Article

Data-driven model reduction for stochastic Burgers equations

Fei Lu [1,†,‡]

[1] Department of Mathematics, Johns Hopkins University; feilu@math.jhu.edu
[*] Correspondence: feilu@math.jhu.edu;
[†] Current address: 3400 N. Charles Street, Baltimore, MD 21218

Abstract: We present a class of efficient parametric closure models for 1D stochastic Burgers equations. Casting it as statistical learning of the flow map, we derive the parametric form by representing the unresolved high wavenumber Fourier modes as functionals of the resolved variables' trajectory. The reduced models are nonlinear autoregression (NAR) time series models, with coefficients estimated from data by least squares. The NAR models can accurately reproduce the energy spectrum, the invariant densities, and the autocorrelations. Taking advantage of the simplicity of the NAR models, we investigate maximal and optimal space-time reduction. Reduction in space dimension is unlimited, and NAR models with two Fourier modes can perform well. The NAR model's stability limits time reduction, with a maximal time step smaller than that of the K-mode Galerkin system. We report a potential criterion for optimal space-time reduction: the NAR models achieve minimal relative error in the energy spectrum at the time step where the K-mode Galerkin system's mean CFL number agrees with the full model's.

Keywords: data-driven modeling, stochastic Burgers equation, closure model, CFL number

1. Introduction

Closure modeling aims for computationally efficient reduced models for tasks requiring repeated simulations such as Bayesian uncertainty quantification[1,2] and data assimilation [3,4]. Consisting of low-dimensional resolved variables, the closure model must take into account the non-negligible effects of unresolved variables. As suggested by the Mori-Zwanzig formalism [5–7], trajectory-wise approximation is no longer appropriate, and the approximation is in a statistical sense. That is, the reduced model aims to generate a process that approximates the target process in distribution, or at least, reproduce the key statistics and dynamics for the quantities of interest. For general nonlinear systems, such a reduced closure model is out of the reach of direction derivation from first principles.

Data-driven approaches, which are based on statistical learning methods, provide useful and practical tools for model reduction. The past decay witnessed revolutionary developments of data-driven strategies, ranging from parametric models (see, e.g.[8–14] and the references therein) to nonparametric and machine learning methods (see, e.g.[15–18]). These developments demand a systematic understanding of model reduction from the perspectives of dynamical systems (see, e.g.[7,19,20]), numerical approximation [21,22], and statistical learning [23].

With 1D stochastic Burgers equation as a prototype model, we aim to bring further the understanding of model reduction from an interpretable statistical inference perspective. More specifically, we consider a stochastic Burgers equation with a periodic solution on $[0, 2\pi]$:

$$\begin{aligned} &u_t = \nu u_{xx} - u u_x + f(x,t)\,, 0 < x < 2\pi, t > 0 \\ &u(0,t) = u(2\pi,t) = 0, \quad u_x(0,t) = u_x(2\pi,t), \end{aligned} \tag{1.1}$$

doi:10.20944/preprints202010.0076.v1

from an initial condition $u(\cdot, 0)$. We consider a stochastic forces $f(x,t)$ that is smooth in space, residing on K_0 low wavenumber Fourier modes, and white in time, given by

$$f(x,t) = \sigma \sum_{m=1}^{K_0} \sin(mx)\dot{W}_m(t) + \cos(mx)\dot{W'}_m(t), \tag{1.2}$$

where $\{W_m, W'_m\}$ are independent Brown motions. We would like to find a discrete-time closure model for the first K Fourier modes, so as to efficiently reproduce the energy spectrum and other statistics of these modes.

We present a class of efficient parametric reduced closure models for 1D stochastic Burgers equations. The key idea is to approximate the discrete-in-time flow map statistically, in particular, to represent the unresolved high wavenumber Fourier modes as functionals of the resolved variables' trajectory. The reduced models are nonlinear autoregression (NAR) time series models, with coefficients estimated from data simply by least squares. We test the NAR models in four settings: reduction of deterministic responses ($K > K_0$) v.s. reduction involving unresolved stochastic force ($K < K_0$), and small v.s. large scales of stochastic force (with $\sigma = 0.2$ and $\sigma = 1$), where K_0 is the number of Fourier modes of the white-in-time stochastic force and σ is the scale of the force. In all these settings, the NAR models can accurately reproduce the energy spectrum, the invariant densities, and the autocorrelations. We also discuss model selection, consistency of estimators, and memory length of the reduced models.

Taking advantage of our NAR models' simplicity, we further investigate a critical issue in model reduction of (stochastic) partial differential equations: maximal space-time reduction. The space dimension can be reduced arbitrarily in our parametric inference approach: NAR models with two Fourier modes performs well. The time reduction is another story. The maximal time step is limited by the NAR model's stability and is smaller than those of the K-mode Galerkin system. Numerical tests indicate that the NAR models achieve the minimal relative error at the time step where the K-mode Galerkin system's mean CFL (Courant–Friedrichs–Lewy) number agrees with the full model's, suggesting a potential criterion for optimal space-time reduction.

One can readily extend our parametric closure modeling strategy to general nonlinear dissipative systems beyond quadratic nonlinearities. Along with [14], we may view it as a parametric inference extension of the nonlinear Galerkin methods [24–27]. But it does not require the existence of an inertial manifold (and the stochastic Burgers equation does not satisfy the spectral gap condition for the existence of an inertial manifold [28]), and it applies to resolved variables of any dimension (e.g., lower than the dimension of the inertial manifold if it exists [14]). Notably, one may use NAR models that are linear in parameters and estimate them by least squares. Therefore, the algorithm is computationally efficient and is scalable for large systems.

The stochastic Burgers equation is a prototype model for developing closure modeling techniques for turbulence (see e.g.[29–35]). In particular, Dolaptchiev etc [35] proposes a closure model for stochastic Burgers equation in a similar setting, based on local averages of finite-difference discretization, reproducing accurate energy spectrum similar to this study. We directly construct a simple yet effective NAR model for the Fourier modes, providing the ground of a statistical inference examination of the model reduction.

The exposition of our study proceeds as follows. Following a brief review of the basic properties of the stochastic Burgers equation and its numerical integration, we introduce in Section 2 the inference approach to closure modeling and compare it with the nonlinear Galerkin methods. Section 3 presents the inference of NAR models: derivation of the parametric form, parameter estimation, and model selection. Examining NAR models' performance in four settings in Section 4, we investigate the space-time reduction. Section 5 concludes our main findings and possible future research.

2. Space-time reduction for stochastic Burgers equations

In this section, we first review basic properties of the stochastic Burgers equation and its numerical integration. Then, we introduce inference-based model reduction and compare it with the nonlinear Galerkin methods.

2.1. The stochastic Burgers equation

A Fourier transform of Eq.(1.1) leads to

$$\frac{d}{dt}\widehat{u}_k = -\nu q_k^2 \widehat{u}_k - \frac{iq_k}{2}\sum_{l=-\infty}^{\infty} \widehat{u}_l \widehat{u}_{k-l} + \widehat{f}_k(t) \tag{2.1}$$

with $q_k = k, k \in \mathbb{Z}$, where $\widehat{u}_k$ are Fourier modes:

$$\widehat{u}_k(t) = \mathcal{F}[u]_k = \frac{1}{2\pi}\int_0^{2\pi} u(x,t)e^{-iq_k x}dx, \;\; u(x,t) = \mathcal{F}^{-1}[\widehat{u}] = \sum_k \widehat{u}_k(t)e^{iq_k x},$$

The system has the following properties. First, it is Galilean invariant: if $u(x,t)$ is a solution, then $u(x-ct,t)+c$, with c an arbitrary constant speed, is a solution. To see this, let $v(x,t) = u(x-ct,t)+c$. Then, $v_t = -cu_x + u_t$, $v_x = u_x$, and

$$v_t = cv_x + u_{xx} + uu_x + f = cv_x + v_{xx} + (v-c)v_x + f = v_{xx} + vv_x + f.$$

Without lost of generality, we set $\int u(x,0)dx = 0$. This implies that $\widehat{u}_0(0) = 0$. In this study, we only consider forces with mean zero, i.e. $\int_0^{2\pi} f(x,t)dx = 0$, therefore from Eq.(2.1), we see that $\widehat{u}_0(t) \equiv 0$, or equivalently, $\int u(x,t)dx \equiv 0$. Second, the zero solution is linearly stable. The diffusion term dissipates energy, while the stochastic force inputs energy, making the system ergodic [29,36]. Third, since u is real, the Fourier modes satisfies $\widehat{u}_{-k} = \widehat{u}_k^*$, where $\widehat{u}_k^*$ is the complex conjugate of $\widehat{u}_k$.

2.2. Galerkin spectral method

We consider the Galerkin spectral method for numerical solutions of the Burgers equation. The system is approximated as follows: the function $u(x,t)$ is represented at grid points $x_i = i\Delta x$ with $i = 0,\ldots,2N-1$ and $\Delta x = \frac{2\pi}{2N}$. The Fourier transform $\mathcal{F}$ is replaced by discrete Fourier transform

$$\widehat{u}_k = \mathcal{F}_{2N}[u]_k = \sum_{n=0}^{2N-1} u(x_i,t)e^{-iq_k x_i}, \;\; u(x_i) = \mathcal{F}_{2N}^{-1}[\widehat{u}]_i = \frac{1}{2N}\sum_{k=-N+1}^{N} \widehat{u}_k e^{iq_k x_i}.$$

Since u is real, we have $\widehat{u}_{-k} = \widehat{u}_k^*$. Noticing further that $\widehat{u}_0 = 0$ due to Galilean invariance, and setting $\widehat{u}_N = 0$, we obtain a truncated system

$$\frac{d}{dt}\widehat{u}_k = -\nu q_k^2 \widehat{u}_k - \frac{ik}{2}\sum_{|k-l|\le N, |l|\le N} \widehat{u}_l \widehat{u}_{k-l} + \hat{f}_k, \text{ with } |k| = 1,\ldots,N. \tag{2.2}$$

We solve Eq. (2.2) using the exponential time differencing fourth order Rouge–Kutta method (ETDRK4) (see [37,38]), with the force term $\widehat{f}_k$ treated as a constant in each time step. Such a mixture preserves both the numerical stability of ETDRK4 and the simplicity of Euler-Maruyama. We will set $\nu = 0.02$, $N = 128$ and $dt = 0.001$. The solution is accurately resolved, with mean Courant–Friedrichs–Lewy (CFL) numbers being 0.139 and 0.045 for $\sigma = 1$ and $\sigma = 0.2$, respectively. Here the mean CFL number is computed as the average along a trajectory with $N_t = 10^5$ steps

$$\text{Mean CFL number} = \frac{1}{N_t}\sum_{n=1}^{N_t} \sup_x |u(x,t_n)| \frac{\Delta t}{\Delta x},$$

where Δt and Δx are the time step and space step, respectively. Furthermore, numerical tests show that the marginal densities converge as trajectory length increases.

