SIAM/ASA J. Uncertainty Quantification
Vol. 10, No. 2, pp. 555–583

Generative Stochastic Modeling of Strongly Nonlinear Flows with Non-Gaussian Statistics*

Hassan Arbabi† and Themistoklis Sapsis†

Abstract. Strongly nonlinear flows, which commonly arise in geophysical and engineering turbulence, are characterized by persistent and intermittent energy transfer between various spatial and temporal scales. These systems are difficult to model and analyze due to combination of high dimensionality and uncertainty, and there has been much interest in obtaining reduced models, in the form of stochastic closures, which can replicate their non-Gaussian statistics in many dimensions. Here, we propose a data-driven framework to model stationary chaotic dynamical systems through nonlinear transformations and a set of decoupled stochastic differential equations (SDEs). Specifically, we use optimal transport to find a transformation from the distribution of time-series data to a multiplicative reference probability measure such as the standard normal distribution. Then we find the set of decoupled SDEs that admit the reference measure as the invariant measure, and also closely match the spectrum of the transformed data. As such, this framework represents the chaotic time series as the evolution of a stochastic system observed through the lens of a nonlinear map. We demonstrate the application of this framework in the Lorenz-96 system, a 10-dimensional model of high-Reynolds cavity flow, and reanalysis climate data. These examples show that SDE models generated by this framework can reproduce the non-Gaussian statistics of systems with moderate dimensions (e.g., 10 and more) and predict super-Gaussian tails that are not readily available from little training data. These findings suggest that this class of models provides an efficient hypothesis space for learning strongly nonlinear flows from small amounts of data.

Key words. measure-preserving chaos, Koopman continuous spectrum, extreme events, spectral proper orthogonal decomposition, mixed spectra, optimal transport

AMS subject classifications. 62G32, 76F20, 49Q22, 60G10, 34H10, 34L05

DOI. 10.1137/20M1359833

1. Data-driven modeling of dynamical systems.

Scientists and engineers are facing increasingly difficult challenges such as predicting and controlling the earth climate, understanding and emulating the human brain, and designing and maintaining social networks. All these challenges require modeling and analysis of systems that have a large number of degrees of freedom, possibly combined with a considerable amount of uncertainty in parameters. In doing so, the classical model-based state-space approach for dynamical systems analysis and control falls short due to the computational size of these problems and lack of an accurate model. As a result, recent decades have seen a great development in the area of model

* Received by the editors August 14, 2020; accepted for publication (in revised form) January 20, 2022; published electronically June 29, 2022.
https://doi.org/10.1137/20M1359833
Funding: This research was partially supported by ARO-MURI grant W911NF-17-1-0306. The work of the second author was also supported by an MIT Sea Grant through Doherty Associate Professorship and AIR Worldwide.

† Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (arbabiha@gmail.com, sapsis@mit.edu).

reduction and data-driven modeling with the focus of making these problems more amenable to computation and analysis.

One of the main challenges in modeling complex systems is centered around strongly nonlinear systems, i.e., chaotic systems with persistent and intermittent energy transfer between multiple scales. Climate dynamics and engineering turbulence give rise to many such problems and have been a major target of model reduction techniques. In particular, using stochastic reduced models has become increasingly popular for these problems because it allows for the representation of turbulent fluctuations, which arise from nonlinear interactions in many degrees of freedom, in the form of simple systems with random forcing. Stochastic models of climate are proposed to estimate climate predictability and response to external output [29, 41], explain the observed energy spectrum and transport rates in the atmosphere [23, 19], facilitate data assimilation [28], and predict weather patterns [11]. Similar models for analysis of turbulent flows have been suggested in [80, 16, 72, 91, 32]. The formal analysis in [51, 52, 49] has established the validity of stochastic closure models under certain assumptions while emphasizing the need to go beyond the commonly used linear models with additive noise. All the above efforts have shown the great utility and promise of stochastic modeling, but they heavily rely on expert intuition and the underlying model of the system (e.g., linearized Navier–Stokes equations), and therefore their application remains limited to well-understood systems and well-chosen variables.

Data-driven methods provide a shortcut for analysis and control of high-dimensional systems by allowing us to discover and exploit low-dimensional structures that may be masked by the governing equations, or to find alternative models with more computational tractability. A key enabler of data-driven analysis for dynamical systems in recent years has been the operator-theoretic formalism, and in particular the Koopman operator theory [36, 54], which describes the evolution of observables, for computation of geometric objects and linear predictors from trajectory data. This approach is accompanied by an arsenal of numerical algorithms that extract dynamical information from large data produced by simulations and experiments [75, 73, 89, 61, 66]. Due to interpretability and connection with the classical theory of dynamical systems, this methodology has become popular in many fields ranging from fluid dynamics [76] and power networks [68] to biological pattern extraction [8] and visual object recognition [22]. The Koopman operator theory is closely related to the Perron–Frobenius operator and Fokker–Planck equations [40, 17, 15]; however, those viewpoints have enjoyed less widespread use in data-driven applications. A central feature of the Koopman operator framework is to use data to find the canonical coordinates which decouple the system dynamics and allow independent and linear representation. This feature is also emphasized in manifold learning methods for data mining which are powered by (diffusion) operator realization and analysis [12, 77]. However, much of the current operator-theoretic framework relies on the discrete spectrum expansion, and the modeling framework for stationary chaotic systems, where the associated operators possess continuous spectrum, is still in a nascent phase [38, 25, 13].

Statistical and machine learning methods offer another large and open-ended avenue for data-driven modeling. Variations of neural networks, including autoencoders, long short-term memory networks, and reservoir computers, are proposed for various tasks like discovering representations of complex systems, closure modeling, and short-time prediction of chaotic systems [e.g., 71, 70, 64, 85, 46, 10]. A big theme of many efforts in this area is to explicitly

encode available physical and mathematical information in the structure of learning to reduce the amount of training data and adhere to known constraints [86, 84, 69]. Furthermore the approximation of operator-theoretic objects from data is naturally connected to the regression problem, and there is a growing body of works that uses machine learning techniques for approximation of the Koopman or Perron–Frobenius operator [31, 93, 81, 90, 43, 46, 62]. All the above studies show the promise of machine learning in modeling dynamical systems, but the methods in this category often accompany important caveats such as lack of interpretability and nonrobust optimization processes.

In this work, we present a data-driven framework for *generative modeling* of strongly nonlinear flows. In contrast to the most data-driven modeling frameworks, which focus on short-time prediction, our goal is to discover models from data that emulate the statistics of a strongly nonlinear flow in the appropriate sense. This problem, which involves matching joint distributions of stochastic processes at arbitrary lags, is computationally prohibitive for strongly nonlinear flows. However, inspired by the spectral theory of dynamical systems [54, 38], we reformulate this problem as finding a model that replicates the invariant measure as well as the power spectral density of the observed time-series data. The cornerstone of our approach is using optimal transport of probabilities [63, 83] to map the data distribution to a multiplicative reference measure and then model each dimension separately using a simple stochastic differential equation (SDE) that mimics the spectrum of the time series. Starting from time-series measurements of observables on a dynamical system, our framework yields a set of randomly forced linear oscillators combined with a nonlinear observation map that produces the same statistics and dynamics as the given time-series data.

Our framework enables a stronger data-driven approach compared to the aforementioned works on stochastic modeling: we start with a general space of stochastic models, i.e., linear stochastic oscillators pulled back under polynomial maps, and then select the model that produces closest spectra to the observed data. This offers an advantage over the previous works on data-driven modeling of strongly nonlinear systems [39, 50] by treating nonlinearities as a feature of the observation map and not the underlying vector field. As a result, we can use dynamic models with guaranteed stable invariant measures as opposed to quadratic models in [39] and avoid using latent variables as opposed to [50]. Despite the generality of our framework, our examples show that this choice of hypothesis space is narrow enough to provide a data-efficient and computationally tractable pipeline for learning statistically accurate models of strongly nonlinear flows with tens of dimensions. This is partially due to formulating a major part of the challenge as a probability transport problem in large dimensions, which has enjoyed great computational progress in recent years.

We use our framework to compute stochastic models for the Lorenz-96 system, high-Reynolds flow in a 2D cavity, and 6-hourly time series of velocity and temperature reanalysis data from the earth atmosphere. In case of cavity flow, we use the spectral proper orthogonal decomposition (SPOD) [82] to obtain the 10-dimensional model of the flow. A theoretical contribution of our work is showing a novel connection between this decomposition and spectral decomposition of the Koopman operator for systems with continuous spectrum. Application of our framework to the above examples generates models that closely match the non-Gaussian features of data such as skewness and heavy tails in large dimensions from relatively little data. In the case of cavity flow, the SDE models not only produce the statistics of modal coordinates

which are used to train the model, but they also lead to accurate approximation of high-order statistics for pointwise flow measurements. The transport-based framework also allows modeling systems that possess heavy tails in their invariant measures. By utilizing this feature in the case of the climate data, we are able to extrapolate the heavy tails of the chaotic reanalysis time series. This contribution is especially important since it allows probabilistic characterization of extreme events from short-time observations, which is an outstanding challenge in modeling strongly nonlinear flows [44].

