Data-driven closures for stochastic dynamical systems

Catherine Brennan[a], Daniele Venturi[a,*]

[a]*Department of Applied Mathematics and Statistics*
University of California Santa Cruz
Santa Cruz, CA 95064

Abstract

In this paper we develop a new data-driven closure approximation method to compute the statistical properties of quantities of interest in high-dimensional stochastic dynamical systems. The new method relies on estimating conditional expectations from sample paths or experimental data, and it is independent of the dimension of the underlying phase space. We also address the important question of whether enough useful data is being injected into the reduced-order model governing the quantity of interest. To this end, we develop a new paradigm to measure the information content of data based on the numerical solution of hyperbolic systems of equations. The effectiveness of the proposed new methods is demonstrated in applications to nonlinear dynamical systems and models of systems biology evolving from random initial states.

1. Introduction

High-dimensional stochastic dynamical systems arise naturally in many areas of engineering, physical sciences and mathematics. Whether it is a physical system being studied in a lab or an equation being solved on a computer, the full state of the system is often intractable to handle in all its complexity. Instead, it is often desirable to reduce such complexity by moving from a full model of the dynamics to a reduced-order model that involves only the observables of interest. Such observables may represent specific features of a stochastic system, e.g., the sensitivity of tumor populations to chemo-treatment in stochastic models tumoral cell growth [1, 10], or the viscous dissipation in intertial range of fully developed turbulence [19, 26]. The dynamics of the observables may be simpler than that of the entire system, although the underlying law by which they evolve in space and time is often quite complex. Nevertheless, approximation of such law can in many cases allow us to avoid performing simulation of the full system and solve directly for the quantities of interest. In this paper, we aim at providing a new general framework to compute the probability density function (PDF) of such quantities of interest. To this end, we will employ formally exact PDF evolution equations and compute the unknown terms based on accurate data-driven closure approximations. To introduce the methodology, consider the following N-dimensional dynamical system evolving on a smooth manifold $\mathcal{M} \subseteq \mathbb{R}^N$

$$\begin{cases} \dfrac{d\boldsymbol{x}}{dt} = \boldsymbol{G}(\boldsymbol{x}), \\ \boldsymbol{x}(0) = \boldsymbol{x}_0(\omega). \end{cases} \tag{1}$$

Here $\boldsymbol{x_0}(\omega)$ is a random initial state with known joint probability density function $p(\boldsymbol{x}_0)$. Non-autonomous systems driven by finite-dimensional random noise can be always written in the form (1), by augmenting

*Corresponding author
Email address: venturi@ucsc.edu (Daniele Venturi)

the number of phase variables (see, e.g., [29]). Suppose we are interested in the dynamics of a real-valued phase space function

$$u(\boldsymbol{x}) = \mathcal{M} \to \mathbb{R} \qquad \text{(observable)}. \tag{2}$$

In models of population biology, this phase space function may be represented by the population of a prey species, e.g., by the first component of the nonlinear predator-prey system (1). In this case we set $u(\boldsymbol{x}(t)) = x_1(t)$. The exact dynamics of any observable in the form (2) can be expressed in terms of a semigroup of linear operators as [5, 8]

$$u(\boldsymbol{x}(t, \boldsymbol{x}_0)) = \exp\left(t\mathcal{K}(\boldsymbol{x}_0)\right) u(\boldsymbol{x}_0), \qquad \text{where} \qquad \mathcal{K}(\boldsymbol{x}_0) = \sum_{k=1}^{N} G_k(\boldsymbol{x}_0) \frac{\partial}{\partial x_{0k}}. \tag{3}$$

Here, $\boldsymbol{x}(t, \boldsymbol{x}_0)$ represents the flow map [34] generated by the system (1). The linear operator $\exp(t\mathcal{K})$ is known as the Koopman operator [15, 17]. Differentiation of (3) with respect to time yields the first-order linear partial differential equation (PDE)

$$\frac{\partial u(t, \boldsymbol{x}_0)}{\partial t} = \mathcal{K}(\boldsymbol{x}_0) u(t, \boldsymbol{x}_0). \tag{4}$$

Example 1.1. By setting $u(\boldsymbol{x}(t, \boldsymbol{x}_0)) = x_i(t, \boldsymbol{x}_0)$ for $i = 1, ..., N$, we obtain the following N-dimensional system of linear PDEs

$$\frac{\partial \boldsymbol{x}(t, \boldsymbol{x}_0)}{\partial t} = \boldsymbol{G}(\boldsymbol{x}_0) \cdot \nabla \boldsymbol{x}(t, \boldsymbol{x}_0), \tag{5}$$

where the gradient is with respect to the variables $\boldsymbol{x}_0$. This system, together with the initial condition $\boldsymbol{x}(0, \boldsymbol{x}_0) = \boldsymbol{x}_0$, allows us to compute the flow map generated by (1).

The dual of the Koopman operator $\exp(t\mathcal{K})$ with respect to the inner product

$$\langle f, g \rangle = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\boldsymbol{x}_0) g(\boldsymbol{x}_0) p(\boldsymbol{x}_0) d\boldsymbol{x}_0 \tag{6}$$

is know as Frobenious-Perron operator. Such operator can be written in the form $\exp(t\mathcal{L})$, where

$$\mathcal{L}(\boldsymbol{x})\phi = -\nabla \cdot (\boldsymbol{G}(\boldsymbol{x})\phi(\boldsymbol{x})). \tag{7}$$

The Frobenious-Perron operator pushes forward in time the joint probability density function of the flow map $\boldsymbol{x}(t, \boldsymbol{x}_0)$, i.e.,

$$p(\boldsymbol{x}, t) = e^{t\mathcal{L}(\boldsymbol{x})} p(\boldsymbol{x}, 0). \tag{8}$$

Differentiation of (8) with respect to time yields the well-known Liouville transport equation [25, 29, 30]

$$\frac{\partial p(\boldsymbol{x}, t)}{\partial t} + \nabla \left(\boldsymbol{G}(\boldsymbol{x}) p(\boldsymbol{x}, t)\right) = 0. \tag{9}$$

Computing the numerical solution of the Liouville equation can be quite challenging due to complications with high-dimensionality, multiple scales, lack of regularity, positivity and conservation properties. From a mathematical viewpoint, (9) is a hyperbolic conservation law in as many variables as the dimension of the system (1).

Remark 1.1. By using the method of characteristics [24], it is easy to obtain the following formal solution to (9)

$$p(\boldsymbol{x}, t) = p_0\left(\boldsymbol{x}_0(\boldsymbol{x}, t)\right) . \exp\left(-\int_0^t \nabla \cdot \boldsymbol{G}\left(\boldsymbol{x}(t, \boldsymbol{x}_0)\right) d\tau\right). \tag{10}$$

Here, $p_0(\boldsymbol{x}) = p(\boldsymbol{x}, 0)$, while $\boldsymbol{x}_0(\boldsymbol{x}, t)$ denotes the inverse flow map generated by (1). This expression provides a representation of the Frobenious-Perron semigroup (8).

This paper is organized as follows. In Section 2 we develop reduced order-PDF equations for arbitrary quantities of interest (2) and discuss their mathematical properties. In particular, we discuss the closure problem arising from the dimension reduction procedure and relate it with the need of computing/estimating conditional expectations. In Section 3 we propose a robust procedure to compute conditional expectations based on sample trajectories of (1) or experimental data. Such procedure opens the possibility to compute data-driven solutions to reduced-order PDF equations and estimate the Mori-Zwanzig memory integrals [28] (Section 5). In Section 4 we develop a new paradigm measure the information content of data. This allows us to infer, in particular, whether we have enough data to accurately close the reduced-order PDF equation for the quantity of interest. In Section 6 we demonstrate the effectiveness of the proposed data-driven closure approximation method in numerical applications to a high-dimensional nonlinear system and a drug resistant malaria propagation model.

2. Reduced-order PDF equations

The Liouville equation (9) describes the exact dynamics of the joint PDF of state variables $\boldsymbol{x}(t)$. In most cases, however, we are only interested in a smaller subset of such variables, or in the observable (2) (phase space function). The probability density function of such observable can be represented as

$$p(z,t) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \delta\left(z - u(\boldsymbol{x})\right) p(\boldsymbol{x}, t) d\boldsymbol{x}, \tag{11}$$

where $\delta(\cdot)$ is the Dirac's delta function (see [14, 28, 22]). Multiplying the Liouville equation by $\delta\left(z - u(\boldsymbol{x})\right)$ and integrating over all phase variables yields

$$\frac{\partial p(z,t)}{\partial t} + \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{ia(z-u(\boldsymbol{x}))} \nabla \cdot \left(\boldsymbol{G}(\boldsymbol{x}) p(\boldsymbol{x}, t)\right) \boldsymbol{x} da = 0. \tag{12}$$

Note that here we employed the Fourier representation of the Dirac delta function. Equation (12) governs the dynamics of the PDF any $u(\boldsymbol{x}(t))$. In general, it is *unclosed* in the sense that there are terms at the right hand side that cannot be computed based on $p(z,t)$ alone. If we set $\boldsymbol{u}(\boldsymbol{x}(t)) = x_k(t)$, i.e., we are interested in the k-th component of the dynamical system (1)) then (12) reduces to[1]

$$\frac{\partial p(x_k,t)}{\partial t} + \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial}{\partial x_k} \left(G_k(\boldsymbol{x}) p(\boldsymbol{x}, t)\right) dx_1 \ldots dx_{k-1} dx_{k+1} \ldots dx_N = 0. \tag{13}$$

The specific form of this equation depends on underlying dynamical system, i.e., on the nonlinear map $\boldsymbol{G}(\boldsymbol{x})$. Let us provide a simple example.

Example 2.1. Consider the Kraichnan-Orszag three-mode problem [21, 33]

$$\dot{x}_1 = x_1 x_3 \qquad \dot{x}_2 = -x_2 x_3 \qquad \dot{x}_3 = -x_1^2 + x_2^2. \tag{14}$$

[1] By using integration by parts and assuming that the joint PDF $p(\boldsymbol{x}, t)$ decays to zero sufficiently fast at infinity we obtain

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \nabla \cdot \left(\boldsymbol{G}(\boldsymbol{x}) p(\boldsymbol{x}, t)\right) dx_1 \ldots dx_{k-1} dx_{k+1} \ldots dx_N =$$
$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial}{\partial x_k} \left(G_k(\boldsymbol{x}) p(\boldsymbol{x}, t)\right) dx_1 \ldots dx_{k-1} dx_{k+1} \ldots dx_N.$$

The associated Liouville equation is

$$\frac{\partial p(\boldsymbol{x},t)}{\partial t} = -\frac{\partial}{\partial x_1}\left(x_1 x_3 p(\boldsymbol{x},t)\right) + \frac{\partial}{\partial x_2}\left(x_2 x_3 p(\boldsymbol{x},t)\right) + \frac{\partial}{\partial x_3}\left((x_1^2 - x_2^2) p(\boldsymbol{x},t)\right). \tag{15}$$

Suppose we are interested in the PDF of the first component of the system, i.e., set $u(\boldsymbol{x}(t)) = x_1(t)$ in equation (2). By integrating (15) with respect to x_2 and x_3 and assuming that $p(\boldsymbol{x},t)$ decays fast enough at infinity, we obtain

$$\frac{\partial p(x_1,t)}{\partial t} = -\frac{\partial}{\partial x_1}\int_{-\infty}^{\infty} x_1 x_3 p(x_1,x_3,t) dx_3. \tag{16}$$

From this equation we see that the evolution of $p(x_1,t)$ depends on an integral involving $p(x_1,x_3,t)$. In other words, to solve an equation of this nature, we must find a way to approximate the term involving $p(x_1,x_3,t)$. To this end, it is convenient to first transform the integral at the right hand side by using conditional probabilities. Specifically, we can write the joint PDF of $x_1(t)$ and $x_3(t)$ at time t as

$$p(x_1,x_3,t) = p(x_1,t)p(x_3|x_1,t), \tag{17}$$

where $p(x_3|x_1,t)$ is the conditional probability density of $x_3(t)$ given $x_1(t)$ [22]. A substitution of (17) into (16) yields

$$\frac{\partial p(x_1,t)}{\partial t} = -\frac{\partial}{\partial x_1}\left(x_1 p(x_1,t)\mathbb{E}[x_3(t)|x_1(t)]\right), \tag{18}$$

where

$$\mathbb{E}[x_3(t)|x_1(t)] = \int_{-\infty}^{\infty} x_3 p(x_3|x_1,t) dx_3 \tag{19}$$

is the conditional expectation of $x_3(t)$ given $x_1(t)$. As we will see in Section 3, $\mathbb{E}[x_3(t)|x_1(t)]$ can be estimated from sample trajectories of (14). Note that the reduced-order PDF equation (18) is a scalar conservation law where the (compressible) advection velocity field is $x_1\mathbb{E}[x_3(t)|x_1(t)]$.

