda_- + \lambda_+}$, then, for any fixed $T > 0$, as $\epsilon \to 0$ the mean and variance of u_T and $\widehat{u}_T$ converge to those of $\bar{u}_T$.*

Proof. This follows from Lemmas 3.1, 3.2, 3.3, using the explicit calculations which are presented after the corollary below. □

Corollary 3.1. *Under the conditions in Theorem 3.1, the mean and variance in the Gaussian filters for $u_n|Y_n$ (defined in Appendix A.3.1) and $\widehat{u}_n|Y_n$, converge to those of $\bar{u}_n|Y_n$, for fixed $n>0$, as $\epsilon\to 0$.*

Proof. This follows from the data assimilation formula of the Kalman filter [1]. □

Let u^ϵ solve

$$du^\epsilon = -\gamma^\epsilon u^\epsilon dt + \sigma_u dB_u \tag{3.1}$$

for a random process γ^ϵ. For $\Gamma_t^\epsilon \equiv \int_0^t \gamma^\epsilon(s)ds$, the integral process of γ^ϵ, the variation-of-constants yields

$$u_T^\epsilon = e^{-\Gamma_T^\epsilon} u_0^\epsilon + \sigma_u \int_0^T e^{-(\Gamma_T^\epsilon - \Gamma_t^\epsilon)} dB_u(t). \tag{3.2}$$

Application of the Itô formula shows that the mean and covariance are given by

$$\begin{aligned} \langle u_T^\epsilon \rangle &= \left\langle e^{-\Gamma_T^\epsilon} u_0^\epsilon \right\rangle \\ \mathrm{Var}(u_T^\epsilon) &= \left\langle \left(e^{-\Gamma_T^\epsilon} u_0^\epsilon\right)^2 \right\rangle - \left\langle e^{-\Gamma_T^\epsilon} u_0^\epsilon \right\rangle^2 + \sigma_u^2 \int_0^T \left\langle e^{-2(\Gamma_T^\epsilon - \Gamma_t^\epsilon)} \right\rangle dt. \end{aligned} \tag{3.3}$$

Here and henceforth, $\langle \cdots \rangle$ denotes the statistical average. Equation (3.3) reveals that the moment generating function (MGF) of the integral process of γ^ϵ is particularly relevant to the first two moments propagation of u^ϵ governed by Equation (3.1).

Lemma 3.1. *Let $\bar{u}_t$ satisfy MSM (2.1). If*

$$\left\langle e^{\alpha(\Gamma_T^\epsilon - \Gamma_t^\epsilon)} \right\rangle \to e^{\alpha\bar{\gamma}(T-t)} \quad \textit{for } \alpha = -1, -2 \textit{ and } 0 \le t \le T \tag{3.4}$$

and if

$$\left\langle \left(e^{-\Gamma_T^\epsilon} u_0^\epsilon\right)^m \right\rangle \to \left\langle \left(e^{-\bar{\gamma}T} \bar{u}_0\right)^m \right\rangle \quad \textit{for } m = 1, 2 \tag{3.5}$$

as $\epsilon\to 0$, then the mean and variance of u_T^ϵ converge to those of $\bar{u}_T$. Further if

$$\left\langle \left(e^{-\Gamma_T^\epsilon} u_0^\epsilon - e^{-\bar{\gamma}T} \bar{u}_0\right)^2 \right\rangle \to 0 \tag{3.6}$$

as $\epsilon\to 0$, then u_T^ϵ converges to $\bar{u}_T$ in $L^2(\Omega;\mathbb{R})$. The convergence rates are determined by those associated with Equations (3.4), (3.5) and (3.6).

Let $\Gamma_t \equiv \int_0^t \gamma(s)ds$ and $\widehat{\Gamma}_t \equiv \int_0^t \widehat{\gamma}(s)ds$ be the integral processes associated to SSM and DSM respectively. Because $\bar{\Gamma}_t \equiv \int_0^t \bar{\gamma}(s)ds = \bar{\gamma}t$ ($\bar{\gamma}(s) = \bar{\gamma}$ is constant) in the case of the MSM, we expect $u_T, \widehat{u}_T \to \bar{u}_T$ provided both integral processes, Γ_t and $\widehat{\Gamma}_t$, behave like the probability distribution $\delta_{\bar{\gamma}t}$ in the small ϵ limit. It turns out this is indeed the case due to averaging. The next two lemmas highlight this behavior.

Lemma 3.2 (SSM). *Let $\gamma_t \sim \rho_+(t)\delta_{\gamma_+} + \rho_-(t)\delta_{\gamma_-}$. Then, for any fixed $t>0$, $\rho_\pm(t) \to \frac{\lambda_\mp}{\lambda_- + \lambda_+}$ as $\epsilon\to 0$. Let $\gamma_\infty \sim \frac{\lambda_-}{\lambda_- + \lambda_+}\delta_{\gamma_+} + \frac{\lambda_+}{\lambda_- + \lambda_+}\delta_{\gamma_-}$ and $\bar{\gamma}_\infty \equiv \langle \gamma_\infty \rangle = \frac{\lambda_-\gamma_+ + \lambda_+\gamma_-}{\lambda_- + \lambda_+}$. Then, for any fixed $T>t>0$, we have $\left\langle e^{\alpha(\Gamma_T - \Gamma_t)} \right\rangle \to e^{\alpha\bar{\gamma}_\infty(T-t)}$ as $\epsilon\to 0$.*

Lemma 3.3 (DSM). *Let $\widehat{\gamma}_\infty \sim \mathcal{N}(\mu, \sigma^2/2\nu)$. Then, for any fixed $t>0$, the mean and variance of $\widehat{\gamma}_t$ converge to those of $\widehat{\gamma}_\infty$ as $\epsilon\to 0$. Furthermore, we have, for any fixed $T>t>0$, $\left\langle e^{\alpha(\widehat{\Gamma}_T - \widehat{\Gamma}_t)} \right\rangle \to e^{\alpha\mu(T-t)}$ as $\epsilon\to 0$.*

3.1.2. Rare-event limit. The main results in this regime are the following theorem and corollary.

Theorem 3.2. *Assume that u_0, $\bar{u}'_0$ and $\widehat{u}'_0$ are identically distributed Gaussian random variables, and assume that (u_0,γ_0) and $(\widehat{u}'_0,\widehat{\gamma}'_0)$ are independent pairs of random variables. Then, for any fixed $T>0$, the mean and variance of $\mathbb{P}(u_T|\gamma_0=\gamma_\pm)$, $\mathbb{P}(\widehat{u}'_T|\widehat{\gamma}'_0=\gamma_\pm)$ converges to those of $\mathbb{P}(\bar{u}'_T|\bar{\gamma}'=\gamma_\pm)$ as $\epsilon\to\infty$.*

Furthermore, let $\gamma_0\triangleq\gamma_\infty$, and suppose $\bar{\gamma}'$ and $\widehat{\gamma}'_0$ are identically distributed with γ_∞. Then the weight, mean and variance of components in the Gaussian mixture approximation for u_T, $\widehat{u}'_T$ converge to those of $\bar{u}'_T$ as $\epsilon\to\infty$.

Proof. This follows from Lemmas 3.4, 3.5, 3.6 below. □

Corollary 3.2. *Under the conditions in Theorem 3.2, the weight, mean and variance of mixture components in the Gaussian mixture filters for $u_n|Y_n$ (defined in Appendix A.3.2) and $\widehat{u}'_n|Y_n$, converge to those of $\bar{u}'_n|Y_n$, for fixed $n>0$, as $\epsilon\to\infty$.*

Proof. This follows from parallel application of the Kalman filter update to the mixture components. □

Lemma 3.4. *Let $\bar{u}'_t$ solve dDSM (2.4). If, for each fixed $T>t>0$,*

$$\left\langle e^{\alpha(\Gamma^\epsilon_T-\Gamma^\epsilon_t)}|\gamma^\epsilon_0=\gamma_\pm\right\rangle\to e^{\alpha\gamma_\pm(T-t)}\quad \textit{for } \alpha=-1,-2 \textit{ and } 0\le t\le T \tag{3.7}$$

and if

$$\left\langle\left(e^{-\Gamma^\epsilon_T}u^\epsilon_0\right)^m|\gamma^\epsilon_0=\gamma_\pm\right\rangle\to\left\langle\left(e^{-\gamma_\pm T}\bar{u}_0\right)^m\right\rangle\quad \textit{for } m=1,2 \tag{3.8}$$

as $\epsilon\to\infty$, then the mean and variance of $u^\epsilon_T|\gamma^\epsilon_0=\gamma_\pm$ converge to those of $\bar{u}_T|\bar{\gamma}'=\gamma_\pm$. The convergence rates are determined by those associated with Equations (3.7), (3.8).

Furthermore, if $\gamma^\epsilon_0\triangleq\bar{\gamma}'$, then the weight, mean and variance of components in the Gaussian mixture approximation for u^ϵ_T converge to those of $\bar{u}_T$ from $\mathbb{P}(u^\epsilon_T)=\mathbb{P}(\gamma^\epsilon_0=\gamma_+)\mathbb{P}(u^\epsilon_T|\gamma^\epsilon_0=\gamma_+)+\mathbb{P}(\gamma^\epsilon_0=\gamma_-)\mathbb{P}(u^\epsilon_T|\gamma^\epsilon_0=\gamma_-)$.

To ensure the convergences of SSM and dDSM to dMSM, as ϵ grows, both Γ_t and $\widehat{\Gamma}'_t\equiv\int_0^t\widehat{\gamma}'(s)ds$ need to converge to $\bar{\Gamma}'_t\equiv\int_0^t\bar{\gamma}'(s)ds\sim\bar{\rho}_+\delta_{\gamma_+t}+\bar{\rho}_-\delta_{\gamma_-t}$.

Lemma 3.5 (SSM). *For fixed $T>t>0$ $\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}|\gamma_0=\gamma_\pm\right\rangle\to e^{\alpha\gamma_\pm(T-t)}$ as $\epsilon\to\infty$.*

Lemma 3.6 (dDSM). *For fixed $T>t>0$ $\left\langle e^{\alpha(\widehat{\Gamma}'_T-\widehat{\Gamma}'_t)}|\widehat{\gamma}'_0=\gamma_\pm\right\rangle\to e^{\alpha\gamma_\pm(T-t)}$ as $\epsilon\to\infty$.*

3.2. Asymptotic matching. The convergence results in the preceding subsection demonstrate that the filtering performances of the approximate filters, and the exact filter, would be similar to one another in that $\bar{\gamma}=\bar{\gamma}_\infty$, $\mu=\bar{\gamma}_\infty$ (when $\epsilon\ll1$) and $\bar{\gamma}'\triangleq\gamma_\infty$, $\widehat{\rho}_\pm\propto\lambda_\mp$, $\mu_\pm=\gamma_\pm$ (when $\epsilon\gg1$). The former result relates to the robustness of the DSM filter inherited from the adaptive parameters $\{\mu,\sigma\}$, demonstrated here when ϵ is small, and demonstrated through extensive numerical simulations in [16, 17].

However, when ϵ deviates considerably from the two extreme values ($\epsilon=0$ and $\epsilon=\infty$), the choice of associated parameters in the filtering models is indeed one critical factor for a successful filtering with model error. The current and next subsections concern the determination of $\Theta\equiv\{\mu,\nu,\sigma\}$ for DSM, and $\Theta'\equiv\{\widehat{\rho}_\pm,\mu_\pm,\nu_\pm,\sigma_\pm\}$ for dDSM. Unlike earlier works in this area where these associated parameters are chosen from a number of parallel direct numerical simulations comparing the original dynamics and its simplifications, our approach will specify the parameters in a systematic analysis-based manner.

3.2.1. Sharp scale-separation regime.

In this parameter regime, because DSM is associated to a nonlinear approximate Kalman filter, we attempt to equate the first- and second-order statistics of SSM and DSM,

$$\begin{cases} \langle u_T \rangle & = \langle \widehat{u}_T \rangle \\ \operatorname{Var}(u_T) & = \operatorname{Var}(\widehat{u}_T) \end{cases} \tag{3.9}$$

for high accuracy. It is worth noticing that, in view of Equation (3.3), if the MGFs agree with one another, that is if

$$\begin{cases} \left\langle e^{\alpha \Gamma_T} \right\rangle = \left\langle e^{\alpha \widehat{\Gamma}_T} \right\rangle & \text{for } \alpha = -1, -2 & \text{(3.10a)} \\ \left\langle e^{\alpha(\Gamma_T - \Gamma_t)} \right\rangle = \left\langle e^{\alpha(\widehat{\Gamma}_T - \widehat{\Gamma}_t)} \right\rangle & \text{for } \alpha = -2 \text{ and } 0 \le t \le T & \text{(3.10b)} \end{cases}$$

and if (u_0, γ_0) and $(\widehat{u}_0, \widehat{\gamma}_0)$ are uncorrelated, and if $u_0 \triangleq \widehat{u}_0$, then Equation (3.9) holds. Motivated by convergence to the common limit, as demonstrated above, we here strive to asymptotically satisfy Equation (3.10) when $\epsilon \ll 1$.

To that end, we derive the approximation

$$\begin{aligned} \left\langle e^{\alpha(\Gamma_T - \Gamma_t)} \right\rangle \simeq \exp \Bigg(& \alpha \bar{\gamma}_\infty (T - t) + \alpha^2 \frac{3}{8} \frac{(\gamma_- - \gamma_+)^2 (\lambda_-^2 + \lambda_+^2)}{\lambda_- \lambda_+ (\lambda_+ + \lambda_-)} (T - t) \epsilon \\ & + \alpha \left(\mathbb{P}(\gamma_0 = \gamma_+) \frac{(\gamma_+ - \gamma_-)}{4\lambda_+} + \mathbb{P}(\gamma_0 = \gamma_-) \frac{(\gamma_- - \gamma_+)}{4\lambda_-} \right) \epsilon + \mathcal{O}(\epsilon^2) \Bigg) \quad \epsilon < T - t \end{aligned} \tag{3.11}$$

in the Appendix A.2.1. We also derive the approximations

$$\left\langle e^{\alpha \widehat{\Gamma}_T} \right\rangle = \exp \left(\alpha \left(\mu T + \langle \widehat{\gamma}_0 - \mu \rangle \frac{\epsilon}{\nu} \right) + \alpha^2 \frac{\sigma^2}{2\nu^2} T \epsilon + \mathcal{O}(\epsilon^2) \right) \quad \epsilon < T \tag{3.12a}$$

$$\left\langle e^{\alpha(\widehat{\Gamma}_T - \widehat{\Gamma}_t)} \right\rangle = \exp \left(\alpha \mu (T - t) + \alpha^2 \frac{\sigma^2}{2\nu^2} (T - t) \epsilon + \mathcal{O}(\epsilon^2) \right) \quad \epsilon < t \tag{3.12b}$$

in the Appendix B.3.1. Importantly, the exponents of MGFs are of the second-order with respect to αT up to $\mathcal{O}(\epsilon)$, indicating that both Γ_T and $\widehat{\Gamma}_T$ are statistically closer to Gaussian in this parameter regime.

From a comparison between the approximations (3.11) and (3.12), we realize Equation (3.10) is asymptotically met provided $\bar{\gamma}_\infty = \mu$ and

$$\frac{3}{8} \frac{(\gamma_- - \gamma_+)^2 (\lambda_-^2 + \lambda_+^2)}{\lambda_- \lambda_+ (\lambda_+ + \lambda_-)} = \frac{\sigma^2}{2\nu^2} \tag{3.13}$$

and

$$\mathbb{P}(\gamma_0 = \gamma_+) \frac{(\gamma_+ - \gamma_-)}{4\lambda_+} + \mathbb{P}(\gamma_0 = \gamma_-) \frac{(\gamma_- - \gamma_+)}{4\lambda_-} = \frac{\langle \widehat{\gamma}_0 - \mu \rangle}{\nu}. \tag{3.14}$$

Equations (3.13), (3.14) can be solved to determine a unique set of $\{\nu, \sigma^2\}$ but might result in $\nu < 0$ which is unphysical. In order to avoid this possibility, we impose the equivalence between variances of stationary processes γ_∞ and $\widehat{\gamma}_\infty$

$$\frac{\lambda_+ \lambda_- (\gamma_+ - \gamma_-)^2}{(\lambda_- + \lambda_+)^2} = \frac{\sigma^2}{2\nu} \tag{3.15}$$

instead of Equation (3.14). From Equations (3.13) and (3.15), we obtain

$$\Theta_{\text{naive}} \equiv \begin{cases} \mu & = \bar{\gamma}_\infty \\ \nu & = \frac{8}{3}\frac{\lambda_-^2\lambda_+^2}{(\lambda_-+\lambda_+)(\lambda_+^2+\lambda_-^2)} \\ \sigma^2 & = \frac{16}{3}\frac{\lambda_-^3\lambda_+^3(\gamma_- - \gamma_+)^2}{(\lambda_-+\lambda_+)^3(\lambda_+^2+\lambda_-^2)} \end{cases} \quad \text{when} \quad \epsilon \ll 1 \tag{3.16}$$

which we term the naive set of DSM parameters, valid when $\epsilon \ll 1$.