2. Stochastic generative modeling of strongly nonlinear flows.

2.1. Problem setup.

Consider a dynamical system given as

$$\begin{aligned} \dot{x} &= f(x), \\ y &= g(x) \end{aligned} \tag{1}$$

with the state variable x in the state space S and the output y in $\mathbb{R}^N$. We use $F^t : S \to S$ to denote the flow map over time t, that is, the mapping that takes the state at time t_0 to the state at time $t_0 + t$. Assume that the trajectories in an open subset of S converge to an attractor Ω which supports an invariant and ergodic probability measure denoted by μ, that is, $\mu(F^{-t}(B)) = \mu(B)$ for all measurable $B \subset \Omega$, $\mu(\Omega) = 1$, and $\mu(\Omega - C) = 0$ or 1 if C is an invariant subset of Ω.

We assume that instead of x, we can only measure y, which is a random variable with distribution $\nu = g_{\#}\mu$, that is, the pushforward of μ under the (measurable) observation map g, defined as

$$\nu(g < a) = \mu\Big(g^{-1}((-\infty, a))\Big). \tag{2}$$

We let U^t be the Koopman operator of this system which describes the evolution of random variables with the dynamics, that is, $U^t g := g \circ F^t$ [36, 54]. Moreover, we assume the observables we study are square integrable with respect to the invariant measure μ so that we can utilize the spectral results developed for the Koopman operator on such Hilbert spaces. The collection of random variables $\{U^t g\}_{t\in\mathbb{R}}$ is a stationary stochastic process, and the time series of y is a realization of this process (see, e.g., [20]).

Given the time series of y, our goal is to construct a dynamical system that generates a statistically similar stochastic process. A *perfect generative model* for this system will generate a stochastic process that has the same joint distributions as this process for an arbitrary combination of lags. To be more precise, let

$$P_{\boldsymbol{\tau},\mathbf{a}}[Z^t] = P(Z^0 > \mathbf{a}_0, Z^{\tau_1} > \mathbf{a}_1, \ldots, Z^{\tau_n} > \mathbf{a}_n) \tag{3}$$

be the joint distribution of the stochastic process $\{Z^t\}_{t\in\mathbb{R}}$, with $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n) \in \mathbb{R}^n$, $\mathbf{a}_i \in \mathbb{R}^N$ for $i = 0, \ldots, n$, and $Z^{\tau_i} > \mathbf{a}_i$ denoting an elementwise inequality between the vectors $\mathbf{a}_i$ and Z^{τ_i}. Then $\{Z^t\}_{t\in\mathbb{R}}$ is a perfect model for our time series if

$$P_{\boldsymbol{\tau},\mathbf{a}}[U^t g] = P_{\boldsymbol{\tau},\mathbf{a}}[Z^t] \tag{4}$$

for all choices of n, $\boldsymbol{\tau}$, and $\mathbf{a}$. Searching for a good model of time series using this characterization is computationally formidable since it requires presentation and matching of many—and in principle an infinite number of—joint distributions.

In this work, we appeal to another characterization of stochastic processes which is motivated by spectral analysis of dynamical systems. Recall that the power spectral density (PSD) of a stationary process is given by the Fourier transform of its covariance; e.g., for the above process we have

$$S_g(\omega) = \int_{-\infty}^{\infty} e^{-i\omega\tau} \mathrm{Cov}(g, U^\tau g) d\tau. \tag{5}$$

It turns out that given the knowledge of the underlying attractor Ω and the invariant measure μ, the PSD of observable g contains all the information about the dynamics of the system that can be gleaned from observing g. More precisely, assume g is square integrable with respect to μ and then define

$$\mathcal{H}_g = \overline{\mathrm{span}\{g, U^\tau g, U^{-\tau} g, U^{2\tau} g, U^{-2\tau} g, \ldots\}}, \tag{6}$$

where τ is the sampling interval. In other words, $\mathcal{H}_g$ is the closure of random variables that can be represented as a linear combination of g and its history. Then with the knowledge of the PSD of g, we can uniquely determine the Koopman operator U restricted to $\mathcal{H}_g$ [38, Proposition 1], implying that we can predict the time evolution of every random variable in $\mathcal{H}_g$. In the case that g is informative enough (or more precisely, $*$-cyclic under the action of U^t [47]), $\mathcal{H}_g$ contains all square-integrable random variables including the state variables.

Guided by the above observation, we reformulate the generative modeling for a stationary stochastic processes as follows: we seek a model that replicates the same first-order distribution (i.e., ν, the invariant measure observed through the lens of g), and the same PSD as the time series of g. We are interested in finding such models for strongly nonlinear flows with many dimensions. For computational tractability in these problems, we propose a hypothesis space which consists of linear stochastic oscillators observed under the inverse of (invertible) multivariate polynomial maps. To find the best model within this space, we first use optimal transport to discover the appropriate observation map that generates the distribution of the target stochastic process. Then we optimize the parameters of the oscillator to best match the PSD of the target stochastic process. Our results and analysis show that this choice of hypothesis space and learning process is data efficient and computationally economic for learning strongly nonlinear flows.

2.2. Modeling with optimal transport and spectral matching. The first step in our framework is to find the mapping $T : \mathbb{R}^N \to \mathbb{R}^N$ that takes the random variable $y \sim \nu$ to another random variable, denoted by q, which has a standard normal distribution, that is,

$$q = T(y) \sim \pi, \tag{7}$$

where π is the standard normal distribution. In other words, we seek the change of variables that makes the attractor look like a normalized Gaussian distribution in each direction. We use the theory of *optimal transport* to find this change of variable. This theory studies the choice of maps which minimize some notion of cost for carrying one probability measure to another

[83, 67]. The existence of such a map generally depends on the regularity of the measure ν. We assume that ν is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^N$ and possesses finite second-order moments, which guarantees the existence of the transport map [83]. For numerical approximation of this map, we use the computational approach developed in [63, 53] (also see [21, 79, 60]). This framework leverages the structure of a specific class of optimal transport maps, known as Knothe–Rosenblatt rearrangement, which leads to advantageous computational properties. The map is assumed to be a lower-triangular, differentiable, and monotone function given as

$$T(y_1, y_2, \dots, y_N) = \begin{bmatrix} T_1(y_1) \\ T_2(y_1, y_2) \\ \vdots \\ T_N(y_1, y_2, \dots, y_N) \end{bmatrix}, \tag{8}$$

where each component is approximated with a multivariate polynomial of a prescribed degree. To ensure the monotonicity, and hence invertibility, of T, we use the integrated squared parameterization, i.e.,

$$T_i(y_1, y_2, \dots, y_i) = c(\mathbf{a}_c; y_1, \dots, y_{i-1}) + \int_0^{y_i} (h(\mathbf{a}_h; y_1, \dots, y_{i-1}, t))^2 dt \tag{9}$$

with

$$c = \mathbf{\Phi}_c(y_1, \dots, y_{i-1})\mathbf{a}_c, \quad h = \mathbf{\Phi}_e(y_1, \dots, y_{i-1})\mathbf{a}_h, \tag{10}$$

where basis functions in $\mathbf{\Phi}_c$ are multivariate Hermite polynomials, and $\mathbf{\Phi}_e$ are Hermite functions extended with constant basis functions [65, 5].

The parameters of T, i.e., expansion coefficients $\mathbf{a}_c$ and $\mathbf{a}_e$, are determined by maximum likelihood estimation: let $\tilde{T}$ be a candidate for T, and consider the pullback of π under this map,

$$\tilde{\nu}(y) = \pi(\tilde{T}(y))|\det \nabla \tilde{T}(y)|. \tag{11}$$

Then the map T is found by minimizing the Kullback–Leibler divergence between this pullback and the data distribution ν, i.e.,

$$T = \arg\min_{\tilde{T}} \mathbb{E}_\nu \left[\log \frac{\nu(y)}{\tilde{\nu}(y)} \right] = \arg\min_{\tilde{T}} \mathbb{E}_\nu \left[-\log \tilde{\nu}(y) \right], \tag{12}$$

where we dropped $\nu(y)$ in the second equality since it is independent of optimizer $\tilde{T}$. The expectation with respect to ν is approximated using the time average of data due to ergodicity. The key feature of this computational setup is that it leads to an optimization problem which can be solved for each T_i separately [53, 63], thereby allowing an efficient computation for large dimensions and long time series.