Remark 2.1. It can be shown that the innocent-looking equation (16) is actually a PDE involving derivatives of $p(x_1,t)$ up to order infinity in the phase variable x_1. In fact, by using Kubo's cumulant expansion [16] of the joint characteristic function of $x_3(t)$ and $x_1(t)$, we can prove that

$$\int_{-\infty}^{\infty} x_3 p(x_3,x_1,t) dx_3 = \mathbb{E}[x_3(t)]p(x_1,t) + \sum_{k=1}^{\infty}(-1)^{k+1}\frac{\langle x_1(t)x_3(t)^k\rangle_c}{k!}\frac{\partial^k p(x_1,t)}{\partial x_1^k}, \tag{20}$$

where $\langle x_1(t)x_3(t)^k\rangle_c$ are classical cumulant averages[2]. A substitution of (20) into (16) yields the infinite-order PDE

$$\frac{\partial p(x_1,t)}{\partial t} = -\mathbb{E}[x_3(t)]\frac{\partial\left(x_1 p(x_1,t)\right)}{\partial x_1} + \sum_{k=1}^{\infty}(-1)^{k+1}\frac{\langle x_1(t)x_3(t)^k\rangle_c}{k!}\frac{\partial^{k+1}\left(x_1 p(x_1,t)\right)}{\partial x_1^{k+1}}. \tag{22}$$

As shown in Figure 1, the rescaled cumulants $\langle x_1(t)x_3(t)^k\rangle_c/k!$ decay slowly with k, suggesting that the cumulant expansion (20) cannot be truncated to a low-order. This implies that any reasonably accurate approximation of the reduced-order PDF equation (22) involves high-order derivatives of $p(x_1,t)$ with respect to x_1. The data-driven cumulant expansion approach we just described relies on computing sample paths

[2]The cumulant averages appearing in equation (20) are defined as

$$\langle x_1(t)x_3(t)^m\rangle_c = \mathbb{E}\left[x_1(t)x_3(t)^m\right] - \mathbb{E}\left[x_1(t)\right]\mathbb{E}\left[x_3(t)^m\right]. \tag{21}$$

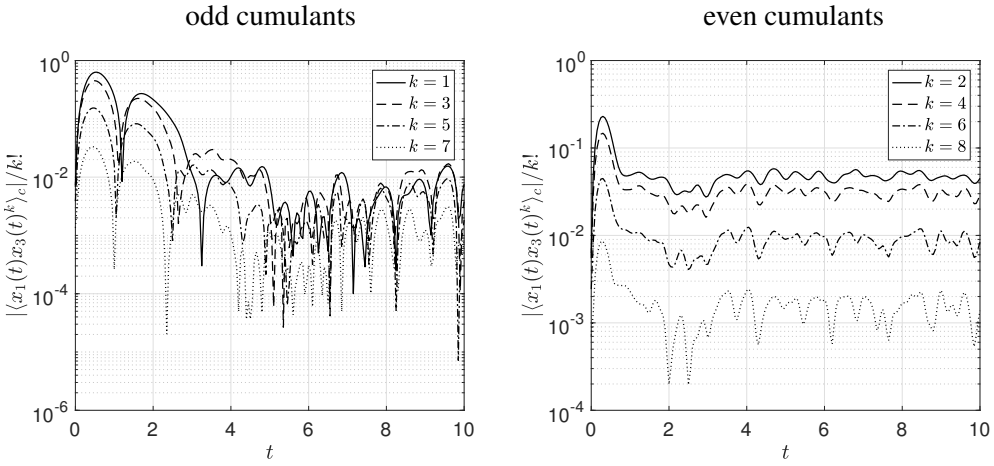

Figure 1: Kraichnan-Orszag three mode problem. Absolute values of the first 8 rescaled cumulants $\langle x_1(t)x_3(t)^k\rangle_c/k!$. The initial condition $x_i(0)$ $(i = 1, 2, 3)$ in (14) is set to be i.i.d. Gaussian with mean and variance 1. We estimated the cumulants numerically by using Monte Carlo (50000 sample paths) and then taking ensemble averages. It is seen that the odd cumulants decay slowly with k, suggesting that the cumulant expansion (20) cannot be truncated to a low order. This implies that any reasonably accurate approximation of the reduced-order equation (22) involves high-order derivatives of $p(x_1, t)$ with respect to x_1.

of (14), estimating the cumulant averages $\langle x_1(t)x_3(t)^k\rangle_c$ using ensemble averaging, and then solving the PDE (22) which potentially involves high-order derivatives of $p(x_1, t)$ with respect to x_1. Clearly this is not practical. A more effective approach relies on estimating the conditional expectation (19) directly from data and then solving the hyperbolic conservation law (18) (first-order PDE).

More generally, if we are interested in the PDF of k-th component of the system (1), then we need to express the right hand side of (13) in terms of conditional expectations, and estimate such expectations from data. If $G_k(\boldsymbol{x})$ is in the form of a sum of separable functions, i.e.,

$$G_k(\boldsymbol{x}) = \sum_{l=1}^{r} \prod_{j=1}^{N} f_{kl}^j(x_j), \tag{23}$$

then we can explicitly write (13) as

$$\frac{\partial p(x_k,t)}{\partial t} + \frac{\partial}{\partial x_k}\left(p(x_k,t) \sum_{l=1}^{r} f_{kl}^k(x_k)\mathbb{E}\left[f_{kl}^1(x_1)...f_{kl}^{k-1}(x_{k-1})f_{kl}^{k+1}(x_{k+1})...f_{kl}^N(x_N)\Big| x_k\right]\right) = 0. \tag{24}$$

Well-known examples of nonlinear dynamical systems with velocity fields in the form (23) are the Kraichnan-Orszag system (14), the Lorenz-63 [31] and the Lorenz-96 [18, 13] systems, and the the semi-discrete form of PDEs with quadratic nonlinearities such as the Burgers' equation and the Kuramoto-Sivashinsky equation. In the next Section, we discuss robust algorithms to compute conditional expectations from data, e.g., sample trajectories of (1).

3. Estimating conditional expectations from data

Computing conditional expectations from data or sample trajectories is a key step in determining accurate closure approximations of reduced-order PDF equations. A major challenge to fitting a conditional expectation is ensuring accuracy and stability. More importantly, the estimator must be flexible and effective for a wide range of numerical applications. Let us briefly recall what conditional expectations are and, more importantly, how to compute them efficiently based on sample paths. To this end, let us provide a simple example.

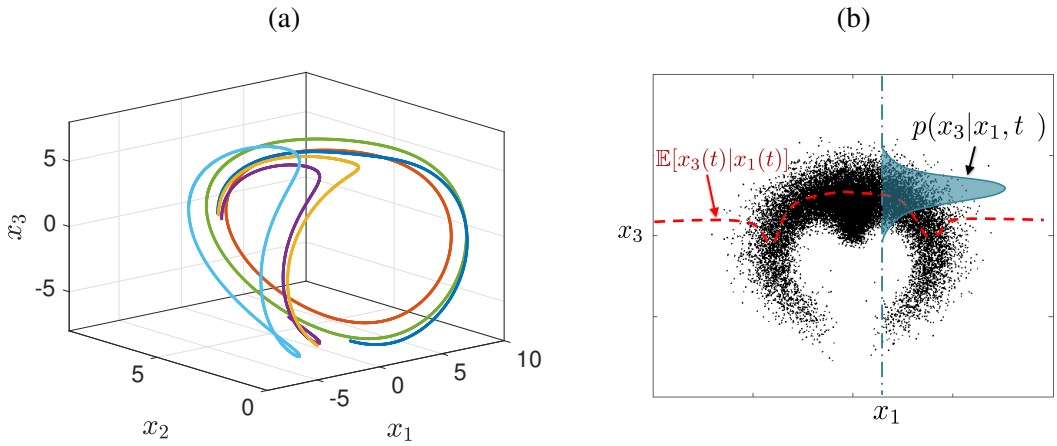

Figure 2: Kraichnan-Orszag three mode problem. (a) Sample trajectories of (14) corresponding to random samples the initial condition $\boldsymbol{x}_0$. (b) Solution samples projected into the plane (x_1, x_3) at time t. For each value of x_1, the conditional PDF $p(x_3|x_1, t)$ can be estimated based on points sitting on or lying nearby the vertical dashed line. The conditional expectation $\mathbb{E}[x_3(t)|x_1(t)]$ is the mean of such conditional PDF.

Example 3.1. Consider the random processes $x_1(t)$ and $x_3(t)$ defined by the dynamical system (14) with random initial state. The conditional expectation of $x_3(t)$ given $x_1(t)$ is defined in (19). The geometric meaning of such conditional expectation is illustrated in Figure 2. We first compute sample trajectories of (14) – see Figure 2(a) – by sampling the initial condition and evolving it in time . We then project the solution samples we obtain at time t into the plane (x_1, x_3), to obtain the scatter plot in Figure 2(b). For each value of x_1, the conditional PDF $p(x_3|x_1, t)$ can be estimated based on all points sitting on or lying nearby the vertical dashed line. The conditional expectation $\mathbb{E}[x_3(t)|x_1(t)]$ is the mean of such conditional PDF.

Hereafter we present two different approaches to estimate conditional expectations from data based on moving averages and smoothing splines. The moving average estimate is obtained by first sorting the data into bins and then computing the average within each bin. With such averages available, we can construct a smooth interpolant using the average value within each bin. Some factors that affect the bin average approximation are the bin size (the number of samples in each bin) and the interpolation method used in the final step. Another approach to estimate conditional expectations uses smoothing splines. This approach seeks to minimize a penalized sum of squares. A smoothing parameter determines the balance between smoothness and goodness-of-fit in the least-squares sense [7]. The choice of smoothing parameter is critical to the accuracy of the results. Specifying the smoothing parameter a priori is generally yields poor estimates [23]. Instead, cross-validation and maximum likelihood estimators can guide the choice the optimal smoothing value for the data set [32]. Such methods can be computationally intensive, especially when the spline estimate is performed at each time step. Other techniques to compute conditional expectations can leverage on recent developments on deep learning [12].

