Reduced-order models for coupled dynamical systems: data-driven methods and the Koopman operator

Santos Gutiérrez, M. ORCID: https://orcid.org/0000-0001-8617-2804, Lucarini, V. ORCID: https://orcid.org/0000-0001-9392-1471, Chekroun, M. D. ORCID: https://orcid.org/0000-0002-4525-5141 and Ghil, M. ORCID: https://orcid.org/0000-0001-5177-7133 (2021) Reduced-order models for coupled dynamical systems: data-driven methods and the Koopman operator. Chaos: An Interdisciplinary Journal of Nonlinear Science, 31 (5). 053116. ISSN 1089-7682 doi: 10.1063/5.0039496 Available at https://centaur.reading.ac.uk/98672/

Chaos

Reduced-order models for coupled dynamical systems: Data-driven methods and the Koopman operator

Cite as: Chaos **31**, 053116 (2021); https://doi.org/10.1063/5.0039496
Submitted: 03 December 2020 . Accepted: 26 April 2021 . Published Online: 17 May 2021

Manuel Santos Gutiérrez, Valerio Lucarini, Mickaël D. Chekroun, and Michael Ghil

Reduced-order models for coupled dynamical systems: Data-driven methods and the Koopman operator

Cite as: Chaos **31**, 053116 (2021); doi: 10.1063/5.0039496
Submitted: 3 December 2020 · Accepted: 26 April 2021 ·
Published Online: 17 May 2021

Manuel Santos Gutiérrez,[1,2,a)] Valerio Lucarini,[1,2] Mickaël D. Chekroun,[3,4] and Michael Ghil[4,5,6]

AFFILIATIONS

[1]Department of Mathematics and Statistics, University of Reading, Reading RG6 6AX, United Kingdom
[2]Centre for the Mathematics of Planet Earth, University of Reading, Reading RG6 6AX, United Kingdom
[3]Department of Earth and Planetary Sciences, Weizmann Institute, Rehovot 76100, Israel
[4]Department of Atmospheric and Oceanic Sciences, University of California at Los Angeles, Los Angeles, California 90095, USA
[5]Geosciences Department and Laboratoire de Météorologie Dynamique (CNRS and IPSL), Ecole Normale Supérieure and PSL University, Paris 75231, France
[6]Institute of Applied Physics of the Russian Academy of Sciences, Nizhny Novgorod 603950, Russia

[a)]**Author to whom correspondence should be addressed:** m.santos@pgr.reading.ac.uk

ABSTRACT

Providing efficient and accurate parameterizations for model reduction is a key goal in many areas of science and technology. Here, we present a strong link between data-driven and theoretical approaches to achieving this goal. Formal perturbation expansions of the Koopman operator allow us to derive general stochastic parameterizations of weakly coupled dynamical systems. Such parameterizations yield a set of stochastic integrodifferential equations with explicit noise and memory kernel formulas to describe the effects of unresolved variables. We show that the perturbation expansions involved need not be truncated when the coupling is additive. The unwieldy integrodifferential equations can be recast as a simpler multilevel Markovian model, and we establish an intuitive connection with a generalized Langevin equation. This connection helps setting up a parallelism between the top-down, equation-based methodology herein and the well-established empirical model reduction (EMR) methodology that has been shown to provide efficient dynamical closures to partially observed systems. Hence, our findings, on the one hand, support the physical basis and robustness of the EMR methodology and, on the other hand, illustrate the practical relevance of the perturbative expansion used for deriving the parameterizations.

Parameterizations aim to reduce the complexity of high-dimensional dynamical systems. Here, a theory-based and a data-driven approach for the parameterization of coupled systems are compared, showing that both yield the same stochastic multilevel structure. The results provide very strong support to the use of empirical methods in model reduction and clarify the practical relevance of the proposed theoretical framework.

I. INTRODUCTION AND MOTIVATION

Multiscale systems are typically characterized by the presence of significant variability over a large range of spatial and temporal scales. This multiscale character is due to a combination of the following factors: the nature of the external forcings; the inhomogeneity of the properties of the system's various components; the complexity of the coupling mechanisms between the components; and the variety of instabilities, dissipative processes, and feedback acting at different scales. In many cases, both the theoretical understanding of such systems and the formulation of numerical models for simulating their properties are based on focusing upon a reduced range of large spatial and long temporal scales of interest, and upon devising an efficient way to effectively capture the impact of the faster dynamical processes acting predominantly in the neglected smaller spatial scales.[24,64]

Given a high-dimensional dynamical system, we are thus interested in reformulating it in such a way that only the variables of interest are resolved. A first guess would be to ignore altogether

the unresolved variables and consider uniquely the filtered evolution laws valid for the targeted range of scales. However, this is well known to be inadequate, because nonlinearity guarantees that the unresolved variables have an impact on the resolved ones, in terms of both the detailed dynamics and its statistical properties. Therefore, the problem of constructing accurate and efficient reduced-order models—or, equivalently, of defining the coarse-grained dynamics—is an essential and fundamental aspect of studying multiscale systems, both theoretically and through numerical simulations.

For the sake of concreteness, let us consider the case of climate science. It is well-nigh impossible, therefore, given our current scientific knowledge and our available or even foreseeable technological capabilities, to create a numerical model able to directly simulate the climate system in all details for all the relevant timescales, which span a range of over 15 orders of magnitude.[30,32,66] Hence, one has to focus on a specific range of scales through suitably developed, approximate evolution equations that provide the basis for the numerical modeling. Such equations are derived from the fundamental laws of climate dynamics through systematic asymptotic expansions that are based on imposing an approximate balance between the forces acting on geophysical flows. These balance relations lead to removing small-scale, fast processes that are assumed to play a minor role at the scales of interest by filtering out the corresponding waves.[38,42,80]

In climate science, parameterization schemes have been traditionally formulated in such a way that one expresses the net impact on the scales of interest of processes occurring within the unresolved scales via deterministic functions of the resolved variables, as in the pioneering work on the parameterization of convective processes by Arakawa and Schubert.[2] More recently, it has been recognized, mostly on empirical grounds, that parameterizations should involve stochastic and non-Markovian components.[4,28,62] Machine learning methods have been proposed as the next frontier of data-driven parameterizations,[29,61,68,84,92] able to deliver a new generation of Earth system models;[73] see, though, the caveats discussed by Refs. 26 and 39.

A. The projection operator formalism of Mori and Zwanzig

Let us reformulate the problem of constructing parameterizations as a projection of the dynamics onto the set of resolved variables. By working at the level of observables, Mori[60] and Zwanzig[90] showed that the evolution laws for the projected dynamics incorporate a deterministic term that would be obtained by neglecting altogether the impact of the unresolved variables, to which a stochastic and a non-Markovian correction had to be added. Chorin *et al.*[20,21] played an important role in developing further these ideas and applying them to several important problems. We briefly recapitulate below the Mori–Zwanzig projection operator approach.

Formally, let Φ denote a generic observable defined on a state space viewed as the product of two finite-dimensional spaces $\mathcal{X} \times \mathcal{Y}$, with variables $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ being the resolved and unresolved variables, respectively. Next, let us define $\mathbb{P}$ to be a projector onto functions depending only on the target variables in $\mathcal{X}$, with the complementary projector on the unresolved variables being defined by $\mathbb{Q} = \mathrm{Id} - \mathbb{P}$.

Given a smooth flow ψ_t on $\mathcal{X} \times \mathcal{Y}$, we consider its action on smooth observables $\Phi = \Phi(\mathbf{x}, \mathbf{y})$ defined by

$$U_t \Phi(\mathbf{x}, \mathbf{y}) = \Phi(\psi_t(\mathbf{x}, \mathbf{y})), \ (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}, \ t \geq 0. \tag{1.1}$$

The operator U_t is the Koopman operator and the family $\{U_t\}_{t \geq 0}$ of Koopman operators indexed by time forms a semigroup; see Sec. II for further details. The Koopman operator describes how functions on the phase space change under the action of the flow; its time evolution obeys the following equation:

$$\partial_t (U_t \Phi) = \mathcal{L} (U_t \Phi). \tag{1.2}$$

The linear operator $\mathcal{L}$ gives the instantaneous rate of change of Φ under the action of the flow ψ_t. Representing U_t formally as the exponential of $\mathcal{L}$, $U_t = e^{t\mathcal{L}}$, is both useful and fully justified by operator semigroup theory.[65] With this representation, $\mathcal{L}$ satisfies the following identities:

$$\begin{aligned} \partial_t (U_t \Phi) &= \mathcal{L} e^{t\mathcal{L}} \Phi \\ &= e^{t\mathcal{L}} \mathcal{L} \Phi \\ &= e^{t\mathcal{L}} (\mathbb{P} + \mathbb{Q}) \mathcal{L} \Phi \qquad (1.3a) \\ &= e^{t\mathcal{L}} \mathbb{P} \mathcal{L} \Phi + e^{t\mathbb{Q}\mathcal{L}} \mathbb{Q} \mathcal{L} \Phi + \int_0^t e^{(t-s)\mathcal{L}} \mathbb{P} \mathcal{L} e^{s\mathbb{Q}\mathcal{L}} \mathbb{Q} \mathcal{L} \Phi \mathrm{d}s, \qquad (1.3b) \end{aligned}$$

where we have employed the Dyson identity[23,25] to obtain Eq. (1.3b), as will be discussed in Sec. II A. The first term in Eq. (1.3b) is the contribution of the resolved variables $\mathbf{x}$ alone to the instantaneous rate of change of Φ. The second term models the fluctuating effects of the unresolved $\mathbf{y}$-variable by itself, while the third and last term represents, via an integral, the time-delayed influence upon $\mathbf{x}$ of its interactions with $\mathbf{y}$.

This formal calculation suggests that any closed model for the $\mathbf{x}$ variables should incorporate a fluctuating term to account for the $\mathbf{y}$ contributions and a memory or integral term for the $\mathbf{y}$–$\mathbf{x}$ interactions. Unfortunately, the Mori–Zwanzig equation (1.3b)—also known as the generalized Langevin equation (GLE)[43,63]—does not provide explicit analytic formulas to determine each of the three summands in the right-hand side (RHS) of Eq. (1.3b). Hence, we need efficient ways to approximate such an equation.

In the limit of perfect timescale separation between the $\mathbf{x}$- and $\mathbf{y}$ variables, the non-Markovian term drops out and the fluctuating term can be represented as a—possibly multiplicative—white-noise term, thus recovering the basic results obtained via homogenization theory;[34,58,64] see the classical derivation by Hasselmann[93] of this result in the context of climate dynamics. When no such separation exists, however, one has to resort to finding an integral kernel beyond the abstract formulation of Eq. (1.3b); see, for instance, the theoretical ansatz based on perturbation expansion presented in Refs. 88 and 89 and discussed later in the paper, and Refs. 43 and 83 for concrete applications.

In parallel with the theoretical approaches to approximate the Mori–Zwanzig equation (1.3b), data-driven methods have been proposed to model fluctuations and memory effects arising as a result of

projecting a large state space onto (much) smaller subspaces. To this end, the work in Ref. 43 provides a rigorous connection between the Mori–Zwanzig equation (1.3b) and multilevel regression models that were initially introduced for climatological purposes in Ref. 48. More recently, the authors in Ref. 51 proposed Nonlinear Auto-Regressive Moving Average with eXogenous input (NARMAX) models as a data-driven methodology that is comparable with the Mori–Zwanzig formalism and applied such models to the deterministic Kuramoto-Sivashinsky equation and to a stochastic Burgers equation; see also Gupta and Lermusiaux.[92] The complementarity of theory-based and data-driven model reduction methods in the absence of timescale separation is very well documented in Ref. 51 as well. A highly complementary recent contribution is aimed at finding a common thread between data-driven methods and the Mori–Zwanzig theoretical framework.[52]

Efforts at using ingeniously selected basis functions as a stepping stone in data-driven methods for model reduction, effective simulation with partial data, and even prediction are multiple and flourishing. Thus, the eigenvalues and eigenfunctions of the Koopman and transfer operators have been used to capture the modes of variability of the underlying flow, regardless of the latter being deterministic or stochastic; see, respectively, Refs. 3 and 19. Dynamic mode decomposition (DMD)[72] allows one to reconstruct from observations the eigenvectors and eigenvalues of the Koopman operator for observables of interest even in high-dimensional dynamical systems.[59,69] The latter approach is complementary to the one presented herein because we shall use the eigenvectors of the Koopman operator to build the projected dynamics for the observables of interest, which can then be rewritten as a multilevel Markovian stochastic model.

Examples of other types of selection of dynamically interesting and effective bases are multichannel singular-spectrum analysis (MSSA)[1,5,31] and data-adaptive harmonic decomposition (DAHMSLM).[10,44] Both methodologies have been used extensively and successfully in the simulation, as well as the prediction, of complex phenomena.[31,41,44]

Our main theoretical result, Theorem 2.1, establishes how such spectral elements determine in turn the constitutive elements of data-driven non-Markovian closure of partially observed complex systems, when rewritten as a multilevel Markovian stochastic model. This theorem highlights, in particular, new bridges with Koopman modes and the DMD,[59,69,72] as well with other kinds of projections onto spectral bases.[10,31,37,47,67,78] A more thorough discussion of the complementary approaches involved is beyond the scope of this paper.

B. This paper

Many approaches to constructing theoretically rigorous parameterizations have been devised. These can be broadly divided into top-down and data-driven approaches: Top-down approaches aim at deriving the parameterizations by applying suitable approximations to the equations describing the dynamics of the whole systems, for instance,[35,57,83,86,88,89] while data-driven parameterizations are built by constructing a statistical-dynamical model of the impact of unresolved scales on the scales of interest. In fact, partial observations of a time-evolving system can be used to deduce the fluctuating and delayed effects of the unobserved processes, as shown in Refs. 43, 45, 48, and 85.

In this paper, we will discuss and compare the properties of the Wouters–Lucarini (WL) top-down parameterization[83,88,89] and of the empirical model reduction (EMR) data-driven parameterization.[43,45,46,48,49] We will also see when and how the integrodifferential equation occurring in the WL parameterization can be recast into a set of Markovian stochastic differential equations (SDEs). In other words, we investigate the quasi-Markovianity of the latter parameterization.[63]

The two aforementioned methodologies are conceptually and practically distinct, even though the ultimate goal of both is to provide a computationally practical approximation for the Mori–Zwanzig or GLE integrodifferential equation. In other words, both approaches—top-down and data-driven—provide fluctuations in the form of stochastic noise and memory effects determined by an integral kernel. On the one hand, the WL approach assumes prior knowledge about the decoupled hidden dynamics but no information about the statistical properties of the coupled system. The empirical approach, on the other hand, samples the observed variables evolving according to the latter. The structure of the multilevel stochastic models (MSMs) that generalize EMRs[43] allows one, moreover, to derive explicit formulas for the fluctuating and memory correction terms that parameterize the influence of hidden processes.

The overall goal of this paper is to provide a conceptual and analytical link between these two approaches, aiming, on the one hand, to buttress the practical relevance of the WL perturbative approach and, on the other hand, to provide further insight into the well-documented robustness of the EMR method. Moreover, we will clarify how multilevel systems arise from both the *top-down* (WL) and the *bottom-up* (EMR) approaches. The paper explores the complete set of boxes and explains all the arrows in Fig. 1. The diagram in the figure shows that, starting from the top box, one can arrive at a memory equation, like (1.3b), via *top-down* or *bottom-up* methods, as indicated by the left and right sequence of arrows, respectively.

The paper is structured as follows. Section II revisits the derivation of the WL parameterization method by applying the Dyson expansion to the Koopman operator associated with two weakly coupled dynamical systems. We show that such an expansion need not be truncated for additively coupled models and consider more general coupling laws than those in Ref. 89.

Furthermore, we study the problem of finding Markovian representations of the memory equation in the WL approach based on the spectral decomposition of the Koopman semigroup. Specifically, Theorem 2.1 shows, in the case of a scalar observable, how to recast the stochastic integrodifferential equation arising in the WL parameterization as a multilevel Markovian stochastic system involving explicitly the spectral elements of the (uncoupled) Koopman operator, and we point out in Remark 2.5 how such a Markovianization extends to the multidimensional setting. Section III provides new insights into the Markovian representation adopted in the MSM framework; these insights help one to determine, in particular, the number of levels required for EMR to converge.

Section IV presents a comparison of the data-driven and top-down parameterization approaches using a simple conceptual stochastic climate model. Finally, we discuss the conclusions obtained from this investigation in Sec. V.

Appendixes are included in order to avoid diverting the attention of the reader from the main message of the paper. Appendix A provides the proof of the key Theorem 2.1. Appendix B briefly revisits the spectral decomposition of correlation functions and provides a criterion to quantify the loss of the semigroup property that is useful in identifying non-Markovian effects. Appendix C discusses the stochastic Itô integration of the elementary form of an MSM. Appendix D shows how EMR approaches can capture the dynamical properties of partially observed systems; in it, we consider a simple climate model, obtained by coupling the Lorenz atmospheric model[54] (L84) and the Lorenz convection model[53] (L63).

II. REVISITING THE WEAK-COUPLING-LIMIT PARAMETERIZATION

To study dynamical systems in which one can separate the variables into two groups, with weak coupling between the two, one often resorts to the so-called parameterizations of the effects of one group on the other. In the weak-coupling limit, the coupling itself can be treated as a perturbation of the main dynamics.[56,88,89] Granted such an assumption and some degree of structural stability of the system, one can apply response theory to derive explicit stochastic and memory terms to describe the impact of the variables we want to neglect on the variables of interest, in the Mori–Zwanzig spirit. Note that, to do so, no assumption on timescale separation between the two groups of variables is necessary. This point is particularly relevant in fields like climate sciences, where no clear timescale separation is observed so that asymptotic expansions of the kind used in homogenization theory are of limited utility.

A. Deriving the WL approximation for the Mori–Zwanzig formalism

Here, using a perturbative approach, we review the derivation of the parameterization presented in Refs. 56, 88, and 89. Formally, we want to couple two dynamical systems generated independently by two vector fields $\mathbf{F}:\mathcal{X}\subseteq\mathbb{R}^{d_1}\longrightarrow\mathcal{X}$ and $\mathbf{G}:\mathcal{Y}\subseteq\mathbb{R}^{d_2}\longrightarrow\mathcal{Y}$ with possibly $d_1\neq d_2$ and typically $d_1\ll d_2$. We study a broad class of systems of the form

$$\dot{\mathbf{x}}(t)=\mathbf{F}(\mathbf{x}(t))+\epsilon\mathbf{C}_{\mathbf{x}}^{\mathbf{x}}(\mathbf{x}(t)):\mathbf{C}_{\mathbf{y}}^{\mathbf{x}}(\mathbf{y}(t)), \tag{2.1a}$$

$$\dot{\mathbf{y}}(t)=\mathbf{G}(\mathbf{y}(t))+\epsilon\mathbf{C}_{\mathbf{x}}^{\mathbf{y}}(\mathbf{x}(t)):\mathbf{C}_{\mathbf{y}}^{\mathbf{y}}(\mathbf{y}(t)). \tag{2.1b}$$

The operation indicated by the colon $\mathbf{x}:\mathbf{y}$ denotes the Hadamard product that multiplies vectors or matrices component-wise. Here, four new vector fields have been introduced to model the coupling law, namely, $\mathbf{C}_{\mathbf{x}}^{\mathbf{x}}:\mathcal{X}\longrightarrow\mathcal{X}$, $\mathbf{C}_{\mathbf{x}}^{\mathbf{y}}:\mathcal{X}\longrightarrow\mathcal{Y}$, $\mathbf{C}_{\mathbf{y}}^{\mathbf{x}}:\mathcal{Y}\longrightarrow\mathcal{X}$, and $\mathbf{C}_{\mathbf{y}}^{\mathbf{y}}:\mathcal{Y}\longrightarrow\mathcal{Y}$.

The real parameter ϵ controls the strength of the coupling between the two groups of variables, $\mathbf{x}(t)$ and $\mathbf{y}(t)$, so that the $\mathbf{x}$ and $\mathbf{y}$ variables are uncoupled for $\epsilon=0$. We assume that the vector fields $\mathbf{F}$ and $\mathbf{G}$, as well as the coupling laws in Eqs. (2.1a) and (2.1b), are such that the system (2.1) possesses a global attractor. Furthermore, we assume throughout this article that this global attractor supports an physical invariant probability measure μ that describes the distribution of trajectories onto the global attractor.

The WL parameterization views the coupling as an ϵ-perturbation of the otherwise independent $\mathbf{x}$- and $\mathbf{y}$-processes, with $\mathbf{x}$ being the observed and $\mathbf{y}$ being the hidden variables. One next assumes that the impacts of perturbations applied to these processes can be addressed using response theory[70,71] so that response formulas can be used to derive an effective equation for the $\mathbf{x}$ variables.

Taking the Mori–Zwanzig[60,90] point of view, we wish to calculate the evolution of observables that depend on the observed variables $\mathbf{x}$ alone, $\Phi=\Phi(\mathbf{x}(t))$. The idea, following Ref. 89, is to perform a perturbative expansion of the differential operator $\mathcal{L}$ governing the evolution of $\Phi(\mathbf{x}(t))$ under the action of the flow associated with Eq. (2.1). Denoting by $u(\mathbf{x},\mathbf{y},t)$ the time evolution of a smooth observable Φ in $\mathcal{C}^\infty(\mathcal{X}\times\mathcal{Y})$, the first step of this Dyson-like operator expansion reads as follows:

$$\partial_t u=\mathcal{L}u=(\mathcal{L}_0+\epsilon\mathcal{L}_1)u. \tag{2.2}$$

Here, $\mathcal{L}_0$ and $\mathcal{L}_1$ account for the advective effects of the uncoupled and coupling terms, respectively, that compose the RHS of Eq. (2.1), namely,

$$\mathcal{L}_0=\begin{bmatrix}\mathbf{F}(\mathbf{x})\\ \mathbf{G}(\mathbf{y})\end{bmatrix}\cdot\begin{bmatrix}\nabla_{\mathbf{x}}\\ \nabla_{\mathbf{y}}\end{bmatrix}, \tag{2.3a}$$

$$\mathcal{L}_1=\begin{bmatrix}\mathbf{C}_{\mathbf{x}}^{\mathbf{x}}(\mathbf{x}):\mathbf{C}_{\mathbf{y}}^{\mathbf{x}}(\mathbf{y})\\ \mathbf{C}_{\mathbf{x}}^{\mathbf{x}}(\mathbf{x}):\mathbf{C}_{\mathbf{y}}^{\mathbf{x}}(\mathbf{y})\end{bmatrix}\cdot\begin{bmatrix}\nabla_{\mathbf{x}}\\ \nabla_{\mathbf{y}}\end{bmatrix}, \tag{2.3b}$$

in Eq. (2.3), $\nabla_{\mathbf{x}}$ and $\nabla_{\mathbf{y}}$ denote the vector differential operators with respect to the variables $\mathbf{x}$ and $\mathbf{y}$.

Recalling Eq. (1.2), the solution operator of Eq. (2.2) is the Koopman operator. Formally, its dual acts on densities and it is the so-called *transfer operator*.[3,50] Equation (2.2) is thus a transport equation, where the physical quantity or observable is advected by the vector field on the RHS of Eq. (2.1).

Note that the operator formalism presented here in the deterministic dynamical systems setting—and the associated semigroup

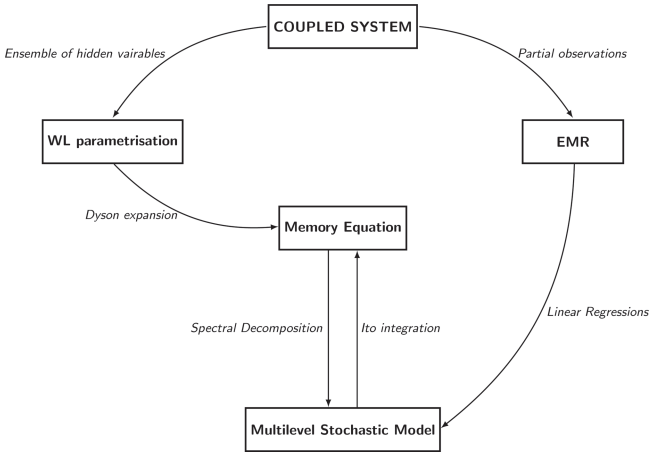


FIG. 1. Schematic view of the two complementary approaches studied in this paper. The arrows on the left-hand side indicate top-down, perturbative parameterizations; on the right, they refer to bottom-up, empirical parameterizations.

theory—extends to Markov diffusion processes driven by a stochastic forcing.[19] In the latter case, the transport equation (2.2) becomes the so-called backward Kolmogorov equation that describes the evolution of the expected value of observables. Loosely speaking, the corresponding extension amounts to adding a Laplacian-like operator to the advection operator $\mathcal{L}$.[19,63] See Ref. 36 for the appropriate context in testing the applicability of response theory to the independent $\mathbf{x}$ and $\mathbf{y}$ processes, when $\epsilon = 0$.

