Journal of Computational Physics 424 (2021) 109864

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism

Kevin K. Lin[a,*], Fei Lu[b]

[a] Department of Mathematics, University of Arizona, Tucson, AZ 85721, USA
[b] Department of Mathematics, Johns Hopkins University, Baltimore, MD 21218, USA

ARTICLE INFO

Article history:
Received 16 January 2020
Received in revised form 3 August 2020
Accepted 19 September 2020
Available online 24 September 2020

Keywords:
Model reduction
Koopman operators
Mori-Zwanzig formalism
Nonlinear time series analysis
System identification

ABSTRACT

Model reduction methods aim to describe complex dynamic phenomena using only relevant dynamical variables, decreasing computational cost, and potentially highlighting key dynamical mechanisms. In the absence of special dynamical features such as scale separation or symmetries, the time evolution of these variables typically exhibits *memory effects*. Recent work has found a variety of data-driven model reduction methods to be effective for representing such non-Markovian dynamics, but their scope and dynamical underpinning remain incompletely understood. Here, we study data-driven model reduction from a dynamical systems perspective. For both chaotic and randomly-forced systems, we show the problem can be naturally formulated within the framework of Koopman operators and the Mori-Zwanzig projection operator formalism. We give a heuristic derivation of a NARMAX (Nonlinear Auto-Regressive Moving Average with eXogenous input) model from an underlying dynamical model. The derivation is based on a simple construction we call *Wiener projection*, which links Mori-Zwanzig theory to both NARMAX and to classical Wiener filtering. We apply these ideas to the Kuramoto-Sivashinsky model of spatiotemporal chaos and a viscous Burgers equation with stochastic forcing.

1. Introduction

Unsteady fluid flow, fluctuations in power grids, neural activity in the brain: these and many other complex dynamical phenomena arise from the interaction of a large number of degrees of freedom across many orders of magnitude in space and time. But, in these and many other systems, only a relatively small subset of the dynamical variables are of direct interest or even observable. Reduced models, i.e., models that use only relevant dynamical variables to reproduce dynamical features of interest on relevant timescales, are thus of great potential utility, especially in tasks requiring repeated model runs like uncertainty quantification, optimization, and control. Moreover, relevant dynamical mechanisms are often easier to glean and understand in reduced models.

Many approaches to model reduction — also known as the closure problem in physics and reduced-order modeling in engineering — have been proposed. On one hand, a variety of analytical and computational methods have been proposed based on dynamical systems theory and statistical mechanics. These have been especially successful in situations with special dynamical features like sharp scale separation, low dimensional attractors, or symmetries; see, e.g., [1–4]. However,

[*] Corresponding author.
E-mail addresses: klin@math.arizona.edu (K.K. Lin), feilu@math.jhu.edu (F. Lu).

https://doi.org/10.1016/j.jcp.2020.109864
0021-9991/

not all scientific and engineering applications exhibit these features, and in such cases reduced models must account for memory and noise effects (see, e.g., [5,6] as well as Sect. 2). On the other hand, while purely data-driven approaches, i.e., those based on fitting generic statistical models to simulation data or physical measurements, have been quite successful in a variety of settings without sharp scale separation (see, e.g., [7–12], the dynamical systems interpretation for these methods is often unclear, and as a result their scope of applicability remain incompletely understood. In addition, a systematic understanding from the nonlinear dynamical systems point of view would provide a framework for analyzing and improving these methods.

This paper is the first step in our effort to bridge this gap; for different perspectives and approaches to similar questions, see, e.g., [13–16]. First, using Koopman operators, the Mori-Zwanzig formalism, and Wiener filtering, we propose a simple mathematical formulation of data-driven model reduction. The resulting framework links dynamical systems theory and data-driven modeling, and can serve as a starting point for systematic approximations in model reduction. In particular, we show that a variant of the NARMAX (Nonlinear Auto-Regressive Moving Average with eXogenous input) representation of stochastic processes, widely used in time series analysis and data-driven modeling (see [17,18,8] and references therein), can be derived via a construction we call "Wiener projections," which is equally applicable to either deterministic chaotic or random dynamical systems. Another consequence of our work is that for problems with time-stationary statistics, classical Wiener filtering can provides an alternative to Mori-Zwanzig as a framework for model reduction.

Organization. In Sect. 2, we recall relevant dynamical systems theory background, including a discrete-time version of the Mori-Zwanzig formalism and the NARMAX representation of stochastic processes. We also formulate the problem of data-driven driven model reduction considered in this paper. Sect. 3 describes the Wiener projection and its basic properties, and shows how it can be used to derive a variant of NARMAX. Sect. 4 is concerned with numerical implementation details, and Sect. 5 examines the application of these ideas to the Kuramoto-Sivashinsky partial differential equation (PDE) and to a stochastic Burgers equation. For the convenience of readers, we have included appendices on an alternate derivation of the Mori-Zwanzig equation (which sheds some light on its interpretation); a summary of classical Wiener filtering and the z-transform; and detailed numerical results on our two examples.

2. Data-driven model reduction in discrete time

2.1. Problem formulation and dynamical setting

We assume the full system of interest is a discrete-time dynamical system

$$X_{n+1} = F(X_n). \tag{2.1}$$

The states X_n are points in a space $\mathbb{X}$, which can be a vector space, a manifold, or a more general space. We refer to Eq. (2.1) as the *full model*. The dynamical variables of interest, or *relevant variables*, are defined by $x = \pi(X)$, π being a given function mapping points in $\mathbb{X}$ to points in d-dimensional Euclidean space $\mathbb{R}^d$, generally with $d \ll \dim(\mathbb{X})$. (The choice of π is dictated in part by the application, in part by dynamical considerations such as scale separation.) Eq. (2.1) can accommodate continuous time systems by letting F be the time-Δt solution map (for some $\Delta t > 0$) or a Poincaré map. We focus on discrete-time reduction because (i) observations are always discrete in time, and (ii) discrete-time reduced models avoid the numerical errors that come from integrating continuous time reduced models, which can be significant in chaotic systems [8,19].

By data-driven model reduction, we mean using data to construct models that use only the relevant variables. We are interested in reduced models that can (i) forecast x_n given its past history, and (ii) reproduce long-time statistics, e.g., correlations and marginal distributions. In general, parametric model reduction methods begin with a family of models with unknown parameters and observations $\widetilde{x}_n = \pi(\widetilde{X}_n)$, where $(\widetilde{X}_n)_{n=0}^N$ is a trajectory (or multiple trajectories) of the full model. One then estimates the parameters by fitting the model to the data, usually by minimizing a suitable loss function. Methods differ in their choice of models and loss functions, which can impact both model fitting and the performance of the reduced model.

A useful approach to the statistical properties of dynamical systems is to view the space of observables on $\mathbb{X}$ as forming a Hilbert space $\mathbb{H} = L^2(\mu)$ with inner product $\langle f, g\rangle = \int fg\, d\mu$. The probability distribution μ describes the long-time statistics of typical solutions of Eq. (2.1), and is *invariant*, i.e., if X_0 has distribution μ, then so does X_n for all $n > 0$. Inner products are thus naturally interpreted as steady-state correlations. Many dynamical systems of interest possess multiple "natural" invariant measures; the choice of a suitable measure is dictated in part by the application, in part by computational tractability. For example, in molecular dynamics, one may consider microcanonical, canonical, or grand canonical ensembles; for dissipative chaotic systems, relevant invariant probability measures are often singular distributions supported on strange attractors that nevertheless reflect the statistics of a set of initial conditions with positive phase space volume.

In principle, the measure μ need not be invariant. But invariance significantly simplifies the problem of data-driven model reduction, and in addition guarantees many convenient mathematical properties. Without the invariance assumption, far more data would be needed. For these reasons, we focus on stationary processes in this paper. Equivalently, we assume X_0 has distribution μ, so that (X_n) is stationary.

Recall that the *Koopman operator* is the operator M defined by $M\varphi(X) = \varphi(F(X))$. The Koopman operator advances observables forward in time: $M\varphi(X)$ gives the value of φ at the next step if the current state is X. The Koopman operator and its adjoint, the Perron-Frobenius transfer operator, describes the dynamics from the function space point of view. Much is known about their properties as operators on $\mathbb{H}$ and on other relevant Banach spaces, see, e.g., [20] or [21]. Both the Koopman and Perron-Frobenius operators have been used extensively in computational nonlinear dynamics; see, e.g., [22,23].

We will use extensively two properties:

(i) The Koopman operator is invertible when F is invertible, and $M^{-1}\varphi = \varphi \circ F^{-1}$.
(ii) With the inner product $\langle \cdot, \cdot \rangle$ above, Koopman operators are Hilbert space isometries (i.e., $\langle Mu, Mv \rangle = \langle u, v \rangle$) and unitary ($MM^* = M^*M = I$) when F is invertible [21,20].

We note that property 2 relies on the invariance of μ.

One of the uses of Koopman operators (and the Mori-Zwanzig formalism introduced in the next section) is to turn nonlinear dynamics questions into questions involving linear operators, for which mathematical analysis and formal manipulation are often easier. We will take advantage of this in Sect. 3.

2.2. Discrete-time Mori-Zwanzig formalism

The MZ formalism originally arose in classical statistical mechanics [6,5], and has been used in physical applications ranging from fluid dynamics to materials science and molecular dynamics (see, e.g., [5,24,10,25–33]). As we will discuss in Sect. 3.3, it also applies to systems with random forcing and/or (bounded) delays. Here we review a discrete MZ theory [34].

The starting point of Mori-Zwanzig formalism is the *Mori-Zwanzig equation*, which asserts that there exists a sequence of functions $\xi_1, \xi_2, \cdots : \mathbb{X} \to \mathbb{R}^d$ such that for $n \geq 0$,

$$x_{n+1} = PF(x_n) + \sum_{k=1}^{n} \Gamma_k(x_{n-k}) + \xi_{n+1}(X_0) \tag{2.2a}$$

with

$$\Gamma_k = P(\xi_k \circ F) \qquad \text{and} \qquad P\xi_n = 0. \tag{2.2b}$$

In Eq. (2.2), P can be any projection operator. The first, "Markov" term is then the "best" approximation of F by functions in the range $\mathbb{V}$ of the projection P (more on this below). The second, "memory" term captures all non-Markovian effects representable in $\mathbb{V}$. The last, "noise" term represents errors at each step, and are orthogonal to functions in $\mathbb{V}$. We present a derivation below, and an alternate derivation via a dual equation in Appendix A.

To make use of the Mori-Zwanzig equation, one must first choose a projection operator P. A common choice is the conditional expectation $(P\varphi)(x) = \mathbb{E}_\mu[\varphi(X)|\pi(X) = x]$. Another is *finite rank projection:* fix a collection of linearly independent functions $\psi_1(x), \cdots, \psi_\nu(x)$ of x, then take P to be orthogonal projection onto their linear span, i.e.,

$$P\varphi(x) = \Psi(x) \cdot \langle \Psi, \Psi \rangle^{-1} \cdot \langle \Psi, \varphi \rangle, \tag{2.3}$$

where $\langle f, g \rangle = \int f^T \cdot g \, d\mu$ for matrix-valued f and g, and the columns of $\Psi(x) = [\psi_1(x) \quad \cdots \quad \psi_\nu(x)]$ span $\mathbb{V}$. With P as in Eq. (2.3), we can write $PF = \Psi \cdot c_0$ and $\Gamma_k = \Psi \cdot c_k$ for coefficient vectors c_k. Eq. (2.2) then becomes

$$x_{n+1} = \sum_{k \geq 0} \Psi(x_{n-k}) \cdot c_k + \xi_{n+1}. \tag{2.4a}$$

Eq. (2.2b) now take the form

$$c_k = \langle \Psi, \Psi \rangle^{-1} \cdot \langle \Psi, \xi_k \circ F \rangle \tag{2.4b}$$

and

$$\langle \xi_n, \Psi \circ \pi \rangle = 0. \tag{2.4c}$$

In this paper, we will mainly consider finite rank projections and a closely related "Wiener projection" in Sect. 3. See, e.g., [5,35,6,25] for discussions of the conditional expectation and other choices.

The Mori-Zwanzig equation is an exact description of the dynamics of x_n. Without further approximation, it does not represent a reduction in model complexity. The equation does, however, highlight the interdependence of the projection P and the noise (ξ_n). To arrive at closed equations of motion for the relevant variables x_n, it is necessary to choose P so that the noise terms (ξ_n) can be effectively modeled. A common approach is to choose P to be a projection onto the slow variables. One then appeals to scale separation and other physical considerations to justify modeling (ξ_n) by a stochastic process η_n, e.g., a stationary Gaussian process. The coefficients (c_k) can be approximated by, e.g., perturbation techniques. The power spectrum of the noise and the memory kernel are related by so-called fluctuation-dissipation relations, of which Eq. (2.4b) is an example [36,6].

As a physical example, one may consider the motion of heavy particle suspended in a fluid, a problem originally studied by Smoluchowski and Einstein [37]. The "system" consists of the heavy particle and the water molecules making up the surrounding fluid. Projecting onto the particle degrees of freedom, the Markov term is given by equation of motion for a free particle, the memory term gives rise to drag due to the fluid, and the noise term represents random forces due to thermal fluctuation of the surrounding fluid.

Orthogonality conditions (e.g., Eq. (2.4c)) play a key role in MZ theory and in Wiener filtering: they are equivalent to optimality in the least squares sense. In using reduced models to generate predictions, one often assumes the driving noise (i.e., the ξ_n in Eq. (2.4)) is independent of x_m for $n > m$. Orthogonality conditions provide partial justification for this (standard) procedure. Eq. (2.4c) comes from $P\xi_n = 0$, but does *not* imply $\Psi(x_m)$ is uncorrelated with ξ_n for $n > m$. More on this in Sect. 3.

Derivation of the Mori-Zwanzig equation. The MZ equation can be driven as follows. We start with the *Dyson formula*

$$M^{n+1} = \sum_{k=0}^{n} M^{n-k} PM(QM)^k + (QM)^{n+1}, \tag{2.5}$$

where, as before, M is the Koopman operator and P is a projection on $\mathbb{H}$ whose range $\mathbb{V}$ are functions that depend *only* on the relevant variables x; and $Q = I - P$ is the orthogonal projection. Eq. (2.5) is readily proved by induction. To see how Eq. (2.4) follows from Eq. (2.5), apply both sides of Eq. (2.5) to the observation function π and evaluate at X_0, yielding

$$\underbrace{(M^{n+1}\pi)(X_0)}_{\text{(I)}} = \underbrace{\sum_{k=0}^{n}(M^{n-k}PM(QM)^k\pi)(X_0)}_{\text{(II)}} + \underbrace{((QM)^{n+1}\pi)(X_0)}_{\text{(III)}}. \tag{2.6}$$

Define $\xi_n = (QM)^n\pi$, so that $P\xi_n = 0$ for $n \geq 1$. For Term (I), the definition of the Koopman operator M gives $\pi(F^{n+1}(X_0)) = \pi(X_{n+1}) = x_{n+1}$. For Term (III), we have (by definition) $\xi_{n+1}(X_0)$. For Term (II), we have

$$(M^{n-k}PM(QM)^k\pi)(X_0) = (PM(QM)^k\pi)(X_{n-k})$$

as before. Since $M(QM)^k\pi = M\xi_k = \xi_k \circ F$ and the range of P consists of functions of $x = \pi(X)$, we get

$$(M^{n-k}PM(QM)^k\pi)(X_0) = P(\xi_k \circ F)(x_{n-k}).$$

Combining all these and $PQ = 0$ yields Eq. (2.2).

2.3. NARMAX modeling

Whereas MZ theory seeks systematic derivations of reduced models, NARMAX (Nonlinear Auto-Regressive Moving Average with eXogenous input) is a generic approach to parametric data-driven modeling of time series [18,38,17]. A common version of the NARMAX model is

$$x_{n+1} = f(x_n) + z_n, \tag{2.7a}$$

$$\begin{aligned} z_n + a_{p-1}z_{n-1} + \cdots + a_0 z_{n-p} &= d_q w_n + \cdots + d_0 w_{n-q} \\ &\quad + \Psi(x_n)\cdot c_1 + \cdots + \Psi(x_{n-r})\cdot c_r, \end{aligned} \tag{2.7b}$$

where f and Ψ are given functions, and the w_i are independent identically distributed (IID) random variables, usually assumed to be Gaussian (as we do here). One can view $x_{n+1} = f(x_n)$ as a crude predictor of x_{n+1}, and Eq. (2.7b) a corrector based on a model of the residuals z_n. Note that like the MZ equation, Eq. (2.7) is non-Markovian.

In applications of NARMAX, the main task of the would-be modeler is to first choose the forms of f, Ψ and the orders p, q, r, then determine a_i, b_i, and d_i by minimizing a suitable loss function. One common approach to parameter estimation is least squares regression: let $\widetilde{x}_n$ denote time series obtained from the full model (either by simulation or physical measurement), and define

$$\widehat{x}_{n+1} = f(\widetilde{x}_n) + \widetilde{z}_n, \tag{2.8a}$$

$$\widetilde{z}_n + a_{p-1}\widetilde{z}_{n-1} + \cdots + a_0\widetilde{z}_{n-p} = \Psi(\widetilde{x}_n)\cdot c_1 + \cdots + \Psi(\widetilde{x}_{n-r})\cdot c_r \tag{2.8b}$$

The $\widehat{x}_{n+1}$ is the one-step prediction based on $\widetilde{x}_n, \cdots, \widetilde{x}_{n-r}$. One then tunes (a_i, b_i) to minimize the mean squared error $\sum_n \|x_n - \widehat{x}_n\|^2$, possibly in combination with regularization techniques, e.g., Tikhonov regularization or a sparsity-inducing ℓ^1 term. The moving average coefficients d_n are determined by fitting a stochastic process of the form $d_q w_n + \cdots + d_0 w_{n-q}$ to the residual.

Another approach to parameter estimation is based on maximum likelihood estimation (MLE). In this approach, one assumes the statistics of the noise (w_n), e.g., independent $N(0, I)$ random vectors, and infer the (a_i, b_i) and d_i jointly by maximum likelihood methods and variations thereof.

Whatever the method, we emphasize that the form of Eqs. (2.7) does not, by itself, determine a reduced model or a model reduction procedure. One must either specify the statistics of the noise term, or the loss function to be minimized. (And, for non-convex loss functions, the optimization procedure.) These choices can have a significant impact on the usefulness of the model so obtained.

3. Wiener projections

3.1. Definition and basic properties

We now set aside Mori-Zwanzig for a moment, and consider another way to conceptualize memory effects in model reduction based on Wiener filters [39,40]. Let u_n and v_n be two zero-mean wide-sense stationary processes. The Wiener filter is the sequence (h_n) that minimizes the mean-squared error (MSE):

$$\mathbb{E}\left(\left\|u_n - (v \star h)_n\right\|^2\right), \tag{3.1}$$

where $(v \star h)_n = \sum_k v_{n-k} \cdot h_k$ denotes convolution, with $h_n = 0$ for $n < 0$. (See Appendix B for more details.) It satisfies the orthogonality condition

$$\mathrm{cov}(v_m, \overline{r}_n) = 0\,, \quad n \geq m, \tag{3.2}$$

where r_n is the residual $u_n - \sum_k v_{n-k} \cdot h_k$, i.e., filter errors are uncorrelated with the data on which the filter output is based. Eq. (3.2) is equivalent to the minimum-MSE criterion.

We observe that the Wiener filter can be applied to model reduction as well: with X_n as in Eq. (2.1) and Ψ as before, let h_n be the causal Wiener filter for $u_n = x_{n+1} = \pi(X_{n+1})$ and $v_n = \Psi(x_n)$. We then obtain $x_{n+1} = \sum_{k\geq 0} \Psi(x_{n-k}) \cdot h_k + r_{n+1}$ with $\mathrm{cov}(\Psi(x_m), r_n) = 0$ for $n > m$ with r_n playing the role of the residual r_n in Eq. (3.2).