2.3. Nonlinear Galerkin and inferential model reduction

For simplicity of notation, we write the Burgers equation in an operator form as

$$\partial_t u + Au = B(u) + f, \;\; u(0) = u_0 \tag{2.3}$$

with a linear operator $A : H_0^1(0, 2\pi) \to L^2(0, 2\pi)$ and a nonlinear operator $B : H_0^1(0, 2\pi) \to L^2(0, 2\pi)$

$$A = -\nu\partial_{xx}, \quad B(u) = -(u^2)_x/2,$$

and with f being the stochastic force.

We first decompose the full model into resolved and unresolved variables. Recall that our goal of model reduction is to derive a closed system that can faithfully describe the dynamics of the coefficients $\{\widehat{u}_k(t)\}_{|k|=1}^K$, or equivalently, the low dimensional process $v(x,t) = \sum_{|k|=1}^K \widehat{u}_k(t)e^{iq_k x}$.

Denote by P the projection operator from $H_0^1(0, 2\pi)$ to $\text{span}\{e^{iq_k x}\}_{|k|=1}^K$, and let $Q := I - P$ (and for simplicity of notation, we will also denote them as projections on the corresponding vector spaces of Fourier modes). With $u = Pu + Qu = v + w$, we can write the system (2.3) as

$$\frac{dv}{dt} = -PAv + PB(v) + Pf + [PB(v+w) - PB(v)], \tag{2.4}$$

$$\frac{dw}{dt} = -QAw + QB(v+w) + Qf. \tag{2.5}$$

To find a closed system for v, we quantify the truncation error $PB(v+w) - PB(v)$ in (2.4), which represents the nonlinear interaction between the low and high wavenumber modes, by either a function of v or a functional of the trajectory of v. In particular, in the nonlinear Galerkin method based on inertial manifold theory, see e.g. [24–27]), one aims to represent the high modes w as a function of the low modes v (and hence obtaining an approximate inertial manifold). In the simplest implementation, one neglects the time derivative in equation (2.5) and solves $w = \psi(v)$ from

$$w \approx -(QA)^{-1}[QB(v+w) + Qf]$$

by fixed point iterations: $\psi_0 = 0, \;\; \psi_{i+1} = -A^{-1}[QB(u+\psi_i) + Qf]$. This leads to an approximation of w as a function of v, which exists if K is large enough and if the system satisfies a gap condition (so that an inertial manifold exists). However, among many dissipative systems with global attractor, only a few have been proven to satisfy the gap condition (see [28] for a recent review). More importantly, we can not always expect K to be larger than the dimension of an inertial manifold, which is unknown in general. Therefore, such an nonlinear Galerkin approach works for neither a system without an inertial manifold nor for a K smaller than the dimension of the inertial manifold.

We take a different perspective on the reduction. Unlike the nonlinear Galerkin which aims for a trajectory-wise approximation, we aim for a probabilistic approximation of the distribution of the stochastic process $(v(\cdot,t), t \geq 0)$. The randomness of the process v can come from random initial conditions and/or from the stochastic force. We emphasize that a key is to represent the dependence of the model error $PB(v+w) - PB(v)$ on the process v, not simply constructing a stochastic process with the same distribution as $PB(v+w) - PB(v)$, which may be independent of the process of v.

In a data-driven approach, such a probabilistic approximation leads naturally to the statistical inference of the underlying process, aiming to represent the model error $[PB(v+w) - PB(v)](t)$ as a functional of the past trajectory $(v(\cdot,s), s \leq t)$. This inferential reduction approach works flexibly for general settings: there is no need of an inertial manifold and the dimension K can be arbitrary (e.g. less than the dimension of the inertial manifold, as shown in [14]).

	Full model	Reduced model
State variables	$\widehat{u}_k(t_n)$ or $\widehat{u}(t_n)$ $v(x,t_n)$ or v $w(x,t)$ or w	u_k^n or u^n the vector $(u_{-K}^n,\ldots,u_K^n)$ NA
Stochastic force	$\widehat{f}_k(t_n)-\widehat{f}_k(t_{n-1})$	f_k^n

Table 1. Correspondence of the variables between the full and reduced models.

Space time reduction. To achieve a space-time reduction for practical computation, the reduced model should be a time series model with a time step $\delta > dt$ for time reduction, instead of a differential system. It approximates the flow map (with $t_n = n\delta$)

$$\widehat{u}_k(t_{n+1}) = F(\widehat{u}_\cdot(t_n), \widehat{f}_\cdot([t_n : t_{n+1}]))_k, \quad |k| \le K, \tag{2.6}$$

where $\widehat{u}_\cdot(t_n) = (\widehat{u}_k(t_n), |k| \ge 0)$ is the vector of all Fourier modes, and thus the above map is not a closed system for the low modes. Recall that for $|k| \le K$,

$$\underbrace{\frac{d}{dt}\widehat{u}_k = -\nu q_k^2\widehat{u}_k - \frac{ik}{2}\sum_{\substack{|k-l|\le K,\\ |l|\le K}} \widehat{u}_l\widehat{u}_{k-l}}_{K\text{-mode truncation}} \underbrace{-\frac{iq_k}{2}\sum_{\substack{|k-l|>K\\ \text{or } |l|>K}} \widehat{u}_l\widehat{u}_{k-l}}_{\text{truncation error}} + \widehat{f}_k(t) \tag{2.7}$$

Clearly, the K-mode truncated Galerkin system can provide an immediate approximation to F in (2.6). Making use of it, we we propose a time series model for $\{\widehat{u}_k(t_n)\}_{|k|=1}^K$ in the form of

$$u_k^{n+1} = u_k^n + R_k^\delta(u^n) + f_k^n + g_k^n + \Phi_k^n, \tag{2.8}$$

where $R_\cdot^\delta(u^n)$ is from a one-step forward integrator with time step-size δ of the deterministic K-mode Galerkin, and $f_k^n = \widehat{f}_k(t_n) - \widehat{f}_k(t_{n-1})$ is the increment of the kth Fourier modes of the original stochastic force at time t_n. Here the term $\Phi_\cdot^n$ and the stochastic force $g_\cdot^n$ aim to represent the truncation error, as well as the discretization error. Since the truncation error depends on the past history of the low wavenumber modes, and as suggested by the Mori-Zwanzig formalism [6,7], we make $\Phi_\cdot^n$ depend on the trajectory $u^{1:n}$ of the state process, as well as the trajectories $f^{1:n}$ and $g^{1:n}$:

$$\Phi^n = \Phi(u^{1:n}, f^{1:n}, g^{1:n}). \tag{2.9}$$

We assume the stochastic force g to be iid Gaussian for simplicity.

The right hand side of Eq.(2.8), together with Φ^n defined in Eq.(2.9), aims for a statistical approximation of the discrete-time map (2.6). However, the general form in Eq.(2.9) leads to a high dimensional function to be learned from data, which is intractable by regression methods using either global or local polynomial basis, due to the well-known curse of dimensionality. Fortunately, the physical model provides informative structures to reduce the dimension, and we can obtain effective approximations based on only a few basis functions with finite memory. In the next section, we derive from the physical model a parametric form for the reduced model, whose coefficients can be efficiently estimated from data.

To avoid confusions between notations, we summarize the correspondence of the variables between the full and reduced models in Table 1.

3. Inference of reduced models

We present here the parametric inference of NAR models: derivation of parametric forms, estimation of the parameters, and model selection.

3.1. Derivation of parametric reduced models

We derive parametric reduced models by extracting basis functions from numerical integration of Eq.(2.4). The combination of these basis functions will give us $\Phi(u^{1:n}, f^{1:n}, g^{1:n})$ in (2.9), which approximates the flow maps $\{F(\widehat{u}_{\cdot}(t_n), \widehat{f}_{\cdot}([t_n : t_{n+1}]))_k, |k| \le K\}$ in (2.6) in a statistical sense.

We first write a closed integro-differential system for the low-mode process $(v(\cdot, t), t \ge 0)$. In view of Eq.(2.4), this can be simply done by integrating the equation of the high modes w in Eq.(2.5):

$$\begin{cases} \frac{dv}{dt} & = -PAv + PB(v) + Pf + [PB(v+w) - PB(v)], \\ w(t) & = e^{-QA\tau} w(t-\tau) + \int_{t-\tau}^{t} e^{-QA(t-s)} [QB(v(s) + w(s)) + Qf(s)] ds, \end{cases} \tag{3.1}$$

where $\tau \in [0, t]$. Note that in addition to the trajectories $(v(\cdot, s), s \in [t-\tau, t])$ and $(Qf(s), s \in [t-\tau, t])$, which we can assume to be known, the state $w(\cdot, t)$ also depends on the initial condition $w(\cdot, t-\tau)$. Therefore, this equation is not strictly closed. But as τ increases, the effect of the initial condition decays exponentially, allowing for possible finite time approximate closure. Given $w(\cdot, t-\tau)$ and $(Qf(s), s \in [t-\tau, t])$, the Picard iteration can provide us an approximation of w as a functional of the trajectory of v. That is, the sequence of functions $\{w^{(l)}\}$, defined by

$$w^{(l+1)}(t) = e^{-QA\tau} w^{(l)}(t-\tau) + \int_{t-\tau}^{t} e^{-QA(t-s)} [QB(v(s) + w^{(l)}(s)) + Qf(s)] ds,$$

with $w^{(0)}(s) = 0$ for $s \in [t-\tau, t]$, will converge to w as $n \to \infty$. In particular, the first Picard iteration provides us a closed representation:

$$w^{(1)}(t) = \int_{t-\tau}^{t} e^{-QA(t-s)} [QB(v(s)) + Qf(s)] ds, \tag{3.2}$$

We can now propose parametric numerical reduced models from the above integro-differential equation. In a simple form, we parametrize both the Riemann sum approximation of the first Picard iteration and a numerical scheme of the differential equation to obtain

$$\begin{aligned} v(t_n) &\approx v(t_{n-1}) + a_1 \delta R^{\delta}(v(t_{n-1})) + a_2 \delta Pf(t_{n-1}) + \delta[PB(v+w) - PB(v)](t_{n-1}), \\ w(t_{n-1}) &\approx \sum_{j=0}^{p} c_j e^{-QAj\delta} [QB(v(t_{n-j})) + Qf(t_{n-j})]. \end{aligned}$$

Here $\delta = t_n - t_{n-1}$ denotes the time step-size, the nonlinear function $R^{\delta}(\cdot)$ comes from a numerical integration of the deterministic truncated Galerkin equation $\frac{dv}{dt} \approx -PAv + PB(v)$ at time t_{n-1} and with time step-size δ, and the coefficients (a_i, c_i) are to be estimated by fitting to data in a statistical sense. To distinguish the approximate process in the reduced model from the original process, we denote it by v^n, and write the reduced model as

$$v^n = v^{n-1} + a_1 \delta R^{\delta}(v^{n-1}) + a_2 \delta Pf(t_{n-1}) + \delta[PB(v^{n-1} + w^{n-1}) - PB(v^{n-1})] + g^n, \tag{3.3a}$$

$$w^{n-1} = \sum_{j=0}^{p} c_j e^{-QAj\delta} [QB(v^{n-j}) + Qf(t_{n-j})], \tag{3.3b}$$

where $\{g^n\}$ is a process representing the residual, can be assumed to be stochastic force for simplicity, but can also be assumed to be a moving average part to better capture the time correlation as in [13,39]. The second equation (3.3b) does not have a residual term, as its goal is to provide a set of basis functions for the approximation of the forward map $v(t_n) = F(v(t_{n-1}), w(t_{n-1}), f)$ as in Eq.(2.6), not to model the high modes.

