Conditional Gaussian nonlinear system: A fast preconditioner and a cheap surrogate model for complex nonlinear systems

Cite as: Chaos **32**, 053122 (2022); https://doi.org/10.1063/5.0081668
Submitted: 09 December 2021 • Accepted: 02 May 2022 • Published Online: 17 May 2022

Nan Chen, Yingda Li and Honghu Liu

Conditional Gaussian nonlinear system: A fast preconditioner and a cheap surrogate model for complex nonlinear systems

Cite as: Chaos **32**, 053122 (2022); doi: 10.1063/5.0081668
Submitted: 9 December 2021 · Accepted: 2 May 2022 ·
Published Online: 17 May 2022

View Online Export Citation CrossMark

Nan Chen,[1,a)] Yingda Li,[1,b)] and Honghu Liu[2,c)]

AFFILIATIONS

[1]Department of Mathematics, University of Wisconsin-Madison, Madison, Wisconsin 53705, USA
[2]Department of Mathematics, Virginia Tech, Blacksburg, Virginia 24061, USA

Note: This article is part of the Focus Issue, Theory-informed and Data-driven Approaches to Advance Climate Sciences.
[a)]**Author to whom correspondence should be addressed:** chennan@math.wisc.edu
[b)]yli678@wisc.edu
[c)]hhliu@vt.edu

ABSTRACT

Developing suitable approximate models for analyzing and simulating complex nonlinear systems is practically important. This paper aims at exploring the skill of a rich class of nonlinear stochastic models, known as the conditional Gaussian nonlinear system (CGNS), as both a cheap surrogate model and a fast preconditioner for facilitating many computationally challenging tasks. The CGNS preserves the underlying physics to a large extent and can reproduce intermittency, extreme events, and other non-Gaussian features in many complex systems arising from practical applications. Three interrelated topics are studied. First, the closed analytic formulas of solving the conditional statistics provide an efficient and accurate data assimilation scheme. It is shown that the data assimilation skill of a suitable CGNS approximate forecast model outweighs that by applying an ensemble method even to the perfect model with strong nonlinearity, where the latter suffers from filter divergence. Second, the CGNS allows the development of a fast algorithm for simultaneously estimating the parameters and the unobserved variables with uncertainty quantification in the presence of only partial observations. Utilizing an appropriate CGNS as a preconditioner significantly reduces the computational cost in accurately estimating the parameters in the original complex system. Finally, the CGNS advances rapid and statistically accurate algorithms for computing the probability density function and sampling the trajectories of the unobserved state variables. These fast algorithms facilitate the development of an efficient and accurate data-driven method for predicting the linear response of the original system with respect to parameter perturbations based on a suitable CGNS preconditioner.

Analyzing and simulating complex nonlinear systems is often very challenging due to high-dimensionality, multiscale features, and strong nonlinear interactions between different state variables. Therefore, developing suitable approximate models is a practically important topic to advance the understanding and prediction of these complex systems. In this paper, the focus is on introducing a stochastic nonlinear modeling framework, known as the conditional Gaussian nonlinear system (CGNS), that can be used as suitable approximate models for many complex nonlinear systems. One key feature of the CGNS is that closed analytic formulas are available for solving the conditional statistics, which facilitate the development of rigorous mathematical analysis and efficient numerical algorithms for handling such nonlinear systems. Different from many purely data-driven models, the CGNS preserves the underlying physics to a large extent. In addition, the nonlinear nature of the CGNS allows it to reproduce the observed intermittency, extreme events, and other non-Gaussian features in many complex systems arising from practical applications. In addition to playing the role of a cheap surrogate model, the CGNS can also be served as a fast preconditioner for facilitating many computationally challenging tasks of the original complex nonlinear systems. The advantages of the CGNS as both a cheap surrogate model and a fast preconditioner are demonstrated in several important broad applications, including data assimilation and ensemble forecast, parameter estimation in the presence of only partial observations, and

efficiently predicting the model response due to internal or external perturbations.

I. INTRODUCTION

Complex nonlinear systems are ubiquitous in many areas, including geophysics, climate science, engineering, neuroscience, and material science.[52,117,120,126,131] Mathematical modeling plays an important role in characterizing and discovering the underlying physics of these complex systems.[41,84] With suitable mathematical models in hand, effective parameter inference, state estimation, and data assimilation become fundamental tasks that serve as the prerequisites for analyzing these systems.[3,53,65,74,92] Accurate forecasts of future states and successful prediction of the system response due to external perturbations are also central topics that have many practical implications.[75,81,89,110,124]

However, there exist quite a few mathematical and computational challenges in analyzing and simulating complex nonlinear systems. First, the intrinsic nonlinearity in these complex nonlinear systems often triggers strongly chaotic or turbulent behavior.[39,106,112] As a consequence, intermittency, extreme events, and non-Gaussian probability density functions (PDFs) are some of the typical features in these systems,[44,82,96,98,125] which impede the use of many traditional mathematical tools to analyze the model properties. Second, due to the nonlinear interactions between state variables across different scales, many of these complex nonlinear systems are high dimensional and have multiscale spatiotemporal structures.[85,90,123,131] Therefore, developing new efficient numerical algorithms to accelerate the computational efficiency becomes essential. Particularly, enhancing the computational efficiency by reducing the complexity of these systems via effective stochastic parameterizations is practically important and is widely used in, for example, climate sciences. Third, it is often the case in practice that only partial observations of the state variables are available,[65,73] which result in additional difficulties for model calibration, state estimation, and prediction where systematic uncertainty quantification needs to be addressed.[35,36,40,86,87]

Since many complex dynamical systems are too expensive to be handled directly, it is of practical importance to develop suitable approximate models, which capture certain features of nature and are easier to deal with. Exploiting systematic reduced order modeling strategies and effective (stochastic) parameterizations is often a prerequisite for the development of approximate models. Linear regression models are arguably the simplest class of approximate models,[46,136] which can already provide certain skills for short-term forecasts although they usually suffer in characterizing the underlying nonlinear physics. Physics-constrained regression models are a set of nonlinear approximate models,[59,69,93] which take into account the energy conserving nonlinear interactions in the model development that guarantees the well-posedness of long-term behavior of the system. Another commonly used approach to developing approximate models is to project the starting complex nonlinear system to the leading a few energetic modes in light of the Galerkin proper orthogonal decomposition methods[63] or other empirical basis functions such as the principal interaction patterns[60,72] and the dynamic mode decomposition.[111,116] With a careful design of the closure terms to compensate for the truncation error, these reduced order models are skillful in resolving certain problems in fluids and turbulence.[8,104,121,135]

Meanwhile, many data-driven approximate modeling strategies have recently been developed.[1,14,16,61,76,101,109,118] One of them is the sparse identification of nonlinear dynamical systems (SINDy),[7] which leads to nonlinear regression models with parsimonious structures via sparse regression and compressed sensing. Many other approximate modeling approaches have also been designed for specific scientific purposes. For example, the past noise forecasting method[17] was developed as a data-driven forecast model for stochastic climate processes that exhibit low-frequency variability. Reduced-space Gaussian process regression forecast[129] was designed for data-driven probabilistic forecast of chaotic dynamical systems. Small-scale parameterization based on a data-informed optimal homotopic deformation of invariant manifolds was developed to design low-dimensional models for both deterministic chaotic systems and stochastic systems.[13,15] Physically consistent data-driven weather forecasting techniques were proposed and applied to operational models.[10,11] Recently, a strong link between the stochastic parameterization approach based on perturbation expansions of the Koopman operator[134] and the data-driven empirical model reduction (EMR) methodology[70] was established in Ref. 114. In addition, machine learning methods nowadays have been extensively incorporated into the reduced order models to further improve the approximation and forecast skill.[12,20,99,108,113]

The objective of this paper is to explore the skill of a rich class of nonlinear stochastic models, known as the "conditional Gaussian nonlinear system" (CGNS),[21] as approximate models for complex nonlinear systems. The CGNS includes many physics-constrained nonlinear stochastic models (e.g., the stochastic versions of various Lorenz models, low-order models of Charney–DeVore flows, and a paradigm model for topographic mean flow interaction), quite a few stochastically coupled reaction–diffusion models in neuroscience and ecology [e.g., stochastically coupled FitzHugh–Nagumo (FHN) models and stochastically coupled susceptible-infectious-removed (SIR) epidemic models], and several large-scale dynamical models in engineering and geophysical flows (e.g., the Boussinesq equations with noise and stochastically forced rotating shallow water equation). See Ref. 21 for a gallery of examples of the CGNS. The CGNS has also been applied to modeling and forecasting several important climate phenomena, such as the Madden–Julian oscillation and the monsoon,[26,27] and has been utilized for Lagrangian data assimilation.[28] Yet, most of the previous work focused on perfect model scenarios, where utilizing the CGNS as an approximate model has not been systematically studied.

The CGNS has several unique features that allow it to be distinct from many existing approximate modeling strategies. First, the CGNS aims at preserving the underlying physical mechanism to the greatest extent. Specifically, the nonlinearity involving the large-scale or slow variables is by design retained, which includes not only the self-interactions among the large-scale variables but also the cross-scale interactions between large- and small-scale variables, while suitable approximations are imposed primarily on the nonlinear self-interactions between small-scale, fast-varying or unresolved state variables via effective stochastic parameterizations. This is fundamentally different from many purely data-driven

nonlinear regression models, which may miss certain crucial underlying physics of the original complex nonlinear systems. Second, the stochastic parameterizations of the self-interactions between small-scale variables lead to an important feature of the CGNS. That is, the distribution of the small-scale variables conditioned on the large-scale ones is Gaussian. One remarkable consequence is that the associated conditional Gaussian distribution can be calculated using closed analytic formulas,[77] which considerably facilitate the mathematical analysis and numerical simulations of the CGNS. In fact, the closed analytic formulas of the conditional Gaussian distributions allow the development of efficient and statistically accurate algorithms for parameter estimation, data assimilation, and ensemble forecast in light of only partial observations. Note that, despite the conditional Gaussianity, the joint and marginal distributions of the CGNS remain highly non-Gaussian. Thus, the intermittency, extreme events, and turbulent features can all be preserved in a suitably designed CGNS. Third, the CGNS is also adaptable to many data-driven scenarios. Physics constraints, localizations, and sparse identification together with many other mathematical and computational strategies can be possibly incorporated into the CGNS. Finally, information theory[67,83] can be applied to quantify the uncertainty and the statistical error of the CGNS in approximating the original complex nonlinear systems.

The specific goal of this paper is twofold. First, by taking advantage of its analytically solvable properties, the CGNS can be served as a fast preconditioner for facilitating many computationally challenging tasks associated with the original complex nonlinear system. Important applications include estimating the parameters of the original system in the presence of only partial observations and recovering the non-Gaussian PDFs as a crucial intermediate step for computing the model sensitivity and response. Second, the CGNS is exploited as a surrogate model by exploiting systematic reduced order modeling strategies and suitable stochastic parameterizations, aiming at spending a much lower computational cost to create comparably accurate results as those obtained from the original complex system. This includes, for example, the state estimation of unobserved variables and the statistical forecast. Figure 1 shows a schematic illustration of utilizing the CGNS as a fast preconditioner and a cheap surrogate model for a general nonlinear system.

The rest of the paper is organized as follows. The general mathematical framework of CGNS and its analytic properties are described in Sec. II. Several systematic strategies for the development of the CGNS are included in Sec. III. Sections IV–VI consist of three important tasks in complex nonlinear systems, showing the roles of the CGNS both as a surrogate model and a preconditioner for the original complex system. Specifically, Sec. IV focuses on the data assimilation and ensemble forecast, Sec. V aims at efficient parameter estimation, and Sec. VI illustrates the use of CGNS in facilitating the study of model sensitivity and response theory. The paper is concluded in Sec. VII.

II. GENERAL MATHEMATICAL FRAMEWORK OF THE CGNS

Let us start with the general formulation of the turbulent dynamical systems motivated from fluid and geophysical applications,[65,85,112,126]

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = (L + D)\,\mathbf{u} + B\,(\mathbf{u}, \mathbf{u}) + \mathbf{F}\,(t) + \boldsymbol{\sigma}\,(\mathbf{u}, t)\,\dot{\mathbf{W}}\,(t)\,, \qquad (1)$$

where the state variable $\mathbf{u} \in \mathbb{C}^N$ is in a high dimensional phase space. In (1), the first two components, $(L + D)\,\mathbf{u}$, represent linear dispersion and dissipation effects, where $L^* = -L$ is a skew-symmetric operator (with $\cdot^*$ being the complex conjugate transpose), and D is a negative-definite matrix. The nonlinear effect is introduced through an energy-conserving quadratic form, $B\,(\mathbf{u}, \mathbf{u})$. In addition, the system is subject to external forcing effects that are decomposed into a deterministic component, $\mathbf{F}\,(t)$, and a stochastic component represented by a Gaussian random process, $\boldsymbol{\sigma}\,(\mathbf{u}, t)\,\dot{\mathbf{W}}\,(t)$, where

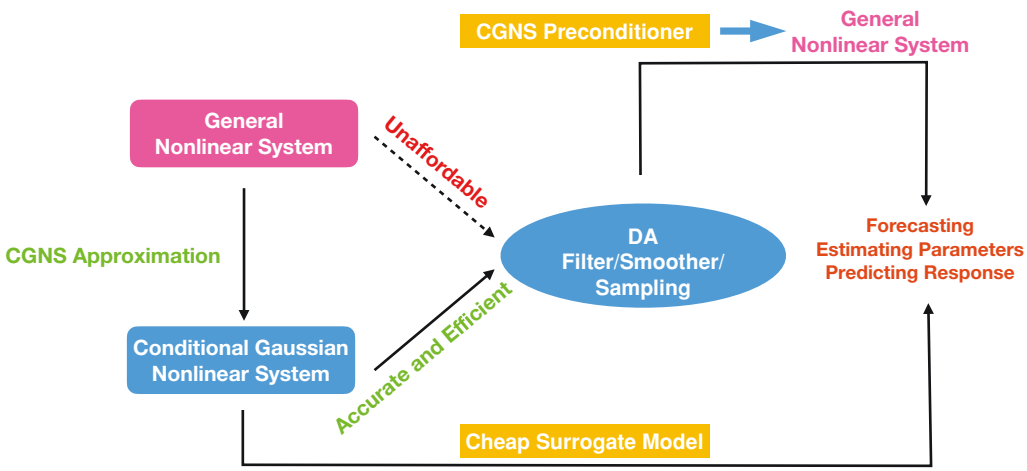

FIG. 1. A schematic illustration of utilizing the CGNS as a fast preconditioner and a cheap surrogate model for a general nonlinear system.

$\boldsymbol{\sigma} \in \mathbb{C}^{N\times K}$ is the noise matrix and $\dot{\mathbf{W}} \in \mathbb{C}^K$ is the white noise. The two components $(L+D)\,\mathbf{u} + B\,(\mathbf{u},\mathbf{u}) + \mathbf{F}\,(t)$ and $\boldsymbol{\sigma}\,(\mathbf{u},t)$ on the right hand side of (1) are also known as the drift part and the diffusion coefficients, respectively.

A. The CGNS

Despite being highly nonlinear and possessing strongly non-Gaussian statistics in both the marginal and joint distributions of the state $\mathbf{u}$, many complex nonlinear dynamical systems (1) have or can be approximated by the following nonlinear system with conditional Gaussian structures. The general mathematical framework of the CGNS is as follows:[21,22,77]

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = [\mathbf{A_0}(\mathbf{X},t) + \mathbf{A_1}(\mathbf{X},t)\mathbf{Y}(t)] + \mathbf{B_1}(\mathbf{X},t)\dot{\mathbf{W}}_\mathbf{1}(t), \tag{2a}$$

$$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}t} = [\mathbf{a_0}(\mathbf{X},t) + \mathbf{a_1}(\mathbf{X},t)\mathbf{Y}(t)] + \mathbf{b_2}(\mathbf{X},t)\dot{\mathbf{W}}_\mathbf{2}(t), \tag{2b}$$

where the original model state $\mathbf{u}$ is decomposed into multi-dimensional state variables $\mathbf{X} \in \mathbb{C}^{N_1}$ and $\mathbf{Y} \in \mathbb{C}^{N_2}$, with $N_1 + N_2 = N$. In (2), $\mathbf{A}_0, \mathbf{a}_0, \mathbf{A}_1, \mathbf{a}_1, \mathbf{B}_1$, and $\mathbf{b}_2$ are vectors or matrices that can depend nonlinearly on the state variables $\mathbf{X}$ and time t, while $\dot{\mathbf{W}}_\mathbf{1}$ and $\dot{\mathbf{W}}_\mathbf{2}$ are independent white noise sources that can have different dimensions from $\mathbf{X}$ and $\mathbf{Y}$. Typically, the decomposition of $\mathbf{u}$ into $\mathbf{X}$ and $\mathbf{Y}$ is organized so that $\mathbf{X}$ is the projection of $\mathbf{u}$ onto some suitable subspace that captures the large-scale dynamics of $\mathbf{u}$, and $\mathbf{Y}$ denotes the small-scale dynamics that is orthogonal to $\mathbf{X}$.

The name "conditional Gaussian" comes from the fact that once a time series of $\mathbf{X}(s)$ for $s \le t$ is given, then the conditional distribution $p(\mathbf{Y}(t)|\mathbf{X}(s \le t))$ is Gaussian. This can be seen by noticing that, with a given $\mathbf{X}$, the process of $\mathbf{Y}$ is linear (with respect to the variable $\mathbf{Y}$ itself since $\mathbf{X}$ has been given) with Gaussian white noises. It is worthwhile to highlight that, from the general form of the complex turbulent system (1) to the CGNS (2), the nonlinear self-interactions of $\mathbf{X}$ and the cross-interactions between $\mathbf{X}$ and $\mathbf{Y}$ in (1) are both completely retained. The only simplification in the CGNS is to approximate the nonlinear self-interactions between $\mathbf{Y}$ by a combination of nonlinear functions of $\mathbf{X}$, conditional linear functions of $\mathbf{Y}$, and effective stochastic noises. Nevertheless, if $\mathbf{Y}$ represents small-scale or fast variables, then such a manipulation is expected to be an effective approximation that preserves the underlying physics to a large extent.

It should be noted that the CGNS in (2) is still highly nonlinear due to the nonlinearity in $\mathbf{A}_0, \mathbf{a}_0, \mathbf{A}_1$, and $\mathbf{a}_1$ as well as the nonlinear coupling between the latter two with $\mathbf{Y}$. Such nonlinearities preserve the non-Gaussian statistics in (1) and allow us to reproduce many observed features in nature such as extreme events with the more tractable conditional Gaussian structure. A gallery of examples of the CGNS, including many physics-constrained nonlinear stochastic models, quite a few stochastically coupled reaction–diffusion models in neuroscience and ecology, and some large-scale dynamical models in engineering and geophysical flows can be found in Ref. 21.

Despite being highly nonlinear and non-Gaussian, one of the important features of the CGNS (2) is that the conditional distribution of $\mathbf{Y}$ given one realization of the time series $\mathbf{X}$ can be solved via closed analytic formulas. Such a unique analytic property significantly facilitates the analysis and calculations of state estimation, data assimilation, and forecast. This feature also makes the CGNS to be quite different from general nonlinear or non-Gaussian systems. For the latter, particle methods have to be applied for state estimation,[9,62,105] in which many empirical tunings are required to mitigate the numerical sampling errors.

Before presenting the aforementioned closed analytic formulas, it is also worth pointing out some similarities and differences between the CGNS framework and the closure modeling approaches. While both approaches aim to provide surrogate models that are computationally more efficient than the original high- or infinite-dimensional nonlinear system, there is a fundamental difference in how the small-scale $\mathbf{Y}$ is handled.

A closure model of the resolved large-scale variable $\mathbf{X}$ is a closed system for $\mathbf{X}$ that does not depend on $\mathbf{Y}$. For such models, the nonlinear coupling between $\mathbf{X}$ and $\mathbf{Y}$ in the original system are approximated/parameterized using "diagnostic" terms that can involve past values of $\mathbf{X}$ as well as noise forcing. A theoretical underpinning of closure modeling is the Mori–Zwanzig (MZ) formalism originated from statistical mechanics[100,137] and later extended to non-Hamiltonian systems as well.[30,31] Nevertheless, how to efficiently construct the terms appearing in the MZ formalism is still not clear, and several different approaches have been proposed that aim to approximate its constituent terms (see e.g., Refs. 30, 58, 119, 132, 133, 69, 32, 80, 18, 79, 114, 127, and references therein).

Instead of using diagnostic terms, the CGNS keeps a simplified prognostic equation for $\mathbf{Y}$. It takes the form of a linear forced equation where both the coefficients in the draft part and the diffusion coefficient can still depend on $\mathbf{X}$ in a highly nonlinear way as encapsulated in the $\mathbf{a_0}(\mathbf{X},t)$, $\mathbf{a_1}(\mathbf{X},t)$ and $\mathbf{b_2}(\mathbf{X},t)$ terms in (2b). The CGNS also requires $\mathbf{Y}$ to enter the $\mathbf{X}$-equation linearly, through the term $\mathbf{A_1}(\mathbf{X},t)\mathbf{Y}(t)$ in (2a). Since the CGNS does not allow memory terms to appear in its vector field in order to have access to closed analytic formulas of the conditional statistics to be presented next, it is expected that the CGNS would be very effective when the memory terms in the exact MZ closure can be replicated using the prognostic Eq. (2b).

B. Closed analytic formulas for computing the conditional statistics and data assimilation

1. Nonlinear filter

For the CGNS (2), given one realization of the time series $\mathbf{X}(s)$ for $s \in [0,t]$, the conditional distribution

$$p(\mathbf{Y}(t)|\mathbf{X}(s), s \le t) \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{f}(t), \mathbf{R}_\mathbf{f}(t)) \tag{3}$$

becomes Gaussian, where the conditional mean $\boldsymbol{\mu}_\mathbf{f}$ and the conditional covariance $\mathbf{R}_\mathbf{f}$ are given by the following explicit formulas:[77]

$$\frac{\mathrm{d}\boldsymbol{\mu}_\mathbf{f}}{\mathrm{d}t} = (\mathbf{a_0} + \mathbf{a_1}\boldsymbol{\mu}_\mathbf{f}) + (\mathbf{R}_\mathbf{f}\mathbf{A}_\mathbf{1}^*)(\mathbf{B}_\mathbf{1}\mathbf{B}_\mathbf{1}^*)^{-1}\left(\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} - (\mathbf{A_0} + \mathbf{A_1}\boldsymbol{\mu}_\mathbf{f})\right), \tag{4a}$$

$$\frac{\mathrm{d}\mathbf{R}_\mathbf{f}}{\mathrm{d}t} = \mathbf{a_1}\mathbf{R}_\mathbf{f} + \mathbf{R}_\mathbf{f}\mathbf{a}_\mathbf{1}^* + \mathbf{b_2}\mathbf{b}_\mathbf{2}^* - (\mathbf{R}_\mathbf{f}\mathbf{A}_\mathbf{1}^*)(\mathbf{B}_\mathbf{1}\mathbf{B}_\mathbf{1}^*)^{-1}(\mathbf{A_1}\mathbf{R}_\mathbf{f}), \tag{4b}$$

with $\cdot^*$ being the complex conjugate transpose. The subscript "$\mathbf{f}$" in the conditional mean $\boldsymbol{\mu}_\mathbf{f}$ and conditional covariance $\mathbf{R}_\mathbf{f}$ is an abbreviation for "filter." The explicit formulas in (3)–(4) correspond to the optimal nonlinear filter solution of the state variable $\mathbf{Y}(t)$ given a realization of the observed time series $\mathbf{X}(s)$ for $s \in [0, t]$. Thus, $\boldsymbol{\mu}_\mathbf{f}$ and $\mathbf{R}_\mathbf{f}$ in (4) are also known as the filter posterior mean and the filter posterior covariance. The classical Kalman–Bucy filter[64] is the simplest special example of (4).