How is this Wiener filter view related to the MZ formalism? We now sketch an argument showing that Wiener-based model reduction is in fact a special case of the MZ equation, one with some attractive properties. Let $\Psi_n = \Psi(x_n)$, and assume F is invertible so that M is invertible and unitary. Let P_W be orthogonal projection onto the subspace

$$W \ = \ \mathrm{span}(\Psi \cup M^{-1}\Psi \cup M^{-2}\Psi \cup \cdots), \tag{3.3}$$

where $M^{-k}\Psi$ is a short-hand for $\{M^{-k}\psi_1, \cdots, M^{-k}\psi_\nu\}$. Note $P_W = P_W^*$, i.e., P_W is self-adjoint. Since $M^{-1}v \in W$ for all $v \in W$, we have

$$M^{-\ell}P_W = P_W M^{-\ell} P_W\ , \quad \ell \geq 0. \tag{3.4}$$

This implies $Q_W M^{-\ell} P_W = 0$ or, upon taking adjoints, $P_W M^{\ell} Q_W = 0$. The Dyson formula (2.5) for P_W thus simplifies:

$$M^{n+1} = \sum_{k=0}^{n} M^{n-k} P_W M (Q_W M)^k + (Q_W M)^{n+1} \tag{3.5a}$$

$$= M^n P_W M + (Q_W M)^{n+1} \tag{3.5b}$$

since $PM(QM)^k = 0$ for $k \geq 1$. Applying both sides of Eq. (3.5b) to π, we obtain

$$x_{n+1} = \sum_{k\geq 0} \Psi(x_{n-k}) \cdot h_k + \xi_{n+1}, \tag{3.6a}$$

$$\langle \xi_n, \Psi_m \rangle = 0\ , \quad n > m. \tag{3.6b}$$

Though Eq. (3.6a) and Eq. (2.4a) are formally identical, the orthogonality relation (3.6b) is strictly stronger than Eq. (2.4c). The reason is that for the finite rank projection in Sect. 2.2, the orthogonality relation means $\int \xi_n(X)^T \cdot \Psi(\pi X)\ d\mu(X) = 0$, i.e., the noise functions ξ_n are orthogonal to a finite dimensional subspace of $\mathbb{H}$. In contrast, in Eq. (3.6b), the orthogonality relation $\mathbb{E}_{X_0 \sim \mu}[\xi_n(X_0)^T \cdot \Psi(\pi X)] = 0$ means the ξ_n is (in general) orthogonal to an infinite dimensional subspace of $\mathbb{H}$, and is analogous to Eq. (3.2), where the expectation is with respect to the stationary measure μ on a suitably defined path space. The orthogonality (3.6b) is significant for two reasons. First, in stochastic models like NARMAX (2.7), one typically assumes the driving noise w_m is independent of x_n for $m > n$. While natural, this is not guaranteed by the MZ equation. Eq. (3.6b) does not imply such independence, either, but comes a step closer.[1] Second, orthogonality relations like (3.6b) are

[1] For the analogous construction with P being conditional expectation (rather than finite rank projection), one can show that the (ξ_n) are martingale differences.

equivalent to optimality in the sense of least squares. The MZ equation does not guarantee the stronger orthogonality (3.6b) because it does not guarantee optimal estimation of x_{n+1} using $\Psi(x_n), \Psi(x_{n-1}), \cdots$.

We refer to the projection P_W and the associated decomposition (3.6) as the *Wiener projection*. Two comments: first, the lack of (explicit) memory terms in Eq. (3.5b) is not surprising because we have simply incorporated all relevant memory effects in the definition of P_W itself, and also assumed the availability of that entire past history at the initial time $n = 0$, so there is nothing more for a memory term to capture. Second, though the subspace W is defined in terms of M^{-1} and its powers, in practice one does not need to compute M^{-1} or F^{-1} in working with W as one can simply keep track of the (recent) history in stepping forward the reduced model. So our formalism can be safely applied to dissipative dynamical systems, for which F^{-1} may be extremely unstable.

In addition to the orthogonality (3.6b), the Wiener projection has the following properties:

(i) Eq. (3.4) implies the existence of $h_0, h_1, \cdots$ such that Eqs. (3.6) hold, and if the vectors $\cup_{k\geq 0} M^{-k}\Psi$ are linearly independent, then the coefficients (h_k) are unique. (The coefficients h_n may be ill-conditioned functions of the data if the basis functions are nearly degenerate. This is an important but nontrivial issue, which we plan to explore in future work.)

(ii) The correlation matrices $\langle \xi_m, \Psi_n\rangle$ and $\langle \xi_m, \xi_n\rangle$ are functions of $m-n$, i.e., ξ_m and Ψ_n are jointly wide sense stationary. (The process (Ψ_n) is stationary by assumption.)

The first claim is a direct consequence of the preceding discussion. For the second claim, first we show that $\xi_n = (Q_W M)^n \pi$ is wide sense stationary: by taking adjoints in Eq. (3.4), we get $P_W M^\ell = P_W M^\ell P_W$. A short calculation[2] yields

$$M^\ell Q_W = Q_W M^\ell Q_W\ , \quad \ell \geq 0. \tag{3.7}$$

Repeated application of Eq. (3.7) yields

$$\begin{aligned}
(Q_W M)^n \pi &= Q_W M Q_W M Q_W \cdots Q_W M Q_W M \pi && (3.8a)\\
&= M^{n-1} Q_W M \pi && (3.8b)\\
&= M^{n-1}\xi_1 && (3.8c)\\
&= \xi_1 \circ F^{n-1}\ , \quad n = 1, 2, \cdots . && (3.8d)
\end{aligned}$$

Thus, $\langle \xi_m, \xi_n\rangle = \langle \xi_1 \circ F^{m-1}, \xi_1 \circ F^{n-1}\rangle$. Since the probability distribution μ is F-invariant, we have

$$\begin{aligned}
\langle \xi_1 \circ F^{m-1}, \xi_1 \circ F^{n-1}\rangle &= \int \xi_1(F^{m-1}(x)) \cdot \xi_1(F^{n-1}(x))^T \, d\mu(x)\\
&= \int \xi_1(F^{m-n}(x)) \cdot \xi_1(x)^T \, d\mu(x)\\
&= \langle \xi_1 \circ F^{m-n}, \xi_1\rangle,
\end{aligned}$$

i.e., $\xi_1, \xi_2, \cdots$ is wide sense stationary.

To see that $\langle \xi_m, \Psi_n\rangle$ is also a function of $m-n$, observe

$$\begin{aligned}
\langle \xi_m, \Psi_n\rangle &= \langle \xi_1 \circ F^{m-1}, \Psi_0 \circ F^n\rangle && (3.9a)\\
&= \langle \xi_1 \circ F^{m-n-1}, \Psi_0\rangle, && (3.9b)
\end{aligned}$$

using Eq. (3.8) and the invariance of μ. This can also be established by a more "operator-theoretic" argument: observe

$$\begin{aligned}
P_W M^{-m} (Q_W M)^n &= P_W M^{-m} M^{n-1} Q_W M && (3.10a)\\
&= P_W M^{n-m-1} Q_W M. && (3.10b)
\end{aligned}$$

(Eq. (3.10a) follows by repeated use of Eq. (3.7) with $\ell = 1$.) Using $\xi_n = M^{n-1}\xi_1$ (see Eq. (3.8c)) and the definition of P_w, we see that $\langle \Psi_m, \xi_n\rangle$ is a function of $m-n$.

[2] Since $P_W M^\ell = P_W M^\ell (P_W + Q_W) = P_W M^\ell P_W + P_W M^\ell Q_W$. Combined with $P_W M^\ell = P_W M^\ell P_W$, we have $P_W M^\ell Q_W = 0$. From this, we get $M^\ell Q_W = (P_W + Q_W) M^\ell Q_W = P_W M^\ell Q_W + Q_W M^\ell Q_W = Q_W M^\ell Q_W$.

3.2. Deriving NARMAX via rational approximations

Eq. (3.6) would not reduce computational cost unless the sum in k can be truncated. Simply keeping a small number of terms, however, may not provide a good approximation. Put another way, to use Eq. (3.6) as the basis for model reduction, it is necessary to find an effective way to parametrize the space of filters (h_n). To do this, we use an idea from MZ theory [6]. Let

$$H(z) = \sum_{n\geq 0} h_n z^{-n} \tag{3.11}$$

denote the *z-transform* of (h_n). This is the discrete-time analog of the Laplace transform; its properties are summarized in Appendix B. The z-transforms $X(z)$, $\Psi(z)$, and $\Xi(z)$ of (x_n), (Ψ_n), and (ξ_n), respectively, are similarly defined. Then using the convolution property of the z-transform (see Appendix B), we have the formal relation

$$X(z) = \Psi(Z)\cdot H(z) + \Xi(z). \tag{3.12}$$

In applications of MZ theory to, e.g., statistical physics, rational approximations of the transfer function $H(z)$ are frequently effective [36,6]. This suggests the (uncontrolled) approximation

$$H(z) \approx B(z)/A(z), \tag{3.13a}$$

with

$$A(z) = z^p + a_{p-1}z^{p-1} + \cdots + a_0 \qquad \text{and} \qquad B(z) = b_r z^r + \cdots + b_0\,. \tag{3.13b}$$

Neglecting convergence and other mathematical issues for now, if we substitute the *ansatz* $H(z) = B(z)/A(z)$ into Eq. (3.12), we obtain $X(Z) = \Psi(z)\cdot B(z)/A(z) + \Xi(z)$. This relation among z-transforms is equivalent to a recurrence relation. To see this, define $y_n = \sum_{n\geq 0} \Psi_{n-k}\cdot h_k$. Then $Y(z) = \Psi(z)\cdot H(z)$, so that $A(z)Y(z) = \Psi(z)\cdot B(z)$. Inverting the z-transform yields $y_n + a_{p-1}y_{n-1} + \cdots + a_0 y_{n-p} = \Psi_{n-p+r}\cdot b_r + \cdots + \Psi_{n-p}\cdot b_0$. Summarizing, this suggests Eq. (3.6a) with the *ansatz* $H(z) = B(z)/A(z)$ can be written

$$x_{n+1} = y_n + \xi_{n+1}, \tag{3.14a}$$

$$y_n + a_{p-1}y_{n-1} + \cdots + a_0 y_{n-p} = \Psi_{n-p+r}\cdot b_r + \cdots + \Psi_{n-p}\cdot b_0. \tag{3.14b}$$

If we set one column of Ψ to be f in Eq. (2.7), Eq. (3.14) is essentially Eq. (2.7).

Modulo transients, Eq. (3.14b) will correctly compute y_n provided the recursion is stable in the sense that bounded Ψ_n lead to bounded y_n. This holds if and only if the roots of the polynomial $A(z)$ all lie strictly within the unit circle. In this paper, we refer to the condition $h_n \to 0$ as the *decaying memory condition.* Decaying memory is necessary for Eq. (3.6) to be meaningful, for otherwise the reduced model would be sensitive to information in the distant past. We note decaying memory is necessary but not sufficient for the overall numerical stability of the reduced model.

If the decaying memory condition can be enforced, Eq. (3.14b) provides an efficient way to compute the convolution in Eq. (3.6a), at a cost of not satisfying Eq. (3.5b) exactly. As a result, there may be additional memory-like corrections. A detailed analysis of this is left for future work.

Eq. (3.12) is purely formal in that in our context, where the (x_n), (Ψ_n), and (ξ_n) are stationary time series, the z-transforms do not converge for any $z\in\mathbb{C}$. A more careful treatment uses the idea of power spectra. As this is useful later in the paper, we recall the notion here.

For a stationary stochastic process (u_n), its *spectral power density* (or simply *power spectrum*) is the function $S_{uu}(\theta) = \sum_{n=-\infty}^{\infty} C_{uu}(n)e^{in\theta}$, where $C_{uu}(n)$ is the autocovariance function (ACF) $\operatorname{cov}(u_n, \overline{u_0})$. Similarly, for two stationary stochastic processes (u_n) and (v_n), their cross power spectrum $S_{uv}(\theta)$ is defined by $\sum_{n=-\infty}^{\infty} C_{uv}(n)e^{in\theta}$, where $C_{uv}(n) = \operatorname{cov}(u_n, \overline{v_0})$ is the cross correlation function (CCF). In our context, we can view x_n, $\Psi_n = \Psi(x_n)$, and ξ_n are (possibly matrix-valued) zero-mean (wide-sense) stationary time series satisfying $x_{n+1} = \sum_{k\geq 0}\Psi_{n-k}\cdot h_k + \xi_n$ with $\operatorname{cov}(x_m, \xi_n) = \operatorname{cov}(\Psi_m, \xi_n) = 0$ for all $n > m$. Then, using the properties of power spectra and z-transforms, one can show

$$S_{xx}(\theta) = H^*(e^{-i\theta})S_{\Psi\Psi}(\theta)H(e^{-i\theta}) + H^*(e^{i\theta})S_{\Psi\xi}(\theta) + S_{\xi\Psi}(\theta)H(e^{-i\theta}) + S_{\xi\xi}(\theta). \tag{3.15}$$

Eq. (3.11) typically does not converge for all $z\in\mathbb{C}$; we assume the domain of convergence contains the unit circle, so that Eq. (3.15) makes sense.

Loss function and nonlinear regression. Eq. (3.14) does not, by itself, fully specify a dynamical model: to have a well-defined model, one needs to specify, e.g., the statistics of the (ξ_n). For example, we can approximate ξ_n by a moving average of the form $d_q w_n + \cdots + d_0 w_{n-q}$, where the w_n are independent $N(0, I)$ random vectors; this then gives a NARMA(X) representation (2.7). Alternatively, one can prescribe the properties of the (ξ_n) implicitly by specifying the loss function to be minimized, which we now discuss. We observe that the rational approximation above implies $p = q$, simplifying order selection.

Since Mori-Zwanzig aims to minimize the difference between the full and reduced models with respect to the L^2 norm, a natural choice is to minimize the mean squared error

$$\mathcal{E}(a,b) = \frac{1}{N}\sum_{n=0}^{N-1} \left\| \widetilde{x}_{n+1} - \widehat{x}_{n+1}\big(\widetilde{\Psi}_1, \cdots, \widetilde{\Psi}_n; a, b\big)\right\|^2 \tag{3.16}$$

$a = (a_{p-1}, \cdots, a_0)$ and $b = (b_q, \cdots, b_0)$ are the coefficients of $A(z)$ and $B(z)$ in Eq. (3.13a), $(\widetilde{x}_n)$ are data obtained from the full model (say by simulation), and $\widetilde{\Psi}_n = \Psi(\widetilde{x}_n)$, and where the one-step prediction $\widehat{x}_n$ is here defined by

$$\widehat{x}_{n+1}(\widetilde{\Psi}_1, \cdots, \widetilde{\Psi}_n) = \sum_{k\geq 0} \widetilde{\Psi}_{n-k} \cdot h_k \,. \tag{3.17}$$

Because of the parametrization $H(z) = B(z)/A(z)$, the mean squared error $\mathcal{E}(a,b)$ depends *nonlinearly* on a and b. This leads to two[3] possible approaches:

- *Nonlinear regression*, i.e., tuning a and b to minimize $\mathcal{E}(a,b)$ in Eq. (3.16).
- Finding h_n directly by solving a (potentially very large) linear programming problem, then finding a good rational approximation $H(z) \approx B(z)/A(z)$.

In either case, we then fit a noise model to the residuals from the nonlinear regression.

For high dimensional problems, the second approach is computationally more challenging. In this paper, we use the nonlinear regression approach. Numerical details are described in Sect. 4.

Multistep form and linear regression. Modulo transients, Eq. (3.14) is equivalent (see Appendix B) to the multistep recursion

$$x_{n+p+1} + a_{p-1}x_{n+p} + \cdots + a_0 x_{n+1} = \Psi(x_{n+r}) \cdot b_r + \cdots + \Psi(x_n) \cdot b_0 + \overline{\xi}_{n+p+1}, \tag{3.18}$$

where $\overline{\xi}_{n+p+1} = \xi_{n+p+1} + a_{p-1}\xi_{n+p} + \cdots + a_0\xi_{n+1}$. Unlike Eq. (3.14), this formulation does not introduce any auxiliary variables. The noise $(\overline{\xi}_n)$ in Eq. (3.18) is related to the (ξ_n) in Eq. (3.14) by $S_{\overline{\xi}\overline{\xi}}(\theta) = |A(e^{i\theta})|^2 S_{\xi\xi}(\theta)$. This means there is no simple orthogonality relation between $\overline{\xi}_n$ and Ψ_n. For these reasons, Eq. (3.18) is less convenient than Eq. (3.14) for model fitting. Both require p vectors $x_1, \cdots, x_p \in \mathbb{R}^d$ as initial conditions. In practice, these initial conditions can have a measurable impact on noise models; we discuss this and other implementation issues in Sect. 4.

Eq. (3.18) suggests an alternative loss function: compute the one-step predictions using

$$\widehat{x}_{n+p+1} + a_{p-1}\widetilde{x}_{n+p} + \cdots + a_0\widetilde{x}_{n+1} = \Psi(\widetilde{x}_{n+r}) \cdot b_r + \cdots + \Psi(\widetilde{x}_n) \cdot b_0, \tag{3.19}$$

and minimizing the left and right hand sides, i.e.,

$$\mathcal{E}_*(a,b) = \frac{1}{N}\sum_{n=1}^{N-1-p} \left\| \widetilde{x}_{n+p+1} - \sum_{j=0}^{p-1} a_j \widetilde{x}_{n+j+1} - \sum_{j=0}^{r} b_j \widetilde{\Psi}_{n+j}\right\|^2. \tag{3.20}$$

One can then fit the residual by a noise model, e.g., by a power spectrum method (see Sect. 4.4) or a moving average model. The difference between minimizing $\mathcal{E}(a,b)$ in Eq. (3.16) and $\mathcal{E}_*(a,b)$ above is that the latter entails only linear regression, which can be computed very quickly when the number of time lags is not large. Also, whereas Eq. (3.17) depends on the all available past history, Eq. (3.19) depends only on the past r steps. However, minimizing $\mathcal{E}_*(a,b)$ may produce such effective models because it neglects long-range correlations in the data.

Finally, we observe that in Eq. (3.18), if the sequence (ξ_n) is assumed to be IID Gaussian, the resulting model is what is often referred to as the NARMA model in time series analysis (see, e.g., [41,38]). In this case, one can infer the coefficients a and b by the conditional maximal likelihood method, which entails minimizes the cost function

$$\mathcal{E}(a,b \mid \xi_1, \cdots, \xi_p) = \frac{1}{N}\sum_{n=1}^{N-1-p} \left\| \widetilde{x}_{n+p+1} - \sum_{j=0}^{p-1} a_j (\widetilde{x}_{n+j+1} - \widetilde{\xi}_{n+j+1}) - \sum_{j=0}^{r} b_j \widetilde{\Psi}_{n+j}\right\|^2. \tag{3.21}$$

In the above, the sequence $(\widetilde{\xi}_n)_{n>p}$ can be computed recursively from data for each given pair of (a,b). This cost function is similar to $\mathcal{E}(a,b)$, and the optimization is similar to the nonlinear regression above: instead of using (h_n) above, one computes the sequence $(\widetilde{\xi}_n)_{n>p}$ in each optimization step (see [8] for more details).

[3] In standard approaches to Wiener filtering, one makes use of the power spectra S_{xx}, $S_{x\psi}$, and $S_{\psi\psi}$ and their meromorphic continuations and solves the filtering problem by Wiener-Hopf techniques (see, e.g., [40]). In the context of data-driven modeling, direct minimization of $\mathcal{E}(a,b)$ is more attractive because of the various sources of statistical error.

3.3. Random dynamical systems and systems with delays

Model reduction techniques are routinely applied to both deterministic and random dynamical systems, as well as systems with delays. The MZ formalism applies to both random dynamical systems and to discrete-time systems with bounded delays, as we now explain. Our construction here is related to the "shift operator" discussed in [42].

We first explain how the MZ formalism applies to random dynamical systems. Consider the Euler-Maruyama discretization[4] of a stochastic differential equation (SDE) of the form $\dot{u}_t = f(u_t) + \dot{w}_t$:

$$u_{n+1} = u_n + f(u_n)\triangle t + \sqrt{\triangle t}\, w_n\ , \tag{3.22}$$

where the w_n are independent $N(0, I)$ random vectors. The above has the general form

$$u_{n+1} = F(u_n, w_n). \tag{3.23}$$

Let $\underline{w} = (\cdots, w_{-1}, w_0, w_1, \cdots)$ denote the entire history of the forcing. A standard way to rewrite Eq. (3.22) as an autonomous dynamical system (Eq. (2.1) above) is to augment the state u_n with the history of the forcing $\underline{w}$. In dynamical systems language, such constructions are known as "skew products." Here we sketch the key ideas, and refer interested readers to, e.g., [43–45] for mathematical details (see also [46,47] for extensions to stochastic differential equations).

Given a forcing sequence $\underline{w}$, we define $\sigma(\underline{w})$ to be the sequence whose nth entry is w_{n+1}, i.e., $\sigma(\underline{w})_n = w_{n+1}$. In other words, $\sigma(\underline{w})$ is sequence $\underline{w}$ shifted by 1 in time. If we shift n times, then w_n is moved into position 0, so that $\pi_0(\sigma^n(\underline{w}) = w_n$, where $\pi_0(\underline{w}) = w_0$.