3.2.2. Extremely rare-event regime. Using a similar argument to that employed in the case of DSM, we set $\rho_\pm = \widehat{\rho}_\pm = \frac{\lambda_\mp}{\lambda_-+\lambda_+}$ and attempt to satisfy

$$\begin{cases} \langle u_T | \gamma_0 = \gamma_\pm \rangle & = \langle \widehat{u}_T | \widehat{\gamma}'_0 = \gamma_\pm \rangle \\ \mathrm{Var}(u_T | \gamma_0 = \gamma_\pm) & = \mathrm{Var}(\widehat{u}_T | \widehat{\gamma}'_0 = \gamma_\pm) \end{cases},$$

hence

$$\begin{cases} \left\langle e^{\alpha \Gamma_T} | \gamma_0 = \gamma_\pm \right\rangle = \left\langle e^{\alpha \widehat{\Gamma}'_T} | \widehat{\gamma}'_0 = \gamma_\pm \right\rangle & \text{for } \alpha = -1, -2 \qquad (3.17\text{a}) \\ \left\langle e^{\alpha(\Gamma_T - \Gamma_t)} | \gamma_0 = \gamma_\pm \right\rangle = \left\langle e^{\alpha(\widehat{\Gamma}'_T - \widehat{\Gamma}'_t)} | \widehat{\gamma}'_0 = \gamma_\pm \right\rangle & \text{for } \alpha = -2 \text{ and } 0 \le t \le T \quad (3.17\text{b}) \end{cases}$$

for dDSM.

In the case $\epsilon \gg 1$, we derive

$$\left\langle e^{\alpha \Gamma_T} | \gamma_0 = \gamma_\pm \right\rangle \simeq \exp\left(\alpha \gamma_\pm T - \frac{1}{\epsilon} \lambda_\pm T \right) \tag{3.18}$$

in the Appendix A.2.2 and

$$\left\langle e^{\alpha \widehat{\Gamma}'_T} | \widehat{\gamma}'_0 = \gamma_\pm \right\rangle \simeq \exp\left(\alpha \left(\gamma_\pm T - \frac{1}{2\epsilon}(\gamma_\pm - \mu_\pm)\nu_\pm T^2 \right) + \frac{\alpha^2}{2}\frac{(\sigma_\pm)^2}{3\epsilon} T^3 \right) \tag{3.19}$$

in the Appendix B.3.2. Note the exponents in the approximations (3.18) and (3.19) are of different forms, indicating both Γ_T and $\widehat{\Gamma}'_T$ are distant from Gaussian in this parameter regime.

Differently from the case of DSM, we here manage to asymptotically satisfy Equation (3.17a) alone, yielding

$$\Theta'_{\text{naive}} \equiv \begin{cases} \widehat{\rho}_\pm = \frac{\lambda_\mp}{\lambda_-+\lambda_+} \\ \mu_\pm = 2T\frac{\lambda_+\lambda_-(\gamma_+ - \gamma_-)^2}{(\lambda_-+\lambda_+)^2} + \gamma_\pm \\ \nu_\pm = \frac{3\lambda_\pm}{2T^2}\frac{(\lambda_-+\lambda_+)^2}{\lambda_+\lambda_-(\gamma_+ - \gamma_-)^2} \\ (\sigma_\pm)^2 = \frac{3\lambda_\pm}{T^2} \end{cases} \quad \text{when} \quad \epsilon \gg 1 \tag{3.20}$$

which we term the naive set of dDSM parameters, valid when $\epsilon \gg 1$. Unlike Equation (3.16), due to the dependence on T, the set of parameters (3.20) is valid only for fixed-time prediction. The Gaussian mixture from dDSM with Θ'_{naive} leads to accurate mean approximations but the accuracy of the variance approximation is not guaranteed in view of Equation (3.3) where integration over $[0,T]$ is involved.

3.3. Minimizing sum-of-squares. In the parameter regime $\epsilon \sim O(1)$, due to the absence of small or large parameters allowing for asymptotic analysis, we invoke a minimization principle to determine the set of parameters Θ and Θ'.

3.3.1. Imprecise scale-separation regime. When $\epsilon < 1$, we aim to find Θ which minimizes the sum-of-squares

$$J(\epsilon) \equiv \kappa \left| \langle u_T \rangle - \langle \widehat{u}_T \rangle \right|^2 + \left| \mathrm{Var}(u_T) - \mathrm{Var}(\widehat{u}_T) \right|^2 \tag{3.21}$$

where $\kappa \geq 0$ is introduced to ensure appropriate scaling of the two terms in the objective function. To be more precise, given $(\widehat{u}_0, \widehat{\gamma}_0)$, Equation (3.21) is an algebraic relation in terms of Θ once we impose $u_0 := \widehat{u}_0$ and $\gamma_0 := \gamma_\infty$ (see Appendices A and B). Note that a minimizer of $J(\epsilon)$ comes as close as possible to fulfilling Equation (3.9). It is worth mentioning that, differently from the MFG matching (3.10) for which $(\widehat{u}_0, \widehat{\gamma}_0)$ should be at most weakly correlated for the approach to be valid, the minimization methodology can be used irrespective of their potentially strong correlation.

We identify a (local) minimizer by taking Θ_{naive} as an initial starting point, and applying an optimizer such as gradient descent. This minimization can be performed using continuation in ϵ, starting from $\epsilon \ll 1$ where the initial guess will be accurate. Because the solution of this minimization is computed at each assimilation time step we name it *dynamic calibration* and denote the resulting time-dependent parameters by Θ_{dynamic}. Of course the key issue in sequential filtering that we are addressing is to maintain an accurate description of the evolving probability distribution with reasonable computational cost. In this context it is impractical to compute Θ_{dynamic} at every observation time. In practice, one can take a time average of a range of dynamic calibrations. We refer to this as static calibration and denote the resulting parameter by Θ_{static}.

3.3.2. Moderately rare-event regime. As for the extremely rare-event regime, we carry out the same procedure for each stable and unstable Gaussian kernel. As for the imprecise scale-separation regime we also minimize an expression analogous to Equation (3.21) in which the conditioned mean and covariance are used instead. We first find Θ'_{dynamic} from Θ'_{naive}, and next find Θ'_{static} from Θ'_{dynamic}. Unlike the method based on matching MGF asymptotics, where the potential inaccuracy of variance approximations are present, this method simultaneously accounts for accuracy in both the mean and covariance approximations.

4. Numerical simulations

Having obtained three different versions of adaptive parameters (naive set, static calibration, dynamic calibration) for DSM and dDSM, we here investigate the filtering performances of the suggested models using numerical simulations.

Very importantly, one distinguished advantage of the framework we are currently adopting lies in the analytic tractability of the state space model. In Appendix A.1, we derive the closed form solution (when $\lambda_+ = \lambda_-$) and the series solution (when $\lambda_+ \neq \lambda_-$) for MGFs of the SSM integral process. In Appendix A.3, we use them to design the Gaussian filter (suitable when ϵ is small) and the Gaussian sum filter (suitable when ϵ is large) for SSM, with the assumption that the filtered variables of u and γ are independent at every observation time. Those results from the direct filtering of SSM are then to be used as the reference solutions in subsequent experiments. We emphasize that the presence of these reference probability distributions enables very careful examination of filter accuracy in our numerical experiments, beyond measuring the distance between a realization of the truth signal and the mean of an approximate filtering solution and beyond what is seen in most other works concerning the computational evaluations of filters; this in turn gives further depth to our demonstrations.

In all our experiments, we use the following parameter values to specify the SSM truth model: $\sigma_u = 0.1549$, $\gamma_+ = 2.27$, $\gamma_- = -0.04$, $\lambda_+ = 1$ and $\lambda_- = 2$ (these choices follow those in [16]). Fixing inter-observation time $T = 1$, we study the cases of $\epsilon = 10^{-1}$, 10^0, 10^1, 10^2. Each one is selected as representative of the parameter regimes: sharp scale-separation, imprecise scale-separation, moderately rare-event, extremely rare-event, in the order given. Since $\mathbb{E}(\tau_k) = 1/r$ for $\tau_k \sim \exp(r)$, the reciprocal of ϵ equals the average number of transitions from the stable mode ($\gamma = \gamma_+$) to the unstable mode ($\gamma = \gamma_-$) on the unit time interval. As λ_- is twice λ_+ in this example, the average time spent in the stable mode is twice that spent in the unstable mode.

We take the initial condition of SSM according to $u_0 \triangleq \mathcal{N}(0.1, 0.0016)$ and $\gamma_0 \triangleq \gamma_\infty$, independently from one another. For MSM (dMSM), we take $\bar{u}_0(\bar{u}'_0) \triangleq u_0$. We also take $\bar{\gamma} = \bar{\gamma}_\infty (= 1.5)$ and $\bar{\gamma}' \triangleq \gamma_\infty$. For DSM (dDSM), we take the independent Gaussian $(\widehat{u}_0, \widehat{\gamma}_0)$ (or $(\widehat{u}'_0, \widehat{\gamma}'_0)$) where $\widehat{u}_0 \triangleq u_0$ and $\widehat{\gamma}_0 \triangleq \mathcal{N}(1.2\bar{\gamma}_\infty, \mathrm{Var}(\gamma_\infty))$. We set $\widehat{\rho}^{\pm}(0) = \bar{\rho}_\pm$. For the observational process in Equation (1.2), we use $R_n = 0.25E$ where $E \equiv \sigma_u^2/(2\bar{\gamma})$ (in this case the variance of $\bar{u}_n|Y_n$ is independent of n).

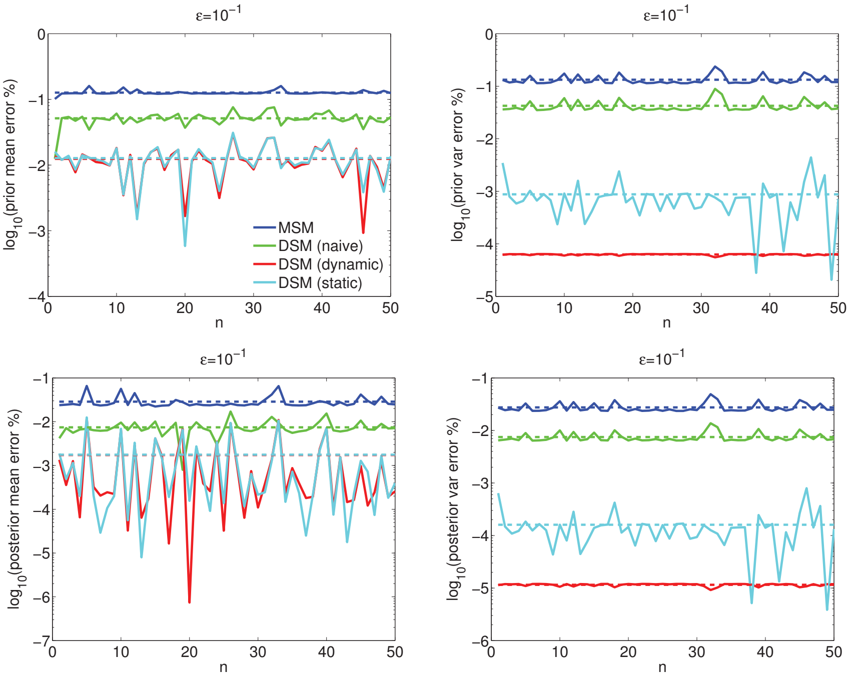

FIG. 4.1. *The relative errors of the mean and variance approximations of the prior $u_n|Y_{n-1}$ (top) and posterior $u_n|Y_n$ (bottom) distributions when $\epsilon = 10^{-1}$. The dashed lines denote the time averages over $0 \le n \le 50$.*

4.1. Performances of simplified filters

4.1.1. Sharp scale-separation regime. We first study the case of $\epsilon = 10^{-1}$. For the implementation of DSM with dynamic calibration, along with Θ_{naive} as a starting point, a local minimizer $\Theta_{\mathrm{dynamic}}$ of Equation (3.21) is solved at every observation time. The choice of κ in $J(\epsilon)$ plays a substantial role in this problem. Here and hereafter, the value of κ is set to zero for simplicity and consistency of presentations; this allows the prior mean from dynamic and static calibrations to be inaccurate but, in filtering,

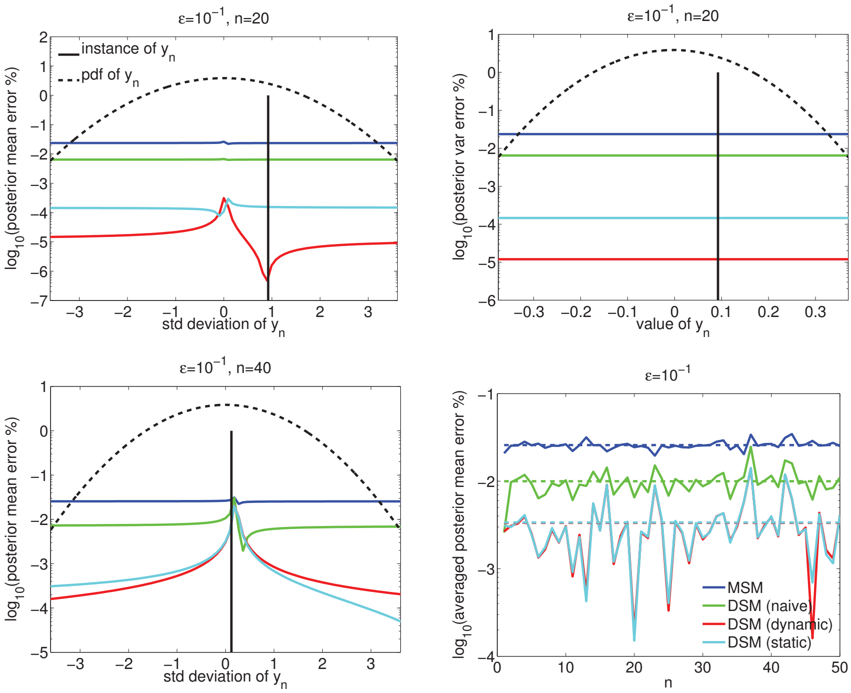

Fig. 4.2. *The relative errors of the approximations of the posterior $u_n|Y_n$ distributions that depend on the realization of $y_n = u_n + \eta_n$ (top and bottom-left), and their statistical averages with respect to the law of y_n (bottom-right) when $\epsilon = 10^{-1}$. In Gaussian filters, the accuracy of the posterior variance does not depend on the instance of y_n (top-right).*

the posterior is the main object of interest. The time average of these parameters for $1 \le n \le 50$ is taken as Θ_{static}.

In addition to DSM filters, we apply Gaussian filters for MSM and SSM. For the latter, due to distinct $\lambda_{\pm}$, we need to truncate the series solution of the MGF. Hereafter, the first 30 terms of the series solution will be kept as this ensures accuracy by virtue of the fact that $\mathbb{P}(N_T > 30) < 10^{-5}$.

In Figure 4.1, we depict the relative errors of the prior and posterior approximations in terms of mean and variance. We see that the approximations of DSM with the parameters tuned by our methods are significantly more accurate than the MSM approximation. As expected, the overall errors of the mean and variance relative to those from SSM filtering solutions are given in the order : DSM (dynamic calibration) $\lesssim$ DSM (static calibration) $<$ DSM (naive set) $<$ MSM. Admittedly, this result is merely for a single realization of the observation process. However we show that the result is indeed robust with respect to the chosen observational data set in the following manner.