After we have found T, we fit a system of stochastic oscillators to the time series of the random variable $q = T(y)$. Since we chose π to be a multiplicative measure, we can do this fitting for each component of q independently. The independent modeling of each coordinate

is a key aspect of this work which enables computational treatment of high-dimensional systems, as opposed to direct discovery of a dynamical system, which would generate the same high-dimensional joint distribution as the data. We consider the systems of forced linear oscillators,

$$\ddot{q}_j + \beta_j \dot{q}_j + k_j q_j = \sqrt{2D_j}\dot{W}_j, \quad \beta_j, k_j, D_j > 0, \quad j = 1, \ldots, N, \tag{13}$$

where W_j's denote mutually independent (generalized) derivatives of the Wiener process (see, e.g., [78]). Each oscillator admits a stationary density given by

$$\rho_j(q = w_1, \dot{q} = w_2) = c_j \exp\left\{ \frac{-\beta_j}{D_j} \left(\frac{w_2^2}{2} + k_j \frac{w_1^2}{2} \right) \right\}, \quad j = 1, \ldots, N, \tag{14}$$

with c_j being the normalization constant [78]. By setting $D_j = k_j \beta_j$, we make sure that the displacement of each oscillator has a standard normal distribution. Next, in order to make the model replicate the *dynamics* of the time series, we optimize the free parameters of each system to match the PSD of q. To be more precise, let S_j denote the PSD of the stochastic process $\{U^t q_j\}$. On the other hand, the PSD of response, q_j, in (13) is given as

$$\tilde{S}_j(\omega) = \frac{2D_j}{|k_j + i\beta_j \omega - \omega^2|^2}, \tag{15}$$

where ω denotes the frequency [78]. We find the pair (k_j, β_j) that minimizes the spectral difference

$$\Delta = \int_0^{\frac{\omega_s}{2}} |\tilde{S}_j(\omega) - S_j(\omega)| d\omega \tag{16}$$

with ω_s being the sampling frequency. We approximate $S_j(\omega)$ from the time series of $q_j = T_j(y)$ using the Welch method [87] and solve this 2D optimization problem using particle swarm optimization. Note that the spectrum of y_j will be affected by the spectrum of $q_1, \ldots, q_j$ due to the coupling of variables in T^{-1}; however, we have found that minimizing the error of q's spectra independently is an efficient way to minimize the error in the spectrum of y—see Figure 11 for a 10-dimensional example. When the time series are not long enough for a robust approximation of PSD, we recommend matching the autocorrelation function up to some finite time lag τ. As discussed in [38], the autocorrelation sequence corresponds to the moments of the spectral measure, and matching those moments in the limit of $\tau \to \infty$ leads to matching of the PSD. At the end, the data-driven stochastic model for the system in (1) consists of the stochastic oscillators in (13) with the observation map T^{-1}.

The crucial finding of our study is that this choice of hypothesis space, i.e., stochastic linear oscillators observed under the inverse of triangular polynomial maps of order two or three, provides an adjustable trade-off between learning capacity and data efficiency in learning models of modal dynamics in strongly nonlinear flows with non-Gaussian statistics. This approach is also advantageous since it is easily scalable to tens of dimensions; for example, the transport maps of the cavity flow example in subsection 3.2, involving 25000 data points in 10 dimensions, are computed in a few minutes on a plain desktop. The potential drawback

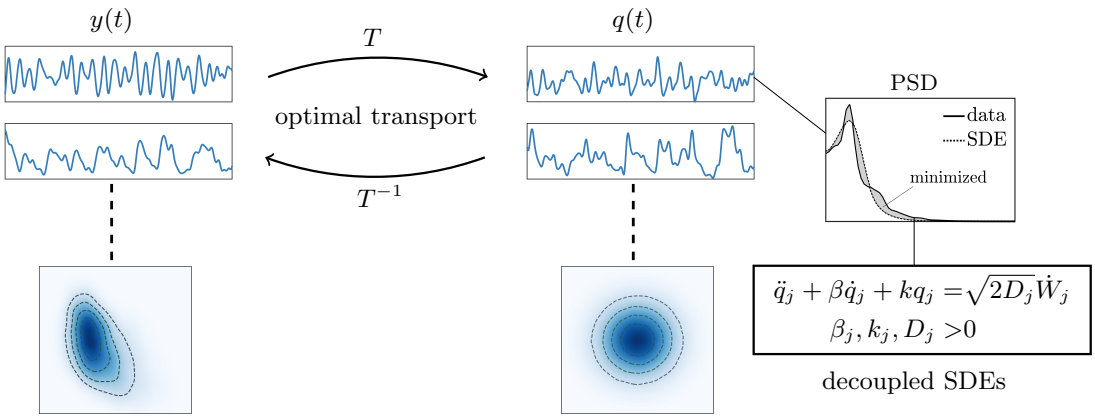


Figure 1. ***Framework for stochastic modeling.*** *A change of variable is found that transports the data distribution to standard normal distribution, and then a stochastic oscillator is built by minimizing the difference between the PSD of transported time series and the SDE model (the shaded area).*

is that our hypothesis space is relatively small (e.g., compared to deep neural networks) and there is no guarantee that it would be suitable for learning other classes of nonlinear systems that may arise in other disciplines. In principle, however, one could use a more sophisticated class of reference measures (e.g., bimodal distributions) and stochastic models (e.g., non-Markovian linear stochastic systems), or even a more inclusive space of transport maps, based on the complexity of the target system and availability of data.

Finally, we note that our framework for modeling, summarized in Figure 1, is connected to other methods of nonlinear systems identification; from the systems and signals perspective, our approach resembles the classical Wiener system identification which approximates nonlinear systems as a linear system with a nonlinear observation map [74, 6]. Our approach can also be considered as a data-driven discovery of a conjugacy [88] between the data coordinates and the space of linear stochastic oscillators. Our conjugacy map (the transport map) is particularly designed so that the obtained conjugate system has statistically independent coordinates. Our approach is similar in spirit to other data-driven methods for discovering conjugate systems with independent components of motion [56] (i.e., phases along each direction of the tori or, equivalently, principal Koopman eigenfunctions). Finally, our method provides a backward technique for fitting solutions of Fokker–Planck equations to data; direct construction of an SDE which gives rise to the data distribution requires solving a Fokker–Planck equation in large dimensions, which is often difficult or impossible, but in this framework the data is mapped to some distribution which is already a solution to a well-known set of models.

3. Results.

3.1. Lorenz-96 system. The Lorenz-96 system is a toy model of atmospheric turbulence and is commonly used in assessment of modeling, closure, and data assimilation schemes for strongly nonlinear systems [48]. This model describes a set of interacting states on a ring which evolve according to

$$\dot{x}_k = x_{k-1}(x_{k+1} - x_{k-2}) - x_k + F, \qquad k = 1, 2, \ldots, K, \tag{17}$$
$$x_0 = x_K, \quad x_{K+1} = x_1, \quad x_{-1} = x_{K-1}.$$

For the standard parameter values of $K = 40$ and $F = 8$, trajectories converge to a chaotic attractor. We assume that we can only observe the first component of the state while the trajectory is evolving on the attractor, i.e.,

$$y = g(\mathbf{x}) = x_1. \tag{18}$$

We record a time series of y with length of 1000 and sampling interval of 0.1. After applying our SDE modeling framework to this time series, we find the stochastic model to be

$$\ddot{q} + 4.73\dot{q} + 26.26q = 15.76\dot{W}, \tag{19}$$

where the state q is related to the observable y through the map

$$q = T(y) \approx 0.001y^3 - 0.010y^2 + 0.279y - 0.570. \tag{20}$$

The probability density function (PDF) and PSD of the observable $y = T^{-1}(q)$ are shown in Figure 2. Due to the inclusion of nonlinear terms in the transport, this model is capable of

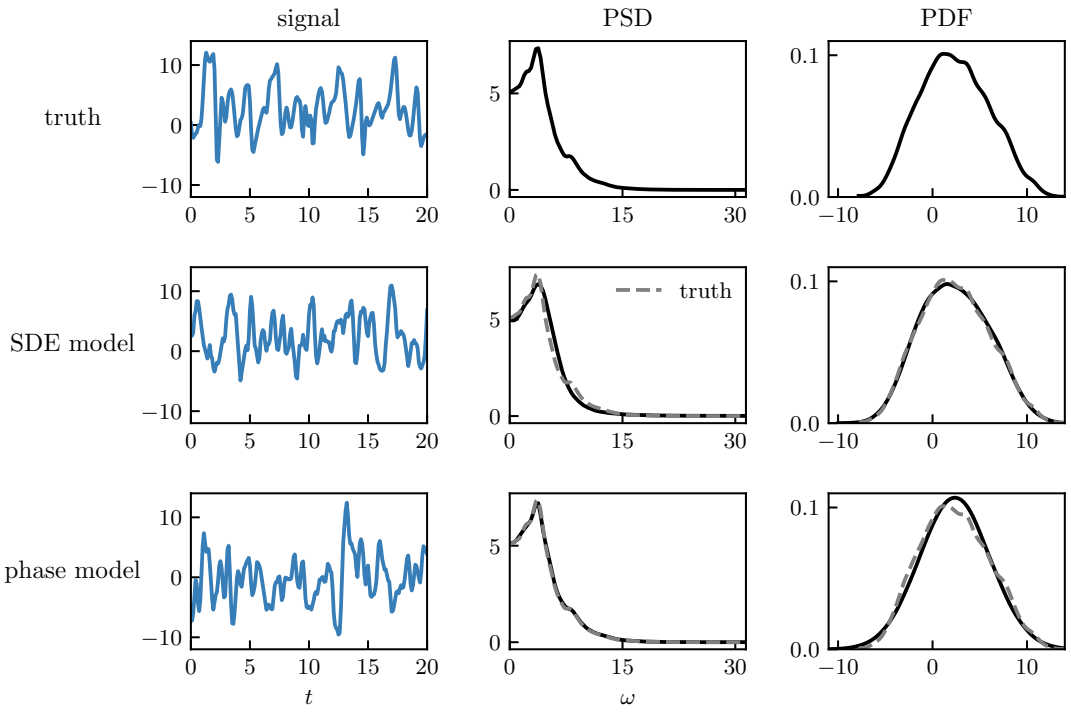


Figure 2. ***Data-driven modeling of Lorenz-96 system.*** *Top: truth observation on a state observable of chaotic Lorenz-96 system; middle: approximation via our stochastic modeling framework; bottom: random phase model approximation. The SDE model captures the skewness in the PDF and closely matches the spectrum. The quasi-periodic phase model, by design, matches the spectrum of data but systematically misses the skewness in the PDF. The PSDs are computed after removing the mean.*

capturing the skewness in the PDF of observable y. This is a crucial feature for models of turbulent systems since skewness gives rise to the nonlinear cascade of energy, and (linear) models with Gaussian distribution are incapable of describing this type of energy transfer between the system components.

