Title
Data-driven non-Markovian closure models

Authors
Kondrashov, D
Chekroun, MD
Ghil, M

Publication Date
2015-03-15

DOI
10.1016/j.physd.2014.12.005

Peer reviewed

DATA-DRIVEN NON-MARKOVIAN CLOSURE MODELS

DMITRI KONDRASHOV, MICKAËL D. CHEKROUN, AND MICHAEL GHIL

Abstract. This paper has two interrelated foci: (i) obtaining stable and efficient data-driven closure models by using a multivariate time series of partial observations from a large-dimensional system; and (ii) comparing these closure models with the optimal closures predicted by the Mori-Zwanzig (MZ) formalism of statistical physics. Multilayer stochastic models (MSMs) are introduced as both a generalization and a time-continuous limit of existing multilevel, regression-based approaches to closure in a data-driven setting; these approaches include empirical model reduction (EMR), as well as more recent multi-layer modeling. It is shown that the multilayer structure of MSMs can provide a natural Markov approximation to the generalized Langevin equation (GLE) of the MZ formalism.

A simple correlation-based stopping criterion for an EMR-MSM model is derived to assess how well it approximates the GLE solution. Sufficient conditions are derived on the structure of the nonlinear cross-interactions between the constitutive layers of a given MSM to guarantee the existence of a global random attractor. This existence ensures that no blow-up can occur for a broad class of MSM applications, a class that includes non-polynomial predictors and nonlinearities that do not necessarily preserve quadratic energy invariants.

The EMR-MSM methodology is applied to a conceptual, nonlinear, stochastic climate model of coupled slow and fast variables, in which only slow variables are observed. It is shown that the resulting closure model with energy-conserving nonlinearities efficiently captures the main statistical features of the slow variables, even when there is no formal scale separation and the fast variables are quite energetic.

Second, an MSM is shown to successfully reproduce the statistics of a partially observed, Lokta-Volterra model of population dynamics in its chaotic regime. The challenges here include the rarity of strange attractors in the model's parameter space and the existence of multiple attractor basins with fractal boundaries. The positivity constraint on the solutions' components replaces here the quadratic-energy–preserving constraint of fluid-flow problems and it successfully prevents blow-up.

1. Introduction and Motivation

1.1. **Background.** Comprehensive dynamical climate models aim at simulating past, present and future climate and, more recently, at predicting it. These models, commonly known as general circulation models or global climate models (abbreviated as GCMs in either case) represent a broad range of time and space scales and use a state vector that is typically constituted by several millions of variables.

While detailed weather prediction out to a few days does require such high numerical resolution, climate variability on longer time scales is dominated by large-scale patterns, which may require only a few appropriately selected modes for their simulation and prediction [1]. For a specific range of frequencies and targeted variables, one may try to formulate low-order models (LOMs) for these purposes. Such models must account in an accurate way (i) for linear and nonlinear self-interactions between a judiciously selected set of resolved high-variance climate components; and (ii) for the cross-interactions between the resolved components and the large

Key words and phrases. Empirical model reduction, inverse modeling, least-mean-square minimization, low-order models, memory effects, nonlinear stochastic dynamics, stochastic closure model.

number of unresolved ones. Although the present article is motivated primarily by the need for LOMs in climate modeling, similar issues arise in many other areas of the physical and life sciences, and LOMs are becoming a key tool in various disciplines as diverse as astrophysics [2], biological neuronal modeling [3], molecular dynamics [4, 5] or pharmacokinetic and pharmacodynamic modeling [6].

In climate dynamics, the prediction of the El Niño-Southern Oscillation (ENSO) has attracted increased attention during the past decades, since ENSO constitutes the dominant mode of interannual climate variability, with major global impacts on temperature, precipitation, tropical cyclones, and human health [7, 8, 9]; it has even been argued recently to have potential impacts on civil conflicts [10]. The earliest successful predictions of ENSO were made using a dynamical model governed by a set of coupled partial differential equations (PDEs) [11] that was itself a highly reduced model by today's standards. Subsequently, stochastically driven linear LOMs — based either on observational data [12, 13, 14] or on dynamical model simulations [15] — have been used for ENSO predictability studies as well as for real-time forecasting [16]. Nowadays, modeling ENSO by such LOMs can be considered as a significant success story, although predicting its extremes is still a challenge; a recent survey on real-time prediction skill of state-of-the art statistical ENSO models compared to comprehensive dynamical climate models is given in [17].

Real-time ENSO predictions based on the *Empirical Model Reduction (EMR)* method introduced in [18, 19] have proven to be highly competitive.[1] Within climate dynamics, the EMR methodology has been successfully applied to the modeling of many multivariate time series on different time scales, whether arising in observed air–sea interactions in the Southern Ocean [20], in the identification and predictability analysis of nonlinear circulation regimes in atmospheric models of intermediate complexity [21, 22, 23], in the modeling of the Madden-Julian Oscillation (MJO) [24] or in the stochastic parameterization of subgrid-scale mid-latitude processes [25].

In these successfully solved climate problems, the key ingredient to modeling and predicting the dynamics of the macroscopic, observed variables from partial and incomplete observations of large systems is the appropriate use of some pre-specified self- and cross-interactions between the macroscopic variables, supplemented by some auxiliary, hidden variables. In an EMR model, these interactions are typically chosen to be quadratic or linear and the associated hidden variables are arranged into a "matrioshka" of layers. Each supplementary layer in this "matrioshka" includes a hidden variable that is less auto-correlated than the one introduced in the previous layer, until some decorrelation criterion is reached; see A. In practice, the unknown coefficients at the main level and the additional ones are learned by means of multilevel regression techniques; we refer to [26] for a review of the EMR methodology and a comparison with other model-reduction methodologies.

Quite recently, a couple of papers [27, 28, 29] pointed out the possibility of undesirable behavior in EMR models, and proposed to add energy-preserving constraints on the quadratic terms in order to prevent such behavior. r. The idea of adding constraints to prevent blow-up in an EMR formulation should clearly not be limited to the class of models that, in the absence of dissipation, possess quadratic energy invariants: many models in climate dynamics do not possess such invariants, e.g. models of the sea surface temperature (SST) field have energy terms

[1] Barnston and colleagues [17] analyzed two dozen ENSO multi-model real-time predictions coordinated by Columbia University's International Research Institute for Climate and Society (IRI) over the 2002–2011 interval and concluded that the "UCLA-TCD prediction (ensemble mean) has the highest seasonally combined correlation skill among the statistical models exceeded by only a few dynamical models [...] as well as one of the smallest RMSE." Note that the former is based on models with a few tens of variables, while the latter have many millions of variables.

that are linear in temperature, rather than quadratic. On the other hand, in certain situations that occur, for instance, in population dynamics [30, 31, 32, 33, 34, 35] or chemical kinetics [36, 37, 38, 39, 40], while the nonlinearities might still be quadratic, introducing such constraints might actually be counter-productive; see Section 7. The work of [28, 29] proposed multi-level regression (MLR) models that did allow for quadratic interactions between the observed and some of the hidden variables and showed that these interactions — given quadratic energy invariants in the flow models to which they were applied — result in a stable, well-behaved reduction of the full flow models. As their very name indicates, though, the fundamental feature of the MLR models is still the multilevel structure proposed a decade ago in the original EMR formulation.

The associated hidden variables in the original EMR formulation depend, due to this multilevel structure, on the past of the observed variables and bring therefore memory effects into the resulting low-dimensional stochastic models, in a fashion that is reminiscent of the closure models in the Mori-Zwanzig (MZ) formalism of statistical mechanics [41] or of related optimal prediction methods [42, 43]. The latter methods also deal with the problem of predicting just a few relevant variables but from a different perspective than the EMR one, i.e. when the equations of the original, full system are available. The connection between the EMR formulation and MZ-type formalisms was first pointed out and illustrated by a simple example in the supporting information of [44].

1.2. **Outline.** The background of this paper is thus provided by (a) the success of the EMR methodology in the modeling and real-time prediction of spatio-temporal climate fields; (b) the recent criticisms in [27, 28, 29] of potential vulnerabilities in the original version of this methodology; and (c) its relationships with the closure methods suggested by the MZ formalism. The purpose of the paper is, therefore, (i) to generalize further both the original EMRs and the MLR models in [27, 28, 29]; (ii) to provide a mathematical analysis of data-based EMR models in their continuous-time limit and of the generalizations thereof; and (iii) to illustrate the insights and additional tools thus obtained by two simple applications.

We call the generalized and rigorously studied continuous-time limit of EMRs multilayer stochastic models (MSMs). In Sec. 2, we formulate the closure problem in the presence of partial observations and consider EMR models as a candidate solution to this problem. In Sec. 3, we introduce MSMs and show that an MSM can be written as a system of stochastic integro-differential equations (see Proposition 3.3). This system can lead in practice to a good approximation of the generalized Langevin equation (GLE) of the MZ formalism, denoted here by (*GLE*) and studied in Sec. 4. In this section, it is shown that the closure obtained by the GLE is theoretically optimal, given a time series of data, rather than a known master equation. Lemma 4.1 supports this statement, subject to the appropriate ergodic assumptions and assuming an infinitely long multivariate time series of partial observations. The difference between the standard way of building the GLE and an approach based on averaging along trajectories, as in Lemma 4.1, is similar to the Eulerian versus the Lagrangian viewpoint in fluid mechanics.

Sections 3 and 4 are fairly theoretical and the hasty reader who might be more motivated by the applications can skip these two sections at a first reading and return to them later, after seeing the usefulness of their theoretical results in Secs. 5–7. In Sec. 5, practical issues of applying the results of Secs. 2–4 to deriving accurate and stable EMRs are considered. In particular, we will see that, under certain circumstances, an MSM can be understood as a Lagrangian approximation of the GLE; see also Proposition 3.3. Numerical results for a conceptual stochastic climate model proposed in [45] are presented in Sec. 6. These highly satisfactory results

demonstrate, among other things, that the η-test formulated in Sec. 5 for the last-level residue of an MSM does provide quite an efficient criterion for the degree of approximation of the GLE solution by the appropriate MSM.

In Sec. 3 we also derive conditions on the cross-interactions between the constitutive layers of a given MSM that guarantee the existence of a global random attractor. This existence ensures that no blow-up can occur for a broad class of MSMs that generalize the class of EMR-like models used so far, including but not restricted to the MLR models of [27, 28, 29]; see Theorem 3.1. This class includes non-polynomial predictors and nonlinearities that do not necessarily preserve quadratic energy invariants, such as assumed in [28, 29]; see Corollary 3.2. The latter results are illustrated in Sec. 7 by solving a closure problem arising in population dynamics that possesses merely linear and quadratic terms, but requires a very different set of constraints to prevent blow-up of the reduced model.

Finally, four appendices provide further details on EMR stopping criteria, on real-time prediction using an MSM, on practical aspects of energy conservation, and on the interpretation of MSM coefficients.

1.3. **Multilayer stochastic models (MSMs) and integro-differential equations.** In this subsection, we take a detour into the deterministic literature of integro-differential equations that will shed some further light on the ability of an MSM to provide an efficient closure model based on partial information on the full model, as derived from a time series. The parallels drawn herein between the two situations yield a broader perspective on the role of an MSM's multilayer structure with respect to its representation as a system of stochastic integro-differential equations, cf. Proposition 3.3 below.

The present remarks demonstrate the underlying relationships between multilayer systems of ordinary differential equations (ODEs) and systems of integro-differential equations, and help one understand why the multilayer structure of an MSM is essential in constructing a class of stochastic differential equations (SDEs) susceptible to approximate a GLE. These observations are actually rooted in older mathematical ideas from the study of models that involve distributed delays [46, 47]; such models arise in theoretical population dynamics and in the modeling of materials with memory [48, 49, 50], as well as in climate dynamics [51, 52].

Motivated by these remarks, we consider now the following system of integro-differential equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i\Big(b_i + \sum_{j=1}^{n} a_{ij}x_j + \sum_{j=1}^{n} \gamma_{ij}\int_{-\infty}^{t} g_{ij}(t-s)x_j(s)\mathrm{d}s\Big),\ i = 1,\ldots,n; \tag{1.1}$$

this system models the population dynamics of a community of n interacting species, where x_i denotes the population density of the i-th species, $b = (b_1, b_2, ..., b_n)$ the vector of intrinsic population growth rates, $A = (a_{ij})$ and $\Gamma = (\gamma_{ij})$ denote the interaction matrices, and g_{ij} the memory kernels that describe the present response of the *per capita* growth rate of a species i to historical population densities x_j. Volterra proposed such a system of integro-differential equations to describe an ecological system of interacting species and investigated it for $n = 2$ [53, Chap. IV].

We wish to show how system (1.1) can be recast into a system of ODEs, and assume for simplicity at first that (1.1) takes the form,

$$\begin{aligned} \frac{\mathrm{d}x_i}{\mathrm{d}t} &= x_i\Big(b_i + \sum_{j=1}^{n} a_{ij}x_j\Big),\ 1 \le i \ne p \le n;\ \text{and} \\ \frac{\mathrm{d}x_p}{\mathrm{d}t} &= x_p\Big(b_p + \sum_{j=1}^{n} a_{pj}x_j + \gamma_{pm}\int_{-\infty}^{t} g_{pm}(t-s)x_m(s)\mathrm{d}s\Big), \end{aligned} \tag{1.2}$$

for some p and m in $\{1, \cdots, n\}$, i.e., that only a single equation exhibits memory effects. Furthermore, the memory kernel g_{pm} is assumed to be given by the Gamma distribution

$$g_{pm}(t) = F_k(t) = \frac{\alpha^k}{(k-1)!}t^{k-1}e^{-\alpha t}, \tag{1.3}$$

for some $\alpha > 0$ and some positive integer $k \ge 1$.

The key step is to note the recursion relation

$$\frac{\mathrm{d}F_k}{\mathrm{d}t} = \alpha F_{k-1} - \alpha F_k \tag{1.4}$$

and to introduce the additional k new variables r_j, with

$$r_j(t) = \int_{-\infty}^{t} x_m(s)F_j(t-s)\mathrm{d}s,\ j = 1, \ldots, k. \tag{1.5}$$

By differentiation we obtain that these auxiliary variables obey the following system of ODEs,

$$\begin{aligned} \frac{\mathrm{d}r_1}{\mathrm{d}t} &= \alpha(x_m - r_1), \\ \frac{\mathrm{d}r_j}{\mathrm{d}t} &= \alpha(r_{j-1} - r_j),\ j = 2, \ldots, k. \end{aligned} \tag{1.6}$$

This system is driven by x_m. More precisely, the dynamics of the auxiliary variable r_1 is directly slaved to that of x_m, while the other r_j-variables are indirectly slaved to x_m, since each variable r_j in a given layer $2 \le j \le k$ interacts with r_{j-1} in the previous layer, thus sharing a multilayer structure reminiscent of the one in the original EMR formulation [18, 19].

These remarks allow us to recast the system of integro-differential equations (1.2) as the following system of ODEs:

$$\begin{aligned} \frac{\mathrm{d}x_i}{\mathrm{d}t} &= x_i\Big(b_i + \sum_{j=1}^{n} a_{ij}x_j\Big),\ 1 \le i \ne p \le n, \\ \frac{\mathrm{d}x_p}{\mathrm{d}t} &= x_p\Big(b_p + \sum_{j=1}^{n} a_{pj}x_j + \gamma_{pm}r_k\Big), \\ \frac{\mathrm{d}r_j}{\mathrm{d}t} &= \alpha(r_{j-1} - r_j),\ j = 2, \ldots, k, \\ \frac{\mathrm{d}r_1}{\mathrm{d}t} &= \alpha(x_m - r_1). \end{aligned} \tag{1.7}$$

The expansion procedure outlined above for the single memory effect in the simplified system of Eq. (1.2) can obviously be carried out for any number of memory terms in any number of equations, subject to the addition of a suitable number of linear equations; hence it can also be applied to the general system of integro-differential equations in Eq. (1.1).

One concludes that, for such a system integro-differential equations — if the kernels are weighted sums of Gamma distributions, or more generally, if these kernels are solutions to a linear system of ODEs with constant coefficients — then the original system can be transformed into a system of ODEs. This transformation is known as the "linear-chain trick" [46, 54, 55]. Of course, it is important to be able to go in the other direction as well. If one finds an interesting solution of the ODE system (1.7), e.g. a periodic solution, then one wants to know if it does solve (1.2) as well. In fact, it is not difficult to prove that any solution of (1.7) that is bounded on the entire real line is also a solution of the integro-differential equation (1.2); see [55, Prop. 7.3]. Since the resulting system of ODEs (1.7) does not involve the knowledge of the past of the x_i-variables, one can say that a "Markovianization" of the original system has been performed by suitably augmenting the number of variables; this augmentation procedure is actually well-known in the rigorous study of systems with distributed delays, such as those that arise in the modeling of materials with memory, for instance, cf. [48, 49, 50] and references therein.

This detour via a class of systems of integro-differential equations provides some general guidance on how to "Markovianize" a broad class of GLEs of the type predicted by MZ closure procedures. Such GLEs take necessarily the form of systems of stochastic integro-differential equations. It is natural, then, to seek approximations to such MZ closures in the form of an augmented system of SDEs whose main, observed variables are supplemented by appropriate auxiliary hidden variables, and one expects the latter variables to interact with the main ones and among themselves in a fashion suggested by the ODE system (1.7).

Of course, the corresponding interactions have to take a specific form, depending on the applications. For dissipative systems, it is shown below that — given a natural energy that has to be dissipated — simple estimates allow one to identify permissible interactions that ensure the existence of dissipative MSMs; see Theorem 3.1 and Corollary 3.2. Within this class of interactions, Proposition 3.3 ensures that a simple correlation-based criterion formulated in Section 5 does address the problem of approximating the GLE by such MSMs.

The approach proposed in this article complements, therewith, more traditional techniques for the Markovian approximation of the GLE. Typically, the latter approaches rely on a continued-fraction expansion of the Laplace transform of the autocorrelation functions of the noise in the GLE, as introduced by Mori [56], or on related approximations by rational functions of linear GLE memory kernels [57, 58]. Still, the applicability of these Markov approximation techniques is limited by relying on rather restrictive assumptions, such as systems with a separable, quadratic Kac-Zwanzig Hamiltonian [57] or linear kernels, although non-Gaussian noise in the GLE is allowed [57]. These restrictions led some authors to conclude that MZ models with linear, finite-length kernels form a subclass of autoregressive moving average (ARMA) models [59, 60].

As shown in the body of this paper, our approach to the derivation of Markov approximations to the GLE goes beyond these limitations and allows for nonlinear kernels as well as for non-Gaussian noise; it applies, furthermore, to a broad class of dissipative, rather than Hamiltonian systems. Nevertheless, the considerations in this subsection demonstrate the intuitive relevance of the multilevel structure inherent in the EMR methodology — albeit initially designed from a different perspective [18, 26] — for the derivation of closure models from a multivariate time series of partial observations. This article shows, furthermore, that — when suitably generalized — the multilayer EMR methodology provides an efficient means of deriving such closure models, as well as facilitating their mathematical analysis.

2. The closure problem from partial observations and its EMR solution

As discussed in Secs. 1.1 and 1.2 above, we are motivated by the modeling of geophysical fluid flows — as well as of more complex climate problems and of large-dimensional problems

from other fields of science — based on a series of partial observations. We formulate below the corresponding observation-based closure problem ($\mathfrak{P}$) and recall its EMR candidate solution, such as initially proposed in [18, 19]. Generalizations of such an EMR solution are discussed in Section 3 below.

One can consider the approach presented in this article as complementary to the derivation of deterministic nonlinear dynamics from observations, in the spirit of Mañe-Takens [61, 62] attractor reconstruction by phase-space embedding of a time series: instead of just trying to reconstruct the attractor from a single time series or from a multivariate one [63, 64, 65], we attempt to actually write down equations that will produce a good approximation of the attractor, including both its geometry and invariant measure.

Several approaches can be used for this purpose. Among them are the recent time-lagged polynomial sparse modeling technique for the nonuniform embedding of time series [66], or the more traditional approaches based on ARMA models [67]. The nonlinear version of the latter [68, 69] is somewhat closer to the EMR methodology, due to the combined presence of noise and memory effects[2]. However, the EMR models differ in their parameter estimation by the top-to-bottom multilevel procedure recalled below and subject to the stopping criterion described in A. As we will see from the theoretical results of Section 3, the multilevel structure intrinsic to EMR allows for a great flexibility in specifying various linear and nonlinear interactions between the main-level and hidden-level variables, in order to design MSM generalizations of EMR models; see Theorem 3.1 and Corollary 3.2 below. The same multilevel structure allows us furthermore to relate MSMs to MZ models, as explained in Sections 4 and 5.

As mentioned earlier, MZ models provide optimal solutions to closure problems such as the observation-based problem ($\mathfrak{P}$) formulated below. As we will see, EMR or their MSM generalizations of Section 3 provide approximate solutions to such optimal solutions; see Section 6 and 7 for applications.