The closed analytic formula (4) provides an efficient algorithm for the nonlinear data assimilation of the CGNS, which avoids using the ensemble or particle methods that may suffer from sampling errors. In Sec. IV, the closed analytic data assimilation formula (4) will be used for an accurate state estimation of the initial value that facilitates effective ensemble forecast. It also allows the development of an efficient Gaussian mixture algorithm for calculating the non-Gaussian PDFs of the CGNS (see Sec. II C), which overcomes the curse of dimensionality. Such non-Gaussian PDFs are crucial in analyzing the model sensitivity and predicting the system response, the details of which will be discussed in Sec. VI.

2. Nonlinear smoother

Filtering exploits the observational information up to the current time instant for an online state estimation. On the other hand, given the observational time series within an entire time interval, the state estimation can become more accurate. This is known as the smoother.[115]

Given one realization of the observed variable $\mathbf{X}(t)$ for $t \in [0, T]$, the optimal smoother estimate $p(\mathbf{Y}(t)|\mathbf{X}(s), s \in [0, T])$ of the CGNS (2) is also Gaussian,[19]

$$p(\mathbf{Y}(t)|\mathbf{X}(s), s \in [0, T]) \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{s}(t), \mathbf{R}_\mathbf{s}(t)), \tag{5}$$

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

$$\frac{\overleftarrow{\mathrm{d}\boldsymbol{\mu}_\mathbf{s}}}{\mathrm{d}t} = -\mathbf{a}_\mathbf{0} - \mathbf{a}_\mathbf{1}\boldsymbol{\mu}_\mathbf{s} + (\mathbf{b}_2\mathbf{b}_2^*)\mathbf{R}_\mathbf{f}^{-1}(\boldsymbol{\mu}_\mathbf{f} - \boldsymbol{\mu}_\mathbf{s}), \tag{6a}$$

$$\frac{\overleftarrow{\mathrm{d}\mathbf{R}_\mathbf{s}}}{\mathrm{d}t} = -(\mathbf{a}_\mathbf{1} + (\mathbf{b}_2\mathbf{b}_2^*)\mathbf{R}_\mathbf{f}^{-1})\mathbf{R}_\mathbf{s} - \mathbf{R}_\mathbf{s}(\mathbf{a}_\mathbf{1}^* + (\mathbf{b}_2\mathbf{b}_2^*)\mathbf{R}_\mathbf{f}) + \mathbf{b}_2\mathbf{b}_2^*, \tag{6b}$$

with $\boldsymbol{\mu}_\mathbf{f}$ and $\mathbf{R}_\mathbf{f}$ given by (4). Here, the subscript "$\mathbf{s}$" in the conditional mean $\boldsymbol{\mu}_\mathbf{s}$ and conditional covariance $\mathbf{R}_\mathbf{s}$ is an abbreviation for "smoother," which should not be confused by the time variable s in $\mathbf{X}(s)$. The notation $\overleftarrow{\mathrm{d}\cdot}/\mathrm{d}t$ corresponds to the negative of the usual derivative, which means that system (6) is solved backward over $[0, T]$ with $(\boldsymbol{\mu}_\mathbf{s}(T), \mathbf{R}_\mathbf{s}(T)) = (\boldsymbol{\mu}_\mathbf{f}(T), \mathbf{R}_\mathbf{f}(T))$ with the starting value of the nonlinear smoother $(\boldsymbol{\mu}_\mathbf{s}(T), \mathbf{R}_\mathbf{s}(T))$ being the same as the filter estimate $(\boldsymbol{\mu}_\mathbf{f}(T), \mathbf{R}_\mathbf{f}(T))$.

The nonlinear smoother plays an important role for an unbiased state estimation and postprocessing of the data. It is also able to quantify the uncertainty in the unobserved variables in the parameter estimation given only partial observations, which will be a topic to be studied in Sec. V. In addition, the nonlinear smoother is the basis to the development of a nonlinear sampling formula, which will be shown below and is a necessary step in analyzing the model sensitivity and predicting the system response in Sec. VI.

3. Nonlinear sampling formula

Associated with the nonlinear smoother, a nonlinear conditional sampling formula can be derived. In addition to satisfying the point-wise optimal estimate (6), i.e., a distribution formed by conditional mean and conditional distribution at each time stamp, the conditional sampled trajectories further take into account the pathwise temporal correlation. These sampled trajectories in the CGNS framework can be regarded as the analogs of the ensemble members in the ensemble Kalman smoother,[43] but the former can be obtained via a closed analytic formula.

Conditioned on one realization of the observed variable $\mathbf{X}(s)$ for $s \in [0, T]$, the optimal strategy of sampling the trajectories associated with the unobserved variable $\mathbf{Y}$ satisfies the following explicit formula:[25]

$$\frac{\overleftarrow{\mathrm{d}\mathbf{Y}}}{\mathrm{d}t} = \frac{\overleftarrow{\mathrm{d}\boldsymbol{\mu}_\mathbf{s}}}{\mathrm{d}t} - \big(\mathbf{a}_\mathbf{1} + (\mathbf{b}_2\mathbf{b}_2^*)\mathbf{R}_\mathbf{f}^{-1}\big)(\mathbf{Y} - \boldsymbol{\mu}_\mathbf{s}) + \mathbf{b}_2\dot{\mathbf{W}}_\mathbf{Y}(t), \tag{7}$$

where $\dot{\mathbf{W}}_\mathbf{Y}(t)$ is a random noise that is independent from $\dot{\mathbf{W}}_2(t)$ in (2). The conditional sampling formula is another necessary component in analyzing the model sensitivity and predicting the system response in Sec. VI.

C. Semi-analytic and statistically accurate formulas for solving the non-Gaussian PDFs via mixtures

The closed analytic formula (4) in calculating the conditional distribution $p(\mathbf{Y}(t)|\mathbf{X}(s), s \le t)$ in (3) also provides an extremely useful way to compute the marginal distribution $p(\mathbf{Y}(t))$. In fact, assuming there are L trajectories of $\mathbf{X}(s \le t)$, denoted by $\mathbf{X}_i^{\mathrm{obs}}$ $(s \le t)$ for $i = 1, \ldots, L$, then in the limit with $L \to \infty$, the marginal distribution $p(\mathbf{Y}(t))$ is given by

$$p(\mathbf{Y}(t)) = \lim_{L\to\infty} \frac{1}{L}\sum_{i=1}^{L} p\Big(\mathbf{Y}(t)|\mathbf{X}_i^{\mathrm{obs}}(s \le t)\Big). \tag{8}$$

See Ref. 24 for the detailed derivation of (8). While the above identity is in the asymptotic form with $L \to \infty$, it has been shown that the error bound in approximating $p(\mathbf{Y}(t))$ with a finite L does not depend on the dimension of $\mathbf{Y}$.[29] In other words, fundamentally different from the traditional Monte Carlo simulations, the method in (8) avoids the curse of dimensionality. If it is further assumed that the dimension of $\mathbf{X}$ is low, then the following efficient and statistically accurate approach can be utilized to compute the joint PDF at any transient phase $p(\mathbf{X}(t), \mathbf{Y}(t))$:[23]

$$p(\mathbf{X}(t), \mathbf{Y}(t)) = \lim_{L\to\infty} \frac{1}{L}\sum_{i=1}^{L} \Big(K_\mathbf{H}(\mathbf{X}(t) - \mathbf{X}_i^{\mathrm{obs}}(t))\, p(\mathbf{Y}(t)|\mathbf{X}_i^{\mathrm{obs}}(s \le t))\Big). \tag{9}$$

Here, the distribution of $\mathbf{X}$ is approximated by a mixture distribution that is solved via a kernel density estimation, which is a non-parametric way to estimate the probability density function of a random variable. The mixture distribution is the probability distribution of a random variable that consists of different simple components. One of the simplest choices, which is also the one used in this paper, is that each mixture component is a Gaussian function centered at $\mathbf{X}_i^{\mathrm{obs}}(t)$. The same bandwidth $\mathbf{H}$, which is essentially

the covariance matrix of each mixture component, is utilized for different Gaussian mixture components and it is determined via the "solve-the-equation" method,[5] which is a method that is designed for dealing with non-Gaussian PDFs. In addition to the advantage of applying (9) in solving high-dimensional PDF (especially when the dimension of $\mathbf{Y}$ is large), the semi-analytic formula in (9) also allows a smoothed PDF, which reduces the sampling error compared with other approaches even for systems with moderate or low dimensions. Note that the assumption of the low-dimensionality of $\mathbf{X}$ is needed here, since otherwise the kernel density estimation suffers from the curse of dimensionality. See Ref. 29 for the detailed error analysis.

It is important to note that if the underlying system contains model error, then the PDF provided by (9) is very different from the one created by simply running the imperfect model. This is because the model error in the marginal distribution $p(\mathbf{Y}(t))$ is mitigated with the help of the observations $\mathbf{X}^i(s \le t)$. Such a unique feature plays a crucial role in improving the results for computing the model response and sensitivity analysis, where the perfect model is seldom known in practice. The details will be illustrated in Sec. VI. As a final remark, only the equilibrium PDF is required in many applications, including the study of the model response. Therefore, if the system is ergodic, then only a single (sufficiently long) trajectory of $\mathbf{X}(0 \le t \le T)$, denoted by $\mathbf{X}^{\text{obs}}(0 \le t \le T)$, is needed in computing the equilibrium distribution $p_{\text{eq}}(\mathbf{X}, \mathbf{Y})$,

$$p_{\text{eq}}(\mathbf{X}, \mathbf{Y}) = \lim_{J\to\infty} \frac{1}{J} \sum_{j=1}^{J} \Big(K_{\mathbf{H}}(\mathbf{X} - \mathbf{X}^{\text{obs}}(t_j))\, p(\mathbf{Y}|\mathbf{X}^{\text{obs}}(s \le t_j)) \Big), \quad (10)$$

where $[t_1, \ldots, t_J]$ is a partition of the time interval $[T_0, T]$ with some burn-in time T_0.

III. STRATEGIES OF DEVELOPING CGNS

The goals of developing approximate models for solving different problems in practice are often distinct to each other. For example, some applications require a skillful forecast model while others seek for a suitable model to reproduce certain statistics. While there is not a universal criterion to build the "optimal" CGNS as an approximate model for all applications, a few potentially useful strategies are provided below for constructing CGNS, which will be applied in Secs. III A–III C.

A. Fast wave averaging

Recall the general form of the complex systems with quadratic nonlinearity (1). Writing it into the form of state variables $(\mathbf{X}, \mathbf{Y})$ yields

$$\begin{aligned} \frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} &= L_{11}\mathbf{X} + L_{12}\mathbf{Y} + B_{11}^1(\mathbf{X},\mathbf{X}) + B_{12}^1(\mathbf{X},\mathbf{Y}) + B_{22}^1(\mathbf{Y},\mathbf{Y}) + \mathbf{F}_1(t) \\ &\quad + \boldsymbol{\sigma}_1(\mathbf{X},\mathbf{Y},t)\dot{\mathbf{W}}_1(t), \\ \frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}t} &= L_{21}\mathbf{X} + L_{22}\mathbf{Y} + B_{11}^2(\mathbf{X},\mathbf{X}) + B_{12}^2(\mathbf{X},\mathbf{Y}) + B_{22}^2(\mathbf{Y},\mathbf{Y}) + \mathbf{F}_2(t) \\ &\quad + \boldsymbol{\sigma}_2(\mathbf{X},\mathbf{Y},t)\dot{\mathbf{W}}_2(t). \end{aligned} \quad (11)$$

In (11), L_{ij} are constant matrices while B_{ij}^k are vector functions with the entries being quadratic functions of the state variables. In some applications, there exists a scale separation of the state variables, where $\mathbf{X}$ and $\mathbf{Y}$ represent slow and fast variables, respectively. In such a case, it is natural to apply a fast wave average such that the terms representing the self-interaction between the fast variables, i.e., $B_{22}^1(\mathbf{Y},\mathbf{Y})$ and $B_{22}^2(\mathbf{Y},\mathbf{Y})$, are approximated by stochastic damping and noise.[94,95] By further approximating the diffusion coefficients $\boldsymbol{\sigma}_1(\mathbf{X},\mathbf{Y})$ and $\boldsymbol{\sigma}_2(\mathbf{X},\mathbf{Y})$ by functions of only $\mathbf{X}$, the resulting system becomes

$$\begin{aligned} \frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} &= \widetilde{L}_{11}\mathbf{X} + \widetilde{L}_{12}\mathbf{Y} + B_{11}^1(\mathbf{X},\mathbf{X}) + B_{12}^1(\mathbf{X},\mathbf{Y}) + \mathbf{F}_1(t) \\ &\quad + \widetilde{\boldsymbol{\sigma}}_1(\mathbf{X},t)\dot{\mathbf{W}}_1(t), \\ \frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}t} &= \widetilde{L}_{21}\mathbf{X} + \widetilde{L}_{22}\mathbf{Y} + B_{11}^2(\mathbf{X},\mathbf{X}) + B_{12}^2(\mathbf{X},\mathbf{Y}) + \mathbf{F}_2(t) \\ &\quad + \widetilde{\boldsymbol{\sigma}}_2(\mathbf{X},t)\dot{\mathbf{W}}_2(t), \end{aligned} \quad (12)$$

which belongs to the CGNS (2). Note that, if the scale separation is not strong enough to apply the fast wave averaging, then $B_{22}^1(\mathbf{Y},\mathbf{Y})$ and $B_{22}^2(\mathbf{Y},\mathbf{Y})$ can be approximated by additional closure terms[102,113] that include nonlinear functions of $\mathbf{X}$ and bilinear functions of $\mathbf{X}$ and $\mathbf{Y}$ to fit the CGNS framework.

B. Stochastic parameterizations

The fast wave averaging or closure approximations are suitable approaches to build CGNS if the starting complex nonlinear system is completely known. However, in many practical applications, the information of the perfect model is not entirely available. Specifically, while the large-scale dynamics of nature is often accessible, the details of the small- or unresolved-scale features are not fully understood in many applications. In such a situation, suitable stochastic parameterizations can be adopted to approximate the processes of the unobserved variables $\mathbf{Y}$ such that the feedback from small/unresolved to large/resolved scales are well characterized and the parameterized system follows the CGNS structure.

One of the simplest strategies is to apply a linear stochastic model with Gaussian white noise to describe each component of the hidden processes of $\mathbf{Y}$ while the processes of $\mathbf{X}$ remain highly nonlinear. This parameterization strategy has been utilized in data assimilation and short-term statistical prediction.[6,48,49] The model with such a simple stochastic parameterization automatically fits the CGNS as there is no quadratic function of $\mathbf{Y}$ involved. A more sophisticated stochastic parameterization is to incorporate the physics-constrained nonlinear regression model[59,93] into the CGNS. This allows a more accurate way in characterizing the nonlinear dynamics of the small-scale features in $\mathbf{Y}$, influenced by the large-scale variables $\mathbf{X}$. In addition, the coupled system with physics constraints also prevents finite-time blow up of the solutions and facilitates a skillful medium- to long-range forecast.

C. System augmentation

Another strategy to derive approximate CGNS is via a simple system augmentation technique detailed below. As has been

discussed in Sec. III A that the most significant difference between the general nonlinear system (1) and the CGNS (2) is the quadratic nonlinear self-interactions of $\mathbf{Y}$, namely, the terms $B_{22}^1(\mathbf{Y},\mathbf{Y})$ and $B_{22}^2(\mathbf{Y},\mathbf{Y})$ appearing in (11). After suitable regrouping of the terms in (11) and assuming for simplicity that the diffusion coefficients depend only on X, we can rewrite this system (11) into the following form:

$$\frac{d\mathbf{X}}{dt} = \left[\mathbf{A_0}(\mathbf{X},t) + \mathbf{A_1}(\mathbf{X},t)\mathbf{Y} + B_{22}^1(\mathbf{Y},\mathbf{Y})\right] + \boldsymbol{\sigma}_1(\mathbf{X},t)\dot{\mathbf{W}}_1(t), \tag{13a}$$

$$\frac{d\mathbf{Y}}{dt} = \left[\mathbf{a_0}(\mathbf{X},t) + \mathbf{a_1}(\mathbf{X},t)\mathbf{Y} + B_{22}^2(\mathbf{Y},\mathbf{Y})\right] + \boldsymbol{\sigma}_2(\mathbf{X},t)\dot{\mathbf{W}}_2(t), \tag{13b}$$

which differs from the CGNS (2) by the two quadratic terms $B_{22}^1(\mathbf{Y},\mathbf{Y})$ and $B_{22}^2(\mathbf{Y},\mathbf{Y})$. Instead of approximating these two terms directly via fast wave averaging as proposed in Sec. III A, we consider here the situation that the scale separation between $\mathbf{X}$ and $\mathbf{Y}$ is not pronounced.

We will still handle $B_{22}^2(\mathbf{Y},\mathbf{Y})$ in (13b) via suitable stochastic parameterization in terms of $\mathbf{X}$, leading to an approximate equation for $\mathbf{Y}$ of the form

$$\frac{d\mathbf{Y}}{dt} = \left[\widetilde{\mathbf{a}}_0(\mathbf{X},t) + \widetilde{\mathbf{a}}_1(\mathbf{X},t)\mathbf{Y}\right] + \widetilde{\boldsymbol{\sigma}}_2(\mathbf{X},t)\dot{\mathbf{W}}_2(t). \tag{14}$$

The term $B_{22}^1(\mathbf{Y},\mathbf{Y})$ is then dealt with through a system augmentation strategy as explained below. Denote by $\mathbf{Z} = ((Y_1)^2, Y_1Y_2, Y_1Y_3, \ldots)^\top$ the $N_2(N_2+1)/2$-dimensional vector whose components consist of all possible quadratic monomials involving the components of the N_2-dimensional small-scale variable $\mathbf{Y}$. Using Itô's formula[47] and (14), we can obtain the corresponding equation for $\mathbf{Z}$, which takes the following form:

$$\frac{d\mathbf{Z}}{dt} = \left[\mathbf{c_0}(\mathbf{X},t) + \mathbf{c_1}(\mathbf{X},t)\mathbf{Y} + \mathbf{c_2}(\mathbf{X},t)\mathbf{Z}\right] + \boldsymbol{\sigma}_3(\mathbf{X},\mathbf{Y},t)\dot{\mathbf{W}}_2(t). \tag{15}$$

Note that the quadratic term $B_{22}^1(\mathbf{Y},\mathbf{Y})$ in (13a) can be rewritten as a linear function of $\mathbf{Z}$, denoted by $L\mathbf{Z}$, thanks to the very choice of $\mathbf{Z}$. Thus, if we further approximate $\mathbf{Y}$ in the diffusion coefficient $\boldsymbol{\sigma}_3(\mathbf{X},\mathbf{Y},t)$ of (15) by, e.g., its global mean $\overline{\mathbf{Y}}$, the augmented system for $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ consisting of (13a), (14), and (15) fits into the form of CGNS given by (2) with $(\mathbf{Y},\mathbf{Z})$ here playing the role of $\mathbf{Y}$ in (2), after replacing $B_{22}^1(\mathbf{Y},\mathbf{Y})$ in (13a) by $L\mathbf{Z}$ and approximating $\boldsymbol{\sigma}_3(\mathbf{X},\mathbf{Y},t)$ in (15) by $\boldsymbol{\sigma}_3(\mathbf{X},\overline{\mathbf{Y}},t)$.

With the CGNS structure available through this augmented system, the conditional distribution such as $p(\mathbf{Y}(t)|\mathbf{X}(s\le t))$ can now be approximated by first computing the conditional distribution $p(\mathbf{Y}(t),\mathbf{Z}(t)|\mathbf{X}(s\le t))$ for the augmented system using formulas provided in Sec. II B and then marginalizing it over $\mathbf{Z}$. The essence of this simple approach is, thus, to increase the dimension of the system in exchange of analytic formulas of the conditional distributions, which would otherwise be computationally expensive to compute even for models with moderate dimensions.

When the dimension, N_2, of $\mathbf{Y}$ is very high, one may prefer to identify only a subset of $\mathbf{Y}$ to apply the system augmentation technique in order not to inflate too much the number of variables since the auxiliary variable $\mathbf{Z}$ has dimension $N_2(N_2+1)/2$. Such an extension goes beyond the scope of the current article and will be addressed in a separate communication. Note also that if $B_{22}^2(\mathbf{Y},\mathbf{Y}) \equiv 0$, then Eq. (14) is reduced to (13b). In this case, there is no approximation involved in the drift part of (15).

We demonstrate now the efficiency of this approach in the context of data assimilation and ensemble forecast using a low-dimensional truncation of a stochastic Burgers-type equation.

IV. DATA ASSIMILATION AND ENSEMBLE FORECAST

Data assimilation concerns the problem of estimating the state variables of a given, usually nonlinear and possibly stochastic, dynamical system when observations of certain related output variables are available.[43,65,74,92] One major challenge in data assimilation is the strong nonlinearity and the associated non-Gaussian statistics in the underlying dynamics, in which a direct application of the particle methods may be inaccurate, especially in the high dimensional situations. The development of cheap and effective approximate models that capture the main characteristics of the underlying dynamics is thus an important topic in state estimation and data assimilation. Since the data assimilation solution corresponds to the initialization of the subsequent forecast, an efficient and accurate data assimilation scheme is also essential to advancing the forecast skill. Note that there is usually a stronger demand in developing suitable approximate models for data assimilation than the subsequent short- or medium-range forecast since the former often involves many numerical or sampling issues in the presence of strong nonlinearity and non-Gaussianity.

In this section, a particular type of approximate models for this purpose obtained by the method of system augmentations presented in Sec. III C is studied. The resulting approximate model has the form of a CGNS. Thus, the associated data assimilation solutions can be calculated using the closed analytic formula (4) as was discussed in Sec. II B. To simplify the presentation, the idea is illustrated using a low-dimensional stochastic differential equation (SDE) with energy-conserving quadratic terms. The data assimilation results from the CGNS, which is an approximate model, will be compared with that by applying a classical ensemble data assimilation method directly to the perfect model. The goal is to illustrate the efficiency and accuracy of the data assimilation scheme using the CGNS, especially in avoiding the potential sampling and other numerical issues that appear in the ensemble-based approaches.

A. A truncated stochastic quadratic system and its CGNS approximation through system augmentation

The model considered here is the following three-dimensional SDEs with energy-conserving quadratic nonlinear terms and subject to additive white noise forcing:

$$\frac{dx}{dt} = \beta_x x + \alpha xy + \alpha yz + \sigma_x \dot{W}_x, \tag{16a}$$

$$\frac{dy}{dt} = \beta_y y - \alpha x^2 + 2\alpha xz + \sigma_y \dot{W}_y, \tag{16b}$$

$$\frac{dz}{dt} = \beta_z z - 3\alpha xy + \sigma_z \dot{W}_z. \tag{16c}$$

Here, the coefficients for the linear terms are chosen such that β_x is positive to introduce linear instability into the system, while β_y and β_z are negative, representing linear damping effects. The coefficient $\alpha > 0$ controls the strength of the nonlinearity; and the noise strength coefficients σ_x, σ_y, and σ_z are positive constants. This system can for instance be obtained as a Fourier-Galerkin projection of the stochastic Burgers–Sivashinsky equation

$$\frac{\partial u}{\partial t} = \big(\nu \partial_{xx} u + \lambda u - u \partial_x u\big) + \dot{W}(t,x)$$

posed on a bounded interval $x \in (0, L)$ subject to homogeneous Dirichlet boundary conditions. In this context, β_x, β_y, and β_z are simply the three largest eigenvalues of the linear operator and α is linked to the domain size L via $\alpha = \pi/(\sqrt{2}L^{3/2})$ (see, e.g., Chap. 6 of Ref. 15).