In Figure 3 we compare the performance of the moving average and smoothing splines approaches in approximating the conditional expectation of two jointly Gaussian random variables. Specifically, we consider the joint distribution

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)}{\sigma_1^2}\frac{(x_2-\mu_2)}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2)}\right]\right) \tag{25}$$

with parameters $\rho = 3/4$, $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = 1$ $\sigma_2 = 2$. As is well known [22], the conditional

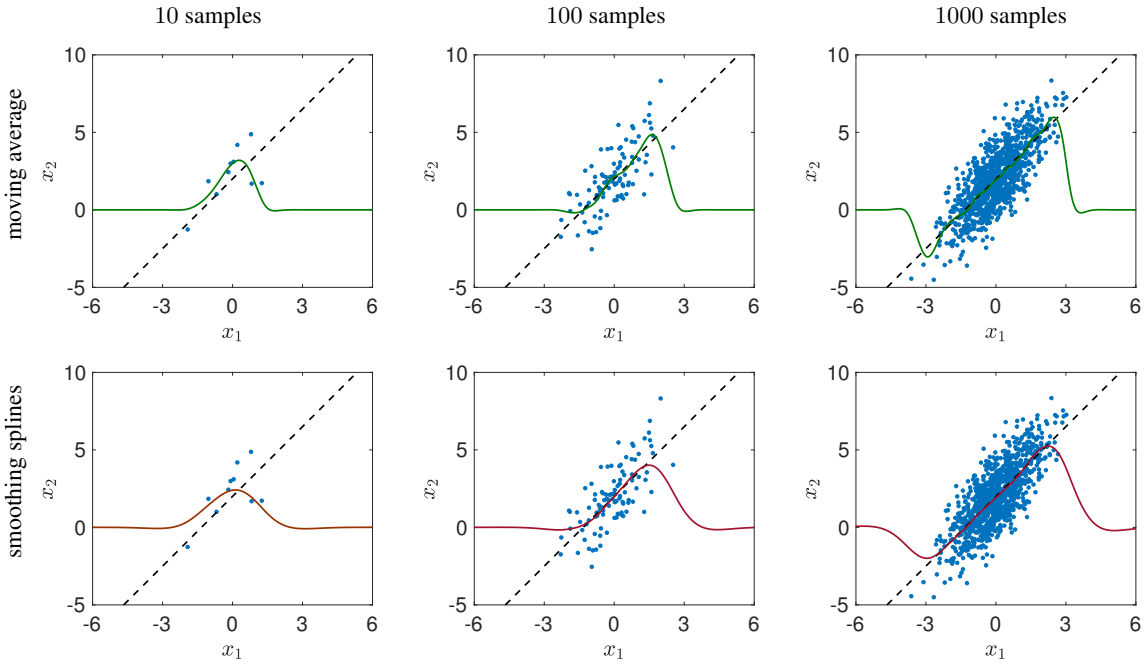

Figure 3: Numerical estimation of the conditional expectation (27) for different number of samples of (25). Shown are results obtained with moving averages (first row) and cubic smoothing splines (second row). It is seen that both methods converge to the correct conditional expectation in the active region as we increase the number of samples.

expectation of x_2 given x_1 can be expressed as[3]

$$\mathbb{E}[x_2|x_1] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1) = 2 + \frac{3}{2}x_1. \tag{27}$$

Such conditional expectation is plotted in Figure 3 (dashed line), together with the plots of the conditional average estimates we obtain with the moving average and the smoothing spline approaches for different numbers of samples. It is seen that both methods converge to the correct conditional expectation as we increase the number of samples. Note, however, that convergence is achieved in regions where the PDF (25) is not small (see the subsequent Remark 3.2). Both estimators require setting suitable parameters to compute expectations, e.g., the width of the moving average window in the moving average approach, or the smoothing parameter in the cubic spline approximant.

Remark 3.1. If the joint PDF of x_1 and x_2 is not compactly supported, then the conditional expectation is defined in the whole real line. It is computationally challenging to estimate the expectation (27) in regions where the joint PDF is very small [3]. At the same time, if we are not interested in rare events (i.e., tails of probability densities), then resolving the dynamics in such regions of small probability is not needed. This means that if we have available a sufficient number of sample trajectories [4] then we can identify the

[3]Given two random variables with joint PDF $p(x_1, x_2)$, the conditional expectation of x_2 given x_1 is defined as

$$\mathbb{E}[x_2|x_1] = \int_{-\infty}^{\infty} x_2 p(x_2|x_1)dx_2 = \frac{1}{p(x_1)}\int_{-\infty}^{\infty} x_2 p(x_1, x_2)dx_2, \tag{26}$$

where $p(x_1)$ is the marginal of $p(x_1, x_2)$ with respect to x_2.

[4]In Section 4, we will quantify what a sufficient number of trajectories is, and propose a new way to measure the information content of data based on PDEs.

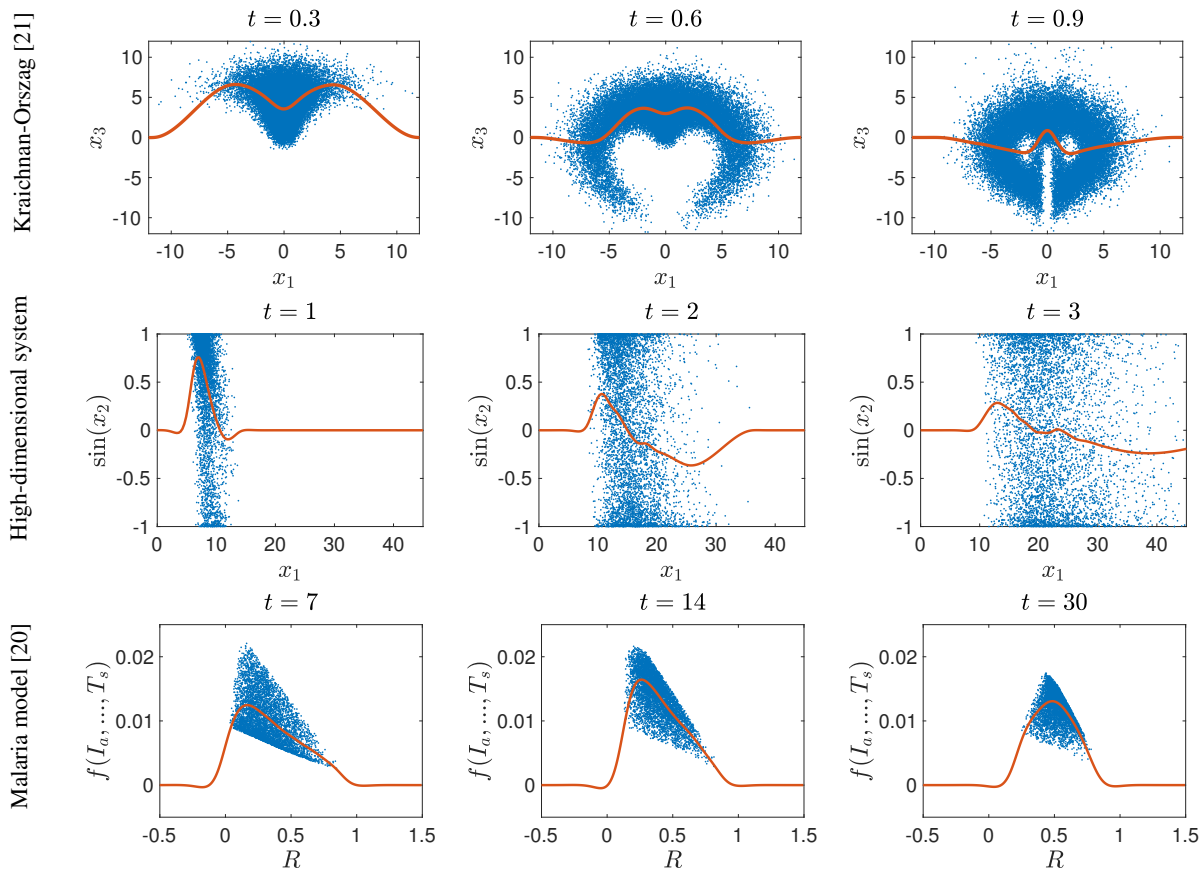

Figure 4: Data-driven smoothing spline estimation of the conditional expectations arising in the study of the dynamical systems we study in Section 6.

active regions where the dynamics are happening with high probability, and approximate the conditional expectation only within such regions [6]. Outside the active regions, we set the expectation equal to zero.

Remark 3.2. If the joint PDF of x_1 and x_2 is compactly supported, e.g. uniform in the square $[0, 1]^2$, then the conditional expectation is undefined outside the support of the joint PDF. This means, in principle, that we are not allowed to set any value for the conditional expectation outside the domain where it exists. However, a quick look at the structure of the reduced-order PDF equations we are considering in this paper, e.g., equation (24), suggests that the conditional expectation plays the role of a velocity field advecting the reduced-order PDF. Therefore, setting such velocity vector equal to zero in the regions where the reduced order PDF is very small or even undefined, does not affect the PDF propagation process. On the other hand, setting the conditional expectation equal to zero in low- or zero-probability regions greatly simplifies the mathematical discretization of PDEs in the form (24).

In Figure 4, we summarize the results we obtain by applying the smoothing spline conditional expectation estimator to several dynamical systems studied in detail in Section 6. Such systems have polynomial-type nonlinearities. The quantity of interest in each case is indicated in the x-axis of the plots. When using this approach, we must be careful to provide enough samples for the estimator to adequately capture the support of the underlying PDF. If we do not have enough samples, the estimator will not be consistent with the true conditional expectation.

4. Measuring the information content of data with PDEs

In this section, we address the important question of whether enough useful data is being injected into the reduced-order PDF equation for the purpose of computing an accurate numerical solution. To this end, we develop a new framework based on hyperbolic systems that allows us to measure a posteriori the information content of data. The key idea relies on developing evolution equations for the unclosed terms (conditional expectations) appearing in the reduced-order PDF equations. As we will see, such equations are generally unclosed, i.e., they require data to be solved, but they have the important property that their solution can be compared with data. This allows us to measure quantitatively whether we have enough data to reliably compute the closure approximation. In other words, we can measure the information content of data by solving a hyperbolic system of PDEs.

To describe the method, let us consider again the Kraichnan-Orszag system (14). As before, suppose we are interested in the phase space function $u(\boldsymbol{x}) = x_1$ (first-component of the system). We have seen that the dynamics of the PDF of x_1 is governed by the unclosed transport equation (18), where the conditional expectation $\mathbb{E}[x_3(t)|x_1(t)]$ can be estimated directly from sample trajectories. Recall that such conditional expectation is defined in (19). By differentiating such expression with respect to time we obtain

$$\frac{\partial \mathbb{E}[x_3(t)|x_1(t)]}{\partial t} = -\frac{1}{p(x_1,t)}\frac{\partial p(x_1,t)}{\partial t}\mathbb{E}[x_3(t)|x_1(t)] + \frac{1}{p(x_1,t)}\int_{-\infty}^{\infty} x_3 \frac{\partial p(x_1,x_3,t)}{\partial t} dx_3. \tag{28}$$

By substituting (18) and the reduced-order PDF equation for $p(x_1, x_3, t)$ obtained by integrating the Liouville equation (15) with respect to x_2 into (28), we find

$$\begin{aligned}\frac{\partial \mathbb{E}[x_3(t)|x_1(t)]}{\partial t} =& \frac{\mathbb{E}[x_3(t)|x_1(t)]}{p(x_1,t)}\frac{\partial}{\partial x_1}\left(x_1 p(x_1,t)\mathbb{E}[x_3(t)|x_1(t)]\right) - \cdots \\ & \frac{1}{p(x_1,t)}\frac{\partial}{\partial x_1}\left(x_1 p(x_1,t)\mathbb{E}[x_3^2(t)|x_1(t)]\right) - x_1^2 + \mathbb{E}[x_2^2(t)|x_1(t)].\end{aligned} \tag{29}$$

This is the formally exact evolution equation for the conditional expectation of $x_3(t)$ given $x_1(t)$ in the Kraichnan-Orszag system. The solution to the nonlinear PDE system (18)-(29) can be computed in a data-driven setting by estimating $\mathbb{E}[x_3^2(t)|x_1(t)]$ and $\mathbb{E}[x_2^2(t)|x_1(t)]$ from sample trajectories of (14) as we discussed in Section 3. The conditional expectation $\mathbb{E}[x_3(t)|x_1(t)]$ we obtain by solving the system (18)-(29) can be then compared with its data-driven estimate. This provides an indication of whether we have sufficient data to compute an accurate closure of (18). This procedure is illustrated in Figure 5. To avoid numerical issues, we solve equation (29) only in the regions where the PDF $p(x_1, t)$ is larger than a threshold, hence the jump in the black dashed line observed in the rightmost plots of Figure 5. As we will see in the next Section, this issue can be avoided if we change the coordinates appropriately.