More precisely, one associates with the solution $u(\mathbf{x}, \mathbf{y}, t)$ of Eq. (2.2) unfolding from an observable $\Phi = \Phi(\mathbf{x}, \mathbf{y})$ at time $t = 0$, a family of linear Koopman operators indexed by time $\{U_t\}_{t\geq 0}$ such that $u(\mathbf{x}, \mathbf{y}, t) = U_t\Phi(\mathbf{x}, \mathbf{y})$, for any $t \geq 0$ and $(\mathbf{x}, \mathbf{y})$ in $\mathcal{X} \times \mathcal{Y}$. These operators are defined—as mentioned already in connection with introducing the GLE (1.3b)—as exponentials of the operator $\mathcal{L}$, i.e., $U_t = e^{t\mathcal{L}}$. This notation is formal as the operator $\mathcal{L}$ is unbounded; it is, however, usable as $\{U_t\}_{t\geq 0}$ satisfies the semigroup property, i.e., $U_{t+s} = U_tU_s$, $t, s \geq 0$, as for a standard exponential. Over the appropriate function space [Such a space can be chosen, for instance, as $D_p = \{\Phi \in L^p_\mu(\mathcal{X} \times \mathcal{Y}) \,|\, A\Phi = \lim_{t\to 0}\ t^{-1}(U_t\Phi - \Phi) \text{ exists}\}$ for some $p \in [1, \infty]$, with μ denoting a relevant invariant measure of the system (2.1) supported by the global attractor, while the limit is taken in the sense of strong convergence.[25]] of observables Φ, this family actually forms a strongly continuous contracting semigroup.[25]

The action of the flow on an observable Φ becomes thus more transparent thanks to the operator U_t, according to the equation

$$U_t\Phi(\mathbf{x}_0, \mathbf{y}_0) = e^{t\mathcal{L}}\Phi(\mathbf{x}_0, \mathbf{y}_0) = \Phi(\mathbf{x}(t; \mathbf{x}_0), \mathbf{y}(t; \mathbf{y}_0)), \tag{2.4}$$

where $(\mathbf{x}(t; \mathbf{x}_0), \mathbf{y}(t; \mathbf{y}_0))$ denotes the system's solution at time t emanating from the initial state $(\mathbf{x}_0, \mathbf{y}_0)$ at time $t = 0$. In what follows, we omit the subscript 0 in $(\mathbf{x}_0, \mathbf{y}_0)$ but still take it as an initial state.

The semigroup $\{U_t\}_{t\geq 0}$ is known as the Koopman semigroup and for each t, U_t is Koopman operator mentioned above; see also Ref. 43, Sec. 4. When the coupling parameter ϵ in system (2.1) is small, one can use formal perturbation expansions of the Koopman semigroup to better isolate and assess the coupling effects at the level of observables. To do so, we follow here the perturbation expansion first introduced by Freeman J. Dyson in the context of quantum electrodynamics[23] and later formulated rigorously in mathematical terms in Ref. 33. Formally, this expansion reads as follows:

$$\begin{aligned} U_t\Phi(\mathbf{x}, \mathbf{y}) &= e^{t\mathcal{L}}\Phi(\mathbf{x}, \mathbf{y}) \\ &= e^{t\mathcal{L}_0 + t\epsilon\mathcal{L}_1}\Phi(\mathbf{x}, \mathbf{y}) && \text{(2.5a)} \\ &= e^{t\mathcal{L}_0}\Phi(\mathbf{x}, \mathbf{y}) + \epsilon\int_0^t e^{s\mathcal{L}}\mathcal{L}_1 e^{(t-s)\mathcal{L}_0}\Phi(\mathbf{x}, \mathbf{y})\mathrm{d}s && \text{(2.5b)} \\ &= e^{t\mathcal{L}_0}\Phi(\mathbf{x}, \mathbf{y}) + \epsilon\int_0^t e^{(t-s)\mathcal{L}_0}\mathcal{L}_1 e^{s\mathcal{L}}\Phi(\mathbf{x}, \mathbf{y})\mathrm{d}s, && \text{(2.5c)} \end{aligned}$$

and it yields the following expansion of the Koopman operator in ϵ:

$$U_t\Phi(\mathbf{x}, \mathbf{y}) = e^{t\mathcal{L}_0}\Phi(\mathbf{x}, \mathbf{y}) + \epsilon\int_0^t e^{(t-s)\mathcal{L}_0}\mathcal{L}_1 e^{s\mathcal{L}_0}\Phi(\mathbf{x}, \mathbf{y})\mathrm{d}s + \mathcal{O}\left(\epsilon^2\right). \tag{2.6}$$

This identity shows that the evolution of a generic observable can be described as an ϵ-perturbation of its decoupled evolution according to $\mathcal{L}_0$. We note that these expansions are purely formal and, in particular, it is not clear in which sense this expansion might converge. For a bounded perturbation operator $\mathcal{L}_1$, it would be straightforward to prove boundedness of the resulting perturbed semigroup. However, $\mathcal{L}_1$ here is a differential linear operator, for which direct estimates are more laborious. Leaving aside the functional analysis framework that would make such an expansion rigorously convergent, we shall use nevertheless the expansion (2.6) throughout this article.

The objective now is, using this operator expansion, to derive an effective reduced-order model for the evolution of the $\mathbf{x}$-variable without having to resolve the $\mathbf{y}$-process. We start observing the system at $t = 0$, but assume that it has already attained a steady state. Since we are only concerned with observables depending solely on the $\mathbf{x}$ variables, we formulate now an evolution equation for such observables. To do so, we consider first the Liouville equation for a generic $\mathbf{y}$-independent observable Φ, this is, $\Phi(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x})$, for every $\mathbf{x}$ and $\mathbf{y}$.

For such an observable, at the time we start observing the coupled system, Eq. (2.2) reduces to

$$\partial_t\left(U_t\Phi\right)|_{t=0} = \left[\mathbf{F}(\mathbf{x}) + \epsilon\mathbf{C}^{\mathbf{x}}_{\mathbf{x}}(\mathbf{x}) : \mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y})\right] \cdot \nabla_{\mathbf{x}}\Phi, \tag{2.7}$$

where $\cdot$ denotes an inner product in $\mathcal{X}$. Equation (2.7) illustrates the trivial fact that the time evolution in Eq. (2.1) of an $\mathbf{x}$-dependent physical quantity is also affected by the $\mathbf{y}$ variables.

Following Refs. 88 and 89, the decoupled equations are assumed to have been evolving for some time prior to the coupling. Hence, we have to formally parameterize the evolution of the $\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y})$-contribution to the vector field which is, ultimately, a vector-valued observable.

We do so by introducing an extended version of the Koopman operators that act on vectors component-wise, rather than just on real-valued observables. Consider $\mathbf{v} : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}^d$, for some positive integer d, and define the action of the Koopman operator $e^{t\mathcal{L}}$ on $\mathbf{v}$ as

$$\left[e^{t\mathcal{L}}\mathbf{v}(\mathbf{x}, \mathbf{y})\right]_i = e^{t\mathcal{L}}\left[\mathbf{v}(\mathbf{x}, \mathbf{y})\right]_i \tag{2.8}$$

for every $i = 1, \ldots, d$. The definition (2.8) will allow us to use the semigroup notation for observables of possibly different dimensions, all of which take their inputs in the phase space $\mathcal{X} \times \mathcal{Y}$. Ultimately, this is a component-wise evaluation of our extended Koopman operator family, and its generator can be obtained analogously. As mentioned above, we have to model the effects of the coupling vector field $\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y})$, whose state at time $t = 0$ is the product of the evolution from time $-t$ to 0. We then have, with the dynamics starting at time $-t$ and initial state $(\mathbf{x}_0, \mathbf{y}_0)$,

$$\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}) = e^{t\mathcal{L}}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0, -t) = e^{t\mathcal{L}}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0). \tag{2.9}$$

Now, by using the perturbative expansions in Eqs. (2.5b)–(2.6), we obtain

$$\begin{aligned}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}) &= e^{t\mathcal{L}}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)\\ &= e^{t\mathcal{L}_0+t\epsilon\mathcal{L}_1}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0) && (2.10a)\\ &= e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)+\epsilon\int_0^t e^{s\mathcal{L}}\mathcal{L}_1 e^{(t-s)\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)\,ds && (2.10b)\\ &= e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)+\epsilon\int_0^t e^{(t-s)\mathcal{L}_0}\mathcal{L}_1 e^{s\mathcal{L}}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)\,ds && (2.10c)\\ &= e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)+\epsilon\int_0^t e^{(t-s)\mathcal{L}_0}\mathcal{L}_1 e^{s\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)\,ds+\mathcal{O}(\epsilon^2). && (2.10d)\end{aligned}$$

Plugging the identity in Eq. (2.10c) into Eq. (2.7), we find the following expression:

$$\partial_t(U_t\Phi)|_{t=0} = \left[\mathbf{F}(\mathbf{x})+\epsilon\mathbf{C}^{\mathbf{x}}_{\mathbf{x}}(\mathbf{x}) : \left\{e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0) + \epsilon\int_0^t e^{s\mathcal{L}}\mathcal{L}_1 e^{(t-s)\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)\,ds\right\}\right]\cdot\nabla_{\mathbf{x}}\Phi. \quad (2.11)$$

This equation is an exact reformulation of the problem induced by Eq. (2.7). This reformulation demonstrates that memory effects enter at second order in powers of the coupling parameter. Notice, though, that even if Eq. (2.11) reduces the dimensionality of the problem from d_1+d_2 to d_1, it does not constitute an approximation for the evolution of Φ as an observable of $\mathbf{x}$ alone, since it depends on the evolution of the $\mathbf{y}$ variables in the coupled regime by means of the action of $e^{s\mathcal{L}}$ onto $\mathcal{L}_1$.

Therefore, we need to perform a further approximation by considering Eq. (2.10d) instead, which leads to

$$\partial_t(U_t\Phi)|_{t=0} \simeq \left[\mathbf{F}(\mathbf{x})+\epsilon\mathbf{C}^{\mathbf{x}}_{\mathbf{x}}(\mathbf{x}) : \left\{e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0) + \epsilon\int_0^t e^{(t-s)\mathcal{L}_0}\mathcal{L}_1 e^{s\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)\,ds\right\}\right]\cdot\nabla_{\mathbf{x}}\Phi, \quad (2.12)$$

where the terms of order ϵ^3 have been dropped. Equation (2.12) is our equivalent of the Dyson approximation in quantum electrodynamics; it approximates the evolution of the $\mathbf{x}$ variables with no need for the evolution of the $\mathbf{y}$ variables in the coupled regime.

This result amounts to saying that—by observing only the statistical properties of the decoupled dynamics of the $\mathbf{y}$-process, obtained with $\epsilon = 0$—one can construct a Markovian contribution

$$\mathbf{C}^{\mathbf{x}}_{\mathbf{x}}(\mathbf{x}) : \left\{e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)\right\},$$

and a non-Markovian contribution

$$\mathbf{C}^{\mathbf{x}}_{\mathbf{x}}(\mathbf{x}) : \left\{\int_0^t e^{(t-s)\mathcal{L}_0}\mathcal{L}_1 e^{s\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)\,ds\right\},$$

to the dynamics of the $\mathbf{x}$ variables that is able to describe the impact of the coupling.

Expanding the kernel $\tilde{\mathcal{K}}$ of the memory contribution, we get

$$\begin{aligned}\tilde{\mathcal{K}}(t,s,\mathbf{x}_0,\mathbf{y}_0) :&= e^{(t-s)\mathcal{L}_0}\mathcal{L}_1 e^{s\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0) && (2.13a)\\ &= e^{(t-s)\mathcal{L}_0}\left(\left[\mathbf{C}^{\mathbf{x}}_{\mathbf{x}}(\mathbf{x}_0) : \mathbf{C}^{\mathbf{y}}_{\mathbf{y}}(\mathbf{y}_0)\right]\cdot\nabla_{\mathbf{x}}\right.\\ &\quad\left.+\left[\mathbf{C}^{\mathbf{y}}_{\mathbf{x}}(\mathbf{x}_0) : \mathbf{C}^{\mathbf{y}}_{\mathbf{y}}(\mathbf{y}_0)\right]\cdot\nabla_{\mathbf{y}}\right)e^{s\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0) && (2.13b)\\ &= \left[e^{(t-s)\mathcal{L}_0}\left(\mathbf{C}^{\mathbf{y}}_{\mathbf{x}}(\mathbf{x}_0) : \mathbf{C}^{\mathbf{y}}_{\mathbf{y}}(\mathbf{y}_0)\right)\right]\cdot\nabla_{\mathbf{y}}e^{s\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0). && (2.13c)\end{aligned}$$

Note that the leading-order Koopman operator $e^{s\mathcal{L}_0}$ models the evolution of the observables in the uncoupled regime. Since there is no prior knowledge on initializing the coupled system at time $-t$, the initial state $\mathbf{y}_0$ in the hidden variables should be drawn from an ensemble, according to a probability density function. At this stage, there is freedom in the choice of such a prior. However, since we are assuming that the coupled system was initialized at time $-t$, it is natural to draw $\mathbf{y}_0$ according to the invariant measure ν associated with the dynamical system generated by the vector field $\mathbf{G}$ from Eq. (2.1).

We wish to sample initial conditions from the coupled steady state, but do not assume any prior knowledge of the coupled statistics. As discussed in Refs. 88 and 89, we can take advantage of response theory to address this situation. Indeed, for any sufficiently smooth observable Ψ, we have

$$\langle\Psi\rangle_\epsilon = \langle\Psi\rangle_{\epsilon=0} + \sum_{k=1}^{\infty}\epsilon^k\delta_k[\Psi],$$

where $\langle\Psi\rangle_\epsilon$ is the expectation value of Ψ in the coupled system (2.1), $\langle\Psi\rangle_{\epsilon=0}$ is the expectation value of Ψ according to the statistics generated by the uncoupled $\mathbf{y}$ process obtained by setting $\epsilon = 0$ in Eq. (2.1b), and $\epsilon^k\delta_k[\Psi]$ is the kth-order response. In what follows, we remove the subscripts for the averages when $\epsilon = 0$. Therefore, we have that the expected value of the coupling function reads as

$$\left\langle\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}\right\rangle_\epsilon = \left\langle\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}\right\rangle + \sum_{k=1}^{\infty}\epsilon^k\delta_k[\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}]. \quad (2.14)$$

Likewise, we can calculate the average of such function at time t,

$$\left\langle e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}\right\rangle_\epsilon = \left\langle e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}\right\rangle + \sum_{k=1}^{\infty}\epsilon^k\delta_k[e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}]. \quad (2.15)$$

Now, by letting $\tilde{\eta}(t,\mathbf{y}_0) = e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)$, we find that in order for the approximate statistics to agree up to second order in ϵ with the exact one, we only need to impose the following conditions upon the first two moments of the parameterized noisy fluctuations (see also Ref. 88):

$$\left\langle\tilde{\eta}(t,\mathbf{y}_0)\right\rangle = \int\nu(d\mathbf{y}_0)\,\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0), \quad (2.16a)$$

$$\left\langle\tilde{\eta}(t,\mathbf{y}_0)\tilde{\eta}^\top(0,\mathbf{y}_0)\right\rangle = \int\nu(d\mathbf{y}_0)e^{t\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)\left(\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0)\right)^\top, \quad (2.16b)$$

where $(\cdot)^\top$ stands for the transpose of a vector or a matrix. It follows that any stochastic noise $\eta(t)$ that satisfies the two conditions

above will be suitable for parameterizing the fluctuations in the **y**-dynamics tied to the lack of knowledge in the initial state. Each of the entries in the correlation matrix given by Eq. (2.16b) is the correlation function between the components of the vector field $\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}$ and these will become explicit provided a suitable spectral decomposition is at hand. Such a decomposition will be provided later in Sec. II C, although the reader is referred at this point to Appendix B for clarity.

In the memory term, though, we neglect ϵ-corrections to its statistics since memory effects are of order ϵ^2 already. Thus, we have by averaging the kernel $\tilde{\mathcal{K}}(t,s,\mathbf{x}_0,\mathbf{y}_0)$ in Eq. (2.13b) with respect to the **y** variables,

$$\mathcal{K}(t,s,\mathbf{x}) := \int \nu(\mathrm{d}\mathbf{y}_0)\left[e^{(t-s)\mathcal{L}_0}\left(\mathbf{C}^{\mathbf{y}}_{\mathbf{x}}(\mathbf{x}_0):\mathbf{C}^{\mathbf{y}}_{\mathbf{y}}(\mathbf{y}_0)\right)\right]\cdot\nabla_{\mathbf{y}}e^{s\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0) \tag{2.17a}$$

$$= \int \nu(\mathrm{d}\mathbf{y}_0)\left[e^{(t-s)\mathcal{L}_0}\mathbf{C}^{\mathbf{y}}_{\mathbf{x}}(\mathbf{x}_0):e^{(t-s)\mathcal{L}_0}\mathbf{C}^{\mathbf{y}}_{\mathbf{y}}(\mathbf{y}_0)\right]\cdot\nabla_{\mathbf{y}}e^{s\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0) \tag{2.17b}$$

$$= \int \nu(\mathrm{d}\mathbf{y}_0)\left[\mathbf{C}^{\mathbf{y}}_{\mathbf{x}}(\mathbf{x}(t-s)):e^{(t-s)\mathcal{L}_0}\mathbf{C}^{\mathbf{y}}_{\mathbf{y}}(\mathbf{y}_0)\right]\cdot\nabla_{\mathbf{y}}e^{s\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}_0). \tag{2.17c}$$

This way, the memory kernel only depends on the **x** variables. Hence, we find a self-consistent evolution of the **x** variables, subject to the influence of unobserved variables **y**, in the form of a stochastic integrodifferential equation (SIDE) resembling the GLE (1.3b),

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}) + \epsilon\mathbf{C}^{\mathbf{x}}_{\mathbf{x}}(\mathbf{x}):\eta(t) + \epsilon^2\mathbf{C}^{\mathbf{x}}_{\mathbf{x}}(\mathbf{x})\cdot\int_0^t \mathcal{K}(t,s,\mathbf{x})\mathrm{d}s, \tag{2.18}$$

where $\eta(t)$ is a stochastic forcing that agrees with the mean and correlation properties stated in Eq. (2.16).

We emphasize that the solution $\mathbf{x}(t)$ of the original system of ordinary differential equations (2.1) does not satisfy Eq. (2.18): it is just the proposed reduced-order model for the **x** variables. The closure provided by expressing the corrections in the second and third term on the RHS of Eq. (2.18) as functions of **x** alone is typically called a parameterization of the effect of the unobserved **y** variables in the climate sciences.[32]

Note that there is considerable freedom in choosing the noise, since we only require agreement up to the second moment. However, a direct consequence of this weak-coupling parameterization is that realizations of the noise can be produced by directly integrating the decoupled hidden variables or by representing it using simple autoregressive models.[82] We are assuming here that the uncoupled dynamics leads to a noisy signal; this can be achieved either by the presence of stochastic forcing in the hidden variables[82,87] or by their uncoupled dynamics being chaotic.[83]

To summarize, the weak-coupling limit allows one to develop a parameterization of the hidden variables for a system of coupled equations where no separation of timescales is assumed. Moreover, this approach provides explicit approximate expressions for the deterministic, stochastic, and non-Markovian terms in the Mori–Zwanzig formalism of Eq. (1.3b).

There are two sources of error in the parameterization proposed in Eq. (2.18). First, the truncation performed in the Dyson expansion neglects higher-order effects, which are weighted by the third power of the coupling parameter ϵ. Second, averaging over the statistics of the uncoupled dynamics can also introduce errors. Furthermore, the nature of the stochastic correction is not fully determined except for its lagged correlation.

The perturbation operator approach taken here is analogous to that of Ref. 89, who only considered the independent or additive-coupling cases; the latter is expanded upon in Sec. II B. Here, though, we generalize further the parameterization formulas that can be obtained via perturbative expansion of linear operators. In fact, the present approach can also be extended to weakly coupled systems of the form

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) + \epsilon\mathbf{C}^{\mathbf{x}}\left(\mathbf{y}(t)\right), \tag{2.19a}$$

$$\dot{\mathbf{y}}(t) = \mathbf{G}(\mathbf{y}(t)) + \epsilon\mathbf{C}^{\mathbf{y}}\left(\mathbf{x}(t),\mathbf{y}(t)\right), \tag{2.19b}$$

where $\mathbf{C}^{\mathbf{y}}$ encodes interactions that need not be separable between the **x** and **y** variables in the hidden layer of the model. Note that the full parameterization of arbitrary couplings was discussed by the two authors of Ref. 89 in the previous work,[88] in which they used a response-theoretic approach.

B. The additive-coupling case

The approximate Dyson expansion given in Eq. (2.12) is exact in the case of additive coupling. Such systems take the form

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) + \epsilon\mathbf{C}^{\mathbf{x}}\left(\mathbf{y}(t)\right), \tag{2.20a}$$

$$\dot{\mathbf{y}}(t) = \mathbf{G}(\mathbf{y}(t)) + \epsilon\mathbf{C}^{\mathbf{y}}\left(\mathbf{x}(t)\right). \tag{2.20b}$$

Indeed, letting $\mathbf{C}^{\mathbf{y}}(\mathbf{x},\mathbf{y}) = \mathbf{C}^{\mathbf{y}}(\mathbf{x})$ in Eq. (2.19b) and using Eq. (2.10b) allow us to avoid the truncation of the Dyson expansion and yield the following expression for the memory term:

$$\tilde{\mathcal{K}}(t,s,\mathbf{x},\mathbf{y}_0) = e^{s\mathcal{L}}\mathcal{L}_1 e^{(t-s)\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}\left(\mathbf{y}_0\right) \tag{2.21a}$$

$$= e^{s\mathcal{L}}\left(\mathbf{C}^{\mathbf{x}}(\mathbf{y}_0)\cdot\nabla_{\mathbf{x}} + \mathbf{C}^{\mathbf{y}}(\mathbf{x})\cdot\nabla_{\mathbf{y}}\right)e^{(t-s)\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}\left(\mathbf{y}_0\right) \tag{2.21b}$$

$$= e^{s\mathcal{L}}\left(\mathbf{C}^{\mathbf{y}}(\mathbf{x})\right)\cdot\nabla_{\mathbf{y}}e^{(t-s)\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}\left(\mathbf{y}_0\right), \tag{2.21c}$$

which is exact. Next, taking averages with respect to ν, we obtain

$$\mathcal{K}(t,s,\mathbf{x}) = \int \nu(\mathrm{d}\mathbf{y}_0)e^{s\mathcal{L}}\mathbf{C}^{\mathbf{y}}(\mathbf{x})\cdot\nabla_{\mathbf{y}}e^{(t-s)\mathcal{L}_0}\mathbf{C}^{\mathbf{x}}\left(\mathbf{y}_0\right). \tag{2.22}$$

Hence, the parameterization in this additive-coupling case is exact, as no terms proportional to ϵ^k, $k \geq 3$ are present. The only assumption made is that the statistics in the **y** variables have reached a steady state according to the unperturbed system. Finally, the full SIDE in this case takes the form

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}) + \epsilon\eta(t) + \epsilon^2\int_0^t \mathcal{K}(t,s,\mathbf{x})\mathrm{d}s, \tag{2.23}$$

where the stochastic process η has the mean and correlation properties given by Eq. (2.16). This equation is, thus, exactly the one obtained in Ref. 89.

Memory effects represented by integral terms seem unavoidable unless the memory kernels vanish quickly with respect to time. Infinite timescale separation between the two sets of variables leads, though, to the vanishing of the associated integral expressions.[64] Here, we are not assuming no such property in the coupled dynamical system under study; see Eqs. (2.1) and (2.20). On the other hand, reduced phase spaces can help explain the statistics of the dynamical system without resorting to delayed effects that entail the integrals in Eqs. (2.18) and (2.23). Following Ref. 19, we briefly review in Appendix B a criterion based on Koopman operators—and, more generally, Markov operators—that enables one to decide whether memory effects can help explain the dynamics and statistics in reduced phase spaces.