($\mathfrak{P}$) Let $\{u(t_k, \xi_j)\}$ be a set of discrete observations of a given, spatio-temporal, scalar field, where $\{t_k = k\delta t\}$[3], and the grid points $\{\xi_j\}$ live in a spatial domain $\mathcal{D}$, which can be two- or three-dimensional. The amount of data is always assumed to be finite, *i.e.* $k \in \{1, ..., N\}$ and $j \in \mathcal{J}$, where $\mathcal{J}$ is a finite set of multi-indices; often, the data set is rather limited, and thus N is not as large as we would like. The goal is to find a model that not only describes the evolution of the $\{u(t_k, \xi_j)\}$ observed so far, but also possesses good prediction skills at future times, $k > N$, and at the same sites $\{\xi_j\}$.

When equations are available to describe the evolution of $\{u(t_k, \xi_j)\}_{j\in\mathcal{I}}$ in the domain $\mathcal{D}$, with $\text{card}(\mathcal{I}) >> \text{card}(\mathcal{J})$, this problem is related to closure problems of a type that has been widely studied in continuum physics, and a variety of approaches is available for dealing with such problems [75]. In the general case, though, no equations are available for the evolution of the full field $\mathcal{U}_\mathcal{I} := \{u(t_k, \xi_j)\}_{j\in\mathcal{I}}$, while the subfield $\mathcal{U}_\mathcal{J} := \{u(t_k, \xi_j)\}_{j\in\mathcal{J}}$ is assumed to consist of *partial observations* from the field $\mathcal{U}_\mathcal{I}$ that contains subgrid information not resolved by the field $\mathcal{U}_\mathcal{J}$.

The dynamics of the unobserved variables $\{u(t_k, \xi_j)\}_{j\in\mathcal{I}\setminus\mathcal{J}}$ is thus lacking and the main issue is to derive an efficient parameterization of the interactions between the resolved and unresolved — i.e., roughly speaking, the observed and unobserved — variables in order to derive equations that model the evolution of the field $\mathcal{U}_\mathcal{J}$ with reasonable accuracy. Furthermore, the scalar field

[2]Note that, in some sense, the EMR methodology can also be viewed as an extension of hidden Markov models (HMMs) [70, 71] or of artificial neural networks (ANNs) [72, 73, 74], since the latter are generally nonlinear but do not involve the memory effects inherent in the EMR methodology; see [24, 44].

[3]Here δt denotes the sampling interval of the data. For the sake of simplicity, we consider in this article data that are uniformly sampled with a constant δt.

u may interact with other fields that are not taken into account or not observed, which makes an accurate modeling of the field $\mathcal{U}_{\mathcal{J}}$ even more difficult. Typically, two-point statistics — such as correlations or the joint probability density function (PDF) of the observed variables — are used to assess the effectiveness and accuracy of the proposed closure model.

Besides these modeling aspects, the prediction requirement in the problem ($\mathfrak{P}$) above is clearly challenging, in theory as well as in practice. It is well known that an inverse model may approximate the data accurately over the training interval, while exhibiting poor prediction skill on the validation interval during hindcast experiments, and performing even worse in real-time prediction. Nonetheless and as mentioned in Sec. 1.1, the data-driven EMR procedure — described in Steps 1–3 below— has proven to be quite skillful in real-time ENSO prediction [17]. For example, although the EMR-ENSO model of [19] is based on a leading subset of principal components of the SST field that constitute only a small fraction of the total set of degrees of freedom of the ENSO phenomenon; still, this EMR-ENSO model "has the highest seasonally combined correlation skill among the statistical models, exceeded by only a few dynamical models"; see footnote[1].

As initially formulated in [18], after appropriate compression of the available data, cf. Step 1 below, the key idea consists in seeking forced-dissipative models of the form

$$\frac{\delta x_k}{\delta t} = -Ax_k + B(x_k, x_k) + F + \text{"noise."} \tag{2.1}$$

Here $x_k = x(t_k)$ is the d-dimensional column vector of the relevant variables to be modeled, $\delta x_k/\delta t$ represents its rate of change in time, F is a column vector, A is a square $d \times d$ matrix, and $B(x_k, x_k)$ accounts for the bilinear interactions that influence the evolution of $\delta x_k/\delta t$. The noise term accounts typically for the effects of the unresolved variables — i.e., of subgrid-scale processes or of other unobserved variables — on the dynamics of the observed variables, as modeled by Eq. (2.1).

Often, this noise is simply assumed to be Gaussian and white in time but, in practice, its amplitude may depend on the fluctuating variable x itself. It is, therefore reasonable to model such a noise as *state-dependent*, and allow for it to possess non-vanishing correlations in time as well as in space. Actually, Markov representation theorems ensure the existence of such a state-dependent noise, as long as the original, possibly large-dimensional, deterministic, time-discrete, observed dynamical system, possesses a relevant invariant measure that is *physical* in the sense recalled in Section 4 below, cf. Eq. (4.3); see also Corollary B in the Supporting Information of [76]. In the EMR methodology, this state-dependent noise is modeled by means of successive regressions against the coarse observed variable x, until the time correlations of the residual noise satisfy suitable decorrelation criteria.[4]

The main steps of the EMR procedure can be summarize as follows:

(i) **Step 1: Data compression.** Let $\mathcal{B} = \{E_1, ..., E_d\}$ be an orthogonal basis[5] that spans a finite-dimensional Euclidean space of dimension d with associated norm $\|\cdot\|$, so that

[4]For the sake of clarity, the decorrelation criterion is recalled in A. Roughly speaking, the residual noise has to have vanishing correlations at lag δt, i.e., at the sampling interval of the available data. The successive minimizations, described in Steps 1–3 below, have been shown, in practice, to reach such a limit, within a reasonable error, after solving only finitely many minimization problems; see [18, 19, 21, 26] and Remark 1 below.

[5]This basis is determined from the data; empirical orthogonal function (EOFs) are often used in EMR methodology, but other possibilities do exist [25, 26]; see also [77] for a comparison of different bases used in model reduction.

the following orthogonal decomposition of $u(t_k, \xi_j)$ onto $\mathcal{B}$ can be easily determined:

$$u(t_k, \xi_j) = \sum_{i=1}^{d} x_i(t_k) E_i(\xi_j). \tag{2.2}$$

In what follows, $x(t_k)$ is the d-dimensional column vector[6] $x(t_k) := (x^1(t_k), ..., x^d(t_k))^{\mathrm{T}}$. Given this decomposition, the problem $(\mathfrak{P})$ is now restricted to the modeling and eventual prediction of the time evolution of $x(t)$ or even of just a few of its components. A data-driven approach to this problem proceeds in the two following steps.

(ii) **Step 2: Main-level regression.** Given the multivariate time series $\{x(t_k) : k = 1, \ldots, N\}$, one seeks a deterministic column vector F_N, a square matrix A_N, and a quadratic form $B_N(x, x)$ to achieve the following l_2 minimization:

$$[F_N, A_N, B_N] = \underset{F,A,B}{\operatorname{argmin}} \Big(\sum_{k=1}^{N-1} \Big\| \frac{\delta x_k}{\delta t} - F + A x_k - B(x_k, x_k) \Big\|^2 \Big); \tag{2.3}$$

here $\delta x_k := x_{k+1} - x_k$, and $x_k = x(t_k)$ represents the evolution in time of $u(t_k, \cdot)$ as decomposed onto the basis $\mathcal{B}$. Clearly, F_N, A_N and B_N in Eq. (2.3) are meant to approximate F, A and B in Eq. (2.1) so as to optimize in an l_2-sense the evolution of x_k with respect to the data $u(t_k, \xi_j)$. The subscript N emphasizes the fact that F_N, A_N and B_N are estimated from the finite-length, multivariate time series $\{x_k\}_{k\in\{1,\ldots,N\}}$.

(iii) **Step 3: Multilevel regression.** Let $r_k^{(0)} := r^{(0)}(t_k) \in \mathbb{R}^d$ be the d-dimensional regression residual associated with the l_2-minimization problem of Step 2. We seek now a $d \times 2d$ rectangular matrix $L_N^{(1)}$ cross-interactions, associated with the main-level residual $r_k^{(0)}$ and the main-level variable x, so that:

$$L_N^{(1)} = \underset{L\in\mathbb{R}^{d\times 2d}}{\operatorname{argmin}} \Big(\sum_{k=1}^{N-1} \Big\| \frac{\delta r_k^{(0)}}{\delta t} - L\big[(x_k)^T, (r_k^{(0)})^T\big]^T \Big\|^2 \Big). \tag{2.4}$$

This main-level step is followed by solving a sequence of l_2-minimization problems: for each m of interest, we seek a $d \times (m+1)d$ rectangular matrix $L^{(m)}$ of cross-interactions, which is given by:

$$L_N^{(m)} = \underset{L\in\mathbb{R}^{d\times (m+1)d}}{\operatorname{argmin}} \Big(\sum_{k=1}^{N-1} \Big\| \frac{\delta r_k^{(m-1)}}{\delta t} - L\big[(x_k)^T, (r_k^{(0)})^T, ..., (r_k^{(m-1)})^T\big]^T \Big\|^2 \Big). \tag{2.5}$$

This recursive sequence of minimizations is stopped when, for some $m = p$, the residual $r_k^{(p)} \delta t$ has vanishing autocorrelation at lag δt, according to the stopping criterion described in A.

From a practical point of view, the solution of the successive minimization problems described above is subject to the general problem of *multicolinearity*, which arises from finite sample size of the data set [18]. Multicolinearity can lead to statistically unstable estimates, in the sense that the latter can be very sensitive to small changes in the data set. Various regularization techniques exist to deal with this problem [78]; they rely mainly on penalizing the l_2-functionals in Steps 1–3 above, and they have been shown to be effective in yielding stable estimates for the EMR models of various climate fields [18, 24, 26].

[6] If $\mathcal{B}$ denotes the set of EOFs, then the x_i's correspond to the principal components.

Once the time-independent forcing term F_N, the square matrix A_N, the bilinear term B_N, and the rectangular matrices $\{L_N^{(m)} : m = 1, \ldots, p\}$ have been determined, we are left with an EMR model of the form:

$$(EMR) \qquad \boxed{\begin{aligned} x_{k+1} - x_k &= \big[- A_N x_k + B_N(x_k, x_k) + F_N \big]\delta t + r_k^{(0)} \delta t, \quad k \in \{1, ..., N\}, \\ r_{k+1}^{(m-1)} - r_k^{(m-1)} &= L_N^{(m)} \big[(x_k)^T, (r_k^{(0)})^T, ..., (r_k^{(m-1)})^T\big]^T \delta t + r_k^{(m)} \delta t, \quad 1 \le m \le p, \end{aligned}}$$

since, in general, $p \geq 1$. This EMR models the dynamics of the multivariate time series $\{x_k : k = 1, \ldots, N\}$ of length N and should be able to predict it for $k > N$.

Note that, for prediction purposes, the model defined by Eq. (*EMR*) has to be initialized appropriately in the past, along with estimating the *hidden* $r^{(m)}$-variables from the *observed* x, as explained in B. For the moment we turn to a natural continuous-time formulation of Eq. (*EMR*) that will help set up a framework in which the EMR methodology will become theoretically more transparent by benefitting from a natural MZ interpretation; see Secs. 3– 5.

Assuming that F_N, A_N, B_N and $\{L_N^{(m)} : m = 1, \ldots, p\}$ have converged once the amount of data gets large enough[7], we are formally left with an MSM that becomes, in the limit of $\delta t \to 0$ and $N \to \infty$:

$$(2.6) \qquad \begin{aligned} \mathrm{d}x &= \big[- Ax + B(x, x) + F \big]\mathrm{d}t + r^{(0)}\mathrm{d}t, \quad 0 < t \le T^*, \\ \mathrm{d}r^{(m-1)} &= L^{(m)} \big[(x)^T, (r^{(0)})^T, ..., (r^{(m-1)})^T\big]^T \mathrm{d}t + r^{(m)}\mathrm{d}t, \quad 1 \le m \le p. \end{aligned}$$

Here $[0, T^*]$ corresponds to the time interval over which the data are available — also known as the *training interval* — during which the MSM parameters of (2.6) have been estimated.

Note that in the theoretical limit above, the minimization functionals appearing in Steps 1–3 have to be replaced by their continuous analogues; this just means replacing the l_2-functionals by their $L^2(0, T^*)$ version. The rigorous justification of such a limit is of course a nontrivial task for general data. It implies in particular going beyond traditional finite-time convergence theory for numerical methods applied to SDEs [80]. SDEs such as those considered in (2.6) involve only locally Lipschitz coefficients, because of the bilinear term B, and thus require more sophisticated techniques to prove strong convergence results of the solutions of (*EMR*)[8] to the solutions of (2.6), when the former system is interpreted as an Euler-Maruyama discretization of the latter. Such a convergence can still be guaranteed, provided the n^{th} moment of the exact and the numerical solutions are bounded for some $n \geq 3$; see [81, 82]. Leaving such considerations aside in the present article, we focus here on theoretical MSM aspects that will turn out to be useful from a modeling point of view, as shown in Secs. 3 –5.

Remark 1. *Note that convergence results such as those mentioned above have been obtained for SDEs driven by white noise. The system* (2.6) *is a system of random differential equations (RDEs), in which the last-level noise term* $r^{(p)}$ *is not necessarily white according to the stopping criterion recalled in A. Equation* (2.6) *can, however, turn out to be well approximated by a genuine system of SDEs driven by a Wiener process* W*. Let us assume that the residual noise* $r^{(p)}$*, obtained from the last level of Step 3 above, obeys the following scaling law as* $\delta t \to 0$:

$$(2.7) \qquad r^{(p)} \sqrt{\delta t} \underset{\delta t \to 0}{\sim} \mathcal{N}(0, Q),$$

[7] in some appropriate weak sense [79].

[8] in which F_N, A_N, B_N and $L_N^{(m)}$ would be replaced by F, A, B and $L^{(m)}$, respectively.

where $\mathcal{N}(0,Q)$ denotes the normal distribution with zero mean and covariance matrix Q. Then $r^{(p)}\delta t$ is a natural approximation of $Q\,\mathrm{d}W$ as $\delta t \to 0$, where $\mathrm{d}W$ denotes a d-dimensional white noise process.

We emphasize, though, that the scaling law (2.7) may be violated by certain data sets and that other scaling laws may apply to the stochastic process obtained in the limit of $\delta t \to 0$. These laws could lead, for instance, to MSMs driven by Lévy α-stable processes, as argued for certain climate fields [83].

When the scaling of (2.7) is violated, one may wish to consider other types of interactions than quadratic or linear in the learning stage of (*EMR*). Doing so may better explain the state dependence of the main-level residual noise, and lead naturally to the more general class of MSM considered hereafter.

3. MSMs: general formulation and random attractors

Motivated by the interest in more general problems than those posed by geophysical fluid flows, we identify here a general class of MSMs that possess a global random attractor, whose existence prevents the blow-up of model solutions. This class extends the one proposed recently in [28, 29] and does not require the inclusion of energy-preserving nonlinearities to ensure the existence of such an attractor.

3.1. MSMs and global random attractors.

A natural extension of the MSMs introduced in Sec. 2 can be written in compact form as follows:

$$\begin{aligned} \mathrm{d}x &= F(x)\mathrm{d}t + \Pi r\,\mathrm{d}t, \\ \mathrm{d}r &= (Cx - Dr)\mathrm{d}t + \Sigma\,\mathrm{d}W_t. \end{aligned} \tag{3.1}$$

Here $r = (r_1, ..., r_p)^{\mathrm{T}}$ is a $p \times q$-dimensional vector with components that are q-dimensional each, such that $pq \geq d$, while C and D are respectively $pq \times d$ and $pq \times pq$ matrices, and Π is the orthogonal projection from $\mathbb{R}^{pq}$ onto $\mathbb{R}^d$. The matrix Σ is a $(pq) \times q$-rectangular matrix whose last q rows are equal to a positive-definite matrix Q and zero elsewhere.

Another, even more general extension can be written in the following form:

$$\boxed{\begin{aligned} \mathrm{d}x &= f_1(x)\mathrm{d}t + g_1(u)\mathrm{d}t \\ \mathrm{d}y &= f_2(y)\mathrm{d}t + g_2(u)\mathrm{d}t + \Pi r \mathrm{d}t, \\ \mathrm{d}r &= h(u,r)\mathrm{d}t + \Sigma\,\mathrm{d}W_t. \end{aligned}} \tag{MSM}$$

Here $u = (x, y) \in \mathbb{R}^d \times \mathbb{R}^{d'}$ and Π is the orthogonal projection from $\mathbb{R}^{pq}$ onto $\mathbb{R}^{d'}$; for simplicity, the functions f_1, f_2, g_1, g_2 and h are assumed to be continuous and locally Lipschitz in $\mathbb{R}^d \times \mathbb{R}^{d'}$. The $d \times d'$-dimensional vector y here plays the role of a hidden variable, similar to the d-dimensional vector r_0 introduced in (2.6), except that the y-variable is allowed now to contribute nonlinearly to the dynamics of x via the vector field g_1 and that it is not necessarily of the same dimension as x, i.e. d' may differ from d. Likewise, nonlinear effects are allowed this time for both the hidden variables y and r. The generalization of (2.6) via $d \neq d'$ and nonlinear effects in the second equation of (*MSM*) was proposed recently in [28, 29], but h there was assumed to be linear.

Obviously, the cross-interactions between the observed and hidden variables, modeled here by the terms g_1, g_2 and h, as well as the self-interactions contained in h and carried by f_1 and f_2, have to be designed properly so that the system remains stable. Ref. [28] has addressed this problem for the class of quadratic-energy–preserving nonlinearities, while keeping h linear, by applying geometric ergodic theory to the Markov semigroup associated with such systems. In

this section, we adopt an alternative approach and address the problem of existence of random attractors for the more general class of inverse models governed by (*MSM*).

Remark 2. *Note that random attractors may exist while the Hörmander condition used in [28] is violated. Such a violation may arise, for instance, for stochastically perturbed system of ODEs that exhibit a strange attractor and support a Sinai-Bowen-Ruelle measure μ, when the noise acts transversally to the support of μ, along the stable Oseledets subspaces [84]. In other words, random attractors may still exist for SDEs driven by degenerate noise and our present approach may be viewed as complementary to that of [28]; see Theorem 3.1 and Corollary 3.2 below. Furthermore, the random-attractor arguments used here are not limited to the case of white noise and can be used to ensure the existence of non-degenerate solutions for MSMs driven by more general noises [85], as proposed in [83]; see also Remark 1.*

We proceed by transforming the system of SDEs in (*MSM*) into a system of RDEs that is more amenable to the application of energy estimate or Lyapunov function methods, which we use hereafter; see [86]. Mathematically, the benefits of dealing with the transformed system instead of the original one relies on the Hölder property of Ornstein-Uhlenbeck processes, which allows one to rely on the classical theory of non-autonomous evolution equations, parameterized by the realization ω of the driving noise; see [87, Lemma 3.1 and Appendix A]. To perform this transformation, we consider the following auxiliary $pq-$dimensional Ornstein-Uhlenbeck process, obtained as the stationary solution of

$$\mathrm{d}z = -A_0 z\,\mathrm{d}t + \Sigma\,\mathrm{d}W_t, \tag{3.2}$$

where A_0 is a positive-definite $pq \times pq$ matrix, that will be determined later on. The change of variables $r \leftarrow r - z$ transforms the system (*MSM*) into the aforementioned system of RDEs:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= f_1(x) + g_1(x,y), \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= f_2(y) + g_2(x,y) + \Pi(r+z), \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= h(u,r) + A_0 z, \end{aligned} \tag{3.3}$$

and let us assume, for simplicity, that

$$h(u,r) = -Dr + \widetilde{h}(u,r), \tag{3.4}$$

with D positive definite and $A_0 = D$.

Following the Lyapunov function techniques in [86] for the existence of random attractors, we assume that there exists a map $V : \mathbb{R}^{d+d'} \to \mathbb{R}^+$, which is continuously differentiable and has the property that pre-images of bounded sets are bounded, e.g. V is polynomial. Furthermore, we require that V verify the following conditions:

(H$_1$) $\quad \exists\, \alpha, \beta > 0 \;:\; \langle f_1(x), \nabla_x V(u)\rangle + \langle f_2(y), \nabla_y V(u)\rangle + \alpha \cdot V(u) \le \beta,\ \forall\, u \in \mathbb{R}^{d+d'}$,

(H$_2$) $\quad \exists\, a, b > 0 \;:\; \langle g_1(u), \nabla_x V(u)\rangle + \langle g_2(u), \nabla_y V(u)\rangle \le aV(u) + b,\ \forall\, u \in \mathbb{R}^{d+d'}$,

(H$_3$) $\quad \langle \Pi r, \nabla_y V(u)\rangle + \langle \tilde{h}(u,r), r\rangle \le m_1 V(u) + m_2 \|r\|^2$,

(H$_4$) $\quad \|\nabla_y V(u)\|^2 \le \gamma V(u) + C$,

where $\langle f, g\rangle$ is the Euclidean inner product in $\mathbb{R}^{d+d'}$.