At each observation time step, the posterior distributions of the approximate models are determined by the instance of observation, which is drawn from a Gaussian. In Figure 4.2, we depict the dependence of the corresponding filter accuracy on y_n for $n = 20$ and $n = 40$. It is observed that, for most values of y_n, Gaussian filters for DSM with dynamic and static calibrations significantly outperform MSM, leading to highly accurate posterior approximations. In Figure 4.2, we also depict the statistical average of the posterior error with respect to y_n for each n. There, one can see the ordering of the accuracies is exactly the same as in the single realization experiment.

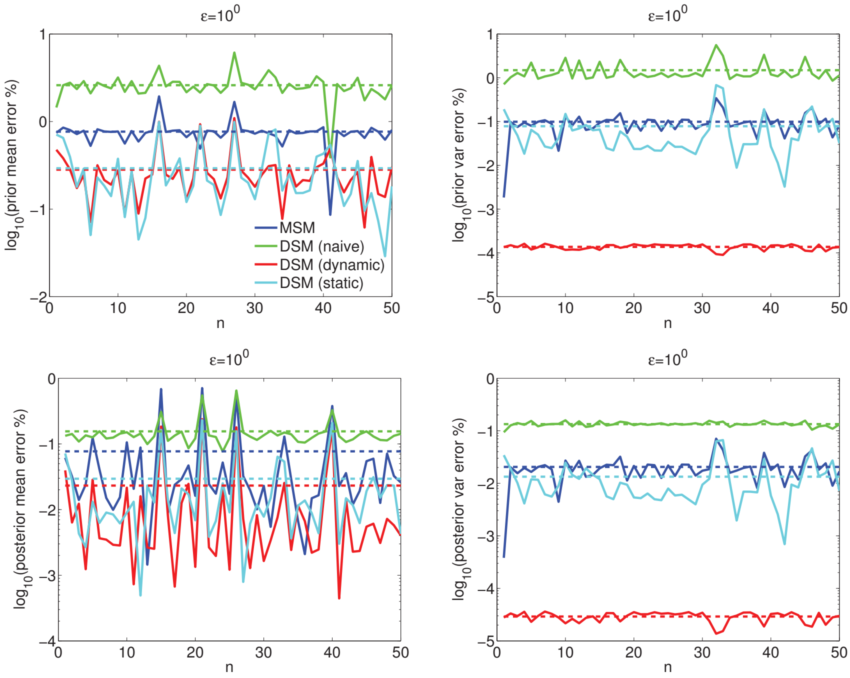

FIG. 4.3. *The relative errors of the mean and variance approximations of the prior $u_n|Y_{n-1}$ (top) and posterior $u_n|Y_n$ (bottom) distributions when $\epsilon = 10^0$.*

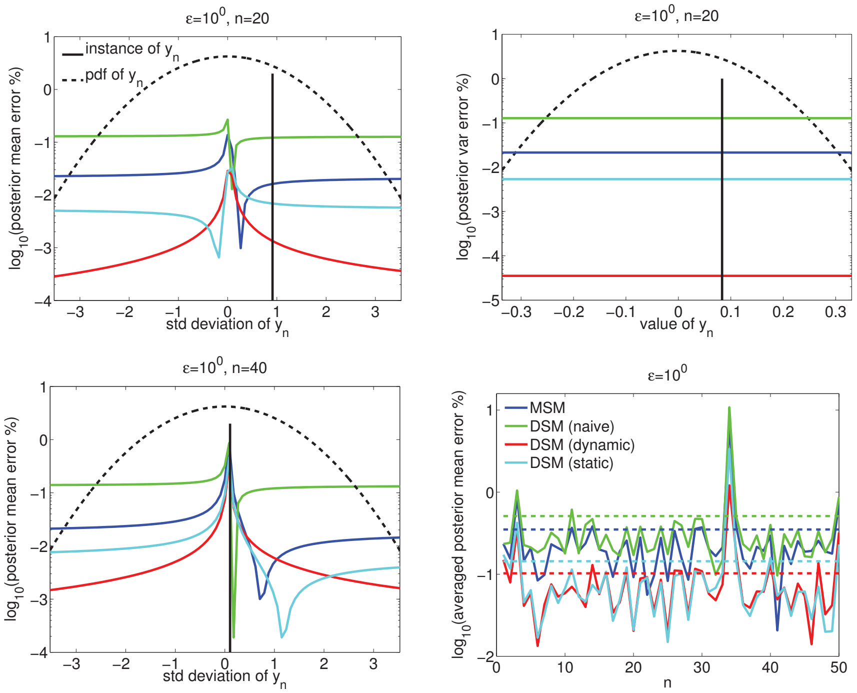

FIG. 4.4. *The relative errors of the approximations of the posterior $u_n|Y_n$ distributions that depend on the realization of $y_n = u_n + \eta_n$, and their statistical averages with respect to the law of y_n when $\epsilon = 10^0$.*

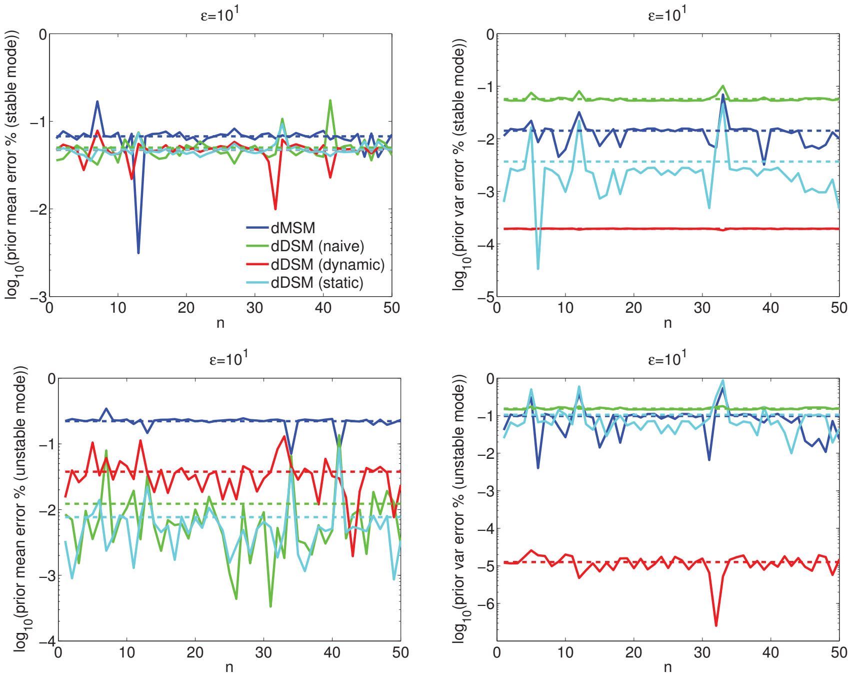

FIG. 4.5. *The relative errors of the mean and variance approximations of each Gaussian kernels of the prior distributions when* $\epsilon = 10^1$.

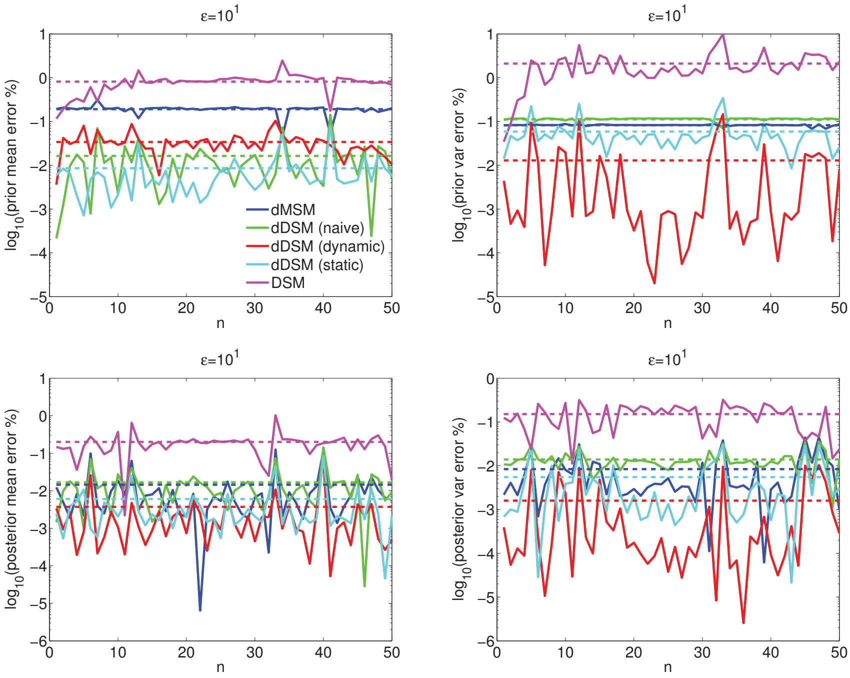

FIG. 4.6. *The relative errors of the mean and variance approximations of the prior* $u_n|Y_{n-1}$ *(top) and posterior* $u_n|Y_n$ *(bottom) distributions when* $\epsilon = 10^1$.

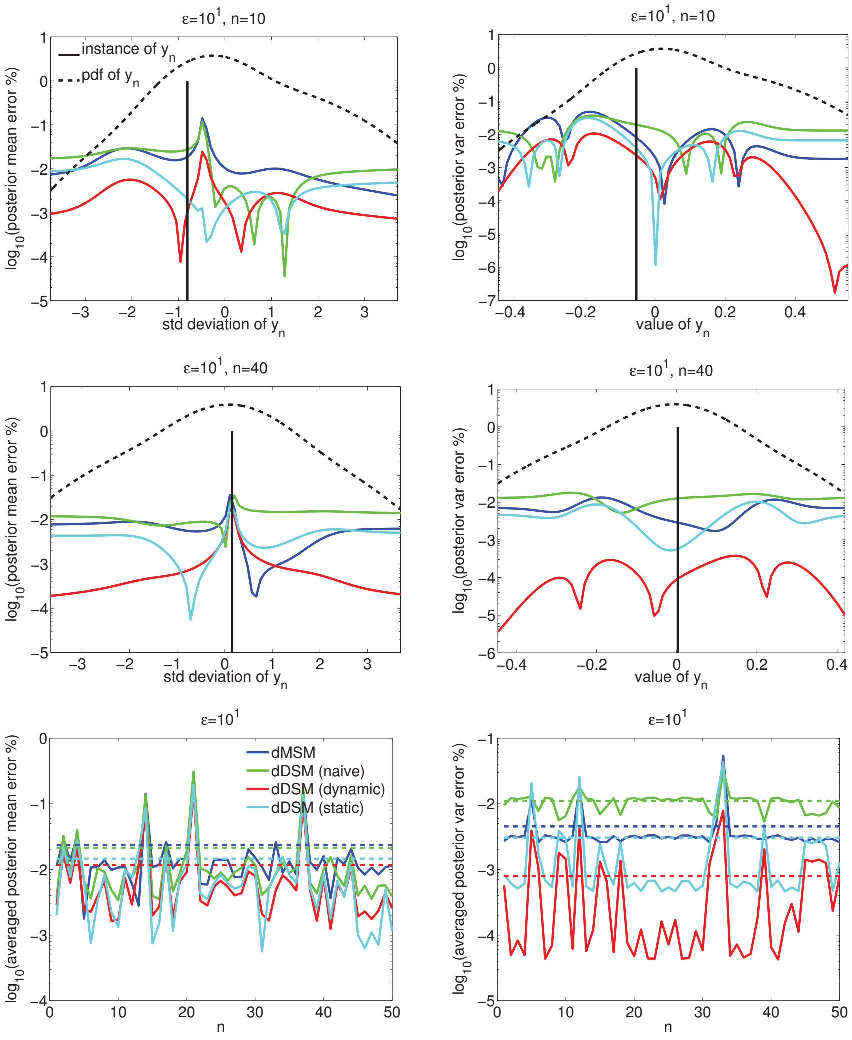

FIG. 4.7. *The relative errors of the approximations of the posterior $u_n|Y_n$ distributions that depend on the realization of $y_n = u_n + \eta_n$, and their statistical averages with respect to the law of y_n when $\epsilon = 10^0$. In Gaussian sum filters, the accuracy of the posterior variance depends on the instance of y_n (top-right and middle-right).*

4.1.2. Imprecise scale-separation regime. Taking $\epsilon = 10^0$, it is not immediately intuitive whether either the Gaussian description or the Gaussian mixture description is a better approximation of the SSM. It turns out that, in this case, the Gaussian filter for SSM is more suitable as the reference solution; our investigation of this issue can be found in Subsection 4.2. Accordingly we find dynamic and static calibrations, and implement Gaussian filters for DSM, MSM and SSM. We depict Figure 4.3 and Figure 4.4, which correspond, respectively, to Figure 4.1 and Figure 4.2. The scenario interpreted from the figures is basically the same as the one arising when $\epsilon = 10^{-1}$, with one exception that the naive DSM is less accurate than the MSM. This is no surprise, because ϵ is no longer small and Θ_{naive} is no longer expected to be valid. Therefore, the overall errors are ordered as: DSM (dynamic calibration) $\lesssim$ DSM (static

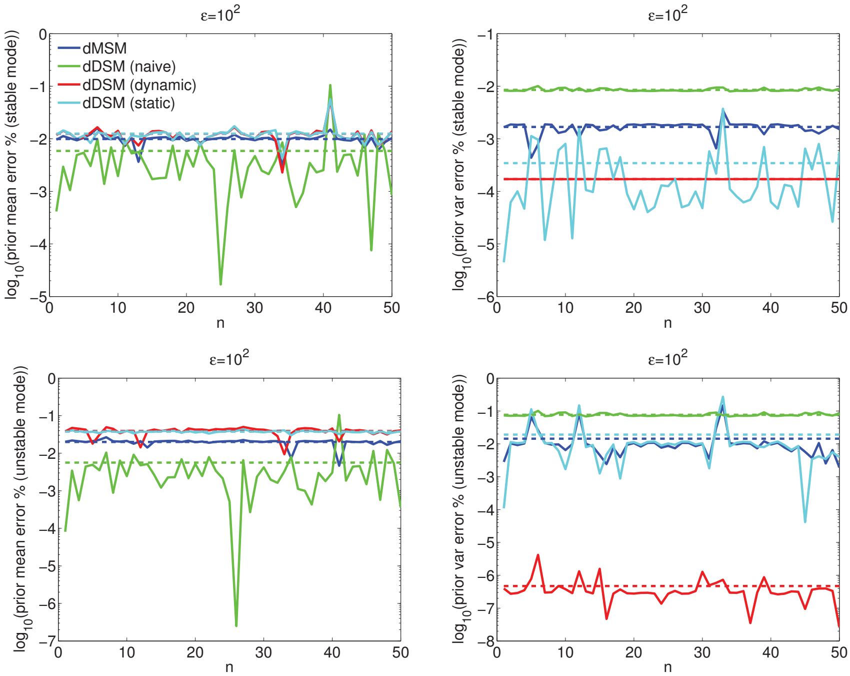

FIG. 4.8. *The relative errors of the mean and variance approximations of each Gaussian kernels of the prior distributions when* $\epsilon = 10^2$.