Note that the time step-size δ can be relatively large, as long as the truncated Galerkin equation $\frac{dv}{dt} \approx -PAv + PB(v)$ of the slow variable v can be reasonably resolved. In general, such a step-size can be much larger than the time step-size needed to resolve the fast process w, because the effect of the unresolved fast process is "averaged" statistically when fitting the coefficients $\{a_j, c_j\}$ to data. Also, the numerical error in the discretization is taken into account statistically.

Theoretically, the right hand side of Eq.(3.3a) is an approximation of the conditional expectation $\mathbb{E}\left[v(t_n)|v(t_{n-p:n-1}), f(t_{n-p:n-1})\right]$, which is the optimal L^2 estimator of the forward map conditional on the information up time t_{n-1}. Here the L^2 is with respect to the joint measure of the vector $(v(t_{\cdot-p:\cdot-1}), f(t_{\cdot-p:\cdot-1}))$, which is approximated by their joint empirical measure when fitting to data.

To avoid nonlinear optimization, the parametric form may be further simplified to be linearly dependent on the coefficients by dropping the terms that are nonlinear in the parameter, which is quadratic. In fact, recall that in the Burgers equation $B(u) = uu_x$ and $PB(v+w) - PB(v) = v_x w + v w_x + w w_x$. By dropping the interaction between the high modes ww_x and approximating

$$PB(v^{n-1} + w^{n-1}) - PB(v^{n-1}) \approx v_x^{n-1} w^{n-1} + v^{n-1} w_x^{n-1}$$

in (3.3a), we obtain a reduced model that depends linearly on the coefficients $\{a_j, c_j\}$.

3.2. The numerical reduced model in Fourier modes.

We now write the reduced model in terms of the Fourier modes as in Eq.(2.8).

As discussed in the above section, the major task is to parametrize the truncation error $PB(v+w)_k - PB(v)_k$. Recall that the operator P projects u to modes with wavenumber $1 \leq |k| \leq K$ and that the bilinear function $PB(v)_k = \sum_l \widehat{u}_l \widehat{u}_{k-l}$ (hereafter, to simplify notation, we also denote P and Q on the corresponding vector spaces of Fourier modes).

$$PB(v+w)_k - PB(v)_k = -\frac{ik}{2} \sum_{|l|>K \text{ or } |k-l|>K} \widehat{u}_l \widehat{u}_{k-l}. \tag{3.4}$$

Since the quadratic term $B(v)$ can only propagate energy from $(\widehat{u}_k, 1 \leq |k| \leq K)$ to modes with wave numbers less than $2K+1$, we get only the high modes with wave numbers $K < |k| \leq 2K$ when we compute w by a single iteration of $QB(v)$. Therefore, in a single iteration approximation, the truncated error will involve the first $2K$ Fourier modes:

$$PB(v+w)_k - PB(v)_k \approx -\frac{ik}{2} \sum_{\substack{K<|k-l|\leq 2K \\ \text{or } K<|l|\leq 2K}} \widehat{u}_l \widehat{u}_{k-l}.$$

Dropping the interaction between the high-modes to avoid nonlinear optimization in parameter estimation, we have

$$PB(v+w)_k - PB(v)_k \approx -\frac{ik}{2} \sum_{\substack{|k-l|\leq K, K<|l|\leq 2K \\ \text{or } |l|\leq K, K<|k-l|\leq 2K}} \widehat{u}_l \widehat{u}_{k-l}.$$

We approximate the high modes $(\widehat{u}_k, K < |k| \leq 2K)$ by a functional of low modes as in (3.3b),

$$\widehat{u}_k(t_{n-1}) \approx \sum_{j=1}^{p} c_{k,j} e^{-QAj\delta} [\widetilde{u}_k(t_{n-j}) + \widehat{f}_k(t_{n-1})], \quad K < |k| \leq 2K$$

where $\widetilde{u}_k$ is the high modes of the nonlinear function $B(v)$:

$$\widetilde{u}_k = QB(v)_k = -\frac{ik}{2} \sum_{|l|\leq K, |k-l|\leq K} \widehat{u}_l \widehat{u}_{k-l}, \text{ for } K < |k| \leq 2K.$$

Here $QB(v)$ only represents the modes up to wavenumber $2K$, due to that the quadratic nonlinearity only involves interaction between double wave-numbers. One can reach higher wave numbers by iterations of the quadratic interaction.

The truncation error term can now be linearly parametrized as

$$[PB(v+w)-PB(v)]_k(t_n)) \approx -\frac{iq_k}{2}\sum_{j=0}^{p} c_{k,j}e^{-QAj\delta} \sum_{\substack{|k-l|\leq K, K<|l|\leq 2K \\ \text{or } |l|\leq K, K<|k-l|\leq 2K}} \widetilde{u}_l(t_n)\widetilde{u}_{k-l}(t_{n-j}), \tag{3.5}$$

where we also denote $\widetilde{u}_k = \widehat{u}_k$ for $|k| \leq K$ for simplicity of notation.

We have now reached a parametric numerical reduced model for the Fourier modes. Denote $u^n = (u_k^n, |k| \leq K) \in \mathbb{C}^K$ the low-modes in the reduced model that approximates the original low-modes $(\widehat{u}_k(t_n), |k| \leq K)$. The reduced model is

$$u_k^n = u_k^{n-1} + \delta[R^\delta(u_\cdot^{n-1}) + f_k^n + \Phi_k^n] + g_k^n, \quad 1 \leq k \leq K, \tag{3.6a}$$

$$\Phi_k^n = \sum_{j=1}^{p}\left[c_{k,j}^v u_k^{n-j} + c_{k,j}^R R^\delta(u_\cdot^{n-j}) + c_{k,j}^f f_k^{n-j} + c_{k,j}^w \sum_{\substack{|k-l|\leq K, K<|l|\leq 2K \\ \text{or } |l|\leq K, K<|k-l|\leq 2K}} \widetilde{u}_l^{n-1}\widetilde{u}_{k-l}^{n-j}\right] \tag{3.6b}$$

with the convention that $u_{-k}^n = (u_k^n)^*$ (with the sup-script * denoting complex conjugate), and where the notion $\widetilde{u}_l^{n-j}$ represents the high modes and is defined by

$$\widetilde{u}_k^{n-j} = \begin{cases} u_k^{n-j}, & 1 \leq k \leq K; \\ \frac{iq_k}{2}e^{-\nu q_k^2 j\delta}\sum_{|l|\leq K, |k-l|\leq K} u_{k-l}^{n-j}u_l^{n-j}, & K < k \leq 2K. \end{cases} \tag{3.7}$$

The reduced model is in form of a nonlinear auto-regression moving average (NARMA) model:

- The map $R^\delta(\cdot): \mathbb{C}^K \to \mathbb{C}^K$ is the 1-step forward of the deterministic K-mode Galerkin truncation equation $\frac{dv}{dt} = -PAv + PB(v)$ using a numerical integration scheme with a time step-size δ, i.e. $v^{n+1} = v^n + \delta\mathbb{R}^\delta(v^n)$. We use the ETDRK4 scheme.
- The term f_k^n denotes the increment of the k-th Fourier modes of the original stochastic force in the time interval $[t_{n-1}, t_n]$, scaled by $1/\delta$, and it is separated from R^δ so that the reduced model can linearly quantify the response of the low-modes to the stochastic force.
- The function $\Phi_k^n := \Phi_k^n(u^{n-p:n-1}, f^{n-p:n-1})$ is a function $\mathbb{C}^{Kp+Kp} \to \mathbb{C}^K$ with parameters $\theta = (c^v, c^R, c^f, c^w) \in \mathbb{R}^{4Kp}$ to be estimated from data. In particular, the coefficients $c_{k,1}^v$ and $c_{k,1}^R$ act as a correction to the integration of the truncated equation.
- The new stochastic force term $\{g^n \in \mathbb{C}^K\}$ are assumed for simplicity to be stochastic force and independent of the original stochastic force (f^n). That is, we assume that $\{g^n\}$ is a sequence of independent identically distributed (iid) Gaussian random vectors, with independent real and imaginary parts, distributed as $\mathcal{N}(0, \mathrm{Diag}(\sigma_k^g))$, with σ_k^g to be estimated from data. In general, one can also assume other distributions for g^n, or other structures such as moving average $\{g^n := \xi_n + \sum_{j=1}^{q} c_j^g \xi_{n-j}\}$ with $\{\xi_n\}$ being a stochastic force sequence [13,39].

We remark that the right hand side of Eq.(3.6a) is an approximation of the conditional expectation function $\mathbb{E}\left[v(t_n)|v(t_{n-p:n-1}), f(t_{n-p:n-1})\right]$, which is the optimal least squares estimator of the forward map of the dynamics.

3.3. Data generation and Parameter estimation

We estimate the parameters of the NAR model by maximizing the likelihood of the data.

Data for the NAR model. To infer a reduced model in form of Eq.(3.6), we generate relevant data from a numerical scheme that sufficiently resolve the system in space and time as introduced in Section 2.2. The data relevant are trajectories of the low-modes of the state and the stochastic force, i.e. $\{\widehat{u}_k(t_n), \widehat{f}_k(t_n)\}$ for $|k| \leq K$ and $n \geq 0$, which are taken as $\{u_k^n, f_k^n\}$ in the reduced model. Here the time instants are $t_n = n\delta$, where δ can be much larger than the time step-size dt needed to resolve the system. Also, the data does not include the high modes. In short, the data are generated by a downsampling, in both space and time, of the high-resolution solutions of the system.