We compare our model to a data-driven approximation of the Koopman operator for the Lorenz system. The Koopman operator is usually approximated via the extended dynamic mode decomposition (EDMD) algorithm [89, 35, 37]; however, the choice of observable dictionary in EDMD, which plays a critical role in the approximation, is not generally well resolved. The most popular choice is to include delay-embeddings of the measurements [1], but it is known that for measure-preserving chaotic systems the finite-dimensional approximations possess only eigenvalues with negative real part which lead to trivial statistics [38]. Here, we use the phase model approximation of the observable evolution in the form of

$$y(t) = \sum_{j=1}^{n} \alpha_{j,y} e^{i(\omega_j t + \zeta_j)}, \tag{21}$$

where ω_j's are randomly chosen frequencies from the support of the Koopman spectral measure for y, α_j's are determined from spectral projections of y, and ζ_j's are initial phases randomly drawn from $\mathcal{U}[0, 2\pi)$ (see Appendix C for more details). It is shown in [38] that, as $n \to \infty$, such an approximation converges to the true evolution of y over *finite time intervals*. A realization of this quasi-periodic phase model for the Lorenz state for $n = 200$ is shown in Figure 2. This model, by construction, accurately reproduces the spectrum of the data but fails to capture the skewness in the PDF. In fact, the phase model, which is commonly used for ocean wave modeling, is shown to converge to Gaussian statistics as $n \to \infty$ [34]. This example highlights the fact that the current data-driven approximations of the Koopman operator in the form of finite-dimensional linear systems cannot reproduce complicated statistics observed in measure-preserving chaotic systems. We note that there have been a few recent works that aim to find nonlinear or stochastic representation of the Koopman operator that does not have the limitations of the quasi-periodic models [9, 57], but application of those approaches to strongly nonlinear flows remains to be explored.

3.2. Chaotic cavity flow and spectral proper orthogonal decomposition (SPOD). Fluid flows at large length and velocity scales are canonical examples of high-dimensional chaotic behavior. We discuss the application of our framework to such flows using a numerical model of the chaotic lid-driven cavity flow. This flow consists of an incompressible fluid in a 2D square domain, $\mathcal{D} = [-1, 1]^2$, with solid walls, where the steady sliding of the top wall, given by

$$u(y = 1) = (1 - x^2)^2, \quad -1 \le x \le 1, \tag{22}$$

induces a circulatory fluid motion inside the cavity. The Reynolds number of this flow is defined as $Re = 2/\kappa$, where κ is the kinematic viscosity of the fluid in the numerical simulation. The cavity flow at $Re = 30{,}000$ converges to a measure-preserving chaotic attractor exhibiting a purely continuous Koopman spectrum [2]. Here, we model the evolution of the post-transient cavity flow in two steps: first, we use the modes obtained by SPOD [45, 82] as a spatial basis for description of the flow evolution. We justify the use of SPOD modes through its connection

with the spectral expansion of the Koopman operator with continuous spectrum. In the second step, we use the framework based on measure transport and spectral matching to find the stochastic model that captures the evolution of flow in the SPOD coordinates.

Consider a nonhomogeneous stationary turbulent flow. The velocity field at each point is a random variable defined on the underlying measure-preserving attractor, but due to nonhomogeneity, velocity at different points have different statistics and spectra. However, since all these variables arise from the same underlying attractor, we can define characteristic spatial fields that connect the spectra of various variables through the notion of Koopman spectral measure. Let $\mathbf{u}_{\mathbf{x}}$ and $\mathbf{u}_{\mathbf{x}'}$ denote velocity field at location $\mathbf{x}$ and $\mathbf{x}'$ in the flow domain. The spectral expansion of the Koopman operator for these two random variables is

$$\langle \mathbf{u}(\mathbf{x}), U^{\tau}\mathbf{u}(\mathbf{x}')\rangle_{\mu} = \int_0^{2\pi} e^{i\omega\tau}\rho_{\mathbf{x},\mathbf{x}'}(\omega)d\omega, \tag{23}$$

where $\rho_{\mathbf{x},\mathbf{x}'}$ is the Koopman cross spectral density of $\mathbf{u}_{\mathbf{x}}$ and $\mathbf{u}_{\mathbf{x}'}$, and $\langle\cdot,\cdot\rangle_{\mu}$ is the inner product with respect to the invariant measure on the attractor [54]. The SPOD modes of the flow at frequency ω, denoted by $\psi(\mathbf{x},\omega)$, are defined as solutions of the eigenvalue problem

$$\int_{\Omega} \rho_{\mathbf{x},\mathbf{x}'}(\omega)\psi(\mathbf{x}',\omega)d\mathbf{x}' = \lambda\psi(\mathbf{x},\omega). \tag{24}$$

As such, SPOD modes are intrinsic dynamical properties that do not depend on the choice of flow realization, as opposed to dynamic modes or Fourier modes, and therefore provide a robust choice of basis for description of the flow evolution. A detailed discussion of Koopman spectral density and its connection with SPOD is given in Appendix B.

We compute the SPOD modes of cavity flow using the algorithm in [82]. We define the SPOD coordinates to be the projection of the flow onto the 10 most energetic SPOD modes at different frequencies, which contain $\sim 50\%$ of the turbulent kinetic energy, i.e.,

$$y_j(t) = \langle \mathbf{u}(\mathbf{x},t), \psi_j\rangle_{\mathcal{D}}, \qquad j = 1,2,\ldots,10, \tag{25}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{D}}$ denotes the spatial inner product over the flow domain.

As the training data for our SDE model, we use a single time series of SPOD coordinates with length of 2,500 seconds and sampling rate of 10 Hz. We use a polynomial map of (total) degree 3 to compute the transport between the distribution of SPOD coordinates and the standard normal distribution, and we identify the corresponding SDE. After generating a 10,000-second-long trajectory of the SDE, we compute the statistics of that trajectory, observed under the inverse of the polynomial map. Figure 3 (a),(b) shows the excellent agreement between the true marginals and the ones obtained from the SDE model. Next, we use the SPOD data generated by the model to construct a new flow trajectory and compare the statistics of pointwise velocity measurements with the original data projected to the 10-dimensional subspace of SPOD modes. The results, shown in Figure 3(c), confirm the accurate recovery of pointwise statistics, hence indicating the high skill of our model in capturing the 10-dimensional joint PDF of flow modal data. Moreover, the match between log PDFs shows the utility of polynomial transport for data-driven extrapolation of PDF tails which we will explore in the next example. More details on modeling of cavity flow including the structure of SPOD modes and all 2D marginals can be found in Appendix D.