For a given realization ω of the noise, we denote by $V(t,\omega)$ the values of V along a trajectory $u(t,\omega) = (x(t,\omega), y(t,\omega))$, and obtain, on the one hand, from the x- and y-equations in (3.3),

$$\mathrm{d}V(t,\omega) = \Big(\sum_{i=1}^{2}\langle f_i(x_i(t,\omega)), \nabla_{x_i}V(t,\omega)\rangle + \langle g_i(x(t,\omega)), \nabla_{x_i}V(t,\omega)\rangle\Big)\mathrm{d}t, \tag{3.5}$$

where we used the notations $x_1 = x,\ x_2 = y$; on the other, after multiplication of the r-equation by r in (3.3), we get

$$\frac{1}{2}\mathrm{d}\|r\|^2 + \lambda\|r\|^2\mathrm{d}t \le \Big(\langle \Pi(r+z), \nabla_y V(t,\omega)\rangle + \langle \tilde{h}(u,r), r\rangle\Big)\mathrm{d}t, \tag{3.6}$$

where $\lambda > 0$ denotes the smallest eigenvalue of D.

By adding (3.5) and (3.6), by using (H_1)–(H_3), and by applying the ϵ-Young inequality to the term $\langle \Pi_2 z, \nabla_y V(t,\omega)\rangle$, we obtain that, for any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that

$$\begin{aligned}\mathrm{d}V(t,\omega) + \frac{1}{2}\mathrm{d}\|r\|^2 + (\alpha V(t,\omega) + \lambda\|r\|^2)\mathrm{d}t \le \Big((a + m_1 + \epsilon\gamma)V(t,\omega) + \\ \beta + b + m_2\|r\|^2 + \epsilon C + C_\epsilon\|\Pi_2 z(t,\omega)\|^2\Big)\mathrm{d}t.\end{aligned} \tag{3.7}$$

This inequality, in turn, leads to

$$\mathrm{d}\widetilde{V}(t,\omega) + \kappa\widetilde{V}(t,\omega)\mathrm{d}t \le (\widetilde{\beta} + \xi(t,\omega))\mathrm{d}t, \tag{3.8}$$

with $\widetilde{V}(t,\omega) := V(t,\omega) + \frac{1}{2}\|r\|^2$, $\widetilde{\beta} = \beta + b + \epsilon C$, and

$$\kappa = \min(\alpha - a - m_1 - \epsilon_1\gamma; 2(\lambda - m_2)) \tag{3.9}$$

for some $\epsilon_1 > 0$, where

$$\xi(t,\omega) = C_{\epsilon_1}\|\Pi z(t,\omega)\|^2. \tag{3.10}$$

Since the Ornstein-Uhlenbeck process z is stationary with respect to the canonical shift θ_t, so is the process $\xi(t,\omega)$, which can be written without loss of generality as $\xi(t,\omega) = \xi(\theta_t\omega)$; see [88, Appendix A] and [89]. Recalling that

$$\lim_{t\to\infty}\frac{z(t,\omega)}{t} = 0, \tag{3.11}$$

almost surely (see [87, Lemma 3.1]), we deduce that, for almost all ω, the positive random variable

$$R(\omega) = \frac{\beta}{\kappa} + \int_{-\infty}^{0} e^{\kappa s}\xi(\theta_s\omega)\mathrm{d}s, \tag{3.12}$$

is well-defined, provided that $\kappa > 0$, which we assume hereafter to be the case.

For $\epsilon > 0$, the random set

$$\mathfrak{B}_\epsilon(\omega) := \{v = (x,y,r) \in X : \widetilde{V}(v) \le (1+\epsilon)R(\omega)\}, \tag{3.13}$$

in $X := \mathbb{R}^{d+d'} \times \mathbb{R}^{pq}$ is then, by assumption on V, almost surely bounded. One can conclude forthwith that $\mathfrak{B}_\epsilon(\omega)$ is a random set that is pullback absorbing for the RDS φ generated by (3.3); namely, for any bounded set $\mathfrak{B} \subset X$, there exists a positive random absorption time $t^*(\omega, \mathfrak{B})$ such that, for every $t \ge t^*(\omega, \mathfrak{B})$,

$$\widetilde{V}(\varphi(t, \theta_{-t}\omega)\mathfrak{B}) \subset \mathfrak{B}_\epsilon(\omega), \tag{3.14}$$

i.e. the compact random set $\mathfrak{B}_\epsilon$ is absorbing every bounded deterministic set $\mathfrak{B}$ of X. Standard results from RDS theory (see [90, Theorem 3.11]) allow us then to conclude that a global random attractor exists for the RDS generated by (3.3) and therefore for the one associated with (*MSM*).

We have thus proved the following theorem.

Theorem 3.1. *Let us consider a system (MSM) that generates an RDS. Assume that there exists a map $V : \mathbb{R}^{d+d'} \to \mathbb{R}^+$, such that $V \in C^1(\mathbb{R}^{d+d'}, \mathbb{R}^+)$, which has the property that pre-images of bounded sets are bounded and that the conditions (H_1)–(H_4) hold. Assume, furthermore, that for all $u \in \mathbb{R}^{d+d'}$ and for all $r \in \mathbb{R}^{pq}$,*

$$\widetilde{h}(u, r) = h(u, r) + Dr, \tag{3.15}$$

with D a $pq \times pq$ positive-definite matrix, and that the structural constants involved in (H_1)–(H_4) can be chosen such that

$$\kappa = \min(\alpha - a - m_1 - \epsilon_1\gamma; 2(\lambda - m_2)) > 0 \tag{3.16}$$

for some $\epsilon_1 > 0$, where λ denotes the smallest eigenvalue of D.

Then the system (MSM) possesses a global random attractor which is pullback attracting the deterministic bounded sets of $\mathbb{R}^{d+d'} \times \mathbb{R}^{pq}$.

Interpretation and practical consequences. Condition (H_1) of Theorem 3.1 identifies a natural dissipation condition that the self-interactions terms f_1 and f_2 should satisfy in the absence of coupling between the x- and y-variables and forcing from the r-equation. Condition (H_2) and conditions (H_3, H_4) indicate what the cross-interactions between the x- and y-variables, on the one hand, and the effects of the forcing by the r-variable on the y-equation, on the other, should satisfy for the system (*MSM*) to be pathwise dissipative in a pullback sense [89]. Condition (3.16) dictates the appropriate balance between these combined effects for such dissipativity to occur.

Conditions (H_1)–(H_4) may also prove practically useful in the design of an MSM. For instance, when an appropriate energy V has been identified such that (H_1) is satisfied, the cross-interactions g_1 and g_2 can be designed according to the constraint

$$\langle g_1(u), \nabla_x V(u)\rangle + \langle g_2(u), \nabla_y V(u)\rangle = 0, \tag{3.17}$$

while conditions (H_3) and (H_4) are easy to satisfy when V is quadratic and $\widetilde{h}$ is linear, as in [28, 29].

If $V(u) = \frac{1}{2}(\|x\|^2 + \|y\|^2)$, it is quite natural to satisfy (H_1) simply by seeking f_i's of the form

$$f_i(x_i) = -A_i x_i + B_i(x_i, x_i) + F_i,\ i = 1, 2, \tag{3.18}$$

with $x_1 = x$ and $x_2 = y$, while the linear parts satisfy the dissipation conditions

$$\langle A_i x_i, x_i\rangle \geq \nu_i \|x_i\|^2, \tag{3.19}$$

and the bilinear parts satisfy the energy-preserving conditions

$$\langle B_i(x_i, x_i), x_i\rangle = 0,\ i = 1, 2. \tag{3.20}$$

The constraint (3.17) becomes therewith

$$\langle g_1(x, y), x\rangle + \langle g_2(x, y), y\rangle = 0, \tag{3.21}$$

and, given a linear $\widetilde{h}$ and bilinear g_i's, Eqs. (3.20, 3.21) lead to a class of energy-preserving MSMs considered in [28, 29]. For this class, Theorem 3.1 ensures the existence of a random attractor provided that the structural constants involved satisfy (3.16).

Clearly, the conditions of Theorem 3.1 allow for the design of a considerably broader class of stable MSMs. For instance, a closer look at the constraint (3.17) allows us to formulate the following corollary.

Corollary 3.2. *Let us consider a system (MSM) for which $d = d'$, and that generates an RDS. Assume that there exists a continuously differentiable map $V : \mathbb{R}^{2d} \to \mathbb{R}^+$, such that condition ($H_1$) is satisfied and that pre-images of bounded sets under V are bounded.*

Assume that there exists a continuously differentiable, scalar-valued function G on $\mathbb{R}^{2d}$ such that

$$g_1(u) = \nabla_y G(u), \qquad g_2(u) = -\nabla_x G(u), \tag{3.22}$$

and that G is a first integral of the flow associated with the Hamiltonian V, i.e. that the Poisson bracket of G with V vanishes,

$$\{G, V\} = 0. \tag{3.23}$$

Then, if conditions (H_3, H_4) are satisfied, the conclusion of Theorem 3.1 holds.

Proof. The proof of this corollary boils down to expanding (3.23) in canonical coordinates to yield [91, p. 215]

$$\sum_{i=1}^{d} \frac{\partial G}{\partial y_i}\frac{\partial V}{\partial x_i} - \frac{\partial G}{\partial x_i}\frac{\partial V}{\partial y_i} = 0, \tag{3.24}$$

which, by using (3.22), is nothing else than (3.17). □

This corollary shows that, once the self-interactions terms f_1 and f_2 have been designed to be dissipative, according to (H_1), with respect to some energy V, the cross-interactions between the main level variable x and the next-level hidden variable y can be derived from any constant of motion G of the flow generated by the auxiliary Hamiltonian system

$$\dot{p} = \nabla_q V(p, q), \qquad \dot{q} = -\nabla_p V(p, q). \tag{3.25}$$

The determination of such cross-interactions can thus benefit from efficient symplectic integrators techniques [92, 93] that we leave for future research.

Whatever the level of generality of the MSM formulation, though, the selection of the corresponding allowable self- and cross-interactions should be balanced with the constraint of reduction of the unexplained variance contained in the residuals of the main and subsequent layers, when compared to an MSM written in its original EMR formulation (2.6). This variance-based constraint is also to be balanced against the correlation-based criterion formulated in Section 5 below. The latter criterion relies on a reformulation of an MSM into a system of stochastic integro-differential equations that is described below, a reformulation that allows for a comparison with the optimal closure model one can obtain from a time series of partial observations of a large system; see Section 4.

Remark 3. *In practice, an ensemble Kalman filter can be used to learn MSMs subject to constraints such as (3.20, 3.21) [28, 29]. In C, we describe an alternative approach, in which such constraints can be naturally incorporated into the recursive minimization procedure described in Section 2.*

3.2. **MSMs as systems of stochastic integro-differential equations.** In this section we show how to rewrite an MSM as a system of stochastic integro-differential equations like those that arise in the MZ formalism. We restrict ourselves here to the case of h in (3.4) being given by

$$\begin{aligned} h(u,r) &= -Dr + \tilde{h}(x), \\ g_2(u) &= g_2(x), \end{aligned} \tag{3.26}$$

i.e. we drop the dependence of $\tilde{h}$ (respectively of g_2) on r and y (respectively on y), while D is still a $pq \times pq$ positive-definite matrix.

In this case, the integration of the last equation in (*MSM*) gives, for $r(0) = 0$,

$$r(t) = \int_0^t e^{-(t-s)D}\tilde{h}(x(s))\mathrm{d}s + \int_0^t e^{-(t-s)D}\Sigma\, \mathrm{d}W_s, \tag{3.27}$$

where the last integral is understood as a *stochastic convolution* in the sense of Itō [94, Chap. 5]. Denoting by $S(t)$ the flow associated with $\dot{y} = f_2(y)$, we obtain likewise, for $y_0 = 0$,

$$y(t) = \int_0^t S(t-s)\Big(g_2(x(s)) + \Pi r(s)\Big)\mathrm{d}s, \tag{3.28}$$

with $r(s)$ given by (3.27).

Let us assume, furthermore, that

$$g_1(u) = \beta(x,y) + Cy, \tag{3.29}$$

where C is a $d \times d'$ matrix, while $\beta(x,y)$ does not contain linear terms in the y-variable nor homogeneous terms in the x-variable. Given these assumptions, we can use Eqs. (3.27, 3.28) to rewrite the system of SDEs (*MSM*) as the following *RDE with retarded arguments*:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f_1(x) + \mathbb{G}(t, x_t) + \xi(t,\omega), \tag{3.30}$$

with

$$x_t(s) := x(t+s),\ -t \le s \le 0, \tag{3.31}$$

$$\xi(t,\omega) = C\int_0^t S(t-s')\Big(\Pi\int_0^{s'} e^{-(s'-s)D}\Sigma\, \mathrm{d}W_s(\omega))\Big)\mathrm{d}s'. \tag{3.32}$$

Here $\mathbb{G}(t, x_t)$ accounts for (i) the terms in (3.29) that come from $\beta(x(t), y(t))$, where $y(t)$ has been replaced by its expression (3.28); and (ii) the terms $C\big(\Pi\int_0^t e^{-(t-s)D}\tilde{h}(x(s))\mathrm{d}s + \int_0^t S(t-s)g_2(x(s))\mathrm{d}s\big)$. Note that each of these terms is given by the following type of integral:

$$\int_0^t F(t,s,x(s))\mathrm{d}s; \tag{3.33}$$

the latter can be rewritten, by using (3.31), as follows:

$$\int_0^t F(t,s,x(s))\mathrm{d}s = \int_{-t}^0 F(t,t+s,x(t+s))\mathrm{d}s = \int_{-t}^0 F(t,t+s,x_t(s))\mathrm{d}s := \mathbb{F}(t,x_t), \tag{3.34}$$

which explains the functional dependence in (3.30) on the past history x_t of x.

The abstract reformulation of (*MSM*) as a closed equation (3.30) that describes the evolution of x shows that an MSM written in a closed form involves a functional dependence on the time history of the observed variables. This important characteristic of a time-continuous EMR was already highlighted, in a simpler context, in the Supporting Information of [44].

Note that $\xi(t,\omega)$ has a straightforward interpretation when taking into account the multilayer structure of an MSM. Indeed, if we remember that the r-equation in (*MSM*) has p levels and if we assume, furthermore, that the matrix D is block-diagonal and given by

$$D = \mathrm{diag}(D_1, \cdots, D_p), \tag{3.35}$$

where each D_j, $j = 1, \ldots, p$, is a $q \times q$ matrix, it follows that $\xi(t,\omega)$ is obtained, due to the assumption on Σ, as the following convolution

$$\xi(t,\omega) = C \int_0^t S(t-s)\Pi z_1(s)\mathrm{d}s; \tag{3.36}$$

here the stochastic process z_1 is obtained from the successive integrations of the following equations:

$$\begin{aligned} \mathrm{d}z_p &= -D_p z_p \mathrm{d}t + Q\,\mathrm{d}W, \\ \mathrm{d}z_{p-1} &= -D_{p-1} z_{p-1}\mathrm{d}t + z_p \mathrm{d}t, \\ \mathrm{d}z_{p-2} &= -D_{p-2} z_{p-2}\mathrm{d}t + z_{p-1}\mathrm{d}t, \\ &\vdots \qquad\qquad \vdots \\ \mathrm{d}z_1 &= -D_1 z_1 \mathrm{d}t + z_2 \mathrm{d}t. \end{aligned} \tag{3.37}$$

In other words, the process z_1 can be rewritten in terms of repeated convolutions as follows

$$z_1(t) = \kappa_1 * \cdots * \kappa_{p-1} * z_p(t), \tag{3.38}$$

where the kernels κ_k are given by the following exponential matrices;

$$\kappa_k(t) = e^{tD_k},\ 1 \le k \le p, \tag{3.39}$$

and the κ_k will also be called memory kernels in the sequel.

In more intuitive terms, z_1 is a stochastic process that results from the propagation of a red noise through the successive linear MSM layers, up to the level just preceding the main level, namely the y-equation in (*MSM*). Mathematically, the power spectrum of z_1 is red and its statistics are Gaussian. The interest of system (3.37) is that it facilitates the simulation of such a "reddish" noise. This reddish noise is then convoluted with the possibly nonlinear flow associated with $\dot{y} = f_2(y)$ to give rise to the stochastic process $\xi(t,\omega)$, which may therewith be non-Gaussian.

We can now summarize the above analysis in the following proposition.

Proposition 3.3. *Under the assumptions of Theorem 3.1, and if the conditions (3.26), (3.29) and (3.35) are satisfied, then any solution of (MSM) emanating from $(x_0, 0, 0)$ satisfies the following RDE with retarded arguments,*

$$\boxed{\frac{\mathrm{d}x}{\mathrm{d}t} = \overbrace{f_1(x)}^{(a)} + \overbrace{\mathbb{G}(t, x_t)}^{(b)} + \overbrace{\xi(t,\omega)}^{(c)},} \tag{RDDE}$$

where

$$\xi(t,\omega) = C \int_0^t S(t-s)\Pi z_1(s)\mathrm{d}s, \tag{3.40}$$

z_1 solves (3.37), and $S(t)$ is the flow generated by the vector field f_2; moreover,

$$\mathbb{G}(t, x_t) = \int_0^t S(t-s) g_2(x(s))\mathrm{d}s + C\Pi \int_0^t e^{-(t-s)D}\tilde{h}(x(s))\mathrm{d}s + \beta(x(t), y(t)), \tag{3.41}$$

with $y(t)$ given by (3.28).

Equation (*RDDE*) shows then that an MSM — with cross-interactions between the observed and hidden variables subject to the conditions of Proposition 3.3 — decomposes the dynamics of the observed variables into (a) nonlinear self-interactions, embedded in (c) a reddish background, plus (b) *state-dependent correction terms* represented by convolution integrals. The reddish background (c) stands for the cross-interactions and self-interactions that are not accounted for, respectively, by the convolution terms in (b) and by the nonlinear terms in (a). This noise term also accounts for the lack of knowledge of the full initial state due to partial observations. Stated otherwise, the deterministic terms in (a) provide a Markovian contribution to $\mathrm{d}x/\mathrm{d}t$, while the terms in (b) constitute a non-Markovian contribution *via* the past history of x, as t evolves, and the (c)-term represents the fluctuations that are not modeled by the terms in (a) and (b); the (c) term also accounts for the uncertainty in the knowledge of the full initial state, due to partial observations [43].

We have thus clarified that any system (*MSM*), subject to conditions like those of Proposition 3.3, resembles the closure models derived by applying the MZ formalism; see [42, Eq. (6)] and [42, 95, 96] for further details. A discrete formulation of an MSM — such as proposed by the original formulation of an EMR in [18, 26] — presents, nevertheless, substantial advantages in practice, since such a discrete system is in general much easier to integrate numerically than the system of stochastic integro-differential equations (*RDDE*). This is particularly true in instances where the memory kernel μ_k in (3.38) decay slowly, in the sense that the observed variables x evolve on a time scale comparable to that of the decorrelation time of such kernels.[9] Such a situation is expected to occur when the separation of time scales between the observed and unobserved variables is not as sharp as required, for instance, in applying stochastic homogenization techniques [100]; see however the recent works [101, 102] for milder assumptions.

The observation that an MSM can be recast into a system of stochastic integro-differential equations that resembles the GLE of the MZ formalism raises a natural question with respect to our closure problem ($\mathfrak{P}$) formulated in Section 2, namely to which extent does an MSM provide a good approximation of the GLE predicted by the MZ formalism? The next section sets up the mathematical framework to address this problem, followed by the formulation of a simple correlation-based criterion in Sec. 5 to solve it.

Our criterion is based on Proposition 3.3 and, in particular, on the representation of $\xi(t,\omega)$ provided by (3.40); it helps provide information on the degree of approximation of the GLE by a given MSM, in terms of the correlation between the residual noise term (c) in the (*RDDE*) and the observed variables $x(t)$. Such a criterion will turn out to be useful in applications, given the fact that the GLE constitutes the optimal closure model that can be achieved from an infinite time series of partial observations, as pointed out by Lemma 4.1 below.

4. Generalized Langevin equation (GLE) for optimal closure from a time series

In this section, we assume that the scalar field $u(t_k,\xi_j)$ in d dimensions, as expanded in (2.2), is the spatial coarse-graining of a discrete field $\mathcal{U}_{\mathcal{I}} := \{u(t_k,\xi_j)\}_{j\in\mathcal{I}}$ in n dimensions. Typically,

[9] In such a case, the direct numerical integration of (*RDDE*) becomes prohibitive, due to the cost of evaluating the integral terms. Such numerical difficulties have limited, so far, the application of the MZ formalism in the derivation of efficient closure models of partially observed high-dimensional systems for which the determination of $\mathbb{G}$ becomes a non-trivial task in the case of the so-called intermediate-range memory effects; see [97]. Nevertheless, in the case of long-range memory effects, the MZ framework was successfully applied to the Euler or Burgers equations by deriving a so-called t-model [98, 99].

card($\mathcal{I}$)=$n \gg d$ and an expansion similar to (2.2) holds in n dimensions, that is:

$$u(t_k, \xi_j) = \sum_{i=1}^{n} y_i(t_k) E_i(\xi_j), \quad j \in \mathcal{I},\, k = 1, \ldots, N. \tag{4.1}$$

We also assume that the evolution of y is governed by the ODE system

$$\dot{y} = \mathbf{R}(y). \tag{4.2}$$

The goal of this section is to show how the GLE from the MZ formalism can be theoretically derived from a time series. This mathematical derivation lays the foundations for seeking the Markovian part of an MSM by regression methods in practice, as discussed and applied in Secs. 5–7.