In the following, the largest scale variable x is treated as the observed variable while there is no direct observations for the state variables (y, z). Under this splitting of the state variables, system (16) does not have the conditional Gaussian structure due to the quadratic nonlinear term αyz between the unobserved variables that appears in (16a). Following the idea presented in Sec. III C, in order to obtain a CGNS to approximate the system (16), three auxiliary variables are introduced for the possible quadratic interactions between the two unobserved variables,

$$p = y^2, \quad q = yz, \quad r = z^2. \tag{17}$$

Using (16) and applying the Itô's formula yields

$$\begin{aligned}
\frac{\mathrm{d}p}{\mathrm{d}t} &= (\sigma_y)^2 + 2\big(\beta_y p - \alpha x^2 y + 2\alpha x q\big) + 2\sigma_y y \dot{W}_y, \\
\frac{\mathrm{d}q}{\mathrm{d}t} &= (\beta_y + \beta_z) q - \alpha x^2 z - 3\alpha x p + 2\alpha x r + \sigma_y z \dot{W}_y + \sigma_y y \dot{W}_z, \\
\frac{\mathrm{d}r}{\mathrm{d}t} &= (\sigma_z)^2 + 2\big(\beta_z r - 3\alpha x q\big) + 2\sigma_z z \dot{W}_z.
\end{aligned} \tag{18}$$

Assume that the global mean values of the unobserved variables y and z are accessible (from a period of training data). Then, in combination with (16), and after replacing y and z in the state-dependent noise terms of (18) by their respective global mean, the following augmented system is arrived at

$$\begin{aligned}
\frac{\mathrm{d}x}{\mathrm{d}t} &= \beta_x x + \alpha x y + \alpha q + \sigma_x \dot{W}_x, \\
\frac{\mathrm{d}y}{\mathrm{d}t} &= \beta_y y - \alpha x^2 + 2\alpha x z + \sigma_y \dot{W}_y, \\
\frac{\mathrm{d}z}{\mathrm{d}t} &= \beta_z z - 3\alpha x y + \sigma_z \dot{W}_z, \\
\frac{\mathrm{d}p}{\mathrm{d}t} &= (\sigma_y)^2 + 2\big(\beta_y p - \alpha x^2 y + 2\alpha x q\big) + 2\sigma_y \bar{y} \dot{W}_y, \\
\frac{\mathrm{d}q}{\mathrm{d}t} &= (\beta_y + \beta_z) q - \alpha x^2 z - 3\alpha x p + 2\alpha x r + \sigma_y \bar{z} \dot{W}_y + \sigma_z \bar{y} \dot{W}_z, \\
\frac{\mathrm{d}r}{\mathrm{d}t} &= (\sigma_z)^2 + 2\big(\beta_z r - 3\alpha x q\big) + 2\sigma_z \bar{z} \dot{W}_z,
\end{aligned} \tag{19}$$

where yz in (16a) becomes the state variable q. This augmented system (19) fits into the CGNS form of (2) with now the unobserved variables taken to be $\mathbf{Y} = (y, z, p, q, r)^\top$.

Although the dimension of the approximate system is increased compared with the original system, closed analytic equations are now accessible for the evolution of the corresponding conditional statistics for the data assimilation solutions [see Eq. (4) in Sec. II B]. As will be shown below, the approximate system (19) can provide a significantly more accurate estimation of (y, z) compared with another conditional Gaussian approximation obtained by simply removing the term αyz in (16a), called the bare truncation (BT) system below. The skill of the proposed method is comparable and sometimes even more accurate than the ensemble Kalman–Bucy filter (EnKBF)[4] while being more efficient thanks to the availability of analytic formulas.

B. Dynamical regimes and numerical setup

In the following, we consider two dynamical regimes by varying σ_x,

$$\text{Regime I: } \sigma_x = 1, \quad \text{Regime II: } \sigma_x = 0.1, \tag{20}$$

while keeping the other parameters to be

$$\sigma_y = 1, \sigma_z = 2, \beta_x = 0.1, \beta_y = -0.5, \beta_z = -1, \alpha = \pi/\sqrt{2}.$$

In particular, we have a relatively strong nonlinear effects with $\alpha \approx 2.2$, and a relatively small spectral gap between the observed and the hidden variables with $\beta_x - \beta_y = 0.6$. The two hidden variables y and z are both subject to strong noise perturbations. The two regimes differ only in the value of the noise strength σ_x in the x-equation.

The same numerical setup is adopted for both parameter regimes. The true signal is obtained by integrating the original SDE system (16) for an arbitrarily fixed noise path using the Euler–Maruyama scheme with a uniform time step size $\delta t = 5 \times 10^{-4}$ and initialized at $(x, y, z) = (0, 0, 0)$.

For the state estimation of the unobserved variables (y, z), we compare three methods:

Method 1: Apply the nonlinear filtering formula (4) for the general CGNS (2) to the augmented system (19), with $\mathbf{X} = x$ and $\mathbf{Y} = (y, z, p, q, r)^\top$. This method will be referred as the CG method below.

Method 2: Apply the nonlinear filtering formula (4) to a bare truncation of (16) in which we simply remove the term αyz in (16a) to obtain a CGNS, with $\mathbf{X} = x$ and $\mathbf{Y} = (y, z)^\top$. This method will be referred as the BT method below.

Method 3: Apply the ensemble Kalman–Bucy filtering (EnKBF) method to (16). See (24) below for its formulation.

The data assimilation for each of the above methods is performed over the time window $[0, 400]$ with the same time step size δt as the true signal. For the CG method, the global mean values $\bar{y}$ and $\bar{z}$ in (19) are taken to be the mean values of the corresponding true signal over the interval $[0, 200]$. For both CG and BT, the initial values of the conditional mean and conditional covariance are taken to be zero. For EnKBF, the size of ensemble is taken to be $N = 100$ and the unobserved variables are initialized at $(y, z) = (0, 0)$. For the sake of

clarity, we provide below some details about the EnKBF applied to (16). We introduce the following notations for the drift part of the system (16):

$$\begin{aligned} g(x,y,z) &= \beta_x x + \alpha xy + \alpha yz, \\ f_1(x,y,z) &= \beta_y y - \alpha x^2 + 2\alpha xz, \\ f_2(x,y,z) &= \beta_z z - 3\alpha xy. \end{aligned} \tag{21}$$

Denote by $\mathbf{y} = (y_1, y_2, \ldots, y_N)^\top$ and $\mathbf{z} = (z_1, z_2, \ldots, z_N)^\top$ the collection of all the N ensemble members. We define also

$$\begin{aligned} \mathcal{N}_1(x_{\text{obs}}(t), \mathbf{y}, \mathbf{z}) &= \frac{1}{\sigma_x^2(N-1)} \sum_{j=1}^{N} (y_j - \overline{\mathbf{y}}(t))(g(x_{\text{obs}}(t), y_j, z_j) \\ &\quad - \overline{g}(x_{\text{obs}}(t), \mathbf{y}, \mathbf{z})), \\ \mathcal{N}_2(x_{\text{obs}}(t), \mathbf{y}, \mathbf{z}) &= \frac{1}{\sigma_x^2(N-1)} \sum_{j=1}^{N} (z_j - \overline{\mathbf{z}}(t))(g(x_{\text{obs}}(t), y_j, z_j) \\ &\quad - \overline{g}(x_{\text{obs}}(t), \mathbf{y}, \mathbf{z})), \end{aligned} \tag{22}$$

where

$$\begin{aligned} \overline{\mathbf{y}}(t) &= \frac{1}{N} \sum_{\ell=1}^{N} y_\ell(t), \quad \overline{\mathbf{z}}(t) = \frac{1}{N} \sum_{\ell=1}^{N} z_\ell(t), \\ \overline{g}(x_{\text{obs}}(t), \mathbf{y}, \mathbf{z}) &= \frac{1}{N} \sum_{\ell=1}^{N} g(x_{\text{obs}}(t), y_\ell, z_\ell). \end{aligned} \tag{23}$$

Then, each ensemble member (y_i, z_i), $i = 1, 2, \ldots, N$, of the EnKBF is computed using

$$\begin{aligned} \frac{dy_i}{dt} &= f_1(x_{\text{obs}}(t), y_i, z_i) + \sigma_y \dot{W}_{y,i} \\ &\quad - \mathcal{N}_1(x_{\text{obs}}(t), \mathbf{y}, \mathbf{z})\big[g(x_{\text{obs}}(t), y_i, z_i) - \dot{x}_{\text{obs}}(t) + \sigma_x \dot{W}_{x,i}\big], \\ \frac{dz_i}{dt} &= f_2(x_{\text{obs}}(t), y_i, z_i) + \sigma_z \dot{W}_{z,i} \\ &\quad - \mathcal{N}_2(x_{\text{obs}}(t), \mathbf{y}, \mathbf{z})\big[g(x_{\text{obs}}(t), y_i, z_i) - \dot{x}_{\text{obs}}(t) + \sigma_x \dot{W}_{x,i}\big], \end{aligned} \tag{24}$$

where $W_{x,i}$, $W_{y,i}$, and $W_{z,i}$, $i = 1, 2, \ldots, N$, are all mutually independent one-dimensional Brownian motions, and x_{obs} is the observed signal of x.

For regime II, we will also compare the ensemble forecast skills. The forecast is performed over the time window $[200, 400]$, which is chosen to avoid overlap with the training window $[0, 200]$ from which the global mean values of y and z appearing in (19) are computed. The forecast model is taken to be the true SDE system (16), and the initial conditions (ICs) of (y, z) are drawn from multivariate Gaussian distributions with mean and covariance Estimated, respectively, from BT, CG, and EnKBF described above. For x, its initial value is taken to be that of the true signal at the corresponding time instant. We will also compute the results when the forecast is initialized with the true signal for all the three variables, which serves as the reference of the theoretic forecast/predictability limit and will be referred to as the case with the perfect IC. The time locations at which to issue the forecasts are equally spaced over the chosen time interval, with a gap of 0.01 between two adjacent forecasts, leading, thus, to a total of 2×10^4 forecast locations. Each forecast is computed up to a lead time of 1 time unit, and a total of 40 ensemble members are generated at each forecast location. This procedure is repeated for each of the methods used to construct the IC.

C. Numerical results

We present now the results obtained based on the numerical procedure described above. For the two regimes given by (20), due to the larger noise strength parameter σ_x used in regime I for the observed variable, the corresponding DA exercise is less challenging and will be presented first.

Results for regime I. As is shown in Fig. 2 for regime I, the dynamics of x exhibits intermittent behavior with relatively quiescent episodes punctuated by large excursion events. Due to the relatively small spectral gap, the dynamics of y also exhibits highly nonlinear oscillations sustained by noise. In contrast, the dynamics of z is mainly a damped oscillation sustained by noise due to the relatively strong linear stabilizing effects, and it is the variable that decays the fastest.

The posterior mean states of (y, z) for regime I obtained by CG and EnKBF are fairly close to each other as is shown in Fig. 3. The conditional covariance matrices of (y, z) estimated by these two methods are also close to each other for this regime (not shown). In contrast, for BT, there are prolonged time windows over which the posterior mean of y deviates significantly from the true signal as is shown in panel (a) of Fig. 4. An inspection of the time series of yz shown in panel (b) of Fig. 4 reveals that such deviation typically occurs when the value of yz is large, which is expected, since the omission of the term αyz in (16a) is the only difference between the BT system and the full system. The posterior mean of z obtained by BT is similar to those obtained by CG and EnKBF shown in panel (c) of Fig. 3, which is thus not presented.

Results for regime II. The dynamics of the true system (16) in regime II exhibits similar features as in regime I shown in Fig. 2, although the amplitude of each variable is slightly reduced due to the smaller noise intensity σ_x used for this regime. The shapes of both PDFs and ACFs of all the three variables are similar to those shown in Fig. 2, except that the decorrelation time of x becoming comparable with that of y in this regime. The analog of Fig. 2 for regime II is thus omitted.

For this regime, CG and EnKBF still provide comparable posterior mean state of z as shown in panel (c) of Fig. 5, although the PDF of the posterior mean obtained by EnKBF approximates slightly better the PDF of the true signal of z. However, CG is significantly more skillful in estimating the conditional mean of y [Fig. 5, panels (b) and (d)].

The deterioration of the skill from EnKBF in this regime is associated with a false bimodal behavior appearing in the posterior mean of the y variable [see panel (d) of Fig. 5], whereas the true signal is unimodal, skewed toward negative values. It has also been checked that the bimodality is always there for EnKBF by further increasing the total number of ensemble members N or decreasing the numerical integration time step δt. Such a pathological behavior is associated with the filter divergence,[55,66] which often occurs for the ensemble-based filters when the noise in the observational process

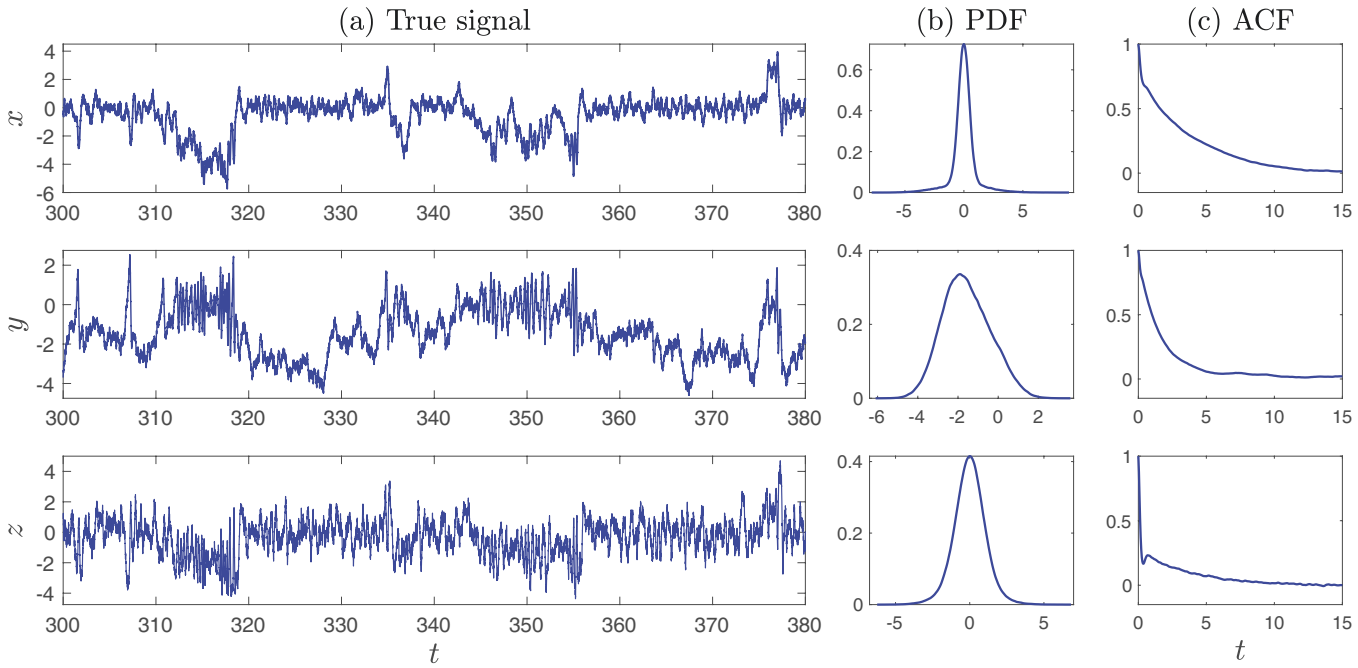

FIG. 2. Panel (a): solution of (16) in regime I given by (20) for one arbitrarily fixed realization of the noise. Panel (b): the probability density functions (PDFs) of the solution. Panel (c): the autocorrelation functions (ACFs) of the solution. This solution trajectory is taken to be the true signal for the DA and prediction experiments presented below. See Sec. IV B for details about the numerical setup. The ACFs and PDFs are estimated based on the solution trajectory over the time window $[0, 10^4]$, corresponding to 2×10^7 data points for the time step used.

is small and the observational process is highly nonlinear. In fact, the small noise in the observational process x makes the filter trusts more toward the information provided by the observations. However, the strong nonlinear and non-Gaussian features of x make it very difficult to accurately recover the states of y and z by inferring mainly from the x process. In contrast, CG tracks well the modulations of the true signal, leading to a much better reproduction of the PDF of the true signal of y [see again panel (d) of Fig. 5]. It is worth pointing out that the original system (16) can also exhibit bimodal dynamics in a broad range of dynamical regimes, even though bimodality is not observed in the true signals of (x, y, z) for neither of the two regimes considered here. This bimodality that can occur in the dynamics of (16) is induced by the additive noise that drives the system to switch from the two locally stable steady states of the corresponding deterministic system produced from a supercritical pitchfork bifurcation, although when occurs, the bimodality is mainly visible in the x variable.

Regarding BT, compared with the corresponding result shown in Fig. 4 for regime I, its performance here is even worse and is, thus, not shown. In particular, the posterior mean state of y not only deviates significantly from the true signal but also has spurious fast oscillations presenting throughout the whole time window. Such fast oscillations also appear in the filtered posterior mean of z, although to a lesser extent.

For this regime, we also compared the skills of ensemble forecast with the forecast model simply taken to be the true SDE system (16), and the initial conditions (IC) of (y, z) drawn from multivariate Gaussian distributions in which the mean and covariance are estimated, respectively, from BT, CG, and EnKBF described above (see Sec. IV B for further details). In addition, we compute the results when the forecast is initialized with the true signal for all the three variables, which serves as the reference of the theoretic forecast limit.

In Fig. 6, we presented the normalized root mean square error (RMSE; normalized by the standard deviation of the true signal) and correlation coefficients of the forecasts for all the methods used. As is expected, the better skills of CG at the DA stage carries over to the ensemble forecast as well. BT performs the worst due to large spurious oscillations appearing in its DA stage. While the RMSE is a convenient way of ranking the performance, to provide a better visualization of the skills, we also show in Fig. 7 the forecasted ensemble mean trajectories at a given lead time, chosen here to be $\tau = 0.4$ time unit, which corresponds roughly to one half of the decorrelation time of the y-variable. The results in Fig. 7 show that the forecast based on initial conditions provided by CG (middle row) actually performs fairly well for all the three variables compared with those when perfect initial conditions are used (top row), although the uncertainty in the y variable is slightly higher for CG. The results for EnKBF are correlated with its performance at the DA stage, with significant error in the x and y variables over the time windows when the filtered posterior mean of y deviates from the true signal as was previously shown in the middle panel of Fig. 5. For BT, large

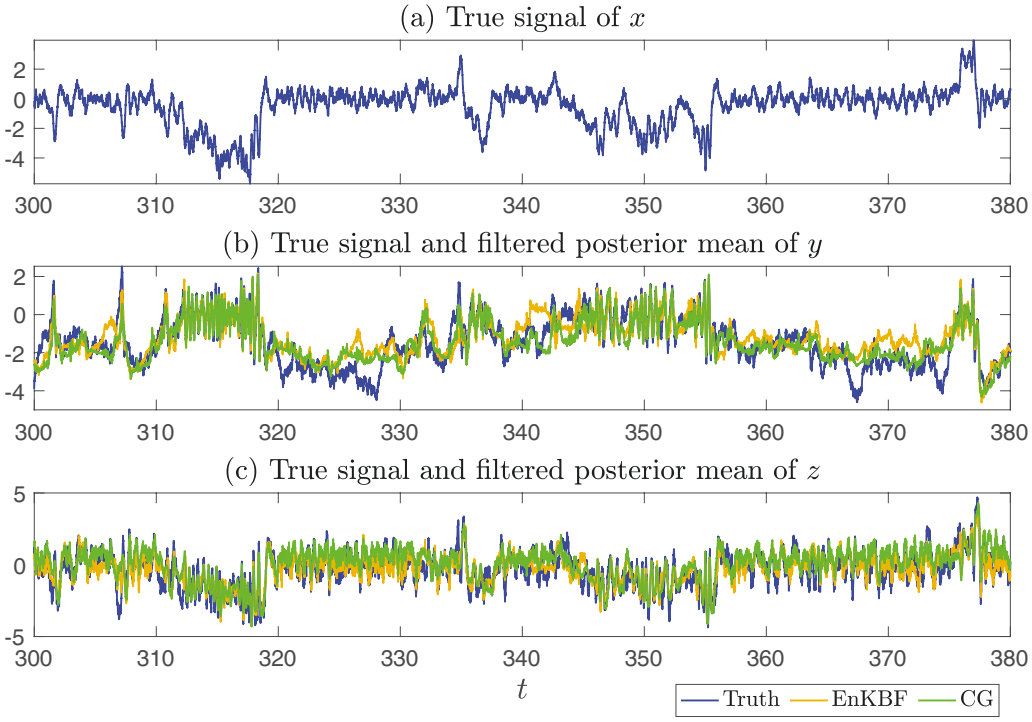

FIG. 3. Panel (a): the true signal of x, which is the same as the x-time series shown in panel (a) of Fig. 2. Panel (b): the filtered posterior mean of y for regime I obtained from CG (green) and EnKBF (orange); the corresponding true signal previously shown in Fig. 2 is plotted in blue. Panel (c): analog of panel (b) for z.

spurious oscillations appear in the forecasted ensemble mean time series for all the variables, especially for x and y. Such oscillations are inherited from those appearing in the assimilated y variables, which propagate to the other two variables due to nonlinear interactions.

V. PARAMETER ESTIMATION

Parameter estimation is an important topic and a necessary precursor for effective state estimation, data assimilation, and prediction. Maximum likelihood estimation (MLE) and maximum *a*

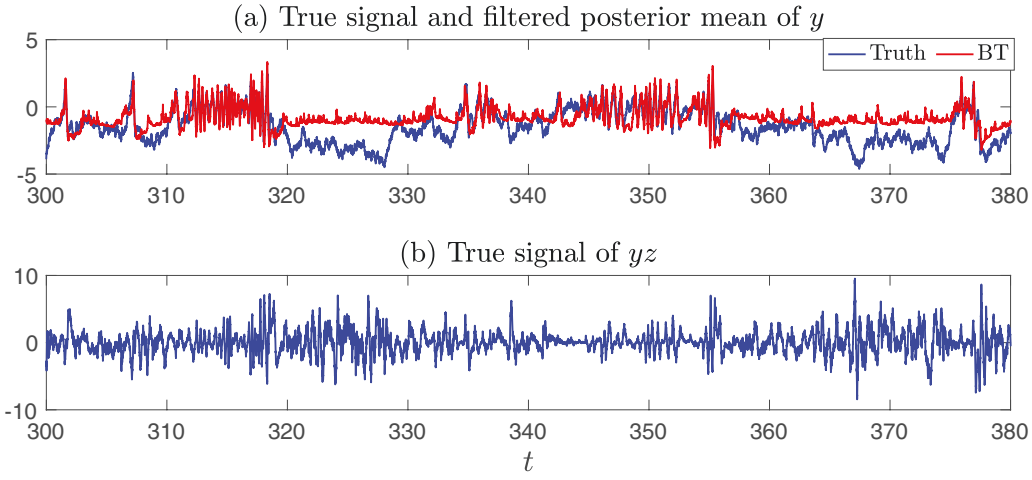

FIG. 4. Panel (a): the filtered posterior mean of y obtained from BT (red) and the true signal of y (blue). Panel (b): The true signal of yz. The deterioration of the estimated y using BT occurs over time windows when the magnitude of yz gets large.