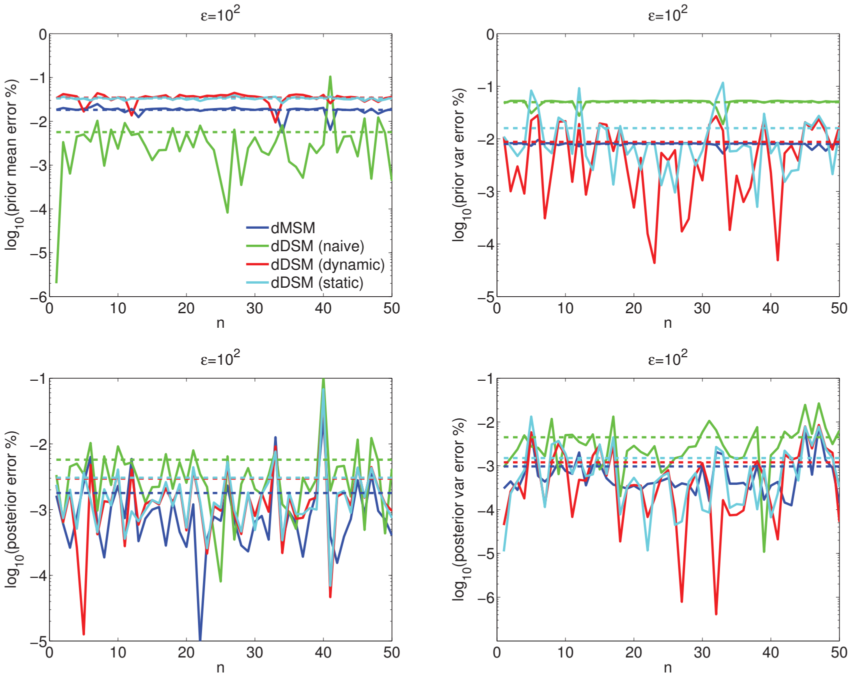

FIG. 4.9. *The relative errors of the mean and variance approximations of the prior* $u_n|Y_{n-1}$ *(top) and posterior* $u_n|Y_n$ *(bottom) distributions when* $\epsilon = 10^2$.

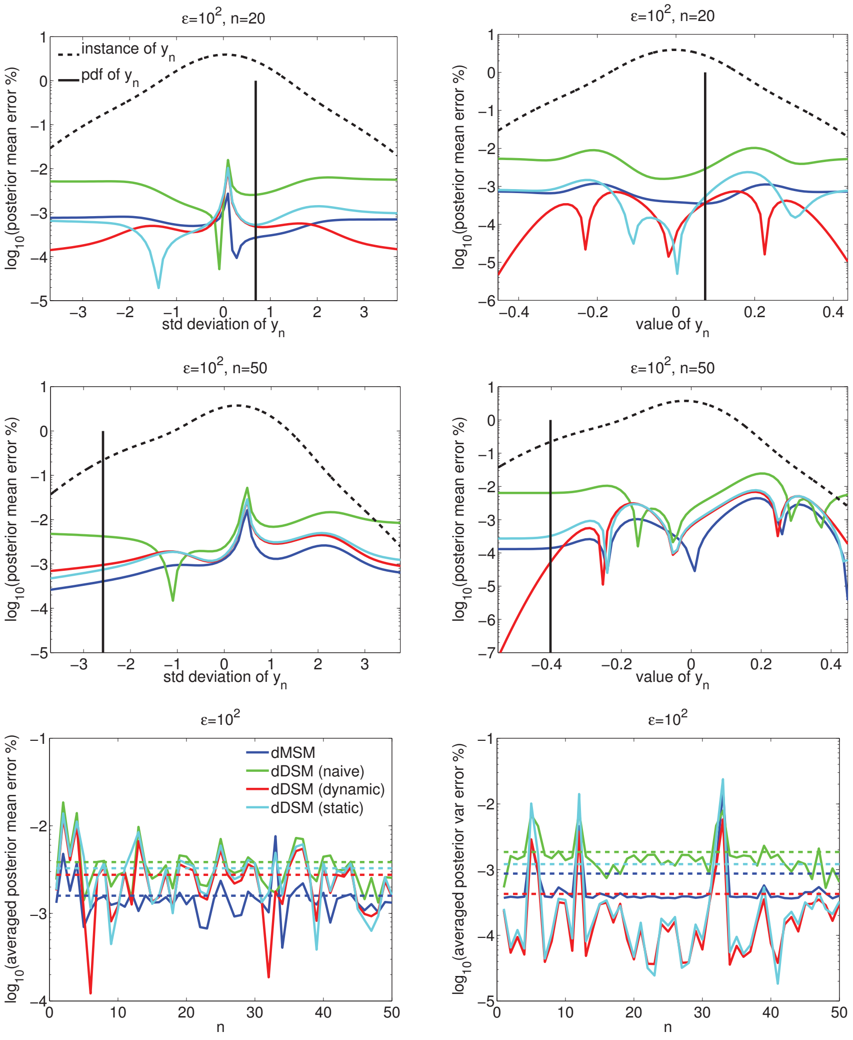

FIG. 4.10. *The relative errors of the approximations of the posterior $u_n|Y_n$ distributions that depend on the realization of $y_n = u_n + \eta_n$, and their statistical averages with respect to the law of y_n when $\epsilon = 10^2$.*

calibration) < MSM < DSM (naive set).

4.1.3. Moderately rare-event regime. When $\epsilon = 10^1$, it is shown in Subsection 4.2 that the Gaussian sum filter for SSM, made efficient by merging the mixture approximation of the posterior into a Gaussian at every observation time, is indeed better than the Gaussian filter for the reference solution. We apply the same kind of Gaussian sum filters for dMSM and dDSM. For the dDSM implementations, taking Θ'_{naive} as a starting point, we solve dynamic calibrations for each of two evolving Gaussian kernels. We then individually average them to obtain a static calibration.

In Figure 4.5, we depict the relative error for each of the Gaussian kernel approximations. Combining these two cases, we plot Figure 4.6 and Figure 4.7, which correspond to Figure 4.3 and Figure 4.4 respectively. Importantly, for comparison, we additionally plot the result from DSM with $(\mu = \bar{\gamma}_\infty, \nu = 0.1\bar{\gamma}_\infty, \sigma = 5\sigma_u)$. These parameters are the

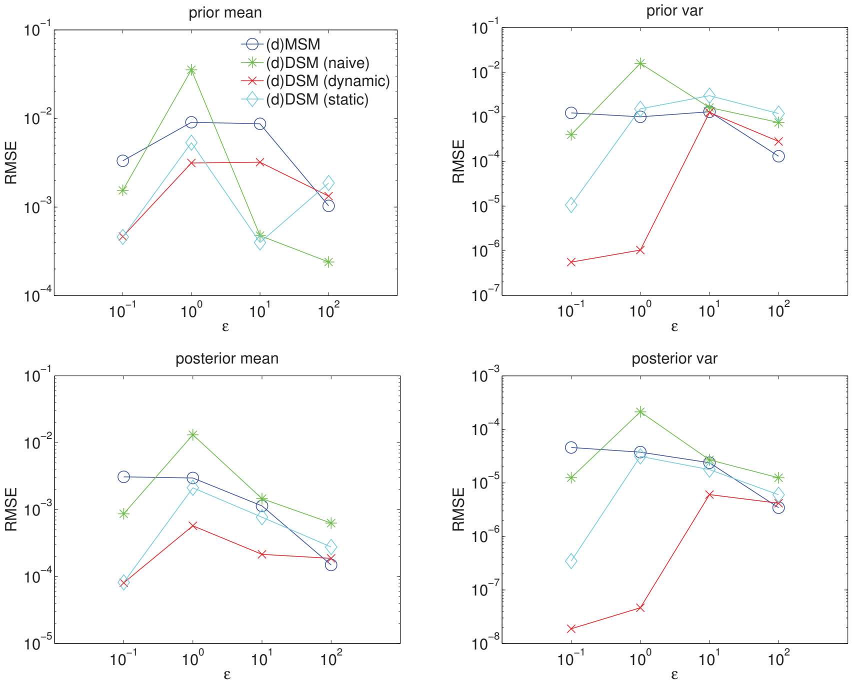

FIG. 4.11. *The root mean square errors between the references from SSM filters and the approximations from MSM and DSM filters.*

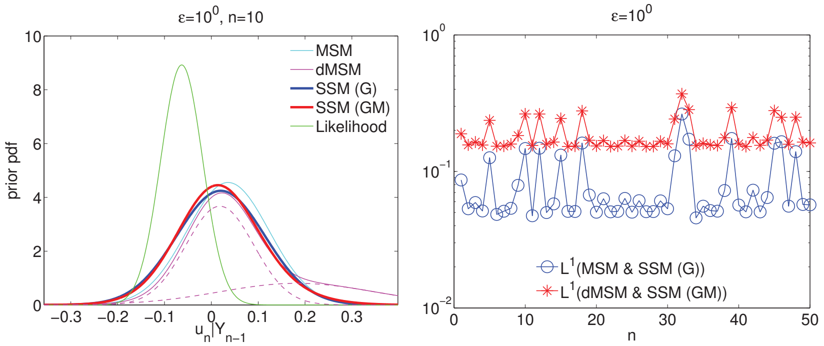

FIG. 4.12. *The Gaussian and Gaussian mixture approximations of SSM (labeled SSM(G) and SSM(GM) respectively) together with MSM and dMSM when* $\epsilon = 10^0$. *The dashed lines represent two Gaussian kernels consisting of dMSM (left).*

ones used in [16]. They are selected as suitable from direct numerical simulations in this parameter regime, and are interestingly very close to Θ_{naive}. Here the DSM appears as a reasonable approximation of SSM, but this Gaussian filter is characterized by significantly less accuracy than the remaining Gaussian sum filters.

Our simulations further indicate, in this case, that the dependency of filter accuracy on the observation is much more complicated than the previous Gaussian filtering examples (Figure 4.7). The overall errors are of the order : dDSM (dynamic calibration) $\lesssim$ dDSM (static calibration) $<$ dMSM $\lesssim$ dDSM (naive set) $<$ DSM (naive set).

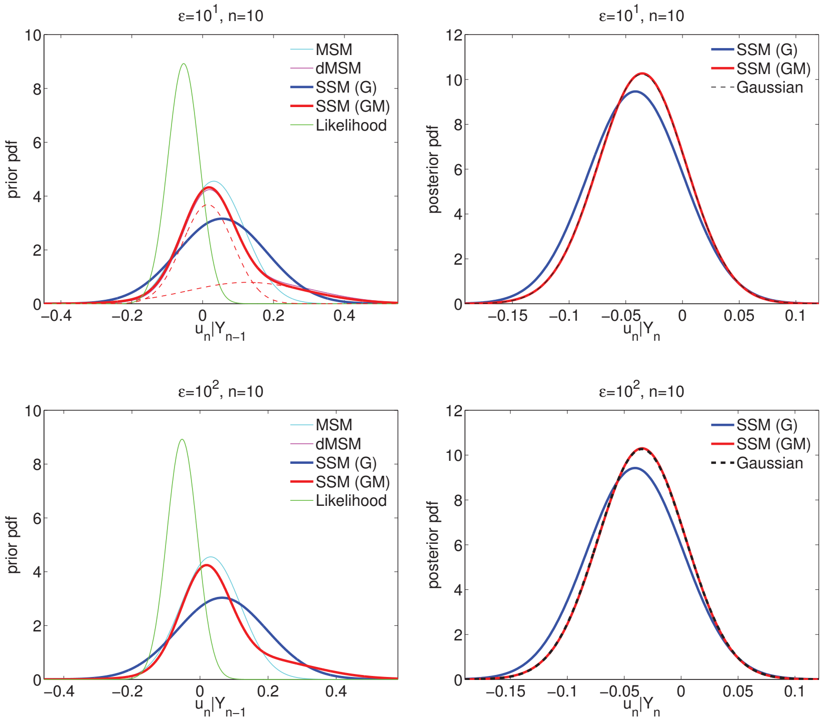

FIG. 4.13. *The Gaussian and Gaussian mixture approximations of SSM prior (left) and Gaussian approximations of SSM posterior (right) when $\epsilon=10^1$ (top) and $\epsilon=10^2$ (bottom). The dashed lines are two Gaussian kernels of SSM approximation (top-left) and Gaussian approximation of Gaussian mixture SSM (right).*

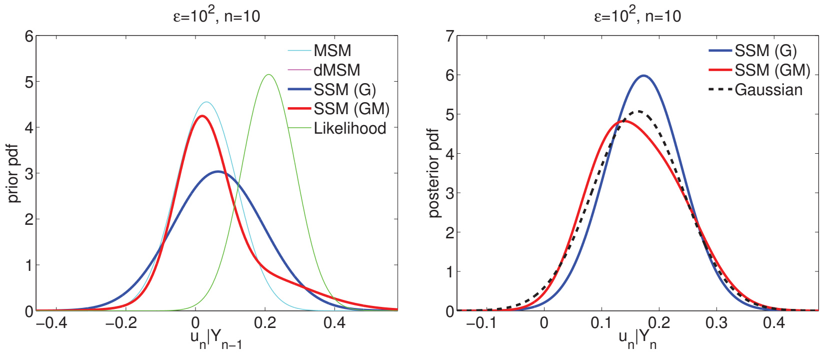

FIG. 4.14. *The Gaussian and Gaussian sum approximations of SSM prior (left) and Gaussian approximations of SSM posterior (right) when $\epsilon=10^2$ and $R=0.75E$.*

4.1.4. Extremely rare-event regime. Like the preceding case, we take as reference the Gaussian sum filter for SSM with projection of posterior into the set of Gaussian distributions. The overall scenario when $\epsilon=10^2$ is similar to the case with $\epsilon=10^1$, except that dMSM becomes more accurate. We plot Figure 4.8, Figure 4.9 and Figure 4.10 that correspond to Figure 4.5, Figure 4.6 and Figure 4.7 respectively. We see the overall errors are of the order : dDSM (dynamic calibration) $\simeq$ dMSM $<$ dDSM (static calibration) $<$ dDSM (naive set). Note dMSM is quite accurate in this case because ϵ is very large.

4.1.5. Summary. To summarize we plot the root mean square errors of mean and variance between the reference and approximations for all four choices of ϵ in Figure 4.11.

4.2. Supplementary analysis. This section discusses our choices, especially in relation to choice of reference solution, made while performing numerical simulations in Subsection 4.1; it can be skipped without harming the understanding of the main messages of the paper.

4.2.1. Imprecise scale-separation regime. In this case where we take $\epsilon = 10^0$, to make sure whether either the Gaussian description or the Gaussian sum description is a better approximation of the SSM, what we do is to compare the similarity/distance between MSM (note the derivation corresponds to $\epsilon = 0$) and Gaussian approximation of SSM, and that between dMSM (that corresponds to $\epsilon = \infty$) and Gaussian sum approximation of SSM.

To that end, we plot the prior distributions from all four cases, when $n = 10$ (and we do the same in the remaining examples), in the left panel of Figure 4.12. We see that the dMSM has a one-sided fat tail, which is due to the contribution by the Gaussian kernel evolved while γ is in the unstable mode. However this feature is not apparent in the mixture approximation of SSM (in fact both Gaussian and Gaussian mixture approximations of SSM are very similar and unimodal). Furthermore, the L^1 distance between MSM and SSM (Gaussian) is significantly smaller than the one between dMSM and SSM (Gaussian mixture), as shown in the right panel of Figure 4.12. The discussion demonstrates that, in this parameter regime, the Gaussian filter for SSM is more suitable as a reference solution than is the Gaussian sum filter.

4.2.2. Moderately rare-event regime. When $\epsilon = 10^1$, we plot the four relevant prior distributions in the top-left panel of Figure 4.13 . While MSM and SSM (Gaussian) are distant from one another, both dMSM and SSM (Gaussian mixture) are characterized by a one-sided fat tail, in contrast to the case of $\epsilon = 10^0$, and further are very close to one another. Therefore SSM with the Gaussian sum filter is chosen as the appropriate reference solution.

We turn our attention to the validity of Gaussian approximation of the Gaussian mixture posterior. The top-right panel of Figure 4.13 depicts the posterior of SSM (Gaussian mixture), which consists of two kernels. The distribution is well approximated by a single Gaussian that has the same mean and variance. This is due to the sharpness of the likelihood we choose (discussed shortly). We may thus approximate the filtering solution by a Gaussian at every observation time, and we can apply Gaussian sum filters in a computationally tractable way without harming accuracy.

4.2.3. Extremely rare-event regime. With regard to SSM filter, the scenario when $\epsilon = 10^2$ is the same as the case with $\epsilon = 10^1$. In the bottom of Figure 4.13, the priors of dMSM and SSM (Gaussian sum) are almost indistinguishable, and the SSM posterior is accurately approximated by a Gaussian.

We conclude the current section with further study of the Gaussian approximation of the posterior. Recall we have fixed $R_n = 0.25E$ thus far. In this case, it is shown that the Gaussian approximation of the posterior can be performed without losing accuracy, but this may not be the case when R_n is bigger. In Figure 4.14, we plot the prior and posterior with $R_n = 0.75E$. Due to the flatter likelihood, the posterior with two kernels significantly deviates from the Gaussian approximation. In this case, the Gaussian approximation of the posterior cannot guarantee the accuracy of the filtering solution.