The methodology we just described for the Kraichnan-Orszag system can be easily generalized and applied to arbitrary nonlinear systems in the form (1), and any observable (2). In Section 6, we study two examples.

4.1. Computational aspects

Solving the nonlinear PDE system (18)-(29) numerically is not straightforward. There are indeed several subtleties and difficulties that are hard to overcome. Perhaps, the most relevant is related to the presence of terms at the right hand side of (29) multiplied by $1/p(x_1, t)$. After simplification, these terms can be written as $\partial \log(p(x_1, t))\partial x_1$. For instance, we have

$$\begin{aligned}\frac{\mathbb{E}[x_3(t)|x_1(t)]}{p(x_1,t)}\frac{\partial}{\partial x_1}\left(x_1 p(x_1,t)\mathbb{E}[x_3(t)|x_1(t)]\right) =& \mathbb{E}[x_3(t)|x_1(t)]\frac{\partial}{\partial x_1}\left(x_1\mathbb{E}[x_3(t)|x_1(t)]\right) + \cdots \\ & x_1\mathbb{E}[x_3(t)|x_1(t)]^2\frac{\partial \log(p(x_1,t))}{\partial x_1}.\end{aligned} \tag{30}$$

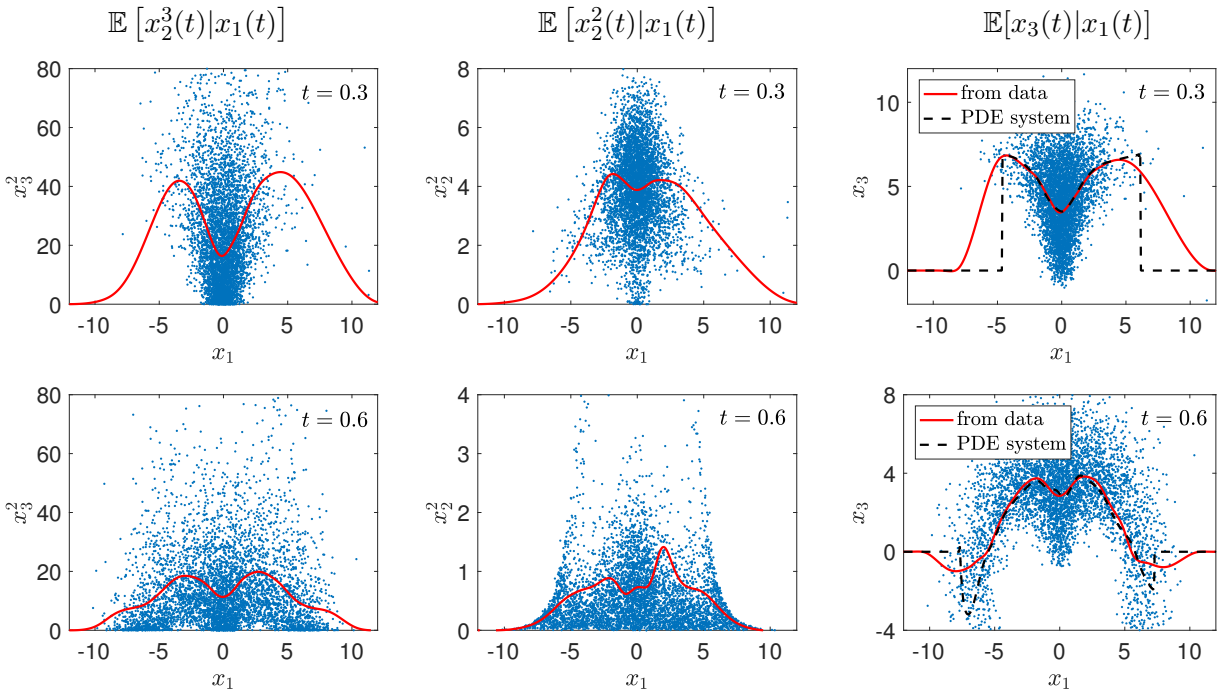

Figure 5: Kraichnan-Orszag system. Data-driven estimates of the conditional expectations $\mathbb{E}[x_3^2(t)|x_1(t)]$ (first column) and $\mathbb{E}[x_2^2(t)|x_1(t)]$ (second column) at different times based on 5000 sample trajectories. In the third column we compare the conditional expectation $\mathbb{E}[x_3(t)|x_1(t)]$ we obtain from data with the one computed by solving the nonlinear PDE system (18)-(29) with $\mathbb{E}[x_3^2(t)|x_1(t)]$ and $\mathbb{E}[x_2^2(t)|x_1(t)]$ estimated from data. It is seen that the two conditional expectations coincide within the active region of the reduced-order PDF $p(x_1, t)$. The error in $\mathbb{E}[x_3(t)|x_1(t)]$ provides a measure of the information content of the sample set obtained from (14).

Clearly, such terms can easily yield numerical overflow in regions where $p(x_1, t)$ is very small. This problem can be mitigated by using adaptive algorithms that can track the support of the PDF $p(x_1, t)$ (see, e.g., [4]). An alternative approach relies on coordinate transformation. In particular, rather than solving equation (29) for $\mathbb{E}[x_3(t), x_1(t)]$, we can solve it for the product of $\mathbb{E}[x_3(t), x_1(t)]$ and $p(x_1, t)$. This product represents the integral of $x_3 p(x_1, x_3, t)$ with respect to x_3. Let us define[5]

$$h(x_1, t) = \int_{-\infty}^{\infty} x_3 p(x_3, x_1, t) dx_3 = p(x_1, t)\mathbb{E}[x_3(t)|x_1(t)]. \tag{33}$$

The evolution equation for $h(x_1, t)$ can be obtained by following the same steps we followed to derive equation (29). This yields,

$$\frac{\partial h(x_1, t)}{\partial t} = -\frac{\partial}{\partial x_1}\left(x_1 p(x_1, t)\mathbb{E}\left[x_3(t)^2 \middle| x_1(t)\right]\right) - x_1^2 p(x_1, t) + \mathbb{E}\left[x_2(t)^2 \middle| x_1(t)\right] p(x_1, t). \tag{34}$$

[5]Recall that

$$\mathbb{E}[x_3(t)|x_1(t)] = \int_{-\infty}^{\infty} x_3 p(x_3|x_1, t) dx_3 = \frac{1}{p(x_1, t)} \int_{-\infty}^{\infty} x_3 p(x_1, x_3, t) dx_3. \tag{31}$$

Therefore,

$$\int_{-\infty}^{\infty} x_3 p(x_1, x_3, t) dx_3 = p(x_1, t)\mathbb{E}[x_3(t)|x_1(t)]. \tag{32}$$

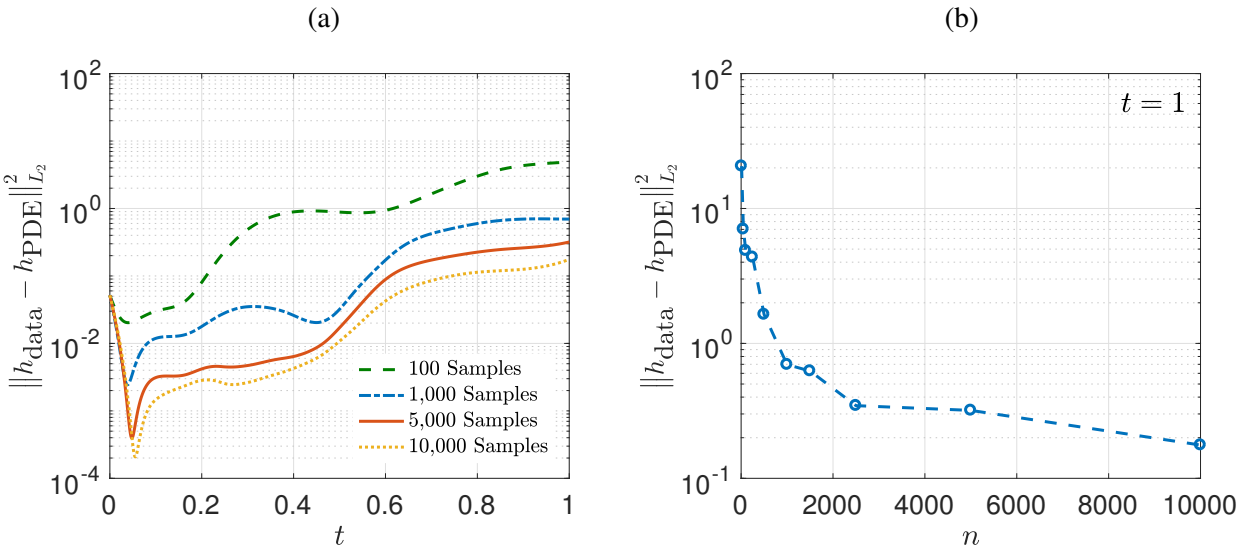

Figure 6: Kraichnan-Orszag three-mode problem. (a) Time-dependent errors in the function (33) plotted for a variety of sample sizes. (b) Decay of the error at $t = 1$ as a function of the number of sample trajectories.

On the other hand, equation (18) can be written in terms of $h(x_1, t)$ as

$$\frac{\partial p(x_1,t)}{\partial t} = -\frac{\partial}{\partial x_1}\left(x_1 h(x_1,t)\right). \tag{35}$$

The hyperbolic system (34)-(35) is linear, with two unclosed terms represented by the conditional expectations $\mathbb{E}[x_3(t)^2|x_1(t)]$ and $\mathbb{E}[x_2(t)^2|x_1(t)]$. The solution can be computed in a data-driven setting by estimating these conditional expectations from sample trajectories by using the methods of Section 3. The advantages of solving (34)-(35) over solving directly (18) rely on the fact that once $h(x_1, t)$ and $p(x_1, t)$ are available, then we can immediately compute the conditional expectation $\mathbb{E}\left[x_3(t)|x_1(t)\right])$ using (33), and compare it with the data-driven estimate. This provides a measure of the information content of sample trajectories for the particular phase space function we are interested in, i.e., $u(\boldsymbol{x}(t)) = x_1(t)$ in this case. In Figure 6, we plot the time-dependent L_2 error between the function (33) we obtain from data and the one we obtain by solving the PDE system (34)-(35). The benchmark solution of $h(x_1, t)$ is obtained by estimating $\mathbb{E}\left[x_3(t)|x_1(t)\right])$ from data and then multiplying it by an accurate kernel density estimate of $p(x_1, t)$.

Remark 4.1. Equation (34) can be immediately integrated in time to obtain

$$\begin{aligned} h(x_1,t) =& h(x_1,0) - \frac{\partial}{\partial x_1}\left(x_1 \int_0^t p(x_1,\tau)\mathbb{E}\left[x_3(\tau)^2 \middle| x_1(\tau)\right] d\tau\right) - \cdots \\ & x_1^2 \int_0^t p(x_1,\tau)d\tau + \int_0^t \mathbb{E}\left[x_2(\tau)^2 \middle| x_1(\tau)\right] p(x_1,\tau)d\tau. \end{aligned} \tag{36}$$

A substitution of (36) into (35) yields a rather complicated integro-differential PDF equation for $p(x_1, t)$. Such equation has exactly the same solution as equation (18), although the unclosed terms (conditional expectations) that need to be estimated from sample trajectories are different.