Using this notation, we can rewrite Eq. (3.23) as $u_{n+1} = F(u_n, \pi_0(\sigma^n(\underline{w})))$, where $\underline{w}$ is a given realization of the forcing sequence. Now denote $\underline{w}^{(n)} = \sigma^n(\underline{w})$; then $\{\underline{w}^{(n)} \mid n \in \mathbb{Z}\}$ is a sequence of forcing sequences, all related to each other by time shifts. Then

$$u_{n+1} = F\Big(u_n, \pi_0(\underline{w}^{(n)})\Big), \tag{3.24a}$$

$$\underline{w}^{(n+1)} = \sigma(\underline{w}^{(n)}). \tag{3.24b}$$

Let $\mathbb{X}$ be the space of all pairs $(u, \underline{w})$, i.e., $\mathbb{X}$ is the state space of the discretized SDE augmented with its forcing history. Then Eq. (3.24) is a dynamical system of the form Eq. (2.1), albeit one with an infinite-dimensional state space $\mathbb{X}$. This does not prevent one from applying the Mori-Zwanzig formalism. In practice, one does not need to (and generally cannot) keep track of the entire forcing history $\underline{w}$, and a fragment of it is often sufficient. Note that within this framework, observation functions Ψ can depend on both the state u_n and the forcing history $\underline{w}^{(n)}$.

Finally, we note that an invariant probability distribution μ, related in a natural way to the stationary distribution of Eq. (3.22), can be constructed on this augmented state space; see, e.g., [43,44].

As for general delay terms, for example terms of the form $\Psi(x_k, x_{k-\ell})$ for $\ell \le L$ (which appear in our model for the Burgers equation later in the paper), one can use a standard construction: as in Eq. (2.1), let F be a given dynamical system with state space $\mathbb{X}$, and replace the state space $\mathbb{X}$ by the $(L+1)$-fold cartesian product $\overline{\mathbb{X}} = \mathbb{X}^{L+1}$, and replace F by a map $\overline{F}$ on $\overline{\mathbb{X}}$ with

$$\overline{F}(\overline{X}) \ = \ \overline{F}(X_0, \cdots, X_L) \ = \ (F(X_0), X_0, \cdots, X_{L-1}) \tag{3.25}$$

for $\overline{X} = (X_0, \cdots, X_L) \in \overline{\mathbb{X}}$. This constructions can be combined with the skew product construction described earlier to handle stochastic systems with delays.

4. Numerical implementation

This section addresses the problem of fitting models of the form (3.14) to data. We take a two-step approach: we first tune the coefficients a and b of the polynomials $A(z)$ and $B(z)$, respectively, to minimize $\mathcal{E}(a, b)$ in Eq. (3.16); we then use a stationary Gaussian process to model the residuals. Sects. 4.1 and 4.2 concern the decaying memory constraint. Sect. 4.3 discusses other details of optimization, and Sect. 4.4 noise modeling.

We have implemented the algorithms described here and the examples of Sect. 5 in Julia version 1.4 [48]. For numerical optimization, we used the `NLopt.jl` package [49]. The source code is being prepared for public release, and will be available at https://github.com/kkylin.

[4] The ideas we introduce here are quite general; we focus on Euler-Maruyama for the sake of simplicity.

4.1. Decaying memory constraint and the second-order cascade

To fit a model of the form Eq. (3.6) to data, we will need to enforce the decaying memory condition $h_k \to 0$ for two reasons. First, the decaying memory condition is necessary for the reduced model to be meaningful. Second, while we can compute one-step predictions directly using Eq. (3.17), either directly or by the fast Fourier transform, the computational cost will be quite high for high dimensional problems. It would be much more efficient if we can implement the convolution indirectly by making use of Eq. (3.14), i.e., compute the one-step prediction by

$$\begin{aligned} \widehat{x}_{n+1} &= y_n, \\ y_n + a_{p-1} y_{n-1} \ + \ \cdots \ + \ a_0 y_{n-p} &= \Psi(\widetilde{x}_{n-p+r}) \cdot b_r \ + \ \cdots \ + \ \Psi(\widetilde{x}_{n-p}) \cdot b_0 . \end{aligned} \tag{4.1}$$

But as discussed earlier, we need the decaying memory condition to ensure these recursions will correctly compute y_n. The challenge is that the loss function $\mathcal{E}(a, b)$ is highly nonlinear in a and b. Because the decaying memory condition involves the roots of $A(z)$ in Eq. (3.13a), it exacerbates the problem. Our general approach is to reformulate Eq. (3.14) so that the decaying memory constraint becomes easier to implement, at the cost of making the cost function highly non-convex. We then fit reduced models to data using this representation by numerical optimization. We have found this to be sufficient for the examples in this paper, though more work needs to be done to ensure its robustness and efficiency for more general problems.

Consider a model of the form Eq. (3.14) given coefficients, and suppose for simplicity that $A(z)$ has real scalar coefficients. We begin with the observation that for a quadratic polynomial $z^2 + \alpha z + \beta$, its roots lie inside the unit disc if and only if (α, β) lies inside the triangle in the $\alpha\beta$-plane with vertices $(\pm 2, 1)$ and $(0, -1)$. That is to say, for such an $A(z)$, the decaying memory condition consists of three *linear* inequalities. To make use of this observation for non-quadratic $A(z)$, we factor $A(z)$ into a product of quadratic factors when $p = \deg(A)$ is even, and quadratic factors and one linear factor if p is odd, i.e.,

$$A(z) = \prod_{i=1}^{p/2} (z^2 + \alpha_i z + \beta_i) \qquad \text{or} \qquad A(z) = (z + \alpha_0) \prod_{i=1}^{\lfloor p/2 \rfloor} (z^2 + \alpha_i z + \beta_i). \tag{4.2}$$

In this form, the decaying memory condition is naturally expressed as a system of linear inequalities, which are easily imposed when performing numerical optimization.

In view of the convolution theorem for z-transforms, the quadratic factorization of $A(z)$ is equivalent to representing the linear filter with transfer function $1/A(z)$ as a *cascade of second-order filters*. To see this, suppose (for simplicity) that $p = 2s$. We introduce auxiliary variables (z_i^n) for $i = 1, \cdots, s$ (these variables z_i^n differ from the z in z-transforms), and suppose they satisfy

$$\begin{array}{lrcl} \text{Stage 1} & z_1^n + \alpha_1 z_1^{n-1} + \beta_1 z_1^{n-2} & = & \Psi_{n-p+r} \cdot b_r + \cdots + \Psi_{n-p} \cdot b_0 \\ \text{Stage 2} & z_2^n + \alpha_2 z_2^{n-1} + \beta_2 z_2^{n-2} & = & z_1^n \\ \vdots & & \vdots & \\ \text{Stage } s & z_s^n + \alpha_s z_s^{n-1} + \beta_s z_s^{n-2} & = & z_{s-1}^n \end{array} \quad . \tag{4.3a}$$

We claim that if

$$x_{n+1} = z_s^n + \xi_{n+1} \tag{4.3b}$$

then Eq. (4.3) is equivalent to Eq. (3.14), modulo transients. To see this, observe that (neglecting initial conditions) we have

$$\begin{aligned} (1 + \alpha_i z^{-1} + \beta_i z^{-2}) Z_i(z) &= Z_{i-1}(z) \, , \qquad i = 2, \cdots, s \\ (1 + \alpha_1 z^{-1} + \beta_1 z^{-2}) Z_1(z) &= z^{-p} \Psi(z) \cdot B(z). \end{aligned}$$

Putting it all together (and remembering $p = 2s$) gives $Z_s(z) = \Psi(z) \cdot B(z) / \Pi_{i=1}^s (z^2 + \alpha_i z + \beta_i) = \Psi(z) \cdot B(z)/A(z)$, and inverting z-transforms yields Eq. (4.3). The equivalence is up to transients because we have neglected initial conditions in this discussion, and the argument is valid only if the decaying memory condition holds. The recursion in Eq. (4.3) is explicit when $p \geq q$. In the notation of Eq. (3.14), the output of the last stage gives y_n, i.e., $y_n = z_s^n$.

Example. For $p = r = 4$, we have two stages:

$$\begin{array}{lrcl} \text{Stage 1} & z_1^n + \alpha_1 z_1^{n-1} + \beta_1 z_1^{n-2} & = & \Psi_n \cdot b_4 + \ \cdots \ + \Psi_{n-4} \cdot b_0 \\ \text{Stage 2} & z_2^n + \alpha_2 z_2^{n-1} + \beta_2 z_2^{n-2} & = & z_1^n \end{array} \tag{4.4}$$

In this case, it is easy to show directly that

$$y_n + a_3 y_{n-1} + \cdots + a_0 y_{n-4} = \Psi_n \cdot b_4 + \cdots + \Psi_{n-4} \cdot b_0 \tag{4.5}$$

where $y_n = z_2^n$ and

$$z^4 + a_3 z^3 + a_2 z^2 + a_1 z + a_0 = (z^2 + \alpha_1 z + \beta_1) \cdot (z^2 + \alpha_2 z + \beta_2)\,, \quad z \in \mathbb{C}. \tag{4.6}$$

The corresponding reduced model can be written as a system

$$\begin{aligned} x_{n+1} &= y_n + \xi_{n+1} \\ y_n &= -\left(a_3 y_{n-1} + \cdots + a_0 y_{n-4}\right) + \left(\Psi_n \cdot b_4 + \cdots + \Psi_{n-4} \cdot b_0\right). \end{aligned}$$

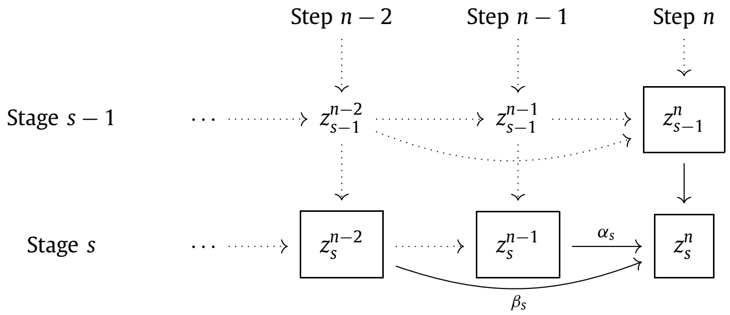
With $p = r = 0$, we have a one-step (Galerkin) recursion $x_{n+1} = \Psi_n \cdot b_0 + \xi_{n+1}$. Similarly, with $p = r = 1$, we have $x_{n+1} = y_n + \xi_{n+1}$ and $y_n = -a_0 y_{n-1} + \Psi_n \cdot b_1 + \Psi_{n-1} \cdot b_0$, and setting $a_0 = 0$ yields $x_{n+1} = \Psi_n \cdot b_1 + \Psi_{n-1} \cdot b_0 + \xi_{n+1}$.

4.2. Initializing and running cascade-form models

We use the cascade-form model to impose the decaying memory condition. This is needed both for running fitted reduced models and, as we explain later, for fitting models to data. Here, we discuss how to initialize and run such models.

Running the model to produce predictions entails carrying out the recursions in Eq. (4.3), at each point computing the predictors $\Psi_n = \Psi(x_n)$ with $x_n = y_{n-1} + \xi_n = z_s^{n-1} + \xi_n$. Though derived from Eq. (3.14), Eq. (4.3) is quite different in form. Here we examine Eq. (4.3) more closely, to clarify the flow of information in the algorithm and other details.

It is useful to first visualize Eq. (4.3) as a computation graph, a fragment of which is shown here:



(For legibility, we have drawn the edges going into z_s^n as solid lines; all others are dotted.) The variable z_s^n at time n and stage s depends on the corresponding variable z_{s-1}^n in the previous stage, as well as the two previous steps (z_s^{n-1} and z_s^{n-2}) in the same stage.

Once we have initial conditions, Eq. (4.3) can be iterated to generate sample paths. The first thing is then to find the initial values z_i^{p-1} and z_i^{p-2} for $i = 1, \cdots, r$ from the given data $\widetilde{x}_1, \cdots, \widetilde{x}_N$. An effective procedure is suggested by the computation graph: we set

$$\widetilde{y}_0 = \widetilde{x}_1\,, \ \ \widetilde{y}_1 = \widetilde{x}_2\,, \ \ \cdots\,, \ \ \widetilde{y}_{p-1} = \widetilde{x}_p \tag{4.7}$$

in the notation of Eq. (3.14) and Eq. (4.3). Assuming the coefficients α_i and β_i have already been determined, the computation graph shows that knowing the values at stage s for $n = 1, 2, \cdots, p$ (which is the same as knowing $y_1, \cdots, y_p$ in Eq. (3.14)) allows one to solve for the values at stage $s-1$ for $n = 3, 4, \cdots, p$. Iterating, this means we can determine z_i^{p-1}, z_i^p for all stages i. From this, it is also straightforward to see that if $y_0 = \cdots = y_{p-1} = 0$, then $z_i^{p-1} = z_i^{p-2} = 0$ for $i = 1, \cdots, r$, so that the initial conditions for Eq. (4.3) are uniquely determined by those of Eq. (3.14).

Once the initial data have been determined and noise generated (as described in Sect. 3), the recurrence relations (4.3) can be iterated to generate predictions from the reduced model.

4.3. Fitting models to data

We now describe our overall optimization strategy:

(i) From the time series $\widetilde{x}_1, \cdots, \widetilde{x}_N$, compute the observations $\widetilde{\Psi}_n = \Psi(\widetilde{x}_n)$.

(ii) For given parameter vectors α, β, b, use the initial values $\widetilde{x}_1, \cdots, \widetilde{x}_p$ to determine the initial values z_i^{p-2}, z_i^{p-1}, $i = 1, \cdots, r$, for Eq. (4.3).

(iii) Generate one-step predictions $\widehat{x}_{n+1}$ by Eq. (4.1) for $n = p, \cdots, N$, where $H(z) = B(z)/A(z)$.

In the cascade representation, the MSE has the form

$$\mathcal{E}'(\alpha,\beta,b) = \frac{1}{N}\sum_{n=p+1}^{N}\left\|\widetilde{x}_{n+1} - \widehat{x}_{n+1}\big(\widetilde{\Psi}_1,\cdots,\widetilde{\Psi}_n;\alpha,\beta,b\big)\right\|^2 \tag{4.8a}$$

with the decaying memory constraints

$$\beta_i \le 1 \qquad \text{and} \qquad \beta_i \ge \pm\alpha_i - 1\,. \tag{4.8b}$$

(This says that (α_i,β_i) lies within a triangle in the α-β plane with vertices $(\pm 2, 1), (0, -1)$. As asserted in Sect. 4.1, one can check that this is equivalent to the roots of $z^2 + \alpha_i z + \beta_i$ lying in the unit disc.) This can be minimized by direct optimization. One then finds the residuals

$$\widetilde{\xi}_n = \widetilde{x}_{n+1} - \widehat{x}_{n+1}\big(\widetilde{\Psi}_1,\cdots,\widetilde{\Psi}_n;\alpha,\beta,b\big) \tag{4.9}$$

and fit a noise model as before.

One can actually further reduce the dimensionality of the optimization problem; this is described below. But first, we note that Step (iii) above is more efficiently implemented by iterating

$$\begin{array}{lrcl} \text{Stage 1} & z_1^n + \alpha_1 z_1^{n-1} + \beta_1 z_1^{n-2} & = & \widetilde{\Psi}_{n-p+r}\cdot b_r + \cdots + \widetilde{\Psi}_{n-p}\cdot b_0 \\ \text{Stage 2} & z_2^n + \alpha_2 z_2^{n-1} + \beta_2 z_2^{n-2} & = & z_1^n \\ \vdots & & \vdots & \\ \text{Stage } s & z_s^n + \alpha_s z_s^{n-1} + \beta_s z_s^{n-2} & = & z_{s-1}^n \\ \text{Output} & \widehat{x}_{n+1} & = & z_s^n \end{array} \tag{4.10}$$

Modulo transients (see "initial conditions" below), this computes the convolutions in Eq. (3.17). Note this iteration can only be carried out if α, β satisfy the decaying memory condition.

To further reduce the dimensionality of the nonlinear optimization problem, we observe that for given (α,β), the function $b \mapsto \mathcal{E}'(\alpha,\beta,b)$ can be minimized by linear regression. For a given (α,β), we thus define $\widehat{b}(\alpha,\beta)$ to be the (unique) minimizer of $b \mapsto \mathcal{E}'(\alpha,\beta,b)$. We then minimize $\mathcal{E}'(\alpha,\beta,\widehat{b}(\alpha,\beta))$ by nonlinear optimization. For the examples in this paper, this is done using the BOBYQA algorithm[50] as implemented in the NLopt package [49].

Initial conditions. The recursions in Eq. (4.10), viewed as a system of non-autonomous linear recurrences with $\widetilde{\Psi}_n$ as time-dependent forcing, have their own initial conditions. Neglecting these "internal" initial conditions during fitting, for example by setting them all to zero, can lead to worse fits. Without accounting for initial conditions, the residuals also exhibit longer transients before approaching stationarity, which can complicate the construction of noise models.

To estimate initial conditions for Eq. (4.10), we exploit the linearity of Eq. (4.10) in the variables z_i^n by decomposing the z_i^n into the sum of a homogeneous solution $z_i^{h,n}$ and a particular solution $z_i^{p,n}$, with $z_i^{p,n}$ satisfying Eq. (4.10) with zero initial conditions and $z_i^{h,n}$ solving Eq. (4.10) with $\widetilde{\Psi}_n \equiv 0$. (In linear systems theory, these are the "zero state response" and "zero input response," respectively.) This leads to a linear regression problem for the initial values $\{z_i^{h,0}, z_i^{h,1} \mid i = 1, \cdots, s\}$, which can be solved jointly with the computation of $\widehat{b}(\alpha,\beta)$ via linear regression.

Remarks on optimization and related issues.

- *Cascade-form models, decaying memory, and optimization.* Eq (4.3) enables us to impose the decaying memory constraint reliably during optimization. However, the decomposition of $A(z)$ into quadratic factors introduces a symmetry: the value of the loss function is invariant when the quadratic factors are permuted. This means there are many equivalent global minima, which introduce many saddles into the landscape. While any of the symmetric global minima will give equivalent reduced models, the presence of the saddles can potentially slow down optimizers.
- *Other optimization strategies.* For simplicity, we have opted for direct nonlinear minimization of $\mathcal{E}'(\alpha,\beta,\widehat{b}(\alpha,\beta))$ in this paper. It may be possible to improve the efficiency of the optimization by exploiting the structure of Eq. (4.3) or the multistep representation (Eq. (3.18) above) by using, e.g., iterative least squares.
- *An implementation detail.* For interested readers, we describe the computation of $\widehat{b}(\alpha,\beta)$ by linear regression. We run the matrix version of the recursion

$$\begin{array}{lrcl} \text{Stage 1} & Z_1^n + \alpha_1 Z_1^{n-1} + \beta_1 Z_1^{n-2} & = & \widetilde{\Psi}_n \\ \text{Stage } i > 1 & Z_i^n + \alpha_2 Z_i^{n-1} + \beta_2 Z_i^{n-2} & = & Z_{i-1}^n, \end{array} \tag{4.11}$$

for $i = 2, \cdots, r$ and setting $Y_n = Z_s^n$; this is a matrix version of Eq. (4.10) with $q = 0$ and $b_0 = I$. The resulting Y_n and Z_i^n are matrix-valued, with the same shape as $\widetilde{\Psi}_n$. By exploiting the commutativity of the convolution operators defined by $B(z)$ and $1/A(z)$, one can show that the desired one step prediction is given by

$$\widehat{x}_{n+1} = Y_{n-p+q} \cdot b_q + \cdots + Y_{n-p} \cdot b_0 \ . \tag{4.12}$$

Combining this with the definition of $\mathcal{E}'(\alpha, \beta, b)$ lets us compute $\widehat{b}(\alpha, \beta)$ via linear regression.

4.4. Noise model

To construct a stochastic process η_n to model the residuals ξ_n, there are a few standard options:

(i) moving average representation, i.e., $\eta_n = d_q w_n + \cdots + d_0 w_{n-q}$ with independent $w_i \sim N(0, I)$;
(ii) estimating the power spectrum of ξ_n and generating a stationary Gaussian process matching the power spectrum;
(iii) constructing a linear SDE and fitting it to ξ_n by, e.g., maximum likelihood.

In earlier work, we have used a moving average representation together with a MLE to infer the coefficients a and b simultaneously with the coefficients of the moving average. In this paper, because we want to compare nonlinear regression with other approaches, the power spectrum method was found to be simpler.

After finding optimal values for a_i, b_i, and the initial y_i, we fit a stationary Gaussian process η_n to the residuals $\widetilde{\xi}_{n+1} = \widehat{x}_{n+1} - \widetilde{x}_{n+1}$, by a random Fourier series approximating a Wiener integral:

$$\eta_n = \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{M-1} f(j\Delta\theta)\, e^{-inj\Delta\theta}\ w_j \sqrt{\Delta\theta} \qquad \xrightarrow[M\to\infty]{\mathcal{D}} \qquad \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} f(\theta)\, e^{-in\theta}\ \dot{W}_\theta\, d\theta, \tag{4.13}$$

where $\Delta\theta = 2\pi/M$, the w_j are independent standard normal random variables (in the complex case, $Re(w_j)$ and $Im(w_j)$ are independent with variance 1/2), $\dot{W}_\theta$ is white noise on the circle S^1, and $f(\theta)$ is a square root of the spectral power density, i.e., $S_{\xi\xi}(\theta) = f(\theta) f(\theta)^*$. When η_n takes on values in $\mathbb{R}^d$, then $S_{\xi\xi}$ and f are $d \times d$ matrices and $\dot{W}_\theta$ is d-dimensional. The power spectrum can be estimated from data by the periodogram method (see, e.g., [51] and references therein). More efficient and accurate sampling methods are available [52], but we have found the random Fourier series above to be sufficient the residuals (ξ_n) are relatively small, as occurs in many examples (including ours). Whatever the method, the resulting reduced models will only satisfy the orthogonality conditions approximately.