5. Conclusions

In this paper we have employed simplified models for the estimation of a partially observed turbulent signal. Our test bed, the switching stochastic model (SSM), is a stochastic differential equation driven by a sign-alternating two-state Markov process. The system is either forced or dissipated depending on the sign of the driving signal, and as a consequence exhibits intermittent turbulent bursting. It is a cheap surrogate for turbulent signal generation, allowing rapid prototyping of a variety of approximate filters—filters with model error. Two approximate models (MSM, DSM) for SSM have been constructed via simplification of the switching process underlying the turbulent bursting, leading to a Gaussian description for the filtering solution. We study the moment generating function (MGF) with respect to the time integral of the switching process to reveal that these two models precisely mimic the SSM behavior when the switching frequency is relatively high. In addition to these two models, based on the same argument, we also build two models (dMSM, dDSM) whose regime of validity is rare switching for the driving signal. We associate these two models with Gaussian sum filters.

We first use the ergodicity of the switching processes to prove MSM (dMSM) is the high (low) switching frequency limit of SSM and DSM (dDSM). Besides verifying the consistency of the proposed approximate models, the convergence results give rise to an analytic determination of DSM (dDSM) parameters when the time-scales of driving input and system output are well separated. We achieve this from the comparison between asymptotics of MGFs in each of two opposing parameter regimes, because their matching implies the lower order moments of the corresponding DSM (dDSM) are very close to those of SSM. The result again gives rise to a determination of DSM (dDSM) parameters when two time-scales are weakly separated. In this case, we numerically find a minimizer of the sum-of-squares error function between the mean and variance of SSM and DSM (dDSM) for which the previous analytic solutions is used as the initial candidate. In our numerical simulations, the filtering results utilizing DSM (dDSM) with the parameters tuned according to our suggestions show significant improvements in accuracy in all the parameter regimes that we examined. Furthermore, when the time-scale separation is weak, the cost of performing the minimizations can be alleviated by averaging the parameters calculated only for a number of observation time steps, while maintaining the accuracy of the filtering solution to a considerable extent.

We have used the tools from three different research areas: limit theorems, asymptotic analysis and computational optimization to complete the whole scenario. These methods are not separate but carefully chained together through a solution cascade to provide a significant step in the analysis and development of filters utilizing approximate models suggested from a rigorous analysis of the underlying system. As the ultimate goal of filtering with model error is to estimate the system state and associated uncertainties of real-world turbulence, at tractable cost, our future work will include the development of these algorithms, and their benchmarking, in the case where the true signal is not generated by SSM, but rather by a real turbulence model.

Acknowledgement. Both authors are supported by the ERC-AMSTAT grant No. 226488. AMS is also supported by EPSRC and ONR.

Appendix A. Switching stochastic model. This section concerns the SSM. Subsection A.1 analytically computes MGFs of integral process of the driving signal, and Subsection A.2 studies their asymptotic behavior. In Subsection A.3, we develop filters for SSM. In the present section, $\lambda_\pm$ will be used in place of $\lambda_\pm/\epsilon$ for notational brevity.

A.1. Moment generating function (MGF) of integral process. Let $\Gamma_t = \int_0^t \gamma(s)ds$ be the time integral of switching process. We calculate

$$\left\langle e^{\alpha\Gamma_t} | \gamma_0 = \gamma_+ \right\rangle \tag{A.1}$$

in Subsection A.1.1 (see Equations (A.3), (A.4), (A.8)), and

$$\begin{gathered} \left\langle e^{\alpha(\Gamma_T - \Gamma_t)} \right\rangle \\ \left\langle e^{\alpha(\Gamma_T - \Gamma_t)} | \gamma_0 = \gamma_\pm \right\rangle \end{gathered} \tag{A.2}$$

in Subsection A.1.2 (see Equations (A.10), (A.11)).

A.1.1. When $\gamma(0) = \gamma_+$, let γ_k (abusing notation with γ_t for economy of notation) denote the value of $\gamma(s)$ after undergoing exactly k transitions, i.e., $\gamma_k = \gamma_+$ for even k and $\gamma_k = \gamma_-$ for odd k. Let $\tau_k \triangleq \tau^{\gamma_+}$ for even k and $\tau_k \triangleq \tau^{\gamma_-}$ for odd k. Let $T_n = \sum_{k=0}^{n-1} \tau_k$ and let $N_t = \max\{n \in \mathbb{N} : T_n \le t\}$ denote the number of transitions of $\gamma(s)$ in the interval $s \in (0,t]$. From $\tau_k = T_{k+1} - T_k$, we have

$$\Gamma_t = \int_0^t \gamma(s)ds = \sum_{k=0}^{N_t - 1} \gamma_k \tau_k + \gamma_{N_t}(t - T_{N_t}) = \sum_{k=0}^{N_t-1} (\gamma_k - \gamma_{N_t})\tau_k + \gamma_{N_t} t.$$

Note $\tau \sim \exp(r)$, so that $\mathbb{E}(e^{\alpha\tau}) = \int_0^\infty e^{\alpha t} r e^{-rt} dt = (1 - \frac{\alpha}{r})^{-1}$ for $\alpha < r$. Since $\{\tau_k\}_{k=0}^{n-1}$ are mutually independent and τ_k is distributed according to the exponential distribution, a formal expansion of the expectation (A.1) is given by

$$\begin{aligned} \left\langle e^{\alpha\Gamma_t} | \gamma_0 = \gamma_+ \right\rangle &= \sum_{n=0}^\infty \mathbb{P}(N_t = n | \gamma_0 = \gamma_+) \left\langle e^{\alpha\left(\sum_{k=0}^{n-1}(\gamma_k - \gamma_n)\tau_k + \gamma_n t\right)} | \gamma_0 = \gamma_+ \right\rangle \\ &= \sum_{\text{n:even}} \mathbb{P}(N_t = n | \gamma_0 = \gamma_+) e^{\alpha\gamma_+ t} \left(1 - \frac{\alpha(\gamma_- - \gamma_+)}{\lambda_-}\right)^{-\frac{n}{2}} \\ &\quad + \sum_{\text{n:odd}} \mathbb{P}(N_t = n | \gamma_0 = \gamma_+) e^{\alpha\gamma_- t} \left(1 - \frac{\alpha(\gamma_+ - \gamma_-)}{\lambda_+}\right)^{-\frac{n+1}{2}} \end{aligned} \tag{A.3}$$

for $\frac{\lambda_-}{\gamma_- - \gamma_+} < \alpha < \frac{\lambda_+}{\gamma_+ - \gamma_-}$.

In what follows, we shall show that, in case of identical $\lambda_+ = \lambda_-$, Equation (A.3) reduces to the closed-form representation

$$\begin{aligned} \left\langle e^{\alpha\Gamma_t} | \gamma_0 = \gamma_+ \right\rangle = e^{-\lambda t} \Bigg(& e^{\alpha\gamma_+ t} \cosh\left(\lambda t \left(1 - \frac{\alpha(\gamma_- - \gamma_+)}{\lambda_-}\right)^{-\frac{1}{2}}\right) \\ & + e^{\alpha\gamma_- t} \sinh\left(\lambda t \left(1 - \frac{\alpha(\gamma_+ - \gamma_-)}{\lambda_+}\right)^{-\frac{1}{2}}\right) \left(1 - \frac{\alpha(\gamma_+ - \gamma_-)}{\lambda_+}\right)^{-\frac{1}{2}} \Bigg) \end{aligned} \tag{A.4}$$

where λ denotes $\lambda_+ = \lambda_-$. When $\lambda_+ \neq \lambda_-$ are distinct, we demonstrate that an accurate approximation of the expectation (A.3) can be obtained with the help of the probability distributions (A.8).

In order to compute $\mathbb{P}(N_t = n|\gamma_0 = \gamma_+)$ and (A.3), we notice

$$\begin{aligned}\mathbb{P}(N_t = n) &= \mathbb{P}(N_t \geq n) - \mathbb{P}(N_t \geq n+1)\\ &= \mathbb{P}(T_n \leq t) - \mathbb{P}(T_{n+1} \leq t)\end{aligned} \tag{A.5}$$

for $t \geq 0$. Let f_n denote the probability density function of T_n, given $\gamma_0 = \gamma_+$. Then Equation (A.5) can be written as

$$\begin{cases}\mathbb{P}(N_t = 0|\gamma_0 = \gamma_+) & = 1 - \int_0^t f_1(s)ds\\ \mathbb{P}(N_t = n|\gamma_0 = \gamma_+) & = \int_0^t f_n(s)ds - \int_0^t f_{n+1}(s)ds, \quad n \geq 1\end{cases} \tag{A.6}$$

Let $f_+(t) := \lambda_+ e^{-\lambda_+ t}$ and $f_-(t) := \lambda_- e^{-\lambda_- t}$, and let $*$ denote the convolution. Then $(f_+)^{*m}(t) = (\lambda_+)^m \frac{t^{m-1}}{(m-1)!} e^{-\lambda_+ t}$, $(f_-)^{*l}(t) = (\lambda_-)^l \frac{t^{l-1}}{(l-1)!} e^{-\lambda_- t}$, and

$$\begin{aligned}f_n(t) &= (f_+)^{*m} * (f_-)^{*l}(t)\\ &= \frac{\lambda_+^m \lambda_-^l}{(m-1)!(l-1)!} \int_0^t s^{m-1} e^{-\lambda_+ s} (t-s)^{l-1} e^{-\lambda_-(t-s)} ds\\ &= \frac{\lambda_+^m \lambda_-^l}{(m-1)!(l-1)!} e^{-\lambda_- t} t^{m+l-1} \int_0^1 s^{m-1} (1-s)^{l-1} e^{-(\lambda_+ - \lambda_-)ts} ds\end{aligned} \tag{A.7}$$

where $n = m+l$ and $l = m-1$ or $l = m$, and $m \geq 1$. Let us consider two different cases of $\lambda_+ = \lambda_-$ in A.1.1.1, and $\lambda_+ \neq \lambda_-$ in A.1.1.2.

A.1.1.1. When $\lambda_+ = \lambda_-$, both are denoted by λ. In this case, from the relationship between the gamma function and the beta function

$$\int_0^1 s^{m-1}(1-s)^{l-1} ds = \frac{(m-1)!(l-1)!}{(m+l-1)!},$$

Equation (A.7) becomes

$$f_n(t) = \lambda^n \frac{t^{n-1}}{(n-1)!} e^{-\lambda t}.$$

Then integration by parts

$$\int_0^t f_n(s) ds = \frac{\lambda^n}{n!} t^n e^{-\lambda t} + \int_0^t f_{n+1}(s) ds$$

and Equation (A.6) ensures that N_t is distributed according to Poisson distribution with parameter λt:

$$\mathbb{P}(N_t = n|\gamma_0 = \gamma_+) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

Therefore, from substituting this into Equation (A.3), Equation (A.4) is derived.

A.1.1.2. When $\lambda_+ \neq \lambda_-$, denoting $B(m,n;\alpha) \equiv \int_0^1 s^{m-1}(1-s)^{n-1}e^{\alpha s}ds$, integration by parts yields

$$B(m,n;\alpha) = \frac{1}{\alpha}\left[(m-1)B(m-1,n;\alpha) - (n-1)B(m,n-1;\alpha)\right].$$

Using $B(m-1,n;\alpha) = e^{\alpha}B(n,m-1;-\alpha)$ and using

$$\int t^n e^{\alpha t}dt = \frac{t^n e^{\alpha t}}{\alpha} - \frac{n}{\alpha}\int t^{n-1}e^{\alpha t}dt,$$

the probability density function f_n, and therefore the probability distribution of N_t, are recursively obtained from Equation (A.6). For instance, the first three of them are given by

$$\begin{cases} \mathbb{P}(N_t = 0|\gamma_0 = \gamma_+) & = e^{-\lambda_+ t} \\ \mathbb{P}(N_t = 1|\gamma_0 = \gamma_+) & = \frac{\lambda_+(e^{-\lambda_+ t} - e^{-\lambda_- t})}{-(\lambda_+ - \lambda_-)} \\ \mathbb{P}(N_t = 2|\gamma_0 = \gamma_+) & = \frac{\lambda_-\lambda_+ \exp(-\lambda_- t) - \lambda_-\lambda_+ \exp(-\lambda_+ t) + t\lambda_-\lambda_+(\lambda_- - \lambda_+)\exp(-\lambda_+ t)}{(\lambda_- - \lambda_+)^2} \\ & \dots \end{cases} \quad \text{(A.8)}$$

The process enables calculation of a truncation of Equation (A.3). The following theorem ensures that it is possible to obtain any desired accuracy of this approximate solution, by appropriate truncation.

THEOREM A.1. *The series in Equation (A.3) is uniformly convergent.*

Proof. When $0 < \lambda_+ < \lambda_-$, the inequality

$$f_n(t) \leq \left(\frac{\lambda_-}{\lambda_+}\right)^l (f_+)^{*n}(t)$$

holds from $e^{-(\lambda_+ - \lambda_-)ts} \leq e^{-(\lambda_+ - \lambda_-)t}$ for $s \in [0,1]$, where $n = m + l$ and, $l = m-1$ or $l = m$. Using

$$\int_0^t \lambda^n \frac{s^{n-1}}{(n-1)!}e^{-\lambda s}ds \leq \int_0^t \lambda^n \frac{s^{n-1}}{(n-1)!}ds = \frac{(\lambda t)^n}{n!},$$

when $n = 2m$ is even, we obtain

$$\begin{aligned} \mathbb{P}(N_t = n|\gamma_0 = \gamma_+) &= \left|\int_0^t (f_n(s) - f_{n+1}(s))\,ds\right| \\ &\leq \int_0^t \left(\left(\frac{\lambda_-}{\lambda_+}\right)^m (f_+)^{*n}(s) + \left(\frac{\lambda_+}{\lambda_-}\right)^m (f_+)^{*(n+1)}(s)\right)ds \\ &\leq \left(\frac{\lambda_-}{\lambda_+}\right)^{\frac{n}{2}} \left(\frac{(\lambda_+ t)^n}{n!} + \frac{(\lambda_+ t)^{n+1}}{(n+1)!}\right). \end{aligned}$$

When $n = 2m-1$ is odd, it satisfies

$$\mathbb{P}(N_t = n|\gamma_0 = \gamma_+) \leq \left(\frac{\lambda_-}{\lambda_+}\right)^{\frac{n-1}{2}} \frac{(\lambda_+ t)^n}{n!} + \left(\frac{\lambda_-}{\lambda_+}\right)^{\frac{n+1}{2}} \frac{(\lambda_+ t)^{n+1}}{(n+1)!}.$$

Then, like the case of Equation (A.4), the series of Equation (A.3) is bounded above by a linear combination of hyperbolic sine and cosine. Uniform convergence follows from an application of the Weierstrass M-test.