In Figure 7 we plot the PDF dynamics we obtain by solving (18) with an accurate Fourier spectral method. The conditional expectation is estimated based on 5000 sample trajectories. In Figure (8) we compare the function (33) we obtain from data (30000 sample trajectories) to the numerical solution of the system of equations (34)-(35). The unclosed terms were estimated with 5000 sample paths.

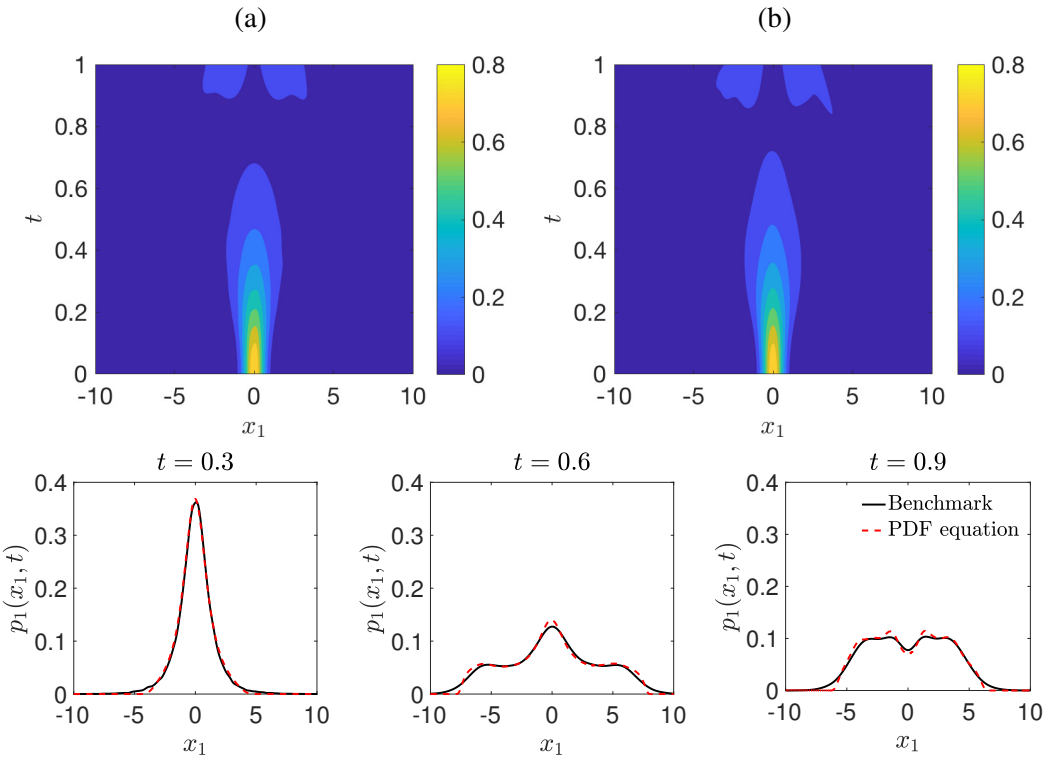

Figure 7: Kraichnan-Orszag three-mode problem. (a) Accurate kernel density estimate of $p_1(x_1, t)$ based on 30000 sample trajectories. (b) Numerical solution of (18) obtained by estimating $\mathbb{E}[x_3(t)|x_1(t)]$ with 5000 sample trajectories.

5. Data-driven estimation of the Mori-Zwanzig memory integral

The Mori-Zwanzig (MZ) formulation is a technique of irreversible statistical mechanics that allows us to formally integrate out an arbitrary number of phase variables in nonlinear dynamical systems. In doing so, we obtain exact evolution equations for quantities of interest such as macroscopic observables in high-dimensional phase spaces [27, 28, 5, 8]. To describe the method, consider the dynamical system (1) and assume that the components of the random initial state $\boldsymbol{x}_0(\omega)$ are statistically independent[6]. Furthermore, suppose we are interested in the PDF of the first component of the system, i.e., $p(x_1, t)$. To study the dynamics of such PDF, we define the following projection operator

$$Pf(\boldsymbol{x}) = \langle f(\boldsymbol{x}) \rangle \prod_{j=2}^{n} p(x_j, 0) \qquad \langle f \rangle = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\boldsymbol{x}) \prod_{j=2}^{n} dx_j \tag{37}$$

as well as the complementary projection $Q = I - P$. Note that P sends the joint PDF $p(\boldsymbol{x}, t)$ into the separated state

$$Pp(\boldsymbol{x}, t) = p(x_1, t)p(x_2, 0) \cdots p(x_N, t). \tag{38}$$

Moreover, $p(x_1, t) = \langle Pp(\boldsymbol{x}, t) \rangle$. Applying P to the Liouville equation (9) and formally integrating out the orthogonal dynamics Qp yields the Mori-Zwanzig (MZ) equation [28, 27]

$$\frac{\partial p(x_1, t)}{\partial t} = \langle PLP \rangle \, p(x_1, t) + \int_0^t \left\langle PLe^{(t-s)QL}QL \right\rangle p(x_1, s) ds. \tag{39}$$

[6]The general case where the components of $\boldsymbol{x}_0(\omega)$ are not statistically independent can be treated similarly.

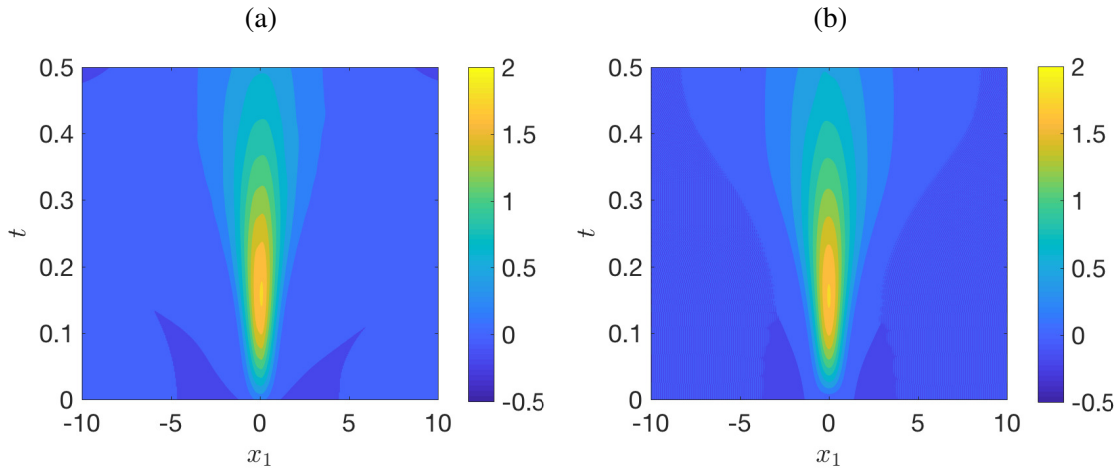

Figure 8: Kraichnan-Orszag three-mode problem. Comparsion between the time evolution of the function (33) we obtain from data (30000 sample paths) (a), and from the numerical solution of the system of equations (34)-(35) (b). In the latter case, the unclosed conditional expectations in (34) were estimated with smoothing splines and 5000 samples.

As demonstrated in [8], there exists a duality between the MZ-PDF equation (39) and the more classical MZ formulation in the space of observables (see [11, 5]). Such duality is the same that pairs the Frobenious-Perron and the Koopman operators we discussed in Section 1. Equation (39) is very challenging to solve. One of the main mathematical difficulties is the evaluation of the memory integral (second term at the right hand side). This term arises from purely formal mathematical manipulations, i.e., by using the variation of constants formula or the Dyson formula (see [35, 28, 5]). Hence, the memory integral does not incorporate any information about the *structure* of the dynamical system (1), i.e., it holds for any Liouville operator L.

At this point, we notice that a comparison between equations (39) and (13) allows us to write the MZ memory integral as[7]

$$\int_0^t \left\langle PLe^{(t-s)QL}QL \right\rangle p(x_1,s)ds = -\left\langle PLP \right\rangle p(x_1,t) - \frac{\partial}{\partial x_1}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(G_1(\boldsymbol{x})p(\boldsymbol{x},t)\right)dx_2\cdots dx_N. \tag{40}$$

The streaming term $\langle PLP\rangle\, p(x_1,t)$, can be generally expressed as

$$\langle PLP\rangle\, p(x_1,t) = -\frac{\partial}{\partial x_1}\left(p(x_1,t)\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}G_1(\boldsymbol{x})\prod_{j=2}^{N}p(x_j,0)dx_j\right). \tag{41}$$

This implies that the MZ memory integral (39) can be expressed as

$$\int_0^t \left\langle PLe^{(t-s)QL}QL \right\rangle p(x_1,s)ds = -\frac{\partial}{\partial x_1}\left(p(x_1,t)M(x_1,t)\right), \tag{42}$$

where the function $M(x_1,t)$ is

$$M(x_1,t) = \mathbb{E}\left[G_1(\boldsymbol{x}(t))|x_1(t)\right] - \mathbb{E}\left[G_1(\boldsymbol{x}(0))|x_1(0)\right]. \tag{43}$$

Clearly, $\mathbb{E}\left[G_1(\boldsymbol{x}(0))|x_1(0)\right]$ is known, since the statistical properties of the initial state are assumed to be known. On the other hand, the term $\mathbb{E}\left[G_1(\boldsymbol{x}(t))|x_1(t)\right]$, i.e., the conditional expectation of $G_1(\boldsymbol{x}(t))$ given

[7]This key observation allows us to represent the MZ memory integral in a way that depends on the specific dynamical system under consideration. In particular, if we have available an estimate of $p(x_1,t)$ and the conditional expectation $\mathbb{E}[G_1(\boldsymbol{x}(t))|x_1(t)]$ (see equation (42)), e.g., from sample paths, then we can easily estimate the MZ memory.

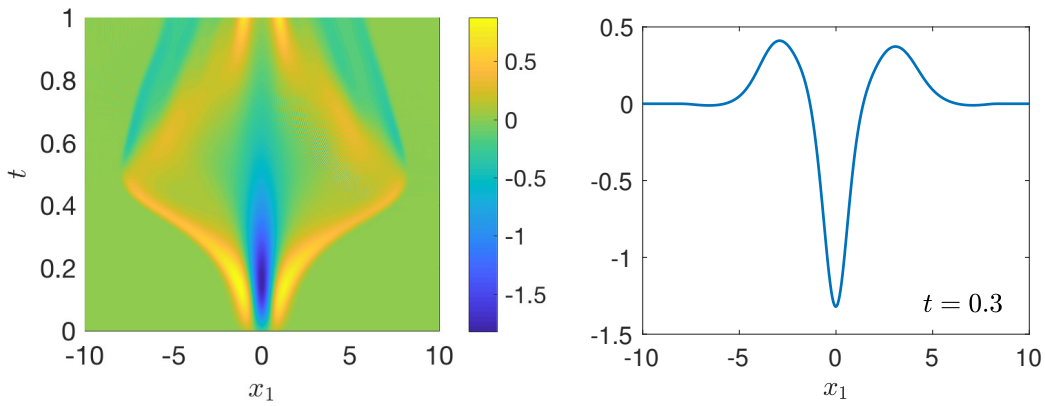

Figure 9: Kraichnan-Orszag three mode problem. Mori-Zwanzig memory integral appearing in the MZ-PDF equation (39). The memory is computed based on 5000 sample trajectories and equations (42)-(44).

$x_1(t)$ is usually not known, and it cannot be computed based on the PDF $p(x_1, t)$ alone. However, it can be estimated from sample trajectories of (1) as we discussed in Section 3. The specific form of $M(x_1, t)$ depends on the structure of the dynamical system, in particular on the form of $G_1(\boldsymbol{x})$. Let us provide a simple example.