5. Examples

We now consider two concrete examples. In addition to illustrating the methods described in earlier sections, there are two specific questions we would like to address:

- How effective is the model reduction method based on nonlinear regression (as described in Sect. 3.2)?
- How does the nonlinear regression compare to the linear regression described in Sect. 3.2?

We would also like to see how the least squares based nonlinear regression compare to the MLE used in [53].

5.1. Kuramoto-Sivashinsky (KS) PDE

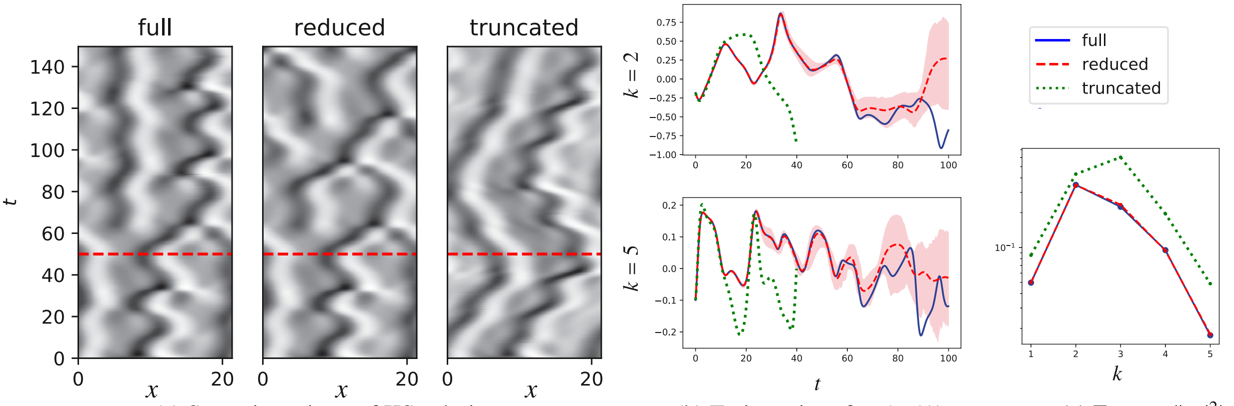
The KS equation

$$U_t + UU_x + U_{xx} + U_{xxxx} = 0 \tag{5.1}$$

is a prototypical model of spatiotemporal chaos. Here, we consider Eq. (5.1) with $0 \le x \le L$ and periodic boundary conditions. In Fourier variables $u_k(t)$, Eq. (5.1) is

$$\dot{u}_k = -\frac{i\lambda_k}{2} \sum_{\ell} u_\ell u_{k-\ell} + (\lambda_k^2 - \lambda_k^4) u_k \ , \quad \lambda_k = \frac{2\pi k}{L}. \tag{5.2}$$

The lowest $\approx L/2\pi$ modes are linearly unstable. This long-wave instability and its interaction with the quadratic nonlinearity lead to sustained chaotic behavior, with positive Lyapunov exponents and exponential decay of correlations [54]. NARMAX modeling of Eq. (5.1) was studied in [53], using likelihood-based parameter estimation and a slightly different form of NARMAX. Here, we use the least squares procedure. Following [53], we set $L \approx 21.55$, leading to 3 linearly unstable modes and a maximum Lyapunov exponent of ≈ 0.04 (Lyapunov time ≈ 25). In this regime, time correlation functions exhibit



(a) Spacetime views of KS solutions (b) Trajectories of $Re(u_k(t))$ (c) Energy $\langle |u_k|^2 \rangle$

Fig. 1. KS solutions. Panel (a) shows results computed using the 108-mode truncation ($\Delta t = 10^{-3}$) *(left)*, the 5-mode reduced model ($\Delta t = 0.1$) *(middle)*, and the 5-mode truncation ($\Delta t = 10^{-3}$) *(right)*. In (b), we plot two Fourier modes as functions of time, with 90% confidence intervals for the reduced model. Panel (c) shows the energy spectrum. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

complex oscillations instead of the simple exponential decay often seen in strongly chaotic systems (Fig. 2(a)), providing a nontrivial testbed for model reduction.

Eq. (5.1) is readily solved by truncating the Fourier series, provided the cutoff is large enough. Here, we take as full model the 108-mode truncation; numerical tests show that KS statistics are insensitive to the cutoff beyond this. Fig. 1(a) shows a sample solution of Eq. (5.1) using this 108-mode truncation ("full"). By comparison, the 5-mode truncation with the same initial conditions ("truncated") diverges rapidly, and fails to reproduce the energy spectrum (Fig. 1(c)).

Reduced model. To construct a reduced model using the lowest $K = 5$ Fourier modes, we follow the procedure outlined in Sect. 3. The first step is to generate data from the full model, which we do by numerically integrating the 108-mode truncation using a 4th-order exponential time-differencing Runge-Kutta (ETDRK4) method [55,56] with timestep $\vartriangle t = 10^{-3}$, for 10^8 steps. We observe the first $K = 5$ Fourier modes at every 100 steps; the observation interval $\delta = 0.1$ is the timestep for the reduced model. We drop the first half of the data to ensure stationarity.

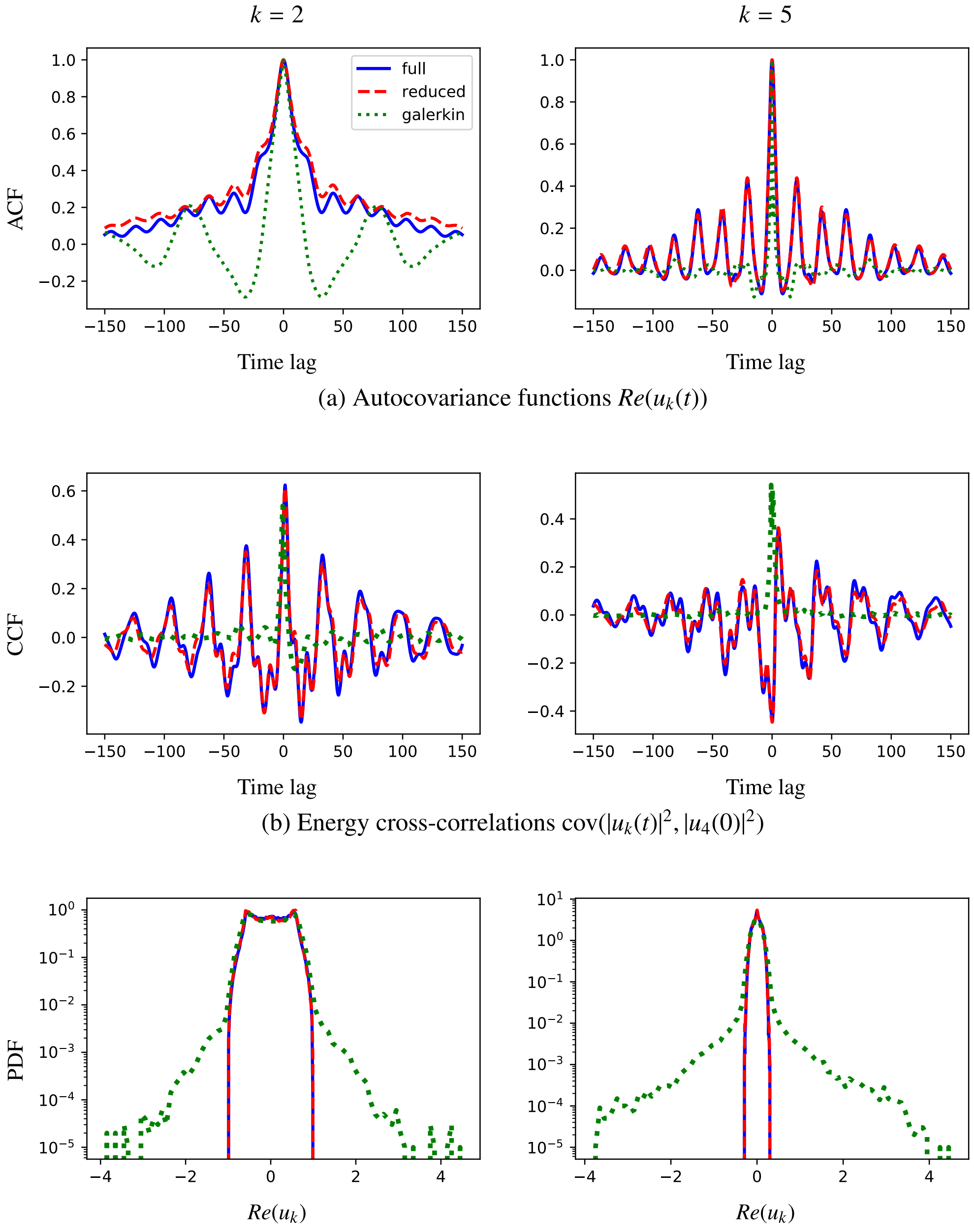
We use the form of the reduced model in Eq. (3.14) with x_n corresponding to $u^n = (u_1^n, u_2^n, \ldots, u_K^n)$; see Appendix C.1 for a detailed description of the model. To select the orders p and r, we tried a variety of small values until a combination is found that produces a stable reduced model. For the function $\Psi(u)$, we use three groups of functions:

$$\begin{aligned} \Psi^a_{n-j} &= u^{n-j} \,, \\ \Psi^b_{n-j} &= R^{\Delta t}(u^{n-j}) \,, \\ \Psi^c_{n-j,k} &= \sum_{\substack{|k-l|\le K,\, K<|l|\le 2K \\ \text{or } |l|\le K,\, K<|k-l|\le 2K}} \widetilde{u}_l^{n-1}\overline{\widetilde{u}_{k-l}^{n-j}} \text{ for } k = 1, \cdots, K. \end{aligned} \tag{5.3}$$

Here the first two groups Ψ^a and Ψ^b come from the Galerkin truncation. The third group in form of Ψ^c represents that interaction between the unresolved high modes and the resolved low modes, in which the high modes $\widetilde{u}$, defined in Appendix C.1d, is motivated by the theory of approximate inertial manifolds. In terms of the formalism of Sect. 3.1, the observation function $\Psi(u)$ is a $K \times (2K + K^2)$ matrix whose entries consist of the terms given above, where K is the number of relevant Fourier modes. Here, we use $K = 5$.

Finally, the reduced model is fit to data by the procedure outlined in Sect. 4. As was found in [53], not all combinations of p and r lead to stable reduced models. Indeed, we have experimented with "replaying" the residuals, i.e., compute the residuals $\widetilde{\xi}_n$ as in Sect. 4, then running the reduced model with $\widetilde{\xi}_{n+1}$ in place of the noise term. In the absence of round-off, one would simply obtain $x_n = \widetilde{x}_n$, i.e., reconstruct the original time series. Instead, for some choices of (p, r), round-off errors were rapidly amplified. Here, we use the pair $p = r = 3$, which is found to strike a balance between accuracy and efficiency. As measured by the product of the mode and step counts, the reduced model represents an over 100-fold reduction in computational cost.

Results. Fig. 1(a) compares the full model ("full"), the reduced model with $p = r = 3$ ("reduced"), and the 5-mode truncation with $\vartriangle t = 10^{-3}$ ("truncated"). As one can see, the reduced model reproduces the full solution up to $t \gtrsim 50$, about 1.8× the Lyapunov time, consistent with [53]. In contrast, the 5-mode truncation is accurate for a fraction of that time. Fig. 1(b) takes a closer look at selected Fourier modes. For the reduced model, 100 independent realizations are run, and the resulting ensemble is used to estimate confidence intervals. Shown is the mean (dashed, red), and 90% confidence intervals. Though the noise terms have amplitudes $\le 10^{-4}$, they are rapidly amplified by exponential separation of trajectories due to the



(a) Autocovariance functions $Re(u_k(t))$

(b) Energy cross-correlations $\mathrm{cov}(|u_k(t)|^2, |u_4(0)|^2)$

(c) Marginal distributions

Fig. 2. KS statistics. In all panels, solid blue is the full model, dashed red is the reduced model, and dotted green the 5-mode truncation. Panel (a) shows autocovariance functions for two Fourier modes $Re(u_k(t))$. In (b), we show cross correlation functions for the energies $|u_k(t)|^2$ and $|u_4(0)|^2$ for $k = 2, 5$. In (c), distributions of $Re(u_k)$ are shown.

long-wave instability in KS. Consistent with Fig. 1(a), the mean follows the true trajectory up to $t \approx 40$, at which point they begin to diverge. In contrast, the 5-mode truncation diverges by $t \approx 20$. Moreover, even when the confidence interval starts to widen, it continues to provide useful bounds for some time. Eventually the ensemble approaches statistical steady state, and the ensemble mean converges toward its expected value. Fig. 1(c) compares the energy spectra $\langle |u_k|^2 \rangle$: while the reduced model correctly predicts the spectrum, the 5-mode truncation produces fluctuations that are too large.

We note that while the noise terms are small in amplitude (see Fig. 4), we could not have constructed the confidence intervals in Fig. 1 without them. Moreover, we conducted numerical experiments without the noise terms. The results (data not shown) show that the reduced models do considerably worse at all tasks, and at least for some choices of (p, r) the solutions converge quickly to 0.

In Fig. 2, we examine long-time statistics. In (a), we compare the autocovariance functions of selected Fourier modes. Unlike the 5-mode truncation, the reduced model is able to reproduce quite complex features in the ACFs. Fig. 2(b) shows cross correlation functions for the energy of the kth mode with the energy of the 4th mode, i.e., $\mathrm{cov}(|u_k^n|^2, |u_4^0|^2)$ as a function of the time lag $n\Delta t$; such cross correlation functions can be viewed as a measure of energy transfer between modes. The reduced model correctly predicts these 4th moments, showing that the reduced model captures genuinely nonlinear effects in KS dynamics. Panel (c) shows the reduced model is able to reproduce marginal distributions, whereas the 5-mode truncation produces marginals that are too wide (compare with Fig. 1(c)). We conclude that both in terms of short-time forecasting and long-time statistics, the reduced model effectively captures KS dynamics. These findings are consistent with [53], suggesting the likelihood-based estimator used in [8,53] and the least squares estimator above are comparable, and the NARMAX model in [53] nearly optimal in the least squares sense. Numerical tests show that slightly different models (with different time lags p and r) may have similar statistical properties (such as consistency) and comparable performance in prediction. This suggests that there may be multiple reduced models fitting the data.

Linear vs. nonlinear regression. Sect. 3.2 emphasizes that choice of loss function should be viewed as part of the model reduction procedure. In particular, for our *ansatz*, the MZ formalism suggests a least squares approach leading to nonlinear regression (the nonlinearity arising from the way we parametrize the transfer function $H(z)$ by a rational approximation). An alternative is to infer the coefficients by linear regression, by minimizing the one step predictions in Eq. (3.19). Though the resulting reduced model Eq. (3.18) is formally equivalent to Eq. (3.14), the coefficients and the statistics of the residuals are different. We emphasize that both models are nonlinear and share the same functional form, and differ only in how model coefficients are inferred. For both models, the residuals are computed via Eq. (4.3) and a stationary Gaussian process fitted using the procedure outlined in Sect. 4.4.

Overall, using linear regression, we found far fewer combinations of (p, r) for which the reduced models is stable. Unfortunately, for the range of relatively low order models we tested ($0 \le p, r \le 3$), we did not find any combinations of p and r for which both procedures produced stable models. Thus, we did not conduct a direct comparison between the two. The closest pair of parameters we found were $p = r = 1$ using nonlinear regression, and $p = 1, r = 0$ using linear regression. This means our nonlinear regression example uses a reduced model of the form $x_{n+2} + a_0 x_{n+1} = \Psi(x_{n+1}) \cdot b_1 + \Psi(x_n) \cdot b_0 + \overline{\eta}_n$, while our linear regression model has the form $x_{n+2} + a_0 x_{n+1} = \Psi(x_n) \cdot b_0 + \overline{\eta}_n$.

Fig. 3 shows the results. Though both models are fairly low order, the nonlinear regression model has performance comparable to the higher-order ($p = r = 3$) model discussed above. In contrast, the linear regression model has significantly worse forecasting performance, and was unable to reproduce the auto-correlation or cross correlation functions accurately. However, it does reproduce marginal distributions and energy spectra (not shown) reasonably well.

To compare the statistical properties of the reduced models produced by linear and nonlinear regression, Fig. 4 shows the power spectra $S_{xx}(\theta)$ and $S_{\xi\xi}(\theta)$ for the relevant variables x_n and the residuals ξ_n, for the linear regression model and the two nonlinear regression model, for the $k = 3$ Fourier mode. (The other modes show similar trends.) As far as these power spectra are concerned, the two nonlinear regression models have nearly identical behavior. For nonlinear regression, the residuals (ξ_n) have broader and flatter power spectra than that of (x_n), indicating that the effect of the approximate Wiener projection here is to capture the relatively slower dynamics. The residual is, however, far from white, suggesting the need for more refined noise models than white noise forcing.

In contrast, linear regression produces much larger residuals, with a flat but less broad power spectrum. It appears that linear regression could not fit the data nearly as well, but the addition of a suitable noise model was able to correct for some of the defects of the reduced model, e.g., marginal distributions and energy spectra. Temporal statistics appear to be more delicate, however, and the linear regression model did not faithfully capture the details of autocovariance functions.

Overall, these results suggest that for the KS equation, linear regression results in considerably worse performance than nonlinear regression. This is consistent with our expectation (see Sect. 3.2) that linear regression may have worse performance because it neglects longer-range correlations in the data. For "static" quantities like energy spectra and marginal distributions, it appears that the noise model was able to compensate for this, but unable to generate correct temporal statistics.