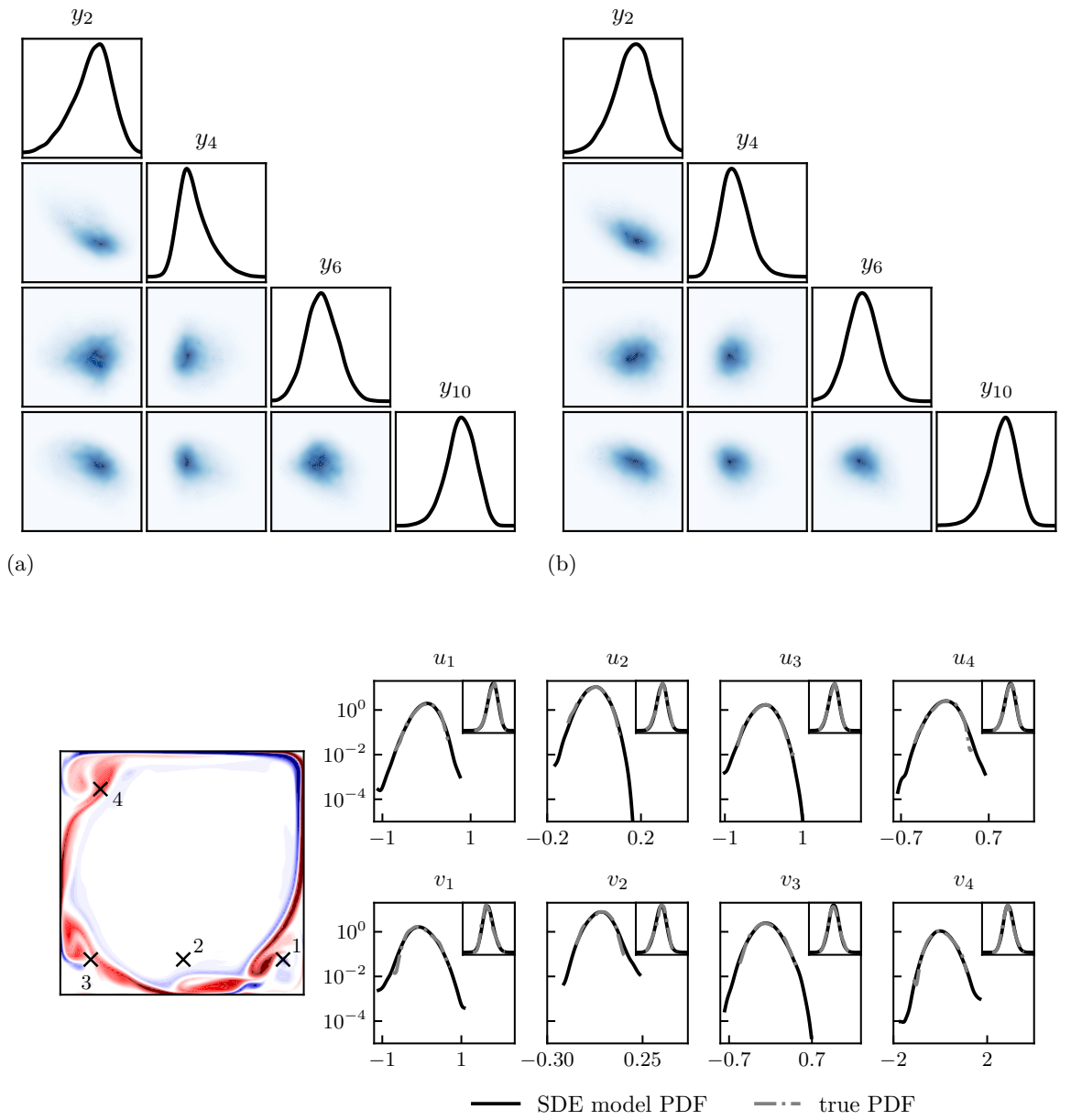


Figure 3. ***Modeling of chaotic cavity flow at*** $\mathbf{Re} = \mathbf{30{,}000}$**.** (a) *Single and pairwise marginal PDFs of* 4 *(out of* 10*) SPOD coordinates from data.* (b) *Same marginals by the SDE model. The quantile axes limits are* $(-0.026, 0.026)$. (c) *Location of sensors for velocity measurements and a vorticity snapshot of the cavity flow (left) and the log PDFs of u- and v-velocity generated by the SDE model and the* 10*-dimensional representation of the flow in SPOD coordinates (labeled as truth). The insets show the PDFs on a linear ordinate scale. The matching between the pointwise statistics indicates the skill of the SDE model in capturing the* 10*-dimensional joint PDF of SPOD coordinates.*

3.3. Reanalysis climate data and tail extrapolation. Statistical characterization of extreme events, i.e., probabilistically rare events that arise from combination of dynamics and randomness, is an important and challenging task. The importance stems from the significant and disruptive effect of extreme events such as earthquakes, rogue waves, and extreme weather patterns. Yet finding probabilities of extreme events, which reside at the tails of the PDF, is challenging because it requires a very large number of data points. The theoretical setup for the study of such events is the extreme value theory (e.g., [14]), which has been recently extended to chaotic dynamical system with some success [24, 44], but remains limited to systems with specific mixing properties. In recent years, there has been a number of methods that use model reduction or machine learning for data-efficient quantification of extreme events [58, 59].

Our modeling framework can be used for characterization of extreme events using relatively little data. The key idea is that we can model a large space of probability measures with heavy tails by using transport to a reference measure like standard normal distribution. By discovering such a transport map, we identify our data distribution with a pullback of Gaussian probability under that map, and the resulting distribution provides an approximation of the tails of the real data distribution. Although analytic description of the tails in these distributions is not available when the transport map is a high-degree multivariate polynomial map, we can easily generate a large sample from the reference measure and pass them through the transport map inverse to construct the tail approximation. In the following, we consider the application of this method to obtain a short extrapolation of super-Gaussian tails in climate data.

Our data is based on 6-hourly reanalysis of velocity and temperature in the earth atmosphere recorded at the 100-mbar iso-pressure surface from 1981 to 2017 [4]. These global fields are then expanded in a spherical wavelet basis described in [42]. Figure 4 shows the time series of the expansion coefficients for a level-1 wavelet envelope centered on top of the North Pole. These time series exhibit a combination of periodic and chaotic oscillations. Therefore,

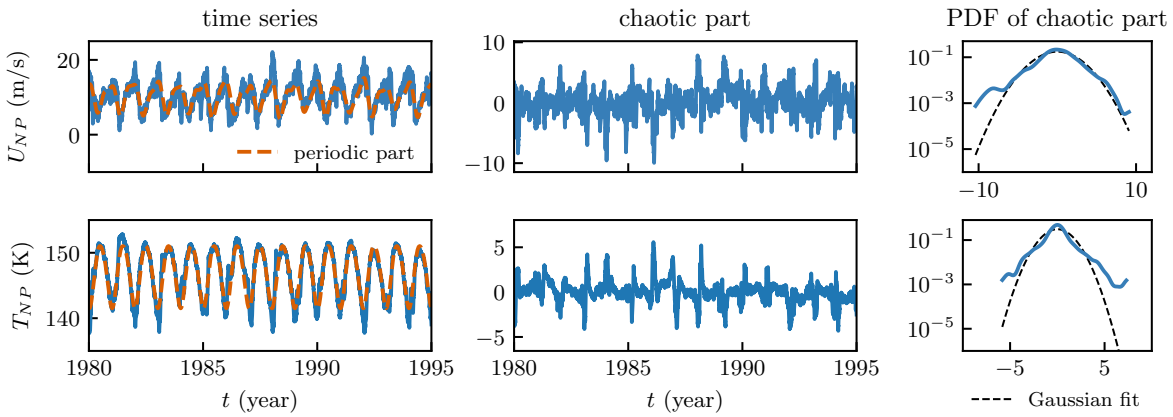


Figure 4. ***Velocity and temperature at the North Pole:*** *The time series showing wavelet coefficients of u -velocity and temperature field at the North Pole (left). We extract the periodic component consisting of Fourier modes with at least* 1% *of signal variance to isolate the chaotic part (middle), which shows super-Gaussian tails (right).*

we model them as observation on a stationary system with mixed spectra, i.e., a system that possesses both discrete and continuous Koopman spectra (see, e.g., [7, 54]). In the first step, we extract the periodic component of data (i.e., Koopman modes) by removing Fourier modes of time series that correspond to more than 1% of fluctuation energy. We apply our stochastic modeling framework to the chaotic remainders with heavy tails.

We construct a sequence of stochastic models for each of the chaotic time series obtained from measurements at the North Pole. In each model, we add an extra random variable from velocity and temperature measurements at other locations to capture the statistical dependence between those variables and the target variable (i.e., chaotic part of temperature or u-velocity). These covariates are chosen from the set of all wavelet coefficients ($> 8{,}000$ variables) in the order of decreasing absolute linear correlation with the target variables. The geographic position of the covariates is shown in Figure 5(a). Importantly, we exploit the triangular structure of transport in (8) and place the target variable on the bottom of each map, so that its mapping to the reference measure is informed by all the covariates. For training the models we use only the time series from 1981 and 1982 and polynomials of total degree 2.

After discovering each model, we generate a 74-year long trajectory of the SDE model and pull it back to the space of observations. The PDFs of the observed map provide a long-time extrapolation to the PDFs of the training data. As shown in Figure 5, those PDFs are in good agreement with the original reanalysis data of 37 years. In particular, the approximation of tails with good training data (i.e., left tails in the top and bottom rows) shows consistency with the addition of covariates to the model.