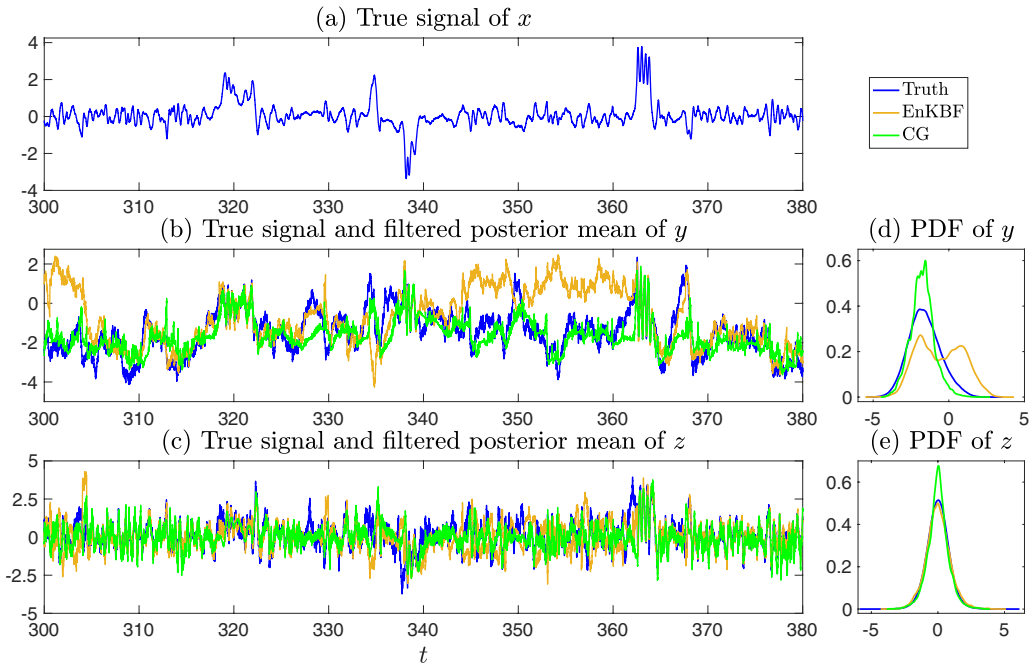

FIG. 5. Panel (a): the true signal of x, which is the x-component of the solution to (16) in regime II (i.e., $\sigma_x = 0.1$) initialized from $(x, y, z) = (0, 0, 0)$ for an arbitrarily fixed realization of the noise. Panel (b): the filtered posterior mean of y for regime II obtained from CG (green) and EnKBF (orange) with the corresponding true signal shown in blue. Panel (c): the analog of panel (b) for z. Panel (d): PDFs of the filtered posterior mean of y from CG (green) and EnKBF (orange) compared with that of the true signal. Panel (e): the analog of panel (d) for z.

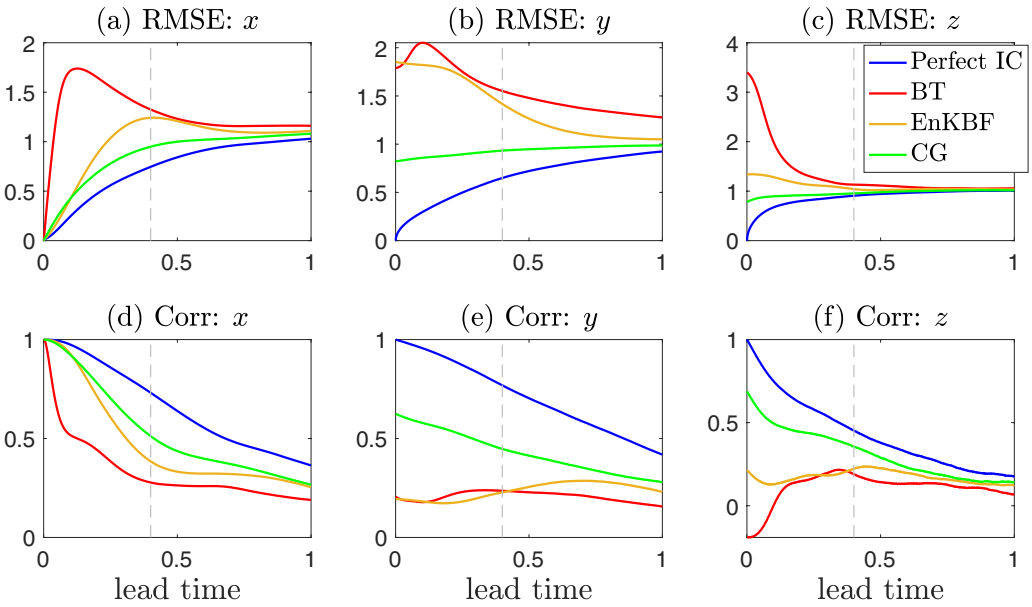

FIG. 6. Panels (a)–(c): RMSE of the forecast skills for regime II when the ICs are chosen either to be the perfect values from the true signals, or are drawn from multivariate Gaussian distributions in which the mean and covariance are estimated, respectively, from BT, CG, and EnKBF; the RMSE is normalized by dividing by the standard deviation of the respective true signal. Panels (d)–(f): the correlation coefficients of the forecast skills. The vertical dashed line corresponds to lead time $\tau = 0.4$ for which the corresponding forecasted ensemble mean time series are shown in Fig. 7.

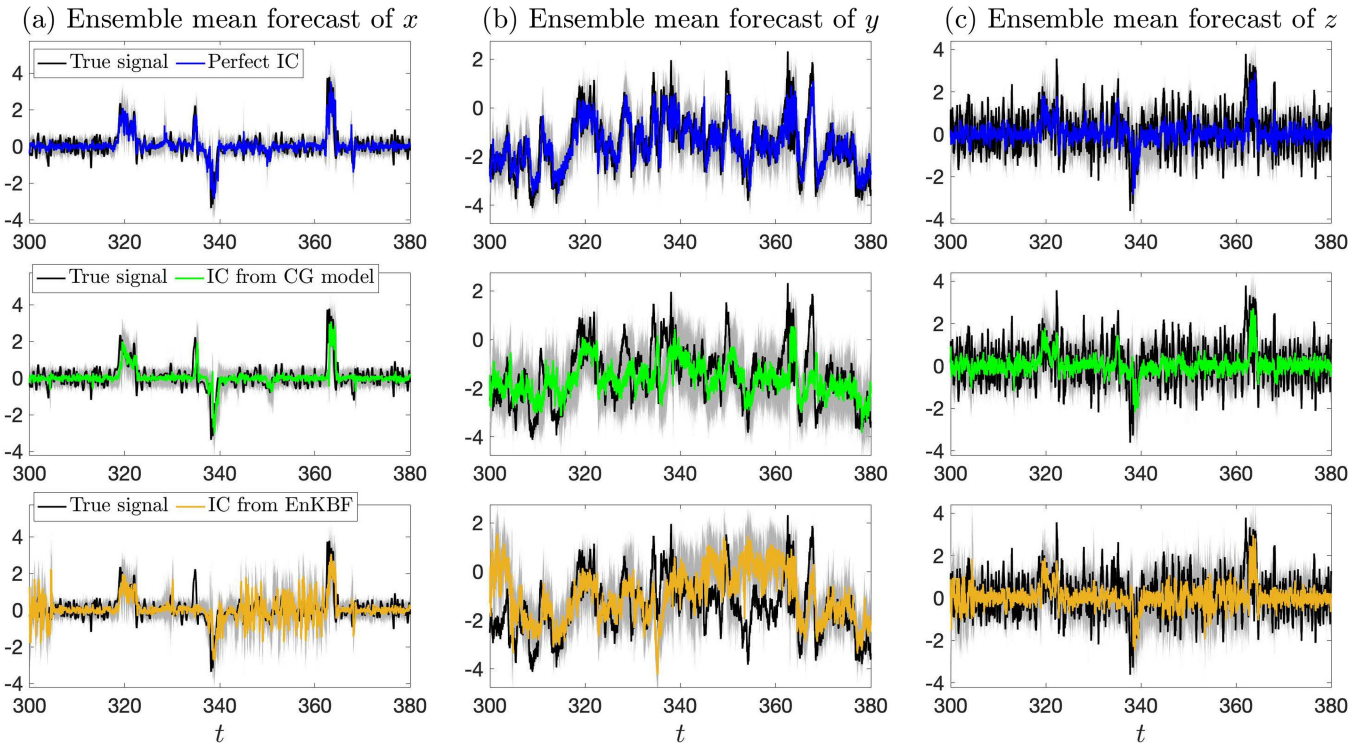

FIG. 7. Panels (a)–(c): the forecasted ensemble mean time series of x, y, and z, respectively, for regime II at lead time $\tau = 0.4$ time unit. The ICs are from either the true signal (top row), or assimilated from the CG method (middle row), or the EnKBF (bottom row). The gray area on each panel marks the spread between 5 and 95 percentile of the corresponding ensemble forecast. The true signals are plotted in black. For BT, large spurious oscillations appear in the forecasted ensemble mean time series for all the variables; and the corresponding results are, thus, not shown here.

posteriori (MAP) estimation are often adopted to infer the parameter values if the observed time series for all the state variables are accessible.[33,71,103] However, only partial observations are available in many complex nonlinear systems. In such a situation, data augmentation is widely used to simultaneously estimate the model parameter and recover the unobserved state variables.[122] In particular, data augmentation has been extensively incorporated into the Markov Chain Monte Carlo (MCMC) algorithms for improving the Bayesian inference.[42,54,107,130] In contrast to targeting the global optimal solution based on the MCMC, many other parameter estimation algorithms seek locally optimal solutions. One of the widely used local optimal parameter estimation approaches is the expectation-maximization (EM) algorithm.[38,50,51,97] The EM algorithm is an iteration method that aims to find the parameter values that maximize the likelihood function and compute certain statistical expectations of the unobserved state variables in an alternating fashion. Note that, due to the local optimality property, the EM algorithm often requires fewer iterations than the MCMC algorithm. Unfortunately, both methods are computationally expensive for general complex nonlinear systems since neither the data augmentation in MCMC nor the computation of the statistical expectation in the EM can be easily obtained.

Unlike the general complex nonlinear systems, the closed analytic formulas offered by the CGNS facilitate the acceleration of the computational efficiency in parameter estimation. In particular, the conditional sampling formula (7) has the potential to allow a rapid data augmentation in the MCMC method while the analytic state estimation formula (6) offers an exact and accurate way to compute the statistical expectation, which is an essential component in the EM algorithm. Note that appropriately incorporating the CGNS into the MCMC may require several additional manipulations, which deserves a separate topic to study. Therefore, the focus below is on applying the CGNS to accelerate the parameter estimation utilizing the EM algorithm, where detailed mathematical justifications are more accessible. The CGNS in this entire procedure acts as a preconditioner to seek a suitable approximation of the statistical expectation of the unobserved state variables given the observational time series via the EM algorithm. With such a statistical expectation available from the CGNS, the original complex nonlinear system is only essential in computing the maximum likelihood solution at the last iteration step, leading thus to a significant speedup of the computational time required.

A. Accelerating the EM algorithm with a CGNS preconditioner

As before, denote by $(\mathbf{X}, \mathbf{Y})$ the state variables of a complex nonlinear system of the form (11), where only a time series of $\mathbf{X}$ is

observed while there is no direct observations for the state variable $\mathbf{Y}$. Let $\boldsymbol{\theta}$ be a vector consisting of all the model parameters. Denote by $\widehat{\mathbf{X}} = \{\mathbf{X}^0, \ldots, \mathbf{X}^j, \ldots, \mathbf{X}^J\}$ and $\widehat{\mathbf{Y}} = \{\mathbf{Y}^0, \ldots, \mathbf{Y}^j, \ldots, \mathbf{Y}^J\}$ a discrete approximation of the continuous time series of $\mathbf{X}$ and $\mathbf{Y}$, respectively, within the time interval $t \in [0, T]$, where $T = J\Delta t$, $\mathbf{X}^j = \mathbf{X}(t_j)$ and $\mathbf{Y}^j = \mathbf{Y}(t_j)$, with $t_j = j\Delta t$.

The goal is to seek an optimal estimation of the unknown parameters $\boldsymbol{\theta}$ by maximizing the log-likelihood function. Since only the time series of $\mathbf{X}$ is observed (denoted by the discrete sequence $\widehat{\mathbf{X}}$ above), the log-likelihood estimate is obtained by averaging over the state variable $\mathbf{Y}$ at the corresponding instants,

$$\mathcal{L}(\boldsymbol{\theta}) = \log q(\widehat{\mathbf{X}}|\boldsymbol{\theta}) = \log \int_{\widehat{\mathbf{Y}}} p(\widehat{\mathbf{X}}, \widehat{\mathbf{Y}}|\boldsymbol{\theta})\, \mathrm{d}\widehat{\mathbf{Y}}. \tag{25}$$

Due to the unknown state variable $\mathbf{Y}$, there is, in general, no simple formula for a direct calculation of the log-likelihood. To find the optimal parameter $\boldsymbol{\theta}$ that maximizes $\mathcal{L}$, the EM iteration algorithm[19,38,68] operates on the following function instead which is obtained as a lower bound of $\mathcal{L}(\boldsymbol{\theta})$ by applying Jensen's inequality (cf. Sec. 3.1 of Ref. 19)

$$\int_{\widehat{\mathbf{Y}}} Q(\widehat{\mathbf{Y}}) \log p(\widehat{\mathbf{Y}}, \widehat{\mathbf{X}}|\boldsymbol{\theta})\, \mathrm{d}\widehat{\mathbf{Y}} - \int_{\widehat{\mathbf{Y}}} Q(\widehat{\mathbf{Y}}) \log Q(\widehat{\mathbf{Y}})\, \mathrm{d}\widehat{\mathbf{Y}}, \tag{26}$$

where $Q(\widehat{\mathbf{Y}})$ is a distribution over the unobserved variable $\widehat{\mathbf{Y}}$. It alternates between performing an expectation (E) step to update $Q(\widehat{\mathbf{Y}})$ and a maximization (M) step to update $\boldsymbol{\theta}$ until the estimated parameter $\boldsymbol{\theta}$ converges. Denote by $\boldsymbol{\theta}_k$ the updated parameters after the kth iteration. The EM algorithm at step $k+1$ is the following:

E-Step. Computing the conditional distribution $p(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k)$ using the previously estimated parameters $\boldsymbol{\theta}_k$. In fact, the maximization in the E-Step is reached when $Q(\widehat{\mathbf{Y}})$ is exactly the conditional distribution of $\widehat{\mathbf{Y}}$ corresponding to the smoother estimates.

M-Step. Updating the parameters $\boldsymbol{\theta}_{k+1}$ by maximizing the cost function $\mathcal{Q}$ defined by

$$\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}_k) = \int_{\widehat{\mathbf{Y}}} p(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k) \log p(\widehat{\mathbf{Y}}, \widehat{\mathbf{X}}|\boldsymbol{\theta})\, \mathrm{d}\widehat{\mathbf{Y}}. \tag{27}$$

That is,

$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}_k). \tag{28}$$

Note that in (27), $p(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k)$ is treated as a known distribution that is solved in the E-step. Therefore, $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}_k)$ is a function of $\boldsymbol{\theta}$ only. The distribution $p(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k)$ can be regarded as the weight function for computing the average (i.e., the integration) of $\log p(\widehat{\mathbf{Y}}, \widehat{\mathbf{X}}|\boldsymbol{\theta})$.

In many situations, the M-Step usually involves solving a quadratic optimization problem, the analytic formula of which is available. However, for general nonlinear systems, the conditional distribution $p(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k)$ in the E-Step is extremely difficult to solve. Note that such a conditional distribution is precisely the smoother estimate of the complex nonlinear system. Particle methods can be applied. Yet, repeatedly using these particle methods through the iteration procedure can be computationally expensive, and careful tuning is required, especially for systems with large dimensions.

To overcome this most significant barrier of computing the conditional distribution $p(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k)$ in the above EM algorithm, we exploit a suitable CGNS model as a preconditioner to accelerate this calculation in the E-Step. To that end, denote by $\boldsymbol{\theta}^M$ the collection of the parameters in the chosen CGNS model. In stark contrast to $p(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k)$, we can now use the closed analytic formula in (6) to efficiently compute the conditional distribution $p^M(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k^M)$ for the CGNS at each EM iteration. When $\boldsymbol{\theta}_k^M$ is converged after a sufficient number of EM iterations based on the CGNS, to obtain an approximation of the optimal $\boldsymbol{\theta}$ for the original system, we simply perform another EM cycle, but this time using $\log p(\widehat{\mathbf{Y}}, \widehat{\mathbf{X}}|\boldsymbol{\theta})$ instead of $\log p^M(\widehat{\mathbf{Y}}, \widehat{\mathbf{X}}|\boldsymbol{\theta}^M)$ in the M-Step. The above procedure is summarized in Algorithm 1; and we refer to Appendix A for further technical details. Throughout this section, a quantity with the superscript M indicates that it is a quantity related to the approximate CGNS model instead of the original model (see also Ref. 138 with the exception of the notation $\widehat{\mathbf{Y}}$). Note also that $\boldsymbol{\theta}^M$ can be different from $\boldsymbol{\theta}$, since the CGNS can involve new parameters not appearing in the original system and vice versa.

Note that $\boldsymbol{\theta}_{\text{opt}}$ obtained from Algorithm 1 should be viewed as an approximation of the true optimal parameters for the original nonlinear system. Its quality relies obviously on the quality of the CGNS in approximating the true dynamics. It is also worth mentioning that additional (physics) constraints can be naturally incorporated into the CGNS for parameter estimation while still preserving the closed analytic formulas in the corresponding EM algorithm; and under certain conditions, a block decomposition of the conditional covariance can be devised to further reduce the computational efforts when high dimensional systems are considered.[19] We illustrate now the approach on the classical two-layer Lorenz 1996 model.

B. A multiscale turbulent test model

In this subsection, a two-layer inhomogeneous Lorenz model is utilized to demonstrate that a suitable CGNS can serve as both a preconditioner and a surrogate model in parameter estimation to speed up the computation. First, we show that a simple approximate model that belongs to CGNS can accelerate the EM algorithm as a preconditioner by following Algorithm 1. Then, we show that this approximate model itself can be used as a surrogate model for prediction once the involved parameters, $\boldsymbol{\theta}^M$, are optimized through the EM cycles in lines 4–6 of Algorithm 1.

1. The perfect model

The two-layer Lorenz 96 (L96) model[78] is a conceptual representation of geophysical turbulence that is commonly used as a testbed for various stochastic parameterization and dimension reduction techniques.[2,32,34,45,56,57,90,91,128] The model mimics a coarse discretization of atmospheric flow on a latitude circle. It supports complex wave-like and chaotic behavior, and the two-layer structure schematically depicts the interactions between small-scale fluctuations and large-scale motions. The stochastic version of the model

ALGORITHM 1. EM algorithm for nonlinear systems with CGNS as a preconditioner.

1 Start with a given realization of the (partial) observations $\widehat{\mathbf{X}}$ for (11);
2 Propose an approximate CGNS model of the form (2);
3 Assign an initial guess for the CGNS model's parameters $\boldsymbol{\theta}_0^M$;
4 **for** $k = 1 : K$ **do**
5 E-step: compute the conditional distribution $p(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_{k-1}^M)$ using the previously estimated parameters $\boldsymbol{\theta}_{k-1}^M$;
6 M-step: update the parameters $\boldsymbol{\theta}_k^M$ with $\boldsymbol{\theta}_k^M = \arg\max_{\boldsymbol{\theta}^M} \widetilde{\mathcal{Q}}(\boldsymbol{\theta}^M; \boldsymbol{\theta}_{k-1}^M)$, where $\widetilde{\mathcal{Q}}(\boldsymbol{\theta}^M; \boldsymbol{\theta}_{k-1}^M) = \int_{\widehat{\mathbf{Y}}} p^M(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_{k-1}^M) \log p^M(\widehat{\mathbf{Y}}, \widehat{\mathbf{X}}|\boldsymbol{\theta}^M)\, \mathrm{d}\widehat{\mathbf{Y}}$.
7 E-step: compute the conditional distribution $p^M(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_K^M)$;
8 M-step: compute the optimal parameters $\boldsymbol{\theta}_{\text{opt}}$ for (11) with $\boldsymbol{\theta}_{\text{opt}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}_K^M)$, where $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}_K^M) = \int_{\widehat{\mathbf{Y}}} p^M(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_K^M) \log p(\widehat{\mathbf{Y}}, \widehat{\mathbf{X}}|\boldsymbol{\theta})\, \mathrm{d}\widehat{\mathbf{Y}}$.

subject to additive noise forcing reads

$$\frac{du_i}{dt} = \left(-u_{i-1}\left(u_{i-2} - u_{i+1}\right) - u_i + f - \frac{hc_i}{J}\sum_{j=1}^{J} v_{i,j}\right) + \sigma_{u_i}\dot{W}_{u_i}, \quad i = 1, \ldots, I, \tag{29a}$$

$$\frac{dv_{i,j}}{dt} = \left(-bc_i v_{i,j+1}\left(v_{i,j+2} - v_{i,j-1}\right) - c_i v_{i,j} + \frac{hc_i}{J}u_i\right) + \sigma_{v_{i,j}}\dot{W}_{v_{i,j}}, \quad j = 1, \ldots, J, \tag{29b}$$

where I denotes the total number of large-scale variables, J the number of small-scale variables corresponding to each large-scale variable, f, h, c_i, b, σ_{u_i}, and $\sigma_{v_{i,j}}$ are given scalar parameters while $\dot{W}_{u_i}$ and $\dot{W}_{v_{i,j}}$ are white noise. The large-scale variables u_i are periodic in i with $u_{i+I} = u_{i-I} = u_i$. The corresponding small-scale variables $v_{i,j}$ are periodic in i with $v_{i+I,j} = v_{i-I,j} = v_{i,j}$ and satisfy the following cyclic conditions in j: $v_{i,j+J} = v_{i+1,j}$ and $v_{i,j-J} = v_{i-1,j}$.

The model discussed here uses variables u_i to describe large-scale or slow movements, which are resolved; small scales or rapid fluctuations represented by $v_{i,j}$ are often unresolved ones. The coupling of fast and slow variables is regulated by the parameter h. The parameter f controls the magnitude of external large-scale forcing, while b determines the amplitude of nonlinear interactions between the fast variables. The parameter c_i specifies how quickly the fast variables are damped in comparison to the slow variables. We take $I = 40$, corresponding to a discretization of the latitude circle into a total of 40 sites/sectors, and choose $J = 4$ small-scale variables associated with each u_i. The constant forcing is set to be $f = 4$, which makes the system chaotic for the parameter regime chosen here. The parameters h, c_i, and b are chosen in such a way that the small-scale variables have a comparatively significant impact on the large-scale ones. In other words, the perfect model only has a weak scale separation. The reason that we consider such a weak scale separation is that it better mimics the real atmosphere with chaotic/turbulent behavior. The parameter c_i varies across the spatial sites, which aims to mimic the fact that the coupling across the variables above the ocean is weaker than that above the land since the latter usually have stronger friction or dissipation. In this sense, the model is inhomogeneous. Finally, additional stochastic noise is added to the system, representing the contribution of the variables that are not explicitly modeled. The noise also interacts with the deterministic part via nonlinear terms, introducing additional complexity that mimics nature. To summarize, the parameters used in the perfect model (29) are as follows:

$$\begin{gathered} I = 40, \quad J = 4, \quad h = 2, \\ c_i = 2 + 0.7\cos(2\pi i/I), \quad b = 2, \quad f = 4, \\ \sigma_{u_i} = \sigma_u = 0.2, \quad \sigma_{v_{i,j}} = \sigma_v = 1. \end{gathered} \tag{30}$$

2. The approximate model

Since in general the perfect model is not always fully known, or it is too complicated to be used in practice, it is essential to develop a simple and computationally tractable approximate model, which is nevertheless able to capture the key nonlinear feedback from the unobserved variables ($v_{i,j}$ here) to the observed variables (u_i here). As was discussed in Sec. III, stochastic parameterization is widely used in describing chaotic signals,[22] which replaces the nonlinear eddy terms by quasilinear stochastic processes on formally infinite embedded domains where the stochastic processes are Gaussian conditional to the large scale mean flow. In addition, physics constraints are adopted in designing approximate models, which also include the effects from the large-scale u_i to the small-scale variables $v_{i,j}$. Therefore, such approximate models can potentially be used as surrogate models of the perfect model.