C. Markovian representation through leading Koopman eigenfunctions

In the context of Langevin dynamics, there are known conditions on memory kernels that allow one to recast certain stochastic integrodifferential equations into a Markovian SDE by means of extended variables; see Ref. 63, Sec. 8.2. The stochastic processes that allow such a procedure are called quasi-Markovian.[63] (Such a Markovianization procedure is actually not limited to stochastic processes and it relies on the same type of ideas in other contexts; see Refs. 7, 9, 22, and 43, Sec. 1.3.)

This Markovianization theory can be formulated in the setting of near-equilibrium statistical mechanics, where one uses fluctuation-dissipation-like relations that link the decay properties of the memory kernel and the decorrelation rates of the fluctuations. Here, we follow the approach in Ref. 63 but without making any assumptions on the Hamiltonian behavior of the $\mathbf{y}$ variables. We need, though, to make assumptions on the spectral properties of the generator of the $\mathbf{y}$-dynamics, as explained below.

We define the generator $\mathcal{L}_0^{\mathbf{y}}$ of the Koopman semigroup associated with the $\mathbf{y}$-dynamics by

$$\mathcal{L}_0^{\mathbf{y}}\Phi(\mathbf{y}) = \mathbf{G}(\mathbf{y}) \cdot \nabla\Phi(\mathbf{y}), \tag{2.24}$$

for every real-valued observable $\Phi \in \mathcal{C}^\infty\left(\mathbb{R}^{d_2}\right)$ and we denote the associated Koopman operator at time t by $U_t = e^{\mathcal{L}_0^{\mathbf{y}} t}$; the subscript $\mathbf{y}$ has been dropped from the ∇ operator herewith, for notational clarity. Recall that the spectrum of such operators provides useful insights into the statistical properties of the system; this topic is beyond the scope of the present paper but it is treated in detail in Ref. 19.

It suffices to show below that the spectrum of U_t allows one to characterize the constitutive ingredients of the WL parameterization (2.18) and (2.23), subject to natural assumptions. Even though we have clarified in Eq. (2.8) how the Koopman operator acts on vector-valued observables of any dimension, we restrict now its action for simplicity to scalar real-valued observables, as in Eq. (2.24). In this case, along the lines of the methodology of dynamic mode decomposition,[59,69,72] we can (formally) decompose the operator as

$$e^{\mathcal{L}_0^{\mathbf{y}} t} = \sum_{j=1}^{N} e^{t\lambda_j}\Pi_j + \mathcal{R}(t), \tag{2.25}$$

where $\{\lambda_j\}_{j=1}^N$ are the eigenvalues that form the point spectrum of $\mathcal{L}_0^{\mathbf{y}}$ and Π_j is the spectral projector onto the eigenspace spanned by the eigenfunction ψ_j. Here, $\mathcal{R}(t)$ is the residual operator associated with the essential spectrum of $\mathcal{L}_0^{\mathbf{y}}$ and its norm is controlled by a decaying exponential. We assume furthermore that the spectrum of $\mathcal{L}_0^{\mathbf{y}}$ lies in the complex left half-plane and that, in particular, $\mathfrak{Re}\lambda_j \leq 0$ for any j.

Such a spectral decomposition and its properties can be rigorously justified for a broad class of differential equations perturbed by small noise disturbances; see Ref. 19, Theorem 1 and Appendix A.5. Based on these rigorous results, we assume, roughly speaking, that these properties survive in a certain small-noise limit and concentrate here on vector fields $\mathbf{y}$ given by $\mathbf{G}$ in (2.24) for which a decomposition such as (2.25) holds and a spectral gap does exist in the appropriate functional space.

In the following lines, we examine the expression of the memory kernel $\mathcal{K}$ appearing in Eq. (2.23) using the eigendecomposition proposed in Eq. (2.25). In particular, we study such an integral kernel $\mathcal{K}$ component-wise,

$$[\mathcal{K}(t,s,\mathbf{x})]_i = \mathbf{C}^{\mathbf{y}}(\mathbf{x}(s)) \cdot \left\langle \nabla \sum_{j=1}^{N} e^{\lambda_j(t-s)}\alpha_{i,j}\psi_j(\mathbf{y}) \right\rangle + \mathcal{R}(t-s)[\mathbf{C}^{\mathbf{x}}]_i \tag{2.26}$$

$$\approx \mathbf{C}^{\mathbf{y}}(\mathbf{x}(s)) \cdot \left\langle \nabla \sum_{j=1}^{N} e^{\lambda_j(t-s)}\alpha_{i,j}\psi_j(\mathbf{y}) \right\rangle \tag{2.27}$$

$$= \mathbf{C}^{\mathbf{y}}(\mathbf{x}(s)) \cdot \sum_{j=1}^{N} e^{\lambda_j(t-s)}\alpha_{i,j}\left\langle \nabla\psi_j(\mathbf{y}) \right\rangle, \tag{2.28}$$

for $i = 1, \ldots, d_1$, where

$$\alpha_{i,j} = \left\langle \psi_j^*, [\mathbf{C}^{\mathbf{x}}]_i \right\rangle = \int \nu(\mathrm{d}\mathbf{y})\overline{\psi_j^*(\mathbf{y})}[\mathbf{C}^{\mathbf{x}}(\mathbf{y})]_i,$$

and we have neglected the contribution coming from the essential spectrum. The $(\cdot)^*$ superscript is used to denote the dual eigenfunction.

This decomposition highlights the fact that the leading eigenvalues of the operator governing the evolution of observables in the uncoupled $\mathbf{y}$-dynamics set the timescale for the memory kernel. Furthermore, this spectral approximation implies that the correlation functions of the noise have the same decay properties, as will become apparent later in the proof of Theorem 2.1. It follows that the correspondence between the noise and integral timescales allows us to recast the SIDE in the WL equation (2.23) into a fully Markovian version with linearly driven hidden variables that are forced by the observed variables through a functional dependence that can be nonlinear. More exactly, we have the following theorem.

Theorem 2.1: *Consider the system (2.20) where Eq. (2.20a) is, instead, a scalar equation for a real-valued variable $x(t)$. Let ν be the physical invariant measure associated with the equation*

$$\dot{\mathbf{y}} = \mathbf{G}(\mathbf{y}), \tag{2.29}$$

i.e., with the flow determined by the vector field $\mathbf{G}$ in system (2.20), for $\epsilon = 0$. Moreover, let $\mathcal{L}_0^{\mathbf{y}}$ be the (uncoupled) Koopman operator associated with (2.29) as defined in Eq. (2.24).

The point spectrum of $\mathcal{L}_0^{\mathbf{y}}$ is assumed to be constituted of N simple eigenvalues, whose corresponding eigenpairs $\{(\lambda_j, \psi_j),\ j = 1, \ldots, N\}$ are ordered as follows: $0 \geq \mathfrak{Re}\lambda_j \geq \mathfrak{Re}\lambda_{j+1}$ and $\lambda_j = \overline{\lambda_{j+1}}$ when $\mathfrak{Im}\lambda_j > 0$, for j in $\{1, \ldots, N\}$.

We assume that $\mathbf{C}^{\mathbf{x}}$ in (2.20) lies in the span $\{\psi_j,\ j = 1, \ldots, N\}$ and has ν-mean zero.

Then, the WL equation (2.23) associated with system (2.20) admits a Markovianization of the form

$$\dot{x}(t) = \mathbf{F}(x(t)) + \epsilon\Lambda \cdot \mathbf{Z}(t), \tag{2.30a}$$

$$d\mathbf{Z}(t) = (\epsilon\mathbf{R}(x(t)) + \mathrm{D}\mathbf{Z}(t))\, dt + \Sigma dW_t, \tag{2.30b}$$

where Λ and $\mathbf{Z}(t)$ lie in $\mathbb{C}^N$ for every t, while the inner product $\Lambda \cdot \mathbf{Z}(t)$ is real. Here, $\mathbf{R}$ is mapping $\mathbb{R}$ into $\mathbb{C}^N$, W_t is a (real-valued) N-dimensional Wiener process with covariance matrix Σ, and D is an $N \times N$ matrix with complex entries, as specified below.

More precisely,

$$\Lambda = \left[\alpha_1^{1/2}\beta_1^{1/2}, \ldots, \alpha_N^{1/2}\beta_N^{1/2}\right]^{\top}, \tag{2.31}$$

where

$$\alpha_j = \left\langle \psi_j^*, \mathbf{C}^{\mathbf{x}} \right\rangle = \int \nu(d\mathbf{y})\overline{\psi_j^*(\mathbf{y})}\mathbf{C}^{\mathbf{x}}(\mathbf{y}), \tag{2.32a}$$

$$\beta_j = \left\langle \mathbf{C}^{\mathbf{x}}, \psi_j \right\rangle = \int \nu(d\mathbf{y})\mathbf{C}^{\mathbf{x}}(\mathbf{y})\psi_j(\mathbf{y}). \tag{2.32b}$$

The $\mathbb{C}^N$-valued mapping $\mathbf{R}$ is defined as

$$\mathbf{R}(x) = \left(\mathbf{C}^{\mathbf{y}}(x) \cdot \frac{\alpha_1^{1/2}}{\beta_1^{1/2}}\left\langle \nabla\psi_1(\mathbf{y}) \right\rangle, \ldots, \mathbf{C}^{\mathbf{y}}(x) \cdot \frac{\alpha_N^{1/2}}{\beta_N^{1/2}}\left\langle \nabla\psi_N(\mathbf{y}) \right\rangle\right), \tag{2.33}$$

where $\mathbf{C}^{\mathbf{y}}$ is defined in (2.20b), and $\langle\cdot\rangle$ denotes averaging with respect to the invariant measure ν.

Finally, $\mathrm{D} = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_N)$ and the covariance matrix Σ is given by

$$\Sigma = (-(\mathrm{D} + \mathrm{D}^*))^{1/2}\,\mathrm{H}, \tag{2.34}$$

where H is an $N \times N$ matrix whose entries are defined as follows: If λ_j is real, then $\mathrm{H}_{j,j} = 1$, and if $\lambda_j = \overline{\lambda_{j+1}}$, then

$$\begin{aligned} \mathrm{H}_{j,j} &= 1, \\ \mathrm{H}_{j+1,j+1} &= 0, \\ \mathrm{H}_{j+1,j} &= 1, \end{aligned} \tag{2.35}$$

while all other entries are zero.

The full proof appears in Appendix A.

Remark 2.1: Note that the Koopman operator of interest here is the one associated with the **y**-subspace $\mathcal{Y} \subseteq \mathbb{R}^{d_2}$ and not with the entire $(\mathbf{x}, \mathbf{y})$-space $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{d_1+d_2}$. Other techniques, like the DMD mentioned in Sec. I A, aim at extracting the modes of variability of the full system by means of studying the Koopman operator in the entire phase space through suitably defined observables. To this end, the latter methods employ projections of observables onto the eigenfunctions of the Koopman operator to obtain the so-called Koopman modes, which are susceptible of capturing the underlying dynamics. Notice that in Theorem 2.1, instead, we are using the Koopman eigenfunctions to identify the closure model, while projections only come into play in the definition of the coefficients α_j and β_j; see Eqs. (2.32a) and (2.32b), respectively.

Remark 2.2:

(i) Assumptions on **F**, **R**, and Λ that ensure that (2.30) possesses a global random attractor—and thus a stable asymptotic behavior in the pullback sense—appear in Ref. 43, Theorem 3.1 and Corollary 3.2.

(ii) Note that Theorem 2.1 can be viewed as a generalization of other Markovianization results for GLEs that appeared in the literature; see Ref. 63. For instance, the scalar GLE in $\mathbb{R}$ reduces to

$$\dot{x} = F(x(t)) - \int_0^t K(t-s)x(s)\mathrm{d}s + \eta(t), \tag{2.36}$$

where λ is in $\mathbb{R}^n$, M is a positive definite $n \times n$ matrix, and $K(t-s) = \left(e^{M(t-s)}\lambda\right) \cdot \lambda$ determines the autocorrelation of the process $\eta(t)$. In this setting, Eq. (2.36) is equivalent to the following SDE:

$$\begin{aligned} \dot{x} &= F(x(t)) + \lambda \cdot z, \\ \mathrm{d}z &= (x\lambda - Mz)\mathrm{d}t + \Sigma\mathrm{d}W_t, \end{aligned} \tag{2.37}$$

with $\Sigma\Sigma^* = M + M^*$. Theorem 2.1 allows for nonlinear dependence on x in the z-equation, and thus for memory kernels that are more complicated than in (2.36). Such a generalization is of practical importance since the process z can then have a more complex correlation dependence on the observed variable x than the one afforded by linear memory terms.

Remark 2.3: When $\mathcal{L}_0^{\mathbf{y}}$ is self-adjoint—in a suitable Hilbert space, as outlined in Appendix B, i.e., when $\mathcal{L}_0^{\mathbf{y}} = \mathcal{L}_0^{\mathbf{y}^*}$—the eigenvalues are real and the eigenvectors are mutually orthogonal. Self-adjointness thus implies that there are no oscillations in the correlation functions of the noise or, equivalently, peaks in their power spectrum. With respect to Theorem 2.1, the matrix H in this case would be the identity, since the eigenvalues and eigenfunctions are real and the Itô solutions of (2.30b) are, hence, real as well.

Remark 2.4: The resulting system given by Eq. (2.30) is now fully Markovian and the only sources of error with respect to the original SIDE (2.18) lie (i) in the effects of the essential spectrum, which are neglected herein, and (ii) the assumptions about the coupling terms. Neglecting the essential spectrum is only valid for Koopman operators with a point spectrum capable of capturing the correlations in the decoupled **y**-system; the latter might only hold in the case of Markovian diffusion processes and not for deterministic ones. Also, the assumption that the coupling functions project solely on the point spectrum might not hold in general.

From a practical perspective, though, a suitable choice of dominant eigendirections can reduce the number of extra dimensions needed to integrate the system. Such a suitable choice boils down to neglecting particular eigendirections and this can be done according to two handy criteria,

(i) The weight determined by the α_j and β_j coefficients defined in Eqs. (2.32a) and (2.32b) is small and

(ii) The eigenvalues λ_j of $\mathcal{L}_0^{\mathbf{y}}$ satisfy $\mathfrak{Re}\lambda_{j_0} \ll \lambda^\dagger$, for some $j_0 \in \{1,\dots,N\}$, in which case $e^{\lambda_{j_0} t}$ decays rapidly as t grows; here $\lambda^\dagger < 0$ and $|\lambda^\dagger|$ is some characteristic inverse time for the deterministic system $\mathbf{F}$. In addition, if $\alpha_{j_0} > \alpha_j$ for $j = 1,\dots,N$ and $j \neq j_0$, both the memory and the noise correlations die out fast. Hence, one can neglect the integral terms and perform a fully Markovian parameterization, which is possible in the presence of white noise.

Remark 2.5: Theorem 2.1 is stated for x scalar for the sake of simplicity, but this result extends to the d_1-dimensional Eq. (2.20) for the (observed) variables $\mathbf{x}$. In this remark, we sketch the main elements that permit such a generalization.

Aside from the obvious generalization of the assumptions in Theorem 2.1 to a multidimensional setting, the main hypothesis consists of assuming that the now vector-valued coupling function $\mathbf{C}^{\mathbf{x}}$ in Eq. (2.20) has components $\{[\mathbf{C}^{\mathbf{x}}]_i : i = 1,\dots,d_1\}$ that project onto the N simple eigenspaces of the decoupled Koopman operator introduced in Eq. (2.24). In this case, the construction of a multilevel Markovianization like Eq. (2.30) can be done in the following fashion:

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) + \epsilon \Lambda \mathbf{Z}(t), \tag{2.38a}$$

$$\mathrm{d}\mathbf{Z}(t) = (\epsilon \mathbf{R}(\mathbf{x}(t)) + \mathcal{D}\mathbf{Z}(t))\,\mathrm{d}t + \mathcal{S}\mathrm{d}W_t. \tag{2.38b}$$

Here W_t is a d_1N-dimensional Wiener process, $\mathbf{Z}(t)$ is a d_1N-dimensional vector, Λ is a matrix of size $d_1 \times d_1 N$, $\mathbf{R} : \mathbb{R}^{d_1} \longrightarrow \mathbb{C}^{d_1 N}$, and $\mathcal{D}$ and $\mathcal{S}$ are $d_1 N \times d_1 N$ block-diagonal matrices given by

$$\mathcal{D} = \begin{bmatrix} \mathrm{D}_1 & & \\ & \ddots & \\ & & \mathrm{D}_N \end{bmatrix} \text{ and } \mathcal{S} = \begin{bmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_N \end{bmatrix}, \tag{2.39}$$

where D_j and Σ_j are $d_1 \times d_1$ diagonal matrices with every (nonzero) element being equal to λ_j or $\sqrt{-2\mathfrak{Re}\lambda_j}$, respectively. More importantly, the vectors $\mathbf{Z}(t)$ and $\mathbf{R}(\mathbf{x})$ are split into N column vectors $\mathbf{z}_j(t)$ and $\mathbf{r}_j(\mathbf{x})$ of length d_1 with j in $\{1,\dots,N\}$. This way, $\mathbf{Z}(t) = \left[\mathbf{z}_1^\top(t),\dots,\mathbf{z}_N^\top(t)\right]^\top$ and $\mathbf{R}(\mathbf{x}) = \left[\mathbf{r}_1^\top(t),\dots,\mathbf{r}_N^\top(t)\right]^\top$.

Therefore, Eq. (2.38) can be written as

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) + \epsilon \Lambda \mathbf{Z}(t), \tag{2.40a}$$

$$\mathrm{d}\mathbf{z}_1(t) = (\epsilon \mathbf{r}_1(\mathbf{x}(t)) + \mathrm{D}_1\mathbf{z}_1(t))\,\mathrm{d}t + \Sigma_1 \mathrm{d}W_t^{(1)}, \tag{2.40b}$$

$$\vdots$$

$$\mathrm{d}\mathbf{z}_N(t) = (\epsilon \mathbf{r}_N(\mathbf{x}(t)) + \mathrm{D}_N\mathbf{z}_N(t))\,\mathrm{d}t + \Sigma_N \mathrm{d}W_t^{(N)}, \tag{2.40c}$$

where $W_t^{(j)}$ is a d_1-dimensional Wiener process. The vectors $\mathbf{r}_j(\mathbf{x})$ are given by

$$\mathbf{r}_j(\mathbf{x}) = \left[\mathbf{C}^{\mathbf{y}}(\mathbf{x}) \cdot \gamma_{1,j}\langle \nabla\psi_1(\mathbf{y})\rangle, \dots, \mathbf{C}^{\mathbf{y}}(\mathbf{x}) \cdot \gamma_{d_1,j}\langle \nabla\psi_1(\mathbf{y})\rangle\right]^\top, \tag{2.41}$$

where $\gamma_{i,j}$ are defined in terms of the parameters (α_j, β_j) introduced in Eqs. (2.32a) and (2.32b), respectively. Here, we do not give the explicit expression of Λ, but its role is to provide suitable weights, in the spirit of Eq. (2.31), to the levels in Eq. (2.40) so that (a) the correlation functions match those of the coupling function $\mathbf{C}^{\mathbf{x}}$ in the uncoupled regime and (b) the resulting term $\Lambda\mathbf{Z}(t)$ is real.

The system equation (2.40) has the general structure one would obtain if the coupling function $\mathbf{C}^{\mathbf{x}}$ projected along all the eigendirections in the point spectrum. This might not be true in general, but the drift matrix $\mathcal{D}$ can be rearranged so that only the relevant modes of variability are modeled—following criteria (i) and (ii), as formulated in Remark 2.4—and still afford a reduction of the number of levels N. Notice that the Nth level variables $\mathbf{z}_N$ described by Eq. (2.40c) decorrelate the fastest, the rest, since their exponential decorrelation rate is given by $|\mathfrak{Re}\lambda_N| \geq |\mathfrak{Re}\lambda_j|$, for all $j = 1,\dots,N-1$.

The advantages of the Markovian system of Eqs. (2.30) and (2.40) over the original WL equation (2.23) are twofold. First, we identify situations in which the WL equation can be Markovianized by introducing extended, hidden variables. This idea was already introduced in a preliminary application of the WL parameterization,[87] in which the authors resorted to a Markovian system to perform their simulations. In fact, one of their examples is studied in the present framework; see Sec. II D.

Second, memory equations contain nonlocal terms that are cumbersome and computationally expensive to integrate, as well as requiring much larger storage for the full history of the system's variables. The efficient Markovianization of evolution equations with memory terms is an active field of research in diverse areas of mathematics and the applied sciences; these areas include the study of bifurcations of delay differential equations,[8,11] the reduction of stochastic partial differential equations to stochastic invariant manifolds,[14,15] and material sciences,[22] among many others.

D. Preliminary example

As seen earlier in Theorem 2.1, if the coupling function is resonant with the Koopman operator associated with the $\mathbf{y}$-dynamics, one can identify the dominant exponential rates of decay of the memory term and the characteristic decorrelation time of the noise. As a consequence, one can Markovianize the parameterization and greatly facilitate the numerical integrations involved.

To illustrate the above statement, we revisit here the preliminary application of the WL parameterization in the context of multiscale triads.[87] In that work, the authors implemented the parameterization for a collection of three-dimensional models that do exhibit time scale separation and compare the corresponding outputs to those obtained via homogenization. The results are encouraging, since the parameterizations in Ref. 87 were obtained only from the decoupled hidden dynamics, in the lines of the present paper as well; see derivation of Eq. (2.18).

One of the first multiscale triads studied in Ref. 87 is the following:

$$\dot{x}(t) = \epsilon B^{(0)} y_1 y_2, \tag{2.42a}$$

$$\dot{y_1}(t) = \epsilon B^{(1)} x y_2 - \gamma_1 y_1 + \sigma_1 \mathrm{d}W_t^{(1)}, \tag{2.42b}$$

$$\dot{y_2}(t) = \epsilon B^{(2)} x y_1 - \gamma_2 y_2 + \sigma_2 \mathrm{d}W_t^{(2)}. \tag{2.42c}$$

Here, we require that $\sum_j B^{(j)} = 0$, $\mathrm{d}W_t^{(1)}$ and $\mathrm{d}W_t^{(2)}$ are scalar Brownian increments, and the parameter ϵ indicates both the timescale separation and the coupling strength. Notice that when the system is decoupled, i.e., when $\epsilon = 0$, the fast dynamics evolve according to an Ornstein-Uhlenbeck (OU) process whose steady-state statistics are governed by Gaussian distributions with explicit mean and variance (see, e.g., Ref. 63). Hence, by virtue of the previous formulas or by following Ref. 87, the WL parameterization yields the following scalar SIDE:

$$\dot{x}(t) = \epsilon\eta(t) + \epsilon^2 \int_0^t \mathcal{K}(s, x(t-s))\mathrm{d}s, \tag{2.43}$$

here,

$$\langle \eta(t) \rangle = 0, \tag{2.44a}$$

$$\langle \eta(t+s)\eta(s) \rangle = \left(B^{(0)}\right)^2 e^{-(\gamma_1+\gamma_2)t} \frac{\sigma_1^2}{2\gamma_1}\frac{\sigma_2^2}{2\gamma_2}, \tag{2.44b}$$

$$\mathcal{K}(s,x) = \begin{bmatrix} x \\ x \end{bmatrix} \cdot \left\langle \begin{bmatrix} B^{(1)} y_2 \\ B^{(2)} y_1 \end{bmatrix} : \nabla_{\mathbf{y}} y_1(s) y_2(s) \right\rangle, \tag{2.44c}$$

where the angular brackets refer to the ensemble averages according to the already mentioned Gaussian distributions arising from the decoupled model. Expanding these averages, Eq. (2.44c) leads to

$$\mathcal{K}(s,x) = x e^{-(\gamma_1+\gamma_2)} \left\langle B^{(1)} y_2^2 + B^{(2)} y_1^2 \right\rangle \tag{2.45a}$$

$$= x B^{(0)} e^{-(\gamma_1+\gamma_2)s} \left(B^{(1)} \frac{\sigma_2^2}{2\gamma_2} + B^{(2)} \frac{\sigma_1^2}{2\gamma_1} \right). \tag{2.45b}$$

The timescales are indicated by the exponents in the formulas above and they are the same for the noise and the memory kernel. This equality suggests the possibility of Markovianizing the memory equation into the following two-dimensional system:

$$\dot{z}_1(t) = \epsilon B^{(0)} z_2, \tag{2.46a}$$

$$\dot{z}_2(t) = -(\gamma_1+\gamma_2) z_2 + \frac{\sigma_1\sigma_2}{2\gamma_1\gamma_2} \{2(\gamma_1+\gamma_2)\}^{1/2} \mathrm{d}W_t + \epsilon \left(B^{(1)} \frac{\sigma_2^2}{2\gamma_2} + B^{(2)} \frac{\sigma_1^2}{2\gamma_1} \right) z_1. \tag{2.46b}$$

Clearly, performing a numerical integration of this system is easier than for a memory equation like Eq. (2.43).