To simplify the presentation, we assume that the vector field $\mathbf{R}$ is *continuously differentiable* on $Y := \mathbb{R}^n$, endowed with the basis $\mathcal{B}' = \{E_i : i = 1, \ldots, n\}$, such that the flow $\{T_t\}$ associated with it is well-defined for all t and possesses a compact global attractor $\mathcal{A} \subset Y$ [103]. We assume, furthermore, that $\{T_t\}$ possesses an invariant probability measure μ, which is *physically relevant* [84, 89], *i.e.*:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \varphi(T_t(y))\mathrm{d}t = \int_{\mathcal{A}} \varphi(y)\mathrm{d}\mu(y), \tag{4.3}$$

for almost all $y \in Y$ (in the sense of Lebesgue measure) and for any observable $\varphi \in L^1_\mu(Y)$. Recall that, like all measures invariant under T_t, an invariant measure that satisfies (4.3) is supported by the global attractor $\mathcal{A}$; see, for instance, [50, Lemma 5.1].

We briefly outline now how the MZ formalism helps derive an abstract *closure model* for the description of the dynamics of $x = Py$, where P is the projection from Y onto $X := \mathbb{R}^d$ and the latter is endowed with the basis $\mathcal{B} = \{E_i : i = 1, \ldots, n\}$, $d \ll n$. By a closure model we mean a model that describes the evolution of $x(t) = Py(t)$ as a function of the x-variable only, plus some possible forcing terms that model the information lost by applying P, *i.e.*, due to the partial character of the available observations. Our approach thus follows [42, 95], who rely on the use of conditional expectation with respect to a meaningful invariant measure such as the μ introduced above. This framework will allow us, further below, to prove the main result of this section, formulated as Lemma 4.1. Practical consequences of this Lemma for MSMs are discussed in Section 5 below.

For convenience, we denote by $v : Y \to X$ the projection P so that, in particular,

$$v_i(y) := (Py)_i\,, \quad i = 1, \ldots, d.$$

By differentiating $v_i(T_t(y))$ with respect to time, we obtain

$$\begin{aligned} \frac{\partial}{\partial t} v_i(T_t(y)) &= \nabla_y (v_i(T_t y)) \cdot \frac{\mathrm{d}T_t(y)}{\mathrm{d}t} \\ &= \nabla_y (v_i(T_t y)) \cdot \mathbf{R}(T_t(y)) = \mathcal{L} v_i(T_t(y)); \end{aligned} \tag{4.4}$$

here $\mathcal{L}$ is the Lie derivative, acting on continuously differentiable functions h, along the vector field $\mathbf{R}$ given by:

$$\mathcal{L}h(y) := \nabla h(y) \cdot \mathbf{R}(y), \tag{4.5}$$

Note that, by definition, $v_i(T_t(y_0))$ is just the i^{th}-component $x_i(t)$ of $x(t; y_0)$, namely the projection onto X of the solution of (4.2) emanating from $y_0 \in Y$.

By introducing — at this point only formally — the time-dependent family of *Koopman operators*[10]:

$$(U_t u)(y_0) := u(T_t y_0),\ y_0 \in Y, \tag{4.6}$$

defined for observables $u : Y \to X$ that live in some appropriate functional space[11], the evolution of $x(t; y_0) = v(T_t y_0)$ is governed by the following system of linear, hyperbolic PDEs with non-constant coefficients in $n + 1$ variables:

$$\frac{\partial}{\partial t} U_t v(y_0) = \mathfrak{L}(U_t v)(y_0),\ y_0 \in Y, \tag{4.7}$$

where $\mathfrak{L}\Psi := (\mathcal{L}\Psi_1, \ldots, \mathcal{L}\Psi_n)$, with $\mathcal{L}$ given by (4.5). Within the appropriate functional setting, it can be shown that $\{U_t\}_{t\geq 0}$ forms a genuine semigroup, *i.e.* it describes not only the dynamics of v but of other observables as well, and its generator is given by $\mathfrak{L}$, *i.e.* $U_t = e^{t\mathfrak{L}}$. In other words, $\mathfrak{L}$ can be interpreted as the "rate of change" of U_t; see *e.g.* [107, Chap. 7.6], or [110] for a more advanced treatment in the dual case of Perron-Frobenius semigroups.

The Liouville-type equation (4.7) can be rewritten in a more convenient form for our purposes; in particular, we will see that the existence of an invariant measure μ satisfying (4.3) plays an essential role in connecting the MSMs introduced in the previous section with the MZ formalism. Let Z be the complement of X in Y, i.e.

$$Y = X \oplus Z, \tag{4.8}$$

and let $\Psi : Y \to X$ be a continuous function. The decomposition of Y allows us to split any $y \in Y$ as the sum $x + z$, with $x \in X$ and $z \in Z$; here both x and z are uniquely determined by the projection $v : Y \to X$ and its complementary projection, $\mathrm{Id}_Y - v$, respectively.

The existence of an invariant measure μ allows us next to define the *conditional expectation*[12] of Ψ for each $v(x)$,

$$\mathbb{E}[\Psi|v](x) := \int_Z \Psi(y) \mathrm{d}\mu_x(z). \tag{4.9}$$

Here μ_x is the probability measure on the unobserved factor space Z obtained by disintegration of μ above x, *i.e.*, for all Borel sets B and F of Z and X, respectively, we have that:

$$\mu(B \times F) = \int_F \mu_x(B) \mathrm{d}\mathfrak{m}(x), \tag{4.10}$$

where $\mathfrak{m}$ is the *push-forward* of the measure μ by v, *i.e.* $\mathfrak{m}(F) = \mu(v^{-1}(F))$, for any Borel set F of X.

[10]For further details about Koopman semigroups and operators, see [104, 105, 106, 107] and references therein. For this article, it suffices to note that U_t describes the action of the flow T_t on observables $u : Y \to X$.

[11]Such a space could be chosen to be for instance $D_p = \{u \in L^p_\mu(Y; X) \mid Au := \lim_{t\to 0}\ t^{-1}(U_t u - u) \text{ exists}\}$ for some $p \in [1, \infty]$, where the limit is taken in the sense of strong convergence [108, 109]. We do not consider in this article the delicate question of the choice of functional spaces that characterize the mixing properties of the flow T_t. Such considerations require typically spaces that take into account the stable and unstable manifolds of the attractor, when the latter supports a Sinai-Bowen-Ruelle measure; see, for instance, [110] in the case of Anosov flows. For a rigorous treatment in the context of Hamiltonian systems, see [111].

[12]Note that $\int_Z \Psi(y) \mathrm{d}\mu_x(z)$ in Eq. (4.9) is finite by assuming that Ψ continuous, since the support of the invariant measure μ, $\mathrm{supp}(\mu)$ is compact. Actually, by the Fubini theorem [112], it suffices to assume that $\Psi \in L^1_\mu(Y; X)$ for this integral to be well defined.

The existence of μ_x such that (4.10) holds is ensured by the *disintegration theorem of probability measures*; see for instance [113, Section 10.2, pp. 341–351] or [114, Chapter 5][13]. The probability measure μ_x can be interpreted as reflecting the statistics of the unobserved variables z when x has been observed; see [76] for further details.

Remark 4. *From the definitions of $\mathfrak{L}$ and v, we see that $\mathfrak{L}v$ corresponds to the vector field*

$$\begin{aligned} v \circ \mathbf{R} : X &\to X, \\ y &\mapsto (R_1(y), ..., R_d(y)), \end{aligned} \tag{4.11}$$

where $\mathbf{R}$ is the vector field governing the evolution of the full system (4.2). Thus $\mathfrak{L}v$ is but the first d components of the vector field $\mathbf{R}$ and (4.7) corresponds then to the functional formulation of

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v(\mathbf{R}(x+z)),\ x \in X,\ z \in Z, \tag{4.12}$$

where $Y = X \oplus Z$.

The truncated vector field $v \circ \mathbf{R}$ is still a function of y and, in particular, it depends on the unobserved variables z. It is then reasonable to seek the vector field in X that best approximates $\mathfrak{L}v$ in a least-square sense, as weighted by the invariant measure μ, i.e. in $L^2_\mu(Y;X)$. In other words, we seek an X-valued function of the observed, d-dimensional x only, which best approximates the X-valued function $\mathfrak{L}v$ of the full, n-dimensional y. It is exactly this approximation that provides the conditional expectation $\mathbb{E}[\mathfrak{L}v|v]$ corresponding to the vector field

$$\overline{v \circ \mathbf{R}}(x) := \int_Z v(\mathbf{R}(x+z))\mathrm{d}\mu_x(z),\ x \in X,\ z \in Z. \tag{4.13}$$

The averaging with respect to the unobserved variables z that occurs in Eqs. (4.9) and (4.13) becomes therewith intuitively clear.

Recalling now the stated purpose of deriving a closure model for $x(t)$, Remark 4 above leads naturally to decompose $v \circ \mathbf{R}$ into its averaged part, given by (4.13), and a fluctuating part,

$$v(\mathbf{R}(x+z)) = \overline{v \circ \mathbf{R}}(x) + (v(\mathbf{R}(x+z)) - \overline{v \circ \mathbf{R}}(x)). \tag{4.14}$$

The parametrization of the fluctuating part $v(\mathbf{R}(x+z)) - \overline{v \circ \mathbf{R}}(x)$ as a function of the x-variable is at the core of the MZ-formalism. This task is achieved through the perturbation theory of semigroups [109]. Leading up to this task, let us first note that, since $\mathfrak{L}$ and U_t commute, $\mathfrak{L}U_t = U_t\mathfrak{L}$ [42], we can rewrite (4.7) as

$$\frac{\partial}{\partial t}U_t v = U_t \mathfrak{L} v;$$

this, in turn, can be rewritten as:

$$\frac{\partial}{\partial t}U_t v = U_t \mathbb{E}[\mathfrak{L}v|v] + U_t(\mathfrak{L}v - \mathbb{E}[\mathfrak{L}v|v]). \tag{4.15}$$

Equation (4.15) is the functional formulation of (4.12), once (4.14) has been applied. The interest of this formulation is that the *variation-of-constants formula*, in its Miyadera-Voigt form [109, Section 3c][14], yields a decomposition of U_t into two terms; as we will see, these two terms have useful interpretations in statistical mechanics [116, 117, 118, 119, 120].

[13] See also [90, §4] and [88, 89, 115] for disintegration of probability measures arising in RDS theory; in this theory, the probability measure $\mathfrak{m}$ on the right-hand side of (4.10) is typically replaced by the probability measure $\mathbb{P}$ associated with the driving system.

[14] The Miyadera-Voigt form of the variation-of-constants formula is also known as the Dyson formula in the physics literature [42].

The decomposition of U_t is achieved by considering the operator B, acting on appropriate observables $\Psi : Y \to X$,

$$B\Psi := \mathbb{E}[\mathfrak{L}\Psi|v], \tag{4.16}$$

as a perturbation of the operator $\mathfrak{L}$, and writing

$$A\Psi = \mathfrak{L}\Psi - \mathbb{E}[\mathfrak{L}\Psi|v]. \tag{4.17}$$

Using the notation V_t for the semigroup generated by A, the above-mentioned variation-of-constants formula leads (formally) to:

$$U_t\Psi = V_t\Psi + \int_0^t U_{t-s}BV_s\Psi \mathrm{d}s. \tag{4.18}$$

It follows that the evolution of an observable $\Psi : Y \to X$ under V_s in (4.18) is described by the following PDE in $n+1$ variables:

$$\frac{\partial}{\partial t}V_t\Psi = A\Psi = \mathfrak{L}V_t\Psi - \mathbb{E}[\mathfrak{L}V_t\Psi|v]. \tag{4.19}$$

Remark 5. *Note that*

$$\frac{\partial}{\partial t}\mathbb{E}[V_t\Psi|v] = \mathbb{E}[\frac{\partial}{\partial t}(V_t\Psi)|v] = \mathbb{E}\Big[\mathfrak{L}V_t\Psi - \mathbb{E}\big[\mathfrak{L}V_t\Psi|v\big]\Big|v\Big] = 0, \tag{4.20}$$

so that, if Ψ is orthogonal to the space spanned by functions of $X = Im(v)$ alone, then $\mathbb{E}[V_t\Psi|v] = 0$ for all t. For this reason, equation (4.19) is known as providing the orthogonal dynamics in the MZ terminology.

According to this remark, by taking

$$\Psi = \mathfrak{L}v - \mathbb{E}[\mathfrak{L}v|v], \tag{4.21}$$

we have thus that $V_s\Psi$ in (4.18) evolves in the subspace that is orthogonal to the space spanned by functions of X. The term $BV_s\Psi$ corresponds to the average vector field on X given by $\overline{(V_s\Psi)} \circ \mathbf{R}$, where the average is taken over Z as in (4.13), so that $\int_0^t U_{t-s}BV_s\Psi \mathrm{d}s$ is a function that depends on $x \in X$ only. If x is given by the time-dependent function $v(T_ty_0)$, then $\int_0^t U_{t-s}BV_s\Psi \mathrm{d}s$ is a function of the past values of x, i.e., a memory term.

By applying (4.18) with $\Psi = \mathfrak{L}v - \mathbb{E}[\mathfrak{L}v|v]$ and using the expression of B given in (4.16), the term $U_t([\mathfrak{L}v - \mathbb{E}[\mathfrak{L}v|v])$ in (4.15) may be rewritten accordingly, and we arrive thus at the following *generalized Langevin equation (GLE)* :

$$\boxed{\frac{\partial}{\partial t}U_tv = \mathcal{R}(U_tv) + \int_0^t U_{t-s}\mathcal{G}(v;\eta_s)\mathrm{d}s + \eta_t.} \tag{GLE}$$

Here

$$\mathcal{R}(v) = \mathbb{E}[\mathfrak{L}v|v], \quad \mathcal{G}(v;\eta_s) = \mathbb{E}\big[\mathfrak{L}\eta_s\big|v\big], \tag{4.22}$$

and

$$\eta_t = V_t(\mathfrak{L}v - \mathbb{E}[\mathfrak{L}v|v]); \tag{4.23}$$

see [42] for further details. As explained above, the integral term in (*GLE*) constitutes the non-Markovian contribution to the evolution of U_tv, and thus of $v(T_ty_0)$; this contribution depends on U_s for the past interval $0 \le s \le t$.

The term η_t is a source of fluctuations related to the partial knowledge x_0 of the full initial state $y_0 = x_0 + z_0$. Recall that the conditional expectation $\mathbb{E}[\Psi|v]$, as defined in (4.9), of a function Ψ in $L^2_\mu(Y;X)$ can be seen as the orthogonal projection $\mathcal{P}$ onto the space of X-valued

functions that depend only on the x-variable; see Remark 6 below. Thus, by introducing the complementary projector

$$\mathcal{Q} = (\mathrm{Id}_{L^2_\mu(Y;X)} - \mathcal{P}), \tag{4.24}$$

we obtain from (4.19) that η_t solves the following initial value problem:

$$\begin{cases} \dfrac{\partial}{\partial t}\eta_t = \mathcal{Q}\mathfrak{L}\eta_t \quad \text{in } Y, \\ \eta_0(x_0, z_0) = v \circ \mathbf{R}(x_0 + z_0) - \overline{v \circ \mathbf{R}}(x_0), \quad x_0 \in X,\ z_0 \in Z. \end{cases} \tag{4.25}$$

In other words, η_t gives the evolution in time, according to the *orthogonal dynamics*, of a deviation $\eta_0(x_0, z_0) := v \circ \mathbf{R}(x_0 + z_0) - \overline{v \circ \mathbf{R}}(x_0)$ at time $t = 0$, with respect to the conditional expectation (4.13).

Since z_0 is distributed according to the disintegrated measure μ_{x_0}, the fluctuating term η_t can be then interpreted as an X-valued random variable. Since $\mathbb{E}[\eta_0|v] = 0$, we deduce from Remark 5 that $\mathbb{E}[\eta_t|v] = 0$ for all $t > 0$, *i.e.* the conditional expectation of η_t remains zero as t evolves, so that η_t is uncorrelated with any function of v [43, 121].

Remark 6. *Recall that we may regard $\mathbb{E}[\Psi|v]$ as the orthogonal projection of Ψ belonging to*

$$L^2_\mu(Y;X) := \{g : Y \to X,\ \text{measurable and such that } \int_Y \|g(y)\|^2 \mathrm{d}\mu(y) < \infty\}, \tag{4.26}$$

onto the space of functions of X, since for all function $f : X \to X$ such that $f \circ v \in L^2_\mu(Y;X)$,

$$\mathbb{E}[\|\Psi - \mathbb{E}[\Psi|v]\|^2] \leq \mathbb{E}[\|\Psi - f \circ v\|^2], \tag{4.27}$$

where the expectation $\mathbb{E}(\Phi)$ is taken here with respect to μ, that is:

$$\mathbb{E}(\Phi) = \int_{\mathcal{A}} \Phi(y)\mathrm{d}\mu(y),\ \Phi \in L^1_\mu(Y). \tag{4.28}$$

Equations (*GLE*) are fewer in number than in (4.2), i.e. $d \ll n$, but this advantage is outweighed by the need to find the fluctuating term η_t as a solution of the orthogonal dynamics (4.25), along with its requisite statistical properties, in order to simulate (*GLE*) accordingly. What equation (*GLE*) does provide is a theoretical "master equation," which can serve as a basis for the design of various practical closure strategies.

With this purpose in mind, the following fundamental lemma shows that actually the reduced, Markovian vector field $\mathcal{R}(v)$ in (*GLE*) can, in principle, be approximated from a time series, when the latter represents partial observations drawn from a physical invariant measure.

Lemma 4.1. *Assume that the main system (4.2) possesses an invariant measure μ that satisfies the "physicality condition" of (4.3). Let $v : Y \to X$ denote the projection onto $X = \mathrm{span}\{E_1, ..., E_d\}$, and let $\mathcal{E}(v, \mu)$ be the closure in $L^2_\mu(Y;X)$ of the set of functions $f : X \to X$ such that $f \circ v \in L^2_\mu(Y;X)$.*

Then, for almost all initial data $y_0 \in Y$, in the sense of Lebesgue measure on Y,

$$\boxed{\underset{f \in \mathcal{E}(v,\mu)}{\mathrm{argmin}} \Big(\lim_{T\to\infty} \frac{1}{T} \int_0^T \Big\| \frac{\mathrm{d}x}{\mathrm{d}t} - f(x(t;y_0)) \Big\|^2 \mathrm{d}t \Big) = \mathbb{E}[\mathfrak{L}v|v]} \tag{4.29}$$

holds, where $\mathbb{E}[\mathfrak{L}v|v]$ is the conditional expectation defined in (4.9), while $x(t;y_0) := v(y(t;y_0))$ with $y(t;y_0)$ denoting the solution $y(t;y_0)$ of (4.2) emanating from y_0.

Proof. The proof uses the two facts that (i) μ is a physical invariant measure in the sense of (4.3), and that (ii) $\mathbb{E}[\mathcal{L}v|v]$ is the projection, in $L^2_\mu(Y;X)$, of the rate of change of $U_t v$ onto X. Let $\varphi : Y \to \mathbb{R}$ denote the observable defined by $\varphi(y) = \|\mathfrak{L}v(y) - f(v(y))\|^2$.

Note that the functional φ thus defined lives in $L^1_\mu(\mathcal{A})$. To see this, first note that — from (4.5) and the definition of $\mathfrak{L}$ — it is not difficult to show that there exists $C > 0$ such that

$$\|\mathfrak{L}v(y)\| \leq C\|\mathbf{R}(y)\|, \text{ for all } y \in \mathcal{A}. \tag{4.30}$$

This bound, in turn, implies that $\mathfrak{L}v$ lives in $L^2_\mu(Y;X)$ and, since $\mathbf{R}$ is continuous, the global attractor $\mathcal{A}$ is compact [103, Def. 1.3.] and the support of μ is contained in $\mathcal{A}$; see *e.g.* [50, Lemma 5.1] for this latter point. Combined with the assumption on f, we deduce that $\varphi \in L^1_\mu(\mathcal{A})$ and that it constitutes therewith an admissible observable.

Applying (4.3) along with the definition of the Koopman operator in (4.6) yields

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \|\mathfrak{L}U_t v(y) - f(U_t v(y))\|^2 \mathrm{d}t = \int_{\mathcal{A}} \|\mathfrak{L}v(y) - f(v(y))\|^2 \mathrm{d}\mu(y), \tag{4.31}$$

for almost all $y \in Y$. From (4.7) and (4.28), we obtain next that, for almost all $y \in Y$,

$$\begin{aligned}\inf_{f\in\mathcal{E}(v,\mu)}\Big(\lim_{T\to\infty}\frac{1}{T}\int_0^T \|\frac{\mathrm{d}x}{\mathrm{d}t} - f(x(t;y))\|^2\mathrm{d}t\Big) &= \inf_{f\in\mathcal{E}(v,\mu)}\Big(\int_{\mathcal{A}} \|\mathfrak{L}v(y) - f(v(y))\|^2\mathrm{d}\mu(y)\Big),\\ &= \mathbb{E}\Big[\big\|\mathfrak{L}v - \mathbb{E}\big[\mathfrak{L}v|v\big]\big\|^2\Big],\end{aligned} \tag{4.32}$$

where (4.27) has been used to get the last equality.