The data can be either a long trajectory or many independent short trajectories. We denote the data consisting of M independent trajectories by

$$\text{Data:} \quad \{u_k^{1:N_t,m}, f_k^{1:N_t,m}\}_{m,k=1}^{M,K} \text{ with } u_k^{1:N_t,m} = \widehat{u}_k(t_{1:N_t})^{(m)}, f_k^{1:N_t,m} = \widehat{f}_k(t_{1:N_t})^{(m)}, \tag{3.8}$$

where m indexes the trajectories, $t_n = n\delta$ with δ being the time interval between two observations, and N_t denotes the number of steps for each trajectory,

Parameter estimation. The parameters in the discrete-time reduced model Eq.(3.6) is estimated by maximum likelihood methods. Our discrete-time reduced model has a few attractive features: (i) the likelihood function can be computed exactly, avoiding possible approximation error that could lead to biases in estimators; (ii) the maximum likelihood estimator (MLE) may be computed by least squares under the assumption that the process $\{g^n\}$ is white noise, avoiding time-consuming nonlinear optimizations.

Under the assumption that $\{g^n\}$ is white noise, the parameters can be estimated simply by least squares, because the reduced model in Eq.(3.6) depends linearly on the parameters. More precisely, the log-likelihood of the data $\{u^{1:N_t,m}, f^{1:N_t,m}\}_{m=1}^{M}$ in (3.8) can be written as

$$l(\theta, \sigma^g) = -\sum_{|k|\leq K} \left[\log \sigma_k^g + \sum_{n,m=1}^{T,M} \frac{|u_k^{n,m} - u_k^{n-1,m} + \delta R^\delta(u_k^{n-1,m}) + \delta f_k^{n,m} + \delta \Phi_k^{n,m}(\theta)|^2}{2MT\sigma_k^g} \right], \tag{3.9}$$

where $|\cdot|$ denotes the absolute value of a complex number, $\theta = (c^v, c^R, c^f, c^w) \in \mathbb{R}^{4Kp}$ and $\sigma^g = (\sigma_1^g, \cdots, \sigma_K^g) \in \mathbb{R}^K$. To compute the maximum likelihood estimator (MLE) of the parameter (θ, σ^g), we note that $\Phi_k^n(\theta)$ in (3.6b) depends linearly on the parameter θ. Therefore, the estimators of θ and σ^g can be analytically computed by finding a zero of the gradient of the likelihood function. More precisely, denoting

$$\Phi_k^n(\theta) = \sum_{j=1}^{4p} \theta_j \Phi_{k,j}^n$$

with $\Phi_{k,j}^n$ denoting the parameterized terms in (3.6b), we compute the MLE as

$$\begin{aligned} \widehat{\theta}_k &= (\mathbf{A}_k)^{-1}\mathbf{b}_k, \quad 1 \leq k \leq K, \\ \widehat{\sigma}_k^g &= \frac{1}{MT} \sum_{n,m=1}^{T,M} \|u_k^{n,m} - u_k^{n-1,m} + \delta R^\delta(u_k^{n-1,m}) + \delta f_k^{n,m} + \delta \Phi_k^{n,m}(\widehat{\theta})\|^2 \end{aligned} \tag{3.10}$$

where the normal matrix $\mathbf{A}_k$ and vector $\mathbf{b}_k$ are defined by

$$\begin{aligned} \mathbf{A}_k(j', j) &= \frac{\delta}{MT} \sum_{n,m=1}^{T,M} \langle \Phi_{k,j'}^{n,m}, \Phi_{k,j}^{n,m} \rangle, \quad 1 \leq j', j \leq 4p, \\ \mathbf{b}_k(j) &= \frac{1}{MT} \sum_{n,m=1}^{T,M} \langle u_k^{n,m} - u_k^{n-1,m} + \delta R^\delta(u_k^{n-1,m}) + \delta f_k^{n,m}, \Phi_{k,j}^{n,m} \rangle. \end{aligned} \tag{3.11}$$

We assume for simplicity that the stochastic force g has independent components, so that the coefficients can be estimated by simple least square regression. One may further improve the NAR model by considering spatial correlation between the components of g or by using moving average models [13,39] to account for the memory in the stochastic force.

3.4. Model selection

The parametric form in Eq.(3.6b) leaves a family of reduced models with many freedoms underdetermined, such as the time lag p and possible redundant terms. To avoid overfitting and redundancy, we proposed to select the reduced model by the following criterion.

- Cross validation: the reduced model should be stable and can reproduce the distribution of the resolved process, particularly the main dynamical-statistical properties. We will consider the energy spectrum, the marginal invariant densities, and temporal correlations:

$$\begin{aligned}\text{Energy spectrum: } \mathbb{E}|\widehat{u}_k|^2 &= \lim_{N_t M\to\infty} \frac{1}{N_t M}\sum_{m,n=1}^{M,N_t} |\widehat{u}_k(t_n)^{(m)}|^2;\\ \text{Invariant density of } \mathrm{Re}(\widehat{u}_k)\text{: } p_k(z)dz &= \lim_{N_t M\to\infty} \frac{1}{N_t M}\sum_{m,n=1}^{M,N_t} \mathbf{1}_{(z,z+dz)}(\mathrm{Re}(\widehat{u}_k(t_n)^{(m)})),\\ \text{Temporal correlations: } C_k^2(h) &= \mathbb{E}[\mathrm{Re}\widehat{u}_k(t+h)\mathrm{Re}\widehat{u}_k(t)];\end{aligned} \quad (3.12)$$

  for $k = 1, \ldots, K$.
- Consistency of the estimators. If the model is perfect and the data are either independent trajectories or a long trajectory from an ergodic measure, the estimators should converge as the data size increases (see e.g.,[40,41]). While our parametric model may not be perfect, the estimators should also become less oscillatory as the data size increases, so that the algorithm is robust and can yield similar reduced models from different data sets.
- Simplicity and sparsity. When there are multiple reduced models performing similarly, we prefer the simplest model. We remove the redundant terms and enforce sparsity by LASSO (least absolute shrinkage and selection operator) regression [42]. Particularly, a singular normal matrix (3.11) indicates the redundancy of the terms and the need to remove strongly correlated terms.

These criteria are by no means exhaustive. Other methods include Bayesian information criterion (BIC, see, e.g. [43]) and the error reduction ratio [44] may be applied, but in our experience, they provide limited help for the selection of reduced models [7,14,39].

In view of statistical learning of the high-dimensional nonlinear flow map in (2.6), each linear-in-parameter reduced model provides an optimal approximation to the flow map in the function space spanned by the proposed terms. A possible future direction is to select adaptive-to-data hypothesis spaces in a nonparametric fashion [23] and analyze the distance between the flow map and the hypothesis space spanned by these proposed terms [45,46].

4. Numerical study on space-time reduction

We examine the inference and performance of NAR models for the stochastic Burgers equation in (1.1)-(1.2). We will consider two settings of the full model: the stochastic force has a scale either $\sigma = 1$ or $\sigma = 0.2$, representing that the stochastic force either dominates or subordinates to the dynamics, respectively. We will also consider two settings for reduction: the number of the Fourier modes of the reduced model is either $K > K_0$ or $K < K_0$, representing a reduction of the deterministic responses and a reduction involving stochastic force, respectively.

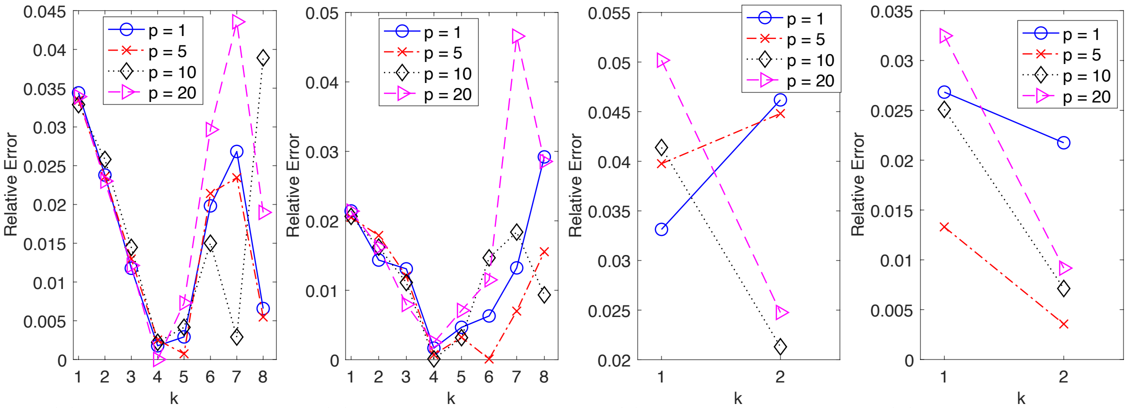
4.1. Settings

As reviewed in Section 2.2, we integrate the equation (2.2) of N Fourier modes by ETD-RK4 with a time-stepping dt that the solution is resolved accurately. We call this discretized system the full model and its configuration is specified in Table 2. We will consider two different scales for the stochastic

doi:10.20944/preprints202010.0076.v1

Table 2. Settings of the full and reduced models

Full model	$\nu = 0.02, L = 1$	viscosity, interval length of the equation
	$N = 256, dt = 0.001$	number of modes, time step-size
	$K_0 = 4$	number of modes in the stochastic force
	$\sigma = 1$ or 0.2	standard deviation of the stochastic force
Reduced models	$K = 8$ or 2	number of modes in the reduced model
	$\delta = dt \times \text{Gap}$	observation time interval
	$\text{Gap} \in \{5, 10, 20, 30, 40, 50, 80, 160\}$	gap of time steps



(a) $K = 8$, $\sigma = 1$ (b) $K = 8$, $\sigma = 0.2$ (c) $K = 2$, $\sigma = 1$ (d) $K = 2$, $\sigma = 0.2$

Figure 1. Relative error in energy spectrum reproduced by the NAR models with different memory lengths p, in four settings of (K, σ). As the time lag p increases, the relative error tends to first decrease and then increase, particularly in (b) and (d) with $\sigma = 0.2$.

force, with standard deviations $\sigma = 1$, leading to a dynamics dominated by the stochastic force, and $\sigma = 0.2$, representing a dynamics dominated by the deterministic drift.

We generate data in (3.8) from the full model as described in Section 3.3. We generate an ensemble of initial conditions by first integrating the system for 10^4 time units from an initial condition $u_0(x) = \sin(x) + 2\cos(x)$ and draw 10^3 samples uniformly from this long trajectory. Then we generate either a long trajectory or an ensemble of trajectories starting from randomly picked initial conditions, and we save data with the time-stepping δ. Numerical tests show that the invariant densities and the correlation functions vary little when the data are generated from different initial conditions.