The results of Sec. II C allow us to carry out the dimension reduction of the multiscale triad by analyzing the spectral properties of the Koopman operator associated with the decoupled $\mathbf{y}$-dynamics. Since the $\mathbf{y}$ variables evolve stochastically, the Koopman operator becomes the *backward-Kolmogorov* equation, which governs the evolution of the expectation values of the observables. Thus, for a generic observable Ψ in the $\mathbf{y}$ phase space, the evolution of its expectation value is given by

$$\begin{aligned} \partial_t \Psi(y_1, y_2) &= \mathcal{L}_0^{\mathbf{y}} \Psi(y_1, y_2) \\ &= \begin{bmatrix} -\gamma_1 y_1 \\ -\gamma_2 y_2 \end{bmatrix} \cdot \nabla \Psi(y_1, y_2) \\ &\quad + \sigma_1^2 \partial_{y_1}^2 \Psi(y_1, y_2) + \sigma_2^2 \partial_{y_2}^2 \Psi(y_1, y_2). \end{aligned} \tag{2.47}$$

Now, let $\Psi(y_1, y_2) = y_1 y_2$ be the coupling function of the triad system (2.42), for which we find that

$$\mathcal{L}_0^{\mathbf{y}} \Psi(y_1, y_2) = -(\gamma_1+\gamma_2)\Psi(y_1, y_2). \tag{2.48}$$

The above equation is an eigenvalue problem, showing that this particular Ψ is an eigenfunction of the Koopman operator associated with the eigenvalue $e^{-(\gamma_1+\gamma_2)}$. This is no surprise since y_1 and y_2 are, respectively, the Hermite polynomial eigenfunctions of the backward-Kolmogorov equation of the scalar OU process.[79] Hence, the product $y_1 y_2$ is also an eigenfunction of the same equation for the joint process. Therefore, we can immediately re-Markovianize the parameterization according to Eqs. (2.30), where $\mathrm{D} = \gamma_1 + \gamma_2$ and

$$\mathbf{R}(x(t)) = \left(B^{(1)} \frac{\sigma_2^2}{2\gamma_2} + B^{(2)} \frac{\sigma_1^2}{2\gamma_1} \right) x(t). \tag{2.49}$$

III. MULTILEVEL STOCHASTIC MODELS AND EMPIRICAL MODEL REDUCTION (EMR)

A. Multilevel stochastic models (MSMs)

MSMs are a general class of SDEs that were introduced in Ref. 43 and are, by their layered structure, susceptible to provide a good approximation of the GLE (1.3b) formulated by Mori and Zwanzig when a high-dimensional system is partially observed; see Ref. 43, Proposition 3.3 and Sec. 5. The MSM framework allows one to provide such approximations that are accompanied by useful dynamical properties, such as the existence of random attractors (Theorem 3.1 in Ref. 43). The conditions on the high-dimensional system's coupling interactions between the resolved and hidden variables are also well understood (Corollary 3.2 in Ref. 43). (Moreover, random attractors with fractal structures that survive highly degenerate noise[17]—a situation that might occur when approximating deterministic chaotic dynamics by stochastic pathwise dynamics—can still be present in MSMs; see Ref. 43, Sec. 7.)

As discussed in Ref. 43, MSMs arise in a variety of data-driven protocols for model reduction that typically use successive regressions from partial observations; see Sec. III B. The general form of an MSM is given by Ref. 43, Eq. (**MSM**); we only use herein its most basic version, which has the following structure:

$$\mathrm{d}\mathbf{x}(t) = \Big(\mathbf{F}(\mathbf{x}(t)) + \epsilon \Pi \mathbf{y}(t) \Big) \mathrm{d}t, \tag{3.1a}$$

$$\mathrm{d}\mathbf{y}(t) = \Big(\epsilon \mathrm{C}\mathbf{x}(t) - \mathrm{D}\mathbf{y}(t) \Big) \mathrm{d}t + \Sigma \mathrm{d}\mathbf{W}_t. \tag{3.1b}$$

Here, the observed vector variable $\mathbf{x}(t)$ lies in $\mathbb{R}^{d_1}$ and, for $\epsilon = 0$, the hidden variables $\mathbf{y}(t) \in \mathbb{R}^{d_2}$ evolve in time independently. Otherwise, the dynamics of the $\mathbf{x}$ variables is linearly coupled to that of the $\mathbf{y}$ variables, which act upon (3.1a) as a stochastic forcing, via the

canonical projection $\Pi : \mathbb{R}^{d_2} \longrightarrow \mathbb{R}^{d_1}$, while $\mathbf{W}_t$ in (3.1b) is a d_1-dimensional Wiener process. Clearly, Eq. (3.1) is closely related to Eq. (2.38) discussed above.

The matrix C in $\mathbb{R}^{d_2 \times d_1}$ models the feedback of the $\mathbf{x}$-process onto the $\mathbf{y}$ variables. In the case of $C \equiv 0$, $\mathbf{y}$ would evolve according to an OU process with drift matrix D and covariance matrix $\Sigma\Sigma^*$. For the sake of simplicity, we restrict ourselves to the case $d_1 = d_2$ so that the projection Π reduces to the identity.

The more general MSM with nonlinear coupling considered in Ref. 43 was shown to be equivalent to a SIDE with explicit expressions for the memory kernels and stochastic forcing being obtained; see Ref. 43, Proposition 3.3. The noise term there results from successive convolutions of the homogeneous solutions of the lower levels of the system with an OU process. In particular, using the Itô stochastic calculus, one readily obtains a SIDE that is equivalent to an MSM; see Ref. 43, Sec. 3.2 and Appendix C.

We show next that the same SIDE can actually be obtained by using the operator formalism presented in Sec. II. One might object that an MSM is a stochastic system, due to the presence of white noise in the hidden layer, whereas the theory presented above applies to deterministic dynamics. However, as clarified below, the operator formalism applies equally well to the MSM case.

In fact, given a smooth, $\mathcal{C}^\infty$ observable

$$\Phi : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}, \quad (\mathbf{x}, \mathbf{y}) \mapsto \Phi(\mathbf{x}, \mathbf{y}), \tag{3.2}$$

its expected value along a stochastic trajectory $X_t = (\mathbf{x}(t), \mathbf{y}(t))^\top$ solving Eq. (3.1), namely, $\mathbb{E}(\Phi(X_t))$, defines a Markov semigroup P_t by

$$P_t\Phi = \mathbb{E}(\Phi(X_t)), \tag{3.3}$$

which solves the backward Kolmogorov equation[19] associated with Eq. (3.1),

$$\partial_t (P_t\Phi) = \begin{bmatrix} \mathbf{F}(\mathbf{x}) + \epsilon\mathbf{y} \\ \epsilon C\mathbf{x} - D\mathbf{y} \end{bmatrix} \cdot \nabla P_t\Phi + \frac{1}{2} \begin{bmatrix} 0 \\ \Sigma\Sigma^* \nabla_{\mathbf{y}}^2 P_t\Phi \end{bmatrix}, \tag{3.4}$$

the only difference with respect to the transport equation (2.2) lies in the presence of a second-order differential operator induced by the white noise.

We introduce the operators

$$\mathcal{L}_0 = \begin{bmatrix} \mathbf{F}(\mathbf{x}) \\ -D\mathbf{y} \end{bmatrix} \cdot \nabla + \begin{bmatrix} 0 \\ \Sigma\Sigma^* \nabla_{\mathbf{y}}^2 \end{bmatrix}, \tag{3.5a}$$

$$\mathcal{L}_1 = \begin{bmatrix} \mathbf{y} \\ C\mathbf{x} \end{bmatrix} \cdot \nabla, \tag{3.5b}$$

which play a role that is analogous to their deterministic relatives in Eq. (2.3) of Sec. II. Again, the operator $\mathcal{L}_1$ is viewed as a perturbation to the operator $\mathcal{L}_0$ due to the coupling.

If one considers observables $\Phi = \Phi(\mathbf{x})$, Eq. (3.4) becomes at time $t = 0$,

$$\partial_t (P_t\Phi)\,|_{t=0} = \left[\mathbf{F}(\mathbf{x}) + \epsilon\mathbf{y}\right] \cdot \nabla_{\mathbf{x}}\Phi, \tag{3.6}$$

and we apply now, as in Sec. II, the Dyson perturbative expansion. By virtue of the formula (2.18), the parameterization leads to a reduced equation of the form

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) + \epsilon\eta(t) + \epsilon^2 \int_0^t \mathcal{K}(s, \mathbf{x}(t-s))\mathrm{d}s, \tag{3.7}$$

where the hidden variables in the decoupled regime are governed by an OU process with invariant measure $\mu_{\mathbf{y}}$. The properties of the stochastic noise $\eta(t)$ are given by

$$\langle \eta(t)\eta^\top(0) \rangle = \int \mathrm{d}\mu_{\mathbf{y}}(\mathbf{y}_0) e^{t\mathcal{L}_0} \mathbf{y}_0\mathbf{y}_0^\top \tag{3.8a}$$

$$= \int \mathrm{d}\mu_{\mathbf{y}}(\mathbf{y}_0) \mathbb{E}\left(\mathbf{y}(t)|\mathbf{y}_0\right) \left(\mathbb{E}\left(\mathbf{y}(0)|\mathbf{y}_0\right)\right)^\top \tag{3.8b}$$

$$= \int \mathrm{d}\mu_{\mathbf{y}}(\mathbf{y}_0) e^{-t\mathrm{D}} \mathbf{y}_0\mathbf{y}_0^\top \tag{3.8c}$$

$$= \int \mathrm{d}\mu_{\mathbf{y}}(\mathbf{y}_0) e^{-t\mathrm{D}} \mathbf{y}_0\mathbf{y}_0^\top = e^{-t\mathrm{D}}\Sigma\Sigma^*, \tag{3.8d}$$

where $\mathbf{y}$ is a function analogous to the coupling function $\mathbf{C}_{\mathbf{y}}^{\mathbf{x}}$ in Sec. II and the initial condition $\mathbf{y}_0$ is assumed to be normally distributed with zero mean and variance $\Sigma\Sigma^*$. The memory kernel is given by

$$\mathcal{K}(s, \mathbf{x}(t-s)) = \int \mathrm{d}\mu_{\mathbf{y}}(\mathbf{y}_0) C\mathbf{x}(t-s) \cdot \nabla_{\mathbf{y}_0} \mathbb{E}\left(\mathbf{y}(s)|\mathbf{y}_0\right) \tag{3.9a}$$

$$= \int \mathrm{d}\mu_{\mathbf{y}}(\mathbf{y}_0) C\mathbf{x}(t-s) \cdot \nabla_{\mathbf{y}_0} e^{-s\mathrm{D}} \mathbf{y}_0 \tag{3.9b}$$

$$= C\mathbf{x}(t-s) \cdot e^{-s\mathrm{D}} \tag{3.9c}$$

$$= e^{-s\mathrm{D}} C\mathbf{x}(t-s). \tag{3.9d}$$

Using the intermediate steps above, the explicit parameterization finally becomes

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) + \epsilon\eta(t) + \epsilon^2 \int_0^t e^{-s\mathrm{D}} C\mathbf{x}(t-s)\mathrm{d}s. \tag{3.10}$$

The integrodifferential equation above is the same as Eq. (C2) one obtains using the Itô integration described in Appendix C. This similarity of results occurs because we are considering the case of additive coupling, and the Dyson expansion can be truncated after the memory term proportional to ϵ^2, cf. Sec. II B.

B. Empirical model reduction (EMR)

As discussed in Secs. I and II, and illustrated in Fig. 1, the evolution of the resolved variables is forced by fluctuating terms and the effects of the previous state of the system. It is desirable, therefore, to construct a full model of the system even when only capable to partially observe it. The EMR methodology[43,45,48,49] aims at achieving this goal; we discuss it below in the broader context of MSMs. Note that EMR provides a solution for the dynamical closure of partially observed systems and thus it differs from the methodology recently proposed in Ref. 6, which requires one to fully observe the system for the data-driven discovery of its underlying equations to work.

Having a set of reduced d_1-dimensional observations $\{\mathbf{x}_i : i = 1, \ldots, n\}$ every $\mathrm{d}t$ time units, one seeks to regress the tendencies $\{\mathrm{d}\mathbf{x}_i : i = 1, \ldots, n\}$ of the data onto a quadratic function of the

form

$$\mathbf{F}(\mathbf{x}) = \mathbf{f} + \mathbf{b}\cdot\mathbf{x} + \mathbf{Q}(\mathbf{x}), \tag{3.11}$$

where $\mathbf{b}$ in $\mathbb{R}^{d_1\times d_1}$ describes dissipative processes and $\mathbf{Q}$ is a quadratic form describing self-interaction between the $\mathbf{x}$ variables. The ith component of the quadratic form is given by

$$[\mathbf{Q}]_i = \mathbf{x}^\top \mathbf{A}_i \mathbf{x}, \tag{3.12}$$

where $\mathbf{A}_i$ in $\mathbb{R}^{d_1\times d_1}$. The function $\mathbf{F}$ is expected to approximate the vector field driving the dynamics in the absence of hidden external influences. Of course, performing regressions yields an error called *residual* $\{\mathbf{y}_i : i = 1,\dots,n\}$. Hence, the evolution of the $\mathbf{x}$ variables satisfies the equation

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}(\mathbf{x}) + \mathbf{y}. \tag{3.13}$$

At this point, one can study the properties of the residual time series $\{\mathbf{y}_i\}_{i=1}^n$ and construct a model able to reproduce its main statistical features. However, we know that if it is possible to sample all the variables of the dynamical system of interest, one expects that the residuals are explained by the errors committed exclusively in the regression algorithm. If some sort of subsampling is done, whether spatial or temporal, the residuals are also due to the delayed influence of unresolved processes that are involved in the coupling.

Allowing for the main level variables $\mathbf{x}$ to be linearly coupled with the residual $\mathbf{y}$, we are creating a model that is able to incorporate memory effects as well. Hence, for each component i of $\mathbf{x}$, we have

$$\frac{\mathrm{d}[\mathbf{x}]_i}{\mathrm{d}t} = [\mathbf{f}]_i + \mathbf{b}_i^{(0)}\cdot\mathbf{x} + \mathbf{x}^\top\mathbf{A}_i\mathbf{x} + [\mathbf{y}^{(0)}]_i, \tag{3.14a}$$

$$\frac{\mathrm{d}[\mathbf{y}^{(0)}]_i}{\mathrm{d}t} = \mathbf{b}_i^{(1)}\cdot[\mathbf{x},\mathbf{r}^{(0)}] + [\mathbf{y}^{(1)}]_i, \tag{3.14b}$$

$$\frac{\mathrm{d}[\mathbf{y}^{(1)}]_i}{\mathrm{d}t} = \mathbf{b}_i^{(2)}\cdot[\mathbf{x},\mathbf{r}^{(0)},\mathbf{r}^{(1)}] + [\mathbf{y}^{(2)}]_i, \tag{3.14c}$$

$$\vdots \tag{3.14d}$$

$$\frac{\mathrm{d}[\mathbf{y}^{(l)}]_i}{\mathrm{d}t} = \mathbf{b}_i^{(l+1)}\cdot[\mathbf{x},\mathbf{r}^{(0)},\mathbf{r}^{(1)},\dots,\mathbf{r}^{(l)}] + [\mathbf{y}^{(l+1)}]_i, \tag{3.14e}$$

where we have introduced new matrices $\mathbf{b}^{(j)} \in \mathbb{R}^{d_1\times(j+2)d_1}$ that model the linear coupling. The residual at the last level $[\mathbf{y}^{(l+1)}]_i$ is assumed to obey the Wiener process for which the correlation matrix is obtained from the last residual time series. The choice of stochastic process in the last step can only be done if the decorrelation of $[\mathbf{y}^{(l+1)}]_i$ is sufficiently fast according to the timescale set by $\mathrm{d}t$. This motivates the problem of choosing the optimal number l of levels.

Several criteria have been established to determine the optimal number of levels l. The basic idea is that the resulting $(l+1)$-residual in Eq. (3.14e) should be well approximated by Gaussian white noise.[43,49] One has, therefore, to test whether the residual variables decorrelate at lag $\mathrm{d}t$ and whether the lag-0 covariance matrix is invariant in the last levels.

Therefore, regression on the tendency of the optimal level $\mathbf{y}^{(l)}$ should yield

$$\mathbf{y}^{(l+1)} - \mathbf{y}^{(l)} \simeq -\mathbf{y}^{(l)} + \gamma^{(l)}, \tag{3.15}$$

where $\gamma^{(l)}$ is the residual of the previous regression and is approximately equal to $\mathbf{y}^{(l+1)}$. Hence, $\gamma^{(l)}$ would become a lagged version of $\mathbf{y}^{(l+1)}$. Subject to this assumption, it is possible to estimate the optimal value of the coefficient of determination R^2,

$$R^2 = 1 - \frac{\sum_k \gamma_l^2}{\sum_k \left(\mathbf{y}^{(l+1)} - \mathbf{y}^{(l)}\right)^2} \simeq 1 - \frac{\sum_l \mathbf{y}^{(l+1)^2}}{\sum_l \mathbf{y}^{(l+1)^2} + \mathbf{y}^{(l)^2}} \simeq 0.5. \tag{3.16}$$

This means that, when the amount of unexplained variance of the last regression is 50%, one has reached the optimal number of levels. It is worth stressing that the empirical model (3.14) has the structure (3.1) of an MSM, as discussed in Ref. 43. It can, therewith, be integrated to transform it into an integrodifferential equation with explicit formulas for the fluctuating noise and memory kernel, cf. Ref. 43, Proposition 3.3; see also Sec. III A for such a transformation from another perspective.

Finally, note that the aforementioned stopping criterion for EMR—namely, $R^2 \simeq 0.5$, see Ref. 43, Appendix A—is based on decorrelation times and it is also present in the multilevel WL equation (2.40). We noted, in fact, in Remark 2.5 of Sec. II C that the last level modeled by Eq. (2.40c) decorrelates the fastest with respect to the rest. Ultimately, making these points amounts to saying that a low number of levels is expected to arise in the EMR method, provided most of the eigenvalues λ_j in Theorem 2.1 are located far away from the imaginary axis, except for a very few of them. Conversely, if the Koopman eigenvalues cluster near the imaginary axis or do not exhibit a spectral gap located at a, suitably defined, small negative real part, many levels are expected to be needed to capture the hidden dynamics; see again Remark 2.5.

IV. NUMERICAL EXPERIMENTS

In Secs. II and III, we have shown that both the WL top-down approach and the EMR data-driven method yield a set of multilevel equations for the variables of interest in a multi-scale system. In particular, both approaches give explicit formulas for the fluctuation term and memory kernel in the GLE (1.3b) of the Mori–Zwanzig formalism. Furthermore, their Markovian representation share that the hidden layers are linearly driven with a white noise background [see Eqs. (2.38) and (3.1)]. We now compare the two approaches to model reduction in a simple, conceptual stochastic climate model.

Since the modeling of geophysical flows is the primary motivation for this research, we consider a set of SDEs proposed in Ref. 27 among others as a physically consistent climate "toy" model. In such a model, the main $\mathbf{x}$ variables are slow and weakly coupled to the fast $\mathbf{y}$ variables. The latter correspond to weather fluctuations and carry, in fact, most of the system's variance. The model's governing equations are

$$\begin{aligned}\mathrm{d}x_1 = \big\{&-x_2\left(L_{12} + a_1x_1 + a_2x_2\right) - d_1x_1 + F_1\\ &+ \epsilon\left(L_{13}y_1 + c_{134}y_1y_2\right)\big\}\,\mathrm{d}t,\end{aligned} \tag{4.1a}$$

$$\mathrm{d}x_2 = \left\{x_1\left(L_{21} + a_1x_1 + a_2x_2\right) - d_2x_2 + F_2 + \epsilon L_{24}y_2\right\}\mathrm{d}t, \quad (4.1b)$$

$$\mathrm{d}y_1 = \left\{\epsilon\left(-L_{13}x_1 + c_{341}y_2x_1\right) + F_3 - \frac{\gamma_1}{h}y_1\right\}\mathrm{d}t + \frac{\sigma_1}{\sqrt{h}}\mathrm{d}W_t^{(1)}, \quad (4.1c)$$

$$\mathrm{d}y_2 = \left\{-\epsilon\left(L_{24}x_2 + c_{413}y_1x_2\right) + F_4 - \frac{\gamma_2}{h}y_2\right\}\mathrm{d}t + \frac{\sigma_2}{\sqrt{h}}\mathrm{d}W_t^{(2)}, \quad (4.1d)$$

where $W_t^{(1)}$ and $W_t^{(2)}$ are two independent Wiener processes.

These equations describe the evolution of four real variables $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$; their timescale separation is determined by the parameter h and the coupling strength is controlled by ϵ. The parameter values used herein are $c_{134} = c_{341} = 0.25$, $c_{413} = -0.5$, $L_{12} = L_{21} = 1$, $L_{24} = -L_{13} = 1$, $a_1 = -a_2 = 1$, $d_1 = 0.2$, $d_2 = 0.1$, $F_1 = -0.25$, $F_2 = F_3 = F_4 = 0$, $\gamma_1 = 2$, $\gamma_2 = 1$, and $\sigma_1 = \sigma_2 = 1$. The timescale separation and the coupling strength are $h = 0.1$ and $\epsilon = 0.5$, respectively.

A. WL approximation

Notice that the hidden variables evolve according to a decoupled OU process. Taking advantage of this fact, we calculate the weak-coupling–limit parameterization of the model, according to the formulas presented in Sec. II for the separable coupling functions given by

$$\mathbf{C}^{\mathbf{x}}(\mathbf{x},\mathbf{y}) = \mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}) = \begin{bmatrix} L_{13}y_1 + c_{134}y_1y_2, \\ L_{24}y_2 \end{bmatrix}, \quad (4.2a)$$

$$\mathbf{C}^{\mathbf{y}}(\mathbf{x},\mathbf{y}) = \mathbf{C}^{\mathbf{y}}_{\mathbf{x}}(\mathbf{x}) : \mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}) = \begin{bmatrix} -L_{13}x_1 + c_{341}y_2x_1 \\ -L_{24}x_2 + c_{413}y_1x_2 \end{bmatrix}. \quad (4.2b)$$

The coupling function $\mathbf{C}^{\mathbf{x}}$ in the slow equation (4.2a) is independent of the $\mathbf{x}$ variables, indicating that the noise correction can be additively incorporated and implemented by examining the decoupled hidden process. Note that the functional form of $\mathbf{C}^{\mathbf{y}}$ implies that the WL parameterization cannot be exact in ϵ, as noted in Eqs. (2.11) and (2.12). This indicates that the WL reduced model will not only introduce an error in averaging over the decoupled steady state, but also that the Dyson expansion Eq. (2.5) has to be truncated at ϵ^3, rather than merely at ϵ^2 where no memory effects would be included.