Finally, by applying in the Hilbert space $L^2_\mu(Y;X)$ the classical projection theorem onto a closed convex set [108, Theorem 5.2], we conclude that $\mathbb{E}\big[\mathfrak{L}v\big|v\big]$ is the unique minimizer of

$$\int_{\mathcal{A}} \|\mathfrak{L}v(y) - f(v(y))\|^2, \tag{4.33}$$

and, therefore, of

$$\lim_{T\to\infty}(1/T)\int_0^T \|\mathfrak{L}U_t v(y) - f(U_t v(y))\|^2\mathrm{d}t, \tag{4.34}$$

when $f \in \mathcal{E}(v,\mu)$. Formula (4.29) is thus proved. □

Remark 7. *One may want to assume the existence of a bounded non-wandering set Λ[76], instead of a global attractor $\mathcal{A}$, for the system (4.2) and thus relax condition (4.3) to hold only for y in the system's basin of attraction $\mathcal{B}(\Lambda)$. In this more general case, the conclusion of Lemma 4.1 still holds for $y_0 \in \mathcal{B}(\Lambda)$, i.e., for time series that relax towards the corresponding statistical equilibrium μ whose support is contained in Λ. MSMs can still be efficiently derived in such a context as shown in Sec. 7.*

5. MSMs as closure models from time series: the η-test

A classical approach in MZ modeling consists of assuming something reasonable about the statistics of the unobserved variables — *e.g.*, on the basis of previous observations — and then turn tothe formulation of Eq. (*GLE*) to determine an analytic or numerical approximation of the terms that appear on the right-hand side; the resulting prediction methods based on the MZ formalism go under the name of *optimal prediction* [42]. The main objective in such an approach is to calculate the conditional expectation based on the assumptions made regarding the statistics of the unobserved variables. When no analytical formula is available for the probability measure that describes the distribution of the unobserved variables, empirical estimation methods are

typically used; these empirical estimations often rely on a large set of trajectories integrated over a short time interval.

For instance, maximum likelihood estimation techniques [122] can be used to find an approximation ν of the density of the unobserved variables as Gaussian mixtures; see [97] for an application of such techniques to the non-Hamiltonian case of the Kuramoto-Sivashinsky equation. In certain cases, the Markovian term in (*GLE*) can be computed explicitly by relying on the estimated ν, while other assumptions — such as the *short-memory approximation* or the *t-model* [98, 99] — can then be used to deal with the non-Markovian term in (*GLE*) through simulations of the full system; see [43, 97].

As mentioned already in Sec. 1.2, the difference in viewpoint between this classical approach and an approach based on averaging along trajectories, such as supported by Lemma 4.1, is similar to the Eulerian versus the Lagrangian viewpoint in fluid mechanics. In this analogy, the approach advocated in this article corresponds to the Lagrangian viewpoint: measurements are made along trajectories and the numerical construction of an MZ model relies on fewer initial states, but requires longer runtimes, for large models, or longer data records, for observational data. Recalling that the MZ formalism is built on the decomposition of the Koopman semigroup given in (4.18), an approach based on averaging along trajectories is furthermore consistent with other Koopman operator techniques developed recently for the spectral analysis of time series [104].

The MSM approach to stochastic inverse modeling, as described above, can thus provide an efficient way of deriving approximation of (*GLE*) in practice, by relying exclusively on available time series, even when no prior knowledge about the full model is available. A key point to achieving this, based on available time series alone, is the quality of the approximation by the vector field f_1 in (*RDDE*) of the genuine Markovian contribution in (*GLE*), on the one hand, and the quality of approximation by the terms collectively labeled (b) in Eq. (*RDDE*) of the non-Markovian contribution $\int_0^t U_{t-s}\mathcal{G}(v;\eta_s)\mathrm{d}s$ in (*GLE*), on the other. Recall that the former model the self-interactions among the observed variables and that the latter model the cross-interactions between the observed and unobserved variables that occurred at the past times s, $0 \leq s \leq t$.

A priori error estimates that may be useful in practice are difficult to establish at this level of generality. However, the remaining term (c) in Eq. (*RDDE*) — when compared to the "residue" η_t in (*GLE*) — can serve to formulate a correlation-based criterion to help determine how well an MSM given by (*RDDE*) approximates (*GLE*). Indeed, by construction of the GLE, the noise term η_t is uncorrelated with the observed time series $x(t)$, due to the orthogonality property of the dynamics in (4.25); see also Remark 5. Therefore, the corresponding term (c) in (*RDDE*) can naturally serve for testing whether an MSM derived from the time series $x(t)$ alone provides a good approximation of the GLE.

The resulting *η-test* can be summarized as follows:

($\mathfrak{T}$) The more Pearson's correlation coefficient[15] between $\xi(t,\omega)$ in (*RDDE*) and the observed variable $x(t)$ is close to zero, the better the approximation of the GLE (*GLE*) associated with $x(t)$.

In particular, to assert that an MSM model constitutes a good approximation of the (*GLE*) associated with a given multivariate time series, the noise terms labeled (c) in Eq. (*RDDE*) should not exhibit any x-dependence, since the x-dependence of the fluctuating terms is supposed to be taken into account solely in the non-Markovian terms (b).

[15] This coefficient is defined as the covariance of the two variables divided by the product of their standard deviations.

Interestingly, the multilayer structure of an MSM provides a simple way to compute the (c)-term $\xi(t,\omega)$ of Eq. (*RDDE*), as provided by the representation formula (3.40) in Proposition 3.3. In that respect, the process z_1 can be easily simulated by integration of (3.37), followed by an integration of

$$\dot{\xi} = f_2(\xi) + \Pi z_1. \tag{5.1}$$

Doing so provides a natural estimation of $\xi(t,\omega)$, up to multiplication by the matrix C from (3.29), and thus allows for an easy estimation of Pearson's correlation coefficient in the η-test ($\mathfrak{T}$) formulated above.

It is important to keep in mind that the η-test only provides information on the x-dependences of this residue that are not fully captured by the (a)- and (b)-terms in (*RDDE*). Hence this test is more useful as an indicator in the design of the relevant constitutive parts of (*MSM*) such as the f_i's and the g_i's, $i = 1, 2$, rather than providing an ultimate criterion to assess the modeling performance of a given MSM.

Situations may indeed arise where the corresponding Pearson's correlation coefficient is not necessarily close to zero, while the corresponding MSM still performs very well in simulating the main statistical properties of the observed variables. In other words, although some x-dependences may not be fully resolved by the deterministic terms, both Markovian and non-Markovian, of an MSM, the contributions of these terms to simulating the main observed statistics could prove to be negligible.

Such a situation is identified in Sec. 6 below for a conceptual stochastic climate model in the presence of weak time scale separation; see panels (g) and (h) of Figs. 2 and 3, and panel (d) of Fig. 5 in the following section. At the same time, the quality of reproduction of the observed statistics is improved as Pearson's correlation coefficient gets closer to zero; see panels (a)–(f) of Figs. 2 and 3, and panels (a)–(c) of Fig. 5. Part of the reason for this success lies in the energy-conserving EMR formulation discussed at the end of Sec 3.1 and in Appendix C. Once adopted, this formulation makes the practical, discrete-time EMR consistent with the full model's structural features, as layed out in Sec. 6.

6. A conceptual stochastic climate model: Numerical results

6.1. **Model formulation.** We illustrate here the MSM approach to stochastic inverse modeling, as described in Secs. 2–5, by deriving a closure model from partial observations of a slow-fast system in which only the nominally slow variables are observed. The time-scale separation between the nominally slow and fast variables ranges from a strong to a weak separation; in the latter case, some of the slow and fast variables actually evolve on a similar time scale. This example will show, in particular, the usefulness of the η-test introduced in Sec. 5, as discussed at the end of this section. For simplicity and for the sake of reproducibility of the results, we use here a simple conceptual climate model proposed in [45]; see also [123, 124]. Three features of this four-dimensional model are of interest with respect to our closure problem ($\mathfrak{P}$) of Sec. 2. First, the model is stochastic, which introduces *de facto* noisy observations. Second, the variables that will be taken as unobserved, carry in fact most of the variance in this model; while this is not the case in actual observations of atmospheric low-frequency variability (LFV) [125, 126], it does present a challenging difficulty to the MSM approach. Third, the model exhibits natural energy-preserving constraints that will be taken into account in the MSM formulation below.

The model obeys the following system of SDEs:

$$\mathrm{d}x_1 = \{-x_2(L_{12} + a_1x_1 + a_2x_2) - d_1x_1 + F_1 + \mathbf{L_{13}y_1} + \mathbf{b_{123}x_2y_1} + \mathbf{c_{134}y_1y_2}\}\mathrm{d}t, \tag{6.1a}$$

$$\mathrm{d}x_2 = \{x_1(L_{21} + a_1x_1 + a_2x_2) - d_2x_2 + F_2 + \mathbf{L_{24}y_2} + \mathbf{b_{213}x_1y_1}\}\mathrm{d}t, \tag{6.1b}$$

$$\mathrm{d}y_1 = \{-L_{13}x_1 + b_{312}x_1x_2 + c_{341}y_2x_1 + F_3 - \frac{\gamma_1}{\epsilon}y_1\}\mathrm{d}t + \frac{\sigma_1}{\sqrt{\epsilon}}dW_1 \tag{6.1c}$$

$$\mathrm{d}y_2 = \{-L_{24}x_2 + c_{413}y_1x_2 + F_4 - \frac{\gamma_2}{\epsilon}y_2\}\mathrm{d}t + \frac{\sigma_2}{\sqrt{\epsilon}}dW_2. \tag{6.1d}$$

The parameter ϵ explicitly controls the time-scale separation between the model's slow and fast variables, namely the (x_1, x_2)-variables, and the (y_1, y_2)-variables, respectively. These variables are coupled linearly, through the skew-symmetric terms, as well as nonlinearly; the nonlinear coupling involves the triple coefficients b_{ijk} and c_{ijk}. The linear and nonlinear coupling can be understood as additive and multiplicative noise forcing the slow-mode evolution, respectively.

The model set-up here follows [123], namely:

$$\begin{aligned} b_{123} = b_{213} = 0.25,\ b_{312} = -0.5, \\ c_{134} = c_{341} = 0.25,\ c_{413} = -0.5; \end{aligned} \tag{6.2}$$

$$\begin{aligned} L_{12} = L_{21} = 1,\ L_{24} = -L_{13} = 1, \\ a_1 = -a_2 = 1.0,\ d_1 = 0.2,\ d_2 = 0.1; \end{aligned} \tag{6.3}$$

and

$$\begin{aligned} F_1 = -0.25,\ F_2 = F_3 = F_4 = 0; \\ \gamma_1 = \gamma_2 = 1, \text{ and } \sigma_1 = \sigma_2 = 1. \end{aligned} \tag{6.4}$$

Note that, in agreement with the energy-conserving constraints of the EMR formulation in Eqs. (C.1)–(C.5) of C, this toy model has a quadratic nonlinear part that conserves energy; for example, the triple coefficients sum to zero, i.e.,

$$b_{123} + b_{213} + b_{312} = 0 \text{ and } c_{134} + c_{341} + c_{413} = 0,$$

as in Eq. (C.3), while the linear part has pairwise skew-symmetric terms given by the values of the coefficients L_{12}, L_{21}, L_{24} and L_{13}, as in Eq. (C.4). Furthermore, the negative-definite contributions of γ_1, γ_2, d_1 and d_2, as in Eq. (C.5), ensure the model's dissipativity, cf. [86] and [127, Sec. 5.4].

Finally, certain quadratic terms are absent from the model, as per Eq. (C.1), while the values for a_1 and a_2 are set so that Eq. (C.2) is satisfied. The terms in bold characters in the first two equations will be discussed further below.

6.2. **Numerical results.** We integrated Eqs. (6.1a)–(6.1d) for 10^4 time units by using the fourth-order Runge–Kutta scheme for the deterministic part and the Euler-Maruyama scheme for the stochastic part, with a time step of $\Delta t = 0.001$. Only the slow model variables x_1 and x_2 are stored here, with a sampling rate of 0.05 time units. We applied an energy-preserving version of (*EMR*), by using the constraints of Eqs. (C.1)–(C.5), in order to model the corresponding multivariate time series of (x_1, x_2).

As the scale-separation parameter ϵ increases, and the scale separation decreases therewith in the model, the decorrelation times steadily increase for the fast modes and decrease for the slow modes; x_1 and y_1, in particular, exhibit the most pronounced changes; see Figs. 1(a)–(d). For $\epsilon = 1.5$, the autocorrelation functions for y_1 and x_1 become very similar, cf. Fig. 1(d), and so there is no longer any formal separation of scales. Note that, in this case, the main level of our resulting EMR model for evolving x_1 and x_2 does not include explicitly the unresolved,

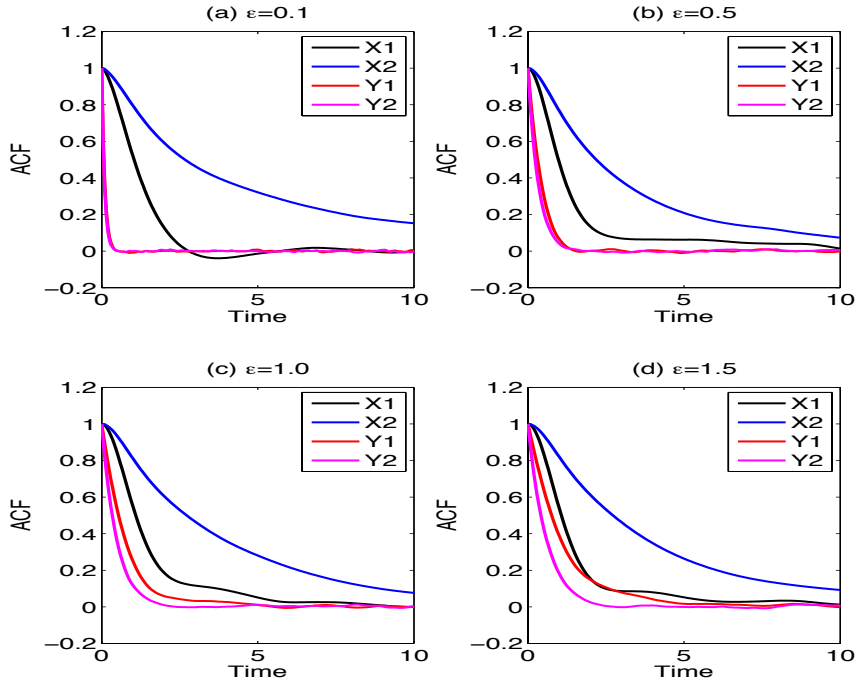

FIGURE 1. Autocorrelation functions of the model variables for Eq. (6.1). Panels (a)–(d) correspond to the ϵ-values $0.1, 0.5, 1.0$ and 1.5; see the color coding for the variables (x_1, x_2, y_1, y_2) in the legend of each panel.

linear and nonlinear interactions marked in bold in Eqs. (6.1a, 6.1b). These contributions to the dynamics of the observed variables (x_1, x_2), which result from their cross-interactions with the unobserved variables (y_1, y_2), need to be properly parameterized by the EMR model's hidden variables $\{r_t^{(m)} : 0 \le m \le p\}$, in order to reproduce the statistical behavior of (x_1, x_2) in terms of their PDFs and autocorrelations.

The energy-preserving EMR model, fitted solely on the x_1 and x_2 time series, has two additional levels ($p = 2$) for all values of ϵ, according to the stopping criterion of A. Figures 2 and 3 present a comparison of the one- and two-dimensional (1-D and 2-D) PDFs, respectively, for slow modes obtained by the energy-preserving EMR and the full model. The two figures show clearly that the energy-preserving EMR model reproduces quite accurately both the univariate and bivariate PDFs.

Figures 1 and 4 show how the autocorrelations and the variance, respectively, of the slow and fast variables change when the scale-separation parameter ϵ varies between 0.1 and 1.5. For $\epsilon = 0.1$ (black line in Fig. 4), most of the variance is carried by the fast modes, whose decorrelation time is much smaller in this case than that of the slow modes, while x_2 has the slowest autocorrelation decay, cf. Fig. 1(a). For $\epsilon \ge 0.5$ (colored lines in Fig. 4), the variance in this model is distributed roughly equally between slow and fast modes, while observations of atmospheric LFV show the slow, barotropic modes, on the 10–100-day scale, to be in fact considerably more energetic than the synoptic-scale, baroclinic modes of 1–10 days [125, 126].

The 1-D PDF is mostly Gaussian for x_1 (left column of Fig. 2) and strongly non-Gaussian for x_2 (right column of Fig. 2); neither changes noticeably with scale separation. The 2-D PDFs in Fig. 3 are quite non-Gaussian, all four of them, and they do not change much in shape or orientation with ϵ either. Moreover, Fig. 5 shows that the energy-preserving EMR model

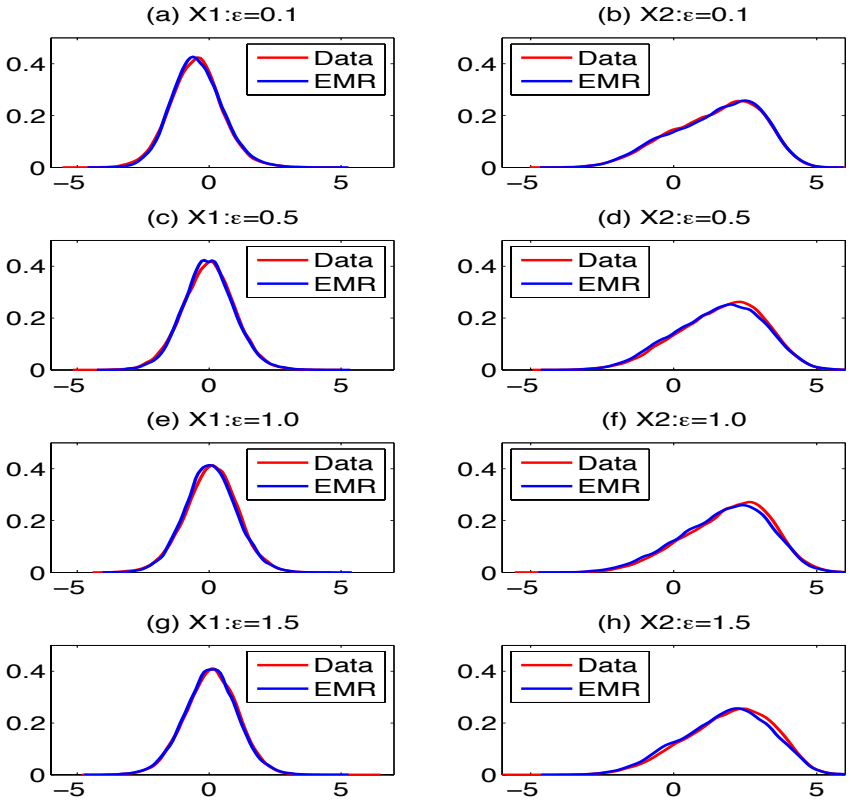

FIGURE 2. One-dimensional (1-D) probability density functions (PDFs) of the resolved variables (x_1, x_2), as modeled by the EMR-reduced model (blue line) *vs.* the simulation by the full coupled system in Eq. (6.1) (red line); the panels differ, from top to bottom, by the model's varying scale separation: $\epsilon = 0.1$ in panels (a) and (b); 0.5 in panels (c) and (d); 1.0 in panels (e) and (f); and 1.5 in panels (g) and (h). Left column (a, c, e, g) PDFs for x_1; and right column (b, d, f, h) PDFs for x_2.

reproduces with very high accuracy changes in the autocorrelation function of x_2, the model's slowest mode, for all values of ϵ. The energy-preserving EMR model does equally well for x_1, as long as the scale separation is sufficiently large, i.e. for $\epsilon = 0.1$ and 0.5; see Figs. 5(a, b). For $\epsilon = 1.0$ and 1.5, there is no pronounced scale separation in the full model, and thus there is not much difference in decorrelation time between the slow mode x_1 and the fast mode y_1. Hence it is not surprising that the autocorrelation for x_1 is reproduced somewhat less accurately, but still reasonably well, cf. Figs. 5(c,d).

In summary, in the partially observed situation studied in this section, a discrete-version of an MSM given by (*EMR*) and subject to the energy-preserving constraints described in C performs very well when the variability of the discarded variables is much faster than that of the slow variables. The EMR model performance is still remarkably good when the variability of the excluded variables is similar in amplitude and time scale to that of the retained variables.