5.2. Stochastically-forced Burgers equation

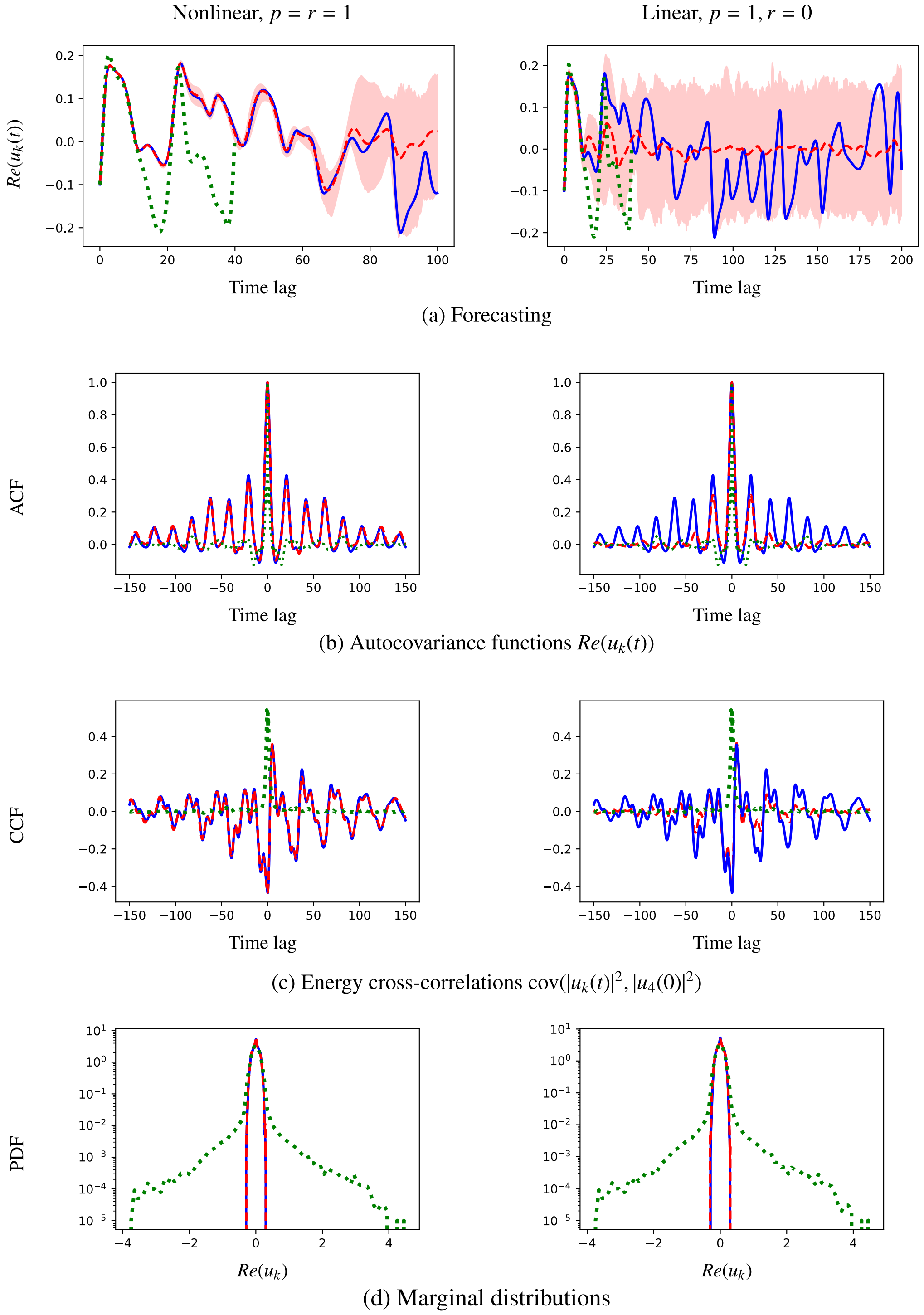
Now consider a stochastically-forced viscous Burgers equation

$$U_t + UU_x = \nu U_{xx} + \zeta \tag{5.4}$$

with $\zeta(t, x)$ white in t and smooth in x, and $U(t, x)$ 2π-periodic in x. More precisely, in Fourier variables,

$$\dot{u}_k = -\frac{i\lambda_k}{2} \sum_{\ell} u_\ell u_{k-\ell} - \nu\lambda_k^2 u_k + \sigma_k \dot{w}_k, \tag{5.5}$$

where $\sigma_k = 1$ for $|k| \le 4$ and $\sigma_k = 0$ for $|k| > 4$, $\dot{w}_k$ is white noise, and $\lambda_k = k$. In contrast to the KS equation, which is deterministic and exhibits self-sustained chaos, the viscous Burgers equation is dissipative: without forcing, all solutions converge to the steady state $u \equiv 0$ as $t \to \infty$. Stationary statistics of $u(x, t)$ thus reflect a balance between the forcing ζ and



(a) Forecasting

(b) Autocovariance functions $Re(u_k(t))$

(c) Energy cross-correlations $\mathrm{cov}(|u_k(t)|^2, |u_4(0)|^2)$

(d) Marginal distributions

Fig. 3. Linear vs nonlinear regression. Left: results from nonlinear regression with $p = r = 1$. Right: results from linear regression with $p = 1, r = 0$. Note as explained in the text, we did not find any orders (p, r) for which both procedures produced useful models.

dissipation through viscosity. We note that the stochastic Burgers equation has the so-called "one force one solution" (1F1S) property [57]: for a given realization of $\zeta_t, t \geq 0$, all initial conditions lead asymptotically to the same (time dependent) solution. Put another way, modulo transients, solutions of Eq. (5.5) are determined by the forcing.

In view of the 1F1S property, a natural question is: given a specific realization of the forcing ζ_t, can a reduced model correctly predict the response of the system? To test this, we compare a fully-resolved, 128-mode truncation of Eq. (5.5)

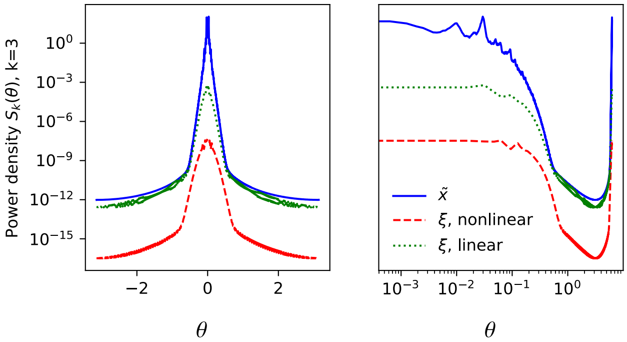


Fig. 4. KS power spectra. The left panel shows the spectral power density $S_{xx}(\theta)$ for $x = u_3$, the $k = 3$ Fourier mode of the KS equation. The right panel shows the same power spectrum on a log-log scale to better exhibit the structure near $\theta = 0$. The solid blue curve is the power spectrum of the Fourier mode u_3 from the full model, the dashed red curve is the power spectrum of the residuals ξ resulting from nonlinear regression with $p = r = 1$; the dotted green curve is the power spectrum of the residual resulting from linear regression with $p = 1, r = 0$. Modes with $k \neq 3$ behave similarly and are not shown.

with an under-resolved 9-mode truncation and a 9-mode reduced model inferred from data. Throughout, $\nu = 0.05$. (See [58] for an alternate view of this problem.)

Data-driven reduced model. To generate data from the full model, we solve Eq. (5.5) using a scheme of the form

$$u_k^{n+1} = G_k(u^n, \Delta t) + \sqrt{\Delta t}\,\sigma_k\, w_k^n, \tag{5.6}$$

where $G_k(u, \Delta t)$ is the result of applying ETDRK4 to the deterministic part of Eq. (5.5), $u_k^n = u_k(n\Delta t)$, $u^n = (u_1^n, \cdots, u_K^n)$, and w_k^n independent $N(0,1)$ random variables. Like the standard Euler-Maruyama scheme, Eq. (5.6) has weak order 1, but is more stable [59]. We solve the full system with timestep $\Delta t = 0.00125$ and observe every 8th step, so the reduced model has timestep $\delta = 0.01$. Except for minor differences, this has the form of Eq. (3.22).

To account for the forcing, we modify Eq. (3.14) to obtain

$$x_{n+1} = y_n + \xi_{n+1}, \tag{5.7a}$$
$$y_n + a_{p-1}y_{n-1} + \cdots + a_0 y_{n-p} \tag{5.7b}$$
$$= \Psi_{n-p+r} \cdot b_r + \cdots + \Psi_{n-p} \cdot b_0 + \tag{5.7c}$$
$$c_q \overline{w}_{n+q} + \cdots + c_0 \overline{w}_n. \tag{5.7d}$$

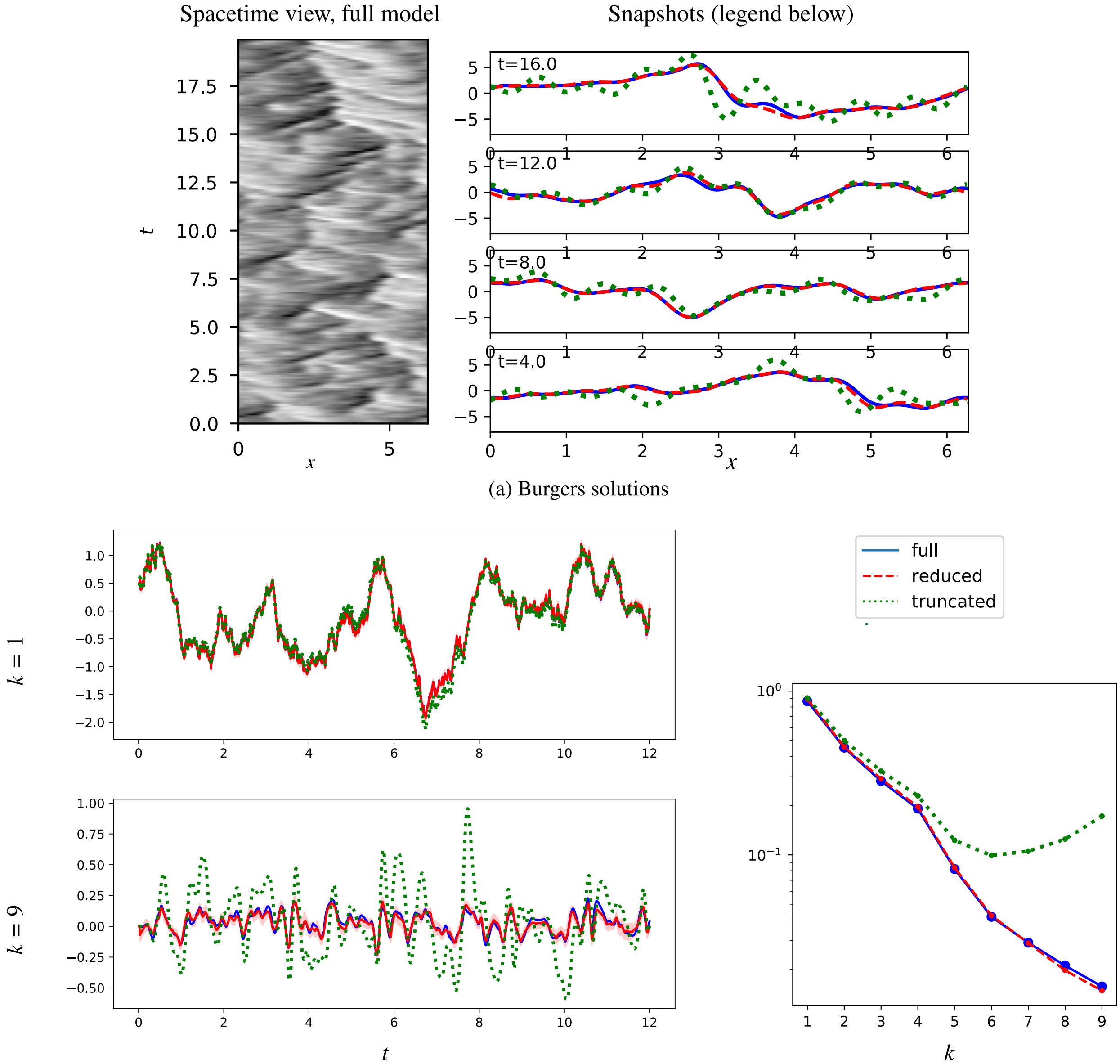
The $\overline{w}_n$ in the moving average (5.7d) are related to the forcing w^n in Eq. (5.6) by $\overline{w}^n = (w^{8n} + \cdots + w^{8n+7})/\sqrt{8}$; this correlates the full model and the reduced model during fitting. The independent noise term ξ_n is inferred from the residuals as before, and permits one to quantify the uncertainty in response prediction via ensemble forecasting. As noted in Sect. 3.3, random dynamical systems like Eq. (5.5) are encompassed within MZ theory, and Eq. (5.7) can be seen as a special case of the Wiener projection. As before, the orders p and r are selected by trial-and-error.

We have also constructed reduced models of the form (3.14), which do not correlate the reduced and full models through shared forcing. All else being equal, we found the performance of Eq. (5.7) to be strictly better in our tests than Eq. (3.14) because more information is retained. We report results obtained using Eq. (5.7) with $p = r = 1$, leading to a ~ 50-fold reduction in cost.

The exact form of the predictors $\Psi(\cdot)$ are given in Appendix D. Interested readers are referred to [60] for further investigation of this and other parametric forms, consistency of estimators, and model selection.

Results. Fig. 5(a) shows sample solutions. The 1F1S property suggests that the low modes in the full, reduced, and the 9-mode truncation models will all be strongly correlated, as confirmed in the snapshots. However, one also sees that the 9-mode truncation exhibits significant deviations from the full model, unlike the reduced model. Panels (b) and (c) shows this behavior in more details: because of the forcing, the low modes of all 3 models stay close over time, but the 9-mode truncation shows relatively large deviations from the full model. As before, Fig. 5(b) shows 90% confidence intervals for the reduced model, computed using an ensemble of 100 trajectories. As expected, the forced modes are tightly entrained to each other, whereas the 9-mode truncation shows significant deviation in higher modes. Because of the 1F1S property, the reduced model can be expected to correctly forecast the response for as long as information about the forcing is available. As for the KS equation, the reduced model here also reproduces long-time statistics; see Fig. 5(c) for the energy spectrum and Appendix D for other statistics.

Finally, we note that while accurate response forecasting will clearly become more difficult for larger observation intervals, the reduced model can nevertheless capture long-time statistics for much larger observation times. Indeed, we have tested the reduced model for much larger observation intervals, up to 0.1 (see Appendix D).



(a) Burgers solutions

(b) Trajectories of $Re(u_k(t))$

(c) Energy $\langle |u_k|^2 \rangle$

Fig. 5. Stochastic Burgers solutions. Panel (a) shows results computed using the 128-mode truncation with $\Delta t = 0.00125$ *(left)*, and snapshots of the full model, the 9-mode reduced model ($\Delta t = 0.01$), and the 9-mode truncation ($\Delta t = 0.00125$). In (b), we plot two Fourier modes as functions of time, with 90% confidence intervals for the reduced model. Panel (c) shows the energy spectrum.

Linear vs. nonlinear regression. In contrast to the KS equation, linear and nonlinear regression produced essentially identical results for the Burgers equation. In particular, our tests show that linear regression can produce marginal distributions and ACFs comparable to the nonlinear regression model, and has nearly identical forecasting skill; see Fig. D.14 in Appendix D.

Fig. 6 compares the spectral power densities $S_{xx}(\theta)$ and $S_{\xi\xi}(\theta)$ for the relevant variables x_n and the residuals ξ_n, for the $k = 3$ Fourier mode. (The other modes are similar.) Unlike the KS equation, here the linear and nonlinear regression give essentially identical power spectra of the noise. The residual spectrum is not broader than the spectrum of the Fourier mode itself, likely because the Fourier modes of the Burgers equation are subjected to white noise forcing and therefore contain much higher frequency content than their KS counterparts. As in the KS example, leaving out the noise terms entirely led to much worse results.

In view of the discussion in Sect. 3.2, the remarkable contrast between this and the KS equation may be due to the fact that the Burgers equation is being driven by white noise. The 1F1S property implies that the dynamics is largely determined by the forcing, and hence long-range temporal correlations play less of a role.

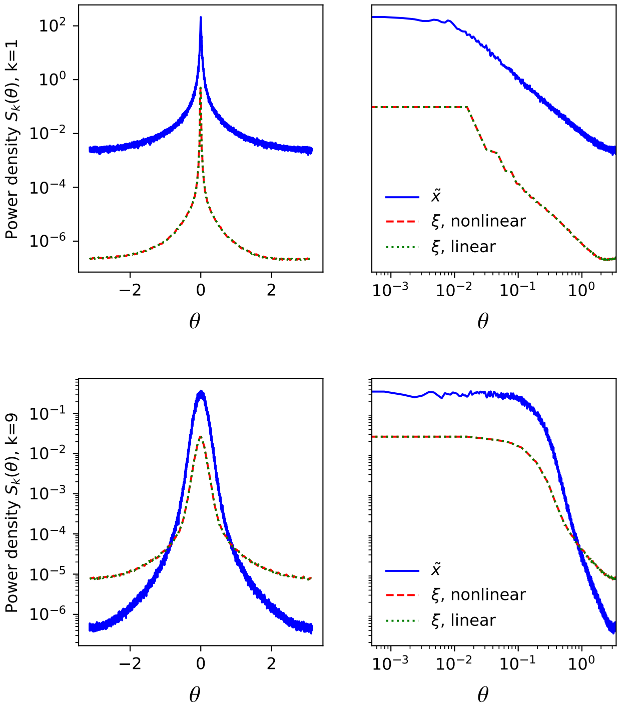


Fig. 6. Burgers power spectra. Left panels show the spectral power density $S_{xx}(\theta)$ for the kth Fourier mode of the Burgers equation. Right panels show the same power spectrum on a log-log scale. Solid blue curves are the power spectrum of the Fourier mode u_k from the full model, dashed red curves that of the residuals ξ from nonlinear regression; and dotted green curves the residual power spectrum from linear regression. The green and red curves essentially coincide.

6. Concluding discussion

Many issues surrounding this topic remain incompletely understood. We mention a few here:

- *Nonparametric modeling.* We have focused on parametric model reduction in this paper. But in principle the observation functions $\Psi(x)$ can be inferred from data using nonparametric methods like delay coordinates, manifold learning, dynamic model decomposition, reservoir computing, and other machine learning techniques [61–67]. Eq. (3.6) still applies in these situations, and the Wiener projection formulation complements and extends existing strategies for data-driven modeling and model reduction by providing a systematic guide to incorporating memory and noise effects, in situations without sharp scale separation. For example, one may infer Ψ by a combination of delayed coordinates and manifold learning, or artificial neural network techniques.
- *Other rational approximations of $H(z)$.* The simple rational approximation $H(z) = B(z)/A(z)$ is used here out of expedience. Experience has shown that other rational approximations, e.g., those based on continued fractions, can sometimes yield effective approximations with relatively few undetermined parameters [6]. These will be investigated in future work.
- *Structure-preserving reduced models.* Most physical systems of interest are characterized by exact or approximate conservation laws and symmetries, and it is important for reduced models to preserve these fundamental physical constraints. Structure-preserving model reduction is an active area of study, and the framework described in this paper may provide a new perspective on this problem.
- *Numerical stability.* In a data-driven approach, one often finds that the estimated reduced model is numerically unstable. Heuristically, this is because (i) reduced models often coarse-grain in both time and space, and the relatively large time steps impose more stringent stability requirements; (ii) most loss functions used in data-driven model reduction reflect the accuracy of the approximation, and one runs the risk of overfitting data. Indeed, our results have shown that the most accurate reduced models (i.e., those with the smallest residuals) are not always the best reduced model. A general understanding of numerical stability in these models is currently lacking. Because our nonlinear regression method always produces linearly stable models, understanding numerical stability will likely require tackling the strong nonlinearities inherent in these models.
- *Quantification of the accuracy of a reduced model.* Data-driven approaches have led to many model reduction strategies that can successfully reproduce key dynamical features such as energy spectrum and correlations. The development of

systematic approaches to quantify, analyze, and compare reduced models to full models remain incomplete. It is our hope that the formalism developed in this paper will provide a new perspective on this fundamental problem.
- *Noise modeling.* For both our examples, the residuals have small amplitude, and we have seen that additive noise models work relatively well. We do not know if this approach will continue to be effective when the residuals have large amplitude, as occurs in, e.g., molecular dynamics at finite temperatures.
- *Relationship to other data-driven modeling approaches?* In recent years, a variety of data-driven modeling and model reduction techniques have been proposed, applicable in different dynamical regimes. These include delay coordinate embedding [68,65,69], manifold learning and kernel regression [70,63], dynamic mode decomposition (DMD) [71,72,14], and many others. The MZ framework should not be viewed as an alternative to these methods. Rather, it is complementary in the sense that it provides a general scaffold into which different model reduction techniques can fit. For example, for problems with low-dimensional attractors in high-dimensional phase spaces, delayed coordinates and extensions like DMD are natural. But when the underlying assumptions (e.g., fast convergence to the low-dimensional attractor, deterministic dynamics) are only satisfied approximately, the MZ formalism may be useful for suggesting corrections.

To conclude, we have shown the Wiener projection provides a framework for data-driven modeling that is grounded in dynamical systems theory. As such, we view it as a step towards bridging the gap between nonlinear dynamics theory and data-driven model reduction. Within this framework, we give a heuristic derivation of a version of the NARMAX model widely used in time series modeling and analysis, providing an interpretation of NARMAX in terms of an underlying dynamical system and evidence that it may be nearly optimal in the sense of least squares. In addition to giving a dynamical basis for NARMAX, this framework may provide a starting point for systematic data-driven model reduction, Using the KS and stochastic Burgers equations, we have demonstrated the flexibility and effectiveness of this view of model reduction for deterministic chaotic and random dynamics.

CRediT authorship contribution statement

KL and FL contributed equally to this paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the Mathematics Group at Lawrence Berkeley National Lab for its support of this work; to Alexandre Chorin for many useful comments on the manuscript; and to Xiantao Li for encouraging us to study the discrete Mori-Zwanzig formalism. KL was supported by NSF grant DMS-1821286, and FL was supported by NSF grant DMS-1821211.

Appendix A. The dual equation and Mori-Zwanzig closure

In this section, we give an alternate derivation of the MZ equation (2.2) that makes use of a dual equation describing the evolution of conditional probability distributions. Though longer, it gives some additional insights into the meaning of the MZ equation.

As before, let F be a dynamical system with state space $\mathbb{X}$. Suppose an initial condition X_0 is drawn from the distribution ρ_0. Let ρ_n denote the distribution of X_n; then $\rho_{n+1} = L\rho_n$, where L is the *transfer operator*, defined by

$$\int \varphi \circ F \, d\rho = \int \varphi \, d(L\rho) \tag{A.1}$$

for suitable test functions φ. The above is equivalent to

$$\int (M\varphi) \, d\rho = \int \varphi \, d(L\rho), \tag{A.2}$$

i.e., the operator L is the adjoint of the Koopman operator M, where the adjoint of an operator T acting on functions is the operator $T^\dagger$ acting on distributions defined by $\int (T\varphi) \, d\rho = \int \varphi \, d(T^\dagger \rho)$. With this, and with P and Q as before, we have

$$\begin{aligned} Q^\dagger \rho_{n+1} &= Q^\dagger L \rho_n \\ &= Q^\dagger L (P^\dagger + Q^\dagger) \rho_n \\ &= Q^\dagger L Q^\dagger \rho_n + Q^\dagger L P^\dagger \rho_n, \end{aligned}$$

using $P^\dagger + Q^\dagger = I$. Solving the recurrence for $Q^\dagger\rho_n$ gives

$$Q^\dagger\rho_n = (Q^\dagger L)^n Q^\dagger\rho_0 + \sum_{k=0}^{n-1}(Q^\dagger L)^{n-k}P^\dagger\rho_k\,. \tag{A.3}$$

From this it follows that

$$\begin{aligned}\rho_{n+1} &= L\rho_n\\ &= LP^\dagger\rho_n + LQ^\dagger\rho_n\\ &= LP^\dagger\rho_n + L(Q^\dagger L)^n Q^\dagger\rho_0 + L\sum_{k=0}^{n-1}(Q^\dagger L)^{n-k}P^\dagger\rho_k\,;\end{aligned}$$

in the last line we just substituted Eq. (A.3).