When $0<\lambda_- <\lambda_+$, from $e^{-(\lambda_+ - \lambda_-)ts}<1$ for $s\in[0,1]$, we obtain

$$f_n(t)\le \left(\frac{\lambda_+}{\lambda_-}\right)^m (f_-)^{*n}(t)$$

when $n=m+l$ where $l=m-1$ or $l=m$. Using an argument similar to the previous case, we are done. □

A.1.2. Because $\rho_\gamma(t)\equiv(\mathbb{P}(\gamma_t=\gamma_+),\mathbb{P}(\gamma_t=\gamma_-))^t$ satisfies $d\rho_\gamma(t)/dt=L^t\rho_\gamma(t)$, where superscript t denotes transpose and

$$L\equiv\begin{pmatrix}-\lambda_+ & \lambda_+\\ \lambda_- & -\lambda_-\end{pmatrix},$$

the solution is given by

$$\begin{pmatrix}\mathbb{P}(\gamma_t=\gamma_+)\\ \mathbb{P}(\gamma_t=\gamma_-)\end{pmatrix}=\frac{1}{\lambda_+ + \lambda_-}\begin{pmatrix}\lambda_- + \lambda_+ e^{-(\lambda_- + \lambda_+)t} & \lambda_- - \lambda_- e^{-(\lambda_- + \lambda_+)t}\\ \lambda_+ - \lambda_+ e^{-(\lambda_- + \lambda_+)t} & \lambda_+ + \lambda_- e^{-(\lambda_- + \lambda_+)t}\end{pmatrix}\begin{pmatrix}\mathbb{P}(\gamma_0=\gamma_+)\\ \mathbb{P}(\gamma_0=\gamma_-)\end{pmatrix}. \tag{A.9}$$

Then, from substituting Equation (A.9) into

$$\begin{aligned}\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}\right\rangle &=\mathbb{P}(\gamma_t=\gamma_+)\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}|\gamma_t=\gamma_+\right\rangle+\mathbb{P}(\gamma_t=\gamma_-)\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}|\gamma_t=\gamma_-\right\rangle\\ &=\mathbb{P}(\gamma_0=\gamma_+)\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}|\gamma_0=\gamma_+\right\rangle+\mathbb{P}(\gamma_0=\gamma_-)\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}|\gamma_0=\gamma_-\right\rangle,\end{aligned} \tag{A.10}$$

the following equation may be deduced:

$$\begin{aligned}\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}|\gamma_0=\gamma_+\right\rangle=\frac{1}{\lambda_+ + \lambda_-}\Bigg(&\left(\lambda_- + \lambda_+ e^{-(\lambda_- + \lambda_+)t}\right)\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}|\gamma_t=\gamma_+\right\rangle\\ &+\left(\lambda_+ - \lambda_+ e^{-(\lambda_- + \lambda_+)t}\right)\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}|\gamma_t=\gamma_-\right\rangle\Bigg).\end{aligned} \tag{A.11}$$

Note that, provided the expectation (A.1) is solved, a formula for $\left\langle e^{\alpha\Gamma_t}|\gamma_0=\gamma_-\right\rangle$ is immediately derived from the parameter exchange $(\lambda_+,\gamma_+)\leftrightarrow(\lambda_-,\gamma_-)$. Therefore, using $\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}|\gamma_t=\gamma_\pm\right\rangle=\left\langle e^{\alpha\Gamma_{T-t}}|\gamma_0=\gamma_\pm\right\rangle$, one can compute the expectation (A.11). This in turn allows for a computation of the expectation (A.10), hence those of Equations (A.2).

A.2. Asymptotics for MGF of integral process. We derive asymptotic formulae for MGFs (A.1), (A.2) when $\epsilon\ll 1$ (scale-separation regime) and $\epsilon\gg 1$ (rare-event regime). The results are Equations (A.12), (A.13), (A.14).

A.2.1. Scale-separation regime. When λ_+ and λ_- are distinct, and ϵ is small, the probability density function of T_n is given by the convolution of f_+ and f_-, each one $\frac{n}{2}\left(\pm\frac{1}{2}\right)$ times respectively. Let $\bar\lambda\equiv\frac{2\lambda_+\lambda_-}{\lambda_+ + \lambda_-}$ be the harmonic mean of λ_+ and λ_-. Then from

$$\left(1-\frac{i\alpha}{\lambda_-}\right)^{-\frac{n}{2}}\left(1-\frac{i\alpha}{\lambda_+}\right)^{-\frac{n}{2}}\simeq\left(1-\frac{i\alpha}{\bar\lambda}\right)^{-n}$$

for large n, T_n is approximately distributed according to the gamma distribution with rate $\bar{\lambda}$, as in A.1.1.1:

$$\mathbb{P}(N_t = n | \gamma_0 = \gamma_+) \simeq e^{-\bar{\lambda} t} \frac{(\bar{\lambda} t)^n}{n!}.$$

Substituting this into Equation (A.4), we obtain

$$\begin{aligned}\left\langle e^{\alpha \Gamma_t} | \gamma_0 = \gamma_+ \right\rangle \simeq e^{-\bar{\lambda} t} \Bigg(& e^{\alpha \gamma_+ t} \cosh\left(\bar{\lambda} t \left(1 - \frac{\alpha(\gamma_- - \gamma_+)}{\lambda_-}\right)^{-\frac{1}{2}} \right) \\ & + e^{\alpha \gamma_- t} \sinh\left(\bar{\lambda} t \left(1 - \frac{\alpha(\gamma_+ - \gamma_-)}{\lambda_+}\right)^{-\frac{1}{2}} \right) \left(1 - \frac{\alpha(\gamma_+ - \gamma_-)}{\lambda_+}\right)^{-\frac{1}{2}} \Bigg) \\ \simeq \exp & \left(\alpha \bar{\gamma}_\infty t + \alpha^2 \frac{3}{8} \frac{(\gamma_- - \gamma_+)^2 (\lambda_-^2 + \lambda_+^2)}{\lambda_- \lambda_+ (\lambda_+ + \lambda_-)} t + \frac{\alpha(\gamma_+ - \gamma_-)}{4\lambda_+} \right)\end{aligned}$$

which, after the rescaling $\lambda_\pm \mapsto \lambda_\pm / \epsilon$, reads

$$\left\langle e^{\alpha \Gamma_t} | \gamma_0 = \gamma_+ \right\rangle \simeq \exp\left(\alpha \bar{\gamma}_\infty t + \alpha^2 \frac{3}{8} \frac{(\gamma_- - \gamma_+)^2 (\lambda_-^2 + \lambda_+^2)}{\lambda_- \lambda_+ (\lambda_+ + \lambda_-)} t\epsilon + \frac{\alpha(\gamma_+ - \gamma_-)}{4\lambda_+} \epsilon + \mathcal{O}(\epsilon^2) \right) \tag{A.12}$$

where the identities

$$(1+x)^k = 1 + kx + (k(k-1)/2!)x^2 + \cdots$$

and

$$\frac{1}{2} e^{x + \alpha\epsilon} + \frac{1}{2} e^{x + \beta\epsilon} (1 + \gamma\epsilon) \simeq e^{x + ((\alpha + \beta + \gamma)/2)\epsilon}$$

are used.

Furthermore, we use $ce^{A+\alpha\epsilon} + (1-c)e^{A+\beta\epsilon} \simeq e^{A+(c\alpha+(1-c)\beta)\epsilon}$ to obtain

$$\begin{aligned}\left\langle e^{\alpha(\Gamma_T - \Gamma_t)} \right\rangle \simeq \exp \Bigg(& \alpha \bar{\gamma}_\infty (T-t) + \alpha^2 \frac{3}{8} \frac{(\gamma_- - \gamma_+)^2 (\lambda_-^2 + \lambda_+^2)}{\lambda_- \lambda_+ (\lambda_+ + \lambda_-)} (T-t)\epsilon \\ & + \alpha \left(\mathbb{P}(\gamma_0 = \gamma_+) \frac{(\gamma_+ - \gamma_-)}{4\lambda_+} + \mathbb{P}(\gamma_0 = \gamma_-) \frac{(\gamma_- - \gamma_+)}{4\lambda_-} \right) \epsilon + \mathcal{O}(\epsilon^2) \Bigg)\end{aligned} \tag{A.13}$$

when $\epsilon < T - t$.

A.2.2. Rare-event regime. When ϵ is large, we just take the first term in Equation (A.3) to obtain

$$\begin{aligned}\left\langle e^{\alpha \Gamma_T} \right\rangle &= \mathbb{P}(\gamma_0 = \gamma_+) \left\langle e^{\alpha \Gamma_T} | \gamma_0 = \gamma_+ \right\rangle + \mathbb{P}(\gamma_0 = \gamma_-) \left\langle e^{\alpha \Gamma_T} | \gamma_0 = \gamma_- \right\rangle \\ &\simeq \mathbb{P}(\gamma_0 = \gamma_+) \exp\left(-\frac{\lambda_+}{\epsilon} t + \alpha \gamma_+ t \right) + \mathbb{P}(\gamma_0 = \gamma_-) \exp\left(-\frac{\lambda_-}{\epsilon} t + \alpha \gamma_- t \right).\end{aligned} \tag{A.14}$$

A.3. Filters for SSM. Here we define the Gaussian filter and Gaussian sum filter for SSM.

A.3.1. Gaussian filter. We build the filter as an assumed density filter: we assume u_0 to be Gaussian, and further assume the independence of (u_0,γ_0), hence that of (u_0,Γ_T). Then, from Equation (3.3), the equations

$$\begin{aligned}\langle u_T\rangle &= \left\langle e^{-\Gamma_T}\right\rangle \langle u_0\rangle \\ \mathrm{Var}(u_T) &= \left\langle e^{-2\Gamma_T}\right\rangle \left(\mathrm{Var}(u_0)+\langle u_0\rangle^2\right) - \left\langle e^{-\Gamma_T}\right\rangle^2 \langle u_0\rangle^2 + \sigma_u^2 \int_0^T \left\langle e^{-2(\Gamma_T-\Gamma_t)}\right\rangle dt\end{aligned} \tag{A.15}$$

are derived in the case of the SSM. Either from using closed-form solution in case of identical $\lambda_+=\lambda_-$ or from using a truncation of the series solution in case of distinct $\lambda_+\neq\lambda_-$, one can compute

$$\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}\right\rangle$$

for $\alpha=-1,-2$ and $t\in[0,T]$, and compute Equation (A.15). Together with using Equation (A.9) for the prediction of γ_t, the mapping of the first two moments $(u_0,\gamma_0)\mapsto(u_T,\gamma_T)$ has been achieved. To complete the filter, we apply Kalman data assimilation for u_T without updating γ_T, as this is consistent with Bayes' rule when (u_T,γ_T) are independent [12].

A.3.2. Gaussian sum filter. Let u_0 be Gaussian mixture and let the independence of (u_0,γ_0) be assumed. Using

$$\mathbb{P}(u_T)=\mathbb{P}(\gamma_0=\gamma_+)\mathbb{P}(u_T|\gamma_0=\gamma_+)+\mathbb{P}(\gamma_0=\gamma_-)\mathbb{P}(u_T|\gamma_0=\gamma_-)$$

we approximate u_T as Gaussian mixture with the number of Gaussian kernels being doubled. Similarly with Equation (A.15), the mean and variance of $\mathbb{P}(u_T|\gamma_0=\gamma_\pm)$ are determined by

$$\left\langle e^{\alpha(\Gamma_T-\Gamma_t)}|\gamma_0=\gamma_\pm\right\rangle$$

for $\alpha=-1,-2$ and $t\in[0,T]$. Using prior calculations, the conditioned mean and variance of individual kernel are obtained. Then, using Equation (A.9) for the prediction of γ_t, the algorithm of $(u_0,\gamma_0)\mapsto(u_T,\gamma_T)$ is established.

To complete the filter, we apply Kalman data assimilation for each Gaussian kernel of u_T, together with evaluation of the weights, whilst preserving the law of γ_T. Because the latter procedure keeps the number of Gaussian kernels unchanged, a total of 2^n weighted Gaussian kernels describe the posterior distribution after n inter-observation time steps, provided u_0 is Gaussian.

Appendix B. Diffusive stochastic models. This section is concerned with DSM(dDSM). Subsection B.1 presents the moments mapping formulae of DSM. The computation of MGFs of a related integral process, and their asymptotic behaviors, are studied in Subsection B.2 and Subsection B.3, respectively.

B.1. DSM moments mapping. Consider the SDE

$$\begin{cases} du &= -\gamma u\,dt+\sigma_u dB_u \\ d\gamma &= -d_\gamma(\gamma-\bar{\gamma})\,dt+\sigma_\gamma dB_\gamma \end{cases}$$

when (u_0,γ_0) is Gaussian. Let $\Gamma_\gamma(t)\equiv\int_0^t\gamma(s)ds$. Then the path-wise solutions are given by

$$\begin{aligned}u_t&=e^{-\Gamma_\gamma(t)}u_0+\sigma_u\int_0^t e^{-(\Gamma_\gamma(t)-\Gamma_\gamma(s))}dB_u(s)\equiv A_t+B_t\\ \gamma_t&=\bar\gamma+(\gamma_0-\bar\gamma)e^{-d_\gamma t}+\sigma_\gamma\int_0^t e^{-d_\gamma(t-s)}dB_\gamma(s).\end{aligned}$$

Let

$$\begin{aligned}b_\gamma(t)&\equiv(1-e^{-d_\gamma t})/d_\gamma\\ \mathcal{B}_\gamma(t)&\equiv\sigma_\gamma\int_0^t ds\int_0^s e^{-d_\gamma(s-s')}dB_\gamma(s').\end{aligned}$$

Then $\Gamma_\gamma(t)=\bar\gamma(t-b_\gamma(t))+b_\gamma(t)\gamma_0+\mathcal{B}_\gamma(t)$, and we have

$$\begin{aligned}\langle u_t\rangle&=\langle A_t\rangle\\ \langle\gamma_t\rangle&=\bar\gamma+(\langle\gamma_0\rangle-\bar\gamma)e^{-d_\gamma t}\\ \mathrm{Var}(u_t)&=\langle u_t^2\rangle-\langle u_t\rangle^2=\langle A_t^2\rangle+\langle B_t^2\rangle-\langle A_t\rangle^2\\ \mathrm{Var}(\gamma_t)&=e^{-2d_\gamma t}\mathrm{Var}(\gamma_0)+\frac{\sigma_\gamma^2}{2d_\gamma}\left(1-e^{-2d_\gamma t}\right)\\ \mathrm{Cov}(u_t,\gamma_t)&=\langle u_t\gamma_t\rangle-\langle u_t\rangle\langle\gamma_t\rangle=\bar\gamma\left(1-e^{-d_\gamma t}\right)\langle A_t\rangle+e^{-d_\gamma t}\langle A_t\gamma_0\rangle+\langle A_t\dot{\mathcal{B}}_\gamma(t)\rangle-\langle A_t\rangle\langle\gamma_t\rangle\end{aligned}\tag{B.1}$$

where upper dot denotes derivative.

Using

$$\begin{aligned}\langle\Gamma_\gamma(t)-\Gamma_\gamma(s)\rangle&=(b_\gamma(t)-b_\gamma(s))\langle\gamma_0\rangle+\bar\gamma((t-s)-(b_\gamma(t)-b_\gamma(s)))\\ \mathrm{Var}\left(\Gamma_\gamma(t)-\Gamma_\gamma(s)\right)&=(b_\gamma(t)-b_\gamma(s))^2\,\mathrm{Var}(\gamma_0)+\mathrm{Var}\left(\mathcal{B}_\gamma(t)-\mathcal{B}_\gamma(s)\right)\\ \langle\mathcal{B}_\gamma(t)\rangle&=0\\ \mathrm{Var}(\mathcal{B}_\gamma(t))&=-\frac{\sigma_\gamma^2}{2d_\gamma^3}\left(3-4e^{-d_\gamma t}+e^{-2d_\gamma t}-2d_\gamma t\right)\\ \mathrm{Var}(\mathcal{B}_\gamma(t)-\mathcal{B}_\gamma(s))&=-\frac{\sigma_\gamma^2}{d_\gamma^3}\left(1+d_\gamma(s-t)+e^{-d_\gamma(s+t)}\times\left(-1-e^{2d_\gamma s}+\cosh(d_\gamma(s-t))\right)\right)\\ \left\langle e^{-\mathcal{B}_\gamma(t)}\dot{\mathcal{B}}_\gamma(t)\right\rangle&=-\frac{1}{2}\partial_t\left(\mathrm{Var}(\mathcal{B}_\gamma(t))\right)\left\langle e^{-\mathcal{B}_\gamma(t)}\right\rangle\end{aligned}$$

and using

$$\begin{aligned}\langle e^z\rangle&=e^{\langle z\rangle+\frac{1}{2}\mathrm{Var}(z)}\\ \langle e^z x\rangle&=e^{\langle z\rangle+\frac{1}{2}\mathrm{Var}(z)}\left(\langle x\rangle+\mathrm{Cov}(x,z)\right)\\ \langle e^z xy\rangle&=e^{\langle z\rangle+\frac{1}{2}\mathrm{Var}(z)}\Big(\mathrm{Cov}(x,y)+\left(\langle x\rangle+\mathrm{Cov}(x,z)\right)\left(\langle y\rangle+\mathrm{Cov}(y,z)\right)\Big)\end{aligned}$$

where (x,y,z) is joint Gaussian, we can compute

$$\begin{aligned}
\langle A_t\rangle &= \left\langle e^{-\Gamma_\gamma(t)}u_0\right\rangle = e^{-\bar{\gamma}(t-b_\gamma(t))}\left\langle e^{-b_\gamma(t)\gamma_0}u_0\right\rangle\left\langle e^{-\mathcal{B}_\gamma(t)}\right\rangle\\
\langle A_t\gamma_0\rangle &= \langle e^{-\Gamma_T}u_0\gamma_0\rangle = e^{-\bar{\gamma}(t-b_\gamma(t))}\left\langle e^{-b_\gamma(t)\gamma_0}u_0\gamma_0\right\rangle\left\langle e^{-\mathcal{B}_\gamma(t)}\right\rangle\\
\left\langle A_t\dot{\mathcal{B}}_\gamma(t)\right\rangle &= \left\langle e^{-\Gamma_T}u_0\dot{\mathcal{B}}_\gamma(t)\right\rangle = e^{-\bar{\gamma}(t-b_\gamma(t))}\left\langle e^{-b_\gamma(t)\gamma_0}u_0\right\rangle\left\langle e^{-\mathcal{B}_\gamma(t)}\dot{\mathcal{B}}_\gamma(t)\right\rangle\\
\langle A_t^2\rangle &= \langle e^{-2\Gamma_T}u_0^2\rangle = e^{-2\bar{\gamma}(t-b_\gamma(t))}\left\langle e^{-2b_\gamma(t)\gamma_0}u_0^2\right\rangle\left\langle e^{-2\mathcal{B}_\gamma(t)}\right\rangle\\
\langle B_t^2\rangle &= \sigma_u^2\int_0^t\left\langle e^{-2(\Gamma_\gamma(t)-\Gamma_\gamma(s))}\right\rangle ds\\
&= \sigma_u^2\int_0^t e^{-2\langle\Gamma_\gamma(t)-\Gamma_\gamma(s)\rangle+2\mathrm{Var}(\Gamma_\gamma(t)-\Gamma_\gamma(s))}ds
\end{aligned}$$

and thereby Equation (B.1). Here a numerical integration rule (we use the trapezoidal rule) can be employed for computation of $\langle B_t^2\rangle$. As a consequence, the analytic moment-mapping $(u_0,\gamma_0)\mapsto(u_t,\gamma_t)$ is obtained.