The approximate model that we introduce for (29) is as follows:

$$\frac{du_i}{dt} = \left(-u_{i-1}\left(u_{i-2} - u_{i+1}\right) - u_i + \hat{f}_i - \hat{a}_i\sum_{j=1}^{J} v_{i,j}\right) + \hat{\sigma}_{u_i}\dot{W}_{u_i}, \quad i = 1, \ldots, I, \tag{31a}$$

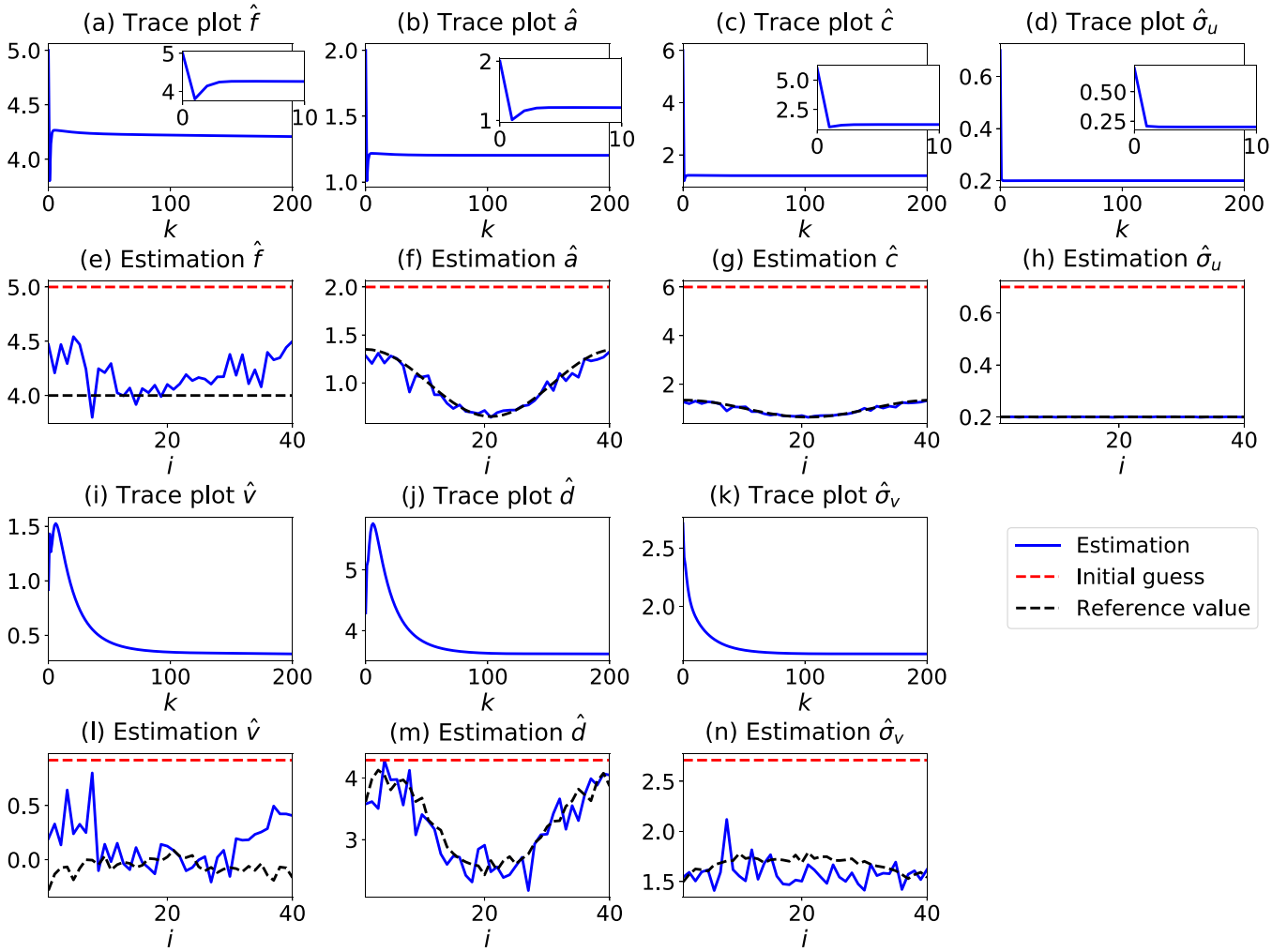

FIG. 8. Learning the parameters of the approximate L96 model (31). The panels (a)–(d) and (i)–(k) in the first and third rows show the trace plots of the estimated parameters of the EM Algorithm 1 at site $i = 2$. The panels (e)–(h) and (l)–(n) in the second and fourth rows show the final estimated parameters θ_K^M (blue) after the Kth step of the EM algorithm with K taken here to be 200; also shown are the initial guesses of the parameters (red) as well as the reference parameter values (back) chosen as follows: $\hat{f}_i$, $\hat{a}_i$ ($=hc_i/J$), $\hat{\sigma}_{u_i}$, and $\hat{c}_i$ are set to be the corresponding true values used in the perfect model, while $\hat{d}_{i,j} = \hat{d}_i$, $\hat{v}_{i,j} = \hat{v}_i$, and $\hat{\sigma}_{v_{i,j}} = \hat{\sigma}_{v_i}$ are calibrated by the true statistics according to (32).

$$\frac{dv_{i,j}}{dt} = -\hat{d}_{i,j}v_{i,j} + \hat{v}_{i,j} + \hat{c}_i u_i + \hat{\sigma}_{v_{i,j}}\dot{W}_{v_{i,j}}, \quad j = 1, \ldots, J, \tag{31b}$$

where $\hat{f}_i$, $\hat{a}_i$, $\hat{\sigma}_{u_i}$, $\hat{d}_{i,j}$, $\hat{v}_{i,j}$, $\hat{c}_i$, and $\hat{\sigma}_{v_{i,j}}$ are unknown constants. Compared with the original system (29), the main simplification here is in the small-scale equations, where we have replaced the small-scale nonlinear interaction and linear damping terms $-bc_i v_{i,j+1}(v_{i,j+2} - v_{i,j-1}) - c_i v_{i,j}$ in (29b) by a simple linear term $-\hat{d}_{i,j}v_{i,j} + \hat{v}_{i,j}$. The equations for u_i are essentially the same as before, although $\hat{f}_i$ is allowed now to vary from site to site. In the numerical experiments below, we will enforce $\hat{a}_i = \hat{c}_i$ for the consistency of the dynamical property with the original system (29).

One desirable feature of this approximate model is that the direct coupling of the state variables only involves u_i and the corresponding $v_{i,j}$ for each fixed i. This is different from the original system (29), where u_i can have direct interactions with $v_{i+1,1}$ and $v_{i-1,J}$ due to the cyclic boundary conditions of the small scale $v_{i,j}$ over j. Such a property allows to use a block decomposition of the covariance matrix of the smoother estimate during both E-step and M-step.[19] The entire state space for all the variables $\{u_i, v_{i,j} | i = 1, \ldots, I, j = 1, \ldots, J\}$ can be decomposed into I subspaces, where each subspace can be dealt with in parallel.

3. Setup of the numerical simulations

The true signal is obtained by integrating the inhomogeneous L96 model (29) using the Euler–Maruyama scheme with the parameters given by (30), a uniform time step size $\delta t = 2 \times 10^{-3}$, and zero initial condition. The same time step size and initial condition are adopted for the approximate model (31) as well as the identified model, where the latter takes the same form as (29) but with estimated parameters. The true signals of u_i with 100 time units are used as the observations for parameter estimation while longer data with 1000 time units are used to compute their statistics. This latter length is also adopted when computing the statistics of the variables in the approximate and the identified models. The total number of the EM loops with CGNS is fixed to be 200.

4. CGNS as a preconditioner for identifying parameters in the perfect model

We first discuss the results of applying Algorithm 1 to estimate the parameters, $\boldsymbol{\theta}^M$, in the approximate model (31). Figure 8 shows that the EM algorithm with the approximate model (31) provides a very accurate approximation of the parameters corresponding to the Gaussian fit of $v_{i,j}$ in the true signal. The quantities shown in Fig. 8 are the trace plots at the site $i = 2$ corresponding to the variables u_2 and $v_{2,j}$ (in the first and the third rows) and the final estimation of the parameters in (31) (in the second and the fourth rows). The trace plots at other spatial locations have similar behavior; the parameters involved in the u_i-equations all converge quickly (within ten iteration steps), and those involved in the $v_{i,j}$-equations converge at a relatively slower speed, but all stabilized after about 100 iterations. The black curves are shown as a reference, where $\hat{f}_i$, $\hat{a}_i$ $(=hc_i/J)$, $\hat{\sigma}_{u_i}$, and $\hat{c}_i$ are taken to be those in the perfect model and $\hat{d}_{i,j} = \hat{d}_i$, $\hat{\nu}_{i,j} = \hat{\nu}_i$, and $\hat{\sigma}_{v_{i,j}} = \hat{\sigma}_{v_i}$ are calibrated by the true statistics, i.e., the mean, the variance, and the decorrelation time, in the perfect model of $v_{i,j}$ averaged over j. More precisely, denoting by μ_i, r_i, and τ_i these averaged mean, variance, and decorrelation time in the perfect model of $v_{i,j}$, the parameters $\hat{\nu}_i$, $\hat{d}_i$, and $\hat{\sigma}_{v_i}$ are then obtained via

$$\tau_i = \frac{1}{\hat{d}_i}, \quad r_i = \frac{\hat{\sigma}_{v_i}^2}{2\hat{d}_i}, \quad \mu_i = \frac{\hat{\nu}_i + hc_i/J\bar{u}_i}{\hat{d}_i}, \tag{32}$$

where $\bar{u}_i$ is the mean value of u_i.

The conditional distribution of the approximate model (31) with the estimated parameters is shown in Fig. 9. Panels (a) and (c) show the true signal of two large-scale modes u_{10} and u_{20}, and panels (c) and (d) show the true signal, and smoother estimate of two hidden modes. The smoother mean is given by the dashed-black curves, and one, two, and three standard derivations of the uncertainty are shown by the light, moderate, and dark shading areas, respectively. The shading areas cover most of the true signal, which indicates appropriate amount of uncertainties are obtained by combining the true observations of the large scale variables u_i and the optimized approximate model. In fact, characterizing appropriate amount of uncertainty plays an important role in the EM algorithm when the hidden process contains large uncertainty. If the uncertainty is totally ignored in computing the optimization in the M-Step, for example, replacing $p(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k^M)$ by its mean state, then the estimated parameters can be very biased. In fact, the solution of the estimation even blows up in the test model used here.

Figure 10 shows the estimated parameters $\boldsymbol{\theta}_{\text{opt}}$ for the perfect model, where the CGNS (31) is utilized as a preconditioner following Algorithm 1. Due to the intrinsic model error of the approximate CGNS model where the hidden variables are fully decomposed, in the sense that the correlation between the small scales corresponding to different large scales are omitted, the estimated value for the bc_i term is zero. However, other parameters, for example, f and c_i, are adjusted accordingly. It is shown in Figs. 11 and 12 that the identified full model with the estimated parameters $\boldsymbol{\theta}_{\text{opt}}$ produces dynamics that resembles the truth to a remarkable extent. Indeed, Fig. 11 shows that the Hovmoller plot of the large scale variables from the identified model [panel (b)] is almost the same as that from the true signal [panel (a)] over the given time window. Time series comparison and PDF and ACF comparisons are also shown in the first four row of Fig. 12. Both the dynamical properties and the statistics are recovered with high accuracy. Of course, the time series from the identified model should not be expected to follow the true time series at all time instants since the original model is placed in a regime with chaotic dynamics.

5. CGNS as a surrogate model

Finally, we mention that the approximate model (31) with the estimated parameters itself can be exploited as a surrogate model of the perfect system, which can be applied for ensemble forecast and other tasks. Indeed, (31) with the optimally estimated parameters $\boldsymbol{\theta}_K^M$ from Algorithm 1 recovers the dynamical properties of the true model to an extent that is almost the same as the full model with the identified parameters $\boldsymbol{\theta}_{\text{opt}}$ as shown in Fig. 11 and the last four rows in Fig. 12.

VI. PREDICTING THE STATISTICAL RESPONSE

Yet, another important topic in studying complex nonlinear systems is to predict the model response to the perturbation of the external forcing. Developing efficient and accurate approaches to study such a key issue facilitates understanding model sensitivity, regime-switching behavior, and nonlinear interactions across different scales. Resolving this problem using advanced mathematical tools also has significant practical implications, such as coping with the climate change scenario. Due to various uncertainties from the internal instability and external forcing, a probabilistic description is more suitable for characterizing the complex turbulent systems (1).

However, there exist several challenges in predicting the statistical response of complex nonlinear systems. First, solving the high-dimensional Fokker–Planck equation is the prerequisite of obtaining the model statistics, which, however, often suffers from the curse of dimensionality. Second, since simulating the perfect system is not always computationally feasible in practice, the predicted model response may become inaccurate when an approximate model is utilized. It is then important to take advantage of a suitable combination of partial observations with the approximate model to mitigate error in calculating the model statistics. Third, for general nonlinear systems, computing the statistical response in terms of the perturbations with different strengths and categories

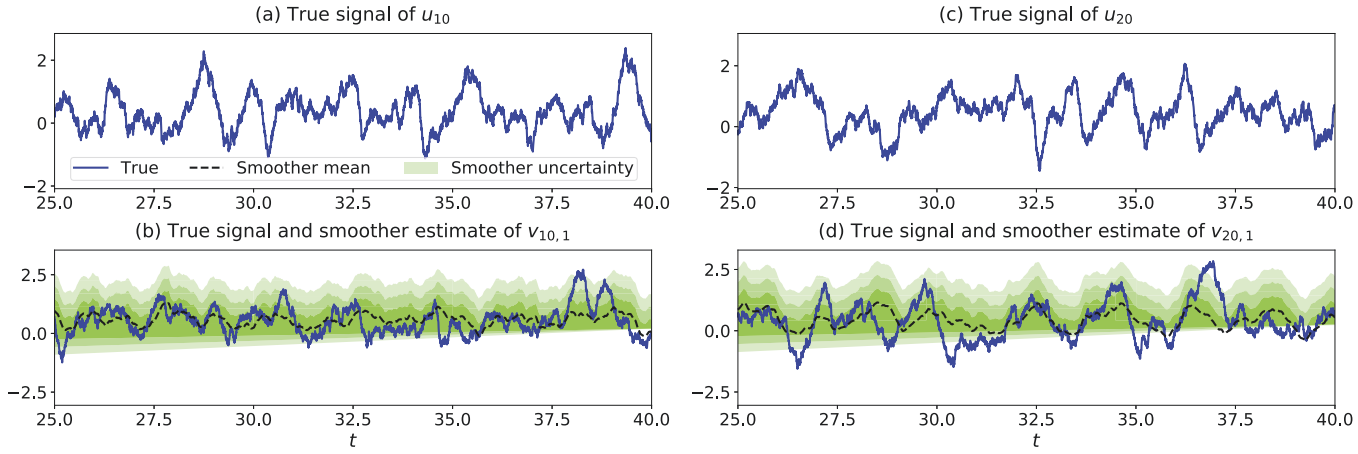

FIG. 9. Smoother estimate of the approximate model in the Kth iteration in Algorithm 1, with $K = 200$. The blue curves: true signals; the black dashed curves: the smoother mean time series of the hidden variable; the light, moderate, and dark shading areas show the one, two, and three standard derivations (STDs) of the uncertainty in the smoother estimate. Panel (a): true signal of u_{10}; Panel (b): true signal of $v_{10,1}$ with one, two, and three, standard derivations of the uncertainty. Panels (c) and (d): same as panels (a) and (b) but for u_{20} and $v_{20,1}$.

requires repeatedly solving the Fokker–Planck equation. As a consequence, even with a relatively fast solver of the Fokker–Planck equation, the total computational cost can still remain significant.

The advantage of the exact and statistically accurate solver of the equilibrium PDF (9) makes the CGNS a natural framework for the development of approximate models for efficiently computing the model statistics and the associated response. One important feature in finding the PDF of the CGNS based on formula (9) is that the available partially observed time series $\mathbf{X}$ is used to compute the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$. As a consequence, the model error in the PDF associated with the direct simulation of the approximate model is mitigated with the help of observations in computing the PDF based on (9). In other words, the resulting PDF from (9) is, in general, closer to the truth than that computed purely based on the approximate model without taking into account any input information from observations.

What remains is the third challenge mentioned above. To overcome such a difficulty, the linear statistical response is utilized as an approximate method for computing the exact statistical response. The linear response only requires a linearization of the statistical equation while there is no linearization involved in the original nonlinear dynamics. Therefore, the nonlinear features of the underlying dynamics is preserved. In addition, the linear response can be computed utilizing the fluctuation–dissipation theorem (FDT),[83] which involves only a single PDF for computing different linear responses. Note that such a PDF is the equilibrium distribution of the unperturbed state, which can be efficiently solved using the formula (10).

A. Computing the linear statistical response via the fluctuation–dissipation theorem (FDT)

Consider the general complex nonlinear systems in (1). Defining $\mathbf{G}(\mathbf{u}, t) = (L + D)\,\mathbf{u} + B\,(\mathbf{u}, \mathbf{u}) + \mathbf{F}(t)$, the model can be written in a concise form

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \mathbf{G}(\mathbf{u}, t) + \boldsymbol{\sigma}\,(\mathbf{u}, t)\dot{\mathbf{W}}. \tag{33}$$

The equilibrium statistics of some functional $A(\mathbf{u})$ associated with (33) is formulated as

$$\langle A(\mathbf{u})\rangle = \int A(\mathbf{u}) p_{\mathrm{eq}}(\mathbf{u})\,\mathrm{d}\mathbf{u}, \tag{34}$$

where $p_{\mathrm{eq}}(\mathbf{u})$ is the equilibrium PDF of $\mathbf{u}$ in (33).

Now consider the dynamics in (33) by a small time separable external forcing perturbation $\delta\mathbf{w}(\mathbf{u})f(t)$, where δ is a small scalar $\mathbf{w}$ is a general nonlinear function of $\mathbf{u}$. The perturbed system reads

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \mathbf{G}(\mathbf{u}, t) + \delta\mathbf{w}(\mathbf{u})f(t) + \boldsymbol{\sigma}\,(\mathbf{u}, t)\dot{\mathbf{W}}. \tag{35}$$

The FDT states that if δ is small enough, then the leading-order correction to the statistics in (34) is given by

$$\delta\langle A(\mathbf{u})\rangle(t) = \delta \int_0^t \mathbf{R}(t - s) f(s)\,\mathrm{d}s, \tag{36}$$

where $\mathbf{R}(t)$ is the linear response operator, which is calculated through correlation functions in the unperturbed dynamics,

$$\mathbf{R}(t) = \langle A(\mathbf{u}(t))\mathcal{B}(\mathbf{u}(0))\rangle, \quad \mathcal{B}(\mathbf{u}) = -\frac{\mathrm{div}_{\mathbf{u}}(\mathbf{w}(\mathbf{u})p_{\mathrm{eq}}(\mathbf{u}))}{p_{\mathrm{eq}}(\mathbf{u})}. \tag{37}$$

See Chap. 2 of Ref. 83 for a rigorous derivation of (36) and (37). In particular, the above procedure of computing the linear response via the FDT does not require the linearization of the underlying complex nonlinear systems. Therefore, the features of the nonlinear dynamics are preserved. Note also that if the functional $A(\mathbf{u})$ in (36) is given by $A(\mathbf{u}) = \mathbf{u}$, then the response computed is for the statistical mean. Likewise, $A(\mathbf{u}) = (\mathbf{u} - \bar{\mathbf{u}})^2$ is used for computing the response in the variance.

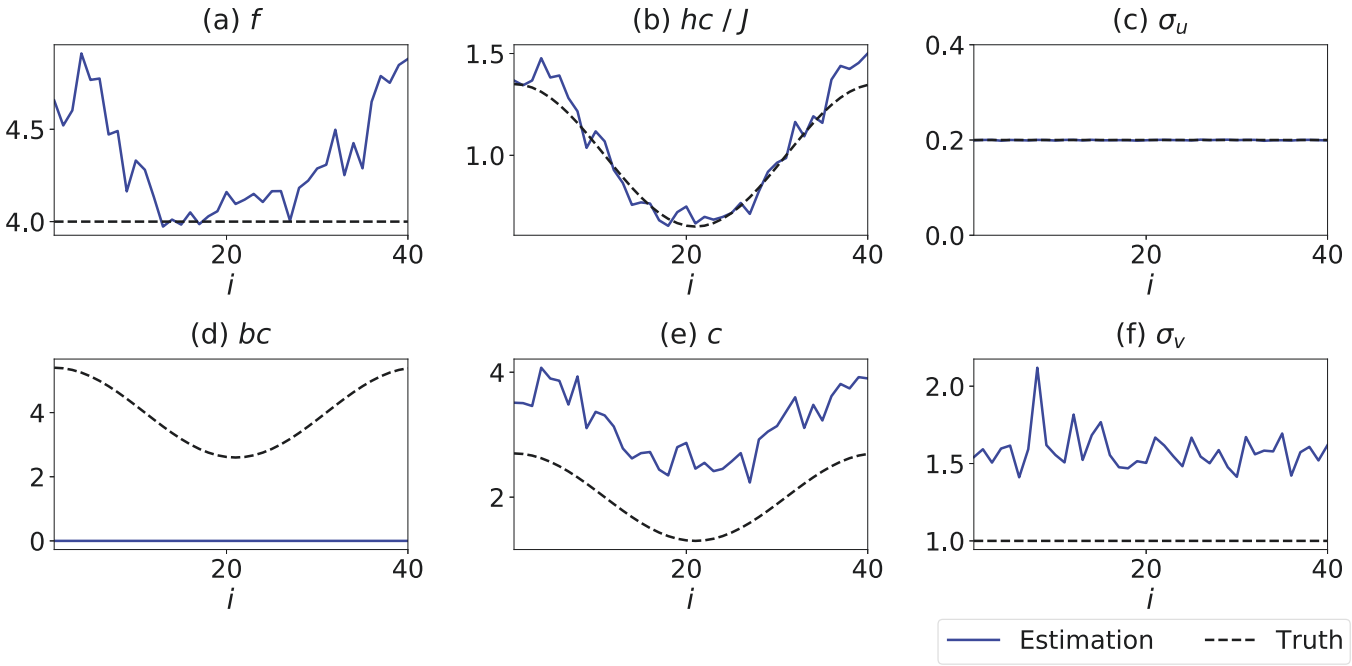

FIG. 10. Panels (a)–(f): the estimated parameter θ_{opt} using Algorithm 1 for the perfect L96 model (29). Here, θ_{opt} is a vector consisting of the parameters f, hc_i/J, σ_u, bc_i, c_i, and σ_v in (29). The true values given by (30) are marked by the dashed black curves, and the estimated parameters are shown by the blue curves.

B. Calculating the linear statistical response via the CGNS preconditioner

According to (37), the calculation of the linear statistical response via the FDT requires the information of

1. the equilibrium PDF $p_{\text{eq}}(\mathbf{u})$,
2. the time series $A(\mathbf{u})$, and
3. the correct formulation of $\mathcal{B}(\mathbf{u})$.

Even if the perfect model is known, directly solving the high-dimensional Fokker–Planck equation is often not computationally affordable. Therefore, a suitable CGNS, serving as a preconditioner, is utilized to find a suitable approximation of the non-Gaussian equilibrium PDF p_{eq} in an efficient way. Specifically, the explicit formula in (10) is utilized to achieve this goal. Besides, the observations also need to be incorporated in computing the equilibrium p_{eq} [and later in recovering the hidden components in $A(\mathbf{u})$] to reduce model error from the approximate model free run. We denote one realization of the observational variable by $\mathbf{X}^{\text{obs}}$, the posterior distribution (filter and smoother) given $\mathbf{X}^{\text{obs}}$ by $p^{M|\text{obs}}$ where the superscript M indicates that an approximate model is used in computing the filter distribution (4), the explicit formula (10) becomes

$$p_{\text{eq}}^{M|\text{obs}}(\mathbf{X},\mathbf{Y}) = \lim_{J\to\infty}\frac{1}{J}\sum_{j=1}^{J}\Big(K_{\mathbf{H}}(\mathbf{X}-\mathbf{X}^{\text{obs}}(t_j))p^{M|\text{obs}}(\mathbf{Y}|\mathbf{X}^{\text{obs}}(s\le t_j))\Big), \tag{38}$$

where $p_{\text{eq}}^{M|\text{obs}}$ is an efficient and effective approximation of the true equilibrium PDF p_{eq}.