Since the number of levels in the EMR model is $p = 2$, the total number of EMR variables is six, and thus it formally exceeds the total number of degrees of freedom of the full model Eq. (6.1), namely four. This is the price to pay, though, for successful orthogonal, multilevel

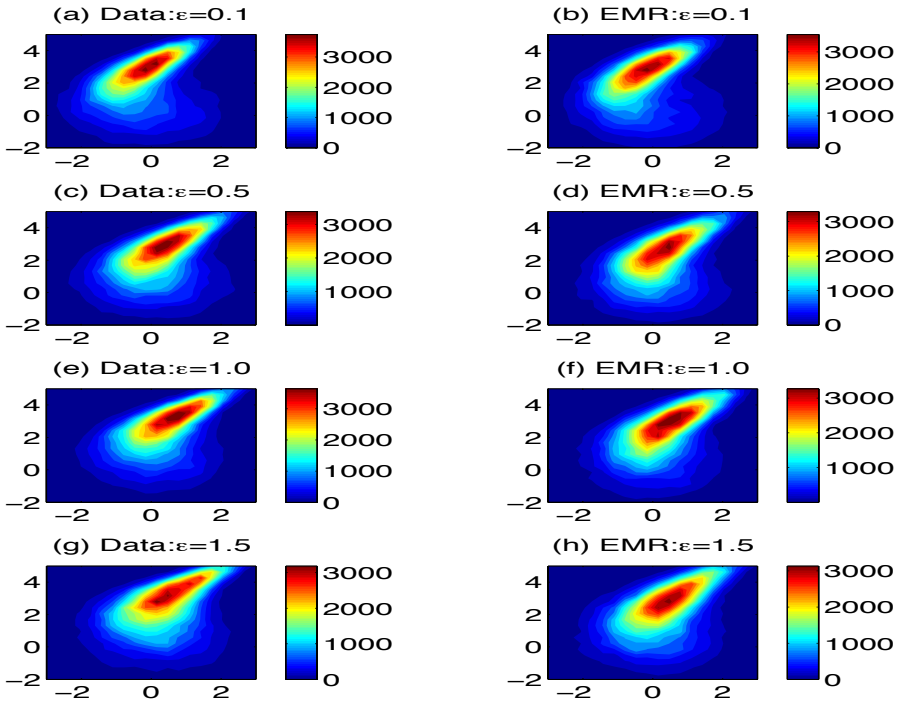

FIGURE 3. Same as Fig. 2, but for the two-dimensional (2-D) PDFs of the resolved variables (x_1, x_2), as simulated by the full coupled system of Eq. (6.1) (left column) *vs.* the 2-D PDFs of (x_1, x_2), as modeled by the EMR-reduced model (right column).

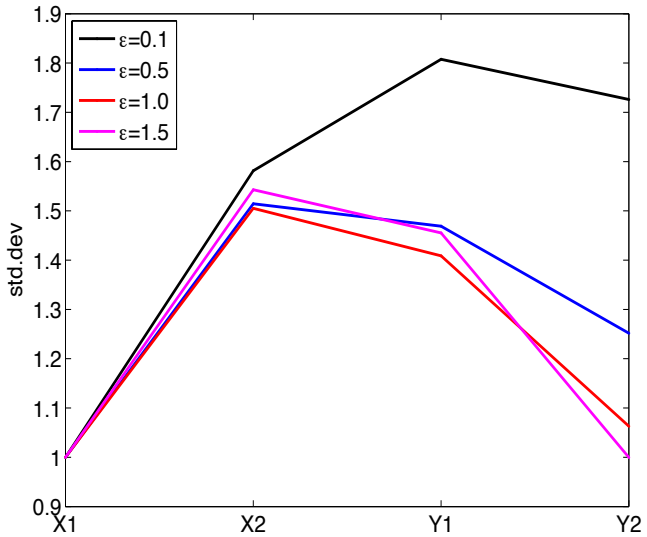

FIGURE 4. Standard deviation of the model variables of the toy model governed by Eq. (6.1), for different values of the scale-separation parameter, namely $\epsilon = 0.1, 0.5, 1.0$ and 1.5. The slow modes (x_1, x_2) and fast modes (y_1, y_2) are on the abscissa; see legend for color code.

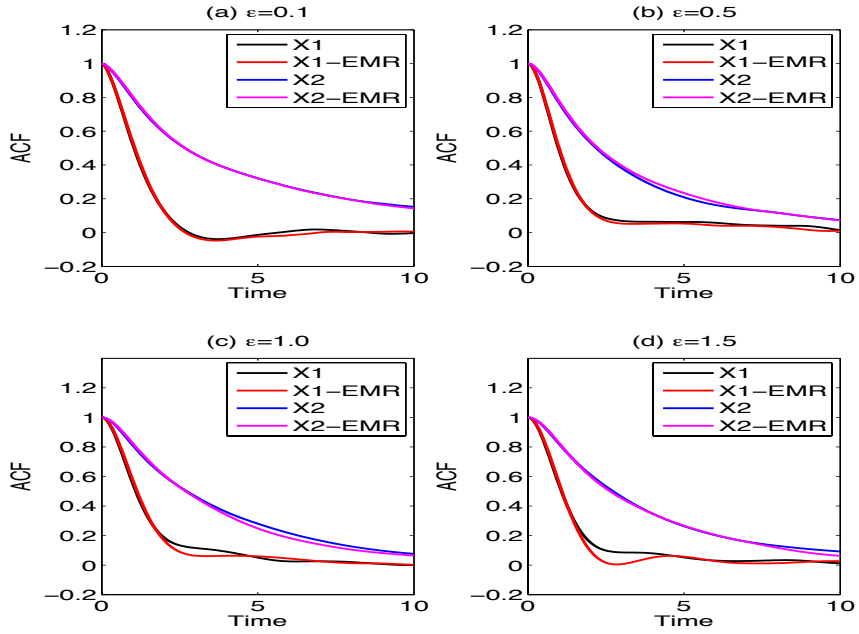

FIGURE 5. Autocorrelation functions of the resolved variables (x_1, x_2) in the full dynamics *vs.* the EMR-reduced dynamics, with varying scale separation $\epsilon = 0.1 - 1.5$ in panels (a)–(d). Color coding appears in each panel legend: the black and blue lines are for the full model's first and second component, while red and purple are for the EMR-reduced model's first and second component.

parameterization of the unresolved processes that were explicitly excluded from the main level of the reduced model.

As discussed in Section 5, it can reasonably be asserted that an MSM written under its form (*RDDE*) represents a good approximation of the GLE — i.e., the optimal closure model predicted by the MZ formalism — as long as the noise term labeled (c) in Eq. (*RDDE*) is weakly correlated in time with x; see the η-test ($\mathfrak{T}$) there.

For the model (6.1) at hand, the maximum absolute values of the corresponding component-wise Pearson's correlation coefficients are $0.11, 0.33, 0.42$ and 0.47, for $\epsilon = 0.1, 0.5, 1.0$ and 1.5, respectively. As a consequence, the η-test allows one to conclude that the more the time-scale separation is reduced, the more some of the (x_1, x_2)-dependences (in terms of correlations) are not well resolved by the Markovian and non-Markovian terms conveyed by our energy-preserving (*EMR*). However, as illustrated here, the contributions of such dependencies to a good reproduction of the main observed statistics may turn out in practice to be negligible. When that is the case, the resulting EMR model may still perform quite well; see panels (g) and (h) of Figs. 2 and 3, and panel (d) of Fig. 5.

Finally, as predicted by the η-test proposed in Sec. 5 here, the reproduction of the observed statistics is improved as Pearson's correlation coefficient gets closer to zero; see panels (a)–(f) of Figs. 2 and 3, and panels (a)–(c) of Fig. 5. Part of the reason for this success, in the case at hand, lies in the energy-conserving formulation discussed at the end of Sec. 3.1 and in C.

Recall, though, as stated already in Sec. 1.1, that conservation of quadratic invariants, like in Eq. (6.1) here, is the exception — rather than the rule — in full climate models that contain thermodynamic variables, precipitation and chemistry, let alone in other areas of the physical and life sciences. In the next section, we consider a population-dynamics model that does

contain quadratic coupling terms between variables but does not possess a quadratic energy to be conserved, in absence of dissipation.

7. Reflected diffusion processes by MSMs and reconstruction of strange attractors

7.1. **The original model.** Our second data-based closure example is the following classical population dynamics model:

$$\frac{\mathrm{d}N_i}{\mathrm{d}t} = b_i N_i\Big(1 - \sum_{j=1}^{n} a_{ij} N_j\Big), \quad 1 \le i \le n; \tag{7.1}$$

here N_i denotes the population size of the i^{th} species relative to its carrying capacity, $b_i > 0$ denotes its intrinsic growth rate, and the a_{ij}'s denote the interaction coefficients: intraspecific when $i = j$, and interspecific when $i \neq j$. We restrict ourselves to the case where $a_{ij} \geq 0$, which corresponds to the well-known *competitive Lotka-Volterra system* [53, 128]. Such systems manifest certain generic features that make their study even more interesting. For instance, they generate flows that — when restricted to the appropriate invariant sets — are topologically equivalent to a broad class of flows generated by systems of first-order ODEs with polynomial right-hand sides, such as the Lorenz system [129, 130].

This class of simple models exhibits a rich variety of dynamics, depending on the parameter values, as long as the initial vector $\mathbf{N}_0$ lies in $\mathbb{R}^n_+ := \{N_i : N_i \geq 0,\, i = 1 \ldots, n\}$, although solution blow-up is possible when a component of $\mathbf{N}_0$ is negative. Smale showed in [30] that — under the aforementioned conditions and for any initial data in the cone $\mathbb{R}^n_+$ — a Lotka-Volterra system (7.1) of five or more species ($n \geq 5$) can exhibit any asymptotic behavior, including steady states, limit cycles, n-tori, or more complicated attractors. This result has had a profound influence on the theory of monotone dynamical systems in showing that competitive systems could display more than simple dynamics; see [31, 32, 33, 34, 35].

In particular, the proof of Smale ensures the existence of a closed invariant set $\mathcal{C}$ which is homeomorphic to the $(n-1)$-dimensional simplex

$$\Delta_{n-1} = \{N_i : N_i \geq 0,\, \sum_{i=1}^{n} N_i = 1\}, \tag{7.2}$$

which is attracting every point (excluding the origin) in the domain $\mathbb{R}^n_+$; see [35, pp. 71-72]. The set $\mathcal{C}$ is known as the carrying simplex since it "carries" all of the asymptotic dynamics associated with (7.1), and its existence significantly reduces the set of possibilities in certain dimensions. For instance, the existence of such a simplex implies that an attractor associated with (7.1) cannot have a dimension greater than $n-1$ so that, in particular, chaos cannot take place when $n = 3$. This situation is obviously in stark contrast with that of three-dimensional ODE systems with quadratic energy-preserving nonlinearities, such as the standard Lorenz system [131, 127], although the dynamics of the latter is still realizable in a higher-dimensional Lokta-Volterra system; see [130, Sect. 4].

In fact, as numerically shown in [132], the smallest dimension for which complicated dynamics takes place on a strange attractor for (7.1) corresponds to $n = 4$. As already conjectured in [132], homoclinic tangencies are responsible for the structural instability of such strange attractors [133]; the latter can even experience sudden changes, and be brutally transformed into a steady state, after a small change in parameter values; see [134, Figs. 3 and 4]. Another interesting feature of (7.1) for $n = 4$ is the rarity of occurrence of chaotic behavior in a 20-dimensional parameter space; see [133] and [134, Fig. 2].

Furthermore, when chaos takes place for a particular set of parameters, the corresponding strange attractor may not attract all the points of $\mathbb{R}^4_+$; in the latter case, it typically coexists with simple local attractors, such as fixed points, and the attractor-basin boundaries are fractal, as in [134, Fig. 5]. Finally, for a solution evolving on a typical strange attractor, the auto-correlation function of each component decays at a nearly identical rate; see, for instance, the red curves in Fig. 6 below.

The aforementioned features — lack of time-scale separation; rarity of strange attractors in the parameter space; existence of fractal boundaries between attractor basins; and positivity of the solutions' components — greatly add to the difficulties in deriving a data-based stochastic closure model able to simulate faithfully the statistics of solutions that evolve on a strange attractor associated with (7.1) for $n = 4$; and especially so when only using a time series that represents partial observations of such a solution. The stumbling block of lack of time-scale separation has already been discussed. The combination of the two facts that (i) in general a time series of partial observations of solutions evolving on a strange attractor can be rigorously represented as a stochastic process that does depend on the unobserved variables (see Theorem A in [76] and Corollary B in its supporting information; and that (ii) chaos takes place over very small regions of the parameter space for (7.1) with $n = 4$, seriously hamper the learning of any inverse stochastic model for the statistics of such partial observations.

Difficulties in estimating the correct statistical behavior can already be observed in the case of full observations of a chaotic attractor [135]: when the corresponding four-dimensional time series is corrupted by a Gaussian noise, the accuracy of the parameter values — as estimated by simple regression — starts to degrade as the data become noisier [135, Table 1] and the estimated behavior may quickly deviate from the original one; see [135, Fig. 3]. This divergence is not surprising and has to be overcome with any other estimation approach, given the rarity of occurrence of chaos in the model's parameter space.

We show in the next subsection that — for the more challenging case of partial observations — it suffices to add a positivity constraint to the standard EMR formulation (*EMR*) for our MSM approach to allow one to derive closure models with very good statistical simulation skill.

7.2. **Numerical results.** The parameters of the system (7.1) for $n = 4$ read as follows:

$$(a_{ij})_{1\leq i,j\leq 4} = \begin{pmatrix} 1 & 1.09 & 1.52 & 0 \\ 0 & 1 & 0.44 & 1.36 \\ 2.33 & 0 & 1 & 0.47 \\ 1.21 & 0.51 & 0.35 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0.72 \\ 1.53 \\ 1.27 \end{pmatrix}. \tag{7.3}$$

These parameter values are those used in [133, 134], for which numerical evidence of chaos is well established. Time series of length $l = 1.5 \times 10^5$ were generated by integration of (7.1), using a standard Euler scheme with time step $\delta t = 0.035$ and initialized at

$$\mathbf{N_0} = (0.5, 0.2, 0.3, 0.7)^T. \tag{7.4}$$

For such an initial state, the time series evolves on an approximation $\widehat{\mathcal{A}}$ of the strange attractor $\mathcal{A}$; see left panel of Fig. 7 below. Only the the first three components (N_1, N_2, N_3) are observed from the integration of this four-dimensional system of ODEs, after removal of the initial transient.

In this population dynamics context, the first property any inverse stochastic model has to satisfy is the positivity of its solutions' components. This constraint can be seen as the counterpart of quadratic-energy conservation in the present context; it leads to the following

natural modification of an MSM in its EMR formulation (*EMR*) for such inverse models, namely

$$
\begin{aligned}
x_{k+1} &= \Pi_{\mathcal{K}_\epsilon}\Big(x_k + \big[-A_N x_k + B_N(x_k,x_k) + F_N\big]\delta t + r_k^{(0)}\delta t\Big), \quad 1 \le k \le l,\\
r_{k+1}^{(m-1)} - r_k^{(m-1)} &= L_N^{(m)}\big[(x_k)^T, (r_k^{(0)})^T, ..., (r_k^{(m-1)})^T\big]^T \delta t + r_k^{(m)}\delta t, \quad 1 \le m \le p.
\end{aligned} \tag{7.5}
$$

We adopt here the notations of Sec. 2, where $x_k = (N_1(t_k), N_2(t_k), N_3(t_k))$, $t_k = k\delta t$, the $r_k^{(m)}$ are three-dimensional vectors, and we recall that p is the number of levels for which the stopping criterion of A is met.

Here $\Pi_{\mathcal{K}_\epsilon}$ denotes the projection onto the convex set

$$
\mathcal{K}_\epsilon := \{(N_1, N_2, N_3) \in \mathbb{R}^3_+ \,:\, N_i \ge \epsilon,\ 1 \le i \le 3\}, \tag{7.6}
$$

for some appropriately chosen $\epsilon > 0$ so that

$$
P\widehat{\mathcal{A}} \subset \mathcal{K}_\epsilon, \tag{7.7}
$$

where $P\widehat{\mathcal{A}}$ denotes the projection of $\widehat{\mathcal{A}}$ onto $\mathbb{R}^3$. Typically, ϵ can be chosen to be any positive number smaller than the minimum of the first three observed components over the simulation interval $0 < t \le l$, which led us to choose $\epsilon = 0.12$.

When the last-level residual in (7.5) is well approximated by an independent and identically distributed (i.i.d.) Gaussian random vector (see Remark 1), the presence of the projection $\Pi_{\mathcal{K}_\epsilon}$ in (7.5) implies that the resulting recursive stochastic process belongs to a class of (discrete) *reflected diffusion processes*; see e.g. [136]. It is the presence of this projector that ensures the vector x_k will have necessarily positive components when it obeys (7.5), in which the last residual has been replaced by an i.i.d. Gaussian noise.

To simplify the estimation procedure of the coefficients in (7.5), a straightforward multilevel regression procedure, such as described in Sec. 2, is performed but the projection is removed at first. The simulation step is then performed by including the projection $\Pi_{\mathcal{K}_\epsilon}$, according to (7.5), in order to avoid the negative values that could lead to blow-up.

Figure 6 show that the autocorrelation functions of the observed variables (N_1, N_2, N_3) are very well reproduced by an MSM (7.5) with $p = 14$ extra layers of hidden variables. The peculiar Shilnikov-like shape of the attractor $\widehat{\mathcal{A}}$ is also captured with relatively good accuracy, as shown by a comparison of panels (a) and (b) in Fig. 7.

The fine, fractal-like structure visible in the MSM attractor reconstruction of Fig. 7b is, as explained below, suggestive of the presence of degenerate noise in the MSM model. A closer look at the numerically estimated coefficients in Eq. (7.5) reveals that the spectrum of the grand linear part of the model — which involves the variables $x, r^{(0)}, \ldots, r^{(p-1)}$, along with the associated matrices — contains three unstable modes, while the noise part forces only a low-dimensional subspace, spanned by a few stable, decaying modes. Such a combination of unstable modes and partial forcing favors a violation of the Hörmander condition. Although we did not formally verify that this is the case, Fig. 7b clearly shows that simple forward integration of Eq. (7.5) does not reveal the symptomatic fuzziness that such a plot typically displays when the SDE's generator is hypoelliptic; see [89]. The fractal-like features observed here in Fig. 7b recall Remark 2 in Sec. 3, to the extent that they argue for overall stable behavior being possible in the presence of linearly unstable modes and of degenerate noise.

The relatively large number of extra layers, $p = 14$, can be significantly reduced by taking into account the particular structure of the right-hand side of (7.1) in formulating the EMR system, i.e. by using some *a priori* knowledge on the f_i, the g_i and the h to be specified in (*MSM*). This number can be even further reduced if the geometric constraint (7.2) is taken into account. The purpose here was merely to show that, even without any *a priori* knowledge

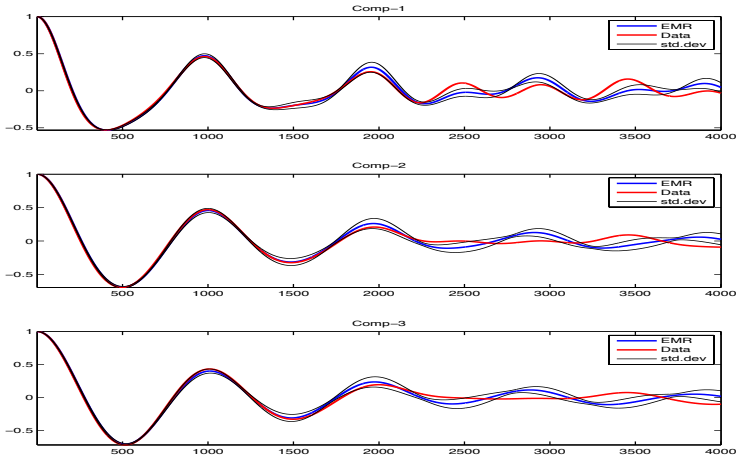

FIGURE 6. Autocorrelation functions of the observed variables (N_1, N_2, N_3) (red curves) obtained by integration of (7.1) with a_{ij} and b given in (7.3), vs. those estimated from the MSM-simulated dynamics (7.5) (blue curves), along with their standard deviations (black curves).

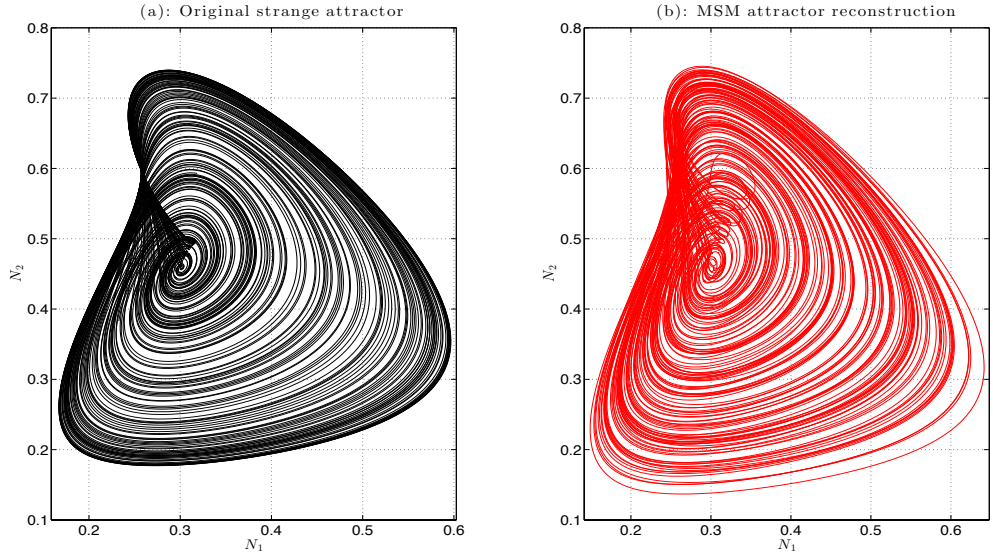

FIGURE 7. Strange attractor of the system (7.1): (a) Original strange attractor of the system; and (b) its MSM reconstruction, both projected onto the (N_1, N_2) plane. Panel (b) is obtained by simple forward integration of (7.5) for an arbitrary realization of the estimated i.i.d. Gaussian noise.

on the dynamics, the learning of an efficient MSM closure is still possible by the use of simple multilevel linear regression techniques, such as those proposed in [18, 19, 26] and described in Sec. 2 here.