B.2. Moment generating function (MGF) of related integral process. Recall the DSM:

$$\textbf{(DSM)}\qquad\begin{cases} d\widehat{u} = -\widehat{\gamma}\widehat{u}\,dt+\sigma_u dB_u\\ d\widehat{\gamma} = -\frac{\nu}{\epsilon}(\widehat{\gamma}-\mu)\,dt+\frac{\sigma}{\sqrt{\epsilon}}dB_\gamma\end{cases}$$

and $\widehat{\Gamma}_t=\int_0^t\widehat{\gamma}(s)ds$. Then, from the results of preceding subsection, we have

$$\begin{aligned}
\langle\widehat{\Gamma}_t\rangle &= \mu t+\langle\widehat{\gamma}_0-\mu\rangle b_\gamma(t)\\
\mathrm{Var}(\widehat{\Gamma}_t) &= \mathrm{Var}(\widehat{\gamma}_0)b_\gamma(t)^2+\mathrm{Var}(\mathcal{B}_\gamma(t))\\
\langle\widehat{\Gamma}_t-\widehat{\Gamma}_s\rangle &= b_\gamma(t-s)\langle\widehat{\gamma}_s\rangle+\mu((t-s)-b_\gamma(t-s))\\
\mathrm{Var}(\widehat{\Gamma}_t-\widehat{\Gamma}_s) &= (b_\gamma(t-s))^2\mathrm{Var}(\widehat{\gamma}_s)+\mathrm{Var}\left(\mathcal{B}_\gamma(t-s)\right)\\
\langle\widehat{\gamma}_t\rangle &= \mu+(\langle\widehat{\gamma}_0\rangle-\mu)e^{-\nu t/\epsilon}\\
\mathrm{Var}(\widehat{\gamma}_t) &= e^{-2\nu t/\epsilon}\mathrm{Var}(\widehat{\gamma}_0)+\frac{\sigma^2}{2\nu}\left(1-e^{-2\nu t/\epsilon}\right)
\end{aligned}$$

where

$$\begin{aligned}
b_\gamma(t) &= \epsilon(1-e^{-\nu t/\epsilon)})/\nu\\
\mathrm{Var}(\mathcal{B}_\gamma(t)) &= -\epsilon^2\frac{\sigma^2}{2\nu^3}\left(3-4e^{-\nu t/\epsilon}+e^{-2\nu t/\epsilon}-2\nu t/\epsilon\right).
\end{aligned}$$

Let $\widehat{\gamma}_0$ be Gaussian so that $\widehat{\Gamma}_t$ is Gaussian as well, and the MGFs

$$\begin{aligned}
\left\langle e^{\alpha\widehat{\Gamma}_t}\right\rangle &= \exp\left(\alpha\langle\widehat{\Gamma}_t\rangle+\frac{\alpha^2}{2}\mathrm{Var}(\widehat{\Gamma}_t)\right)\\
\left\langle e^{\alpha(\widehat{\Gamma}_t-\widehat{\Gamma}_s)}\right\rangle &= \exp\left(\alpha\langle\widehat{\Gamma}_t-\widehat{\Gamma}_s\rangle+\frac{\alpha^2}{2}\mathrm{Var}(\widehat{\Gamma}_t-\widehat{\Gamma}_s)\right)
\end{aligned}\tag{B.2}$$

can be computed.

B.3. Asymptotics of MGFs of related integral process.

B.3.1. Scale-separation regime. For small ϵ, from substituting $b_\gamma(t)=\frac{1}{d}\epsilon+\mathcal{O}(\epsilon^2)$ and $\mathrm{Var}(\mathcal{B}_\gamma(t)-\mathcal{B}_\gamma(s))=\frac{\sigma^2}{d^2}(t-s)\epsilon+\mathcal{O}(\epsilon^2)$ into Equation (B.2), we obtain

$$\left\langle e^{\alpha\widehat{\Gamma}_t}\right\rangle=\exp\left(\alpha\left(\mu t+\langle\widehat{\gamma}_0-\mu\rangle\frac{\epsilon}{\nu}\right)+\alpha^2\frac{\sigma^2}{2\nu^2}t\epsilon+\mathcal{O}(\epsilon^2)\right)\qquad \epsilon<t$$
$$\left\langle e^{\alpha(\widehat{\Gamma}_t-\widehat{\Gamma}_s)}\right\rangle=\exp\left(\alpha\mu(t-s)+\alpha^2\frac{\sigma^2}{2\nu^2}(t-s)\epsilon+\mathcal{O}(\epsilon^2)\right)\qquad \epsilon<s.$$

B.3.2. Rare-event regime. When ϵ is large, we use $b_r(t)=t-\frac{1}{2}\nu t^2/\epsilon+\mathcal{O}(1/\epsilon^2)$ and $\mathrm{Var}(\mathcal{B}_\gamma(t))=\frac{\sigma^2}{3}t^3/\epsilon+\mathcal{O}(1/\epsilon^2)$ to obtain

$$\left\langle e^{\alpha\widehat{\Gamma}_t}\right\rangle=\exp\left(\alpha\left(\mu t+\langle\widehat{\gamma}_0-\mu\rangle\left(t-\frac{\nu}{2}\frac{t^2}{\epsilon}\right)\right)+\frac{\alpha^2}{2}\left(\mathrm{Var}(\widehat{\gamma}_0)\left(t^2-\nu\frac{t^3}{\epsilon}\right)+\frac{\sigma^2}{3}\frac{t^3}{\epsilon}\right)+\mathcal{O}\left(\frac{1}{\epsilon^2}\right)\right)$$

for DSM. Therefore, in case of dDSM, we have

$$\begin{aligned}\left\langle e^{\alpha\widehat{\Gamma}'_t}|\widehat{\gamma}'_0=\gamma_\pm\right\rangle&=\exp\left(\alpha\left(\mu_\pm t+(\gamma_\pm-\mu_\pm)\left(t-\frac{\nu_\pm}{2}\frac{t^2}{\epsilon}\right)\right)+\frac{\alpha^2}{2}\frac{(\sigma_\pm)^2}{3}\frac{t^3}{\epsilon}+\mathcal{O}\left(\frac{1}{\epsilon^2}\right)\right)\\&=\exp\left(\alpha\left(\gamma_\pm t-\frac{1}{2\epsilon}(\gamma_\pm-\mu_\pm)\nu_\pm t^2\right)+\frac{\alpha^2}{2}\frac{(\sigma_\pm)^2}{3\epsilon}t^3+\mathcal{O}\left(\frac{1}{\epsilon^2}\right)\right).\end{aligned}\tag{B.3}$$

Appendix C. Proofs of results. We collect together the proofs of results which underpin our understanding of the algorithms studied in this paper.

C.1. Scale-separation limit. We prove Lemma 3.1.

Proof. **(Proof of Lemma 3.1).** The convergence of the mean and variance follows from Equation (3.3) and the bounded convergence theorem.

To show $L^2(\Omega;\mathbb{R})$ convergence, from Equation (3.2), we obtain

$$u_T^\epsilon-\bar{u}_T=\left(e^{-\Gamma_T^\epsilon}u_0^\epsilon-e^{-\bar{\gamma}T}\bar{u}_0\right)+\sigma_u\int_0^T\left(e^{-(\Gamma_T^\epsilon-\Gamma_t^\epsilon)}-e^{-\bar{\gamma}(T-t)}\right)dB_u(t)$$

and

$$|u_T^\epsilon-\bar{u}_T|^2\le 2\left|e^{-\Gamma_T^\epsilon}u_0^\epsilon-e^{-\bar{\gamma}T}\bar{u}_0\right|^2+2\sigma_u^2\left|\int_0^T\left(e^{-(\Gamma_T^\epsilon-\Gamma_t^\epsilon)}-e^{-\bar{\gamma}(T-t)}\right)dB_u(t)\right|^2.$$

The use of the Itô lemma leads to

$$\langle|u_T^\epsilon-\bar{u}_T|^2\rangle\le 2\left\langle\left|e^{-\Gamma_T^\epsilon}u_0^\epsilon-e^{-\bar{\gamma}T}\bar{u}_0\right|^2\right\rangle+2\sigma_u^2\int_0^T\left\langle\left|e^{-(\Gamma_T^\epsilon-\Gamma_t^\epsilon)}-e^{-\bar{\gamma}(T-t)}\right|^2\right\rangle dt.$$

Note that the term

$$\left\langle e^{-2(\Gamma_T^\epsilon-\Gamma_t^\epsilon)}\right\rangle-2\left\langle e^{-(\Gamma_T^\epsilon-\Gamma_t^\epsilon)}\right\rangle e^{-\bar{\gamma}(T-t)}+e^{-2\bar{\gamma}(T-t)}\tag{C.1}$$

converges to zero as $\epsilon\to 0$. Then the bounded convergence theorem ensures the integration of the term (C.1) also converges to zero as $\epsilon\to 0$. This implies the convergence

$$\langle|u_T^\epsilon-\bar{u}_T|^2\rangle\to 0$$

as $\epsilon \to 0$ holds. □

Now we state and prove Lemma C.1 and Lemma C.2 which will be used to prove Lemma 3.2 and Lemma 3.3.

Lemma C.1. *Let $\mathcal{Y}$ be a Markov chain or a diffusion process associated with generator $\frac{1}{\epsilon}Q_0$. We assume $\mathcal{Y}$ is an ergodic process with invariant measure $\rho_{\mathcal{Y}}^{\infty}$ satisfying Null$(Q_0)=$span$\{\mathbf{1}\}$, Null$(Q_0^*)=$span$\{\rho_{\mathcal{Y}}^{\infty}\}$. Let $\mathcal{X}$ satisfy the ODE*

$$\frac{d\mathcal{X}}{dt}=f(\mathcal{X},\mathcal{Y}),$$

and let the generator of the combined process $(\mathcal{X},\mathcal{Y})$ be of the form

$$Q=\frac{1}{\epsilon}Q_0+Q_1.$$

Let $\bar{\mathcal{X}}$ satisfy the ODE

$$\frac{d\bar{\mathcal{X}}}{dt}=\bar{Q}_1(\bar{\mathcal{X}})=\int f(\bar{\mathcal{X}},\cdot)d\rho_{\mathcal{Y}}^{\infty}(\cdot). \tag{C.2}$$

Then, for any $t>0$, $\mathcal{X}(t)$ converges weakly or in distribution to $\bar{\mathcal{X}}(t)$ as $\epsilon \to 0$ (recall $X_\epsilon \rightharpoonup X$ is said to converge weakly provided $\mathbb{E}(f(X_\epsilon))\to\mathbb{E}(f(X))$ for any bounded continuous function f).

Proof. The first step is to show that the averaged ODE is given by Equation (C.2). Let be Φ be a bounded continuous function and let

$$v(x,y,t)=\mathbb{E}(\Phi(\mathcal{X}_t,\mathcal{Y}_t)|\mathcal{X}_0=x,\mathcal{Y}_0=y).$$

Then the backward equation

$$\partial_t v(x,y,t)=Qv(x,y,t)=\left(\frac{1}{\epsilon}Q_0+Q_1\right)v(x,y,t) \tag{C.3}$$

is satisfied. We seek solution $v=v(x,y,t)$ in the form of the multi-scale expansion

$$v=v_0+\epsilon v_1+\mathcal{O}(\epsilon^2).$$

From substituting the expansion and equating coefficients of equal powers of ϵ to zero, we find

$$\begin{aligned}\mathcal{O}\left(\frac{1}{\epsilon}\right): &\qquad Q_0v_0=0\\ \mathcal{O}(1): &\qquad Q_0v_1=-Q_1v_0+\frac{dv_0}{dt}\end{aligned} \tag{C.4}$$

and we see v_0 is independent of y due to null$(Q_0)=\mathbf{1}$. The operator Q_0 is singular and, for Equation (C.4) to have a solution, the Fredholm alternative implies the solvability condition

$$-Q_1v_0+\frac{dv_0}{dt}\perp \mathrm{Null}(Q_0^*).$$

For arbitrary $c(x)$, we find

$$\int\int c(x)\left(\frac{dv_0}{dt}-Q_1v_0\right)dxd\rho_\infty(y)=\int c(x)\left(\frac{dv_0}{dt}-\bar{Q}_1v_0\right)dx=0$$

which implies

$$\frac{dv_0}{dt}-\bar{Q}_1v_0=0.$$

The second step is to show the weak convergence. A substitution of

$$v=v_0+\epsilon v_1+r$$

into Equation (C.3) leads to

$$\frac{dr}{dt}=\left(\frac{1}{\epsilon}Q_0+Q_1\right)r+\epsilon q$$
$$q=Q_1v_1-\frac{dv_1}{dt}$$

and

$$r(t)=e^{Qt}r(0)+\epsilon\int_0^t e^{Q(t-s)}q(s)ds$$

from the variation-of-constants. Because Φ is bounded, $|e^{Qt}|_\infty\le 1$ is satisfied from $v(t)=e^{Qt}v(0)$. We then have

$$\begin{aligned}|r(t)|_\infty &\le \epsilon|e^{Qt}|_\infty|r(0)|_\infty+\epsilon\int_0^t|e^{Q(t-s)}|_\infty|q(s)|_\infty ds\\ &\le \epsilon|v_1(0)|_\infty+\epsilon\int_0^t|q(s)|_\infty ds\\ &\le \epsilon\left(|v_1(0)|_\infty+t\sup_{0\le s\le t}|q(s)|_\infty\right)\end{aligned}$$

and we obtain

$$|v(t)-v_0(t)|_\infty\le C(T)\epsilon$$

for $0\le t\le T$. □

LEMMA C.2. *Let $F_{X_\epsilon}(\cdot)\equiv\mathbb{P}(X_\epsilon\le\cdot)$ and F_X be the distribution functions of X_ϵ and non-random variable X, respectively. If*

$$\lim_{b\to\infty}e^{\alpha b}(F_{X_\epsilon}(b)-1)\to 0 \tag{C.5a}$$

$$\lim_{a\to-\infty}e^{\alpha a}F_{X_\epsilon}(a)\to 0 \tag{C.5b}$$

$$\lim_{a\to-\infty,b\to\infty}\int_a^b(F_{X_\epsilon}(x)-F_X(x))e^{\alpha x}dx\to 0 \tag{C.5c}$$

as $\epsilon\to 0$, then $\left\langle e^{\alpha X_\epsilon}\right\rangle\to e^{\alpha X}$ follows. The convergence rate is given by the lowest one in Equation (C.5).