Next, in the presence of partial observations, the conditional sampling formula (7) is exploited to calculate the unobserved component of the time series in $A(\mathbf{u})$. Note that the approximate model is used to compute the filter distribution (4), the smoother distribution formula (6), and the conditional sampling formula (7). Note that the partial observations are involved in computing both $p_{\text{eq}}(\mathbf{u})$ and $A(\mathbf{u})$, which aims to mitigate the model error in the approximate model in both equilibrium PDF and time series of the unobserved components. In addition, with $p_{\text{eq}}(\mathbf{u})$ and $A(\mathbf{u})$ obtained from the CGNS preconditioner, the original nonlinear system structure is utilized to form $\mathcal{B}(\mathbf{u})$ to compute the linear response $\mathbf{R}(t)$. The entire procedure of the FDT via CGNS is given in Algorithm 2.

C. A 4D stochastic climate model

This section utilizes a four-mode stochastic model with key features of atmospheric low-frequency variability to show how to use CGNS as a preconditioner incorporated with partial observations to calculate the linear statistical response.

1. The perfect model

The stochastic climate model is designed in such a way that it involves many of the major dynamical properties of comprehensive global circulation models (GCMs) but with only four degree of

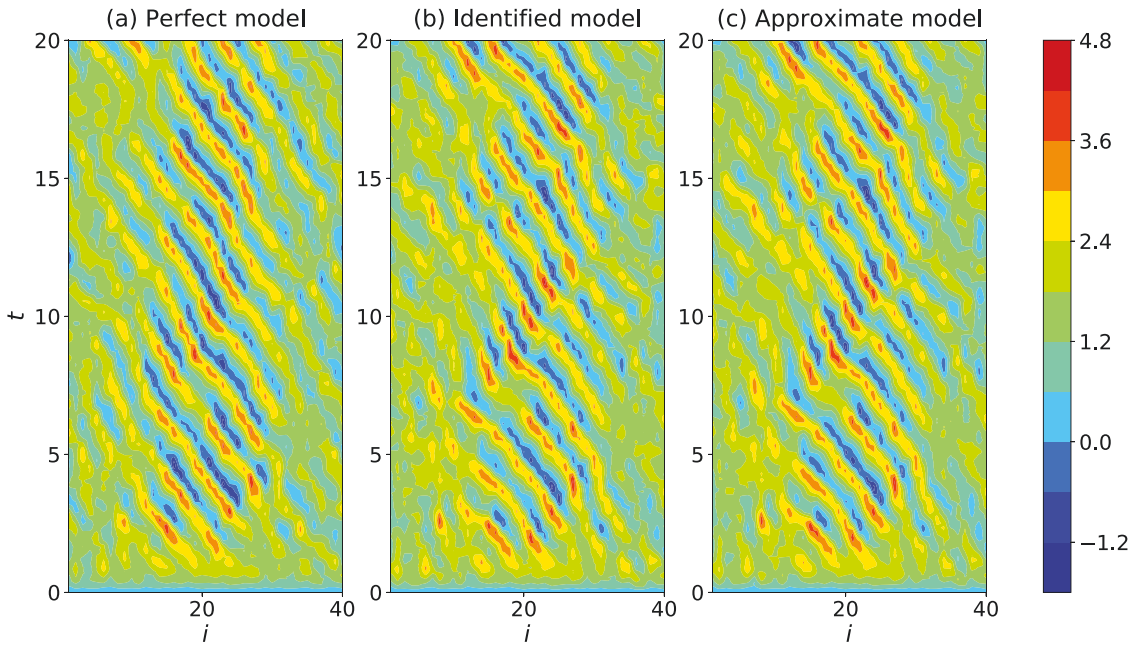

FIG. 11. Hovmoller diagram of the large scale variables u_i from different models. Panel (a): perfect model (29) with parameters (30); panel (b): perfect model (29) with estimated parameters θ_{opt} after line 8 using Algorithm 1; panel (c): approximate model (31) with parameters θ_K^M with $K = 200$.

freedom.[83,88,94,95] The model reads as follows:

$$\frac{dx_1}{dt} = \left(-x_2(L_{12} + a_1x_1 + a_2x_2) - d_1x_1 + F_1 + L_{13}y_1 + b_{123}x_2y_1\right) + \sigma_1\dot{W}_{x_1}, \tag{39a}$$

$$\frac{dx_2}{dt} = \left(+x_1(L_{12} + a_1x_1 + a_2x_2) - d_2x_2 + F_2 + L_{24}y_2 + b_{213}x_1y_1\right) + \sigma_2\dot{W}_{x_2}, \tag{39b}$$

$$\frac{dy_1}{dt} = \left(-L_{13}x_1 + b_{312}x_1x_2 + F_3 - \gamma_1y_1\right) + \sigma_3\dot{W}_{y_1}, \tag{39c}$$

$$\frac{dy_2}{dt} = \left(-L_{24}x_2 + F_4 - \gamma_2y_2\right) + \sigma_4\dot{W}_{y_2}, \tag{39d}$$

where $b_{123} + b_{213} + b_{312} = 0$. Consistent with many geophysical flow models, the model has energy-conserving quadratic nonlinear terms, a linear operator, and external forcing terms. The linear operator contains two parts: one is a skew-symmetric component formally related to the Coriolis effect and topographic Rossby wave propagation; the other is a negative definite symmetric portion conceptually analogous to dissipative processes such as surface drag and radiative damping. The coupling in different variables is through both linear and nonlinear terms, where the nonlinear coupling through b_{ijk} produces multiplicative noise effects. In fact, the strategies described in Sec. III A are applied to y_1 and y_2 that introduce the stochastic noise and damping terms. The variables x_1 and x_2 can be regarded as the climate variables and y_1 and y_2 represent the weather variables. The parameters used to generate the true dynamics are as follows:

$$\begin{gathered} d_1 = 1, \quad d_2 = 0.4, \quad \gamma_1 = 0.5, \quad \gamma_2 = 0.5, \quad L_{12} = 1, \\ L_{13} = 0.5, \quad L_{24} = 0.5, \\ a_1 = 2, \quad a_2 = 1, \quad b_{123} = 1.5, \quad b_{213} = 1.5, \\ \sigma_1 = 0.5, \quad \sigma_2 = 2, \quad \sigma_3 = 0.5, \quad \sigma_4 = 1, \quad F_1 = F_2 = F_3 = F_4 = 0. \end{gathered} \tag{40}$$

One realization of the true signal is shown in black color in Fig. 13. Both climate variable x_1 and weather variable y_1 have intermittent behavior with non-Gaussian PDFs. Note that this stochastic model is CGNS with $\mathbf{X} = (x_1, x_2)^\top$ and $\mathbf{Y} = (y_1, y_2)^\top$. We use a CGNS as the perfect model such that the FDT based on the perfect model can be computed in an accurate fashion, which can be served as a reference solution.

2. The approximate model

In practice, running the entire perfect model is prohibitively costly. As a result, simpler or reduced models are commonly utilized in computing the responses. Linear stochastic models are

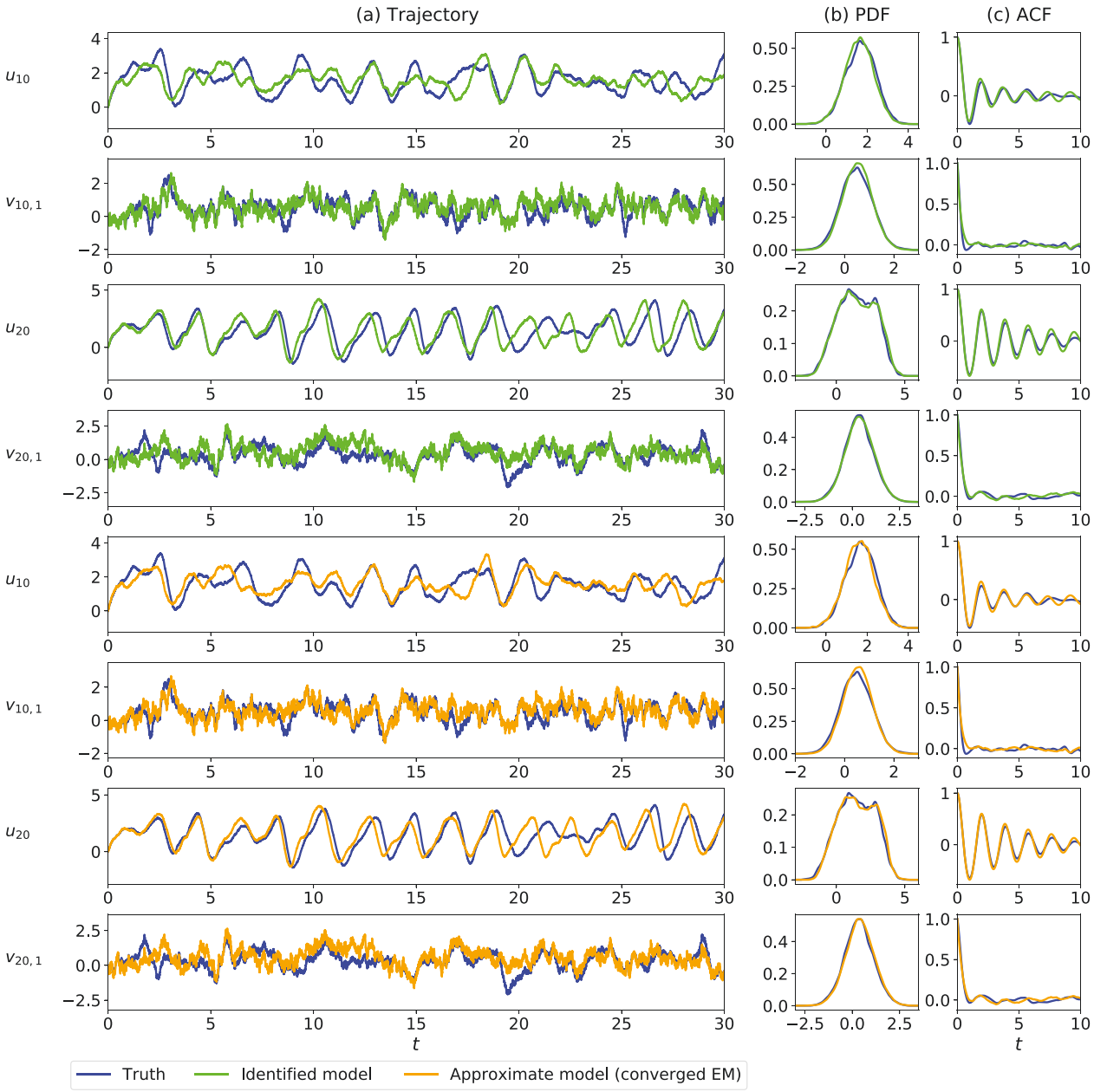

FIG. 12. Comparison of time series and statistics obtained from the full two-layer L96 model (29) with the true parameters (30) (blue), the identified model (29) with the estimated parameters θ_{opt} from Algorithm 1 (green), and the approximate CGNS model (31) with the estimated parameters θ_K^M (orange) after the Kth iteration in Algorithm 1 where $K = 200$. Panels (a)–(c): trajectory, PDF, and ACF, respectively.

widely used as approximate models for the unresolved variables.[37] Therefore, the hidden processes are replaced by two linear Gaussian equations, the parameters of which are calibrated by the true equilibrium statistics, i.e., the mean, the variance, and the decorrelation time. In addition, the parameters in the observed processes are assumed to be the same as those in the perfect model (39). The approximate model reads

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = \big(-x_2(L_{12} + a_1x_1 + a_2x_2) - d_1x_1 + F_1 + L_{13}y_1 + b_{123}x_2y_1\big) + \sigma_1\dot{W}_{x_1}, \tag{41a}$$

ALGORITHM 2. FDT with the CGNS preconditioner.

1 Start with a given realization of the observations $\mathbf{X}^{\text{obs}}$
2 Propose an approximate model that belongs to CGNS (2)
3 Compute the filter posterior distribution $p^{M|\text{obs}}(\mathbf{Y}|\mathbf{X}^{\text{obs}}(s \le t_i))$ via (4)
4 Form the equilibrium $p_{\text{eq}}{}^{M|\text{obs}}$ via Eq. (38)
5 Compute the smoother posterior distribution $p^{M|\text{obs}}(\mathbf{Y}(t)|\mathbf{X}^{\text{obs}}(s)$, $s \in [0, T]$) from (6)
6 Sample one realization of the hidden time series via (7) that is used to approximate the unobserved component of $A(\mathbf{u})$;
7 Compute the response operator $\mathbf{R}(t)$ via (37) and compute the linear response via (36).

$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = \left(+x_1(L_{12} + a_1x_1 + a_2x_2) - d_2x_2 + F_2 + L_{24}y_2 + b_{213}x_1y_1\right) + \sigma_2\dot{W}_{x_2}, \tag{41b}$$

$$\frac{\mathrm{d}y_1}{\mathrm{d}t} = -\hat{\gamma}_3\left(y_1 - \hat{y}_1\right) + \hat{\sigma}_3\dot{W}_{y_1}, \tag{41c}$$

$$\frac{\mathrm{d}y_2}{\mathrm{d}t} = -\hat{\gamma}_4\left(y_2 - \hat{y}_2\right) + \hat{\sigma}_4\dot{W}_{y_1}, \tag{41d}$$

which belongs to the CGNS. Note that despite the simplicity of utilizing linear Gaussian models to approximate the hidden processes, one major issue in (41) is that the physics constraint is no longer satisfied in (41). Therefore, the model in (41) can contain large errors for a long-term simulation.

3. Calculating linear response via FDT

In this example, we aim at calculating the linear response to the perturbation of the external forcing and linear interaction parameters. Specifically, the following two perturbation cases are considered:

Case 1: Perturbing parameters of forcing in the observed processes, i.e., $F_1^\delta = F_2^\delta = 0.3$, $\delta\mathbf{w}(\mathbf{u})f(t) = (0.3, 0.3, 0, 0)^\top$;
Case 2: Perturbing parameters in linear interaction terms, i.e., $L_{13}^\delta = L_{24}^\delta = 0.1$, $\delta\mathbf{w}(\mathbf{u})f(t) = (0.1y_1, 0.1y_2, -0.1x_1, -0.1x_2)^\top$.

Note that in the second case, the parameters appear in both the observed and hidden processes. We compare the linear response in the following models:

Perfect FDT (or perfect model): The equilibrium PDF $p_{\text{eq}}(\mathbf{u})$, the time series $A(\mathbf{u})$, and the formulation of $\mathcal{B}(\mathbf{u})$ are all from the perfect model (39).
Imperfect FDT (or imperfect model): The equilibrium PDF $p_{\text{eq}}(\mathbf{u})$, the time series $A(\mathbf{u})$, and the formulation of $\mathcal{B}(\mathbf{u})$ are all from the approximate model (41).
Concatenate FDT (or concatenate model): The equilibrium PDF $p_{\text{eq}}(\mathbf{u})$ and the time series $A(\mathbf{u})$ are from the simple concatenation of the observations, i.e., the trajectories of observed variables x_1 and x_2 from the perfect model (39), and the trajectories of the hidden variables y_1 and y_2 from the approximate

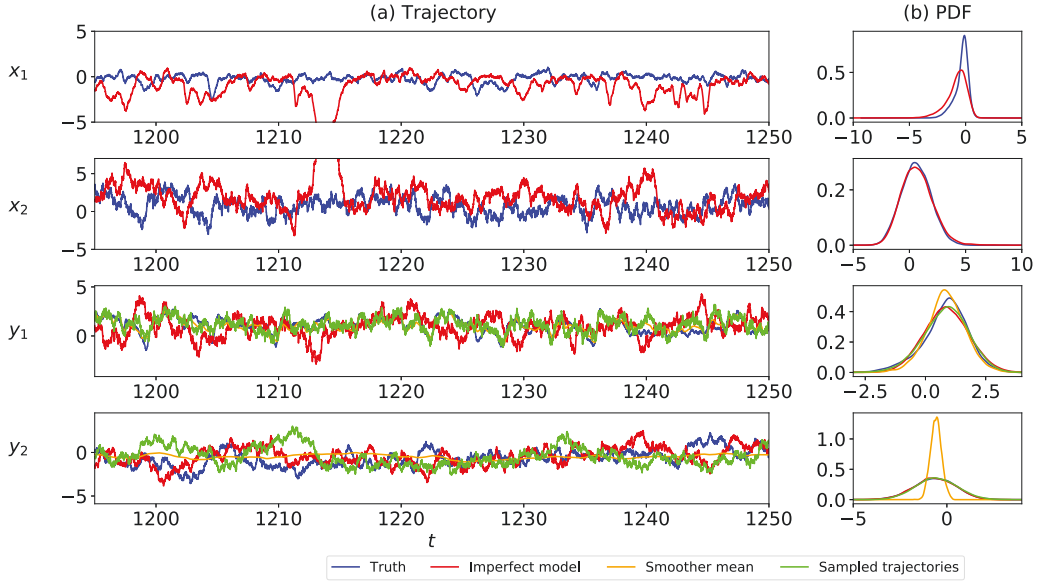

FIG. 13. Comparison of the trajectories from the perfect 4D stochastic climate model (39), approximate model (41), the smoother mean time series, and the sampled trajectories based on the approximate model. Blue curves: perfect model trajectories; red curves: approximate model trajectories; orange curves: the smoother mean time series; green curves: sampled trajectories. Panel (a): trajectories; panel (b): PDFs.

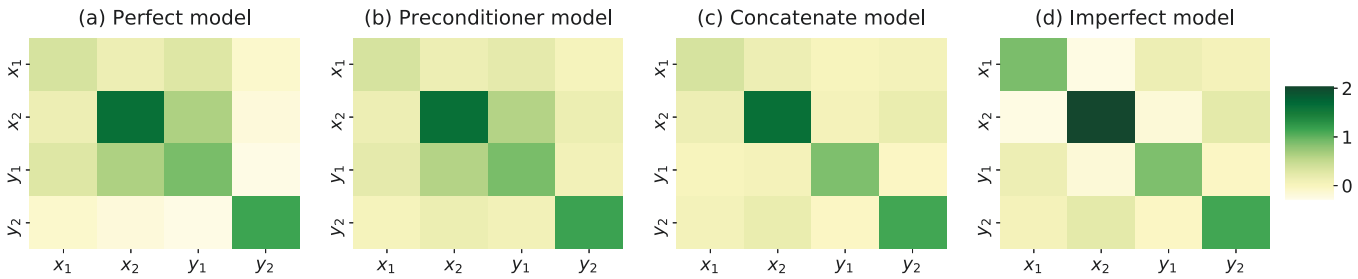

FIG. 14. Comparison of covariance matrices in different scenarios. Panel (a): the perfect model where the equilibrium PDF is from the perfect model (39); panel (b): preconditioner model where equilibrium PDF is from (4) where true observations and approximate model are utilized in computing $K_{\mathbf{H}}(\mathbf{X} - \mathbf{X}^{\text{obs}}(t_i))$ and $p^{M|\text{obs}}(\mathbf{Y}|\mathbf{X}(s \le t_i))$; panel (c): concatenate model where the equilibrium PDF p_{eq} is from the simple concatenation of the observations, i.e., the trajectories of observed variables x_1 and x_2 from the perfect model (39) and the trajectories of the hidden variables y_1 and y_2 from the approximate model (41) free-run; panel (d): the imperfect model where the equilibrium PDF is from the approximate model (41).

model (41) free-run. The formulation of $\mathcal{B}(\mathbf{u})$ is from the approximate model (41).

FDT with preconditioner (or preconditioner model): The equilibrium PDF $p_{\text{eq}}(\mathbf{u})$ is calculated via (10) where true observations and the CGNS approximate model (41) are utilized in computing $K_{\mathbf{H}}(\mathbf{X} - \mathbf{X}^{\text{obs}}(t_i))$ and $p^{M|\text{obs}}(\mathbf{Y}|\mathbf{X}(s \le t_i))$. The time series $A(\mathbf{u})$ are generated by Eq. (7) with the CGNS approximate model and the true observations. The formulation of $\mathcal{B}(\mathbf{u})$ is from the perfect model.

The details of the $\mathcal{B}(\mathbf{u})$'s forms from the perfect model and approximate model can be found in Appendix B. In this experiment, the true

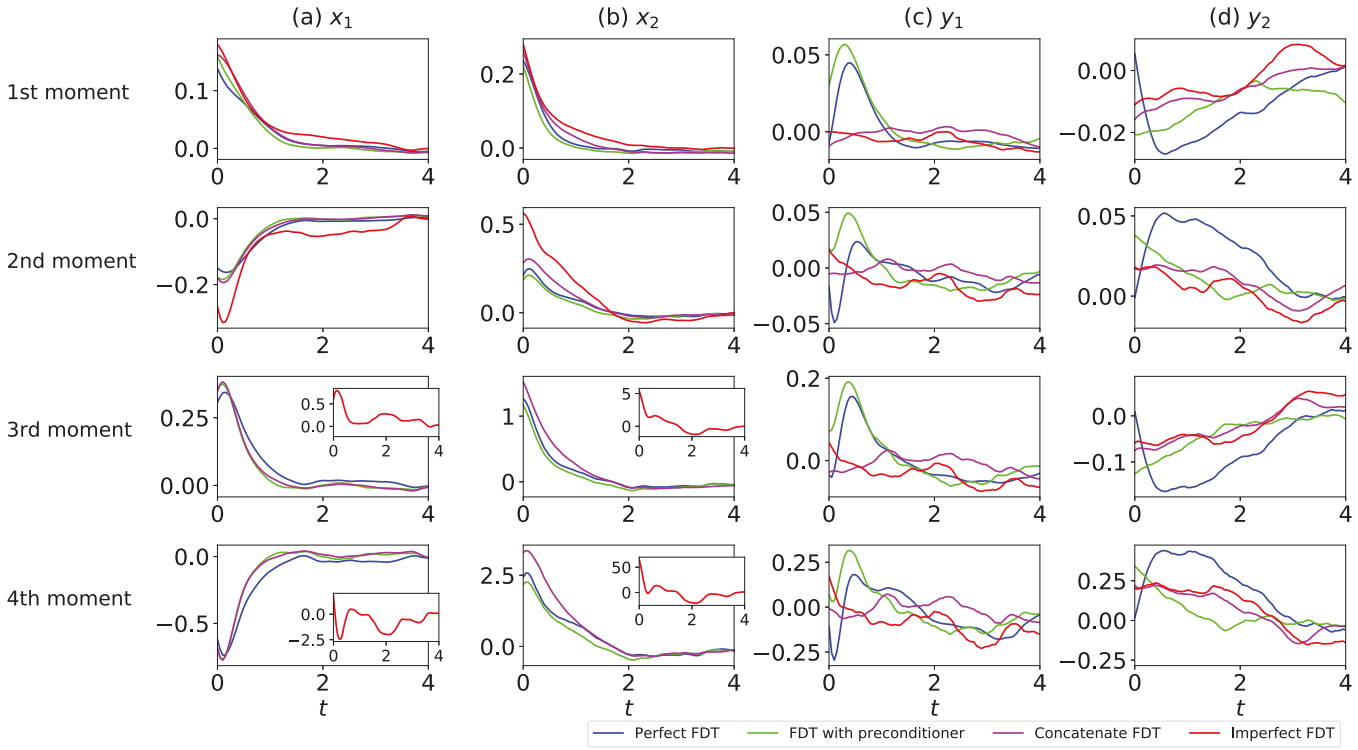

FIG. 15. Response operator $\mathbf{R}(t)$ in (37) for the response of the first four moments of the 4D climate model (39) when perturbing parameters in the observed processes with $F_1^\delta = F_2^\delta = 0.3$. In each panel, the blue and red curves show the linear response from the perfect model (39) and the free-run of the approximate model (41), respectively. The magenta curves show linear response from simply concatenating the trajectories of observed variables x_1 and x_2 from the perfect model (39) and the trajectories of the hidden variables y_1 and y_2 from the approximate model (41). The green curves show the linear response from the procedure discussed in Sec. VI B. Panels (a)–(d): x_1, x_2, y_1, and y_2, respectively.