We then infer NAR models for the first K Fourier modes with a time step δ. We will consider two values for K (recall that K_0 is the number of Fourier modes in the stochastic force)

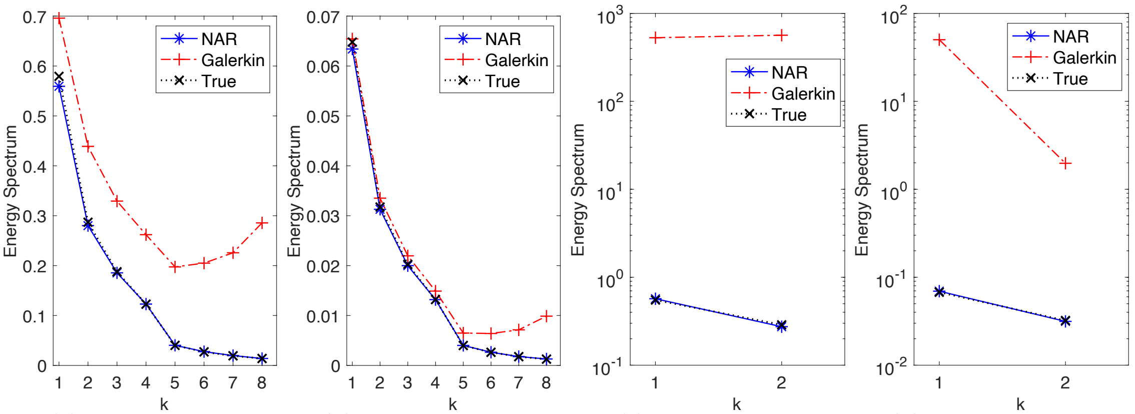
- $K = 8 > K_0 = 4$. In this case, $Qf = 0$, i.e., the stochastic force does not act on the unresolved Fourier modes w in (2.5), so w is a deterministic functional of the history of the resolved Fourier modes. In view of (3.3b), the reduced model mainly quantifies this deterministic map. We call this case "reduction of the deterministic response" and present the results in Section 4.3.
- $K = 2 < K_0$. In this case, $Qf \neq 0$, and w in (2.5) depend on the unobserved Fourier modes of the stochastic force. Thus, the reduced model has to quantifies the effects of the unresolved Fourier modes of both the solution and the stochastic force. We call this case "reduction involving unresolved stochastic force" and present the results in Section 4.4.

In either case, we explore the maximal time step that NAR models can reach by testing time steps $\delta = dt \times \{5, 10, 20, 30, 40, 50, 80, 160\}$.

We summarize the configurations and notations in Table 2.

4.2. Model selection and memory length

We demonstrate model selection and the effect of memory length for reduced models with time step $\delta = 5dt$. We aim to select a universal parametric form of the NAR model for different setting of



(a) $K = 8,\ \sigma = 1$ (b) $K = 8,\ \sigma = 0.2$ (c) $K = 2,\ \sigma = 1$ (d) $K = 2,\ \sigma = 0.2$

Figure 2. Energy spectrum of NAR models with $p = 1$ and the K-mode Galerkin systems in four settings of (K, σ). The time step is $\delta = 5dt$ for the NAR models and is dt for the Galerkin models. The NAR models accurately reproduce the true energy spectrum in all settings.

(K, σ), where $K \in \{8, 2\}$ is the number of Fourier modes in the NAR model and $\sigma \in \{1, 0.2\}$ is the standard deviation of the full model's stochastic force. Such a parametric form will be used later for the exploration of maximal time reduction by NAR models in the next sections.

We select the model according to Section 3.4: for each pair (K, σ), we test a pool of NAR models and select the *simplest* model that best reproduces the statistics and has consistent estimators. The statistics are computed along a long trajectory of $T = 2000$ time units. We say that an NAR is numerically unstable if it blows up (e.g. $|u^n|$ exceeding 10^5) before reaching $T = 2000$ time units.

We estimate the coefficients in (3.6b) for a few time lag p's. Numerical tests show that the normal matrix in regression is almost singular either when the stochastic force f_k^{n-j} presents or when the lag for u_k^{n-j} or $R^\delta(u^{n-j})$ is bigger than two. Thus, for simplicity, we remove them and set:

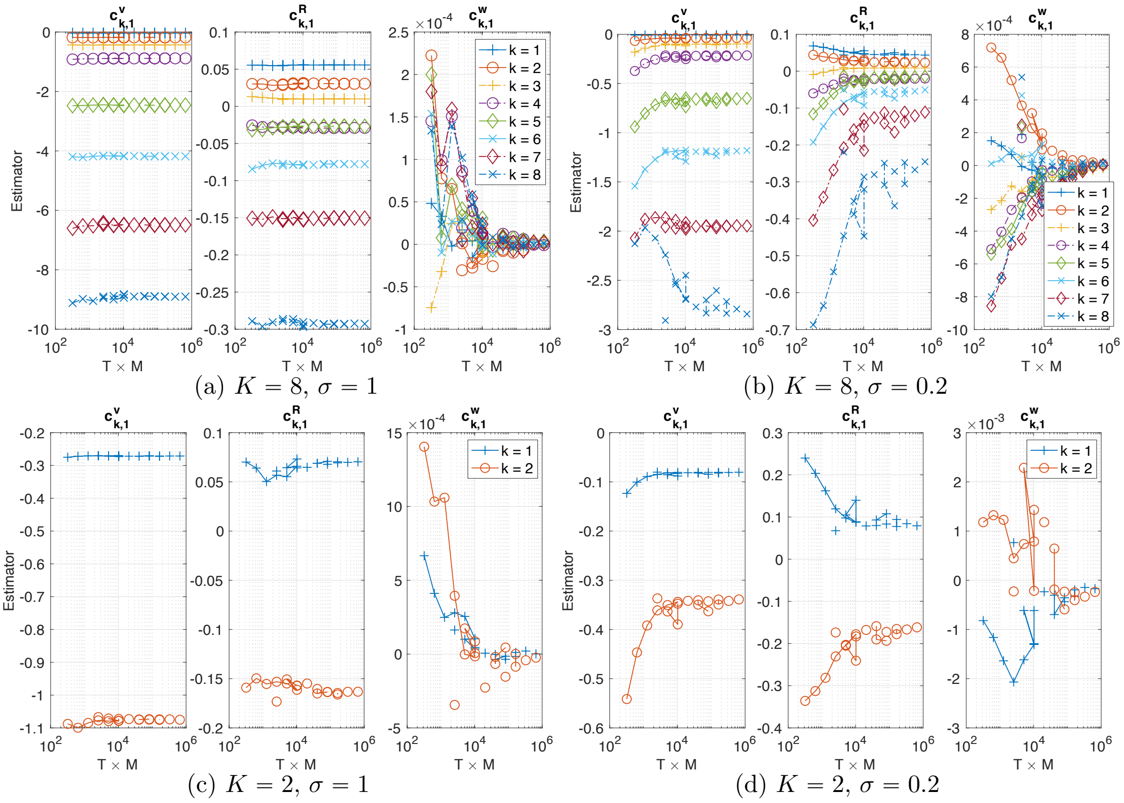
$$\begin{aligned} \text{Removed: } & c^f_{k,j} = 0 \text{ for all } k, j, \quad \text{and } c^v_{k,j} = c^R_{k,j} = 0 \text{ for all } 1 < j \leq p, \\ \text{To be estimated: } & c^v_{k,1},\ c^R_{k,1},\ c^w_{k,j}, 1 \leq j \leq p. \end{aligned} \tag{4.1}$$

That is, in (3.6b), the terms u_k^{n-j} and $R^\delta(u^{n-j})$ have a time lag 1, the stochastic force term f_k^{n-j} is removed, and only the high-order (the fourth) term has a time lag p. The memory length is $p\delta$.

Memory length. To select a memory length, we test NAR models with time lags $p \in \{1, 5, 10, 20\}$ and consider their reproduction of the energy spectrum in (3.12). Figure 1 shows the relative error in energy spectrum of these NAR models. It shows that as p increases: (1) when the scale of the stochastic force is large ($\sigma = 1$), the error oscillates without a clear pattern; (2) when $\sigma = 0.2$, the error first decreases and then increases. Thus, a longer memory does not necessarily lead to a better reduced model when the stochastic force dominates the dynamics; but when deterministic flow dominates the dynamics, a proper memory can be helpful.

In all the four settings, the simplest NAR models with $p = 1$ can consistently reproduce the energy spectrum with relative errors within 5%. Remarkably, the accuracy remains when the true energy spectrum is at the scale of 10^{-2} for the modes with $k = 7, 8$ in Figure 2(a-b) and $k = 2$ in Figure 2(d). Figure 2 also shows that the truncated K-mode Galerkin systems can not reproduce the true energy spectrum in any of these settings, with upward tails due the lack of fast energy dissipation from the high modes. Thus, the NAR model has introduced additional energy dissipation through Φ^n.

Consistency of estimators. The estimator of the NAR models tends to converge as data size increases. Figure 3 shows that the estimated coefficients of NAR with $p = 1$ from data consisting of M trajectories, each with length T, where $M \in \{2, 8, 32, 128, 512\}$ and $T \in \{40, 80, 160, 320, 640, 1280\}$. As $T \times M$



(a) $K = 8$, $\sigma = 1$

(b) $K = 8$, $\sigma = 0.2$

(c) $K = 2$, $\sigma = 1$

(d) $K = 2$, $\sigma = 0.2$

Figure 3. Estimated coefficients $(c^v_{k,1}, c^R_{k,1}, c^w_{k,j})$ in NAR models with $p = 1$ and $\delta = 5dt$ in four settings of (K, σ). The estimators tend to converge fast as the trajectory length T and number M increase: note that the coefficients $c^w_{k,1}$ are at the scale of 10^{-4} or 10^{-3}.

increases, all the estimators tend to converge (note that the coefficients $c^w_{k,1}$ are at the scale of 10^{-4} or 10^{-3}). In particular, they converge faster when $\sigma = 1$ than when $\sigma = 0.2$: the estimators in (a)-(c) oscillate little after $T \times M > 10^3$, indicating that different trajectories lead to similar estimators, while the estimators (take $c^R_{K,1}$ for example) in (b)-(d) oscillate until $T \times M > 10^5$. This agrees with the fact that a larger stochastic force makes the system mix faster, so each trajectory provides more effective samples driving the estimator to converge faster.

Numerical tests also show that an NAR model can be numerically unstable while its coefficient estimator being consistent (i.e., tending to converge as above). Thus, consistency is not sufficient for the selection of an NAR model.