4. Dependence on training data and degree of polynomial mapping. Under our assumptions for modeling, the solution to the transport map is known to exist; however, restricting the search to a specific set of functions, such as polynomials, induces a bias in approximating the statistics of the flow. We are usually interested in evaluating the expected value of a square-integrable random variable h, and it is known [53] that

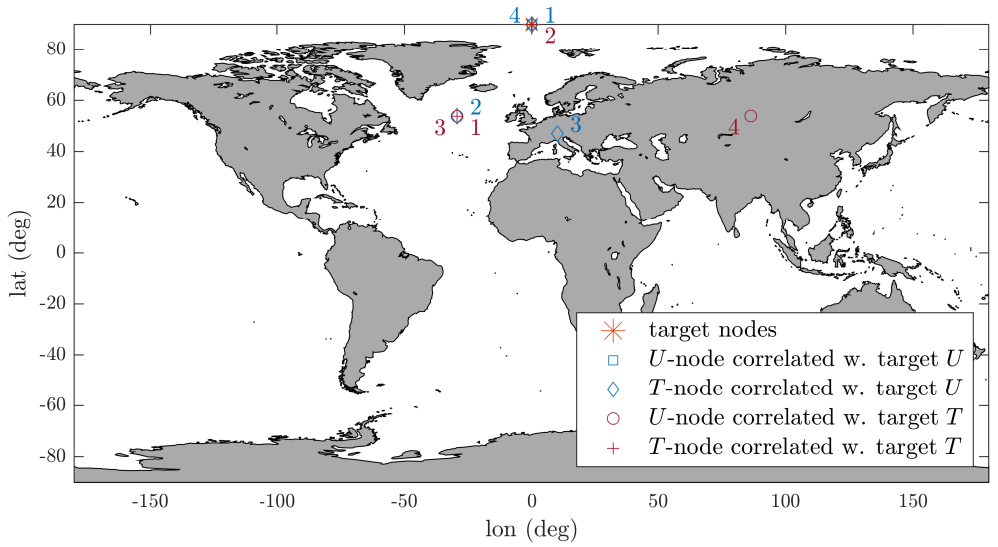
$$\|\mathbb{E}_\nu[h] - \mathbb{E}_{\tilde{\nu}}[h]\| = \sqrt{2(\mathbb{E}_\nu[\|h\|^2] - \mathbb{E}_{\tilde{\nu}}[\|h\|^2]}\sqrt{\mathbb{E}_\nu\left[\log\frac{\nu}{\tilde{\nu}}\right]}, \tag{26}$$

where ν is the target distribution and $\tilde{\nu}$ is the distribution generated by our model. Note that the last term on the right-hand side is the objective of the optimization problem in subsection 2.2, and therefore we can control the bias in estimation of the statistics through our approximation. Rigorous results on convergence for the type of transport maps we use here are sparse (for example, see [92] for a similar setup on bounded domains), and theoretical characterization of the bias in application to strongly nonlinear flows is beyond the scope of this paper. Instead, we perform a numerical study of convergence for examples of the Lorenz system and cavity flow where we have access to extensive data as the ground truth.

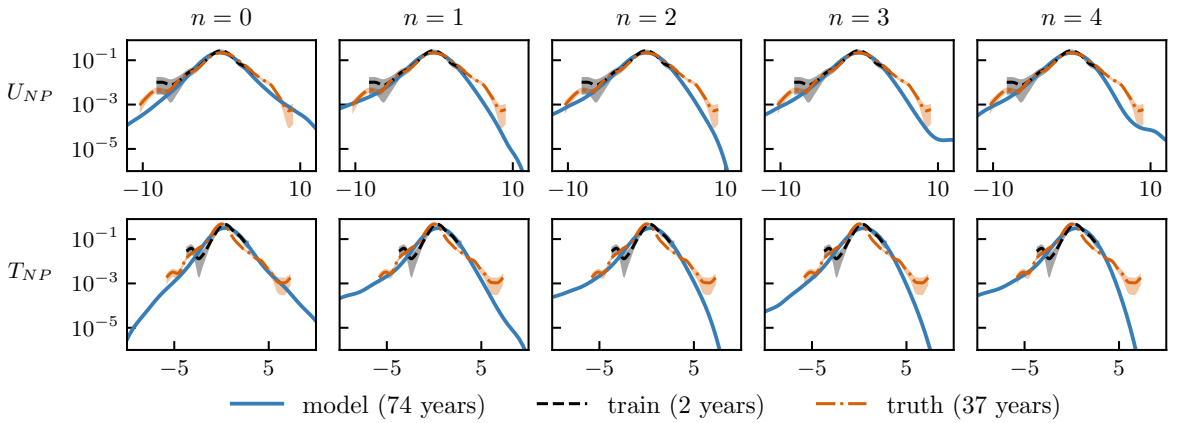
To assess the quality of modeling via transport we use the *variance diagnostic* [53],

$$e = \mathbb{V}\mathrm{ar}_{\rho_{\mathrm{truth}}}\left(\log\frac{\rho_{\mathrm{truth}}}{\rho_{\mathrm{model}}}\right), \tag{27}$$

where ρ_{truth} and ρ_{model} denote the density of the data distribution and the density of the distribution generated by the transport model. In the case of Lorenz-96, we have direct access



(a)



(b)

Figure 5. ***Extrapolation of tails for climate data.*** (a) *Location of covariates used to build models for the target random variables, i.e., u-velocity and temperature at the North Pole.* (b) *approximation of PDF tails by generating a surrogate trajectory of the SDE model and pulling it back under the transport map. n is the number of covariates used to learn the SDE model. The training data is the time series in* 1981–1982 *and the truth data is the time series in* 1981–2017. *The shaded envelopes show the* 95% *(pointwise) confidence interval of PDF estimation for training and truth data.*

to the PDF of the state variable (computed from simulations and taken as truth), and we can evaluate the diagnostic directly. As shown in Figure 6(a), with the increase of the training sample size and the polynomial degree of the transport map, the variance diagnostic rapidly decreases, indicating the scalable accuracy of the transport-based model for the Lorenz-96 system.

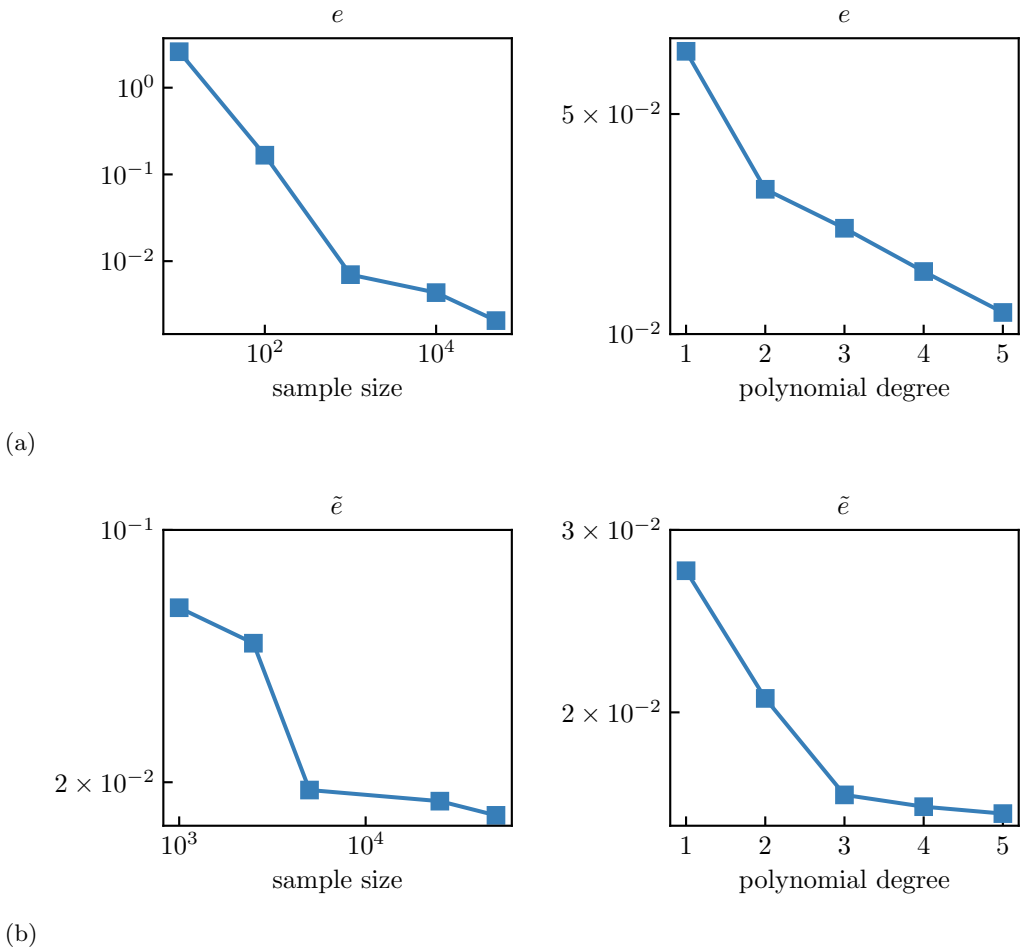


Figure 6. ***Statistical error for modeling via transport***. (a) *Lorenz-96 system: the PDF variance error of state variable (left) versus sample size with fixed polynomial degree* 3 *and (right) versus polynomial degree with fixed sample size of* 1,000. (b) *Cavity flow: the average of PDF variance error for pointwise velocity (right) versus sample size with fixed polynomial degree* 2 *and (right) versus polynomial degree with fixed sample size* 25,000.