Example 5.1. Consider the Kraichnan-Orszag system (14) evolving from a random initial state $\boldsymbol{x}(0)$ with i.i.d. components. A simple calculation shows that the function (43) in this case takes the form

$$M(x_1, t) = x_1 \left(\mathbb{E}[x_3(t)|x_1(t)] - \mathbb{E}[x_3(0)] \right). \tag{44}$$

The conditional expectation $\mathbb{E}[x_3(t)|x_1(t)]$ can be estimated from sample trajectories by using the methods we discussed in Section 3. With estimates of $p(x_1, t)$ and $\mathbb{E}[x_3(t)|x_1(t)]$ available, it is easy to compute the Mori-Zwanzig memory integral (42)-(44). In Figure 9 we plot the results we obtain with 5000 sample trajectories.

6. Numerical examples

In this section we apply the mathematical methods presented in Sections 1-5 to a high-dimensional nonlinear dynamical system and a drug-resistant malaria propagation model [20, 9].

6.1. High-dimensional nonlinear dynamics

Consider the following N-dimensional nonlinear dynamical system

$$\frac{dx_i}{dt} = -\sin(x_{i+1})x_i - Ax_i + F, \qquad i = 1, ..., N, \tag{45}$$

where $x_{n+1}(t) = x_1(t)$ (periodic boundary conditions). Depending on the value of F, A and on the number of phase variables N, this system can exhibit different behaviors. Here we set $F = 10$, $A = 0.2$ and $N = 1000$. The Liouville transport equation associated with (45) is

$$\frac{\partial p(\boldsymbol{x}, t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i} \left[(F - \sin(x_{i+1})x_i - Ax_i)\, p(\boldsymbol{x}, t) \right]. \tag{46}$$

This equation cannot be solved in a tensor product representation because of the high number of phase variables and possible lack of regularity of the solution. The evolution equation for the PDF of each phase

variable $x_i(t)$ can be obtained by integrating (46) with respect to all other variables. This yields the unclosed equation

$$\frac{\partial p(x_i,t)}{\partial t} = -\frac{\partial}{\partial x_i}\int_{-\infty}^{\infty}\left[(F-\sin(x_{i+1})x_i - Ax_i)\,p(x_i,x_{i+1},t)\right]dx_{i+1}. \tag{47}$$

We can write (47) equivalently as

$$\frac{\partial p(x_i,t)}{\partial t} = -F\frac{\partial p(x_i,t)}{\partial x_i} + A\frac{\partial (x_i p(x_i,t))}{\partial x_i} - \frac{\partial}{\partial x_i}x_i\int_{-\infty}^{\infty}\sin(x_{i+1})p(x_i,x_{i+1},t)dx_{i+1}. \tag{48}$$

Note that all equations for $p(x_i,t)$ have the same structure, independently of i. This means that if the random initial state $\boldsymbol{x}_0$ has i.i.d. components, then the evolution of each $p(x_i,t)$ does not depend on i, i.e., it is the same for all $i = 1,...,N$. A similar conclusion holds for the joint distributions $p(x_i,x_{i+1},t)$, which satisfy the equations

$$\begin{aligned}\frac{\partial p(x_i,x_{i+1},t)}{\partial t} =& -\frac{\partial}{\partial x_i}\left[(F-\sin(x_{i+1})x_i - Ax_i)\,p(x_i,x_{i+1},t)\right] - \cdots \\ & \frac{\partial}{\partial x_{i+1}}\int_{-\infty}^{\infty}\left[(F-\sin(x_{i+2})x_{i+1} - Ax_{i+1})\,p(x_i,x_{i+1},x_{i+2},t)\right]dx_{i+2}.\end{aligned} \tag{49}$$

Without loss of generality, let us set $i = 1$ in equation (48) and express the integral in terms of the conditional expectation of $\sin(x_2(t))$ given $x_1(t)$. This yields

$$\frac{\partial p(x_1,t)}{\partial t} = \frac{\partial}{\partial x_1}\left(x_1 p(x_1,t)\mathbb{E}\left[\sin(x_2(t))|x_1(t)\right]\right) + \frac{\partial}{\partial x_1}\left[(Ax_1 - F)p(x_1,t)\right], \tag{50}$$

where

$$\mathbb{E}\left[\sin(x_2(t))|x_1(t)\right] = \int_{-\infty}^{\infty}\sin(x_2)p(x_2|x_1,t)dx_2. \tag{51}$$

The conditional expectation (51) can be estimated from sample trajectories of (45) by using the mathematical techniques discussed in Section 3. The results are summarized in Figure 4. With the conditional expectation available, we can compute the numerical solution of (50) with a Fourier spectral method and compare it with an accurate kernel density benchmark estimate. This is done in Figure 10.

Alongside the data-driven closure of equation (50), we also studied closures based on a system of equation similar to the one we derived in Section 4(a). In this case we obtain the hyperbolic system

$$\frac{\partial p(x_1,t)}{\partial t} = \frac{\partial}{\partial x_1}\left(x_1 h(x_1,t)\right) + \frac{\partial}{\partial x_1}\left[(Ax_1 - F)p(x_1,t)\right], \tag{52}$$

$$\begin{aligned}\frac{\partial h(x_1,t)}{\partial t} =& -\frac{\partial}{\partial x_1}\left(p(x_1,t)\mathbb{E}\left[\sin(x_2(t))\left(F-\sin(x_2(t))x_1(t) - Ax_1(t)\right)|x_1(t)\right]\right) + \cdots \\ & p(x_1,t)\mathbb{E}\left[\cos(x_2(t))\left(F-\sin(x_3(t))x_2(t) - Ax_2(t)\right)|x_1(t)\right],\end{aligned} \tag{53}$$

where

$$h(x_1,t) = \int_{-\infty}^{\infty}\sin(x_2)p(x_1,x_2,t)dx_2. \tag{54}$$

The numerical solution to (52)-(53) allows us to to measure the information content of the sample trajectories we employed to compute the closure approximation (see Section 4(b)). In Figure 11, we plot the time-dependent error between the function (54) we obtain from data and the solution to the system (52)-(53). Observe that the error decreases as we increase the number of sample trajectories, suggesting that the information content increases with sample size. With accurate estimates for $p(x_1,t)$ and $\mathbb{E}[G_1(\boldsymbol{x}(t))|x_1(t)]$ available (here $G_1(\boldsymbol{x}) = \sin(x_2)x_1 - Ax_1 + F$), we can immediately compute the MZ memory integral by using equations (42)-(43). The results we obtain are shown in Figure 12.

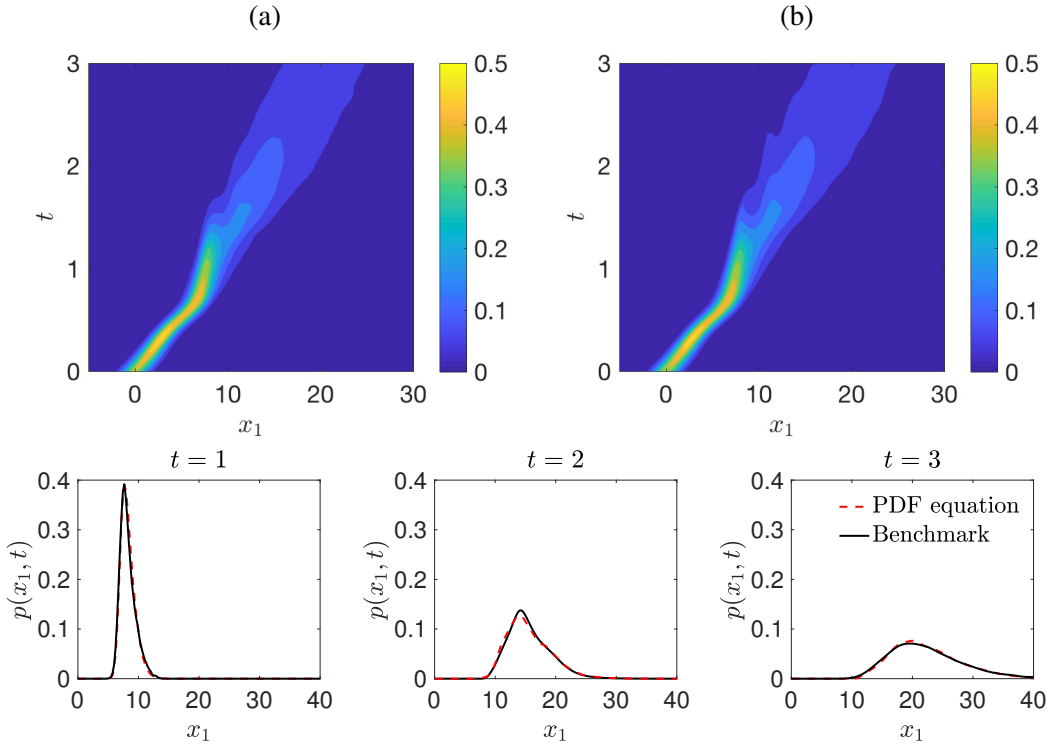

Figure 10: Nonlinear dynamical system (45). (a) Accurate kernel density estimate [2] of $p(x_1, t)$ based on 20000 sample trajectories. (b) Data-driven solution of the transport equation (50). We estimated the conditional expectation $\mathbb{E}\left[\sin(x_2(t))|x_1(t)\right]$ based on 5000 sample trajectories of (45) (see Figure 4).

6.2. Drug resistant malaria propagation model

The following dynamical system was proposed in [9, 20] to model efficacy of intermittent preventative treatment (IPT) for battling drug resistant malaria.

$$
\begin{cases}
\frac{dS}{dt} = \mu_h(1-S) - \beta_h S(M_s + kM_r) - qcS + \sigma(1-\xi)(I_a + J_a) + rT_a(1-b) + rT + wR \\
\frac{dI_s}{dt} = \lambda\beta_h M_s S + \nu I_a - I_s(pa + \sigma + \mu_h) \\
\frac{dI_a}{dt} = \beta_h M_s S(1-\lambda) - I_a(qc + \nu + \sigma + \mu_h) \\
\frac{dJ_s}{dt} = \lambda k\beta_h M_r(S + \tau T_s + \tau T + \tau T_a) + vJ_a - J_s(\sigma + \mu_h) \\
\frac{dJ_a}{dt} = k\beta_h M_r(1-\lambda)(S + \tau T_s + \tau T + \tau T_a) - J_a(\sigma + \nu + \mu_h) \\
\frac{dT_s}{dt} = paI_s - T_s(r + \tau k\beta_h M_r + \mu_h) \\
\frac{dT}{dt} = qcS - T(r + \tau k\beta_h M_r + \mu_h) \\
\frac{dT_a}{dt} = qcI_a - T_a(r + \tau k\beta_h M_r + \mu_h) \\
\frac{dR}{dt} = rT_s + brT_a + \xi\sigma(I_a + J_a) + \sigma I_s + \sigma J_s - R(w + \mu_h)
\end{cases}
\tag{55}
$$

Each phase variable here represents a proportion of the human population that is exposed to the virus (see Table 1). The remaining equations in the system govern the dynamics of the mosquito populations that are

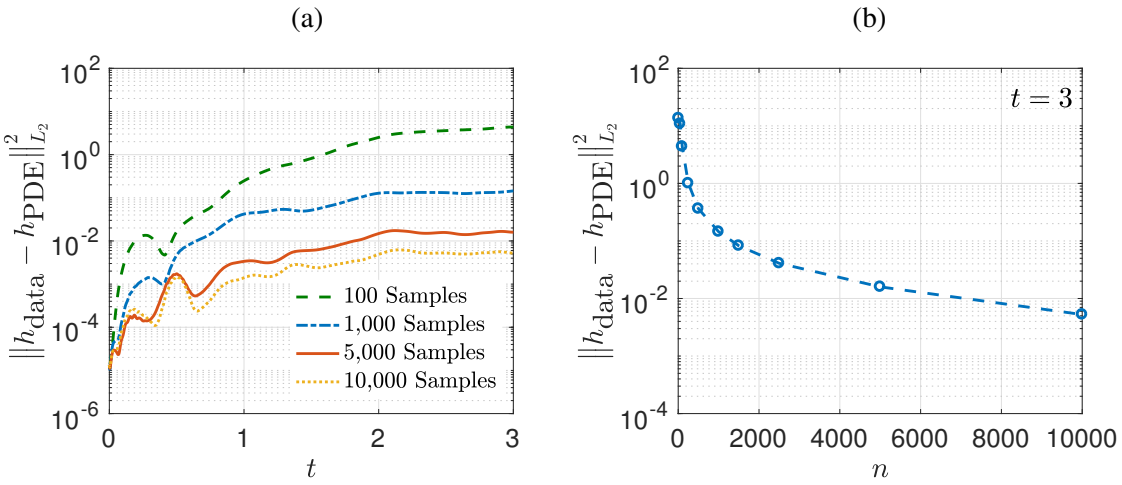

Figure 11: Nonlinear dynamical system (45). (a) time dependent errors in (54) for a variety of sample sizes. (b) Error decay at $t = 3$ versus the number of sample trajectories.