According to the WL parameterization discussed in Sec. II, the fluctuation terms correspond to the decoupled evolution of the coupling function $\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}$, concretely as in Eq. (2.16). This allows to directly compute the correlation function,

$$\left\langle \mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y})\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y}(t))^{\top}\right\rangle = \begin{bmatrix} L_{13}^2 e^{-(\gamma_1/h)t}\frac{\sigma_1^2}{2\gamma_1} + c_{134}^2 e^{-(\gamma_1+\gamma_2)t/h}\frac{\sigma_1^2\sigma_2^2}{2\gamma_1\gamma_2} & 0 \\ 0 & L_{24}^2 e^{-(\gamma_2/h)t} \end{bmatrix}. \quad (4.3)$$

From Eq. (4.3), we deduce that the noise covariance matrix for the given parameter values is given by

$$\left\langle \mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y})\mathbf{C}^{\mathbf{x}}_{\mathbf{y}}(\mathbf{y})^{\top}\right\rangle = \begin{bmatrix} 0.2578\cdots & 0 \\ 0 & 0.5 \end{bmatrix}. \quad (4.4)$$

The memory kernel $\mathcal{K}$, which is a vector of two components $(\kappa_1, \kappa_2)^{\top}$, is given by

$$\mathcal{K}(s,\mathbf{x}) = \begin{bmatrix} \kappa_1(s,\mathbf{x}) \\ \kappa_2(s,\mathbf{x}) \end{bmatrix} = \left\langle \mathbf{C}^{\mathbf{y}}(\mathbf{x},\mathbf{y})\cdot\nabla_{\mathbf{y}}\mathbf{C}^{\mathbf{x}}(\mathbf{x}(s),\mathbf{y}(s))\right\rangle, \quad (4.5)$$

here the brackets $\langle\cdot\rangle$ indicate the averages for the uncoupled equilibrium in the $\mathbf{y}$ variables, which happen to be a set of independent OU processes. Explicitly,

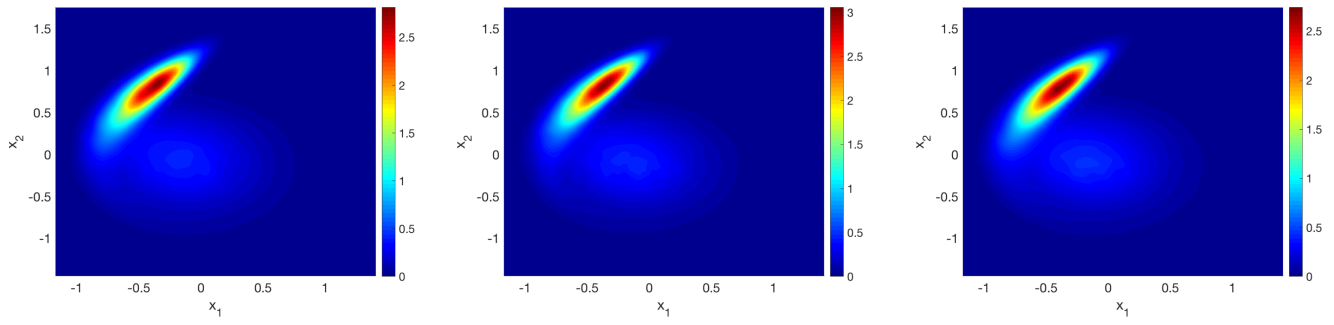
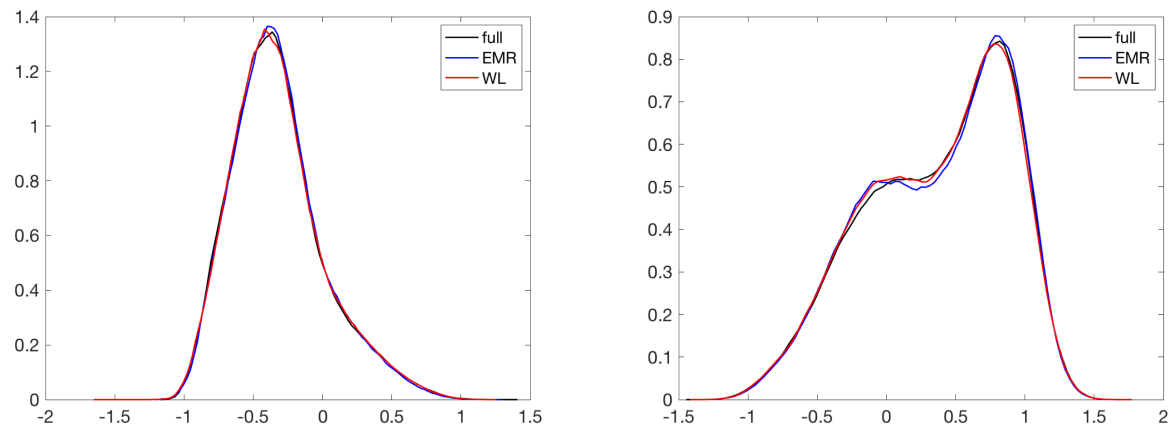


FIG. 2. Two-dimensional probability density functions (PDFs) of the stochastic model (4.1) in the (x_1, x_2) plane, as obtained with (a) the full integration; (b) an integration of the EMR model; and (c) the WL parameterization. The timescale separation parameter used is $h = 0.1$. The PDFs shown here and in Fig. 5 were obtained by using the Matlab R2019a kernel smoothing function *ksdensity*.



(a) x_1-PDF (b) x_2-PDF

FIG. 3. PDFs of (a) the x_1 variable and (b) the x_2 variable. The separation parameter is $h = 0.1$, and the colors used for each method are indicated by the legends inside the panels.

$$\kappa_1(s,\mathbf{x}) = \left\langle \left(-L_{13}x_1 + c_{341}y_2x_1\right)\partial_{y_1}\left(L_{13}y_1(s) + c_{134}y_1(s)y_2(s)\right)\right\rangle \tag{4.6a}$$

$$- \left\langle \left(L_{24}x_2 + c_{413}y_1x_2\right)\partial_{y_2}\left(L_{13}y_1(s) + c_{134}y_1(s)y_2(s)\right)\right\rangle \tag{4.6b}$$

$$= -L_{13}^2 e^{-(\gamma_1/h)s}x_1 + c_{341}c_{134}e^{-(\gamma_1+\gamma_2)s/h}\frac{\sigma_2^2}{2\gamma_2}x_1$$

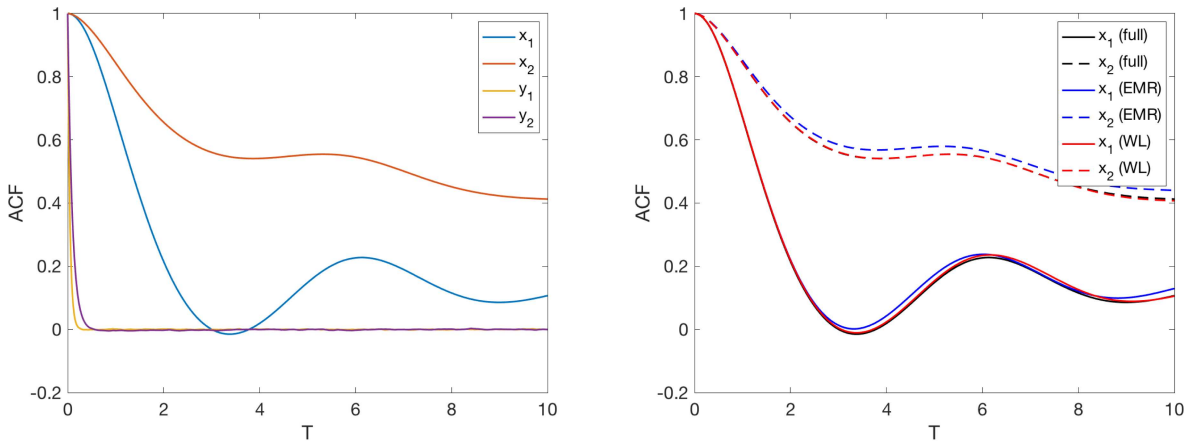
$$+ c_{134}c_{341}e^{-(\gamma_1+\gamma_2)s/h}\frac{\sigma_1^2}{2\gamma_1}x_1, \tag{4.6c}$$

$$\kappa_2(s,\mathbf{x}) = \left\langle \left(-L_{13}x_1 + c_{341}y_2x_1\right)\partial_{y_1}\left(L_{24}y_2(s)\right)\right\rangle$$

$$- \left\langle L_{24}x_2\partial_{y_2}\left(L_{24}y_2(s)\right)\right\rangle \tag{4.6d}$$

$$= -L_{24}^2 e^{-(\gamma_2/h)s}x_2. \tag{4.6e}$$

The reduced-order model obtained herewith does give explicit formulas for the evaluation of the stochastic noise and the memory kernel, independently of the timescale separation h, although these formulas are rather complicated. Still, the scheme remains the same when changing parameter values, so it is flexible in studying different scenarios.



(a) (b)

FIG. 4. Autocorrelation functions for the four variables x_1, x_2, y_1, y_2 obtained (a) from the full model; and (b) the comparison of the corresponding results for x_1, x_2 with the full model, the EMR model, and the WL parameterization. See the legend for the choice of lines; $h = 0.1$.

TABLE I. Empirically estimated EMR model coefficients at the first level, Eq. (3.14a), for $h = 0.1$. First column gives the coefficients for the constant forcing $\mathbf{f}^{(0)}$, the second and third columns indicate the linear component of the vector field $\mathbf{b}^{(0)}$, and the last three columns determine the quadratic form $\mathbf{A}$.

f	x_1	x_2	x_1^2	x_1x_2	x_2^2
−0.314 04	−0.509 54	−0.065 313	0	−1.009 2	0.997 04
−0.153 56	0.123 53	0.219 79	1.0092	−0.997 04	0

B. EMR model and results

1. Basic EMR algorithm implementation

Regarding the data-driven EMR protocol, we integrated the full model with a time step of $\mathrm{d}_\ell t = 10^{-3}$ time units for a duration of $T_\ell = 10^4$ time units in order to learn the model parameters. Then, a separate run was performed in order to examine the ability of the inferred model to reproduce the general statistical features. This time, the EMR system was integrated together with the full model using a time step of $\mathrm{d}_\tau t = 10^{-2}$ time units for a total of $T_\tau = 10^5$ time units. The equations were solved using a fourth-order Runge–Kutta and a Euler–Maruyama method for the deterministic and stochastic components, respectively.

By sampling every time step, we learned an EMR model whose coefficients were explicitly found. The convergence criterion $R^2 \simeq 0.5$ was attained by adding two extra levels, for a total of three. The convergence was not affected by changes in the timescale separation parameter—namely, $h = 0.1$ and $h = 1$ in the case at hand. Probably, the value of h was not that important here because of the low dimensionality and stochastic nature of the hidden process. However, convergence is likely to be altered in more complicated models, as illustrated in Appendix D.

The climatologies of the slow $\mathbf{x}$ variables are obtained using data from the full model, the EMR model, and the WL parameterization. The two-dimensional probability density functions (PDFs) of the stochastic model (4.1) in the (x_1, x_2) plane are shown in Fig. 2. These PDFs were calculated by employing the Matlab R2019a kernel smoothing function *ksdensity*. Their respective marginals are shown in Fig. 3. The agreement between the two methodologies when approximating the clearly non-Gaussian density arising from the full model is clearly excellent.

The timescale separation between the $\mathbf{x}$ variables and the $\mathbf{y}$ variables is clearly depicted in the left panel of Fig. 4, where the fast variables decorrelate almost instantly compared to the slow ones. The approximation of these autocorrelation functions is also obtained using the EMR and WL methods.

TABLE II. Empirically estimated EMR model coefficients at the second level, Eq. (3.14b), for $h = 0.1$. First column gives the coefficients for the constant forcing $\mathbf{f}^{(1)}$, the next two columns indicate the linear coupling to the main level (i.e., the first two columns of $\mathbf{b}^{(1)}$), and the last two columns determine the linear drift for the second level (i.e., the last two columns of $\mathbf{b}^{(1)}$).

$f^{(1)}$	x_1	x_2	$r_1^{(1)}$	$r_2^{(1)}$
0	-5.6397×10^{-4}	-6.9382×10^{-05}	−20.282 3	0.016 165
0	-1.648×10^{-4}	-4.72×10^{-4}	0.086 621	−10.042 6

TABLE III. Empirically estimated EMR model coefficients at the first level, for $h = 1$.

f	x_1	x_2	x_1^2	x_1x_2	x_2^2
−0.311 81	−0.339	−0.429 44	0	−0.934 39	0.979 58
−0.169 25	0.468 33	0.175 13	0.934 39	−0.979 58	0

In general, cf. Ref. 48, the regressions performed in the main level (3.14a) of the EMR model allow one to effectively reconstruct the coefficients of a weakly coupled model; see Appendix D. The EMR methodology, though, only allows for linear coupling between the slow $\mathbf{x}$'s and the fast $\mathbf{y}$'s. The nonlinear coupling between the slow and fast variables in system (4.1) compromises the estimation of the main model parameters in Eq. (3.14a) so that we cannot expect to recover the original, full model's behavior given by (4.1). The EMR model coefficients at the first and second levels are as shown in Tables I and II, respectively.

As discussed in Sec. III B, the EMR has the structure of an MSM and it can be recast into an integrodifferential equation. If one only considers the first added level, the EMR can be readily integrated giving the following equation for the evolution of the slow variables $\mathbf{x} = (x_1, x_2)$:

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) + e^{-\mathrm{D}t}\mathbf{y}(0) + \int_0^t e^{-\mathrm{D}(t-s)}\Sigma \mathrm{d}\mathbf{W}_s + \int_0^t e^{-\mathrm{D}(t-s)}\mathrm{C}\mathbf{x}(s)\mathrm{d}s. \tag{4.7}$$

Here, $\mathbf{W}_s$ is an independent two-dimensional Wiener process and

$$\mathrm{D} = \begin{bmatrix} -19.9982 & 2.1122 \times 10^{-3} \\ -0.775\,28 & -10.116 \end{bmatrix}, \quad \mathrm{C} = \begin{bmatrix} -5 \times 10^{-3} & -5 \times 10^{-4} \\ -1 \times 10^{-3} & -5 \times 10^{-3} \end{bmatrix}, \tag{4.8a}$$

$$\Sigma = \begin{bmatrix} 0.2626 & -0.0014 \\ -0.0014 & 0.5013 \end{bmatrix}. \tag{4.8b}$$

First thing to note is that the matrix C has a small norm and, by virtue of Eq. (4.7), it means that memory effects are going to be very small. On the other hand, the eigenvalues of the matrix D are $\lambda_1 \simeq -20, \lambda_2 \simeq -10$, which are approximatelly the drift coefficients of the uncoupled OU process driving the $\mathbf{y}$ variables. This indicates that the exponential kernel is damping the effects of the $\mathbf{x}$ variables in past times rather quickly. Moreover, the covariance matrix

TABLE IV. Empirically estimated EMR model coefficients at the second level, for $h = 1$.

$f^{(1)}$	x_1	x_2	$r_1^{(1)}$	$r_2^{(1)}$
0	−3.9452e-3	−1.5027e-4	−2.100 1	0.016 509
0	−5.0312e-4	−3.79e-3	0.054 861	−1.121 4

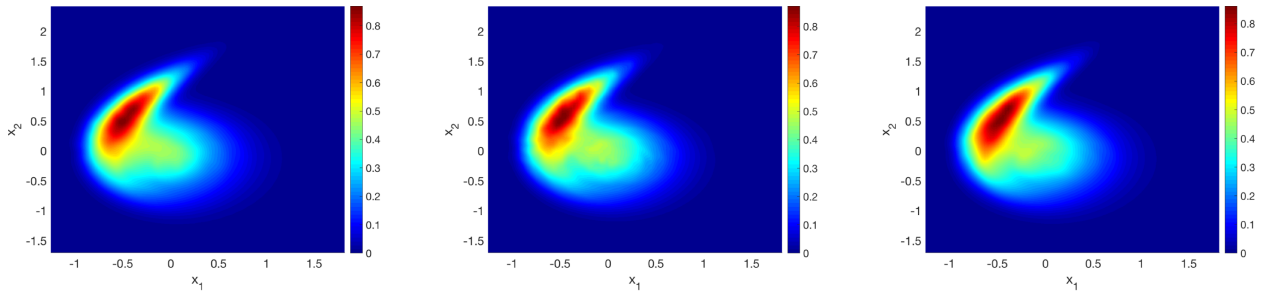


FIG. 5. Two-dimensional smoothed PDFs of the stochastic model (4.1), but with a timescale separation of $h = 1$. Panels (a)–(c) are calculated by integrating the full model, the EMR model, and WL approximation, respectively, as in Fig. 2.

Σ corresponds to that obtained by integrating the $\mathbf{y}$-dynamics independently, according to Eq. (4.4).

Regarding the WL approximation, we stress that the $\mathbf{y}$ variables are no longer present, after taking the averages in its construction. The memory kernel $\mathcal{K}$ in this case differs from the matrix D above, although its dominant terms correspond to its eigenvalues. Note that, if $L_{13} = 0$, the coupling function $\mathbf{C}_{\mathbf{y}}^{\mathbf{x}}$ would project entirely onto the eigenfunction of the OU process associated with the eigenvalue $-(\gamma_1 + \gamma_2)$. The same statement would hold for $c_{134} = 0$, where in this case $\mathbf{C}_{\mathbf{y}}^{\mathbf{x}}$ projects onto the eigenfunctions associated with the eigenvalue $-\gamma_1$.

2. Reduced time scale separation

The parameter h controls the timescale separation in the evolution of the $\mathbf{x}$ and $\mathbf{y}$ variables. Here, we set $h = 1$ so that this separation is reduced by an order of magnitude as illustrated by the autocorrelation functions in Fig. 7. The question to be addressed in this subsection is the effect of such a reduction in the WL and EMR parameterizations and their respective performance.

In the WL parameterization, there is no need to sample the dynamics in order to construct it, since the formulas of Sec. II are explicit and do not depend on h. In the case of model (4.1), the covariance matrix and time correlations of the WL noise correction are thus given by Eq. (4.3) with no reference to h. The memory term, though, is expected to change as the kernel $\mathcal{K}$ will decay more slowly by a factor of 10. Therefore, memory effects are more important, as expected.

The EMR approach, on the other hand, requires a new learning phase for this value of $h = 1$. We used the same numerical integration parameters $\mathrm{d}_\ell t = 10^{-3}$ and $T_\ell = 10^4$ time units as for the previous case. In the first level regression, one observes that the coefficient values listed in Table III are essentially the same from those estimated in the previous case, for $h = 0.1$, and listed

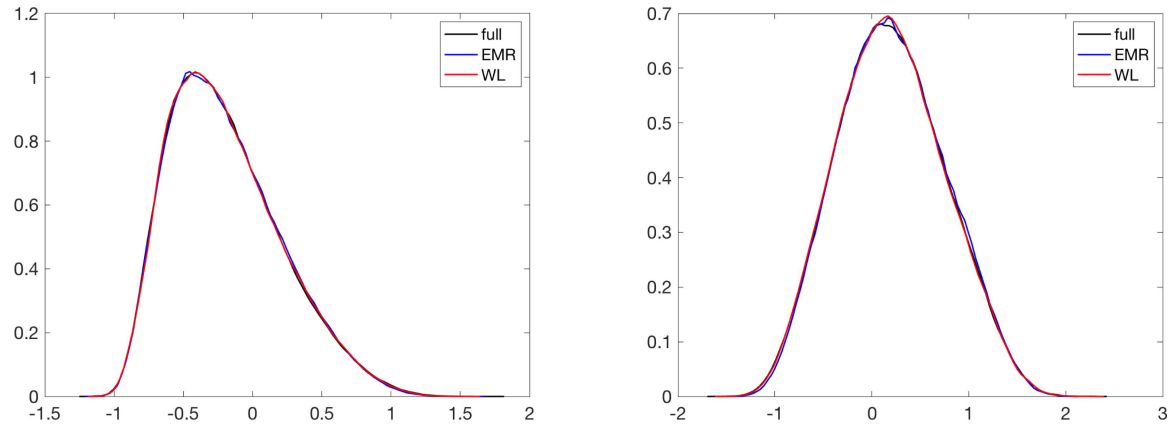


FIG. 6. PDFs of (a) the x_1 variable and (b) the x_2 variable, for a timescale separation of $h = 1$; compare with Fig. 3.

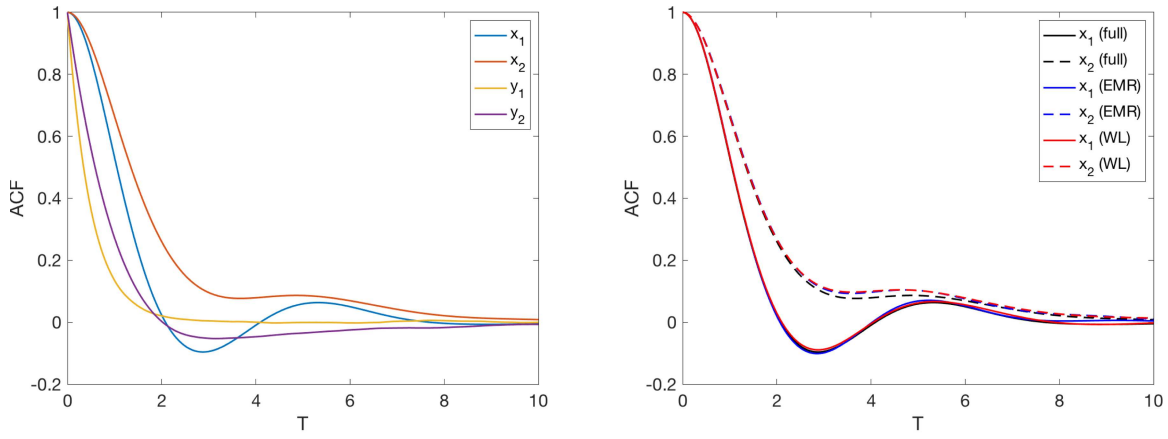


FIG. 7. Autocorrelation functions for the four variables x_1, x_2, y_1, y_2 obtained (a) from the full model; and (b) the comparison of the corresponding results for x_1, x_2 with the full model, the EMR model, and the WL parameterization. See the legend for the choice of lines; $h = 1$.

in Table I, as well as being even more distant from the original ones in the table's first row. The second level coefficients, for $h = 1$, are shown in Table IV.

The covariance matrix Σ of the noise correction is indicated in Eqs. (4.9a) and (4.9b) and it agrees fairly well with the previous values, for $h = 0.1$, as given in Eq. (4.8b). The matrix C indicates that the strength of the memory effects also has a magnitude that is of the same order as that in the previous case of $h = 0.1$, which is rather surprising, given the factor of 10 in timescale separation h; compare Eqs. (4.8a) and (4.9a). This observation tells us that the loss of Markovianity might be intrinsic to the nature of the coupling rather than being due to the time scale separation, even though, in the limit case of infinite scale separation, memory effects will dissapear entirely. The memory kernel, as determined by D, scales almost exactly with the timescale separation and it is expected to change depending on how the coupling functions project onto the eigenspaces of the underlying Orstein–Uhlenbeck process, as discussed more generally earlier in Theorem 2.1.

The performance of both parameterization techniques is summarized in Figs. 5–7. These figures are the exact counterparts of Figs. 2–4 for the reduced timescale separation $h = 1$. First, we note from Fig. 7(a) that for $h = 1$ there is indeed no strict timescale separation, as indicated by the autocorrelation functions obtained from the full model. Second, consideration of Figs. 5(a)–5(c), 6(a), 6(b), and 7(b) shows that neither the WL nor the EMR approach seems to be affected by the timescale reduction

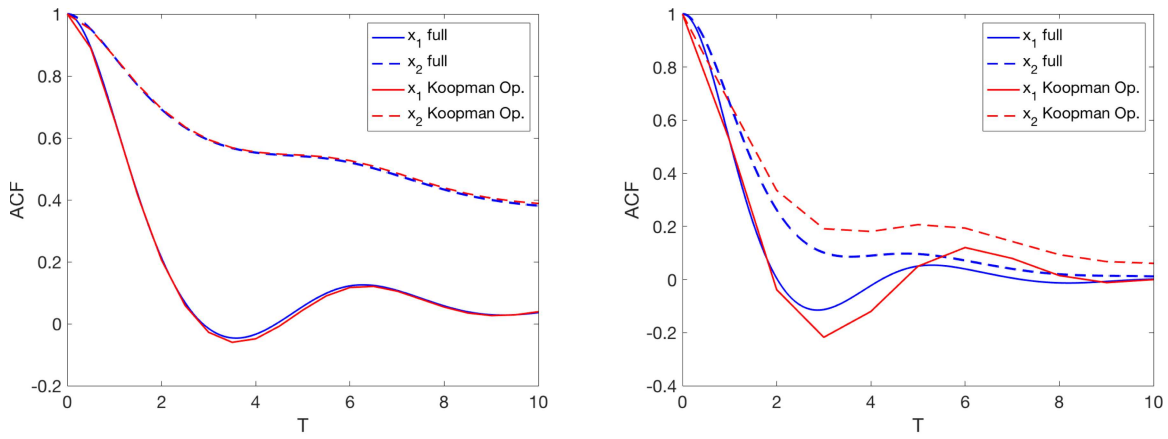


FIG. 8. Autocorrelation functions for the two **x** variables obtained from the full model in the spectral reconstruction using the Koopman operator: (a) $h = 0.1$ and (b) $h = 1$.

$$\mathrm{D} = \begin{bmatrix} -2.1001 & 0.016\,509 \\ 0.054\,861 & -1.1214 \end{bmatrix}, \mathrm{C} = \begin{bmatrix} -4 \times 10^{-3} & -1 \times 10^{-4} \\ -6 \times 10^{-4} & -4 \times 10^{-3} \end{bmatrix}, \tag{4.9a}$$

$$\Sigma = \begin{bmatrix} 0.2618 & 0.0011 \\ 0.0011 & 0.4554 \end{bmatrix}. \tag{4.9b}$$

C. Memory effects

We would like to end this results section by analyzing the role of the memory effects when performing a reduction of the highly idealized model given by Eq. (4.1). For this purpose, we apply the criterion (B4) discussed in the corresponding Appendix B. We thus spectrally approximate the autocorrelation functions of the variables x_1 and x_2 using Eq. (B4) with the Koopman operator $\mathcal{T}_\tau$ estimated using Ulam's method with a transition time of $\tau = 0.5$ time units for the case where $h = 0.1$ and $\tau = 1$ time unit for $h = 1$.