Proof. This follows from

$$\begin{aligned}
\left\langle e^{\alpha X_\epsilon}\right\rangle &= \lim_{a\to-\infty,b\to\infty}\int_a^b e^{\alpha x}dF_{X_\epsilon}(x)\\
&= \lim_{a\to-\infty,b\to\infty} e^{\alpha x}F_{X_\epsilon}(x)|_a^b - \int_a^b F_{X_\epsilon}(x)\alpha e^{\alpha x}dx\\
&= \lim_{b\to\infty} e^{\alpha b}(F_{X_\epsilon}(b)-1) - \lim_{a\to-\infty} e^{\alpha a}F_{X_\epsilon}(a) + e^{\alpha X}\\
&\quad - \lim_{a\to-\infty,b\to\infty}\int_a^b (F_{X_\epsilon}(x)-F_X(x))\alpha e^{\alpha x}dx
\end{aligned}$$

where integration by parts is used. □

Proof. **(Proof of Lemma 3.2.)** For a bounded continuous function Φ, let

$$v(x,y_i,t) = \mathbb{E}(\Phi(\Gamma_t,\gamma_t)|\Gamma_0 = x, \gamma_0 = y_i)$$

where $y_1=\gamma_+$ and $y_2=\gamma_-$. Then the backward equation

$$\partial_t v(x,y_i,t) = \sum_j L_{ij}v(x,y_j,t) + y_i\partial_x v(x,y_i,t)$$

—alternatively

$$\begin{aligned}
&\partial_t v(x,t) = Qv(x,t)\\
&\qquad Q = \frac{1}{\epsilon}Q_0 + Q_1 = \frac{1}{\epsilon}\begin{pmatrix} -\lambda_+ & \lambda_+\\ \lambda_- & -\lambda_-\end{pmatrix} + \begin{pmatrix} y_1\partial_x & 0\\ 0 & y_2\partial_x\end{pmatrix}
\end{aligned}$$

in vector notation—is satisfied. The generator of γ is given by

$$L = \frac{1}{\epsilon}\begin{pmatrix} -\lambda_+ & \lambda_+\\ \lambda_- & -\lambda_-\end{pmatrix}$$

and γ is ergodic process [36].

From Equation (A.9), the time invariant measure of γ is

$$\rho_\gamma^\infty = \frac{1}{\lambda_- + \lambda_+}\begin{pmatrix}\lambda_-\\ \lambda_+\end{pmatrix}$$

or

$$\rho_\gamma^\infty \triangleq \frac{\lambda_-}{\lambda_-+\lambda_+}\delta_{\gamma_+} + \frac{\lambda_+}{\lambda_-+\lambda_+}\delta_{\gamma_-}$$

on $\mathbb{R}$. An averaging of

$$\frac{d\Gamma}{dt} = \gamma$$

yields

$$\frac{d\bar{\Gamma}}{dt} = \int \gamma d\rho_\gamma^\infty = \frac{\lambda_-\gamma_+ + \lambda_+\gamma_-}{\lambda_-+\lambda_+} \equiv \gamma_\infty.$$

Let

$$v_0(x,t) = \mathbb{E}(\phi(\bar{\Gamma}_t)|\bar{\Gamma}_0 = x)$$

where $\phi(\cdot) = \Phi(\cdot, y)$. Then

$$\partial_t v_0(x,t) = \gamma_\infty \partial_x v_0(x,t) = \bar{Q}_1 v_0(x,t)$$

and Lemma C.1 ensures $v(x, y_i, t) \to v_0(x,t)$ as $\epsilon \to 0$. In this case, the weak convergence of Γ_t to $\gamma_\infty t$ implies $\Gamma_T - \Gamma_t \rightharpoonup \gamma_\infty(T-t)$ from Slutsky's theorem, which states that if $X_\epsilon \rightharpoonup X$ and $Y_\epsilon \rightharpoonup Y$ as $\epsilon \to 0$, where Y is non-random, then $X_\epsilon + Y_\epsilon \rightharpoonup X + Y$ as $\epsilon \to 0$.

Let the distribution function of $\Gamma_T - \Gamma_t$ be denoted by $F_{\Gamma_T - \Gamma_t}(x) \equiv \mathbb{P}(\Gamma_T - \Gamma_t \leq x)$. Then

$$F_{\Gamma_T - \Gamma_t}(x) = \left\{ \begin{array}{lll} 0 & \text{for} & x < \gamma_-(T-t) \\ 1 & \text{for} & x \geq \gamma_+(T-t) \end{array} \right. .$$

Taking $a < \gamma_-(T-t)$ and $b > \gamma_+(T-t)$, Equations (C.5a), (C.5b) are satisfied. Note $\Gamma_T - \Gamma_t \rightharpoonup \gamma_\infty(T-t)$ is equivalent to $F_{\Gamma_T - \Gamma_t}(x) \to F_{\gamma_\infty(T-t)}(x)$ for every x that is continuity point of $F_{\gamma_\infty(T-t)}$, given by

$$F_{\gamma_\infty(T-t)}(x) = \left\{ \begin{array}{lll} 0 & \text{for} & x < \gamma_\infty(T-t) \\ 1 & \text{for} & x \geq \gamma_\infty(T-t) \end{array} \right.$$

from the Lévy-Cramér continuity theorem. Then Equation (C.5c) is satisfied from the bounded convergence theorem and Lemma C.2 ensures the convergence of the MGF. The convergence rate of $\left\langle e^{\alpha(\Gamma_T - \Gamma_t)} \right\rangle \to e^{\alpha \gamma_\infty (T-t)}$ is determined by the following convergence:

$$\lim_{\epsilon \to 0} \int_{\gamma_-(T-t)}^{\gamma_+(T-t)} \left(F_{\Gamma_T - \Gamma_t}(x) - F_{\gamma_\infty(T-t)}(x) \right) e^{\alpha x} dx = 0.$$

□

Proof. **(Proof of Lemma 3.3.)** The generator of the system

$$\left\{ \begin{array}{ll} d\widehat{\Gamma} & = \widehat{\gamma} dt \\ d\widehat{\gamma} & = -\frac{1}{\epsilon} \nabla U(\widehat{\gamma}) dt + \frac{1}{\sqrt{\epsilon}} \beta(\widehat{\gamma}) dB_\gamma \end{array} \right.$$

is given by

$$y \partial_x + \frac{1}{\epsilon} \left(-\nabla U(y) \partial_y + \frac{1}{2} \beta(y)^2 \partial_y^2 \right) = Q_1 + \frac{1}{\epsilon} Q_0.$$

If $\widehat{\gamma}$ is an ergodic process with invariant measure $\rho_{\widehat{\gamma}}^\infty$, then Lemma C.1 ensures $\widehat{\Gamma}(t) \rightharpoonup \overline{\widehat{\Gamma}}(t)$, which solves

$$\frac{d\overline{\widehat{\Gamma}}}{dt} = \int \widehat{\gamma} d\rho_{\widehat{\gamma}}^\infty.$$

In the case of DSM, the generator is given by

$$y \partial_x + \frac{1}{\epsilon} \left(-\nu(y - \mu) \partial_y + \frac{\sigma^2}{2} \partial_y^2 \right) = Q_1 + \frac{1}{\epsilon} Q_0$$

and the invariant measure for $\widehat{\gamma}$ is

$$\rho^{\infty}_{\widehat{\gamma}} = \mathcal{N}\left(\mu, \frac{\sigma^2}{2\nu}\right)$$

because it solves $Q_0^* \rho^{\infty}_{\widehat{\gamma}} = 0$. Therefore we obtain $\widehat{\Gamma}_t \rightharpoonup \mu t$ and further $\widehat{\Gamma}_T - \widehat{\Gamma}_t \rightharpoonup \mu(T-t)$ from Slutsky's theorem.

Since $\widehat{\gamma}$ is a Gaussian process, we use the Chernoff bound $F_{\mathcal{N}(0,1)}(x) \leq \frac{1}{2} e^{-x^2/2}$ to meet Equations (C.5a), (C.5b). Note Equation (C.5c) is satisfied from the bounded convergence theorem. As a consequence, the convergence of MGF follows. The analysis of the convergence rate is the same as in the SSM case. □

C.2. Rare-event limit.

Proof. **(Proof of Lemma 3.4.)** This follows from

$$\begin{aligned}
\langle u_t^\epsilon | \gamma_0^\epsilon = \gamma_\pm \rangle &= \left\langle e^{-\Gamma_t^\epsilon} u_0^\epsilon | \gamma_0^\epsilon = \gamma_\pm \right\rangle \\
\mathrm{Var}(u_t^\epsilon | \gamma_0^\epsilon = \gamma_\pm) &= \left\langle \left(e^{-\Gamma_t^\epsilon} u_0^\epsilon \right)^2 | \gamma_0^\epsilon = \gamma_\pm \right\rangle - \left\langle e^{-\Gamma_t^\epsilon} u_0^\epsilon | \gamma_0^\epsilon = \gamma_\pm \right\rangle^2 \\
&\quad + \sigma_u^2 \int_0^t \left\langle e^{-2(\Gamma_t^\epsilon - \Gamma_s^\epsilon)} | \gamma_0^\epsilon = \gamma_\pm \right\rangle ds.
\end{aligned}$$

□

Proof. **(Proof of Lemma 3.5.)** We take $\epsilon \to \infty$ in Equation (A.3) and use Theorem A.1. In view of the approximation (A.14), this corresponds to the case of $t = 0$. We invoke Equation (A.11) to complete the proof. □

Proof. **(Proof of Lemma 3.6.)** We take $\epsilon \to \infty$ in Equation (B.3) for the case of $t = 0$. Direct computation of Equation (B.2) ensures the result. □

REFERENCES

[1] B. Anderson and J. Moore, *Optimal Filtering*, Prentice–Hall Englewood Cliffs, NJ, 1979.

[2] A.F. Bennett, *Inverse Methods in Physical Oceanography*, Cambridge University Press, 1992.

[3] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, American Mathematical Society, 2011.

[4] C.M. Bishop et al., *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[5] T. Bohr, M. Jensen, G. Paladin, and A. Vulpiani, *Dynamical Systems Approach to Turbulence*, Cambridge University Press, 2005.

[6] M. Branicki, B. Gershgorin, and A. Majda, *Filtering skill for turbulent signals for a suite of nonlinear and linear extended Kalman filters*, J. Comput. Phys., 231:1462–1498, 2012.

[7] M. Branicki and A. Majda, *Dynamic stochastic superresolution of sparsely observed turbulent systems*, J. Comput. Phys., 241(5):333–363, 2013.

[8] E. Castronovo, J. Harlim, and A. Majda, *Mathematical test criteria for filtering complex systems: plentiful observations*, J. Comput. Phys., 227:3678–3714, 2008.

[9] R. Chen and J. Liu, *Mixture Kalman filters*, J. R. Stat. Soc. Ser. B., Stat. Methodol., 62:493–508, 2000.

[10] D. Cioranescu and P. Donato, *Introduction to Homogenization*, Oxford University Press, 1999.

[11] D. Crouse, P. Willett, K. Pattipati, and L. Svensson, *A look at gaussian mixture reduction algorithms*, International Conference on Information Fusion, 1–8, July 2011.

[12] A. Doucet, N. De Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer–Verlag, 2001.

[13] G. Evensen, *Data Assimilation: the Ensemble Kalman Filter*, Springer–Verlag, 2009.

[14] U. Frisch, *Turbulence*, Cambridge University Press, 1996.

[15] A. Gelb, *Applied Optimal Estimation*, MIT Press, 1974.

[16] B. Gershgorin, J. Harlim, and A. Majda, *Test models for improving filtering with model errors through stochastic parameter estimation*, J. Comput. Phys., 229:1–31, 2010.

[17] B. Gershgorin, J. Harlim, and A.J. Majda, *Improving filtering and prediction of spatially extended turbulent systems with model errors through stochastic parameter estimation*, J. Comput. Phys., 229:32–57, 2010.

[18] B. Gershgorin and A. Majda, *A nonlinear test model for filtering slow-fast systems*, Commun. Math. Sci., 6:611–649, 2008.

[19] Z. Ghahramani and G.E. Hinton, *Variational learning for switching state-space models*, Neural Comput., 12:831–864, 2000.

[20] N. Gordon, D. Salmond, and A. Smith, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation*, in Radar and Signal Processing, IEE Proceedings F, IET, 140:107–113, 1993.

[21] J. Harlim and A. Majda, *Filtering nonlinear dynamical systems with linear stochastic models*, Nonlinearity, 21(227):1281-1306, 2008.

[22] J. Harlim and A.J. Majda, *Test models for filtering with superparameterization*, Multiscale Model. Simul., 11:282–308, 2013.

[23] A. Jazwinski, *Stochastic processes and filtering theory*, San Diego, California: Mathematics in Science and Engineering, 64(2):1730-1730, 1970.

[24] R. Kalman et al., *A new approach to linear filtering and prediction problems*, Journal of Basic Engineering, 82:35–45, 1960.

[25] R.E. Kalman and R.S. Bucy, *New results in linear filtering and prediction theory*, J. Basic Eng.-T ASME, 83:95–108, 1961.

[26] E. Kalnay, *Atmospheric Modeling, Data Assimilation, and Predictability*, Cambridge University Press, 2003.

[27] S.R. Keating, A.J. Majda, and K.S. Smith, *New methods for estimating poleward eddy heat transport using satellite altimetry*, Mon. Wea. Rev.,14(5): 1703–1722, 2011.

[28] H. Kushner, *Approximations to optimal nonlinear filters*, Automatic Control, IEEE Transactions on, 12:546–556, 1967.

[29] K. Law, A. Stuart, and K. Zygalakis, *Data Assimilation: A Mathematical Introduction*, Springer, 2015.

[30] A.J. Majda and J. Harlim, *Filtering Complex Turbulent Systems*, Cambridge University Press, 2012.

[31] A.J. Majda, J. Harlim, and B. Gershgorin, *Mathematical strategies for filtering turbulent dynamical systems*, Discrete Contin. Dyn. Syst., 27:441–486, 2010.

[32] X. Mao and C. Yuan, *Stochastic differential equations with Markovian switching*, World Scientific, 7(3):275, 2006.

[33] P.S. Maybeck, *Stochastic Models, Estimation, and Control*, Academic Press, 1982.

[34] K.P. Murphy, *Machine Learning: a Probabilistic Perspective*, MIT Press, 2012.

[35] D.S. Oliver, A.C. Reynolds, and N. Liu, *Inverse Theory for Petroleum Reservoir Characterization and History Matching*, Cambridge University Press, 2008.

[36] G. Pavliotis and A. Stuart, *Multiscale Methods: Averaging and Homogenization*, Springer, 2008.

[37] S. Reich and C. Cotter, *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press, 2015.

[38] T.P. Sapsis and A.J. Majda, *A statistically accurate modified quasilinear Gaussian closure for uncertainty quantification in turbulent dynamical systems*, Phys. D, 252(5):34-45, 2013.

[39] H.W. Sorenson and D.L. Alspach, *Recursive bayesian estimation using gaussian sums*, Automatica, 7:465–479, 1971.

[40] A. Stordal, H. Karlsen, G. Nævdal, H. Skaug, and B. Vallès, *Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter*, Comput. Geosci., 15:293–305, 2011.

[41] J. Walter and C. Schütte, *Conditional averaging for diffusive fast-slow systems: a sketch for derivation*, in Analysis, Modeling and Simulation of Multiscale Problems, Springer, 647–682, 2006.

[42] V. Zakharov, V. L'vov, and G. Falkovich, *Kolmogorov spectra of turbulence I. wave turbulence*, Springer–Verlag Series in Nonlinear Dynamics, Springer, Berlin (Germany), 1992.