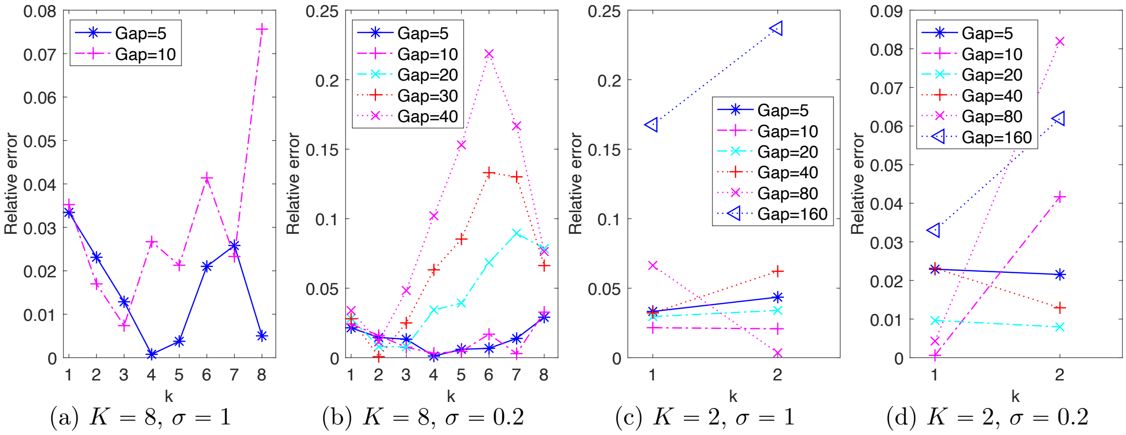
In our tests, sparse regression algorithms such as LASSO (see e.g.,[42]) or sequential thresholding (see e.g., [47,48]) have difficulty in proper thresholding, because the coefficient c^w of the high order terms are small and can vary in scales in different settings, but these high order terms are important for the NAR model.

Since the NAR models with $p = 1$ perform well in all the four settings and since they are the simplest, we use them in the next sections to explore the maximal time reduction.

4.3. Reduction of the deterministic response

We explore in this and the next section the maximal time step δ that the NAR models can reach. We consider only the simplest models with time lag $p = 1$.

We consider first the models with $K = 8$ Fourier modes. Since the stochastic force acts directly only on the first $K_0 = 4$ Fourier modes, the unresolved variable w in (3.1) is a deterministic functional of the path of the K modes, so is the truncation error $PB(v + w) - PB(v)$ in (3.3b). Thus, the NAR model mainly reduces the deterministic response of the resolved variables to the unresolved variables.



(a) $K = 8$, $\sigma = 1$ (b) $K = 8$, $\sigma = 0.2$ (c) $K = 2$, $\sigma = 1$ (d) $K = 2$, $\sigma = 0.2$

Figure 4. Relative error in energy spectrum reproduced by the NAR models with time steps $\delta = dt \times \text{Gap}$ for Gap $\in \{5, 10, 20, 30, 40, 50\}$ in four settings of (K, σ). All NAR models are with time lag $p = 1$. The missing Gap's in (a)-(b) lead to numerically unstable NAR models. Thus, the maximal δ's that an NAR model can reach are $\delta \in [0.01, 0.02)$ and $\delta \in [0.04, 0.05)$ for (a) and (b) respectively, and $\delta \geq 0.16$ for (c)-(d).

.

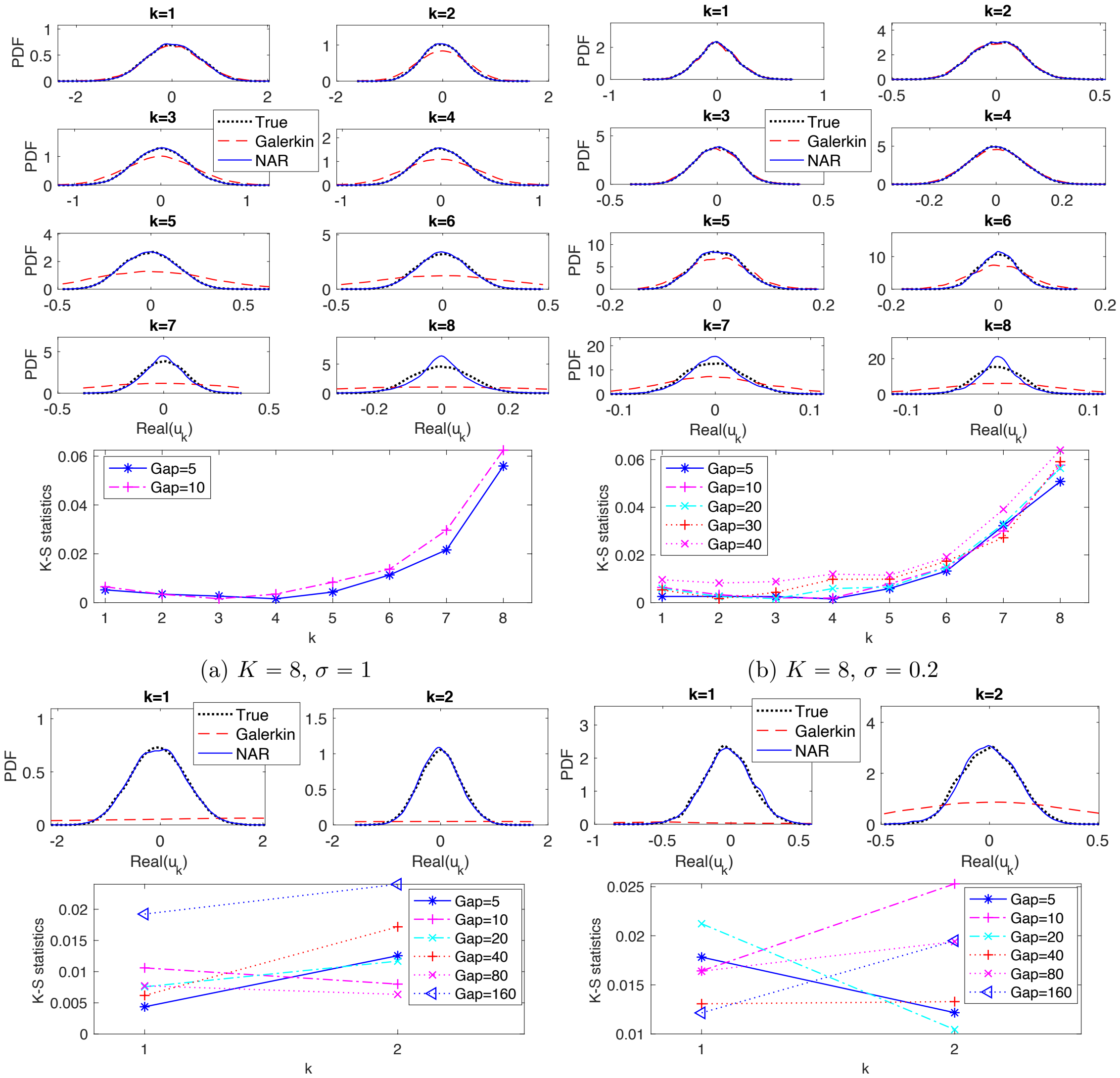
In particular, the term Φ^n in the NAR model (3.6a) optimally approximates this deterministic response on the function space linearly spanned by the terms in (3.6b).

We consider time steps $\delta = dt \times \text{Gap}$ with Gap $\in \{5, 10, 20, 30, 40, 50\}$. For each δ, we first estimate the coefficients $(c^v_{k,1}, c^R_{k,1}, c^w_{k,1})$ of the NAR model from the data with the same time step. We then validate the estimated NAR model by its statistics.

Numerical tests show that the NAR models with Gap ≥ 20 are numerically unstable for the setting $(K = 8, \sigma = 1)$, and the number is Gap $= 50$ for the setting $(K = 8, \sigma = 0.2)$. Figure 4(a-b) show the relative error in energy spectrum reproduced by NAR models with those stable time steps. The relative errors increase as the Gap increases. Note that the relative errors for modes $k = 1, 2$ change little, but those of with $k \in \{3, 4, 5, 6\}$ increase significantly. In particular, note that in (b), the relative error at $k = 8$ are about 8% for Gap $\in \{20, 30, 40\}$, but the relative errors at $k \in \{3, 4, 5, 6\}$ increase sharply to form a peak at $k = 6$ when Gap $= 40$. We will discuss connections with CFL numbers in Section 4.5.

These NAR models reproduce the PDF's and ACF's relatively accurately. Figure 5 shows the marginal PDF's of the real parts of the modes. The top row shows the marginal PDF's for the NAR models with Gap $= 5$, in comparison with those of the full model and the Galerkin truncated system (solved with time step dt). For the modes with wave numbers $k \in \{1, 2, 3, 4\}$, the NAR model captures the shape and spread of the PDF's almost perfectly, improving those of the Galerkin truncated system. For the modes with $k \in \{5, 6, 7, 8\}$, the NAR model still performs well, significantly improving those of the truncated Galerkin system. The discrepancy between the PDF's gets larger as the wavenumber increases, because these modes are affected more by the unresolved modes. The bottom row shows that the Kolmogorov-Smirnov statistics (the maximal difference between the cumulative distribution functions) increases slightly as the Gap increases. Figure 6 shows the ACF's. The top row shows that both the NAR model (with Gap $= 5$) and the Galerkin system can reproduce the ACF's accurately. The bottom row shows that the relative error of the ACF, in $L^2([0, 3])$-norm, increases as Gap increases (particularly in the case $\sigma = 0.2$). Recall that the truncated Galerkin system produces PDF's with support much wider than the truth for the high modes (see Figure 5), and that R^δ becomes less accurate as δ increases. Thus, the terms u and $R^\delta(u)$ in the NAR model (3.6) preserves the temporal correlation, and the high order term helps to dissipative energy and preserve the invariant measure.

In summary, when $K = 8$, the maximal time steps are $\delta \in dt \times [10, 20) = [0.01, 0.02)$ and $\delta = dt \times [40, 50) = [0.04, 0.05)$ when $\sigma = 1$ and $\sigma = 0.2$, respectively, for NAR models with $p = 1$.