The above is equivalent to the operator equation

$$L^{n+1} = LP^\dagger L^n + L(Q^\dagger L)^n Q^\dagger + L\sum_{k=0}^{n-1}(Q^\dagger L)^{n-k}P^\dagger L^k\,. \tag{A.4}$$

Taking adjoints, we get the Dyson formula

$$M^{n+1} = M^n PM + Q\,(MQ\,)^n M + \sum_{k=0}^{n-1} M^k P(MQ\,)^{n-k}M. \tag{A.5}$$

From this, the MZ equation follows as before.

Suppose now P is conditional expectation with respect to μ. Observe that for an observable φ and a probability distribution ρ, we have

$$\begin{aligned}\int P\varphi\, d\rho &= \iint\left[\int \varphi(x,y)\,\mu_{Y|X}(dy|x)\right]\rho(dx,dy')\\ &= \int\left[\int \varphi(x,y)\,\mu_{Y|X}(dy|x)\right]\rho_X(dx)\\ &= \int \varphi(x,y)\int\ \mu_{Y|X}(dy|x)\,\rho_X(dx)\\ &= \int \varphi\, d(P^\dagger\rho).\end{aligned}$$

So the dual $P^\dagger$ to the conditional expectation P is

$$(P^\dagger\rho)(dx,dy) = \rho_X(dx)\cdot\mu_{Y|X}(dy|x)\,. \tag{A.6}$$

That is, for a density ρ, $P^\dagger\rho$ is the product of the X-marginal of ρ and the conditional density $\mu_{Y|X}$. The operator $P^\dagger$ preserves the X-marginals of densities, and $P^\dagger\mu = \mu$. If one were to construct reduced models by keeping only the Markov term in the MZ equation, this corresponds to the closure assumption that the unresolved modes have statistics given by the stationary distribution μ conditioned on the resolved modes. This is the discrete-time analog of the averaging principles for ODEs (see, e.g., [73]).

Appendix B. Brief summary of z-transform and Wiener filters

For the convenience of readers, this Appendix provides a brief non-technical summary of some basic facts about z-transforms and Wiener filtering. See, e.g., [74,40,75,76] for more details.

z-transforms and linear filtering. We first consider (real or complex, scalar or vector) bi-infinite sequences $\cdots, x_{-1}, x_0, x_1, \cdots$ that are *causal*, i.e., $x_n = 0$ for $n < 0$. For simplicity, we assume $(x_n) \in \ell^1$ (though much of what we say below holds as long as the x_n decay sufficiently fast as $n \to \infty$). For a causal sequence, its *z-transform* is the formal series

$$X(z) = \sum_{n\geq 0} x_n z^{-n}. \tag{B.1}$$

In the above expression, z should be viewed as a complex variable, though the series typically does not converge for all $z \in \mathbb{C}$. The ℓ^1 assumption (which covers many examples in applications) means the domain of convergence of $X(z)$ includes the unit circle and $X(e^{-i\theta})$ is a Fourier series with x_n as coefficients. In this case, the z-transform is invertible by

$$x_n = \frac{1}{2\pi}\int_0^{2\pi} e^{-in\theta} X(e^{-i\theta})\, d\theta. \tag{B.2}$$

More generally, the z transform can be inverted by an appropriate application of the Cauchy integral formula.

The z-transform is the analog of the Laplace transform for difference equations. Two key properties include:

(i) *Shifts:* if $y_n = x_{n+1}$ for $n \geq 0$, then $Y(z) = z(X(z) - x_0)$.
(ii) *Convolution:* if $w_n = (x \star y)_n = \sum_{k\geq 0} x_k y_{n-k}$, then $W(z) = X(z) \cdot Y(z)$.

In signal processing and time series analysis, the z-transform is useful for representing the action of "linear filters." That is, suppose we have a signal (x_n). A linear filter is a linear transformation mapping (x_n) to (y_n), with

$$y_n = (x \star h)_n = \sum_{k\geq 0} x_k \cdot h_{n-k}. \tag{B.3}$$

The sequence (h_n), which defines the filter, is known as its *impulse response*, so called because h_n is the response of the filter when (x_n) is the unit impulse, i.e., $x_n = \delta_{n0}$, δ_{mn} being the Kronecker delta function. By the convolution property, we then have $Y(z) = H(z)X(z)$. $H(z)$ is the "transfer function" of the linear filter.

One of the ways in which the z-transform is useful is that the analytic properties of the transfer function encode the asymptotic behavior of the impulse response. For example, if the transfer function $H(z)$ of a scalar filter were meromorphic and all its poles lie strictly inside the unit disc, then Eq. (B.2) tells us h_n is causal and decays exponentially as $n \to \infty$. (If we only know that the restriction of H to the unit circle is continuous, then $h_n \to 0$ is implied by the Riemann-Lebesgue lemma.) In the reverse direction, if $(h_n) \in \ell^1$ (as we assume), then $H(z)$ cannot have any poles outside the unit disc.

An application to NARMAX. In Sect. 3.2, we asserted the equivalence of Eqs. (3.14) and (3.18) modulo transients. Here we show a derivation using z-transforms; an alternative is to use the substitution $y_n = x_{n+1} - \xi_{n+1}$ in Eq. (3.18). One of the advantages of the z-transform method is that it provides an operational calculus for keeping track of indices systematically.

First, take z-transforms of Eq. (3.14), we get

$$z(X(z) - x_0) = Y(z) + z(\Xi(z) - \xi_0) \tag{B.4a}$$

$$A(z)Y(z) + p_0(z) = \Psi(z) \cdot B(z) + q_0(z) \tag{B.4b}$$

where $p_0(z)$ and $q_0(z)$ are polynomials whose coefficients are functions of the initial conditions $x_0, \cdots, x_p$ and $\Psi(x_0), \cdots, \Psi(x_q)$, with $\deg(p) \leq \deg(A)$ and $\deg(q) \leq \deg(B)$, and $p_0 \equiv q_0 \equiv 0$ if the $x_0 = \cdots = x_p = \Psi_0 = \cdots = \Psi_q = 0$. Substituting Eq. (B.4b) into (B.4a) and simplifying, we get

$$zA(z)(X(z) - x_0) + p_0(z) = \Psi(z) \cdot B(z) + q_0(z) + zA(z)(\Xi(z) - \xi_0). \tag{B.5}$$

For comparison, if we transform Eq. (3.18), we get

$$zA(z)X(z) + p_1(z) = \Psi(z) \cdot B(z) + q_1(z) + zA(z)\Xi(z). \tag{B.6}$$

Comparing Eqs. (B.5) and (B.6), we see they are equivalent modulo terms involving initial conditions. If all the zeros of $A(z)$ lie inside the unit circle, then transients will decay as $n \to \infty$, so modulo transients Eqs. (B.5) and (B.6) are equivalent. In particular, the recursions are exactly equivalent for homogeneous initial conditions.

The above argument relies on the z-transform. Because the recursions are driven by the ξ_n, its validity hinges on what we assume about ξ_n: if the ξ_n were, e.g., white noise, then the z-transforms are not well-defined, but if the ξ_n decay sufficiently fast as $n \to \infty$, then the z-transforms are valid. Supposing now that there is a sequence ξ_n such that Eqs. (B.5) and (B.6) are not equivalent for homogeneous initial conditions $x_0 = \cdots = x_p = \Psi_0 = \cdots = \Psi_q = 0$. Then there is a least $N > 0$ for which they would disagree. But then if we set $\xi'_n = \xi_n$ for $n \leq N + p$ and $\xi'_n = 0$ for $n > N + p$, then (because the recursion has order p) the two recursions would differ when driven by ξ'_n.

Correlation functions and power spectra. The preceding discussion of the z-transform only makes sense if the sequences involved decay sufficiently fast as $n \to \infty$. In our context, we are interested in convolving such a sequence (h_n) with stationary stochastic processes. The formal series (B.1) does not make sense.

A standard approach is based on correlation functions. Suppose (X_k) and Y_k are zero-mean stationary stochastic processes taking values in $\mathbb{R}^d$. We define the (matrix-valued) correlation function to be

$$C_{xy}(k) = E\Big(X_k \cdot Y_0^*\Big) \tag{B.7}$$

where "$*$" denotes the conjugate transpose. The corresponding *power spectrum* is

$$S_{xy}(z) = \sum_k z^{-k} C_{xy}(k). \tag{B.8}$$

Note this generalizes the notion of spectrum introduced earlier, and we are abusing notation slightly. The spectrum introduced earlier is $S_{xy}(e^{-i\theta})$. We record some useful properties:

(i) $C_{xx}(0)$ is hermitian positive-semidefinite.
(ii) $C_{xy}(k)^* = C_{yx}(-k)$, in particular $C_{xx}(k)^* = C_{xx}(-k)$.
(iii) $S_{xy}(e^{-i\theta})^* = S_{yx}(e^{-i\theta})$.
(iv) $S_{xx}(e^{-i\theta})^* = S_{xx}(e^{-i\theta})$, i.e., the power spectrum is hermitian for all θ.
(v) If $Y = h \star X$, then

$$C_{yx}(n) = \sum_k h_{n-k} \cdot C_{xx}(k), \tag{B.9}$$

or $C_{yx} = h \star C_{xx}$.
(vi) Taking z-transforms yields

$$S_{yx}(z) = H(z) \cdot S_{xx}(z). \tag{B.10}$$

Note the above identities are valid for both scalar and matrix quantities.
(vii) Similarly,

$$C_{xy}(n) = \sum_k C_{xx}(n+k) \cdot h_k^*. \tag{B.11}$$

The z-transform is now

$$S_{xy}(z) = S_{xx}(z) \cdot H^*(1/z) \tag{B.12}$$

where H^* is the z-transform of the sequence h_n^*.
(viii) Putting these relations together yields $C_{yy} = h \star C_{xx} \star h^*$, or equivalently

$$S_{yy}(z) = H(z) \cdot S_{xx}(z) \cdot H^*(1/z). \tag{B.13}$$

On the unit circle, this simplifies to

$$S_{yy}(e^{-i\theta}) = H(e^{-i\theta}) \cdot S_{xx}(e^{-i\theta}) \cdot H^*(e^{i\theta}). \tag{B.14}$$

In the scalar case, this reduces to $S_{yy}(e^{-i\theta}) = |H(e^{-i\theta})|^2 S_{xx}(e^{-i\theta})$.

These properties also form the basis for the random Fourier representation of stationary stochastic processes in Eq. (4.13).

Wiener filtering. We now record some basic results of Wiener filter theory for interested readers. This material is not used directly in the paper.

The Wiener filter is the linear filter (h_n) that minimizes the MSE

$$\mathbb{E}\Big|X_n - \sum_{k\geq 0} \Psi_{n-k} \cdot h_{-k}\Big|^2. \tag{B.15}$$

Equivalently, if we write

$$X_n = \sum_k h_{n-k} \cdot \Psi_k + \xi_n \tag{B.16}$$

this amounts of choosing (h_n) to minimize the residuals $\mathbb{E}|\xi_n|^2$. One can show that the power spectrum satisfies

$$S_{\xi\xi} \;=\; \underbrace{S_{xx} - S_{x\psi} \cdot S_{\psi\psi}^{-1} \cdot S_{\psi x}}_{\text{(I)}} \;+\; \underbrace{(H \cdot S_{\psi\psi} - S_{x\psi}) \cdot S_{\psi\psi}^{-1} \cdot (S_{\psi\psi} \cdot H^* - S_{\psi x})}_{\text{(II)}} \tag{B.17}$$

where $S_{\cdot}(\cdot)$ denotes power spectra as before, and $H(z)$ is the z-transform of (h_n). Observe $S_{\xi\xi}(e^{-i\theta}) \geq 0$ for all H. If we set

$$H(e^{-i\theta}) = S_{x\psi}(e^{-i\theta}) \cdot S_{\psi\psi}^{-1}(e^{-i\theta}), \tag{B.18}$$

then (II) vanishes. This means (I) is ≥ 0. Since (II) is obviously ≥ 0 as well, we see $Tr(S_{\xi\xi})$ is minimized by Eq. (B.18).

Unfortunately, the linear filter (h_n) defined by Eq. (B.18) may not be *causal,* i.e., h_n may be nonzero for $n < 0$. Such a filter would use future values of Ψ_m with $m > n$ to predict X_n. How, then, do we find a causal filter, i.e., one with $h_n = 0$ for $n < 0$? Let us first look at the special case where $S_{\psi\psi}(z) \equiv I_{d\times d}$, i.e., (Ψ_n) is "white." Then the functional to be minimized is

$$Tr\left((H - S_{x\psi}) \cdot (H^* - S_{\psi x})\right). \tag{B.19}$$

By Plancherel's Theorem, the optimal *causal* solution is given by $H = [S_{x\psi}]_+$, where

$$[S]_+(e^{-i\theta}) = \frac{1}{2\pi} \sum_{n=0}^{\infty} \int_0^{2\pi} e^{in(\theta-\theta')} S(e^{-i\theta'})\, d\theta'. \tag{B.20}$$

Summing over $n \ge 0$ instead of $n \in \mathbb{Z}$ sets the impulse response $s_n = 0$ for $n < 0$, thus making it causal. The $[\cdot]_+$ operator transforms a given function to the time domain, zero out all entries for $n < 0$, then transform back to frequency domain.

Now, in general Ψ_n will not be white. But, since $S_{\psi\psi}(e^{-i\theta}) \ge 0$, there exist C such that $C(e^{-i\theta}) \cdot C^*(e^{i\theta}) = S_{\psi\psi}(e^{-i\theta})$. So if we take $W = C^{-1}$ (as a function on S^1), then $w \star \Psi$ will be white. A remarkable fact is that under very broad conditions, there is a function $W(z)$ such that all its poles *and* zeros lie inside the unit circle, and $W(e^{-i\theta}) = C(e^{-i\theta})$. Such a W defines a causal stable linear filter (w_n) such that $w \star \Psi$ has power spectrum

$$W(e^{-i\theta}) \cdot S_{\psi\psi}(e^{-i\theta}) \cdot W(e^{-i\theta})^* \equiv I_{d\times d}, \tag{B.21}$$

i.e., $w \star \Psi$ is white. (The filter (w_n) is known as a whitening filter.) Using the whitening filter, one can check that

$$H(z) = [S_{x\psi}(z) \cdot W^*(1/z)]_+ \cdot W(z) \tag{B.22}$$

is indeed the causal Wiener filter.

Appendix C. Kuramoto-Sivashinsky equation

Nonlinear terms in the NARMAX model

The Kuramoto-Sivashinsky example in Sect. 5 uses the reduced model from [53]. For the convenience of readers, the full *ansatz* is reproduced here:

$$u_k^{n+1} = u_k^n + R_k^{\Delta t}(u^n)\, \Delta t + z_k^n\, \Delta t, \tag{C.1a}$$

$$z_k^{n+1} = \Phi_k^n + \xi_k^{n+1}, \tag{C.1b}$$

$$\begin{aligned}\Phi_k^n = &\sum_{j=0}^{p} a_{k,j} z_k^{n-j} + \sum_{j=0}^{r} b_{k,j} u_k^{n-j} + c_{k,(K+1)} R_k^{\Delta t}(u^n) \\ &+ i \sum_{j=1}^{K} c_{k,j} \widetilde{u}_{j+K}^n \overline{\widetilde{u}_{j+K-k}^n} + \sum_{j=0}^{q} d_{k,j} \xi_k^{n-j},\end{aligned} \tag{C.1c}$$

where

$$\widetilde{u}_j^n = \begin{cases} u_j^n\,, & 1 \le j \le K \\ i \sum_{\ell=j-K}^{K} u_\ell^n u_{j-\ell}^n\,, & K < j \le 2K. \end{cases} \tag{C.1d}$$

The nonlinear terms in Eqs. (C.1c) and (C.1d) are suggested by inertial manifold theory. See [53] for details.

We compare the above *ansatz* to the model used in this study, of the form (3.14) with predictors in (5.3). It is straightforward to show that the *ansatz* in Eq. (C.1) is equivalent to a model of the form in Eq. (3.18):

$$u_{n+p'+1} + a_{p'-1} u_{n+p'} + \cdots + a_0 u_{n+1} = \Psi'_{n+q'} \cdot b_{q'} + \cdots + \Psi_n \cdot b_0 + \xi'_{n+1}, \tag{C.2}$$

for some choice of orders p', q', coefficients a_i, b_i, functions $\{\Psi'_n\}$ and noise ξ'_n. In addition to the different approaches estimating the parameters (a, b), the models are different in the following aspect:

(i) Here we model the noise by a Gaussian process using power spectrum from the residual $\widetilde{\xi}_n$, whereas Eq. (C.1) models the noise by a moving average process.

(ii) as suggested by the Wiener projection formalism, the model (3.14) in this study contains time-delayed copies of all nonlinear terms, whereas Eq. (C.1) does not.

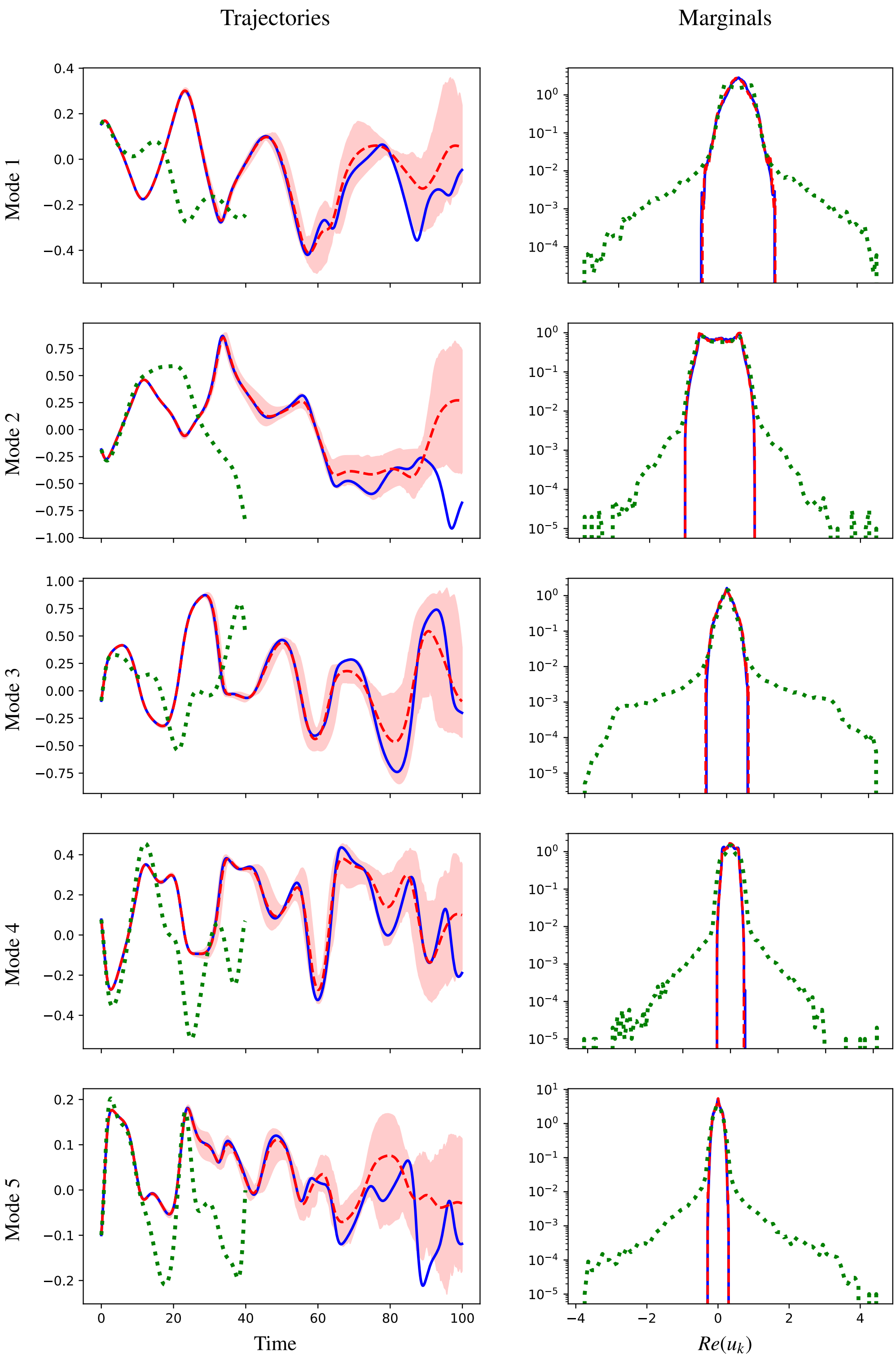


Fig. C.7. Comparison of finite-time forecasts and marginal distributions. In all panels, solid blue line is the full model (108-mode truncation), dashed red line is the 5-mode reduced model, and dotted green line the 5-mode truncation. *Left:* trajectories starting from the same initial conditions. For the reduced model, we show the 5th percentile, mean, and 95th percentile, computed with an ensemble of size 100. The truncated model was terminated at $t = 40$ to reduce clutter. *Right:* marginal densities.