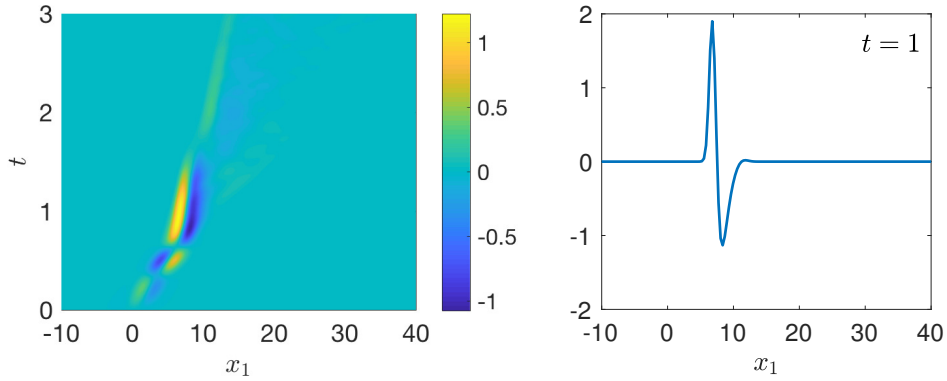

Figure 12: Mori-Zwanzig memory integral associated with the exact PDF equation for the first component of dynamical system (45) in 1000 dimensions. The initial state here is set as i.i.d. Gaussian.

responsible for spreading the virus

$$\begin{cases} \dfrac{dM_s}{dt} = \beta_m(1 - M_s - M_r)(I_a + I_s) - \mu_m M_s, \\ \dfrac{dM_r}{dt} = k\beta_m(1 - M_s - M_r)(J_a + J_s) - \mu_m M_r. \end{cases} \tag{56}$$

M_s here represents the proportion of the mosquito population infected by the sensitive strain of malaria while M_r represents the proportion infected by the resistant strain. We will use a clearance time of 6 days in correspondence with the relatively short half life of popular anti-malarial drug chlorproguanil-dapsone (CPG-DDS) [20].

The dynamics defined by (55)-(56) can be described as follows: Once a human has been infected, they move from the susceptible class, S, to one of the infected classes. The infection can be cleared naturally or via IPT. The parameter ξ represents the proportion of infections that will clear naturally and σ is the rate at which the infection will clear. The efficacy of IPT depends on the length of time the treatment remains effective after administration, r, and the factor by which treatment impacts transmissability, τ. The number of treatments per day, c, is scaled by a factor, q, to account for the fact that typically only children recieve IPT. All of the humans in groups T, T_s, and T_a have been treated with a drug that is effective against weak

Species	Variable	Description
	M_s	Proportion of mosquitoes infected with the weak strain
	M_r	Proportion of mosquitoes infected with the strong strain
	S	Proportion of susceptible humans
	I_s	Proportion of symptomatic humans infected with the weak strain
	I_a	Proportion of asymptomatic humans infected with the weak strain
	J_s	Proportion of symptomatic humans infected with the strong strain
	J_a	Proportion of asymptomatic humans infected with the strong strain
	T_s	Treated symptomatic infectious humans
	T	IPT treated susceptible humans
	T_a	IPT treated asymptomatic infected humans
	R	Proportion humans with temporary immunity

Table 1: Definition of all phase variables appearing in the nonlinear system (55)-(56).

strain of malaria, but not against the strong strain. Once an infection is cleared, an individual can move into the temporarily immune class, R, or return to the susceptible class, S. The rate at which immunity is lost is modeled in the parameter w. When members of the temporarily immune class lose their immunity, they return to the susceptible class. The parameters β_m and β_h are transmission rates from humans to parasites and vice versa, while μ_m and μ_h represent death rates for mosquitoes and humans respectively. The transmission rates are multiplied by a reduction factor, k, to account for the presence of the resistant strain. The quantity $p \in [0, 1]$ measures treatment efficacy, ν is the rate at which an asymptomatic infection progresses to a symptomatic one and λ represents the proportion of infections that will be asymptomatic. Finally, b represents the proportion of the class T_a that will gain temporary immunity, while $(1-b)$ represents the proportion that will return to the susceptible population. A thorough parametric study of (55)-(56) is presented in [9].

Suppose we are interested in a statistical description of the population of humans that have temporary immunity from the virus. Such population is represented by the phase variable $R(t)$. The evolution equation for the PDF of $R(t)$ is

$$\frac{\partial p(R,t)}{\partial t} = \frac{\partial}{\partial R}\left[(w+\mu_h)\,Rp(R,t) - h(R,t)\right], \tag{57}$$

where

$$h(R,t) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left[(T_s+bT_a)r + \sigma(\xi I_a + \xi J_a + I_s + J_s)\right] p(R, I_a, \ldots, T_s)dI_a\cdots dT_s. \tag{58}$$

Clearly, $h(R,t)$ is an unclosed term which can be written as a product of $p(R,t)$ and the conditional expectation of $(T_s(t)+bT_a(t))r + \sigma\left(\xi I_a(t) + \xi J_a(t) + I_s(t) + J_s(t)\right)$ given $R(t)$, i.e.,

$$h(R,t) = p(R,t)\mathbb{E}\left[(T_s(t)+bT_a(t))r + \sigma\left(\xi I_a(t) + \xi J_a(t) + I_s(t) + J_s(t)\right)|R(t)\right]. \tag{59}$$

The evolution equation for $h(R,t)$ can be obtained by differentiating (58) with respect to time and using the

Lioville equation. This yields,

$$\frac{\partial h(R,t)}{\partial t} = -\frac{\partial}{\partial R}\Bigg(p(R,t)\mathbb{E}\left[\Big(r(T_s + bT_a) + \sigma(\xi I_a + \xi J_a + I_s + J_s) \Big)^2 \Big| R(t) \right] - h(R,t)R(\omega + \mu_h) \Bigg)$$
$$+ p(R,t)\mathbb{E}\Big[\xi\sigma\Big((1-\lambda)\beta_h M_s(t)S(t) - I_a(t)(qc + \nu + \sigma + \mu_h) \Big) + \ldots$$
$$\sigma\Big(\lambda\beta_h M_s(t)S(t) + \nu I_a(t) - I_s(t)(ap + \sigma + \mu_h) \Big) + \ldots$$
$$\xi\sigma\Big((1-\lambda)k\beta_h M_r(t)(S(t) + \tau T_s(t) + \tau T(t) + \tau T_a(t)) - J_a(t)(\sigma + \nu + \mu_h) \Big) + \ldots$$
$$\sigma\Big(\lambda k\beta_h M_r(t)(S(t) + \tau T_s(t) + \tau T(t) + \tau T_a(t)) + \nu J_a(t) - J_s(t)(\sigma + \mu_h) \Big) + \ldots$$
$$rb\Big(qcI_a(t) - rT_a(t) - \tau k\beta_h T_a(t)M_r(t) - \mu_h T_a(t) \Big) + \ldots$$
$$r\Big(apI_s(t) - rT_s(t) - \tau k\beta_h T_s(t)M_r(t) - \mu_h T_s(t) \Big) \Big| R(t) \Big]. \tag{60}$$

In previous applications, we have assumed that each state variable had an independent initial condition. Here, we must impose the additional constraints $S + I_s + I_a + J_s + J_a + T_s + T + T_a + R = 1$ and $M_r + M_s \leq 1$, representing conservation of humans and mosquitoes. In doing so, we reduce the degrees of freedom by one, resulting in statistically dependent initial conditions. We set $I_a(0)$, $I_s(0)$, $J_a(0)$, $J_s(0)$, and $S(0)$ to be deterministic, while $R(0)$, $T_a(0)$, $M_s(0)$ and $M_r(0)$ are random. In particular, M_s and M_r evolve from an initial Gamma distribution with shape parameter $1/2$ and scale parameter $1/4$. In Figure 13 we plot the PDF dynamics of the humans with temporary immunity we obtain from data-driven closure approximation, and compare it with an accurate benchmark PDF. A phase space analysis suggests that the PDF $p(R,t)$ is attracted by a stable node. This implies that $p(R,t)$ approaches a Dirac-delta distribution asymptotically in time. The information content of the sample trajectories of (55)-(56) can be measured, as before, by checking the error between the function (59) computed from data or from the solution of the hyperbolic system (57)-(60). The results we obtain are summarized in Figure 14. Finally, the Mori-Zwanzig memory integral associated with the reduced order equation for the PDF $p(R,t)$ (humans with temporary immunity) can be computed using (42)-(43). In Figure 15, we plot the results we obtain with 5000 sample trajectories.

7. Summary

In this paper, we developed a new data-driven method to compute the probability density function of quantities of interest in high-dimensional random systems. The method is based on estimating suitable system-dependent conditional expectations from data, e.g., sample trajectories or experimental data. We also addressed the very important question of whether enough useful data is being injected into the reduced-order PDF equation governing the quantity of interest for the purpose of computing an accurate numerical solution. To this end, we developed a new paradigm which allowed us to measure the information content of data a posteriori by solving systems of hyperbolic PDEs. We applied the proposed mathematical framework to the Kraichnan-Orszag three mode problem, to a high-dimensional nonlinear dynamical system, and to a drug resistant malaria propagation model. In all cases we found that the numerical results are in agreement with the theory we developed and they allow us to compute effectively the PDF of the quantity of interest. A question we did not address in this paper is whether the proposed data-driven method approach has advantages over probability density function estimators purely based on data, e.g., [2]. Such estimators are computationally efficient in low dimensions, but they are agnostic about the dynamics in the phase space, i.e., they do not take into account the law by which the sample trajectories evolve in time. In principle, this opens the possibility to develop new classes of estimators that leverage on the additional *information*

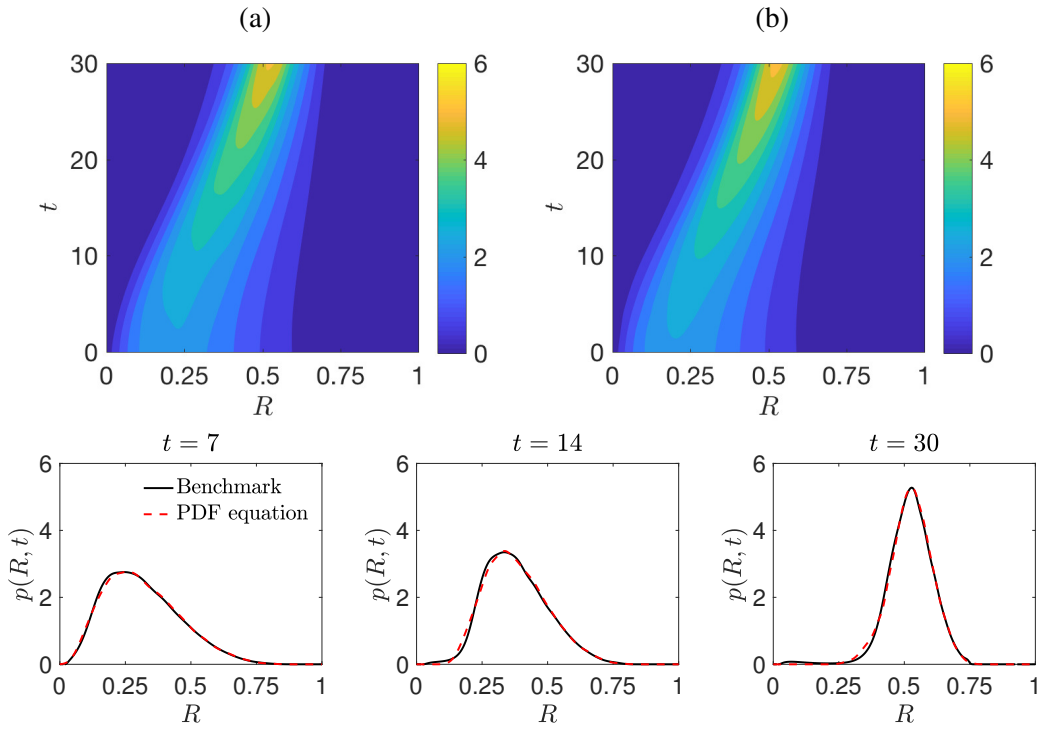

Figure 13: Drug resistant malaria propagation model. (a) Accurate kernel density estimate of $p(R, t)$ (humans with temporary immunity) based on 30000 sample trajectories. (b) Numerical solution of (57)-(58) obtained estimating the conditional expectation in (58) with 5000 sample trajectories. We also plot the time snapshots of the PDF $p(R, t)$ at 7 days, 14 days, and 30 days.