(a) $K = 8$, $\sigma = 1$

(b) $K = 8$, $\sigma = 0.2$

(c) $K = 2$, $\sigma = 1$

(d) $K = 2$, $\sigma = 0.2$

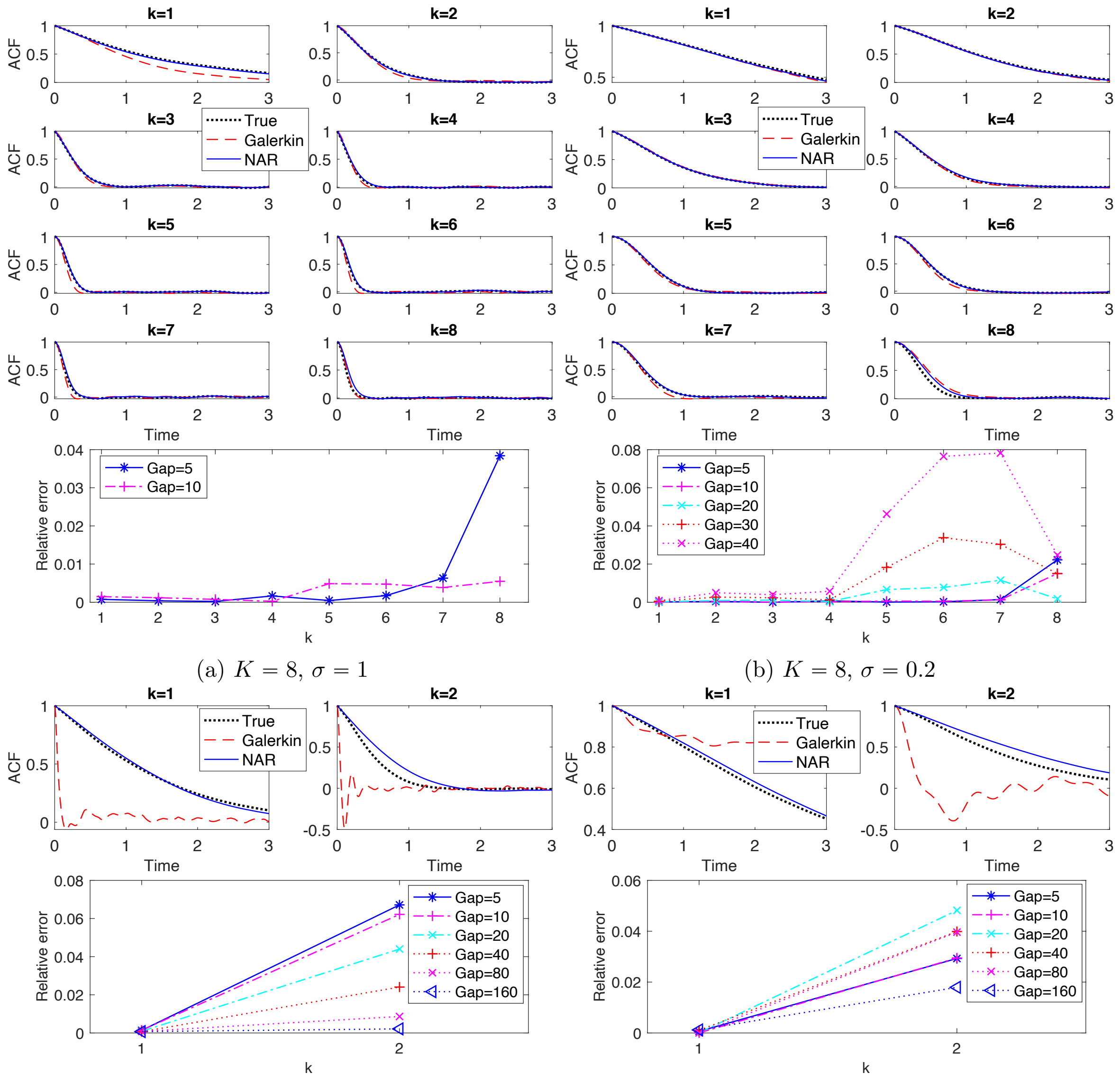
Figure 5. Marginal PDF's and K-S statistics. In each of (a)–(d), the top panels are plots of the empirical marginal PDF's of the real parts of the Fourier modes, from data (True), the K-mode Galerkin system (Galerklin) and the NAR models with $p = 1$ and $\delta = 5dt$. The bottom plots are the Kolmogorov-Smirnov statistics (the maximum difference between the cumulative distribution functions) of True and NAR models with time steps $\delta = dt \times \text{Gap}$.

All these NAR models can accurately reproduce the energy spectrum, the invariant measure and the temporal autocorrelation.

4.4. Reduction involving unresolved stochastic force

We consider next NAR models with $K = 2$. In this case, the unresolved variable w in (3.1) is a functional of both the path of the K modes and the unresolved stochastic force. Thus, in view of (3.3b) and (3.5)–(3.6), the NAR model quantifies the response of the K-modes to both the unresolved Fourier modes and the unresolved stochastic force.

Note first that $K = 2$ is too small for the K-mode Galerkin system to meaningfully reproduce any of the statistical or dynamical properties, see Figure 2(c)-(d) for the energy spectrum, Figure 5(c)-(d) for the marginal PDF's and Figure 6(c)-(d) for the ACF's. On the contrary, the NAR models with $\delta = 5dt$, whose term R^δ comes from the K-mode Galerkin, reproduce these statistics accurately. Remarkably,



(a) $K = 8,\ \sigma = 1$

(b) $K = 8,\ \sigma = 0.2$

(c) $K = 2,\ \sigma = 1$

(d) $K = 2,\ \sigma = 0.2$

Figure 6. ACF (auto correlation functions) when $K = 8$.

the NAR models remain accurate even when the time step is as large as $\delta = 80dt$, with the K-S statistics being less than 0.025 in Figure 5(c)-(d), and with the relative error in ACF's less than 6% in Figure 6(c)-(d).

To explore the maximal time step that NAR models can reach, we consider time steps $\delta = dt \times \text{Gap}$ with Gap $\in \{5, 10, 20, 40, 80, 160\}$. Numerical tests show that the NAR models are numerically stable for all of them in both settings of $\sigma = 1$ and $\sigma = 0.2$. Figure 4(c)-(d) show the relative error in energy spectrum reproduced by NAR models with these time steps. The relative error first decrease and then increase as Gap increases, reaching the lowest when Gap $= 10$ and Gap $= 20$ for the settings $\sigma = 1$ and $\sigma = 0.2$, respectively. In particular, all of these relative errors remain less than 9% except when Gap $= 160$ in the setting $\sigma = 1$.

In summary, when $K = 2$, NAR models can tolerate large time steps. The maximal time steps are at least $\delta = dt \times 80 = 0.08$ and $\delta dt \times 160 = 0.16$ when $\sigma = 1$ and $\sigma = 0.2$, respectively, for the NAR models to reproduce the energy spectrum with relative error less than 9%.

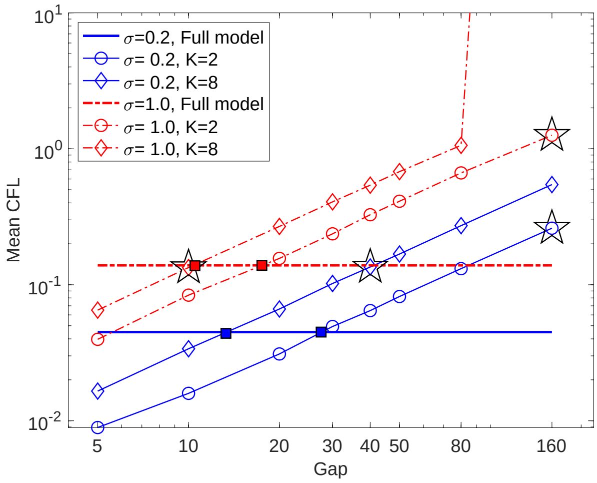


Figure 7. The mean CFL numbers of the full models and the K-mode Galerkin systems. The mean CFL number is computed along a trajectory with 10^5 steps. The time step is $dt = 0.001$ for the full model, and is $\delta = dt \times \text{Gap}$ for the K-mode Galerkin system. When $(\sigma = 1, K = 8)$, the K-mode Galerkin system blows up after Gap > 80, so its CFL number is missing afterwards. The stars are the maximal Gap such that the NAR model is stable. The squares are where the full model's mean CFL numbers agree with those of the K-mode Galerkin systems. The relative errors in energy spectrum in Figure 4(c)-(d) are the smallest when the Gap's are the closest to these squares.

4.5. Discussion on space-time reduction

Since model reduction aims for space-time reduction, it is natural to consider the maximal reduction in space-time, in other words, the minimum "spatial" dimension K and maximum time step $\delta = dt \times \text{Gap}$. We have the following observations from the previous sections:

1. Space dimension reduction, memory length of the reduced model and the stochastic force are closely related. As suggested by the discrete Mori-Zwanzig formalism for random dynamics (see e.g.,[7]), space dimension reduction would lead to non-Markovian closure models. Figure 1 suggests that a proper medium length of the memory leads to best NAR model. It also suggests that the scale of the white in time stochastic force can affect the memory length, and a larger scale of stochastic force leads to shorter memory. We leave it as future work to investigate the relations between memory length, (colored or white in time) stochastic force, and energy dissipation.
2. Maximal time step depends on the space dimension and the scale of the stochastic force, mainly limited by the stability of the nonlinear reduced model. Figure 4 shows that the maximum time step when $K = 2$ is at least $\delta = dt \times \text{Gap}$ with Gap $= 160$, much larger than those of the case of $K = 8$. It also shows that as the scale of stochastic force increases from $\sigma = 0.2$ to $\sigma 1$, the NAR models' maximal time step decreases (because the NAR models either become unstable or have larger errors in energy spectrum). It is noteworthy to mention that these maximal time step of NAR models are smaller than those that the K-mode Galerkin system can tolerate. Figure 7 shows that the K-mode Galerkin system can be stable for time step much larger than those of the NAR models: the maximal time step for the K-mode Galerkin system is when the mean CFL number (which increases linearly) reaches 1, but the maximal time step for the NAR models to be stable is smaller. For example, in the setting $(K = 8, \sigma = 0.2)$, the maximal time gap for the Galerkin system is Gap $= 80$ (the end of the red diamond line), but the maximal time gap for the NAR model is about Gap $= 10$. The increased numerical instability of the NAR model is likely due to the nonlinear terms Φ^n, which are important for the NAR model to preserve energy dissipation and the energy spectrum (see Figure 2 and the coefficients in Figure 3).

Beyond maximal reduction, an intriguing question arises: when does the reduced model perform the best (i.e., the least relative error in energy spectrum)? We call it *optimality of space-time reduction*. It is more interesting and relevant to model reduction than maximal reduction in space-time, because one may achieve a large time step or a small space dimension at the price of a large error in the NAR model, as we have seen in Figure 4. We note that the relative errors in energy spectrum in Figure 4(c)-(d) are the smallest when the Gap's are the closest to the squares in Figure 7, where the full model's mean CFL numbers agree with those of the K-mode Galerkin system. We conjecture that optimal space-time reduction can be achieved by an NAR model when the K-mode Galerkin system preserves the CFL number of the full model.

5. Conclusion

We consider data-driven model reduction for stochastic Burgers equations, casting it as a statistical learning problem on approximating the flow map of low-wavenumber Fourier modes. We derive a class of efficient parametric reduced closure models, based on representing the high modes as functionals of the resolved variables' trajectory. The reduced models are nonlinear autoregression (NAR) time series models, with coefficients estimated from data by least squares. In various settings, the NAR models can accurately reproduce the statistics such as the energy spectrum, the invariant densities, and the autocorrelations.