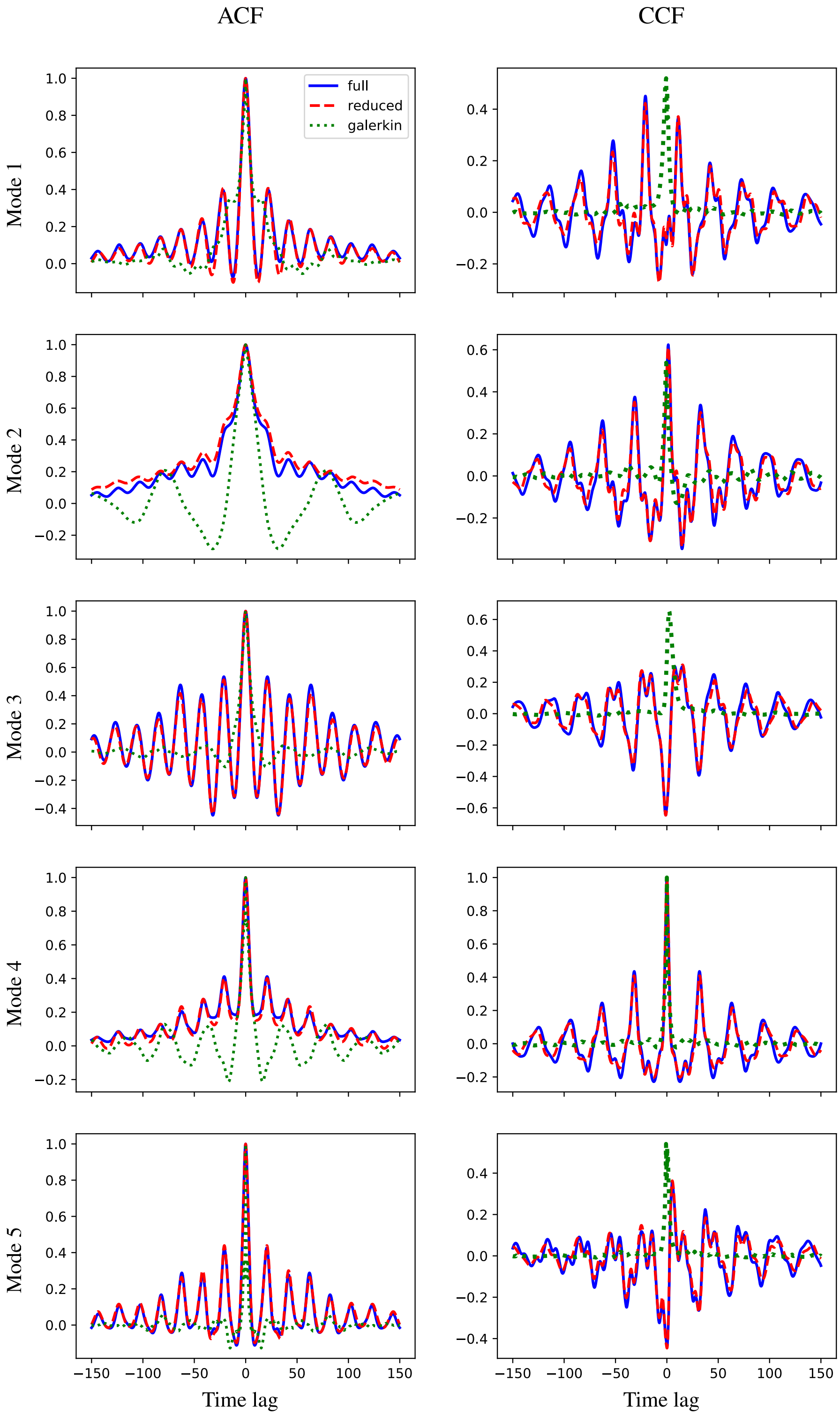


Fig. C.8. Comparison of autocovariance functions (ACFs) and energy cross correlation functions (CCFs). In all panels, solid blue line is the full model (108-mode truncation), dashed red line is the 5-mode reduced model, and dotted green line the 5-mode truncation. *Left:* Autocovariance functions for $Re(u_k(t))$ for $k = 1, \cdots, 5$. *Right:* Cross correlations between $|u_4(t)|^2$ and $|u_k(t)|^2$ for $k = 1, \cdots, 5$.

Detailed numerical results

Figs. C.7 and C.8 are full versions of the numerical results shown in Sect. 5.1.

To further quantify finite-time forecasts as a function of the "lead time" (i.e., time since initial observation), we introduce two standard measures of forecasting "skill," the root mean squared error and the anomaly correlation. Both are based on ensemble forecasts in the following way: let $v(t_n)$ denote the time series data for the full model, and take N_0 short pieces, i.e., $\{(v(t_n), n = n_i, n_i + 1, \dots, n_i + T)\}_{i=1}^{N_0}$ with $n_{i+1} = n_i + T_{lag}/\Delta t$, where $T = T_{lag}/\Delta t$ is the length of each piece and T_{lag} is the time gap between two adjacent pieces. For each short piece $(v(t_n), n = n_i, \dots, n_i + T)$, we generate N_{ens} trajectories of length T from the reduced model, starting all ensemble members from the same initial segment $\big(v(t_{n_i}), v(t_{n_i+1}), \dots, v(t_{n_i+m})\big)$, where $m = 2p + 1$, and denote the sample trajectories by $\big(u^n(i, j), n = 1, \dots, T\big)$ for $i = 1, \dots, N_0$ and $j = 1, \dots, N_{ens}$.

Again, we do not introduce artificial perturbations into the initial conditions, because the exact initial conditions are known, and by initializing from data, we preserve the memory of the system so as to generate better ensemble trajectories.

The *root mean squared error* is

$$\mathrm{RMSE}(\tau_n) := \left(\frac{1}{N_0} \sum_{i=1}^{N_0} \left| \mathrm{Re}\, v(t_{n_i+n}) - \mathrm{Re}\, \overline{u}^n(i) \right|^2 \right)^{1/2}, \tag{C.3}$$

where $\tau_n = n\Delta t$, $\overline{u}^n(i) = \frac{1}{N_{ens}} \sum_{j=1}^{N_{ens}} u^n(i, j)$, and the *anomaly correlation* (see, e.g., [77]) is

$$\mathrm{ANCR}(\tau_n) := \frac{1}{N_0} \sum_{i=1}^{N_0} \frac{\mathbf{a}^{v,i}(n) \cdot \mathbf{a}^{u,i}(n)}{\sqrt{|\mathbf{a}^{v,i}(n)|^2 \, \left|\mathbf{a}^{u,i}(n)\right|^2}}, \tag{C.4}$$

where $\mathbf{a}^{v,i}(n) = \mathrm{Re}\, v(t_{n_i+n}) - \mathrm{Re}\, \langle v \rangle$ and $\mathbf{a}^{u,i}(n) = \mathrm{Re}\, \overline{u}^n(i) - \mathrm{Re}\, \langle v \rangle$ are the anomalies in data and the ensemble mean. Here $\mathbf{a} \cdot \mathbf{b} = \sum_{k=1}^{K} a_k b_k$, $|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a}$, and $\langle v \rangle$ is the time average of the long trajectory of v. Both statistics measure the accuracy of the mean ensemble prediction: the RMSE measures, in an average sense, the difference between the mean ensemble trajectory, and the ANCR shows the average correlation between the mean ensemble trajectory and the true data trajectory. $\mathrm{RMSE} = 0$ and $\mathrm{ANCR} = 1$ would correspond to a perfect prediction, and small RMSEs and large (close to 1) ANCRs are desired.

For our reduced model, we computed the RMSE and ANCR using ensembles of $N_{ens} = 100$ trajectories with independent initial conditions. Fig. C.9 (left) shows the RMSE and ANCR for a range of lead times. As expected, the RMSE increases with lead time, and consistent with Fig. 1(a), it is relatively small compared to its apparent asymptotic value (about 0.6) for lead times < 50. The ANCR in Fig. C.9 (right) corroborates this. The two figures are comparable to Fig. 5 of [53] and show very similar trends.

Role of the noise terms ξ_n. We experimented with running the reduced model with $\xi_n \equiv 0$, i.e., without any noise term. This does not appreciably change the ACF or marginal distributions, nor the forecasting skill of the reduced model. However, the kind of ensemble prediction and uncertainty quantification illustrated in Fig. C.7 cannot be carried out without noise terms calibrated to the reduced model.

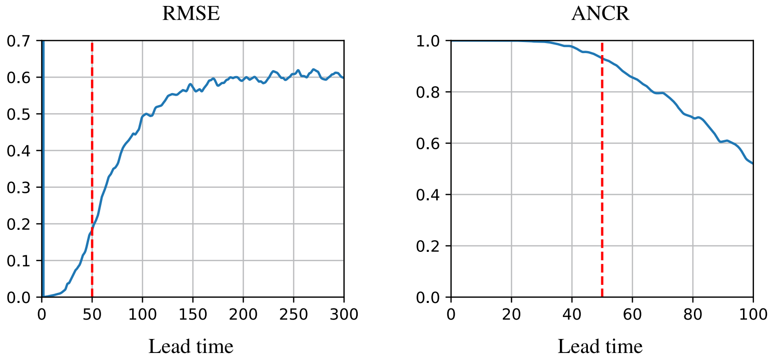


Fig. C.9. Forecasting skill as function of lead time of the reduced model for the KS equation. *Left:* root mean squared error (RMSE). *Right:* anomaly correlation (ANCR). See text for details.

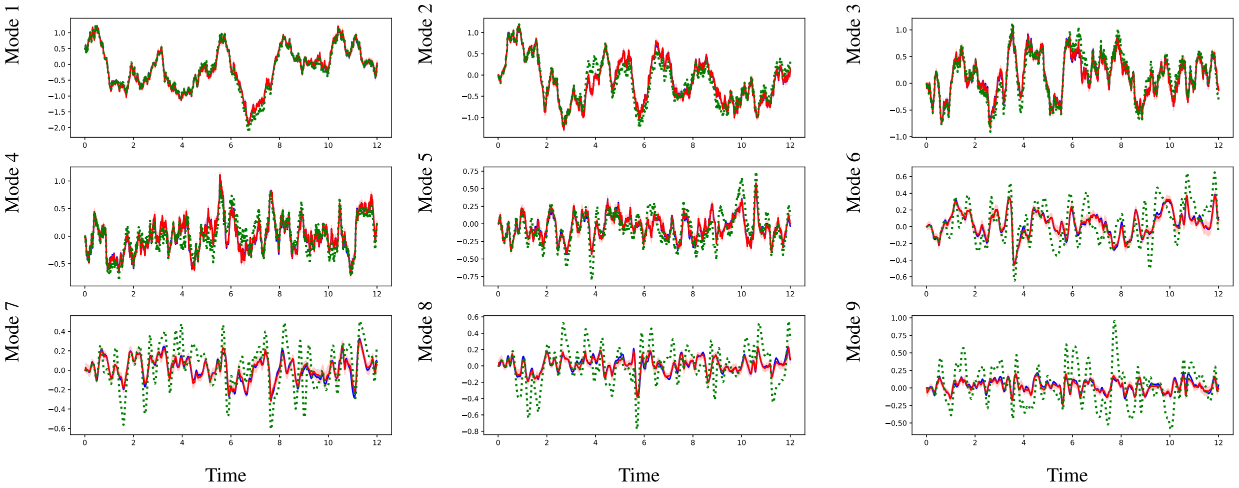


Fig. D.10. Response forecasting for the stochastic Burgers equation. For $k = 1, \cdots, 9$, we plot $Re(u_k(t))$ as functions of t. In all panels, solid blue line is the full model (128-mode truncation), dashed red line is the 9-mode reduced model, and dotted green line the 9-mode Galerkin truncation. Initial transients ($t < 8$) are not shown.

Appendix D. Stochastic Burgers equation

The nonlinear terms $\{\Psi_{n-j}\}$ in Eq. (5.7c) are defined by

$$\Psi^a_{n-j} = u^{n-j}, \quad \Psi^b_{n-j} = R^{\Delta t}(u^{n-j}), \quad \text{and} \quad \Psi^c_{n-j,k} = \sum_{\substack{|k-l|\le K, K<|l|\le 2K \\ \text{or } |l|\le K, K<|k-l|\le 2K}} \widetilde{u}_l^{n-1}\widetilde{u}_{k-l}^{n-j} \text{ for } k = 1, \cdots, K,$$

where the terms $\{\widetilde{u}\}$ are defined as

$$\widetilde{u}_k^{n-j} = \begin{cases} u_k^{n-j}, & 1 \le k \le K; \\ \frac{i\lambda_k}{2} e^{-\nu\lambda_k^2 j\delta} \sum_{\substack{|l|\le K, \\ |k-l|\le K}} u_{k-l}^{n-j}u_l^{n-j}, & K < k \le 2K. \end{cases} \tag{D.1}$$

These terms resemble those in Eq. (C.1d) as they are also introduced to represent the high modes by the low modes. But there is a major difference: they represent the high modes as a functional of the history of the low modes, rather than a function of the current state of the low modes. This is due to the lack of an inertial manifold for the Burgers equation, unlike the KSE. These terms are derived from an Riemann sum approximation of the integral equation for the high modes, with suitable linear parametrization of the quadratic interaction. A detailed derivation of the ansatz is presented in a forthcoming paper.

Figs. D.10–D.13 show numerical results for the stochastic Burgers equation.

References

[1] G. Pavliotis, A. Stuart, Multiscale Methods: Averaging and Homogenization, Springer Science & Business Media, 2008.
[2] I.G. Kevrekidis, G. Samaey, Equation-free multiscale computation: algorithms and applications, Annu. Rev. Phys. Chem. 60 (2009).
[3] A.J. Roberts, Model Emergent Dynamics in Complex Systems, vol. 20, SIAM, 2014.
[4] A. Abdulle, E. Weinan, B. Engquist, E. Vanden-Eijnden, The heterogeneous multiscale method, Acta Numer. 21 (2012) 1–87.
[5] A.J. Chorin, O.H. Hald, Stochastic Tools in Mathematics and Science, 3rd edition, Springer, New York, NY, 2013.
[6] R. Zwanzig, Nonequilibrium Statistical Mechanics, Oxford, 2001.
[7] D. Kondrashov, M.D. Chekroun, M. Ghil, Data-driven non-Markovian closure models, Physica D 297 (2015) 33–55, https://doi.org/10.1016/j.physd.2014.12.005.
[8] A.J. Chorin, F. Lu, Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics, Proc. Natl. Acad. Sci. USA 112 (32) (2015) 9804–9809.
[9] J. Harlim, X. Li, Parametric reduced models for the nonlinear Schrödinger equation, Phys. Rev. E 91 (5) (2015), https://doi.org/10.1103/PhysRevE.91.053306.
[10] H. Lei, N.A. Baker, X. Li, Data-driven parameterization of the generalized Langevin equation, Proc. Natl. Acad. Sci. USA 113 (50) (2016) 14183–14188, https://doi.org/10.1073/pnas.1609587113.
[11] X. Xie, M. Mohebujjaman, L.G. Rebholz, T. Iliescu, Data-driven filtered reduced order modeling of fluid flows, SIAM J. Sci. Comput. 40 (3) (2018) B834–B857, https://doi.org/10.1137/17M1145136.
[12] M.D. Chekroun, D. Kondrashov, Data-adaptive harmonic spectra and multilayer Stuart-Landau models, Chaos 27 (9) (2017) 093110, https://doi.org/10.1063/1.4989400, arXiv:1706.04275.
[13] T. Berry, D. Giannakis, J. Harlim, Bridging data science and dynamical systems theory, arXiv preprint arXiv:2002.07928, 2020.

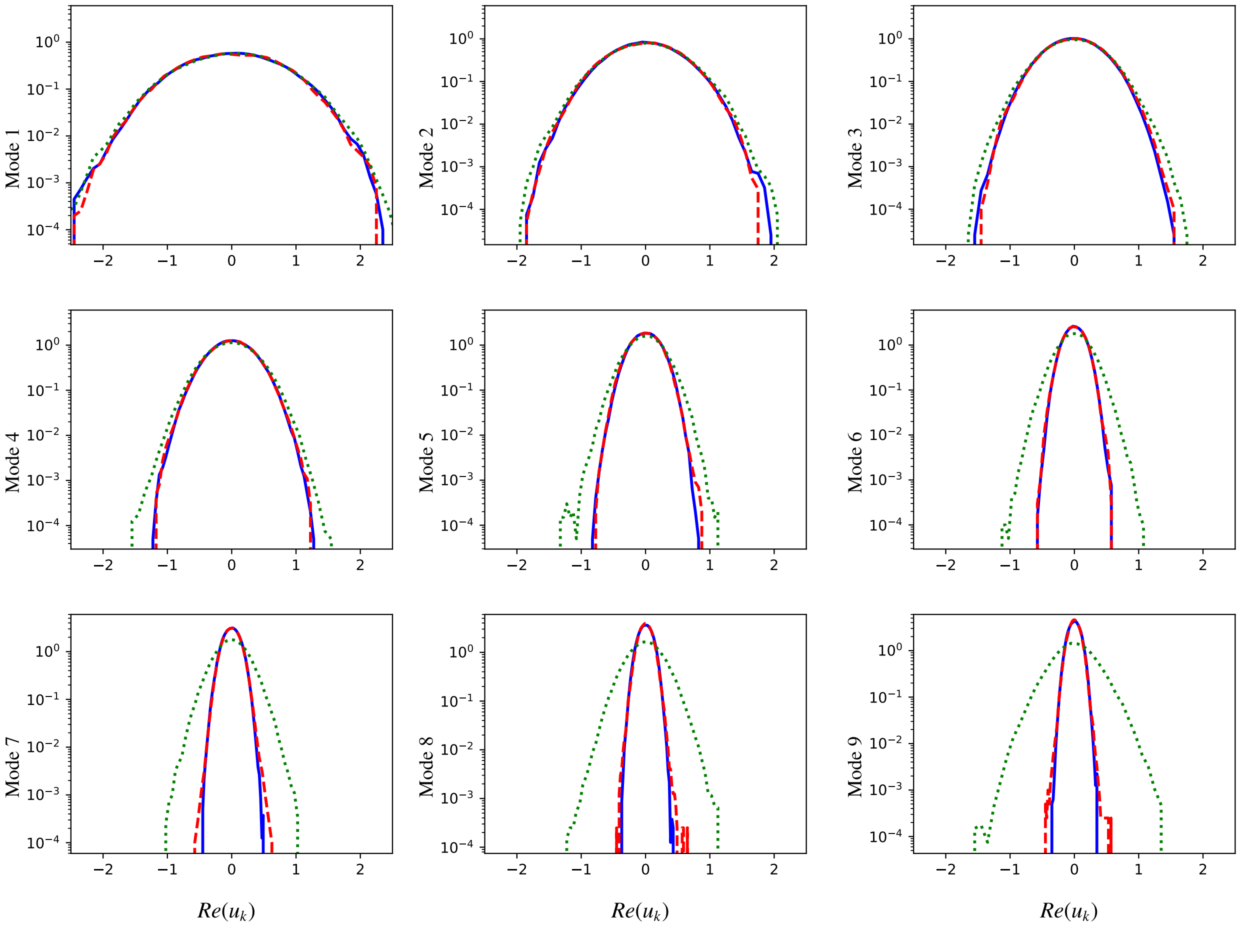


Fig. D.11. Marginal densities for the stochastic Burgers equation. We plot estimated densities for $Re(u_k)$ for $k = 1, \cdots, 9$. In all panels, solid blue line is the full model (128-mode truncation), dashed red line is the 9-mode reduced model, and dotted green line the 9-mode Galerkin truncation.