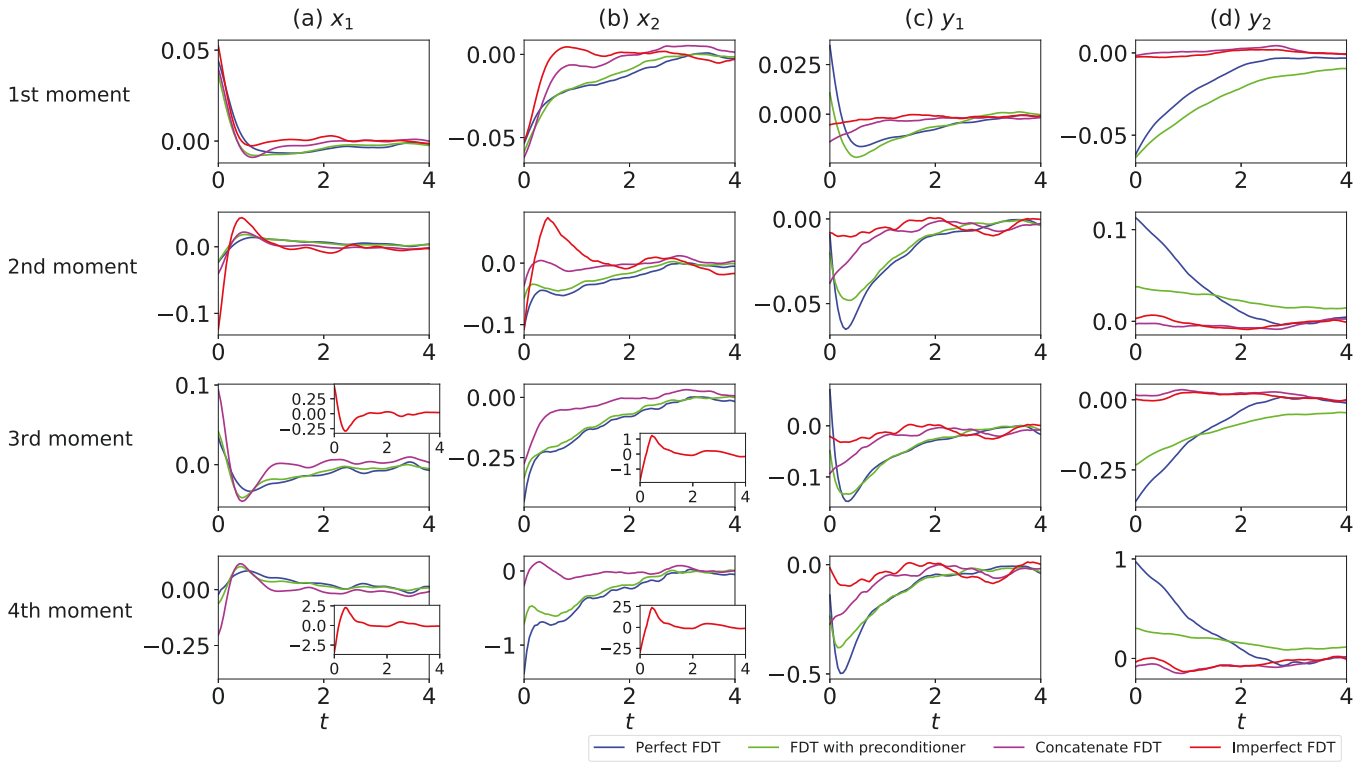

FIG. 16. Similar to Fig. 15 but perturbing parameters in the linear interaction terms $L^{\delta}_{13} = L^{\delta}_{24} = 0.1$. Panel (a)–(d): x_1, x_2, y_1, and y_2, respectively.

signal is obtained using Euler–Maruyama scheme with a uniform time step $\delta t = 5 \times 10^{-3}$ and the total length is 1000 time units.

Before discussing the linear response using CGNS as a preconditioner, we start by showing some prerequisites, the equilibrium covariance, and time series $A(\mathbf{u})$ from different models. The trajectories from the approximate model (41) free-run (red color) are shown in Fig. 13. Note that due to the violation of the energy-conserving constraint of the nonlinear terms in (41), the amplitude of x_1 and x_2 from the approximate model is much larger than the one from the perfect model (39). In addition, the PDF of y_1 is Gaussian by design, which is also different from the skewed PDF as in the perfect model. The model error can also be found in the comparison of the equilibrium covariance matrices of the perfect model [panel (a)] and the approximate model [panel (d)] in Fig. 14. Due to large error caused by the simple parameterization of the hidden processes, one may consider concatenating the observations (from the perfect model) and the trajectories of the hidden variables y_1 and y_2 from the approximate model (41) free-run to approximate the required equilibrium PDF p_{eq} and the unobserved time series. However, it is expected that the correlation between the observations from the perfect model and trajectories from the approximate model free-run is neglected [see panel (c) of Fig. 14].

In contrast, in light of the desired structures of CGNS, the partial observations can be incorporated with both approximating the equilibrium PDF and recovering the unobserved times series. Therefore, the model error is significantly mitigated, and the correlation between the observed and unobserved variables is preserved. The green curves in Fig. 13 show one sampled trajectories of y_1 and y_2 using conditional sampling formula (7) from the approximate model (41) and observations. The overall dynamics of the recovered sampled trajectories are very similar to those in the perfect model. In addition, the skewed PDF of y_1 can be found in the green curve but the appropriate model free run only brings Gaussian statistics by design. More importantly, the correlation between the sampled trajectories and the observations is consistent with that in the perfect model as shown in panel (b) of Fig. 14. A final remark is that the smoother (or filter) mean time series is widely used as a surrogate of the true signals, which, however, underestimate the uncertainty as shown in the orange color in Fig. 13. Therefore, conditional sampling is essential in approximating the time series of $A(\mathbf{u})$.

Figure 15 shows the linear response operator $R(t)$ in (37) for the response of the first four moments when perturbing the external forcing parameters in the observed processes. Here, the blue and red curves show the linear response from the perfect model (39) and the free-run of the approximate model (41), respectively. The magenta curves show the linear response from simply concatenating the trajectories of observed variables x_1 and x_2 from the perfect model (39) and the trajectories of the hidden variables y_1 and y_2 from the approximate model (41). The green curves show

the linear response from the procedure discussed in Sec. VI B. The imperfect FDT contains huge errors in capturing higher moments for the observed variables x_1 and x_2. Concatenate FDT works well for computing the four moments of x_1; however, it is gradually away from the perfect FDT from lower moments to higher moments. For example, the gap between the concatenate FDT and the perfect FDT is obvious in the last row of x_2. This is because the strong correlation between the x_2 and y_1 [shown in panel (a) in Fig. 14] is omitted in this simple concatenation [shown in panel (c) in Fig. 14]. In contrast, the response operator $R(t)$ of the two observed variables from preconditioner FDT is very close to the perfect FDT. Due to the indirect perturbation of the unresolved variables, the response is expected to be small. The preconditioner FDT can still capture the trend of the linear response operator as that using perfect FDT. Figure 16 shows the second perturbation case, i.e., perturbing the linear interaction parameters appearing in both the observed and the hidden processes. In addition to the model error of the approximate model being mitigated and correlation between observations and hidden dimensions of the approximate model being preserved, the perfect model structure is utilized to calculate the $\mathcal{B}(\mathbf{u})$ in FDT with preconditioner. Therefore, the performance of the FDT with preconditioner outperforms concatenate FDT and imperfect FDT.

VII. DISCUSSION AND CONCLUSIONS

In this paper, the skill of a rich class of nonlinear stochastic models, known as the "conditional Gaussian nonlinear system" (CGNS) (2), as both a cheap surrogate model and a fast preconditioner is explored to advance many computationally challenging tasks in complex nonlinear systems. The CGNS not only preserves the main underlying physics of nature but also reproduces the observed intermittency, extreme events and other non-Gaussian features as well. The closed analytic formula of solving the conditional statistics facilitates the development of many mathematical theories and fast numerical algorithms. Three topics are covered in this paper. First, the closed analytic formulas of the conditional statistics of the CGNS allow an efficient and accurate data assimilation scheme. It is shown in Sec. IV that the data assimilation skill of a suitable CGNS approximate forecast model outweighs the EnKBF applying directly even to the perfect model in the presence of strong nonlinear and turbulent features. The latter may suffer from filter divergence when the observational process is highly nonlinear with small observational uncertainties. Second, as is shown in Sec. V, the CGNS allows the development of a fast algorithm for simultaneously estimating the parameters and the unobserved state variables with uncertainty quantification in the presence of only partial observations. Utilizing an appropriate CGNS as a preconditioner significantly reduces the computational cost in accurately estimating the parameters in the original complex system. The same CGNS can also serve as a surrogate model for reproducing the large-scale dynamics and statistics of nature and can be applicable to ensemble forecast. Finally, the CGNS advances rapid and statistically accurate algorithms for both computing the probability density function and sampling the trajectories of the unobserved state variables. As shown in Sec. VI, these fast algorithms facilitate the development of an efficient and accurate data-driven method for predicting the linear response of the original system with respect to parameter perturbations based on a suitable CGNS preconditioner.

Several important topics remain as future work. First, it is crucial to systematically determine the CGNS approximate model. A promising approach is to write down the general structure of (12) and then apply a parameter estimation algorithm together with certain sparse identification method to prevent the overfitting issue. Second, it is of practical significance to develop further pathways that allow us to apply the CGNS to more complicated systems and explore the approximate errors. Incorporating the CGNS into intermediate complicated models or even certain versions of the general circulation models (GCMs), say for climate science or geophysics, can be interesting and practically useful tasks. Finally, the CGNS may have the potential to combine with machine learning algorithms to advance the ensemble forecast. One possible direction is to exploit the analytic formula in (9) for the ensemble forecast, where the complicated nonlinear interactions in solving the conditional distributions can be replaced by a cheaper machine learning architecture.

ACKNOWLEDGMENTS

The authors are grateful to the anonymous reviewers for their valuable feedback that helped significantly improve the manuscript. The research of N.C. was partially funded by the Office of VCRGE at UW-Madison and ONR N00014-21-1-2904. Y.L. was supported as a graduate student under the these grants. The research of Y.L. was also supported in part by NSF Award No. DMS-2023239 through the IFDS at UW-Madison. The work of H.L. was partially funded by NSF Award No. DMS-2108856.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

APPENDIX A: DETAILS OF THE EM ALGORITHM FOR PARAMETERS ESTIMATION

1. The EM algorithm for the CGNS

In this appendix, we provide some technical details about Algorithm 1, which concerns the use of CGNS as a preconditioner in the EM approach for parameters estimation. To fix ideas, we will focus on a special case in which the original nonlinear system (11) is subject to diagonal and additive noise, which is sufficient for the applications considered in Sec. V. For the simplicity of discussions, we will also consider only the real-valued variables and systems. We refer to Ref. 19 for more general settings (see also Ref. 51).

We first generate the discrete partial observations $\widehat{\mathbf{X}}$ using for instance the Euler-Maruyama scheme[47] applied to the original nonlinear system (11), with a sufficiently small time step size Δt. Given a CGNS approximation of the form (2), we explain now how

the conditional distribution $p^M(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}^M_{k-1})$ and the cost function $\widetilde{\mathcal{Q}}(\boldsymbol{\theta}^M; \boldsymbol{\theta}^M_{k-1})$ in lines 5 and 6 of Algorithm 1 are formed at each EM iteration for the CGNS.

Due to the above choice of the noise terms in (11), $\mathbf{B}_1$ and $\mathbf{b}_2$ in its corresponding CGNS approximation (2) are simply diagonal matrices with unknown diffusion coefficient parameters appearing on the corresponding main diagonal. In the following, we also denote by ξ the parameters in the drift part of the CGNS. Then, the collection $\boldsymbol{\theta}^M$ of the parameters in the CGNS is $\boldsymbol{\theta}^M = (\xi, \mathrm{diag}(\mathbf{B}_1), \mathrm{diag}(\mathbf{b}_2))^\top$. The discretization of (2) using the Euler–Maruyama scheme reads

$$\mathbf{X}^{j+1} = \mathbf{X}^j + (\mathbf{A}_0(\mathbf{X}^j, t; \xi) + \mathbf{A}_1(\mathbf{X}^j, t; \xi)\mathbf{Y}^j)\Delta t + \mathbf{B}_1\sqrt{\Delta t}\boldsymbol{\varepsilon}^j_1, \quad \text{(A1a)}$$

$$\mathbf{Y}^{j+1} = \mathbf{Y}^j + (\mathbf{a}_0(\mathbf{X}^j, t; \xi) + \mathbf{a}_1(\mathbf{X}^j, t; \xi)\mathbf{Y}^j)\Delta t + \mathbf{b}_2\sqrt{\Delta t}\boldsymbol{\varepsilon}^j_2, \quad \text{(A1b)}$$

where $j = 0, \ldots, J$ for some fixed positive integer J. Here, $\boldsymbol{\varepsilon}^j_1$ and $\boldsymbol{\varepsilon}^j_2$ are independent and identically distributed Gaussian white noises. They have the same dimensions as $\mathbf{X}$ and $\mathbf{Y}$, respectively, due again to the way the noise forcing is chosen. Assume also all the parameters in the drift part appear as multiplicative prefactors of some functions of $\mathbf{X}^j$ and $\mathbf{Y}^j$.

We introduce now some additional notations in order to put (A1) into a more compact form to be used below. Denote by $\mathbf{M}^j$ the matrix that includes those linear/nonlinear functions in the drift part, which are multiplied by the parameters ξ. Denote also by $\S^j$ those terms that do not involve parameters such as the first terms $\mathbf{X}^j$ or $\mathbf{Y}^j$ in (A1). Finally, let $\mathbf{R}$ be the covariance matrix associated with the noise terms in (A1), namely, the diagonal matrix whose diagonal consisting of those from the diagonal matrices $\mathbf{B}_1\mathbf{B}_1^\top\Delta t$ and $\mathbf{b}_2\mathbf{b}_2^\top\Delta t$. Apparently, there is a one-to-one correspondence between the diagonal of $\mathbf{R}$ and the parameters in the diffusion terms. With these notations, we can rewrite (A1) into

$$\mathbf{u}^{j+1} = \mathbf{M}^j\xi + \S^j + \mathbf{R}^{1/2}\boldsymbol{\varepsilon}^j, \quad j = 0, \ldots, J, \quad \text{(A2)}$$

where $\mathbf{u}^{j+1} = (\mathbf{X}^{j+1}, \mathbf{Y}^{j+1})^\top$ and $\boldsymbol{\varepsilon}^j = (\boldsymbol{\varepsilon}^j_1, \boldsymbol{\varepsilon}^j_2)^\top$. Thus, at each time step, given $\mathbf{M}^j$ and $\S^j$, $\mathbf{u}^{j+1}$ follows a Gaussian distribution with mean $\boldsymbol{\mu}^j = \mathbf{M}^j\xi + \S^j$ and variance given by $\mathbf{R}$,

$$p^M(\mathbf{u}^{j+1}|\mathbf{M}^j, \S^j, \boldsymbol{\theta}^M) = \frac{1}{\sqrt{(2\pi)^N}}|\mathbf{R}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{u}^{j+1} - \boldsymbol{\mu}^j)^\top(\mathbf{R})^{-1}(\mathbf{u}^{j+1} - \boldsymbol{\mu}^j)\right), \quad \text{(A3)}$$

where N is the dimension of the phase space.

At the kth EM iteration for each $k = 1, \ldots, K$, the parameters $\boldsymbol{\theta}^M_{k-1}$ is already computed. We can, thus, use (6) to compute the optimal smoother estimate $p^M(\mathbf{Y}(t)|\mathbf{X}(s), s \in [0, T], \boldsymbol{\theta}^M_{k-1})$ or equivalently its "discretized" form $p^M(\mathbf{Y}^j|\widehat{\mathbf{X}}, \boldsymbol{\theta}^M_{k-1})$ for $j = 0, \ldots, J$ in the E-Step. On the other hand, by exploiting the relationship in (A2) for all j, the M-Step (line 6 in Algorithm 1) is solved via minimizing the following cost function:

$$\widetilde{\mathcal{L}} = \frac{1}{2}\sum_{j=J_1}^{J-1}\left\langle(\mathbf{u}^{j+1} - \mathbf{M}^j\xi - \S^j)^\top(\mathbf{R})^{-1}(\mathbf{u}^{j+1} - \mathbf{M}^j\xi - \S^j)\right\rangle + \frac{J'}{2}\log|\mathbf{R}|, \quad \text{(A4)}$$

where the summation of j starts from a certain non-zero integer J_1 to eliminate the inaccuracy from the burn-in period and $J' = J - J_1$. Note that (A4) corresponds to the negative of $\mathcal{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}_k)$ defined by (27) if $J_1 = 1$, after dropping some constant terms independent of ξ and $\mathbf{R}$. In (A4), $\langle\cdot\rangle$ denotes the expectation over the uncertain component of $\mathbf{u}^j$ and $\mathbf{u}^{j+1}$, namely, $\mathbf{Y}^j$ and $\mathbf{Y}^{j+1}$, while the expectations of the observed component $\mathbf{X}^j$ and $\mathbf{X}^{j+1}$ are simply themselves since the time series of them are given. Since the hidden variables $\mathbf{Y}^j$ appear in a linear way in the matrix $\mathbf{M}^j$ due to the structure of the CGNS, only the quadratic terms of $\mathbf{Y}^j$, namely, $\langle\mathbf{Y}^{j+1}, (\mathbf{Y}^{j+1})^\top\rangle$, $\langle\mathbf{Y}^{j+1}, (\mathbf{Y}^j)^\top\rangle$, $\langle\mathbf{Y}^j, (\mathbf{Y}^j)^\top\rangle$, need to be solved in the expectation in (A4). These terms can be solved via some manipulations of the results from the closed formulas of the smoother estimates (6). Details can be found in Appendix A of Ref. 19. To find the minimum of $\widetilde{\mathcal{L}}$, we set $\frac{\partial\widetilde{\mathcal{L}}}{\partial\xi_i} = 0$ and $\frac{\partial\widetilde{\mathcal{L}}}{\partial R_{\ell\ell}} = 0$ for each component ξ_i of ξ and each diagonal element $R_{\ell\ell}$ of $\mathbf{R}$, which leads to

$$\mathbf{R} = \frac{1}{J'}\sum_{j=J_1}^{J-1}\left\langle(\mathbf{u}^{j+1} - \mathbf{M}^j\xi - \S^j)(\mathbf{u}^{j+1} - \mathbf{M}^j\xi - \S^j)^\top\right\rangle, \quad \text{(A5a)}$$

$$\xi = \mathbf{D}^{-1}, \quad \text{(A5b)}$$

where

$$\mathbf{D} = \sum_{j=J_1}^{J-1}\left\langle(\mathbf{M}^j)^\top\mathbf{R}^{-1}\mathbf{M}^j\right\rangle \quad \text{and} \quad \vec{\mathbf{r}} = \sum_{j=J_1}^{J-1}\left\langle(\mathbf{M}^j)^\top\mathbf{R}^{-1}(\mathbf{u}^{j+1} - \S^j)\right\rangle. \quad \text{(A6)}$$

Note that we solve (A5) based on an iteration method, where ξ is obtained given $\mathbf{R}$ from the previous step and then $\mathbf{R}$ is calculated from the updated ξ.

2. The last M-step in Algorithm 1

Recall Algorithm 1. The last M-Step (line 8 in the algorithm) requires to solve the minimization of the cost function, which is based on the original nonlinear system. The main difference here compared with minimizing the cost function associated with the CGNS (line 6 in the algorithm) is that higher order moments of $\mathbf{Y}$, resulting from the general nonlinear structure of the original complex system, may be involved. If the nonlinearity of the original complex system is up to quadratic, then the expectation in the analog of (A4) for the original system may involve up to the fourth moments of $\mathbf{Y}$.

Note that these moments are calculated based on the smoother estimate from the E-Step (line 7), which still utilizes the CGNS. In other words, the smoother estimate only provides a conditional Gaussian distribution. The higher order moments are, thus, computed based on the quasi-Gaussian closure approximation, which

are represented by the known information from the mean and the variance of the conditional Gaussian smoother estimate. For example, denote by Y_i a scalar component of $\mathbf{Y}$. So does Y_j, Y_k, and Y_m. Then, the third order moment $\langle Y_i Y_j Y_k \rangle$ and the fourth order moment $\langle Y_i Y_j Y_k Y_m \rangle$ can be obtained as follows:

$$\begin{aligned} \langle Y_i Y_j Y_k \rangle &= \mu_i \mu_j \mu_k + \mu_k \sigma_{ij} + \mu_i \sigma_{kj} + \mu_j \sigma_{ik}, \\ \langle Y_i Y_j Y_k Y_m \rangle &= \langle Y_i Y_j Y_k \rangle \mu_m + \mu_i \mu_j \sigma_{km} + \mu_i \mu_k \sigma_{jm} + \mu_k \mu_j \sigma_{im} \\ &\quad + \sigma_{ij}\sigma_{km} + \sigma_{ik}\sigma_{jm} + \sigma_{jk}\sigma_{im}, \end{aligned} \tag{A7}$$

where μ_i and σ_{ij} are mean and covariance of corresponding components.

APPENDIX B: CALCULATING $\mathcal{B}(\mathbf{u})$ IN THE LINEAR RESPONSE OPERATOR

In light of (37) and (38), one has the following explicit expression of $\mathcal{B}(u)$:

$$\begin{aligned} \mathcal{B}(\mathbf{u}) &= -\frac{\mathrm{div}_{\mathbf{u}}(\mathbf{w}(\mathbf{u}) p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u}))}{p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u})} \\ &= -\sum_{i=1}^{N} \frac{\partial}{\partial \mathbf{u}_i} \mathbf{w}_i(\mathbf{u}) - \sum_{i=1}^{N} \mathbf{w}_i \frac{\partial}{\partial \mathbf{u}_i} p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u}). \end{aligned} \tag{B1}$$

When perturbing the parameters of forcing in the observed processes, the forms of $\mathcal{B}(\mathbf{u})$ from the perfect model and the approximate model are the same since the parameters F_1 and F_2 appear exactly the same way as in both the perfect and the approximate model. The $\mathcal{B}(\mathbf{u})$ term reads as follows:

$$\mathcal{B}(\mathbf{u}) = -\frac{\partial}{\partial x_1} p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u}) - \frac{\partial}{\partial x_2} p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u}). \tag{B2}$$

When perturbing the parameters in linear interactions terms that appear in both the observed and hidden processes, the formulation of $\mathcal{B}(\mathbf{u})$ from the perfect and approximate models are different. Given the perturbation vector $\mathbf{w}(\mathbf{u}) = (y_1, y_2, -x_1, -x_2)^\top$, the $\mathcal{B}(\mathbf{u})$ from the perfect model is as follows:

$$\begin{aligned} \mathcal{B}(\mathbf{u}) &= -y_1 \frac{\partial}{\partial x_1} p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u}) - y_2 \frac{\partial}{\partial x_2} p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u}) + x_1 \frac{\partial}{\partial y_1} p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u}) \\ &\quad + x_2 \frac{\partial}{\partial y_2} p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u}). \end{aligned} \tag{B3}$$

However, since there are no L_{13} and L_{24} parameters in the hidden processes of the approximate model, the formulation of $\mathcal{B}(\mathbf{u})$ from the approximate model remains

$$\mathcal{B}(\mathbf{u}) = -y_1 \frac{\partial}{\partial x_1} p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u}) - y_2 \frac{\partial}{\partial x_2} p_{\mathrm{eq}}^{M|\mathrm{obs}}(\mathbf{u}). \tag{B4}$$

REFERENCES

[1] S. E. Ahmed, S. Pawar, O. San, A. Rasheed, T. Iliescu, and B. R. Noack, "On closures for reduced order models—A spectrum of first-principle to machine-learned avenues," Phys. Fluids **33**(9), 091301 (2021).