In the cavity flow example, the target distribution is given by sample data in a 10-dimensional space, and there is no analytic expression for its density. Therefore, as a proxy for the error of modeling, we look at the variance diagnostic in recovering the statistics of a set of observables of the flow. Specifically, we use the statistics of pointwise velocity measurements because they are functions of all the SPOD coordinates, and therefore reveal the performance of the model in emulating the *high-dimensional joint* distribution of the data. Let $e_{u(\mathbf{x})}, e_{v(\mathbf{x})}$ be the variance diagnostic for the pointwise PDF of the, respectively, u- and v-velocity field at position $\mathbf{x}$ in the flow domain. We compute the average of this error over 50 randomly placed sensors in the flow to assess the overall quality of this model:

$$\tilde{e} = \frac{1}{n} \sum_{k=1}^{50} e_{u(\mathbf{x}_k)} + e_{v(\mathbf{x}_k)}. \tag{28}$$

The results in Figure 6(b) are similar: with the increase of training sample size and the polynomial degree, the proxy variance diagnostics decreases. This shows that the space of models considered here contains good candidates for learning models of strongly nonlinear flows from relatively little data.

5. Discussion. We presented a framework for generative modeling of strongly nonlinear flows in the form of decoupled stochastic oscillators with nonlinear observation maps. The key feature of our framework is the use of measure transport to model non-Gaussian invariant measures that arise from statistics of complex chaotic systems. Application to the high-Reynolds cavity flow showed that models generated by our framework can accurately reproduce the 10-dimensional joint pdfs of modal coordinates and hence recover the pointwise statistics in the flow. We also showed the promise of our framework in data-driven characterization of extreme events through an example of reanalysis climate data.

In the context of operator-theoretic approximation, our framework is closest in spirit to the approach in [17, 18] where the Perron–Frobenius operator of the underlying dynamical system is approximated using a Markov chain for transition between the computational cells in the state space (also see [27] for a similar Koopman operator approximation). Although this approach reproduces both the statistics and dynamical behavior, it is not computationally scalable to moderate and high-dimensional systems, and for systems like complex flows one needs to use extreme coarse-graining of the dynamics, e.g., by clustering data into a few discrete representative states [30, 66]. More recent data-driven approaches for generative modeling are the generative adversarial network (GAN) [26] and variational autoencoders (VAEs) [33]. GANs and VAEs have produced exemplary results in emulating high-dimensional and rich data distributions; however, like other deep networks, they offer little interpretability and lack an inherent dynamical character. Training GANs, in particular, often requires nontrivial techniques, and interestingly one of the effective improvements on GAN training, the Wasserstein GAN [3], is based on ideas from measure transport. The framework presented in this paper provides a balanced trade-off for data-driven modeling of strongly nonlinear systems that arise in turbulent flows. Through the use of single-layer triangular polynomial maps, it offers much interpretability in assessing the role of different variables on the observed dynamics, and due to separability of the underlying optimization problem it is easily extendable to tens of dimensions.

One promising direction for extension of our framework is to include physics-based constraints in computation of transport maps and the underlying SDE models. As shown in previous works, incorporating physical information about the target system into the structure of learning can substantially increase the data-efficiency of modeling. In the context of statistical modeling for turbulent flows, statistical laws and tail slopes predicted from the theory provide suitable constraints. In this work, we showed the effectiveness of our approach for learning strongly nonlinear fluid flows, but our framework also has great potential for data-driven modeling in other applications such as nonlinear optics, chemical reaction networks, molecular dynamics, and biological systems. For these systems, it might be necessary to make other judicious choices for the reference measure and the type of transport maps to achieve an efficient learning process.

Appendix A. Data and compilation of PDFs. The data for Lorenz-96 system is generated using direct integration of (18) using fourth-order Runge–Kutta with time step $\Delta t = 0.01$. The velocity field for the lid-driven cavity flow is computed by direct solution of Navier–Stokes equations in NEK5000 [64]. The cavity domain is divided into 25 elements in each direction, and the solution within each element is represented with a polynomial of 7th degree. The stationary solution is recorded after 3,500 transient simulation time units. The processed reanalysis climate data is provided by AIR Worldwide. The PDFs reported for Lorenz, cavity flow, and climate (truth and model) are computed using a 100-bin histogram and, for visualizations, smoothed by convolving with Gaussian kernel with standard deviation of 2 bin widths. The PDFs for the climate training data are computed using 50 bins. The confidence intervals for PDFs of climate are computed using binomial statistics and the adjusted Wald formula in [1].

A Python implementation of our framework and the data for producing the figures in this paper can be found at https://github.com/arbabiha/StochasticModelingwData.

Appendix B. Cross-spectral density of Koopman operator and Spectral Proper Orthogonal Decomposition (SPOD). In this section, we recall the Koopman spectral expansion for chaotic systems and define the cross-spectral density of Koopman operator which is used in the definition of SPOD in subsection 3.2. Consider a deterministic dynamical system with an attractor Ω which supports a physical measure μ, and $\mu(\Omega) = 1$. Let $f, g \in \mathcal{H} := L^2(\Omega, \mu)$ be observables of this system, with the usual inner product

$$\langle f, g \rangle_\mu = \int_A f g^* d\mu. \tag{B.1}$$

Now recall the Koopman operator U^t defined as $Uf = f \circ F^t$, where F^t is the reversible flow of the dynamical system. U^t is a *unitary* operator on $\mathcal{H}$, that is, $(U^t)^* = U^{-t}$. This implies that the spectrum of the Koopman operator lies on the imaginary axis in the complex plane, and the spectral expansion of the Koopman operator [54, 55] is given as

$$U^t f = \sum_{k=0}^{\infty} v_k \phi_k e^{i\omega_k t} + \int_0^{2\pi} e^{i\omega t} dE_\omega f, \tag{B.2}$$

where the countable sum is the Koopman mode decomposition of f associated with the quasi-periodic part of the evolution, $i\omega_j$ is a Koopman eigenvalue (i.e., an element of discrete spectrum) associated with eigenfunction ϕ_k, and E is the spectral measure of the Koopman operator associated with the continuous part of the spectrum. To be more precise, E is a measure on $[0, 2\pi)$ that takes values in the space of projections on H. That is, for every measurable set $B \subset [0, 2\pi)$, E_B is a projection operator, and $E_B f$ is projection of f onto the eigensubspace associated with the part of spectrum residing in B. We are interested in the continuous part of the Koopman spectrum which corresponds to chaotic behavior, and therefore *we assume that there are no quasi-periodic parts, including the part with zero frequency (i.e., mean of the observable) present in the evolution.*

The operator-valued measure E is difficult to characterize using data. Instead, we can define a positive real-valued measure associated only with f and defined as follows [47]:

$$\mu_f(B) = \langle E(B)f, f\rangle_\mu. \tag{B.3}$$

Then we can rewrite the spectral expansion of f as

$$\langle U^t f, f\rangle_\mu = \int_0^{2\pi} e^{i\omega t}\langle dE_\omega f, f\rangle = \int_0^{2\pi} e^{i\omega t} d\mu_f(\omega), \tag{B.4}$$

where ρ_f is the Koopman spectral density of observable f. Similarly, we can define a measure for spectral correlation of two distinct observables. That is,

$$\langle E(B)f, g\rangle_\mu = \langle f, E(B)g\rangle_\mu = \mu_{f,g}(B) \tag{B.5}$$

defines a finite complex-valued measure $\mu_{f,g}$ on $[0, 2\pi)$. Under the assumption of absolute continuity for this measure, we can write

$$\mu_{f,g}(B) = \int_B \rho_{f,g} d\alpha, \tag{B.6}$$

with $\rho_{f,g}$ being the *Koopman spectral density of observables* f *and* g. Consequently, we can write the following expansion for the dynamic evolution of the two observables

$$\langle f, U^\tau g\rangle_\mu = \int_0^{2\pi} e^{i\omega\tau}\rho_{f,g}(\omega)d\omega. \tag{B.7}$$

In view of the duality between measure-preserving deterministic systems and stationary stochastic processes [20], the expansion in (B.7) is the same as the spectral expansion for stochastic processes used in the definition of SPOD [45, 82].

Appendix C. Construction of random phase model for chaotic systems. Let $g : \Omega \to \mathbb{R}$ be an observable on the measure-preserving dynamical system. Given the spectral density of g, Algorithm C.1 generates a random phase model for the evolution of g in time. The main idea is to approximate the spectral measure of g using sum of delta functions, that is, modeling the evolution as rotation on tori. Using ergodicity, we can use a single realization to compute the PDF reported in subsection 3.1.