.

source provided by the law that governs the dynamics of the system. In this setting, PDF estimation can be formulated as a PDE-constrained optimization problem, with the constraint being the exact reduced-order PDF equation for the quantity of interest. Preliminary numerical results we obtained suggest that adding the evolution equation (PDE) in the PDF estimation process can reduce significantly the number of sample trajectories that are necessary to obtain an accurate estimation.

Acknowledgements This work was supported by DARPA grant N66001-15-2-4055 and NSF-TRIPODS grant 81389-444168.

References

[1] S. Benzekry, C. Lamont, A. Beheshti, A. Tracz, J. M. L. Ebos, L. Hlatky, and P. Hahnfeldt. Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Comput. Biol.*, 10(8):e1003800, 2014.

[2] Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *Annals of Statistics*, 38(5):2916–2957, 2010.

[3] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, 2001.

[4] H. Cho, D. Venturi, and G. E. Karniadakis. Adaptive discontinuous Galerkin method for response-excitation PDF equations. *SIAM J. Sci. Comput.*, 5(4):B890B911, 2013.

[5] A. J. Chorin, O. H. Hald, and R. Kupferman. Optimal prediction and the Mori-Zwanzig representation of irreversible processes. *Proc. Natl. Acad. Sci. USA*, 97(7):2968–2973, 2000.

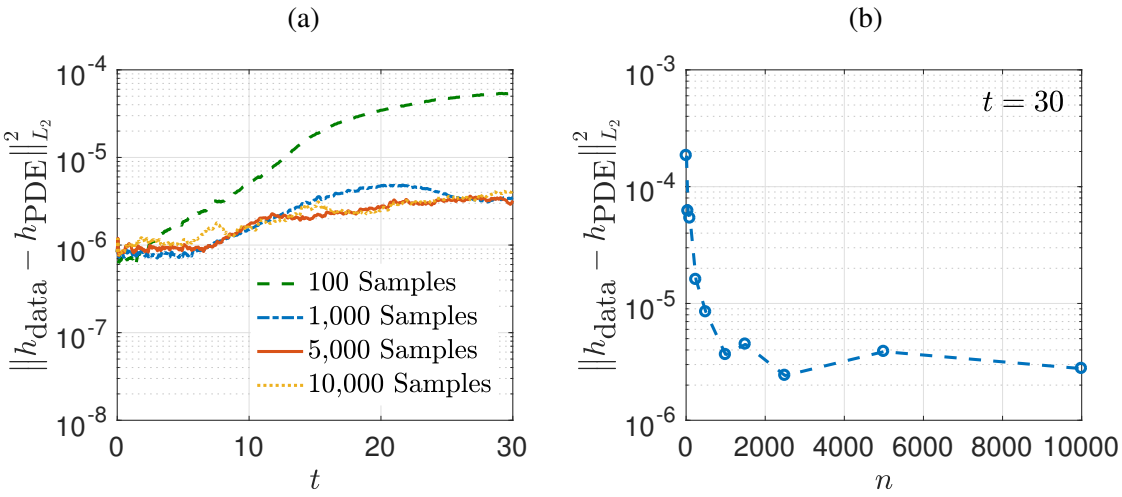

Figure 14: Drug resistant malaria propagation model. (a) Time dependent errors in (59) for a variety of sample sizes. (b) Error decay at $t = 30$ versus the number of sample trajectories.

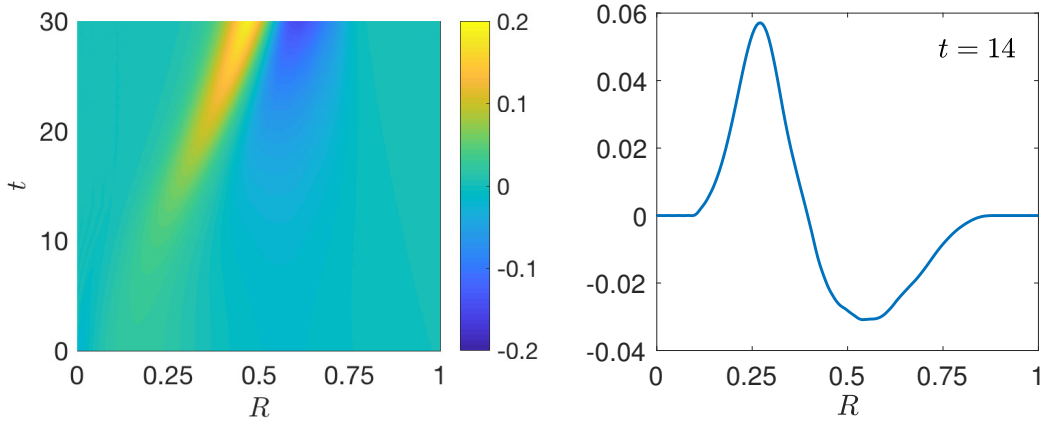

Figure 15: Drug resistant Malaria model (55)-(56). Mori-Zwanzig memory integral associated with the PDF of the phase variable representing the population of humans with temporary immunity.

[6] A. J. Chorin and X. Tu. Implicit sampling for particle filters. *PNAS*, 41:17249–17254, 2009.

[7] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1979.

[8] J. Dominy and D. Venturi. Duality and conditional expectation in the Nakajima-Mori-Zwanzig formulation. *J. Math. Phys.*, 58(082701):1–26, 2017.

[9] M. T. Ewungkem, O. Prosper, K. Gurski, C. Manore, A. Peace, and Z. Feng. Intermittent preventative treatment (ipt) and the spread of drug resistant malaria. In T. Jackson and A. Radunskaya, editors, *Applications of Dynamical Systems in Biology and Medicine*, volume 158, pages 197–233. Springer, 2015.

[10] A. Fiasconaro, B. Spagnolo, A. Ochab-Marcinek, and E. Gudowska-Nowak. Co-occurrence of resonant activation and noise-enhanced stability in a model of cancer growth in the presence of immune response. *Phys. Rev. E*, 74(4):041904 (10pp), 2006.

[11] A. Gouasmi, E. J. Parish, and K. Duraisamy. A priori estimation of memory effects in reduced-order models of nonlinear systems using the MoriZwanzig formalism. *Proc. R. Soc. A*, 473:1–24, 2017.

[12] D. Graupe. *Deep learning neural networks: design and case studies*. World Scientific, 2016.

[13] A. Karimi and M. R. Paul. Extensive chaos in the Lorenz-96 model. *Chaos*, 20(4):043105(1–11), 2010.

[14] A. I. Khuri. Applications of Dirac's delta function in statistics. *Int. J. Math. Educ. Sci. Technol.*, 35(2):185–195, 2004.

[15] B. O. Koopman. Hamiltonian systems and transformation in Hilbert spaces. *Proc. Natl. Acad. Sci. USA*, 17(5):315–318, 1931.

[16] R. Kubo. Generalized cumulant expansion method. *J. Phys. Soc. Jpn.*, 17(7):1100–1120, 1962.

[17] Q. Li, F. Dietrich, E. M. Bolt, and I. G. Kevrekidis. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the Koopman operator. *Chaos*, 10:103111, 2017.

[18] E. N. Lorenz. Predictability - A problem partly solved. In *ECMWF seminar on predictability: Volume 1*, pages 1–18. Reading, 1996.

[19] W. D. McComb. *The Physics of Fluid Turbulence*. Oxford University Press, 1990.

[20] W. P. O'Meara, D. L. Smith, and F. E. McKenzie. Potential impact of intermittent preventive treatment (IPT) on spread of drug-resistant malaria. *PLOS Medicine*, 3(5):633 – 642, 2006.

[21] S. A. Orszag and L. R. Bissonnette. Dynamical properties of truncated Wiener-Hermite expansions. *Physics of Fluids*, 10(12):26032613, 1967.

[22] A. Papoulis. *Probability, random variables and stochastic processes*. McGraw-Hill, third edition, 1991.

[23] S. B. Pope and R. Gadh. Fitting noisy data using cross-validated cubic smoothing splines. *Communications in Statistics - Simulation and Computation*, pages 349–376, 1988.

[24] H.-K. Rhee, R. Aris, and N. R. Amundson. *First-order partial differential equations, volume 1: theory and applications of single equations*. Dover, 2001.

[25] K. Sobczyk. *Stochastic differential equations: with applications to physics and engineering*. Springer, 2001.

[26] D. Venturi. The numerical approximation of nonlinear functionals and functional differential equations. *Physics Reports*, 732:1–102, 2018.

[27] D. Venturi, H. Cho, and G. E. Karniadakis. The Mori-Zwanzig approach to uncertainty quantification. In R. Ghanem, D. Higdon, and H. Owhadi, editors, *Handbook of uncertainty quantification*. Springer, 2016.

[28] D. Venturi and G. E. Karniadakis. Convolutionless Nakajima-Zwanzig equations for stochastic analysis in nonlinear dynamical systems. *Proc. R. Soc. A*, 470(2166):1–20, 2014.

[29] D. Venturi, T. P. Sapsis, H. Cho, and G. E. Karniadakis. A computable evolution equation for the joint response-excitation probability density function of stochastic dynamical systems. *Proc. R. Soc. A*, 468(2139):759–783, 2012.

[30] D. Venturi, D. M. Tartakovsky, A. M. Tartakovsky, and G. E. Karniadakis. Exact PDF equations and closure approximations for advective-reactive transport. *J. Comput. Phys.*, 243:323–343, 2013.

[31] D. Viswanath. The fractal property of the Lorenz attractor. *Physica D*, 190(1-2):115–128, 2004.

[32] G. Wahba. A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics*, 13(4):1378–1402, 1985.

[33] X. Wan and G. E. Karniadakis. Multi-element generalized polynomial chaos for arbitrary probability measures. *SIAM J. Sci. Comput.*, 28(3):901–928, 2006.

[34] S. Wiggins. *Introduction to applied nonlinear dynamical systems and chaos*. Springer, 2003.

[35] Y. Zhu, J. M. Dominy, and D. Venturi. Rigorous error estimates for the memory integral in the Mori-Zwanzig formulation. *arXiv*, (1708.02235):1–32, 2017.