8. Acknowledgments

We thank the reviewers for their very useful and constructive comments. The collaboration with several colleagues at Columbia University has helped motivate the data-driven focus of this article. We thank S. Kravstov for useful discussions and two anonymous reviewers for stimulating comments. Preliminary results of this work were presented by DK at the workshop on "Non-equilibrium Statistical Mechanics and the Theory of Extreme Events in Earth Science," held at the Isaac Newton Institute for Mathematical Sciences in November 2013.

This study was supported by grant N00014-12-1-0911 from the Multi-University Research Initiative of the Office of Naval Research, by the National Science Foundation grants DMS-1049253 and OCE-1243175, and by grant DOE-DE-F0A-0000411 from the Department of Energy.

Appendix A. EMR stopping criteria

The stopping criterion for adding levels to an EMR model is based on empirically testing (i) whether the autocorrelations of the last-level residual $\zeta_t \equiv r_k^{(p)}\delta t$ approach zero; and (ii) whether its covariance matrix at zero lag converges to a constant matrix $\Sigma = \zeta_t^{\mathrm{T}}\zeta_t$. For testing purposes, we can assume without loss of generality that $\delta t = 1$. Part (i) of this stopping test is rooted in a standard Durbin-Watson statistical test [137]. This test can also be understood in terms of determining the variance of the regression residual for the additional level that would be added to the EMR model by considering the increments $\delta\zeta_k := \zeta_{k+1} - \zeta_k$.

When ζ_t is reasonably well approximated by a random Gaussian variable, i.e. $\zeta_t \sim \mathcal{N}(0, Q)$, performing an extra regression for $\delta\zeta_k$, according to (2.5), results typically in regression coefficients that all approach zero, except the one corresponding to ζ_k itself, which approaches -1, namely

$$\zeta_{k+1} - \zeta_k \approx -\zeta_k + \gamma_k; \tag{A.1}$$

this observation is numerically documented, for instance, in [18, Fig. 5]. Thus the regression residual γ_k becomes identical to ζ_{k+1} and is just a lagged copy of ζ_k, with lag 1. The coefficient of determination R_i^2 for the i-th component of ζ_k, with $i = 1, \ldots, d$, then becomes:

$$R_i^2 = 1 - \frac{\sum_k \gamma_{i,k}^2}{\sum_k (\zeta_{i,k+1} - \zeta_{i,k})^2} \approx 1 - \frac{\sum_k \zeta_{i,k+1}^2}{\sum_k (\zeta_{i,k+1}^2 + \zeta_{i,k}^2)} \approx 1 - \frac{\mathrm{var}(\zeta_{i,k})}{2\mathrm{var}(\zeta_{i,k})} = 0.5. \tag{A.2}$$

In other words, the fraction of unexplained variance resulting from the regression (A.1) approaches 0.5.

Convergence of the covariance matrix of $\zeta_t \sim r_k^{(p)}$ to a constant matrix $\Sigma = Q^T Q$ can be checked numerically by computing its eigenvalues at each level of Eq. (*EMR*). This convergence typically coincides with the convergence of R_i^2 to 0.5 in Eq. (A.2), for each of the components of ζ_t.

Appendix B. Real-time prediction from an MSM, and initialization of the hidden variables

This appendix is concerned with the initialization problem for real-time prediction based on a time-discrete version of an MSM. The notations used herein are those of of Sec. 2 and of the previous appendix.

To simplify the presentation, we stick to the case of an EMR in its original formulation, namely a discrete system such as (*EMR*). To integrate (*EMR*) for predictive purposes requires

some attention to the initialization of the hidden $r^{(m)}$ variables, $m = 1, \ldots, p$, on the additional EMR levels.

For illustration purposes, we consider here the simple case of a three-level model, namely the main level and two additional ones, i.e. $p = 2$. This model is assumed to be *trained* on the interval $(0, T^*)$, i.e. the model coefficients are obtained from the multivariate time series of data available for $t = k\delta t$ in this interval. The next point in the time series, $T^* + 1$, belongs to the *validation* interval where we want to initiate prediction into the future. In other words, we place ourselves in the real-time case, in which the model coefficients are no longer updated with new observations after T^*.

The right-hand side of (*EMR*) provides a practical way to initialize the hidden variables. By assuming, as in A and without loss of generality, that $\delta t = 1$, one gets the following recurrence relations:

$$r_k^{(0)} = f_1(x_k) - (x_{k+1} - x_k), \tag{B.1a}$$

$$r_k^{(1)} = L^{(0)}[x_k, r_k^{(0)}] - (r_{k+1}^{(0)} - r_k^{(0)}); \tag{B.1b}$$

here $f_1(x) = -Ax + B(x, x) + F$, and the last-level noise ξ_k has been dropped.

If we assume x_{T^*} to be the last available observed data point, then (B.1) only provides the hidden variables $r_k^{(0)}$ and $r_k^{(1)}$ up to their corresponding last values, namely $r_{T^*-1}^{(0)}$ and $r_{T^*-2}^{(1)}$, respectively. Indeed, if $x_{k+1} = x_{T^*}$, then Eq. (B.1a) only provides $r_{T^*-1}^{(0)}$, from which Eq. (B.1b) can only help determine $r_{T^*-2}^{(1)}$, since $r_{T^*}^{(0)}$ is not available.

On the other hand, if we assume that a new observation x_{T^*+1} becomes available at $T^* + 1$ in the validation interval, the value of the $r^{(1)}$-variable can then be calculated beyond its previously known value, i.e. $r_{T^*-1}^{(1)}$ becomes available according to

$$\begin{aligned} r_{T^*}^{(0)} &= f_1(x_{T^*}) - (x_{T^*+1} - x_{T^*}), \\ r_{T^*-1}^{(1)} &= L^{(0)}[x_{T^*-1}, r_{T^*-1}^{(0)}] - (r_{T^*}^{(0)} - r_{T^*-1}^{(0)}). \end{aligned} \tag{B.2}$$

After such an initialization of the hidden $r^{(1)}$-variable beyond its previously computed value, the model prediction $\widehat{x}_{T^*+2}$ of x_{T^*+2} from the initial data x_{T^*+1} is then obtained by integrating the $r^{(1)}$-variable from the last level to the main one, according to:

$$\begin{aligned} r_{T^*}^{(1)} &= r_{T^*-1}^{(1)} + L^{(1)}[x_{T^*-1}, r_{T^*-1}^{(0)}, r_{T^*-1}^{(1)}] + \zeta_{T^*-1}, \\ r_{T^*+1}^{(0)} &= r_{T^*}^{(0)} + L^{(0)}[x_{T^*}, r_{T^*}^{(0)}] + r_{T^*}^{(1)}, \\ \widehat{x}_{T^*+2} &= x_{T^*+1} + f_1(x_{T^*+1}) + r_{T^*+1}^{(0)}. \end{aligned} \tag{B.3}$$

Note that both $r_{T^*}^{(1)}$ and $r_{T^*+1}^{(0)}$ are now randomized due to the presence of the random variable ζ_{T^*-1} on the last level. Hence the forecast uncertainty is properly accounted for in x_{T^*+2}.

Such an initialization and forecast procedure based on Eq. (1.2) can obviously be carried out for any number of levels of an EMR. More generally, it can be extended to an explicit Euler-Maruyama discretization of any stable MSM, given in its general form and satisfying the hypotheses of Theorem 3.1. The discussion above emphasizes that, for such multilevel systems and for real-time prediction purposes, the forecast has to be started several time steps back into the past; the number of these steps must equal the number of hidden layers in (*MSM*). As an illustration, Eq. (B.2) represents *backward* initialization of hidden variables in the past by going from the main level to the last one, followed by their *forward* integration from the last level to the main one into the future, as in (B.3).

Appendix C. Energy conservation and its practical EMR aspects

In this appendix, we identify linear relations among the coefficients B_{ijk} so that Eq. (3.20) be satisfied. Since it is easier to introduce linear constraints in the least-square estimation of the regression coefficients, we look for such constraints that are, in fact, both necessary and sufficient for Eq. (3.20) to hold. These constraints ensure that *quadratic* nonlinearities in an EMR formulation will preserve the energy $\|x\|^2$ for the Euclidean norm $\|\cdot\|$ associated with a given basis. In another basis, these conditions remain valid, up to a linear change of coordinates. There are three types of such linear equality constraints for B.

First, the coefficient B_{iii} that corresponds to the quadratic term x_i^2 in the equation for the time evolution of x_i is required to be identically zero:

$$B_{iii} = 0, \quad i = 1, \ldots, d. \tag{C.1}$$

In addition, there is a condition that involves the quadratic interactions $x_j x_k$ and x_k^2 in the equation for x_j, and x_j^2 and $x_j x_k$ in the equation for x_k. This condition yields the following skew-symmetric constraints for the two pairs of coefficients:

$$B_{jjk} + B_{kjj} = 0, \quad B_{jkk} + B_{kjk} = 0, \quad 1 \le j \ne k \le d. \tag{C.2}$$

Finally, there are the quadratic interactions of type $x_j x_k$ in the equation for x_i that require the sum of three EMR coefficients to be zero:

$$B_{ijk} + B_{jik} + B_{kij} = 0, \quad 1 \le i, j, k \le d \text{ such that } (i-j)(j-k)(k-i) \ne 0. \tag{C.3}$$

A condition like (3.19) for the matrix A to be positive definite can also be included in the regression procedure, as necessary and appropriate. For particular applications that are consistent with geophysical flow models, such as the one considered in Sec. 6, this condition can be cast as a combination of linear equality and inequality constraints. The equality constraint imposes skew-symmetry for off-diagonal terms:

$$A_{ij} + A_{ji} = 0, \quad i \ne j, \tag{C.4}$$

while the inequality one requires nonnegative diagonal terms:

$$A_{ii} > 0. \tag{C.5}$$

The total set of constraints is determined by a look-up through all possible occurrences of the above types, Eqs. (C.1)–(C.5), and their number p scales as $p \sim d^2$, where d is the dimension of x. For a quadratic EMR model such as (*EMR*), the total number of model coefficients P is also of order d^2. Hence the use of these constraints reduces considerably the number of independent model coefficients, from P to $P - p$.

Note that energy conserving constraints have been used in model reduction techniques before [27, 28, 29], cf. [138, 139]. In both these approaches, however, it is assumed that the full model equations are known and available. In particular, refs. [27, 28, 29] used essentially parameter estimation techniques to obtain the values of the parameters in a model of known structure. The data-driven methodology proposed in this section, however, operates whether the full governing equations that generated the data are known or not. Hence, Eqs. (C.1)–(C.5) are new and of some interest, in particular in the highly realistic and frequently occurring case in which a detailed — physical, chemical, biological or socio-economic — model of the process that generated the data is not known.

In the unconstrained case, the coefficients for each component x_i of x can be estimated sequentially by the EMR methodolgy, while the energy-conserving constraints require estimation of all model coefficients at the main level — *i.e.*, A, B and F — simultaneously, by using a grand matrix of predictors. The least-square minimization with constraints is solved via

quadratic programming with a set of Lagrange multipliers, by using an active-set algorithm or a projection method [140].

Appendix D. Interpretation of MSM coefficients

Here we focus on the question of interpreting coefficients of EMR or, more generally, MSM models that have been constructed in the partial-observing situation. The emphasis is on a simple but essential time-orthogonality property that the observed and hidden variables must satisfy as a consequence of the multilevel regression procedure in the EMR and MSM methodology. This time-orthogonality property typically yields different reduced-model coefficients than those of the original full model, while still allowing the former to simulate the main statistical features of the observed dynamics. For illustration purposes, we use a simple linear-model example borrowed from [27].

First we point out a basic but important property to be satisfied by the $r^{(m)}$-variables once the multilevel regression described in Sec. 2 has been applied. Let us recall that for classical — *i.e.*, unconstrained and nonregularized — least-square minimization, and in continuous time, a regression residual $r^{(0)}(t)$ is necessarily orthogonal in the time domain to the predictor variables x [78]. Likewise, the $r^{(m)}(t)$ variables are estimated so as to be orthogonal in the time domain to the variables from the previous levels $(x, r^{(0)}, \ldots, r^{(m-1)})$,

$$\begin{aligned} &\langle x, r^{(m-1)} \rangle_{L^2} = 0; \\ &\langle r^{(j)}, r^{(m)} \rangle_{L^2} = 0, \quad 1 \leq j, m \leq p-1 \ \text{ with } j \leq m-1, \end{aligned} \tag{D.1}$$

with respect to the $L^2((0,T^*);\mathbb{R}^d)$ inner product defined by:

$$\langle z, y \rangle_{L^2} := \sum_{i=1}^{d} \frac{1}{T^*} \int_0^{T^*} z_i(t) y_i(t) \mathrm{d}t, \tag{D.2}$$

and for any vector-valued functions$z(t) = (z_1(t), \ldots, z_d(t))^{\mathrm{T}}$ and $y = (y_1(t), \ldots, y_d(t))^{\mathrm{T}}$.

The simple two-variable, linear model from [27] that we use is given by

$$\begin{pmatrix} \mathrm{d}x \\ \mathrm{d}y \end{pmatrix} = \begin{pmatrix} a & 1 \\ q & A \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \mathrm{d}t + \begin{pmatrix} 0 \\ \sigma \mathrm{d}W \end{pmatrix}; \tag{D.3}$$

x here stands for a slow, resolved variable, y for a fast, unresolved one, while the model coefficients are $a = -2, q = 1$ and $A = -1$, with $\sigma > 0$. By using the standard Euler-Maruyama scheme for SDEs [80, p. 305], one integrates the following finite-difference version of (D.3):

$$\begin{pmatrix} x_{i+1} - x_i \\ y_{i+1} - y_i \end{pmatrix} = \begin{pmatrix} a & 1 \\ q & A \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} \delta t + \begin{pmatrix} 0 \\ \sigma \xi_i \sqrt{\delta t} \end{pmatrix}, \tag{D.4}$$

where the ξ_i's are real-valued random variables drawn independently from a normal distribution $\mathcal{N}(0,1)$.

In [27], standard least-square techniques (see appendices therein) were used to derive analytically an EMR model able to reproduce statistical features of the dynamics — such as decay of correlations or the PDF — of the x-variable alone, by using only a finite-length time series $\{x_i : i = 1, ..., N\}$ of the resolved variable obtained by integrating the full system given by Eq. (D.4). According to Theorem 4.2 in [27], in the limit of $N \to \infty$ and $\delta t \to 0$, the resulting EMR model's linear part for the $(x, r)^{\mathrm{T}}$ vector has different coefficients but the same eigenvalues as the original (x, y) model of Eqs. (D.4).

We stress that the difference in model coefficients is solely due to the change of basis imposed by the orthogonal dynamics of the "hidden" EMR variable r, according to the conditions given

in (D.1). The related similarity transformation $S : (x, y) \to (x, r)$ of the model's linear part is simply

$$\begin{pmatrix} x \\ y \end{pmatrix} = S \begin{pmatrix} x \\ r \end{pmatrix}, \text{ with } S = \begin{pmatrix} 1 & 0 \\ -a & 1 \end{pmatrix}, \text{ and } S^{-1} = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}. \tag{D.5}$$

In fact, the transformed version of the full model in continuous time, Eq. (D.3), is given by the MSM

$$\begin{pmatrix} \mathrm{d}x \\ \mathrm{d}r \end{pmatrix} = S^{-1} \begin{pmatrix} a & 1 \\ q & A \end{pmatrix} S \begin{pmatrix} x \\ r \end{pmatrix} \mathrm{d}t + \begin{pmatrix} 0 \\ \sigma \mathrm{d}W \end{pmatrix}, \tag{D.6}$$

while the transformed version of the full model's discrete form, Eq. (D.4), is equivalent to the EMR model:

$$\begin{pmatrix} x_{i+1} - x_i \\ r_{i+1} - r_i \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ q - Aa & a + A \end{pmatrix} \begin{pmatrix} x_i \\ r_i \end{pmatrix} \delta t + \begin{pmatrix} 0 \\ \sigma \xi_i \sqrt{\delta t} \end{pmatrix}. \tag{D.7}$$

Thus, the linear part of Eq. (D.7) shares the same eigenvalues as the linear part of Eq. (D.4), while the random forcing in Eq. (D.7) is of the same amplitude, and the same nature, as in Eq. (D.4); hence the model statistics of the simulated variables are identical in the limit of $N \to \infty$ and $\delta t \to 0$. The only difference between Eq. (D.7) and Eq. (D.4) is that the hidden variable r is orthogonal to x — while the original y variable is not.

The highly idealized situation analyzed here can obviously be generalized to a broad class of stochastically forced systems that might, in particular, have a much higher number of resolved as well as hidden variables.

References

[1] M. Ghil and A. W. Robertson, "Solving problems with GCMs: General circulation models and their role in the climate modeling hierarchy," in *General Circulation Model Development: Past, Present and Future* (D. Randall, ed.), p. 285325, Academic Press, San Diego, 2000.

[2] A. L. Wilmot-Smith, P. C. H. Martens, D. Nandy, E. R. Priest, and S. M. Tobias, "Low-order stellar dynamo models," *Mon. Not. R. Astron. Soc.*, vol. 363, pp. 1167–1172, 2005.

[3] M. Bachar, J. J. Batzel, and S. Ditlevsen, *Stochastic Biomathematical Models: with Applications to Neuronal Modeling*. Springer, 2013.

[4] F. Noé and S. Fischer, "Transition networks for modeling the kinetics of conformational change in macromolecules," *Current opinion in structural biology*, vol. 18, no. 2, pp. 154–162, 2008.

[5] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," *J. Chemical Phys.*, vol. 134, no. 17, p. 174105, 2011.

[6] S. Donnet and A. Samson, "A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models," *Advanced drug delivery reviews*, vol. 65, no. 7, pp. 929–939, 2013.

[7] S. J. Camargo, A. W. Robertson, S. J. Gaffney, and M. Ghil, "Cluster analysis of typhoon tracks. Part II: Large-scale circulation and ENSO," *J. Climate*, vol. 20, pp. 3654–3676, 2007.

[8] R. S. Kovats, M. J. Bouma, S. Hajat, E. Worrall, and A. Haines, "El Niño and health," *The Lancet*, vol. 362, pp. 1481–1489, 2003.

[9] C. Ropelewski and M. Halpert, "Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation.," *Mon. Wea. Rev.*, vol. 115, pp. 1606–1626, 1987.
[10] S. M. Hsiang, K. C. Meng, and M. A. Cane, "Civil conflicts are associated with the global climate," *Nature*, vol. 476, pp. 438–441, 2011.
[11] M. Cane, S. Zebiak, and S. Dolan, "Experimental forecasts of El Niño," *Nature*, vol. 321, pp. 827–832, 1986.
[12] C. Penland, "Random forcing and forecasting using principal oscillation pattern analysis," *Mon. Wea. Rev.*, vol. 117, pp. 2165–2185, 1989.
[13] C. Penland and P. D. Sardeshmukh, "The optimal growth of tropical sea surface temperature anomalies," *J. Climate*, vol. 8, pp. 1999–2024, 1995.
[14] C. R. Winkler, M. Newman, and P. D. Sardeshmukh, "A linear model of wintertime low-frequency variability. Part I: Formulation and forecast skill," *J. Climate*, vol. 14, pp. 4474–4494, 2001.
[15] Y. Xue, M. A. Cane, S. E. Zebiak, and M. B. Blumenthal, "On the prediction of ENSO: A study with a low-order Markov model," *Tellus*, vol. 46A, pp. 512–528, 1994.
[16] C. Penland and T. Magorian, "Prediction of Niño-3 sea surface temperatures using iinear inverse modeling," *J. Climate*, vol. 6, pp. 1067–1076, 1993.
[17] A. G. Barnston, M. K. Tippett, M. L. Heureux, S. Li, and D. G. DeWitt, "Skill of real-time seasonal ENSO model predictions during 2002-2011 — is our capability improving?," *Bull. Am. Meteor. Soc.*, vol. 93(5), pp. 631–651, 2012.
[18] S. Kravtsov, D. Kondrashov, and M. Ghil, "Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability," *J. Climate*, vol. 18, pp. 4404–4424, 2005.
[19] D. Kondrashov, S. Kravtsov, A. W. Robertson, and M. Ghil, "A hierarchy of data-based ENSO models," *J. Climate*, vol. 18, pp. 4425–4444, 2005.
[20] S. Kravtsov, D. Kondrashov, and M. Ghil, "An empirical stochastic model of sea-surface temperature and surface wind over the Southern Ocean," *Ocean Sci.*, vol. 8, pp. 1891–1936, 2011.
[21] D. Kondrashov, S. Kravtsov, and M. Ghil, "Empirical mode reduction in a model of extratropical low-frequency variability," *J. Atmos. Sci.*, vol. 63, pp. 1859–1877, 2006.
[22] D. Kondrashov, S. Kravtsov, and M. Ghil, "Signatures of nonlinear dynamics in an idealized atmospheric model," *J. Atmos. Sci.*, vol. 68, pp. 3–12, 2011.
[23] J. M. Peters, S. Kravtsov, and N. T. Schwartz, "Predictability associated with nonlinear regimes in an atmospheric model," *J. Atmos. Sci.*, vol. 69, pp. 1137–1154, MAR 2012.
[24] D. Kondrashov, M. D. Chekroun, A. W. Robertson, and M. Ghil, "Low-order stochastic model and "past-noise forecasting" of the Madden-Julian Oscillation," *Geophys. Res. Lett.*, vol. 40, pp. 5305–5310, 2013. doi:10.1002/grl.50991.
[25] K. Strounine, S. Kravtsov, D. Kondrashov, and M. Ghil, "Reduced models of atmospheric low-frequency variability: Parameter estimation and comparative performance," *Physica D*, vol. 239, pp. 145–166, 2010.
[26] S. Kravtsov, D. Kondrashov, and M. Ghil, "Empirical model reduction and the modeling hierarchy in climate dynamics," in *Stochastic Physics and Climate Modeling* (T. N. Palmer and P. Williams, eds.), pp. 35–72, Cambridge Univ. Press, 2009.
[27] A. J. Majda and Y. Yuan, "Fundamental limitations of ad hoc linear and quadratic multi-level regression models for physical systems," *Discrete Contin. Dyn. Systems*, vol. 17(4), pp. 1333–1363, 2012.
[28] A. J. Majda and J. Harlim, "Physics constrained nonlinear regression models for time series," *Nonlinearity*, vol. 26, no. 1, pp. 201–217, 2013.