The difference between the two τ-values is due to the fact that we expect the coarse-grained phase space to be sensitive to the system's variability. Hence, if the timescale separation is large, a shorter transition time is required in order to capture the influence of the hidden processes. In fact, a range of transition times was tested in the case of $h = 1$ to find that the optimal value was $\tau = 1$. It follows that the methodology is robust in showing the effects of memory in the projected phase space.

We clearly observe in Fig. 8 that the correlation functions can be accurately reconstructed in the case of large timescale separation $h = 0.1$ [Fig. 8(a)] but not so for $h = 1$ [Fig. 8(b)]. This indicates, naturally, that memory effects are negligible in the first case and relevant in the second case.

V. CONCLUSIONS

To formulate accurate and efficient parameterizations for multiscale processes is a crucial challenge in many areas of science and technology for one of the two reasons: either the numerical simulation of all scales active in a given system is computationally unfeasible or there is a mismatch between the model resolution and the granularity and homogeneity of the observations, as in the case in geophysical flows and in the climate system. Moreover, the construction of parameterizations is instrumental to help understand the nature of nonlinear fluxes across scales and the physical processes responsible for cascades, instabilities, and feedbacks.

There are two main approaches for constructing parameterizations: top-down, by deriving the parameterizations directly from the evolution equations governing the system through the use of suitable approximations; and data-driven, in which the parameterizations are constructed through suitable optimization procedures, which are first tuned in a training phase and then actually used in the prediction phase. Both approaches aim to derive the effective dynamics for the variables of interest: formally, this is achieved by applying the Mori–Zwanzig projection operator[60,90] to the full dynamics. The result of doing so is to describe the impact of the hidden variables by formulating a generalized Langevin equation (GLE) (1.3b) for the variables of interest that includes a deterministic, a stochastic, and a non-Markovian component.

Top-down and data-driven approaches are conceptually complementary and have different practical advantages and disadvantages. In this paper, we have shown the fundamental equivalence between a top-down and a data-driven approach that have been formulated and applied in the recent literature. This equivalence was illustrated schematically in Fig. 1.

We first revisited in Sec. II the WL parameterization of Refs. 88 and 89, which relies on an assumption of weak coupling between the hidden and observed variables and have extended the previous results by considering more general coupling classes. We have also shown that the perturbative expansion that yields the WL parameterization is exact when the coupling between the hidden and resolved variables is additive.

The Dyson formalism (2.12) appears to be essential for computing the effects of the hidden processes on the dynamics of the observed variables, when working at the level of the system's observables. This methodology is explicit in the sense that no information about the actual coupled process is needed, because the formal computations are performed by considering the limit in which no coupling is present. Other advantages of this approach are that it can be implemented without the need for any hypothesis on the timescale separation between the hidden variables and the observed ones and that it is also scale adaptive.[83]

We addressed systematically the problem of re-Markovianizing the WL memory equation, which was first pointed out in Ref. 87 and discussed further in Sec. II D. In this example, a system (2.42) with one observed and two hidden variables that yielded a scalar WL parameterization was re-Markovianized to a Markovian system with just two scalar differential equations. Throughout Sec. II, we provided a broader framework for re-Markovianization. This framework was presented for a scalar equation in Theorem 2.1 and described for higher dimensional systems in Remark 2.5. The required assumptions for this treatment boil down to certain spectral properties of the Koopman operator for the hidden variables.

The multilevel structure of the re-Markovianization obtained in Sec. II motivated the comparison with multilayer stochastic models (MSMs) in Sec. III. Such MSMs arise naturally in data-driven reduction methods and they had been shown in Ref. 43 to approximate the GLE predicted by Mori[60] and Zwanzig.[91]

We showed in Sec. III A that a seemless application of the WL parameterization solves the MSM of Eq. (3.1) and coincides with its Itô integration, cf. Ref. 43. Note that an MSM can be obtained from partial observations of the coupled system, which amounts to the special case of the data-driven empirical model reduction (EMR) methodology.[43,45,46,48,49]

The EMR methodology was revisited here in Sec. III B and it is, in principle, dual to the WL parameterization, in the sense that only partial observations of the coupled system are required, without the need for knowing the actual equations of motion. Comparing the multilevel structure of Eq. (2.40) with that of Eq. (3.14) suggests that the Koopman eigenvalues λ_j highlighted in Theorem 2.1 may help provide insights into the number of levels needed for EMR to converge. This practical role of the λ_j's deserves, therewith, a more careful examination in further work.

Additionally, we considered in Sec. IV a conceptual climate model to which we applied both of the methodologies revisited

herein. Since both the MSM and the WL parameterization yield a memory equation that involves integrals and stochastic noise, we were able to compare their structure, as well as their statistical outputs. We found that both methodologies produced equivalent numerical results and that the memory kernel and noise predicted in the WL parameterization agreed with what was found using the data-driven EMR approach.

Concluding, our viewpoint is complementary to the dynamic mode decomposition[59,69,72] as it uses the basis of eigenvectors of the Koopman operator to construct the projected—in the sense of Mori–Zwanzig—dynamics of the observables of interest, which is then recast in the Markovian form using the multilevel Markovian model framework, where the number of levels corresponds to the number of eigenvectors of the Koopman operators one considers in the reconstruction of the dynamics. In a nutshell, our findings support, on the one hand, the physical basis and robustness of the EMR methodology and, on the other hand, illustrate the practical relevance of the WL perturbative expansion used for deriving the parameterizations.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge useful exchanges with Georg Gottwald, Jeroen Wouters, and Gabriele Vissio. The authors would like to thank both reviewers for their encouraging comments and insightful suggestions. V.L. acknowledges the support received from the European Union's Horizon 2020 research and innovation program through the project CRESCENDO (Grant Agreement No. 641816). This article is TiPES contribution No. 56; the TiPES (Tipping Points in the Earth System) project has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 820970. This research was partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center, by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 810370), and by the Office of Naval Research (ONR) Multidisciplinary University Research Initiative (MURI) under Grant No. N00014-20-1-2023. This paper has also been partially supported by the EIT Climate-KIC; EIT Climate-KIC is supported by the European Institute of Innovation & Technology (EIT), a body of the European Union, under Grant Agreement No. 190733, and by the Russian Science Foundation (Grant No. 20-62-46056).

APPENDIX A: PROOF OF THEOREM 2.1

Proof. The aim is to show that under the assumptions of this theorem—which require that the coupling function $\mathbf{C}^{\mathbf{x}} : \mathbb{R} \longrightarrow \mathbb{R}$ projects entirely onto span $\{\psi_j,\ j = 1, \ldots, N\}$—the memory and noise terms of the WL equation (2.23) are obtained from the term $\epsilon\Lambda \cdot \mathbf{Z}(t)$ in Eq. (2.30a), after integration of Eq. (2.30b).

Step 1. In this step, we expand the memory term and the lag correlations of the noise in the WL equation (2.23) in terms of the leading eigenelements of the uncoupled Koopman operator $\mathcal{L}_0^{\mathbf{y}}$. These expansions will serve us in Step 2 to compare the noise and memory terms of the WL equation with those produced after integration of Eq. (2.30b).

Let the λ_j be sorted as in the hypotheses of the theorem. Hence, to distinguish real and complex conjugate eigenvalues and for notational convenience, we introduce the set of indices I_r and I_+ defined as

$$I_r = \{j \in \{1, \ldots, N\} : \lambda_j \text{ is real}\}, \tag{A1a}$$

$$I_+ = \{j \in \{1, \ldots, N\} : \Im\mathrm{m}\lambda_j > 0\}. \tag{A1b}$$

An immediate consequence that is used several times throughout the proof is that the sum of the eigenvalues is real and may be split as follows:

$$\sum_{j=1}^{N} \lambda_j = \sum_{j\in I_r} \lambda_j + \sum_{j\in I_+} \lambda_j + \sum_{j\in I_+} \overline{\lambda_j} = \sum_{j=1}^{N} \Re\mathfrak{e}\lambda_j \in \mathbb{R}. \tag{A2}$$

As previously stated, we expand the mean and correlation functions of the scalar noise term η in WL equation (2.23) in terms of the eigenpairs. The mean is zero by assumption, but the autocorrelation function can be expanded as follows, based on Eq. (B2):

$$\langle \eta(t)\eta(0)\rangle = C_{\mathbf{C}^{\mathbf{x}},\mathbf{C}^{\mathbf{x}}}(t) = \sum_{j=1}^{N} e^{\lambda_j t}\alpha_j\beta_j, \tag{A3}$$

herein, α_j and β_j are as defined in (2.32a) and (2.32b), respectively. The expansion of the correlation function in Eq. (A3) is a finite sum by virtue of the assumption that $\mathbf{C}^{\mathbf{x}}$ lies in span $\{\psi_j,\ j = 1, \ldots, N\}$ and therefore there is no contribution from the essential spectrum.

Regarding the complex scalars α_j and β_j defined in (2.32a) and (2.32b), it follows that for each j such that $\lambda_j = \overline{\lambda_{j+1}}$, we get $\alpha_j = \overline{\alpha_{j+1}}$ and $\beta_j = \overline{\beta_{j+1}}$. Indeed,

$$\alpha_j = \int \nu(\mathrm{d}\mathbf{y})\overline{\psi_j^*(\mathbf{y})}\mathbf{C}^{\mathbf{x}}(\mathbf{y}) = \overline{\int \nu(\mathrm{d}\mathbf{y})\psi_j^*(\mathbf{y})\mathbf{C}^{\mathbf{x}}(\mathbf{y})} \tag{A4a}$$

$$= \overline{\int \nu(\mathrm{d}\mathbf{y})\overline{\psi_{j+1}^*(\mathbf{y})}\mathbf{C}^{\mathbf{x}}(\mathbf{y})} = \overline{\alpha_{j+1}}, \tag{A4b}$$

in which we have exploited the fact that $\psi_j = \overline{\psi_{j+1}}$ when λ_j is complex and j in $\{1, \ldots, N\}$.

The same proof can be repeated for β_j. Such a conjugacy relation also holds for the gradients of the eigenfunctions $\nabla\psi_j$, for those j in $\{1 \ldots, N\}$ such that $\lambda_j = \overline{\lambda_{j+1}}$. This is observed by the following equality:

$$\nabla\psi_j(\mathbf{y}) = \nabla\overline{\psi_{j+1}}(\mathbf{y}) = \overline{\nabla\psi_{j+1}(\mathbf{y})}, \tag{A5}$$

since ∇ is a differential operator that only involves here differentiation with respect to a real variable.

Exploiting (A5), the memory kernel $\mathcal{K}$ then expands as [recalling that $\mathcal{R}(t) = 0$]

$$\begin{aligned}\mathcal{K}(t,s,x) &= \mathbf{C}^{\mathbf{y}}(x(s)) \cdot \left\langle \nabla \sum_{j=1}^{N} e^{\lambda_j (t-s)} \alpha_j \psi_j(\mathbf{y}) \right\rangle \\ &= \mathbf{C}^{\mathbf{y}}(x(s)) \cdot \sum_{j \in I_r} e^{\lambda_j (t-s)} \alpha_j \langle \nabla \psi_j(\mathbf{y}) \rangle + \mathbf{C}^{\mathbf{y}}(x(s)) \\ &\quad \cdot \sum_{j \in I_+} e^{\lambda_j (t-s)} \alpha_j \langle \nabla \psi_j(\mathbf{y}) \rangle \\ &\quad + \mathbf{C}^{\mathbf{y}}(x(s)) \cdot \sum_{j \in I_+} \overline{e^{\lambda_j (t-s)} \alpha_j \langle \nabla \psi_j(\mathbf{y}) \rangle}, \end{aligned} \tag{A6}$$

which leads to

$$\mathcal{K}(t,s,x) = \mathbf{C}^{\mathbf{y}}(x(s)) \cdot \sum_{j=1}^{N} \mathfrak{Re}\left(e^{\lambda_j (t-s)} \alpha_j \langle \nabla \psi_j(\mathbf{y}) \rangle \right). \tag{A7}$$

Note that to go from (A6) to (A7), we have made use of the aforementioned conjugacy relations that led to a real-valued memory kernel at the end. With (A3) and (A7) at hand, the noise and memory terms in the WL equation (2.23) are thus characterized in terms of the leading eigenelements of the uncoupled Koopman operator $\mathcal{L}_0^{\mathbf{y}}$.

Step 2. The second step consists of analyzing the noise and memory terms produced by integration of Eq. (2.30b) and to compare these terms with those of the WL equation.

Performing an Itô integration of Eq. (2.30b) leads to

$$\mathbf{Z}(t) = e^{\mathrm{D}t}\mathbf{Z}(0) + \underbrace{\int_0^t e^{\mathrm{D}(t-s)} \Sigma \mathrm{d}W_s}_{\text{noise term}} + \epsilon \underbrace{\int_0^t e^{\mathrm{D}(t-s)} \mathbf{R}(x(s)) \mathrm{d}s}_{\text{memory term}}, \tag{A8}$$

where (for simplicity) we have assumed that the initial condition distributes normally with mean zero and variance equal to the identity matrix, and the function $\mathbf{R} : \mathbb{R} \longrightarrow \mathbb{C}^N$ is as defined in (2.33). The noise and memory contributions of $\mathbf{Z}(t)$ are as indicated by the brackets in (A8).

Let us denote the noisy component of Eq. (A8) as $\mathbf{q}(t)$ in $\mathbb{R}^N$. Then, it is clear that $\mathbf{q}$ has zero mean and the lag cross correlations read as

$$\mathbb{E}\left(\mathbf{q}(t)\mathbf{q}^{\top}(0)\right) = e^{\mathrm{D}t} = \begin{bmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_N t} \end{bmatrix}. \tag{A9}$$

Let Λ be defined as in (2.31) and let us calculate the mean and lag correlations of the one-dimensional stochastic process $\Lambda \cdot \mathbf{q}$, aimed at approximating η in the WL equation. First, note that $\Lambda \cdot \mathbf{q}$ is a zero-mean Gaussian process, with lag correlations given by

$$\mathbb{E}\left((\Lambda \cdot \mathbf{q}(t)) (\Lambda \cdot \mathbf{q}(0)) \right) = \left(e^{\mathrm{D}t}\Lambda\right) \cdot \Lambda. \tag{A10a}$$

Now, expanding $\left(e^{\mathrm{D}t}\Lambda\right) \cdot \Lambda$ in (A10a) shows that we recover the right-hand side (RHS) of (A3).

However, the noise term η in the WL equation is a real-valued stochastic process, and we are dealing with complex scalars, so therefore we still have to show that $\Lambda \cdot \mathbf{q}(t)$ is real for every t in $\mathbb{R}$. To do so, let us denote by $\mathbf{w}^{\top} = [w_1, \ldots, w_N]$ any arbitrary row vector in $\mathbb{R}^N$. In other words, $\mathbf{w}$ is an arbitrary column vector with real entries.

Consider the following inner product:

$$\Lambda \cdot e^{\mathrm{D}t}\Sigma\mathbf{w} = \Lambda \cdot e^{\mathrm{D}t} \begin{bmatrix} \sqrt{-2\mathfrak{Re}\lambda_1} & & \\ & \ddots & \\ & & \sqrt{-\mathfrak{Re}\lambda_N} \end{bmatrix} \mathrm{H}\mathbf{w} \tag{A11a}$$

$$= \Lambda \cdot \begin{bmatrix} e^{\lambda_1 t}\sqrt{-2\mathfrak{Re}\lambda_1} & & \\ & \ddots & \\ & & e^{\overline{\lambda_2} t}\sqrt{-2\mathfrak{Re}\lambda_N} \end{bmatrix} \mathrm{H}\mathbf{w}. \tag{A11b}$$

By construction of the matrix H in Eqs. (2.34) and (2.35), the product $\mathrm{H}\mathbf{w}$ is given component-wise, for $j = 2, \ldots, N$, as

$$[\mathrm{H}\mathbf{w}]_j = \begin{cases} w_j & \text{if } j \in I_r \text{ or, } j \in I_+, \\ w_{j-1} & \text{if } \lambda_j = \overline{\lambda_{j-1}}, \end{cases} \tag{A12}$$

while $[\mathrm{H}\mathbf{w}]_1 = w_1$. This implies, in particular, that $[\mathrm{H}\mathbf{w}]_j = [\mathrm{H}\mathbf{w}]_{j+1}$ whenever $\lambda_j = \overline{\lambda_{j+1}}$. As a consequence, we get

$$\begin{aligned}\Lambda \cdot e^{\mathrm{D}t}\Sigma\mathbf{w} &= \sum_{j \in I_r} \alpha_j^{1/2} \beta_j^{1/2} e^{\lambda_j t} \sqrt{-2\lambda_j} w_j + \sum_{j \in I_+} \alpha_j^{1/2} \beta_j^{1/2} e^{\lambda_j t} \sqrt{-2\mathfrak{Re}\lambda_j} w_j \\ &\quad + \sum_{j \in I_+} \overline{\alpha_j^{1/2} \beta_j^{1/2} e^{\lambda_j t}} \sqrt{-2\mathfrak{Re}\lambda_j} w_j, \end{aligned} \tag{A13}$$

which shows that $\Lambda \cdot e^{\mathrm{D}t}\Sigma$ is a real-valued quantity, and thus for any realization of the N-dimensional Wiener process W_t, the product $\Lambda \cdot e^{\mathrm{D}(t-s)}\Sigma \mathrm{d}W_s$ is real, and hence $\Lambda \cdot \mathbf{q}(t)$ is also real for every t.

Finally, we are left with showing that the memory kernel $\mathcal{K}$ of the WL equation coincides with that of the memory term in Eq. (A8), when multiplied by the vector Λ. To do so, we exploit the expansion (A7) of $\mathcal{K}$ for this comparison, namely, using the expression of $\mathbf{R}$ in (2.33), we observe that

$$\begin{aligned}\Lambda \cdot \int_0^t e^{\mathrm{D}(t-s)} \mathbf{R}(x(s)) \mathrm{d}s &= \mathbf{C}^{\mathbf{y}}(x(s)) \cdot \sum_{j \in I_r} e^{\lambda_j (t-s)} \alpha_j \langle \nabla \psi_j(\mathbf{y}) \rangle \\ &\quad + \mathbf{C}^{\mathbf{y}}(x(s)) \cdot \sum_{j \in I_+} e^{\lambda_j (t-s)} \alpha_j \langle \nabla \psi_j(\mathbf{y}) \rangle \\ &\quad + \mathbf{C}^{\mathbf{y}}(x(s)) \cdot \sum_{j \in I_+} \overline{e^{\lambda_j (t-s)} \alpha_j \langle \nabla \psi_j(\mathbf{y}) \rangle} \\ &= \mathbf{C}^{\mathbf{y}}(x(s)) \cdot \sum_{j=1}^{N} \mathfrak{Re}\left(e^{\lambda_j (t-s)} \alpha_j \langle \nabla \psi_j(\mathbf{y}) \rangle \right), \end{aligned} \tag{A14}$$

which indeed coincides with the expression of $\mathcal{K}$, as desired. The proof is complete. □

APPENDIX B: SEMIGROUP PROPERTY OF THE PROJECTED KOOPMAN OPERATOR FAMILY

It was shown in Ref. 16, Theorem A, that projection onto a reduced state space is closely related with a coarse graining of the

(full) probability transitions on the original system's attractor, while Theorem 2 in Ref. 19 dealt recently with the impact of such a projection in terms of reduction of the Koopman semigroup. In Ref. 19, the authors proposed a criterion based on the spectral theory of Markov semigroups to ascertain whether the reduced state space associated with a given projection can fully explain the statistics of the desired variables. This approach provides potential insights into the need for modeling non-Markovian effects by inspecting the loss of the semigroup property, as explained below.

Moreover, it follows from Ref. 19 that the analysis of correlation functions is not only of physical interest but also of methodological utility. Correlation functions can be defined by means of the Koopman operator or, dually, by means of the transfer operator's providing the solution of the Liouville equation.

Let μ denote an ergodic invariant measure of the system and takes two observables Φ_1 and Φ_2 in the space L^2_μ of zero-mean functions that are square-integrable with respect to μ. Assume furthermore that the spectrum of the operator $\mathcal{L}$ in L^2_μ is a pure point spectrum, given by the eigenvalues $\{\lambda_j\}_{j=1}^\infty$ and their associated eigenfunctions $\{\psi_j\}_{j=1}^\infty$, where the eigenvalues are ordered by their decreasing real parts.

Then, the correlation function $C_{\Phi_1,\Phi_2}(t)$ between the functions Φ_1 and Φ_2 is given by

$$C_{\Phi_1,\Phi_2}(t) = \int \Phi_1 \cdot e^{t\mathcal{L}}\Phi_2 \mathrm{d}\mu = \int e^{t\mathcal{L}^*}\Phi_1 \cdot \Phi_2 \mathrm{d}\mu, \tag{B1}$$

and it can be expanded, formally, as

$$C_{\Phi_1,\Phi_2}(t) = \sum_{j=1}^{\infty} e^{\lambda_j t}\left\langle \Phi_1, \psi_j\right\rangle_\mu \left\langle \psi_j^*, \Phi_2\right\rangle_\mu. \tag{B2}$$

The dual operators in (B1) and the adjoint eigenvectors in (B2) are indicated by the superscript $(\cdot)^*$, while $\langle\cdot,\cdot\rangle_\mu$ denotes the inner product. We refer to Corollary 1 in Ref. 19 for a proof of (B2) in the context of Markov semigroups. The proof actually applies to the case of the Koopman semigroups considered here as long as the Koopman semigroup U_t defined by (2.4) is a strongly continuous semigroup in L^2_μ. The RHS of Eq. (B2) consists of a linear combination of exponential terms whose coefficients are calculated by projecting Φ_1 and Φ_2 onto the corresponding eigenspaces. These coefficients weight each exponential function and they can become exceedingly large if the Koopman operator deviates very much from normality.[77] Note also that the set of eigenvalues λ_j's play a key role in defining the response of the system to perturbations.[55,75]

The interactions between the resolved and hidden variables that are modeled by the Dyson expansion of the Koopman operator in Sec. I A may introduce memory effects into the closed, reduced model for the $\mathbf{x}$ variables, as given by Eqs. (2.18)–(2.22). In certain situations, such memory effects can be neglected, even in the absence of exact slaving relationships between the resolved and hidden variables.[13] But the loss of slaving relationships requires, in general, an explicit representation of memory effects[12] to achieve an efficient model reduction.

Furthermore, it was shown in Refs. 16 and 19 that the reduction of the Koopman semigroup to observables that act only on the reduced state space leads, in most circumstances, to a family of operators that, while Markovian, no longer satisfy the semigroup property. One might then ask to which extent this loss of the semigroup property arising from the reduction, and the related emergence of memory effects, is crucial for providing a faithful reduced model of the observed variables.