[14] J.N. Kutz, S.L. Brunton, B.W. Brunton, J.L. Proctor, Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems, SIAM, Philadelphia, PA, 2016.
[15] I. Mezić, Spectral properties of dynamical systems, model reduction and decompositions, Nonlinear Dyn. 41 (1–3) (2005) 309–325.
[16] M.O. Williams, I.G. Kevrekidis, C.W. Rowley, A data-driven approximation of the Koopman operator: extending dynamic mode decomposition, J. Nonlinear Sci. 25 (6) (2015) 1307–1346.
[17] J.D. Hamilton, Time Series Analysis, Princeton University Press, Princeton, NJ, 1994.
[18] S.A. Billings, Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatiotemporal Domains, John Wiley and Sons, 2013.
[19] F. Lu, K.K. Lin, A.J. Chorin, Comparison of continuous and discrete-time data-based modeling for hypoelliptic systems, Commun. Appl. Math. Comput. Sci. 11 (2) (2016) 187–216.
[20] P. Walters, An Introduction to Ergodic Theory, vol. 79, Springer Science & Business Media, 2000.
[21] M. Reed, B. Simon, Methods of Modern Mathematical Physics, vol. I, revised and enlarged edition, 1980.
[22] G. Froyland, K. Padberg, Almost-invariant sets and invariant manifolds-connecting probabilistic and geometric descriptions of coherent structures in flows, Physica D 238 (16) (2009) 1507–1523.
[23] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schütte, F. Noeé, Data-driven model reduction and transfer operator approximation, J. Nonlinear Sci. 28 (3) (2018) 985–1010.
[24] A.J. Chorin, O.H. Hald, R. Kupferman, Optimal prediction with memory, Physica D 166 (3–4) (2002) 239–257.
[25] L. Ma, X. Li, C. Liu, Coarse-graining Langevin dynamics using reduced-order techniques, J. Comput. Phys. 380 (2019) 170–190, https://doi.org/10.1016/j.jcp.2018.11.035.
[26] H. Cho, D. Venturi, G.E. Karniadakis, Statistical analysis and simulation of random shocks in stochastic Burgers equation, Proc. R. Soc. A, Math. Phys. Eng. Sci. 470 (2171) (2014) 20140080.
[27] Z. Li, H.S. Lee, E. Darve, G.E. Karniadakis, Computing the non-markovian coarse-grained interactions derived from the Mori–Zwanzig formalism in molecular systems: application to polymer melts, J. Chem. Phys. 146 (1) (2017) 014104.
[28] Z. Li, X. Bian, X. Li, G.E. Karniadakis, Incorporation of memory effects in coarse-grained modeling via the Mori-Zwanzig formalism, J. Chem. Phys. 143 (24) (2015) 243128.
[29] A. Panchenko, L.L. Barannyk, R.P. Gilbert, Closure method for spatially averaged dynamics of particle chains, Nonlinear Anal., Real World Appl. 12 (3) (2011) 1681–1697, https://doi.org/10.1016/j.nonrwa.2010.10.021, http://www.sciencedirect.com/science/article/pii/S1468121810002968.
[30] S.C. Venkataramani, R.C. Venkataramani, J.M. Restrepo, Dimension reduction for systems with slow relaxation, J. Stat. Phys. 167 (3–4) (2017) 892–933.
[31] P. Stinis, Stochastic optimal prediction for the Kuramoto–Sivashinsky equation, Multiscale Model. Simul. 2 (4) (2004) 580–612.
[32] E.J. Parish, K. Duraisamy, Non-markovian closure models for large eddy simulations using the Mori-Zwanzig formalism, Phys. Rev. Fluids 2 (1) (2017) 014604.

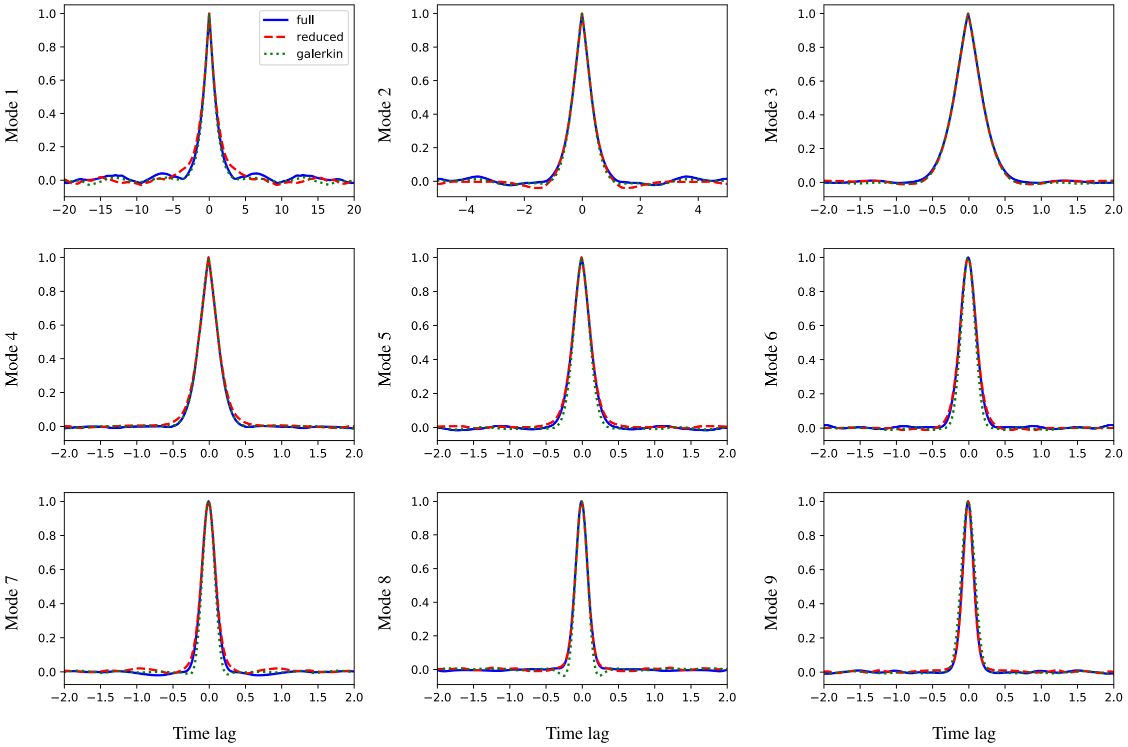


Fig. D.12. Autocovariance functions for the stochastic Burgers equation. We plot autocovariance functions for $Re(u_k)$ for $k = 1, \cdots, 9$. In all panels, solid blue line is the full model (128-mode truncation), dashed red line is the 9-mode reduced model, and dotted green line the 9-mode Galerkin truncation.

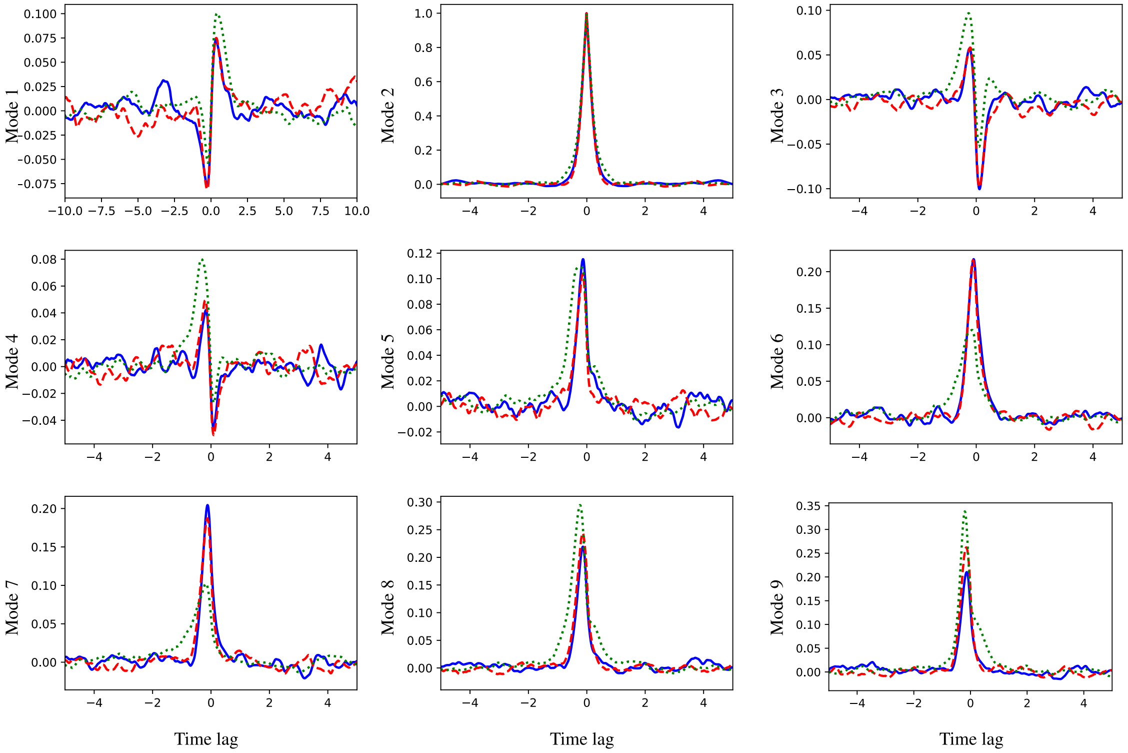
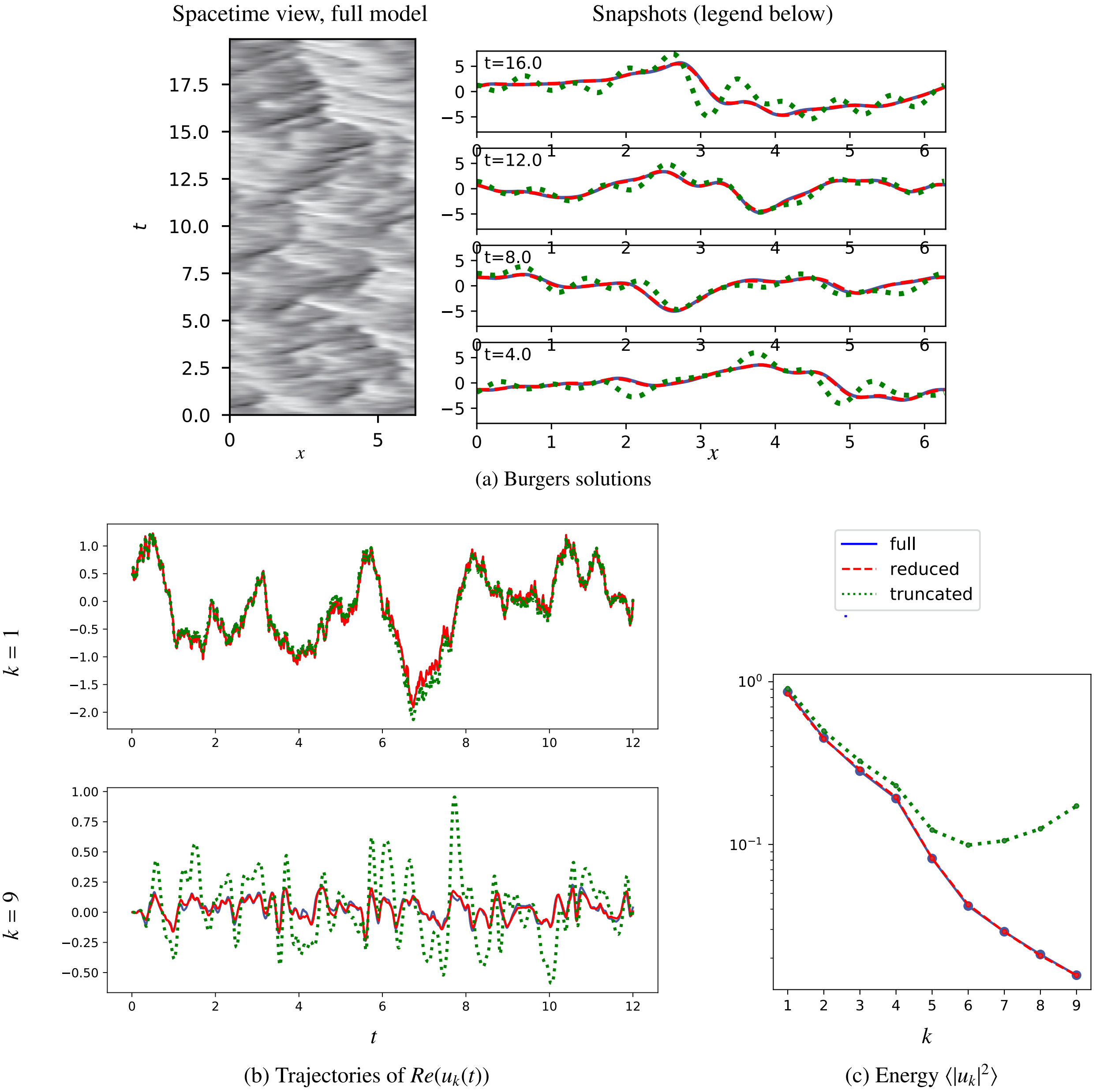


Fig. D.13. Energy cross-correlation functions for the stochastic Burgers equation. We plot cross correlation functions for $|u_2|^2$ and $|u_k|^2$ for $k = 1, \cdots, 9$. In all panels, solid blue line is the full model (128-mode truncation), dashed red line is the 9-mode reduced model, and dotted green line the 9-mode Galerkin truncation.



(a) Burgers solutions

(b) Trajectories of $Re(u_k(t))$

(c) Energy $\langle |u_k|^2 \rangle$

Fig. D.14. The results using a linear regression with $p = 1, r = 1$ as in Eq. (3.19). They are almost identical as those in Fig. 5 from a nonlinear regression using the model in Eq. (3.14) with $p = 1, r = 1$.

[33] S. Wang, Z. Li, W. Pan, Implicit-solvent coarse-grained modeling for polymer solutions via Mori-Zwanzig formalism, Soft Matter (2019), https://doi.org/10.1039/C9SM01211G.
[34] E. Darve, J. Solomon, A. Kia, Computing generalized Langevin equations and generalized Fokker-Planck equations, Proc. Natl. Acad. Sci. USA 106 (2009) 10884–10889.
[35] H. Grabert, Projection Operator Techniques in Nonequilibrium Statistical Mechanics, Springer, 1982.
[36] D. Forster, Hydrodynamic Fluctuations, Broken Symmetry, and Correlation Functions, CRC Press, 2018.
[37] A. Einstein, Investigations on the Theory of the Brownian Movement, Courier Corporation, 1956.
[38] J. Fan, Q. Yao, Nonlinear Time Series: Nonparametric and Parametric Methods, Springer, New York, NY, 2003.
[39] E.J. Hannan, Multiple Time Series, John Wiley and Sons, 1970.
[40] T. Kailath, Lectures on Wiener and Kalman Filtering, Springer, 1981.
[41] P. Brockwell, R. Davis, Introduction to Time Series and Forecasting, Springer, New York, NY, 2002.
[42] T. Berry, D. Giannakis, J. Harlim, Nonparametric forecasting of low-dimensional dynamical systems, Phys. Rev. E 91 (3) (2015) 032915.
[43] F. Ledrappier, L.S. Young, Entropy formula for random transformations, Probab. Theory Relat. Fields 80 (2) (1988) 217–240.
[44] Y. Kifer, Ergodic Theory of Random Transformations, vol. 10, Springer Science & Business Media, 2012.
[45] L. Arnold, Random Dynamical Systems, Springer Science & Business Media, 2013.
[46] P.H. Baxendale, The Lyapunov spectrum of a stochastic flow of diffeomorphisms, in: Lyapunov Exponents, Springer, 1986, pp. 322–337.

[47] H. Kunita, Stochastic Flows and Stochastic Differential Equations, vol. 24, Cambridge University Press, 1997.
[48] J. Bezanson, A. Edelman, S. Karpinski, V.B. Shah, Julia: a fresh approach to numerical computing, SIAM Rev. 59 (2017) 65–98.
[49] S.G. Johnson, The NLopt nonlinear-optimization package, http://ab-initio.mit.edu/nlopt, 2019.
[50] M.J.D. Powell, The BOBYQA algorithm for bound constrained optimization without derivatives, Tech. Rep. NA2009/06, Cambridge University, 2009.
[51] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical recipes, in: The Art of Scientific Computing, 3rd edition, Cambridge University Press, 2007.
[52] C. Cameron, Relative efficiency of gaussian stochastic process sampling procedures, J. Comput. Phys. 192 (2) (2003) 546–569.
[53] F. Lu, K.K. Lin, A.J. Chorin, Data-based stochastic model reduction for the Kuramoto–Sivashinsky equation, Physica D 340 (2017) 46–57.
[54] J.M. Hyman, B. Nicolaenko, The Kuramoto-Sivashinsky equation: a bridge between PDEs and dynamical systems, Physica D 18 (1986) 113–126.
[55] S.M. Cox, P.C. Matthews, Exponential time differencing for stiff systems, J. Comput. Phys. 176 (2) (2002) 430–455.
[56] A.K. Kassam, L.N. Trefethen, Fourth-order time stepping for stiff PDEs, SIAM J. Sci. Comput. 26 (4) (2005) 1214–1233.
[57] W. E, K. Khanin, A. Mazel, Y. Sinai, Invariant measure for Burgers equation with stochastic forcing, Ann. Math. 151 (3) (2000) 877–960.
[58] J. Bunder, A.J. Roberts, Resolution of subgrid microscale interactions enhances the discretisation of nonautonomous partial differential equations, Appl. Math. Comput. 304 (2017) 164–179.
[59] P.E. Kloeden, E. Platen, Numerical Solution of Stochastic Differential Equations, 3rd edition, Springer, Berlin, 1999.
[60] F. LuData-driven model reduction for stochastic Burgers equations. Preprint. 2020.
[61] B.A. Freno, K.T. Carlberg, Machine-learning error models for approximate solutions to parameterized systems of nonlinear equations, Comput. Methods Appl. Mech. Eng. 348 (2019) 250–296.
[62] S.L. Brunton, J.L. Proctor, J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. USA (2016) 201517384.
[63] S.W. Jiang, J. Harlim, Modeling of missing dynamical systems: deriving parametric models using a nonparametric framework, Res. Math. Sci. 7 (3) (2020) 1–25.
[64] D. Mukhin, A. Gavrilov, A. Feigin, E. Loskutov, J. Kurths, Principal nonlinear dynamical modes of climate variability, Sci. Rep. 5 (2015) 15510, https://doi.org/10.1038/srep15510, http://www.nature.com/srep/2015/151022/srep15510/full/srep15510.html.
[65] T. Berry, J.R. Cressman, Z. Greguric-Ferencek, T. Sauer, Time-scale separation from diffusion-mapped delay coordinates, SIAM J. Appl. Dyn. Syst. 12 (2) (2013) 618–649.
[66] C. Ma, J. Wang, W. E, Model reduction with memory and the machine learning of dynamical systems, arXiv:1808.04258, 2018.
[67] J. Pathak, B. Hunt, M. Girvan, Z. Lu, E. Ott, Model-free prediction of large spatiotemporally chaotic systems from data: a reservoir computing approach, Phys. Rev. Lett. 120 (2) (2018), https://doi.org/10.1103/PhysRevLett.120.024102.
[68] D.S. Broomhead, G.P. King, Extracting qualitative dynamics from experimental data, Physica D 20 (2–3) (1986) 217–236.
[69] M.D. Chekroun, I. Koren, H. Liu, Efficient reduction for diagnosing Hopf bifurcation in delay differential systems: applications to cloud-rain models, Chaos 30 (5) (2020) 053130.
[70] J. Duan, W. Wei, Effective Dynamics of Stochastic Partial Differential Equations, Elsevier, 2014.
[71] P.J. Schmid, Dynamic mode decomposition of numerical and experimental data, J. Fluid Mech. 656 (2010) 5–28.
[72] J.H. Tu, C.W. Rowley, D.M. Luchtenburg, S.L. Brunton, J.N. Kutz, On dynamic mode decomposition: theory and applications, J. Comput. Dyn. 1 (2) (2014) 391–421.
[73] M.I. Freidlin, A.D. Wentzell, Random Perturbations of Dynamical Systems, 3rd edition, Grundlehren der Mathematischen Wissenschaften (Fundamental Principles of Mathematical Sciences), vol. 260, Springer, Heidelberg, 2012, translated from the 1979 Russian original by Joseph Szücs.
[74] N. Wiener, Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications, 1949, the Technology Press of the Massachusetts Institute of Technology, Cambridge, MassJohn Wiley & Sons, Inc., New York, N. Y; Chapman & Hall, Ltd., London.
[75] A.M. Yaglom, An Introduction to the Theory of Stationary Random Functions, Revised English edition, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1962. Translated and edited by Richard A. Silverman.
[76] A.M. Yaglom, Correlation Theory of Stationary and Related Random Functions, vol. I, Springer Series in Statistics, Springer-Verlag, New York, 1987, basic results.
[77] D. Crommelin, E. Vanden-Eijnden, Subgrid-scale parameterization with conditional Markov chains, J. Atmos. Sci. 65 (8) (2008) 2661–2675.