[29] J. Harlim, A. Mahdi, and A. Majda, "An ensemble kalman filter for statistical estimation of physics constrained nonlinear regression models," *J. Comput. Phys.*, no. 257, pp. 782–812, 2014.
[30] S. Smale, "On the differential equations of species in competition," *J. Math. Biol.*, vol. 3, no. 1, pp. 5–7, 1976.
[31] M. W. Hirsch, "Systems of differential equations which are competitive or cooperative: I. Limit sets," *SIAM J. Math. Analysis*, vol. 13, no. 2, pp. 167–179, 1982.
[32] M. W. Hirsch, "Systems of differential equations that are competitive or cooperative II: Convergence almost everywhere," *SIAM J. Math. Analysis*, vol. 16, no. 3, pp. 423–439, 1985.
[33] M. W. Hirsch, "Systems of differential equations which are competitive or cooperative: III. Competing species," *Nonlinearity*, vol. 1, no. 1, p. 51, 1988.
[34] M. W. Hirsch, "Systems of differential equations that are competitive or cooperative. iv: Structural stability in three-dimensional systems," *SIAM J. Math. Analysis*, vol. 21, no. 5, pp. 1225–1234, 1990.
[35] H. L. Smith, *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*. American Mathematical Society, Providence, R.I., 2008.
[36] O. E. Rössler, "Chaotic behavior in simple reaction systems," *Zeitschrift Naturforschung Teil A*, vol. 31, pp. 259–264, 1976.
[37] O. E. Rössler, "Chaos in abstract kinetics: Two prototypes," *Bulletin of mathematical biology*, vol. 39, no. 2, pp. 275–289, 1977.
[38] L. A. Segel and M. Slemrod, "The quasi-steady-state assumption: a case study in perturbation," *SIAM review*, vol. 31, no. 3, pp. 446–477, 1989.
[39] U. Maas and S. B. Pope, "Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space," *Combustion and Flame*, vol. 88, no. 3, pp. 239–264, 1992.
[40] L. V. Kalachev and R. J. Field, "Reduction of a model describing ozone oscillations in the troposphere: example of an algorithmic approach to model reduction in atmospheric chemistry," *J. Atmos. Chemistry*, vol. 39, no. 1, pp. 65–93, 2001.
[41] R. Zwanzig, *Nonequilibrium Statistical Mechanics*. Oxford University Press, 2001.
[42] A. J. Chorin, O. H. Hald, and R. Kupferman, "Optimal prediction with memory," *Physica D*, vol. 166, pp. 239–257, 2002.
[43] A. J. Chorin and O. H. Hald, *Stochastic Tools in Mathematics and Science*. No. 147 in Surveys and Tutorials in the Applied Mathematical Sciences, Springer, New York, 2006.
[44] M. D. Chekroun, D. Kondrashov, and M. Ghil, "Predicting stochastic systems by noise sampling, and application to the El Niño-Southern Oscillation," *Proc. Natl. Acad. Sci. USA*, vol. 108, pp. 11766–11771, 2011.
[45] A. J. Majda, R. Abramov, and M. Grote, *Information Theory and Stochastics for Multiscale Nonlinear Systems*. Philadelphia, PA: American Mathematical Society, 2005. 140 pp.
[46] D. M. Fargue, "Reducibilité des systèmes héréditaires à des systèmes dynamiques (régis des équations differentielles ou aux dérivées partielles)," *C. R. Acad. Sci. Paris Ser. B.*, vol. 277, pp. 471–473, 1973.
[47] A. Wörz-Busekros, "Global stability in ecological systems with continuous time delay," *SIAM J. Appl. Math.*, vol. 35, no. 1, pp. 123–134, 1978.
[48] C. M. Dafermos, "Asymptotic stability in viscoelasticity," *Archive for rational mechanics and analysis*, vol. 37, no. 4, pp. 297–308, 1970.
[49] M. D. Chekroun, F. Di Plinio, N. E. Glatt-Holtz, and V. Pata, "Asymptotics of the Coleman-Gurtin model," *Discrete Contin. Dyn. Syst. Ser. S*, vol. 4, no. 2, pp. 351–369,

2011.
[50] M. D. Chekroun and N. E. Glatt-Holtz, "Invariant measures for dissipative dynamical systems: abstract results and applications," *Commun. Math. Phys.*, vol. 316, no. 3, pp. 723–761, 2012.
[51] K. Bhattacharya, M. Ghil, and I. L. Vulis, "Internal variability of an energy-balance model with delayed albedo effects," *J. Atmos. Sci.*, vol. 39, p. 17471773, 1982.
[52] L. Roques, M. D. Chekroun, M. Cristofol, S. Soubeyrand, and M. Ghil, "Parameter estimation for energy balance models with memory," *Proc. Roy. Soc. A*, vol. 470, p. 20140349, 2014.
[53] V. Volterra, *Leçons sur la théorie mathématique de la lutte pour la vie*. Gauthier-Villars, Paris, 1931.
[54] N. McDonald, *Time Lags in Biological Models*. Lecture Notes in Biomathematics, vol. 27, Springer-Verlag, 1978.
[55] H. Smith, *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Springer, 2011.
[56] H. Mori, "A continued-fraction representation of the time-correlation functions," *Prog. Theor. Phys.*, vol. 34, no. 3, pp. 399–416, 1965.
[57] R. Kupferman, "Fractional kinetics in Kac–Zwanzig heat bath models," *J. Stat. Phys.*, vol. 114, no. 1-2, pp. 291–326, 2004.
[58] R. Kupferman and A. Stuart, "Fitting sde models to nonlinear Kac–Zwanzig heat bath models," *Physica D*, vol. 199, no. 3, pp. 279–316, 2004.
[59] I. Horenko, C. Hartmann, C. Schütte, and F. Noé, "Data-based parameter estimation of generalized multidimensional langevin processes," *Physical Review E*, vol. 76, no. 1, p. 016706, 2007.
[60] M. Niemann, T. Laubrich, E. Olbrich, and H. Kantz, "Usage of the mori-zwanzig method in time series analysis," *Physical Review E*, vol. 77, no. 1, p. 011117, 2008.
[61] R. Mañé, "On the dimension of the compact invariant sets of certain non-linear maps," in *Dynamical Systems and Turbulence*, vol. 898 of *Lecture Notes in Mathematics*, pp. 230–242, Springer, Berlin, 1981.
[62] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*, vol. 898 of *Lecture Notes in Mathematics*, pp. 366–381, Springer, Berlin, 1981.
[63] D. S. Broomhead and G. P. King, "Extracting qualitative dynamics from experimental data," *Physica D*, vol. 20, no. 2-3, pp. 217–236, 1986.
[64] D. S. Broomhead and G. P. King, "On the qualitative analysis of experimental dynamical systems," in *Nonlinear Phenomena and Chaos* (S. Sarkar, ed.), pp. 113–144, Bristol, England: Adam Hilger, 1986.
[65] 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.*, vol. 40, no. 1, pp. 1–41, 2002.
[66] C. Nichkawde, "Sparse model from optimal nonuniform embedding of time series," *Physical Review E*, vol. 89, no. 4, p. 042911, 2014.
[67] G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. Boca Raton, Fla.: Holden-Day, 1970.
[68] S. A. Billings, S. Chen, and M. J. Korenberg, "Identification of MIMO non-linear systems using a forward-regression orthogonal estimator," *Intl. J. Control*, vol. 49, no. 6, pp. 2157–2189, 1989.
[69] S. Lu, K. H. Ju, and K. H. Chon, "A new algorithm for linear and nonlinear arma model parameter estimation using affine geometry [and application to blood flow/pressure data],"

Biomedical Engineering, IEEE Transactions on, vol. 48, no. 10, pp. 1116–1124, 2001.
[70] A. Willsky, "Multiresolution markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396–1458, 2002.
[71] A. T. Ihler, S. Kirshner, M. Ghil, A. W. Robertson, and P. Smyth, "Graphical models for statistical inference and data assimilation," *Physica D*, vol. 230, pp. 72–87, 2007.
[72] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
[73] D. Mukhin, E. Loskutov, A. Mukhina, A. Feigin, I. Zaliapin, and M. Ghil, "Predicting critical transitions in enso models, part i: Methodology and simple models with memory." *J. Climate,* under review, 2014.
[74] D. Mukhin, D. Kondrashov, E. Loskutov, A. Gavrilov, A. Feigin, I. Zaliapin, and M. Ghil, "Predicting critical transitions in enso models, part ii: Spatially dependent models." *J. Climate,* under review, 2014.
[75] A. J. Majda and P. R. Kramer, "Simplified models for turbulent diffusion: theory, numerical modelling, and physical phenomena," *Physics Reports*, vol. 314, no. 4, pp. 237–574, 1999.
[76] M. D. Chekroun, J. D. Neelin, D. Kondrashov, J. C. McWilliams, and M. Ghil, "Rough parameter dependence in climate models: The role of Ruelle-Pollicott resonances," *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 5, pp. 1684–1690, 2014.
[77] D. T. Crommelin and A. J. Majda, "Strategies for model reduction: comparing different optimal bases," *J. Atmospheric Sci.*, vol. 61, no. 17, pp. 2206–2217, 2004.
[78] T. Hastie, R. Tibshirani, and J. Friedman, *Linear Methods for Regression*. Springer, 2009.
[79] P. Billingsley, *Convergence of Probability Measures*. John Wiley & Sons, 2009.
[80] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, vol. 23 of *Applications of Mathematics*. Springer-Verlag, Berlin, 1992.
[81] D. J. Higham, X. Mao, and A. M. Stuart, "Strong convergence of Euler-type methods for nonlinear stochastic differential equations," *SIAM J. Numerical Analysis*, vol. 40, no. 3, pp. 1041–1063, 2002.
[82] X. Mao and C. Yuan, *Stochastic Differential Equations with Markovian Switching*. World Scientific, 2006.
[83] C. Penland and B. D. Ewald, "On modelling physical systems with stochastic models: diffusion versus Lévy processes," *Phil. Trans. Roy. Soc. A*, vol. 366, no. 1875, pp. 2455–2474, 2008.
[84] J.-P. Eckmann and D. Ruelle, "Ergodic theory of chaos and strange attractors," *Rev. Modern Phys.*, vol. 57, pp. 617–656, 1985.
[85] B. Gess, W. Liu, and M. Röckner, "Random attractors for a class of stochastic partial differential equations driven by general additive noise," *J. Differential Equations*, vol. 251, no. 4, pp. 1225–1253, 2011.
[86] K. R. Schenk-Hoppé, "Random attractors-general properties, existence and applications to stochastic bifurcation theory," *Discrete Cont. Dyn. Syst.*, vol. 4, pp. 99–130, 1998.
[87] M. D. Chekroun, H. Liu, and S. Wang, *Approximation of Invariant Manifolds: Stochastic Manifolds for Nonlinear SPDEs I*. Springer Briefs in Mathematics, *to appear*, Springer, 2014.
[88] L. Arnold, *Random Dynamical Systems*. New York: Springer-Verlag, 1998.
[89] M. D. Chekroun, E. Simonnet, and M. Ghil, "Stochastic climate dynamics: Random attractors and time-dependent invariant measures," *Physica D*, vol. 240, no. 21, pp. 1685–1700, 2011.

[90] H. Crauel and F. Flandoli, "Attractors for random dynamical systems," *Probab. Theory Relat. Fields*, vol. 100, pp. 365–393, 1994.
[91] V. I. Arnol'd, *Mathematical Methods of Classical Mechanics*, vol. 60. Springer, 1989.
[92] F. Varadi, B. Runnegar, and M. Ghil, "Successive refinements in long-term integrations of planetary orbits," *Astrophys. J.*, vol. 592, pp. 620–630, 2003.
[93] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-preserving Algorithms for Ordinary Differential Equations*. Springer, 2006.
[94] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, vol. 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge: Cambridge University Press, 2008.
[95] A. J. Chorin, O. H. Hald, and R. Kupferman, "Prediction from partial data, renormalization, and averaging," *J. Sci. Comput.*, vol. 28, no. 2-3, pp. 245–261, 2006.
[96] A. J. Chorin and P. Stinis, "Problem reduction, renormalization, and memory," *Commun. Appl. Math. Comput. Sci.*, vol. 1, pp. 1–27, 2006.
[97] P. Stinis, "A comparative study of two stochastic mode reduction methods," *Physica D*, vol. 213, no. 2, pp. 197–213, 2006.
[98] O. H. Hald and P. Stinis, "Optimal prediction and the rate of decay for solutions of the euler equations in two and three dimensions," *Proc. Natl. Acad. Sci. USA*, vol. 104, no. 16, pp. 6527–6532, 2007.
[99] P. Stinis, "Higher-order Mori-Zwanzig models for the Euler equations," *Multiscale Model. & Simul.*, vol. 6, no. 3, pp. 741–760, 2007.
[100] A. J. Majda, I. Timofeyev, and E. Vanden-Eijnden, "A mathematical framework for stochastic climate models," *Commun. Pure Appl. Math.*, vol. 54, pp. 891–974, 2001.
[101] I. Melbourne and A. M. Stuart, "A note on diffusion limits of chaotic skew-product flows," *Nonlinearity*, vol. 24, no. 4, pp. 1361–1367, 2011.
[102] G. A. Gottwald and I. Melbourne, "Homogenization for deterministic maps and multiplicative noise," *Proc. R. Soc. Lond. Ser. A*, vol. 469, no. 2156, pp. 20130201, 16 pp., 2013.
[103] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, vol. 68 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2nd ed., 1997.
[104] M. Budišić, R. Mohr, and I. Mezić, "Applied Koopmanism," *Chaos*, vol. 22, p. 047510, 2012. doi: 10.1063/1.4772195.
[105] M. D. Chekroun and J. Roux, "Homeomorphisms group of normed vector space: Conjugacy problems and the Koopman operator," *Discrete Cont. Dyn. Syst. Series A (DCDS-A)*, vol. 33, pp. 3957–3980, 2013.
[106] I. P. Cornfeld, S. V. Fomin, and Y. G. Sinai, *Ergodic Theory*. Springer-Verlag, Berlin, 1982.
[107] A. Lasota and M. Mackey, *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, vol. 97. Springer, 1994.
[108] H. Brézis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2010.
[109] K.-J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, vol. 194 of *Graduate Texts in Mathematics*. New York: Springer-Verlag, 2000.
[110] O. Butterley and C. Liverani, "Smooth Anosov flows: correlation spectra and stability," *J. Modern Dynamics*, vol. 1, pp. 301–322, 2007.
[111] D. Givon, R. Kupferman, and A. Stuart, "Extracting macroscopic dynamics: model problems and algorithms," *Nonlinearity*, vol. 17, no. 6, pp. R55–R127, 2004.
[112] W. Rudin, *Real and Complex Analysis (3rd ed.)*. New York: McGraw-Hill, 1987.

[113] R. M. Dudley, *Real Analysis and Probability*. Cambridge University Press, 2002.
[114] O. Kallenberg, *Foundations of Modern Probability*. Springer, 2002.
[115] H. Crauel, *Random Probability Measures on Polish Spaces*. Taylor & Francis Inc., 2002.
[116] H. Mori, "Transport, collective motion, and brownian motion," *Prog. Theor. Phys.*, vol. 33, no. 3, pp. 423–455, 1965.
[117] S. Nakajima, "On quantum theory of transport phenomena steady diffusion," *Prog. Theor. Phys.*, vol. 20, no. 6, pp. 948–959, 1958.
[118] I. Prigogine and P. Resibois, "On the kinetics of the approach to equilibrium," *Physica*, vol. 27, no. 7, pp. 629–646, 1961.
[119] R. Zwanzig, "Ensemble method in the theory of irreversibility," *J. Chem. Phys.*, vol. 33, p. 1338, 1960.
[120] R. Zwanzig, "On the identity of three generalized master equations," *Physica*, vol. 30, no. 6, pp. 1109–1123, 1964.
[121] A. J. Chorin, O. H. Hald, and R. Kupferman, "Optimal prediction and the Mori–Zwanzig representation of irreversible processes," *Proc. Natl. Acad. Sci. USA*, vol. 97, no. 7, pp. 2968–2973, 2000.
[122] E. Lehmann and G. Casella, *Theory of point estimation*, vol. 31. Springer, 1998.
[123] 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.*, vol. 64, pp. 3987–4003, NOV 2007.
[124] A. J. Majda, C. Franzke, and B. Khouider, "An applied mathematics perspective on stochastic modelling for climate," *Phil. Trans. Roy. Soc. London A*, vol. 366, pp. 2429–2455, JUL 28 2008.
[125] M. Kimoto and M. Ghil, "Multiple flow regimes in the Northern Hemisphere winter. Part I: Methodology and hemispheric regimes," *J. Atmos. Sci.*, vol. 50, pp. 2625–2643, 1993.
[126] M. Kimoto and M. Ghil, "Multiple flow regimes in the Northern Hemisphere winter. Part II: Sectorial regimes and preferred transitions," *J. Atmos. Sci.*, vol. 50, pp. 2645–2673, 1993.
[127] M. Ghil and S. Childress, *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory and Climate Dynamics*. New York/Berlin: Springer-Verlag, 1987.
[128] R. M. May, *Stability and Complexity in Model Ecosystems*. Princeton University Press, 2001.
[129] M. Peschel and W. Mende, *The Predator-Prey Model: Do We Live in a Volterra World?* Springer,Wien/New York, 1986.
[130] V. Kozlov and S. Vakulenko, "On chaos in Lotka–Volterra systems: An analytical approach," *Nonlinearity*, vol. 26, no. 8, p. 2299, 2013.
[131] E. N. Lorenz, "Deterministic nonperiodic flow," *J. Atmos. Sci.*, vol. 20, no. 2, pp. 130–141, 1963.
[132] A. Arneodo, P. Coullet, J. Peyraud, and C. Tresser, "Strange attractors in Volterra equations for species in competition," *J. Math. Biology*, vol. 14, no. 2, pp. 153–157, 1982.
[133] J. Vano, J. Wildenberg, M. Anderson, J. Noel, and J. Sprott, "Chaos in low-dimensional Lotka–Volterra models of competition," *Nonlinearity*, vol. 19, no. 10, p. 2391, 2006.
[134] L. Roques and M. D. Chekroun, "Probing chaos and biodiversity in a simple competition model," *Ecological Complexity*, vol. 8, no. 1, pp. 98–104, 2011.
[135] E. Voit and I.-C. Chou, "Parameter estimation in canonical biological systems models," *Intl. J. Systems Synthetic Biol.*, vol. 1, pp. 1–19, 2010.
[136] P. Dupuis, "Large deviations analysis of reflected diffusions and constrained stochastic approximation algorithms in convex sets," *Stochastics*, vol. 21, no. 1, pp. 63–96, 1987.

[137] J. Durbin and G. S. Watson, "Testing for serial correlation in least squares regression. i," *Biometrika*, vol. 37, no. 3-4, pp. 409–428, 1950.
[138] F. Kwasniok, "The reduction of complex dynamical systems using principal interaction patterns," *Physica D*, vol. 92, pp. 28–60, 1996.
[139] F. Kwasniok, "Reduced atmospheric models using dynamically motivated basis functions," *J. Atmos. Sci.*, vol. 64, pp. 3452–3474, 2007.
[140] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*. London: Academic Press, 1981.

E-mail address: dkondras@atmos.ucla.edu

(DK,MDC,MG) Department of Atmospheric & Oceanic Sciences and Institute of Geophysics & Planetary Physics, 405 Hilgard Ave, Box 951565, 7127 Math Sciences Bldg. University of California, Los Angeles, CA 90095-1565, U.S.A.

(MG) Geosciences Department and Laboratoire de Météorologie Dynamique (CNRS and IPSL), École Normale Supérieure, F-75231 Paris Cedex 05, FRANCE.