When considering reduced state spaces obtained by projection, along with observables Φ_1 and Φ_2 defined on them, Theorem 2 in Ref. 19 shows the existence of a family of Markov operators $\{\mathcal{T}_t\}_{t\geq 0}$ that satisfies

$$\begin{aligned}\int \Phi_1 \cdot \mathcal{T}_t\Phi_2 \mathrm{d}\mu_{\mathbf{x}} &= \int_0^t [\Phi_1 \circ \pi_{\mathbf{x}}] \cdot e^{\mathcal{L}t}[\Phi_2 \circ \pi_{\mathbf{x}}]\mathrm{d}\mu \\ &= C_{(\Phi_1\circ\pi_{\mathbf{x}},\Phi_2\circ\pi_{\mathbf{x}})}(t),\end{aligned} \tag{B3}$$

for every $t \geq 0$, where $\pi_{\mathbf{x}}$ is the canonical projection onto the reduced subspace and $\mu_{\mathbf{x}}$ is the *disintegrated* or *sample measure* associated with $\pi_{\mathbf{x}}$; see Ref. 19, Remark 3. However, due to the projection, the semigroup property is lost, namely, $\mathcal{T}_s\mathcal{T}_t \neq \mathcal{T}_{t+s}$ for some t, s.

Following the reasoning given above, one can establish a criterion for the need to model a memory contribution when performing the model reduction. Formally, if there exist $\tau > 0$ and $T \in \mathbb{N}$ such that for every $t \in \{k\tau \in \mathbb{R} : 0 \leq k \leq T\}$, we have

$$C_{(\Phi_1\circ\pi_{\mathbf{x}},\Phi_2\circ\pi_{\mathbf{x}})}(t) = \int \Phi_1 \cdot \mathcal{T}_t\Phi_2 \mathrm{d}\mu = \int \Phi_1 \cdot (\mathcal{T}_\tau)^k \Phi_2 \mathrm{d}\mu, \tag{B4}$$

and one can say that the semigroup is preserved, to some extent, depending on how large T can be in Eq. (B4). Other such criteria are available in the context of mutually dual Koopman and transfer operators. Thus, Tantet *et al.*[76] had already considered empirical ways of quantifying the loss of the semigroup property in reduced dimensions.

The interpretation of τ comes from the practical implementation of the methodology and it is usually referred to as the *transition time*. Indeed, numerically, the approximation of such Markov operators is done by Ulam's method, by projecting them onto a finite basis, typically using the characteristic functions of certain domains of phase space. Then, the transitions between domains—after an adequate transition time τ—are counted to obtain matrix estimates of the operator $\mathcal{T}_\tau$ acting on the reduced phase space. Hence, one seeks a suitably small, or large,[75] transition time to obtain the best candidate for applying Eq. (B4), see also Ref. 19, Sec. 3.3. Thus, Eq. (B4) can be implemented in practice this way in order to (potentially) reconstruct correlation functions on the whole phase space. A very simple illustration of such a transfer operator calculation is given in Ref. 81.

APPENDIX C: ITÔ INTEGRATION OF THE MSM

In the main text, we proposed a solution of the MSM given by Eq. (3.1) using the Dyson expansion for the linear operators involved in the backward Kolmogorov equation—advection acting on functions. Therefore, we substituted nonlinear ordinary differential equations for a partial differential equation, for the sake of having linear operators in hand. The same solution can be attained by direct integration of the MSM in the form (3.1). We convolute, in the Itô sense, Eq. (3.1b) to find an explicit solution for $\mathbf{y}(t)$ when

$d_1 = d_2$,

$$\mathbf{y}(t) = e^{-\mathrm{D}t}\mathbf{y}(0) + \int_0^t e^{-\mathrm{D}(t-s)}\Sigma \mathrm{d}W_s + \epsilon \int_0^t e^{-\mathrm{D}(t-s)}\mathrm{C}\mathbf{x}(s)\mathrm{d}s, \quad \text{(C1)}$$

here, $\mathbf{y}(0)$ indicates an initial state that can be assumed to be distributed in a prescribed way. For more general, nonlinear MSMs considered there, see Ref. 43, Proposition 3.3.

By substituting the expression (C1) into Eq. (3.1a), we find an exact expression for the evolution of $\mathbf{x}(t)$,

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t)) + \epsilon e^{-\mathrm{D}t}\mathbf{y}(0) + \epsilon \int_0^t e^{-\mathrm{D}(t-s)}\Sigma \mathrm{d}W_s + \epsilon^2 \int_0^t e^{-\mathrm{D}(t-s)}\mathrm{C}\mathbf{x}(s)\mathrm{d}s, \quad \text{(C2)}$$

in which the memory effects in the fourth term are of second order in ϵ. Note that the ϵ-order terms arise from a noise realization in the decoupled regime, whereas the memory term is exclusive due to the coupling of the main variables with the hidden ones. Hence, the only degree of freedom left is the distribution of the initial states $\mathbf{y}(0)$.

APPENDIX D: THE COUPLED L84–L63 SYSTEM

The EMR methodology's ability to capture the statistics of low-dimensional dynamical systems was illustrated in Ref. 48, where the authors considered the L63 system[53] as a test case in which the phase space can be fully sampled. Moreover, provided that the integration time step is short enough, the parameters of the underlying model can be fully captured with a high degree of confidence.

Here, we repeat the analysis of Ref. 48 to illustrate the effectiveness of EMR in capturing statistical and dynamical properties in an extended system. The model we consider is the result of coupling the $\mathbf{X} = (X, Y, Z)$ variables of the L84 system[54] with the $\mathbf{x} = (x, y, z)$ variables of the L63 system,[53] namely,

$$\dot{X} = -Y^2 - Z^2 - aX + a(F_0 + hx), \quad \text{(D1a)}$$

$$\dot{Y} = XY - bXZ - Y + G, \quad \text{(D1b)}$$

$$\dot{Z} = XZ + bXY - Z, \quad \text{(D1c)}$$

$$\dot{x} = \tau s(y - x), \quad \text{(D1d)}$$

$$\dot{y} = \tau(\rho x - y - xz), \quad \text{(D1e)}$$

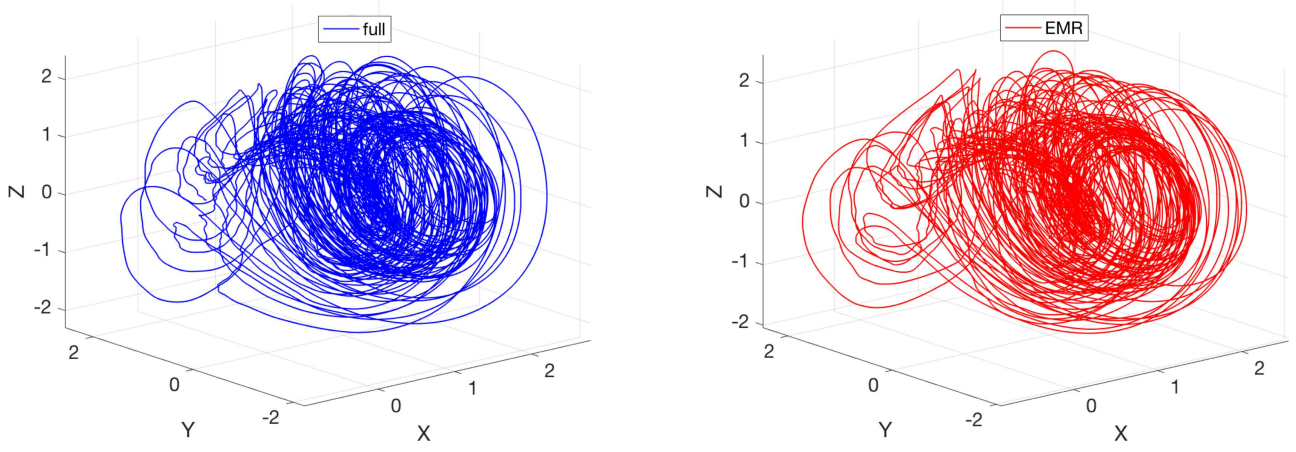
$$\dot{z} = \tau(xy - \beta z), \quad \text{(D1f)}$$

the parameter values are $a = 0.25, b = 4$, $F_0 = 8, G = 1$, and $s = 10$, $\rho = 28, \beta = 8/3$, respectively. The parameter h measures the strength of the coupling, while τ scales the rate of change in the L63 system and, therewith, the timescale ratio between the two subsystems.

This system is a skew-product, in the sense of Ref. 74, since the coupling is one-way only, with the L63 system driving the L84 dynamics. Hence, one has—as noted in Ref. 82—a fully Markovian parameterization of the L63 variables. Furthermore, the correlation function that defines the stochastic noise $\eta(t)$ can be further expanded and simplified, with respect to Eq. (2.16). One can, in fact, write explicitly

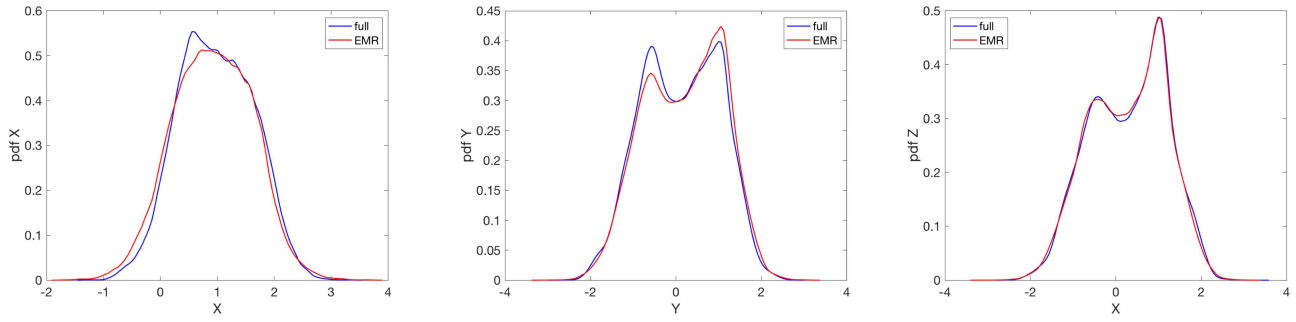
$$C(\eta(0), \eta(t)) = \langle (x(0), 0, 0) \cdot (x(t), 0, 0) \rangle, \quad \text{(D2)}$$

where the angular brackets $\langle \cdot \rangle$ indicate averages with respect to the physical measure associated with the L63 system. Since L84 does not



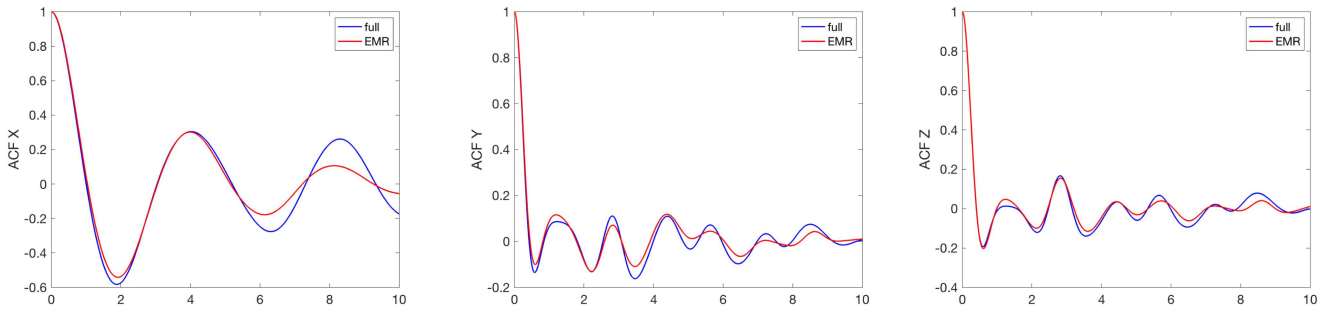
(a) Full model (b) EMR model

FIG. 9. Trajectories of the L84–L63 model in the three-dimensional (X, Y, Z) phase space of the L84 model, for $h = 0.25$ and 200 time units: (a) for the full L84–L63 model governed by Eq. (D1) (blue) and (b) for the EMR model (red).



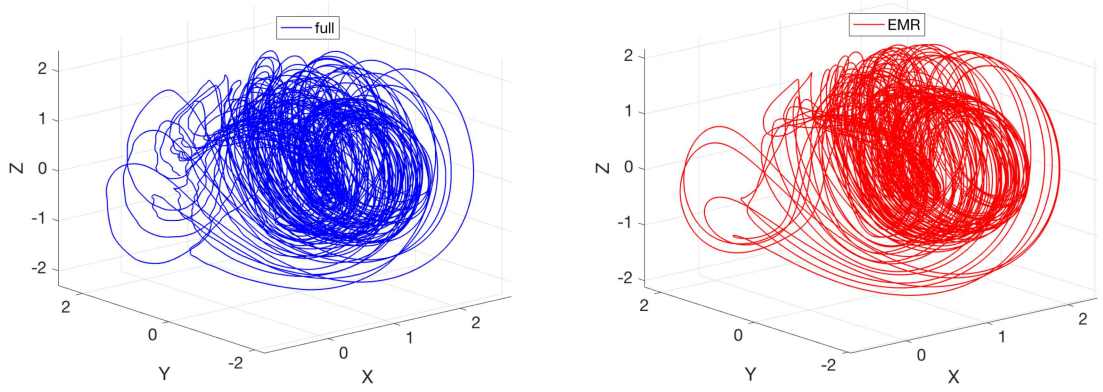
(a) X-PDF (b) Y-PDF (c) Z-PDF

FIG. 10. Smoothed PDFs of the L84–L63 variables (a) X, (b) Y, and (c) Z with a coupling strength of $h = 0.25$. The blue curve corresponds to the full model and the red curve corresponds to the EMR model. These PDFs and those in Fig. 13 were obtained by using the Matlab R2019a kernel smoothing function *ksdensity*.



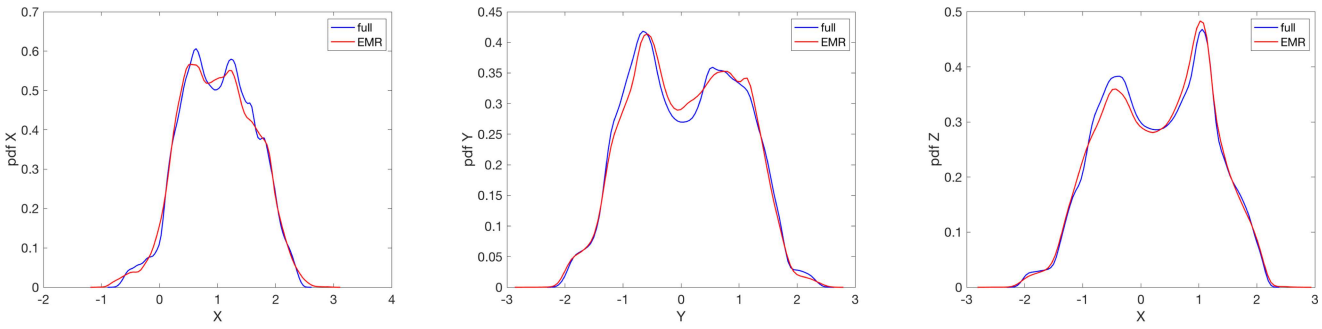
(a) X-ACF (b) Y-ACF (c) Z-ACF

FIG. 11. Autocorrelation functions (ACFs) of the L84–L63 variables (a) X, (b) Y, and (c) Z for a coupling strength of $h = 0.25$.



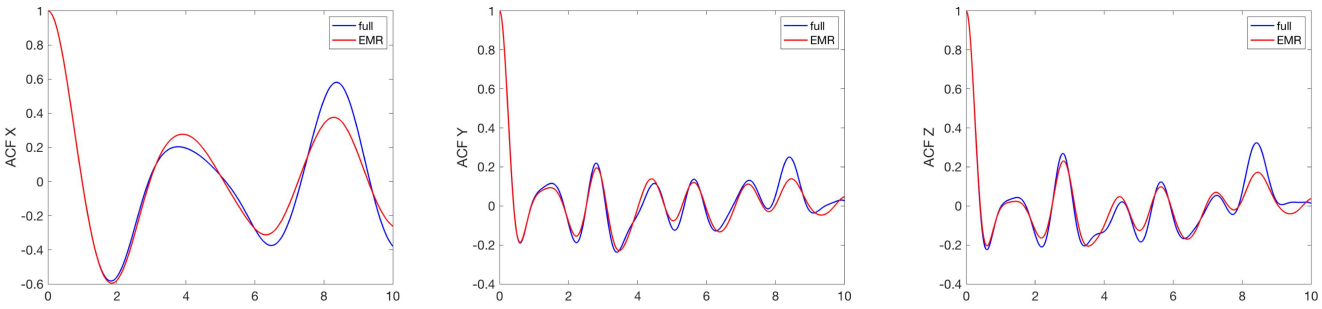
(a) Full model (b) EMR model

FIG. 12. Example trajectories of the L84–L63 model on the (X, Y, Z) domain with a coupling strength of $h = 0.025$ integrated for 200 time units. Subfigure (a) corresponds to the full model (D1) (blue) and subfigure (b) refers to the EMR model (red).



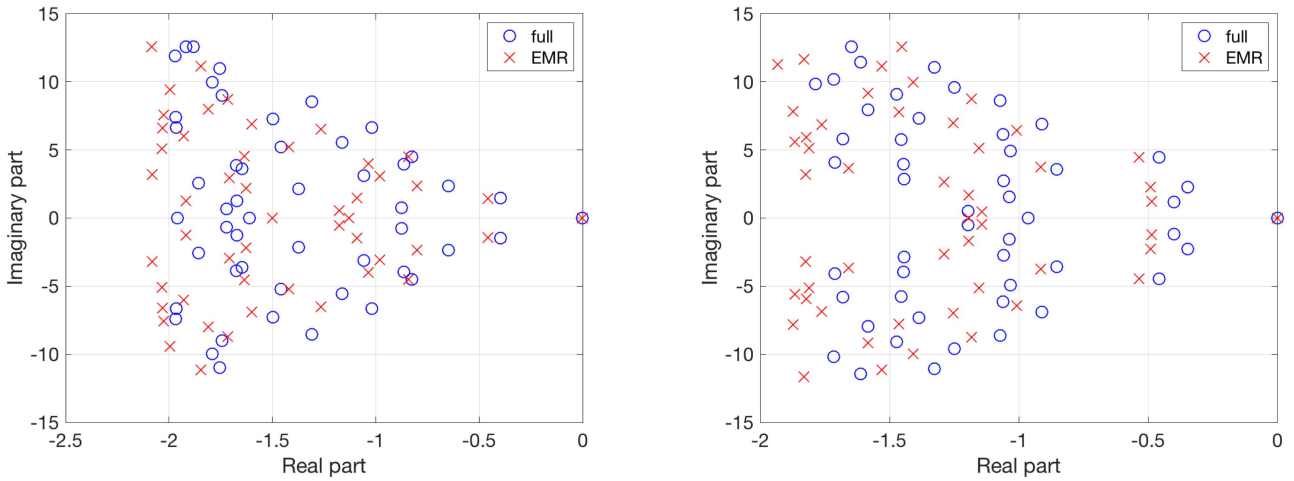
(a) X-PDF (b) Y-PDF (c) Z-PDF

FIG. 13. Smoothed PDFs of the L84–L63 variables (a) X, (b) Y, and (c) Z, with a coupling strength of $h = 0.025$. The blue curve corresponds to the full model and the red curve corresponds to the EMR model.



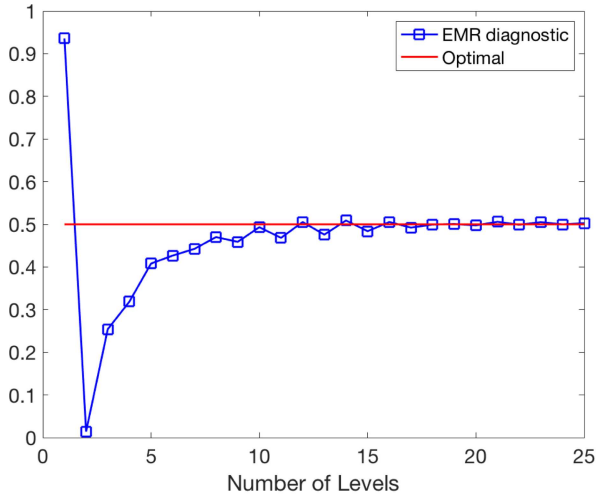
(a) X-ACF (b) Y-ACF (c) Z-ACF

FIG. 14. ACFs of the L84–L63 variables (a) X, (b) Y, and (c) Z for $h = 0.025$. The blue curve corresponds to the full model and the red curve corresponds to the EMR model.

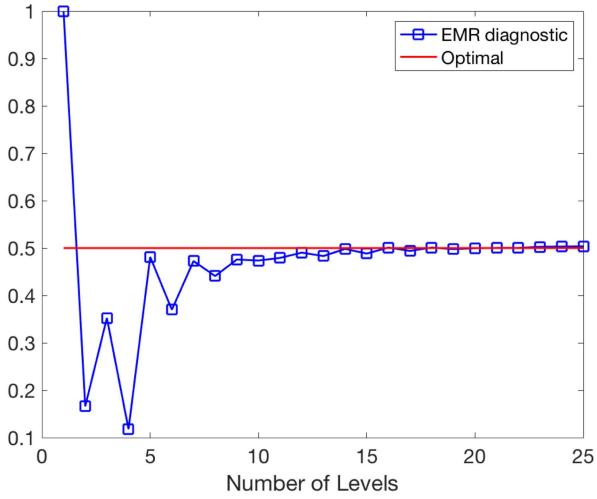


(a) $h = 0.25$ (b) $h = 0.025$

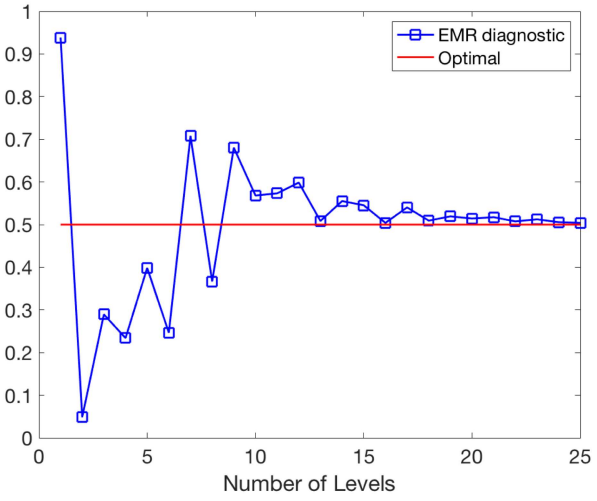
FIG. 15. Leading eigenvalues of the discretized Koopman operator in the L84 model's phase space. The blue open circles correspond to the data obtained by integrating the full model's Eq. (D1) and the red $\times$ symbols correspond to the EMR model.