Using the simplest NAR model, we investigate the *maximal space-time reduction* in four settings: reduction of deterministic responses ($K > K_0$) v.s. reduction involving unresolved stochastic force ($K < K_0$), and small v.s. large scales of stochastic force (with $\sigma = 0.2$ and $\sigma = 1$), where K_0 is the number of Fourier modes of the white-in-time stochastic force, and σ is the scale of the force. Reduction in space dimension is unlimited, and NAR models with $K = 2$ Fourier modes can reproduce the energy spectrum with relative errors less than 5%. The time reduction is another story. Maximal time reduction depends on both the dimension reduction and the stochastic force's scale, as they affect the stability of the NAR model. The NAR model's stability limits the maximal time step to be smaller than those of the K-mode Galerkin system. Numerical tests indicate that the NAR models achieve the minimal relative error at the time step where the K-mode Galerkin system's mean CFL number agrees with the full model's. This is a potential criterion for *optimal space-time reduction*.

The simplicity of our NAR model structure opens various fronts for further understanding of data-driven model reduction. Future directions include: (1) studying the connection between optimal space-time reduction, the CFL number, and quantification of the accuracy of reduced models; (2) investigating the relation between memory length, dimension reduction, the stochastic force, and the energy dissipation of the system; (3) developing post-processing techniques to predict the shocks using the reduced models.

Funding: This research was funded by NSF-1913243 and NSF-1821211.

Acknowledgments: FL is grateful for Alexandre Chorin for introducing this problem. This study is part of our joint project on renormalization group methods. FL would like to thank Kevin Lin, Panos Stinis, John Harlim, Xiantao Li, Mauro Maggioni and Felix Ye for helpful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ETDRK4	exponential time differencing fourth order Rouge–Kutta method
CFL number	Courant–Friedrichs–Lewy number
NAR	nonlinear autoregression
PDF	probability density function
ACF	autocorrelation function

References

1. Stinis, P. Mori-Zwanzig Reduced Models for Uncertainty Quantification II: Initial Condition Uncertainty. *arXiv:1212.6360 [math]* **2012**, [arXiv:math/1212.6360].
2. Li, Z.; Bian, X.; Li, X.; Karniadakis, G.E. Incorporation of Memory Effects in Coarse-Grained Modeling via the Mori-Zwanzig Formalism. *J. Chem. Phys.* **2015**, *143*, 243128. Eng3, doi:10.1063/1.4935490.
3. Lu, F.; Tu, X.; Chorin, A.J. Accounting for Model Error from Unresolved Scales in Ensemble Kalman Filters by Stochastic Parameterization. *Mon. Wea. Rev.* **2017**, *145*, 3709–3723.
4. Lu, F.; Weitzel, N.; Monahan, A. Joint state-parameter estimation of a nonlinear SPDE model from sparse noisy data. *preprint* **2019**.
5. Zwanzig, R. *Nonequilibrium Statistical Mechanics*; Oxford University Press: Oxford ; New York, 2001.
6. Chorin, A.J.; Hald, O.H. *Stochastic Tools in Mathematics and Science*, 3rd ed.; Springer, New York, NY, 2013.
7. Lin, K.K.; Lu, F. Data-driven model reduction, Wiener projections, and the Mori-Zwanzig formalism. *arXiv preprint arXiv:1908.07725* **2019**.
8. Kondrashov, D.; Chekroun, M.D.; Ghil, M. Data-Driven Non-Markovian Closure Models. *Physica D* **2015**, *297*, 33–55. doi:10.1016/j.physd.2014.12.005.
9. Harlim, J.; Li, X. Parametric Reduced Models for the Nonlinear Schrödinger Equation. *Physical Review E* **2015**, *91*. doi:10.1103/PhysRevE.91.053306.
10. Lei, H.; Baker, N.A.; Li, X. Data-Driven Parameterization of the Generalized Langevin Equation. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 14183–14188. doi:10.1073/pnas.1609587113.
11. Xie, X.; Mohebujjaman, M.; Rebholz, L.G.; Iliescu, T. Data-Driven Filtered Reduced Order Modeling of Fluid Flows. *SIAM J. Sci. Comput.* **2018**, *40*, B834–B857. doi:10.1137/17M1145136.
12. Chekroun, M.D.; Kondrashov, D. Data-Adaptive Harmonic Spectra and Multilayer Stuart-Landau Models. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **2017**, *27*, 093110, [1706.04275]. doi:10.1063/1.4989400.
13. Chorin, A.J.; Lu, F. Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 9804–9809.
14. Lu, F.; Lin, K.K.; Chorin, A.J. Data-based stochastic model reduction for the Kuramoto–Sivashinsky equation. *Physica D* **2017**, *340*, 46–57.
15. Pathak, J.; Hunt, B.; Girvan, M.; Lu, Z.; Ott, E. Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. *Phys. Rev. Lett.* **2018**, *120*. doi:10.1103/PhysRevLett.120.024102.
16. Ma, C.; Wang, J.; E, W. Model reduction with memory and the machine learning of dynamical systems. *arXiv:1808.04258* **2018**.
17. Raissi, M.; Perdikaris, P.; Karniadakis, G. Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. *J. Comput. Phys.* **2019**, *378*, 686–707. doi:10.1016/j.jcp.2018.10.045.
18. Parish, E.J.; Duraisamy, K. A Paradigm for Data-Driven Predictive Modeling Using Field Inversion and Machine Learning. *J. Comput. Phys.* **2016**, *305*, 758–774. Eng0, doi:10.1016/j.jcp.2015.11.012.
19. Duan, J.; Wei, W. *Effective dynamics of stochastic partial differential equations*; Elsevier, 2014.
20. Stinis, P. Renormalized Mori-Zwanzig-Reduced Models for Systems without Scale Separation. *Proc. Royal Soc. A* **2015**, *471*, 20140446–20140446. doi:10.1098/rspa.2014.0446.
21. Hudson, T.; Li, X.H. Coarse-Graining of Overdamped Langevin Dynamics via the Mori–Zwanzig Formalism. *Multiscale Modeling & Simulation* **2020**, *18*, 1113–1135.
22. Choi, Y.; Carlberg, K. Space–Time Least-Squares Petrov–Galerkin Projection for Nonlinear Model Reduction. *SIAM Journal on Scientific Computing* **2019**, *41*, A26–A58.
23. Jiang, S.W.; Harlim, J. Modeling of Missing Dynamical Systems: Deriving Parametric Models Using a Nonparametric Framework. *ArXiv190508082 Math Stat* **2020**, [arXiv:math, stat/1905.08082].
24. Marion, M.; Temam, R. Nonlinear Galerkin methods. *SIAM Journal on Numerical Analysis* **1989**, *26*, 1139–1157.
25. Jolly, M.S.; Kevrekidis, I.G.; Titi, E.S. Approximate inertial manifolds for the Kuramoto-Sivashinsky equation: analysis and computations. *Physica D* **1990**, *44*, 38–60.
26. Rosa, R. Approximate inertial manifolds of exponential order. *Discrete Contin. Dynam. Systems* **1995**, *3*, 421–448.

27. Novo, J.; Titi, E.S.; Wynne, S. Efficient methods using high accuracy approximate inertial manifolds. *Numerische Mathematik* **2001**, *87*, 523–554.
28. Zelik, S. Inertial manifolds and finite-dimensional reduction for dissipative PDEs. *P. Roy. Soc. Edinb. A* **2014**, *144*, 1245–1327.
29. E, W.; Khanin, K.; Mazel, A.; Sinai, Y.G. Invariant Measures for Burgers Equation with Stochastic Forcing. *Ann. Math.* **2000**, pp. 877–960.
30. Chorin, A.J. Averaging and Renormalization for the Korteveg-deVries-Burgers Equation. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 9674–9679. doi:10.1073/pnas.1334126100.
31. Chorin, A.J.; Hald, O.H. Viscosity-Dependent Inertial Spectra of the Burgers and Korteweg-deVries-Burgers Equations. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 3921–3923. doi:10.1073/pnas.0500335102.
32. Bec, J.; Khanin, K. Burgers Turbulence. *Physics Reports* **2007**, *447*, 1–66. doi:10.1016/j.physrep.2007.04.002.
33. Beck, M.; Wayne, C.E. Using Global Invariant Manifolds to Understand Metastability in the Burgers Equation With Small Viscosity. *SIAM J. Appl. Dyn. Syst.* **2009**, *8*, 1043–1065. doi:10.1137/08073651X.
34. Wang, Z.; Akhtar, I.; Borggaard, J.; Iliescu, T. Two-Level Discretizations of Nonlinear Closure Models for Proper Orthogonal Decomposition. *J. Comput. Phys.* **2011**, *230*, 126–146. Eng4, doi:10.1016/j.jcp.2010.09.015.
35. Dolaptchiev, S.; Achatz, U.; Timofeyev, I. Stochastic closure for local averages in the finite-difference discretization of the forced Burgers equation. *Theoretical and Computational Fluid Dynamics* **2013**, *27*, 297–317.
36. Da Prato, G. *An introduction to infinite-dimensional analysis*; Springer Science & Business Media, 2006.
37. Cox, S.M.; Matthews, P.C. Exponential time differencing for stiff systems. *J. Comput. Phys.* **2002**, *176*, 430–455.
38. Kassam, A.K.; Trefethen, L.N. Fourth-order time stepping for stiff PDEs. *SIAM J. Sci. Comput.* **2005**, *26*, 1214–1233.
39. Lu, F.; Lin, K.K.; Chorin, A.J. Comparison of continuous and discrete-time data-based modeling for hypoelliptic systems. *Comm. App. Math. Com. Sc.* **2016**, *11*, 187–216.
40. Fan, J.; Yao, Q. *Nonlinear Time Series: Nonparametric and Parametric Methods*; Springer, New York, NY, 2003.
41. Kutoyants, Y.A. *Statistical inference for ergodic diffusion processes*; Springer, 2004.
42. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **1996**, *58*, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.
43. Brockwell, P.; Davis, R. *Introduction to Time Series and Forecasting*; Springer, New York, NY, 2002.
44. Billings, S.A. *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatiotemporal Domains*; John Wiley and Sons, 2013.
45. Györfi, L.; Kohler, M.; Krzyzak, A.; Walk, H. *A distribution-free theory of nonparametric regression*; Springer Science & Business Media, 2006.
46. Lu, F.; Zhong, M.; Tang, S.; Maggioni, M. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 14424–14433.
47. She, Y. Thresholding-Based Iterative Selection Procedures for Model Selection and Shrinkage. *Electron. J. Statist.* **2009**, *3*, 384–415. doi:10.1214/08-EJS348.
48. Quade, M.; Abel, M.; Kutz, N.J.; Brunton, S.L. Sparse Identification of Nonlinear Dynamics for Rapid Model Recovery. *Chaos* **2018**, *28*, 063116. doi:10.1063/1.5027470.