[2] H. M. Arnold, I. M. Moroz, and T. N. Palmer, "Stochastic parametrizations and model uncertainty in the Lorenz'96 system," Phil. Trans. R. Soc. A **371**(1991), 20110479 (2013).

[3] M. Asch, M. Bocquet, and M. Nodet, *Data Assimilation: Methods, Algorithms, and Applications* (SIAM, 2016).

[4] K. Bergemann and S. Reich, "An ensemble Kalman-Bucy filter for continuous data assimilation," Meteorol. Zeitschr. **21**, 213–219 (2012).

[5] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," Ann. Stat. **38**(5), 2916–2957 (2010).

[6] M. Branicki, N. Chen, and A. J. Majda, "Non-Gaussian test models for prediction and state estimation with model errors," Chin. Ann. Math. Ser. B **34**(1), 29–64 (2013).

[7] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," Proc. Natl. Acad. Sci. U.S.A. **113**(15), 3932–3937 (2016).

[8] K. Carlberg, C. Farhat, J. Cortial, and D. Amsallem, "The GNAT method for nonlinear model reduction: Effective implementation and application to computational fluid dynamics and turbulent flows," J. Comput. Phys. **242**, 623–647 (2013).

[9] J. Carpenter, P. Clifford, and P. Fearnhead, "Improved particle filter for nonlinear problems," IEE Proc. Radar Sonar Navigat. **146**(1), 2–7 (1999).

[10] A. Chattopadhyay, M. Mustafa, P. Hassanzadeh, E. Bach, and K. Kashinath, "Towards physically consistent data-driven weather forecasting: Integrating data assimilation with equivariance-preserving spatial transformers in a case study with ERA5," in *Geoscientific Model Development Discussions* (EGU Copernicus Publications, 2021), pp. 1–23.

[11] A. Chattopadhyay, M. Mustafa, P. Hassanzadeh, and K. Kashinath, "Deep spatial transformers for autoregressive data-driven forecasting of geophysical turbulence," in *Proceedings of the 10th International Conference on Climate Informatics* (Association for Computing Machinery, New York, 2020), pp. 106–112.

[12] A. Chattopadhyay, A. Subel, and P. Hassanzadeh, "Data-driven super-parameterization using deep learning: Experimentation with multiscale Lorenz 96 systems and transfer learning," J. Adv. Model. Earth Syst. **12**(11), e2020MS002084 (2020).

[13] M. D. Chekroun, H. Liu, and J. C. McWilliams, "Variational approach to closure of nonlinear dynamical systems: Autonomous case," J. Stat. Phys. **179**, 1073–1160 (2020).

[14] M. D. Chekroun, H. Liu, and J. C. McWilliams, "Stochastic rectification of fast oscillations on slow manifold closures," Proc. Natl. Acad. Sci. U.S.A. **118**(48), 147 (2021).

[15] M. D. Chekroun, H. Liu, and S. Wang, *Stochastic Parameterizing Manifolds and Non-Markovian Reduced Equations: Stochastic Manifolds for Nonlinear SPDEs II*, Springer Briefs in Mathematics (Springer, 2015).

[16] M. D. Chekroun and D. Kondrashov, "Data-adaptive harmonic spectra and multilayer Stuart-Landau models," Chaos **27**(9), 093110 (2017).

[17] M. D. Chekroun, D. Kondrashov, and M. Ghil, "Predicting stochastic systems by noise sampling, and application to the El Niño-southern oscillation," Proc. Natl. Acad. Sci. U.S.A. **108**(29), 11766–11771 (2011).

[18] N. Chen, H. Liu, and F. Lu, "Shock trace prediction by reduced models for a viscous stochastic Burgers equation," arXiv:2112.13840 (2021).

[19] N. Chen, "Learning nonlinear turbulent dynamics from partial observations via analytically solvable conditional statistics," J. Comput. Phys. **418**, 109635 (2020).

[20] N. Chen and Y. Li, "BAMCAFE: A Bayesian machine learning advanced forecast ensemble method for complex turbulent systems with partial observations," Chaos **31**(11), 113114 (2021).

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

[22] N. Chen and A. J. Majda, "Filtering nonlinear turbulent dynamical systems through conditional Gaussian statistics," Mon. Weather Rev. **144**(12), 4885–4917 (2016).

[23] N. Chen and A. J. Majda, "Beating the curse of dimension with accurate statistics for the Fokker–Planck equation in complex turbulent systems," Proc. Natl. Acad. Sci. U.S.A. **114**(49), 12864–12869 (2017).

[24] N. Chen and A. J. Majda, "Efficient statistically accurate algorithms for the Fokker–Planck equation in large dimensions," J. Comput. Phys. **354**, 242–268 (2018).

[25] N. Chen and A. J. Majda, "Efficient nonlinear optimal smoothing and sampling algorithms for complex turbulent nonlinear dynamical systems with partial observations," J. Comput. Phys. **410**, 109381 (2020).

[26]N. Chen, A. J. Majda, and D. Giannakis, "Predicting the cloud patterns of the Madden-Julian oscillation through a low-order nonlinear stochastic model," Geophys. Res. Lett. **41**(15), 5612–5619, https://doi.org/10.1002/2014GL060876 (2014).
[27]N. Chen, A. J. Majda, C. T. Sabeerali, and R. S. Ajayamohan, "Predicting monsoon intraseasonal precipitation using a low-order nonlinear stochastic model," J. Clim. **31**(11), 4403–4427 (2018).
[28]N. Chen, A. J. Majda, and X. T. Tong, "Information barriers for noisy Lagrangian tracers in filtering random incompressible flows," Nonlinearity **27**(9), 2133 (2014).
[29]N. Chen, A. J. Majda, and X. T. Tong, "Rigorous analysis for efficient statistically accurate algorithms for solving Fokker–Planck equations in large dimensions," SIAM/ASA J. Uncertain. Quantif. **6**(3), 1198–1223 (2018).
[30]A. J. Chorin, O. H. Hald, and R. Kupferman, "Optimal prediction with memory," Physica D **166**(3), 239–257 (2002).
[31]A. J. Chorin and O. H. Hald, *Stochastic Tools in Mathematics and Science*, Surveys and Tutorials in the Applied Mathematical Sciences (Springer New York, 2006).
[32]A. J. Chorin and F. Lu, "Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics," Proc. Natl. Acad. Sci. U.S.A. **112**(32), 9804–9809 (2015).
[33]M. C. Coleman and D. E. Block, "Bayesian parameter estimation with informative priors for nonlinear systems," AIChE J. **52**(2), 651–667 (2006).
[34]D. Crommelin and E. Vanden-Eijnden, "Subgrid-scale parameterization with conditional Markov chains," J. Atmos. Sci. **65**(8), 2661–2675 (2008).
[35]J. A. Curry and P. J. Webster, "Climate science and the uncertainty monster," Bull. Am. Meteorol. Soc. **92**(12), 1667–1682 (2011).
[36]T. DelSole, "Predictability and information theory. Part I: Measures of predictability," J. Atmos. Sci. **61**(20), 2425–2440 (2004).
[37]T. DelSole, "Predictability and information theory. Part II: Imperfect forecasts," J. Atmos. Sci. **62**(9), 3368–3381 (2005).
[38]A. Dembo and O. Zeitouni, "Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm," Stochast. Process. Appl. **23**(1), 91–113 (1986).
[39]H. A. Dijkstra, *Nonlinear Climate Dynamics* (Cambridge University Press, 2013).
[40]P. N. Edwards, "Global climate science, uncertainty and politics: Data-laden models, model-filtered data," Sci. Cult. **8**(4), 437–472 (1999).
[41]P. N. Edwards, "History of climate modeling," Wiley Interdiscip. Rev. Clim. Change **2**(1), 128–139 (2011).
[42]B. Eraker, "MCMC analysis of diffusion models with application to finance," J. Bus. Econ. Stat. **19**(2), 177–191 (2001).
[43]G. Evensen, *Data Assimilation: The Ensemble Kalman Filter* (Springer Science & Business Media, 2009).
[44]M. Farazmand and T. P. Sapsis, "Extreme events: Mechanisms and prediction," Appl. Mech. Rev. **71**(5), 050801-1 to 050801-19 (2019).
[45]I. Fatkullin and E. Vanden-Eijnden, "A computational strategy for multiscale systems with applications to Lorenz 96 model," J. Comput. Phys. **200**(2), 605–638 (2004).
[46]D. A. Freedman, *Statistical Models: Theory and Practice* (Cambridge University Press, 2009).
[47]C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, Springer Series in Synergetics Vol. 13 (Springer, New York, 2004).
[48]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**(1), 32–57 (2010).
[49]B. Gershgorin, J. Harlim, and A. J. Majda, "Test models for improving filtering with model errors through stochastic parameter estimation," J. Comput. Phys. **229**(1), 1–31 (2010).
[50]Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," Technical Report CRG-TR-96-2 (University of Toronto, 1996).
[51]Z. Ghahramani and S. T. Roweis, "Learning nonlinear dynamical systems using an EM algorithm," in *Advances in Neural Information Processing Systems* (NeurIPS Proceedings Online Publisher, 1999), pp. 431–437.
[52]M. Ghil and S. Childress, *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics* (Springer Science & Business Media, 2012).
[53]M. Ghil and P. Malanotte-Rizzoli, "Data assimilation in meteorology and oceanography," Adv. Geophys. **33**, 141–266 (1991).
[54]A. Golightly and D. J. Wilkinson, "Bayesian inference for nonlinear multivariate diffusion models observed with error," Comput. Stat. Data Anal. **52**(3), 1674–1693 (2008).
[55]G. A. Gottwald and A. J. Majda, "A mechanism for catastrophic filter divergence in data assimilation for sparse observation networks," Nonlin. Process. Geophys. **20**(5), 705 (2013).
[56]I. Grooms and A. J. Majda, "Stochastic superparameterization in quasi-geostrophic turbulence," J. Comput. Phys. **271**, 78–98 (2014).
[57]I. G. Grooms and A. J. Majda, "Stochastic superparameterization in a one-dimensional model for wave turbulence," Commun. Math. Sci. **12**(3), 509–525 (2014).
[58]O. H. Hald and P. Stinis, "Optimal prediction and the rate of decay for solutions of the euler equations in two and three dimensions," Proc. Natl. Acad. Sci. U.S.A. **104**(16), 6527–6532 (2007).
[59]J. Harlim, A. Mahdi, and A. J. Majda, "An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models," J. Comput. Phys. **257**, 782–812 (2014).
[60]K. Hasselmann, "PIPs and POPs: The reduction of complex dynamical systems using principal interaction and oscillation patterns," J. Geophys. Res.: Atmos. **93**(D9), 11015–11021, https://doi.org/10.1029/JD093iD09p11015 (1988).
[61]S. Hijazi, G. Stabile, A. Mola, and G. Rozza, "Data-driven POD-Galerkin reduced order model for turbulent flows," J. Comput. Phys. **416**, 109513 (2020).
[62]J. D. Hol, T. B. Schon, and F. Gustafsson, "On resampling algorithms for particle filters," in *2006 IEEE Nonlinear Statistical Signal Processing Workshop* (IEEE, 2006), pp. 79–82.
[63]P. Holmes, J. L. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry* (Cambridge University Press, 1996).
[64]R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," J. Basic Eng. **83**(1), 95–108 (1961).
[65]E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability* (Cambridge University Press, 2003).
[66]D. Kelly, A. J. Majda, and X. T. Tong, "Concrete ensemble Kalman filters with rigorous catastrophic filter divergence," Proc. Natl. Acad. Sci. U.S.A. **112**(34), 10589–10594 (2015).
[67]R. Kleeman, "Information theory and dynamical system predictability," Entropy **13**(3), 612–649 (2011).
[68]J. Kokkala, A. Solin, and S. Särkkä, "Expectation maximization based parameter estimation by sigma-point and particle smoothing," in *17th International Conference on Information Fusion (FUSION)* (IEEE, 2014), pp. 1–8.
[69]D. Kondrashov, M. D. Chekroun, and M. Ghil, "Data-driven non-Markovian closure models," Physica D **297**, 33–55 (2015).
[70]S. Kravtsov, D. Kondrashov, and M. Ghil, "Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability," J. Clim. **18**(21), 4404–4424 (2005).
[71]N. R. Kristensen, H. Madsen, and S. B. Jørgensen, "Parameter estimation in stochastic grey-box models," Automatica **40**(2), 225–237 (2004).
[72]F. Kwasniok, "The reduction of complex dynamical systems using principal interaction patterns," Physica D **92**(1-2), 28–60 (1996).
[73]W. K-M Lau and D. E. Waliser, *Intraseasonal Variability in the Atmosphere-ocean Climate System* (Springer Science & Business Media, 2011).
[74]K. Law, A. Stuart, and K. Zygalakis, *Data Assimilation* (Springer, Cham, 2015), p. 214.
[75]M. Leutbecher and T. N. Palmer, "Ensemble forecasting," J. Comput. Phys. **227**(7), 3515–3539 (2008).
[76]K. K. Lin and F. Lu, "Data-driven model reduction, wiener projections, and the Koopman-Mori-Zwanzig formalism," J. Comput. Phys. **424**, 109864 (2021).
[77]R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes II: Applications* (Springer Science & Business Media, 2013), Vol. 6.
[78]E. N. Lorenz, "Predictability: A problem partly solved," in *Proceedings of Seminar on Predictability* (Cambridge University Press, 1996), Vol. 1.
[79]F. Lu, "Data-driven model reduction for stochastic Burgers equations," Entropy **22**(12), 1360 (2020).

[80]F. Lu, K. K. Lin, and A. J. Chorin, "Data-based stochastic model reduction for the Kuramoto–Sivashinsky equation," Phys. D **340**, 46–57 (2017).
[81]V. Lucarini, F. Ragone, and F. Lunkeit, "Predicting climate change using response theory: Global averages and spatial patterns," J. Stat. Phys. **166**(3-4), 1036–1064 (2017).
[82]A. Majda, *Introduction to PDEs and Waves for the Atmosphere and Ocean* (American Mathematical Society, 2003), Vol. 9.
[83]A. Majda, R. V. Abramov, and M. J. Grote, *Information Theory and Stochastics for Multiscale Nonlinear Systems* (American Mathematical Society, 2005), Vol. 25.
[84]A. J. Majda, "Challenges in climate science and contemporary applied mathematics," Commun. Pure Appl. Math. **65**(7), 920–948 (2012).
[85]A. J. Majda, *Introduction to Turbulent Dynamical Systems in Complex Systems* (Springer, 2016).
[86]A. J. Majda and M. Branicki, "Lessons in uncertainty quantification for turbulent dynamical systems," Discr. Contin. Dynam. Syst. A **32**(9), 3133–3221 (2012).
[87]A. J. Majda and N. Chen, "Model error, information barriers, state estimation and prediction in complex multiscale systems," Entropy **20**(9), 644 (2018).
[88]A. J. Majda, C. Franzke, and B. Khouider, "An applied mathematics perspective on stochastic modelling for climate," Philos. Trans. R. Soc. Lond. A **366**(1875), 2427–2453 (2008).
[89]A. J. Majda, B. Gershgorin, and Y. Yuan, "Low-frequency climate response and fluctuation–dissipation theorems: Theory and practice," J. Atmos. Sci. **67**(4), 1186–1201 (2010).
[90]A. J. Majda and I. Grooms, "New perspectives on superparameterization for geophysical turbulence," J. Comput. Phys. **271**, 60–77 (2014).
[91]A. J. Majda and M. J. Grote, "Mathematical test models for superparametrization in anisotropic turbulence," Proc. Natl. Acad. Sci. U.S.A. **106**(14), 5470–5474 (2009).
[92]A. J. Majda and J. Harlim, *Filtering Complex Turbulent Systems* (Cambridge University Press, 2012).
[93]A. J. Majda and J. Harlim, "Physics constrained nonlinear regression models for time series," Nonlinearity **26**(1), 201 (2013).
[94]A. J. Majda, I. Timofeyev, and E. V. Eijnden, "Models for stochastic climate prediction," Proc. Natl. Acad. Sci. U.S.A. **96**(26), 14687–14691 (1999).
[95]A. J. Majda, I. Timofeyev, and E. V. Eijnden, "A mathematical framework for stochastic climate models," Commun. Pure Appl. Math. **54**(8), 891–974 (2001).
[96]P. Manneville and Y. Pomeau, "Intermittency and the Lorenz model," Phys. Lett. A **75**(1-2), 1–2 (1979).
[97]G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions* (John Wiley & Sons, 2007).
[98]H. K. Moffatt, "Extreme events in turbulent flow," J. Fluid Mech. **914**, F1 (2021).
[99]A. Moosavi, R. Stefanescu, and A. Sandu, "Efficient construction of local parametric reduced order models using machine learning techniques," arXiv:1511.02909 (2015).
[100]H. Mori, "Transport, collective motion, and Brownian motion," Prog. Theor. Phys. **33**(3), 423–455 (1965).
[101]C. Mou, B. Koc, O. San, L. G. Rebholz, and T. Iliescu, "Data-driven variational multiscale reduced order models," Comp. Meth. Appl. Mech. Eng. **373**, 113470 (2021).
[102]C. Mou, Z. Wang, D. R. Wells, X. Xie, and T. Iliescu, "Reduced order models for the quasi-geostrophic equations: A brief survey," Fluids **6**(1), 16 (2021).
[103]I. J. Myung, "Tutorial on maximum likelihood estimation," J. Math. Psychol. **47**(1), 90–100 (2003).
[104]B. R. Noack, M. Morzynski, and G. Tadmor, *Reduced-Order Modelling for Flow Control* (Springer Science & Business Media, 2011), Vol. 528.
[105]K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An adaptive color-based particle filter," Image Vision Comput. **21**(1), 99–110 (2003).
[106]T. N. Palmer, "A nonlinear dynamical perspective on climate change," Weather **48**(10), 314–326 (1993).
[107]O. Papaspiliopoulos, G. O. Roberts, and O. Stramer, "Data augmentation for diffusions," J. Comput. Graph. Stat. **22**(3), 665–688 (2013).
[108]S. Pawar, S. E. Ahmed, O. San, and A. Rasheed, "Data-driven recovery of hidden physics in reduced order modeling of fluid flows," Phys. Fluids **32**(3), 036602 (2020).
[109]B. Peherstorfer and K. Willcox, "Dynamic data-driven reduced-order models," Comp. Meth. Appl. Mech. Eng. **291**, 21–41 (2015).
[110]F. Ragone, V. Lucarini, and F. Lunkeit, "A new framework for climate sensitivity and prediction: A modelling perspective," Clim. Dynam. **46**(5-6), 1459–1471 (2016).
[111]C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson, "Spectral analysis of nonlinear flows," J. Fluid Mech. **641**, 115–127 (2009).
[112]R. Salmon, *Lectures on Geophysical Fluid Dynamics* (Oxford University Press, 1998).
[113]O. San and R. Maulik, "Extreme learning machine for reduced order modeling of turbulent geophysical flows," Phys. Rev. E **97**(4), 042322 (2018).
[114]M. S. Gutiérrez, V. Lucarini, M. D. Chekroun, and M. Ghil, "Reduced-order models for coupled dynamical systems: Data-driven methods and the Koopman operator," Chaos **31**(5), 053116 (2021).
[115]S. Särkkä, *Bayesian Filtering and Smoothing* (Cambridge University Press, 2013), Vol. 3.
[116]P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," J. Fluid Mech. **656**, 5–28 (2010).
[117]S. A. Sheard and A. Mostashari, "Principles of complex systems for systems engineering," Syst. Eng. **12**(4), 295–311 (2009).
[118]F. Smarra, A. Jain, T. De Rubeis, D. Ambrosini, A. D'Innocenzo, and R. Mangharam, "Data-driven model predictive control using random forests for building energy optimization and climate control," Appl. Energy **226**, 1252–1272 (2018).
[119]P. Stinis, "Higher-order Mori-Zwanzig models for the Euler equations," Multis. Model. Simul. **6**(3), 741–760 (2007).
[120]S. H. Strogatz, *Nonlinear Dynamics and Chaos with Student Solutions Manual: With Applications to Physics, Biology, Chemistry, and Engineering* (CRC Press, 2018).
[121]K. Taira, M. S. Hemati, S. L. Brunton, Y. Sun, K. Duraisamy, S. Bagheri, S. T. M Dawson, and C.-A. Yeh, "Modal analysis of fluid flows: Applications and outlook," AIAA J. **58**(3), 998–1022 (2020).
[122]M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation," J. Am. Stat. Assoc. **82**(398), 528–540 (1987).
[123]W.-K. Tao, J.-D. Chern, R. Atlas, D. Randall, M. Khairoutdinov, J.-L. Li, D. E. Waliser, A. Hou, X. Lin, C. Peters-Lidard *et al.*, "A multiscale modeling system: Developments, applications, and critical issues," Bull. Am. Meteorol. Soc. **90**(4), 515–534 (2009).
[124]Z. Toth and E. Kalnay, "Ensemble forecasting at NCEP and the breeding method," Mon. Weather Rev. **125**(12), 3297–3319 (1997).
[125]K. E. Trenberth, J. T. Fasullo, and T. G. Shepherd, "Attribution of climate extreme events," Nat. Clim. Change **5**(8), 725–730 (2015).
[126]G. K. Vallis, *Atmospheric and Oceanic Fluid Dynamics* (Cambridge University Press, 2017).
[127]S. C. Venkataramani, R. C. Venkataramani, and J. M. Restrepo, "Dimension reduction for systems with slow relaxation," J. Stat. Phys. **167**(3), 892–933 (2017).
[128]G. Vissio and V. Lucarini, "A proof of concept for scale-adaptive parametrizations: The case of the Lorenz '96 model," Q. J. R. Meteorol. Soc. **144**(710), 63–75 (2018).
[129]Z. Y. Wan and T. P. Sapsis, "Reduced-space Gaussian process regression for data-driven probabilistic forecast of chaotic dynamical systems," Physica D **345**, 40–55 (2017).
[130]G. C. G Wei and M. A. Tanner, "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," J. Am. Stat. Assoc. **85**(411), 699–704 (1990).
[131]D. C. Wilcox, "Multiscale model for turbulent flows," AIAA J. **26**(11), 1311–1320 (1988).
[132]J. Wouters and V. Lucarini, "Disentangling multi-level systems: Averaging, correlations and memory," J. Stat. Mech. **2012**, P03003 (2012).
[133]J. Wouters and V. Lucarini, "Multi-level dynamical systems: Connecting the Ruelle response theory and the Mori-Zwanzig approach," J. Stat. Phys. **151**(5), 850–860 (2013).
[134]J. Wouters and V. Lucarini, "Multi-level dynamical systems: Connecting the Ruelle response theory and the Mori-Zwanzig approach," J. Stat. Phys. **151**(5), 850–860 (2013).

[135]X. Xie, M. Mohebujjaman, L. G. Rebholz, and T. Iliescu, "Data-driven filtered reduced order modeling of fluid flows," SIAM J. Sci. Comput. **40**(3), B834–B857 (2018).
[136]X. Yan and X. Su, *Linear Regression Analysis: Theory and Computing* (World Scientific, 2009).
[137]R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, 2001).
[138]We should have used $p^M(\widehat{\mathbf{Y}}^M|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k^M)$ to denote the conditional distribution for the CGNS given the observation $\widehat{\mathbf{X}}$, where $\widehat{\mathbf{Y}}^M$ is the analog of $\widehat{\mathbf{Y}}$ for the CGNS. We used $p^M(\widehat{\mathbf{Y}}|\widehat{\mathbf{X}}, \boldsymbol{\theta}_k^M)$ instead to avoid excessive notations.