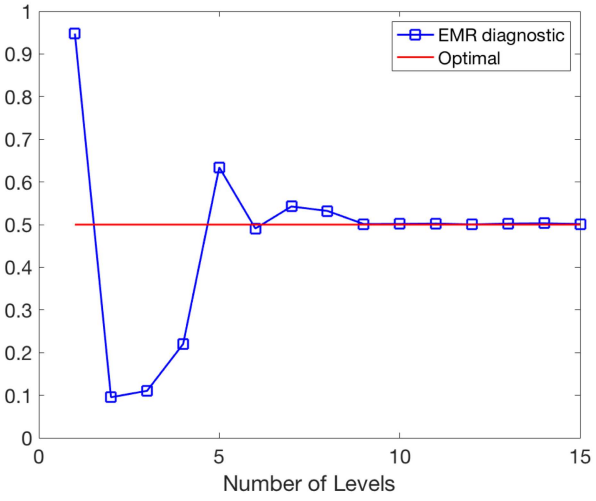
(a) $h = 0.25$ and $\tau = 5$



(b) $h = 0.025$ and $\tau = 5$



(c) $h = 0.25$ and $\tau = 2$



(d) Noisy L63 model with $h = 0.25$ and $\tau = 5$

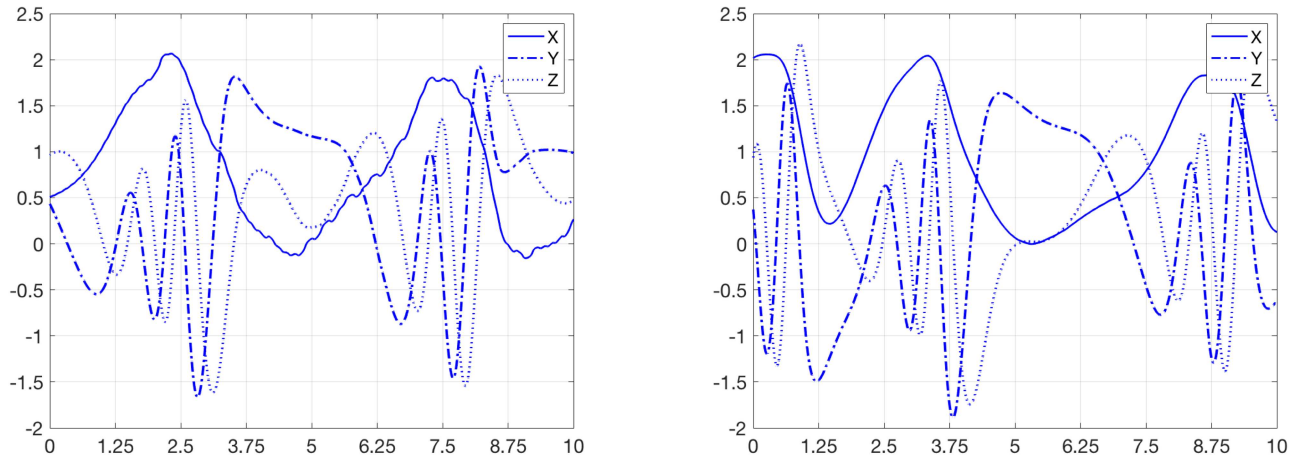
FIG. 16. Determination coefficients R^2 of the EMR method as a function of the number ℓ of levels. (a) $h = 0.25$; (b) $h = 0.025$; and (d) $h = 0.25$ but with the L63 model including additive noise. Panels (a,b,d) all have the timescale separation $\tau = 5$, while in panel (c) $h = 0.25$ and $\tau = 2$.

feed back into L63, the evolution of $\mathbf{x}(t)$ only obeys the dynamics of L63, and thus the decorrelation of the noise scales with τ.

In most of the numerical experiments, the timescale separation between the two systems is $\tau = 5$. The relevance of this timescale parameter was investigated in the previous work.[82] Here, we focus on the effects of the coupling strength h, and we shall study the cases of $h = 0.25$ and 0.025. Partial observations only will be used in these experiments, by sampling the three-dimensional outputs of the L84 system. Then, the observed tendencies are regressed and sequentially layered following the EMR approach, as explained in Sec. III.

1. EMR outputs

The L84–L63 model is integrated for 730 time units that correspond in L84 to 10 natural years, with a time step of 5×10^{-3} time units; two separate runs are made for the coupling strengths $h = 0.25$ and $h = 0.025$. These two full-model runs are used to train the corresponding EMR model versions, both of which use only the slow $\mathbf{X}$ variables and eliminate the fast $\mathbf{x}$ variables. Then, two separate full-model simulations are run, for testing purposes, over 7–300 time units, and the EMR model's output is compared with it, for the two parameter values. Below we show the main statistical outputs of the EMR methodology compared to the two reference integrations



(a) $h = 0.25$ (b) $h = 0.025$

FIG. 17. Time-series of the L84 variables (X, Y, Z) over 10 time units: (a) $h = 0.25$ and (b) 0.025.

TABLE V. Means of the EMR coefficients of the L84–L63 model, estimated from an ensemble of 50 runs over 10 time units for $h = 0.25$.

EMR	1	x	y	z	x^2	xy	y^2	xz	yz	z^2
f_X	1.949	−0.352	−0.002	−0.104	−0.015	0.01	0.052	−0.946	−0.001	−0.921
f_Y	0.999	0.001	−1.001	0.002	0	1.001	−4.003	0	0	0
f_Z	0.002	−0.003	−0.002	−1.001	0.001	4.003	1.001	0	0	0

TABLE VI. Standard deviations of the EMR coefficients of the L84–L63 model, estimated from an ensemble of 50 runs over 10 time units for $h = 0.25$.

EMR	1	x	y	z	x^2	xy	y^2	xz	yz	z^2
f_X	0.543	1.005	0.246	0.291	0.428	0.152	0.188	0.104	0.076	0.099
f_Y	0.001	0.002	0	0.001	0.001	0	0	0	0	0
f_Z	0.001	0.002	0.001	0.001	0.001	0.001	0	0	0	0

TABLE VII. Means of the EMR coefficients of the L84–L63 model, estimated from an ensemble of 50 runs over 10 time units for $h = 0.025$.

EMR	1	x	y	z	x^2	xy	y^2	xz	yz	z^2
f_X	2.006	−0.253	−0.001	−0.003	−0.004	0	0.001	−1.002	0.001	−1.003
f_Y	1	0	−1.001	0.002	0	1.001	−4.003	0	0	0
f_Z	0.001	−0.002	−0.002	−1.001	0.001	4.003	1.001	0	0	0

TABLE VIII. Standard deviations of the EMR coefficients of the L84–L63 model, estimated from an ensemble of 50 runs over 10 time units for $h = 0.025$.

EMR	1	x	y	z	x^2	xy	y^2	xz	yz	z^2
f_X	0.034	0.063	0.019	0.021	0.026	0.01	0.014	0.007	0.008	0.006
f_Y	0.001	0.002	0	0.001	0.001	0	0	0	0	0
f_Z	0.001	0.002	0.001	0.001	0.001	0.001	0	0	0	0

of the full model. The results for the two separate h-values are shown in Figs. 9–11 and 12–14, respectively.

The region of phase space explored by the EMR model clearly coincides with the one visited by the full model, as seen in Figs. 9 and 12, and the relative occupancies within this region—as indicated by the smoothed PDFs shown in Figs. 10 and 13, respectively—agree very well. The timescales are also well captured, as indicated by the good approximation of the autocorrelation functions, cf. Figs. 11 and 14.

Notice that, while the original L84–L63 system is purely deterministic, the EMR model includes white noise acting on the hidden layers of the learned model. This fact could suggest that a smoothing of the invariant measure is inevitable and that the EMR methodology may not be able to capture fractal geometries in phase space, since the EMR model does not satisfy the Hörmander's hypoellipticity condition.[18,40] The numerical evidence in Figs. 9 and 12, however, illustrates a strikingly good approximation of the full model's attractor, including its very fine, and presumably fractal, structure.

Actually, Theorem 3.1 and Corollary 3.2 in Ref. 43 provided sufficient conditions for the existence of a random attractor for a broad class of MSMs that are not subject to a non-degeneracy condition of Hörmander type. In other words, one can have an MSM that possesses a random attractor and is thus dynamically quite stable, while exhibiting in a forward sense an invariant measure of the associated Fokker–Planck equation that is singular with respect to the Lebesgue measure. This mathematically rigorous result helps explain what is observed numerically not only in the present paper for the EMR of the L84–L63 model, but also in the case of the EMR model of the Lotka–Volterra example in Ref. 43, Fig. 7.

Ulam's method was used on the projection of the full $(\mathbf{X}, \mathbf{x})$ phase space onto the $\mathbf{X}$ subspace to approximate the spectrum of the Koopman operator, since it can provide further information on the characteristics of the time series, beyond PDFs and correlation functions. The observed spectra (red $\times$ symbols) using a coarse partition of phase space into 512 nonintersecting boxes showed good agreement with the spectra based on the full model (blue open circles); see Fig. 15. This agreement confirms further that, at this level of coarse graining, the EMR model captures well the characteristics of the full model's solutions.

2. Convergence

Convergence in the EMR approach is determined by the "whiteness" of the last-level residual, as explained in Sec. III B; see Eq. (3.16) and discussion thereof. In Fig. 16, we plotted the mean of the determination coefficients R^2 for the three $\mathbf{X}$ variables and we show that its convergence in the EMR approach depends only mildly on the coupling parameter h. Indeed, for $h = 0.25$ we observe in panel (a) that around 18 levels are necessary before achieving the optimal level, whereas for weaker coupling with $h = 0.025$ convergence is attained in panel (b) already with 15 levels, as one might expect.

Furthermore, as already pointed out in Ref. 43, Sec. 7, on a different example, the results in Fig. 16(c) illustrate that a smaller timescale separation τ can require a higher number of levels for EMR to attain convergence: in the case at hand, around 25 levels are needed. For completeness, Fig. 16(d) shows that including additive white noise in the L63 system can, in fact, accelerate the convergence of the method, with convergence achieved at $\ell = 7$.

3. Model coefficients

We show here that the EMR model coefficients can be efficiently approximated when phase space subsampling is carried out. Here, regressions are performed over 50 short time series of 10 time units each, with a time step of 5×10^{-3}, as in Section 1 of Appendix D. The reason for taking this sample length here is that 10 time units is visually enough for the slow variable $\mathbf{X}$ to go through a cycle, as illustrated in Fig. 17, for both $h = 0.25$ and 0.025.

The estimated coefficients and their standard deviations using the EMR regressions are listed in Tables V and VI for $h = 0.25$ and in Tables VII and VIII for $h = 0.025$. The tables show the coefficients of the linear and quadratic forms at the first level in the EMR regressions: see Eq. (3.14a).

As expected, a stronger coupling of $h = 0.25$ leads to greater uncertainty in the estimation, as indicated by the corresponding standard error. For the fairly complex and chaotic system at hand, we note that no memory effects are artificially introduced in the regressions at the second level. Indeed, we found that the coupling of the main level with the subsequent ones was 0 to the fourth decimal place.

DATA AVAILABILITY

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

REFERENCES

[1] S. M. Alessio, *Digital Signal Processing and Spectral Analysis for Scientists: Concepts and Applications* (Springer Science & Business Media, 2015).
[2] A. Arakawa and W. H. Schubert, "Interaction of a cumulus cloud ensemble with the large-scale environment, part I," J. Atmos. Sci. **31**(3), 674–701 (1974).
[3] V. Baladi, *Positive Transfer Operators and Decay of Correlations* (World Scientific, Singapore, 2000).
[4] J. Berner, U. Achatz, L. Batté, L. Bengtsson, A. D. L. Cámara, H. M. Christensen, M. Colangeli, D. R. B. Coleman, D. Crommelin, S. I. Dolaptchiev, and C. L. Franzke, "Stochastic parameterization: Toward a new view of weather and climate models," Bull. Am. Meteorol. Soc. **98**(3), 565–588 (2017).
[5] D. S. Broomhead, R. Jones, and G. P. King, "Topological dimension and local coordinates from time series data," J. Phys. A **20**(9), L563 (1987).
[6] 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).
[7] M. D. Chekroun, F. Di Plinio, N. E. Glatt-Holtz, and V. Pata, "Asymptotics of the Coleman-Gurtin model," Discrete Contin. Dyn. Syst. S **4**(2), 351–369 (2011).
[8] M. D. Chekroun, M. Ghil, H. Liu, and S. Wang, "Low-dimensional Galerkin approximations of nonlinear delay differential equations," Discrete Contin. Dyn. Syst. A **36**(8), 4133–4177 (2016).
[9] M. D. Chekroun and N. E. Glatt-Holtz, "Invariant measures for dissipative dynamical systems: Abstract results and applications," Commun. Math. Phys. **316**(3), 723–761 (2012).
[10] M. D. Chekroun and D. Kondrashov, "Data-adaptive harmonic spectra and multilayer Stuart-Landau models," Chaos **27**(9), 093110 (2017).
[11] M. D. Chekroun, I. Koren, and H. Liu, "Efficient reduction for diagnosing Hopf bifurcation in delay differential systems: Applications to cloud-rain models," Chaos **40**(8), 053130 (2020).
[12] M. D. Chekroun, H. Liu, and J. McWilliams, "The emergence of fast oscillations in a reduced primitive equation model and its implications for closure theories," Comput. Fluids **151**, 3–22 (2017).

[13]M. D. Chekroun, H. Liu, and J. C. McWilliams, "Variational approach to closure of nonlinear dynamical systems: Autonomous case," J. Stat. Phys. **179**(179), 1073–1160 (2020).
[14]M. D. Chekroun, H. Liu, and S. Wang, *Approximation of Stochastic Invariant Manifolds: Stochastic Manifolds for Nonlinear SPDEs I*, Springer Briefs in Mathematics (Springer, 2015).
[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, J. D. Neelin, D. Kondrashov, J. C. McWilliams, and M. Ghil, "Rough parameter dependence in climate models and the role of Ruelle-Pollicott resonances," Proc. Natl. Acad. Sci. U.S.A. **111**(5), 1684–1690 (2014).
[17]M. D. Chekroun, E. Simonnet, and M. Ghil, "Stochastic climate dynamics: Random attractors and time-dependent invariant measures," Physica D **240**(21), 1685–1700 (2011).
[18]M. D. Chekroun, E. Simonnet, and M. Ghil, "Stochastic climate dynamics: Random attractors and time-dependent invariant measures," Physica D **240**(21), 1685–1700 (2011).
[19]M. D. Chekroun, A. Tantet, H. A. Dijkstra, and J. D. Neelin, "Ruelle–Pollicott resonances of stochastic systems in reduced state space. Part I: Theory," J. Stat. Phys. **179**, 1366–1402 (2020).
[20]A. J. Chorin, O. H. Hald, and R. Kupferman, "Optimal prediction with memory," Physica D **166**((3–4)), 239–257 (2002).
[21]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).
[22]C. M. Dafermos, "Asymptotic stability in viscoelasticity," Arch. Ration. Mech. Anal. **37**(4), 297–308 (1970).
[23]F. J. Dyson, "The radiation theories of Tomonaga, Schwinger, and Feynman," Phys. Rev. **75**(3), 486–502 (1949).
[24]W. E and J. Lu, "Multiscale modeling," Scholarpedia **6**(10), 11527 (2011), revision #91540.
[25]K.-J. Engel and R. Nagel, *A Short Course on Operator Semigroups* (Springer Science & Business Media, 2006).
[26]D. Faranda, M. Vrac, P. Yiou, F. Pons, A. Hamid, G. Carella, C. Langue, S. Thao, and V. Gautard, "Boosting performance in machine learning of turbulent and geophysical flows via scale separation," arXiv:hal-02337839v2 (2019).
[27]C. Franzke, A. J. Majda, and G. Branstator, "The origin of nonlinear signatures of planetary wave dynamics: Mean phase space tendencies and contributions from non-Gaussianity," J. Atmos. Sci. **64**(11), 3987–4003 (2007).
[28]C. L. E. Franzke, T. J. O'Kane, J. Berner, P. D. Williams, and V. Lucarini, "Stochastic climate theory and modeling," Wiley Interdiscip. Rev.: Clim. Change **6**(1), 63–78 (2015).
[29]P. Gentine, M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis, "Could machine learning break the convection parameterization deadlock?," Geophys. Res. Lett. **45**(11), 5742–5751, https://doi.org/10.1029/2018GL078202 (2018).
[30]M. Ghil, "A mathematical theory of climate sensitivity or, How to deal with both anthropogenic forcing and natural variability?," in *Climate Change: Multidecadal and Beyond*, edited by C. P. Chang, M. Ghil, M. Latif, and J. M. Wallace (World Scientific Publishing Co., 2015), pp. 31–51.
[31]M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou, "Advanced spectral methods for climatic time series," Rev. Geophys. **40**(1), 1–41, https://doi.org/10.1029/2000RG000092 (2002).
[32]M. Ghil and V. Lucarini, "The physics of climate variability and climate change," Rev. Mod. Phys. **92**, 035002 (2020).
[33]T. L. Gill, "The Feynman-Dyson view," J. Phys.: Conf. Ser. **845**, 012023 (2017).
[34]G. A. Gottwald and I. Melbourne, "Homogenization for deterministic maps and multiplicative noise," Proc. R. Soc. A **469**(2156), 20130201 (2013).
[35]I. Grooms and A. J. Majda, "Efficient stochastic superparameterization for geophysical turbulence," Proc. Natl. Acad. Sci. U.S.A. **110**(12), 4464–4469 (2013).
[36]M. Hairer and A. J. Majda, "A simple framework to justify linear response theory," Nonlinearity **23**(4), 909–922 (2010).
[37]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).
[38]J. R. Holton and G. J. Hakim, *An Introduction to Dynamic Meteorology*, 4th ed. (Academic Press, San Diego, CA, 2013).
[39]H. Hosni and A. Vulpiani, "Data science and the art of modelling," Lett. Mat. **6**(2), 121–129 (2018).
[40]L. Hörmander, "Hypoelliptic second order differential equations," Acta Math. **119**(0), 147–171 (1967).
[41]C. L. Keppenne and M. Ghil, "Adaptive filtering and prediction of the Southern Oscillation index," J. Geophys. Res.: Atmos. **97**(D18), 20449–20454, https://doi.org/10.1029/92JD02219 (1992).
[42]R. Klein, "Scale-dependent models for atmospheric flows," Annu. Rev. Fluid Mech. **42**(1), 249–274 (2010).
[43]D. Kondrashov, M. D. Chekroun, and M. Ghil, "Data-driven non-Markovian closure models," Physica D **297**, 33–55 (2015).
[44]D. Kondrashov, M. D. Chekroun, and M. Ghil, "Data-adaptive harmonic decomposition and prediction of Arctic sea ice extent," Dyn. Stat. Clim. Syst. **3**(1), 23 (2018).
[45]D. Kondrashov, S. Kravtsov, and M. Ghil, "Empirical mode reduction in a model of extratropical low-frequency variability," J. Atmos. Sci. **63**(7), 1859–1877 (2006).
[46]D. Kondrashov, S. Kravtsov, A. W. Robertson, and M. Ghil, "A hierarchy of data-based ENSO models," J. Clim. **18**(21), 4425–4444 (2005).
[47]D. Kondrashov, M. D. Chekroun, X. Yuan, and M. Ghil, "Data-adaptive harmonic decomposition and stochastic modeling of Arctic sea ice," in *Advances in Nonlinear Geosciences*, edited by A. A. Tsonis (Springer International Publishing, 2018).
[48]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).
[49]S. Kravtsov, D. Kondrashov, and M. Ghil, "Empirical model reduction and the modeling hierarchy in climate dynamics and the geosciences," in *Stochastic Physics and Climate Modelling*, edited by T. N. Palmer and P. Williams (Cambridge University Press, Cambridge, 2010), pp. 35–72.
[50]A. Lasota and M. C. Mackey, *Chaos, Fractals and Noise* (Springer, New York, 1994).
[51]K. K. Lin and F. Lu, "Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism," J. Comput. Phys. **424**, 109864 (2021).
[52]Y. T. Lin, Y. Tian, D. Livescu, and M. Anghel, "Data-driven learning for the Mori–Zwanzig formalism: A generalization of the Koopman learning framework," arXiv:2101.05873 (2021).
[53]E. N. Lorenz, "Deterministic nonperiodic flow," J. Atmos. Sci. **20**, 130–141 (1963).
[54]E. N. Lorenz, "Irregularity: A fundamental property of the atmosphere," Tellus A **36A**(2), 98–110 (1984).
[55]V. Lucarini, "Revising and extending the linear response theory for statistical mechanical systems: Evaluating observables as predictors and predictands," J. Stat. Phys. **173**(6), 1698–1721 (2018).
[56]V. Lucarini and J. Wouters, "Response formulae for n-point correlations in statistical mechanical systems and application to a problem of coarse graining," J. Phys. A: Math. Theor. **50**(35), 355003 (2017).
[57]A. J. Majda, I. Timofeyev, and E. Vanden Eijnden, "A mathematical framework for stochastic climate models," Commun. Pure Appl. Math. **54**(8), 891–974 (2001).
[58]I. Melbourne and A. M. Stuart, "A note on diffusion limits of chaotic skew-product flows," Nonlinearity **24**(4), 1361–1367 (2011).
[59]I. Mezić, "Spectral properties of dynamical systems, model reduction and decompositions," Nonlinear Dyn. **41**(1), 309–325 (2005).
[60]H. Mori, "Transport, collective motion, and Brownian motion," Prog. Theor. Phys. **33**(3), 423–455 (1965).
[61]P. A. O'Gorman and J. G. Dwyer, "Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events," J. Adv. Model. Earth Syst. **10**(10), 2548–2563 (2018).
[62]*Stochastic Physics and Climate Modelling*, edited by T. N. Palmer and P. Williams (Cambridge University Press, Cambridge, 2009).
[63]G. A. Pavliotis, *Stochastic Processes and Applications* (Springer, New York, 2014), Vol. 60.

[64]G. A. Pavliotis and A. M. Stuart, *Multiscale Methods* (Springer, New York, 2008).
[65]A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations* (Springer Science & Business Media, 2012), Vol. 44.
[66]J. P. Peixoto and A. H. Oort, *Physics of Climate* (AIP Press, New York, 1992).
[67]C. Penland, "A stochastic model of IndoPacific sea surface temperature anomalies," Physica D **98**(2–4), 534–558 (1996).
[68]S. Rasp, M. S. Pritchard, and P. Gentine, "Deep learning to represent subgrid processes in climate models," Proc. Natl. Acad. Sci. U.S.A. **115**(39), 9684–9689 (2018).
[69]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).
[70]D. Ruelle, "Nonequilibrium statistical mechanics near equilibrium: Computing higher-order terms," Nonlinearity **11**(1), 5–18 (1998).
[71]D. Ruelle, "A review of linear response theory for general differentiable dynamical systems," Nonlinearity **22**(4), 855–870 (2009).
[72]P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," J. Fluid Mech. **656**, 5–28 (2010).
[73]T. Schneider, S. Lan, A. Stuart, and J. Teixeira, "Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations," Geophys. Res. Lett. **44**(24), 12, 396–12, 417, https://doi.org/10.1002/2016GL071741 (2017).
[74]G. R. Sell, *Topological Dynamics and Ordinary Differential Equations* (Van Nostrand Reinhold, 1971).
[75]A. Tantet, V. Lucarini, and H. A. Dijkstra, "Resonances in a chaotic attractor crisis of the Lorenz flow," J. Stat. Phys. **170**(3), 584–616 (2018).
[76]A. Tantet, F. R. van der Burgt, and H. A. Dijkstra, "An early warning indicator for atmospheric blocking events using transfer operators," Chaos **25**(3), 036406 (2015).
[77]L. N. Trefethen and M. Embree, *Spectra and Pseudospectra* (Princeton University Press, 2005).
[78]J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, "On dynamic mode decomposition: Theory and applications," J. Comput. Dyn. **1**(2), 391–421 (2014).
[79]G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the Brownian motion," Phys. Rev. **36**, 823–841 (1930).
[80]G. K. Vallis, *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-scale Circulation* (Cambridge University Press, Cambridge, 2006).
[81]G. Vissio, V. Lembo, V. Lucarini, and M. Ghil, "Evaluating the performance of climate models based on Wasserstein distance," Geophys. Res. Lett. **47**, e2020GL089385, https://doi.org/10.1029/2020GL089385 (2020).
[82]G. Vissio and V. Lucarini, "Evaluating a stochastic parametrization for a fast-slow system using the Wasserstein distance," Nonlinear Process. Geophys. **25**(2), 413–427 (2018).
[83]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).
[84]J. A. Weyn, D. R. Durran, and R. Caruana, "Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere," J. Adv. Model. Earth Syst. **12**(9), e2020MS002109 (2020).
[85]D. S. Wilks, "Effects of stochastic parametrizations in the Lorenz '96 system," Q. J. R. Meteorol. Soc. **131**(606), 389–407 (2005).
[86]J. Wouters and G. A. Gottwald, "Edgeworth expansions for slow-fast systems with finite time-scale separation," Proc. R. Soc. A **475**(2223), 20180358 (2019).
[87]J. Wouters, S. Iankov Dolaptchiev, V. Lucarini, and U. Achatz, "Parameterization of stochastic multiscale triads," Nonlinear Process. Geophys. **23**(6), 435–445 (2016).
[88]J. Wouters and V. Lucarini, "Disentangling multi-level systems: Averaging, correlations and memory," J. Stat. Mech.: Theor. Exp. **2012**(3), P03003 (2012).
[89]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).
[90]R. Zwanzig, "Memory effects in irreversible thermodynamics," Phys. Rev. **124**(4), 983–992 (1961).
[91]R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, 2001).
[92]A. Gupta and P. F. J. Lermusiaux, "Neural closure models for dynamical systems," arXiv:2012.13869 (2020).
[93]K. Hasselmann, "Stochastic climate models. Part I. Theory," Tellus **28**, 473–485 (1976).