To see the connection with the approximation of the Koopman operator, note that the work in [38] has shown that approximations such as

$$U_m^1 g := \sum_{j=1}^{m} e^{2\pi i\omega_j} E(B_j)g, \tag{C.1}$$

where B_j's are a partition of $[0, 2\pi)$ and $\omega_j \in B_j$, will converge to the true evolution $U^1 g$ in $L^2(\Omega, \mu)$ in the limit of infinitely refined partition. This approximation is in the function space, and the choice of B_j, ω_j's are not unique. Algorithm C.1 generates a realization of this approximation, with $B_j = [e_{j-1}, e_j)$, that is spectrally consistent, i.e.,

$$|a_j|^2 = \int_{B_j} \rho_g(\omega)d\omega. \tag{C.2}$$

Algorithm C.1 Construction of random phase model from spectral density

Initialization: spectral density $\rho(\omega)$, number of intervals on the frequency domain m

1: Draw m random value of cell edges, e_j, from uniform distribution on $[0, 2\pi)$.
2: Set $e_0 = 0$ and $e_{m+1} = 2\pi$.
3: Sort the random cell edges to form the sequence $e_j, j = 0, 1, \ldots, m+1$ with $e_{j+1} > e_j$.
4: **for** $j = 1, \ldots, m$ **do**
5: Let $\omega_j = \frac{1}{2}(e_j + e_{j-1}),\ j = 1, \ldots, m$.
6: Let $\Delta\omega_j = e_j - e_{j-1},\ j = 1, \ldots, m$.
7: Let

$$a_j = \sqrt{\rho(\omega_j)\Delta\omega_j} \approx \left(\int_{e_{j-1}}^{e_j} \rho(\omega) d\omega \right)^{1/2}. \tag{C.3}$$

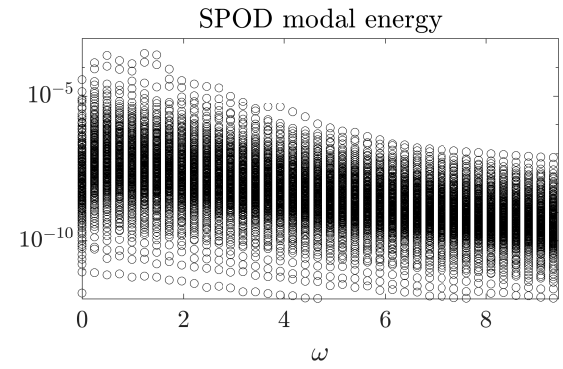
8: **end for**
9: Draw ζ_j randomly and independently for each j from the uniform distribution on $[0, 2\pi)$ and let

$$g_t = \sum_{j=1}^{m} a_k e^{i(\omega_j t + \zeta_j)}. \tag{C.4}$$

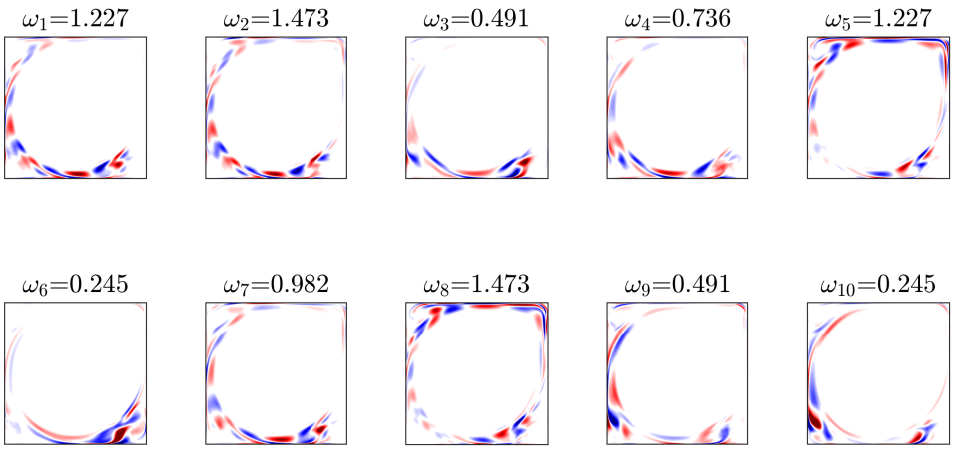
Appendix D. Supplemental information and figures for cavity flow. The data for the cavity flow consists of a single trajectory with the length of 12,000 seconds. We have removed the mean flow from the data and then applied the algorithm in [82] to the first 2,000 seconds to compute the SPOD modes of flow shown in Figure 7. Then we projected the next 10,000 seconds onto the top 10 energetic modes, using (25), to obtain the SPOD coordinates. We have used the first 2,500 second of the SPOD coordinate time series, with sampling rate of 10 Hz, for training of the SDE model. The marginal distributions of flow data (Figure 8) are computed using the whole trajectory. An SDE trajectory of the same length is used to compute the marginal distributions in Figures 3 and 9. To compute the pointwise statistics, we reconstruct the flow field via

$$\tilde{\mathbf{u}}(\mathbf{x}) = \sum_{j=1}^{N} \sum_{k=1}^{N} H_{jk} y_k \boldsymbol{\psi}_j(\mathbf{x}), \tag{D.1}$$

where the matrix H is the inverse of Gramian matrix G defined as $G_{jk} = \langle \boldsymbol{\psi}_j, \boldsymbol{\psi}_k \rangle_{\mathcal{D}}$. The results in this paper are compiled using the real part of SPOD modes for simplicity of presentation.



(a)



(b)

Figure 7. ***Spectral orthogonal decomposition (SPOD) for cavity flow at*** $\mathbf{Re} = \mathbf{30{,}000}$**.** (a) *Distribution of the kinetic energy within the SPOD modes vs. frequency; for each frequency we have* 155 *spatially orthogonal modes.* (b) *The (real part of the) vorticity for the top* 10 *energetic SPOD modes corresponding to the* 10 *highest circles in* (a)*. The red color shows clockwise rotation and the blue counterclockwise.*

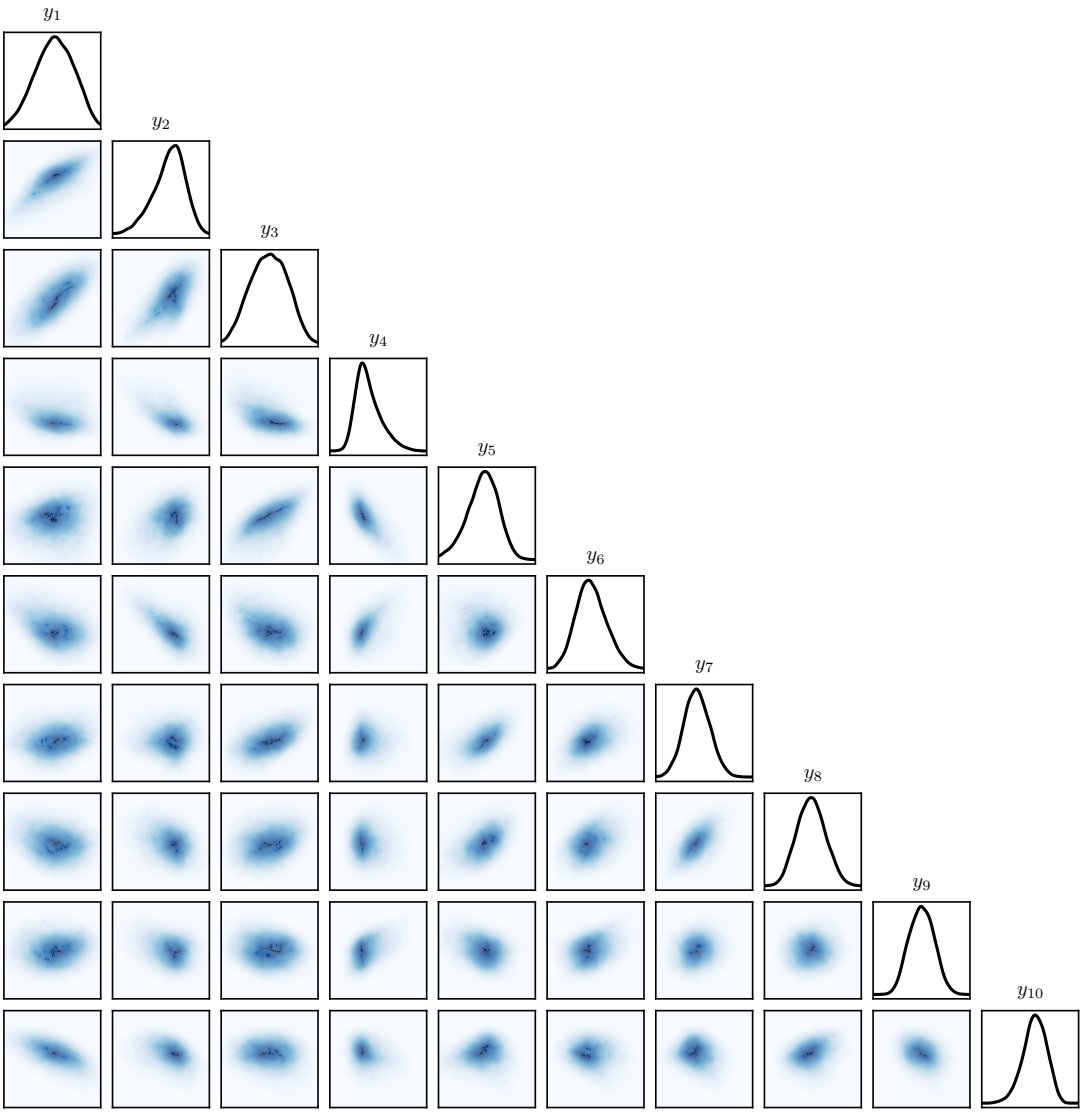


Figure 8. *Marginal distributions of SPOD coordinates in cavity flow from the flow simulation. The quantile axes limits are* $(-0.026, 0.026)$.

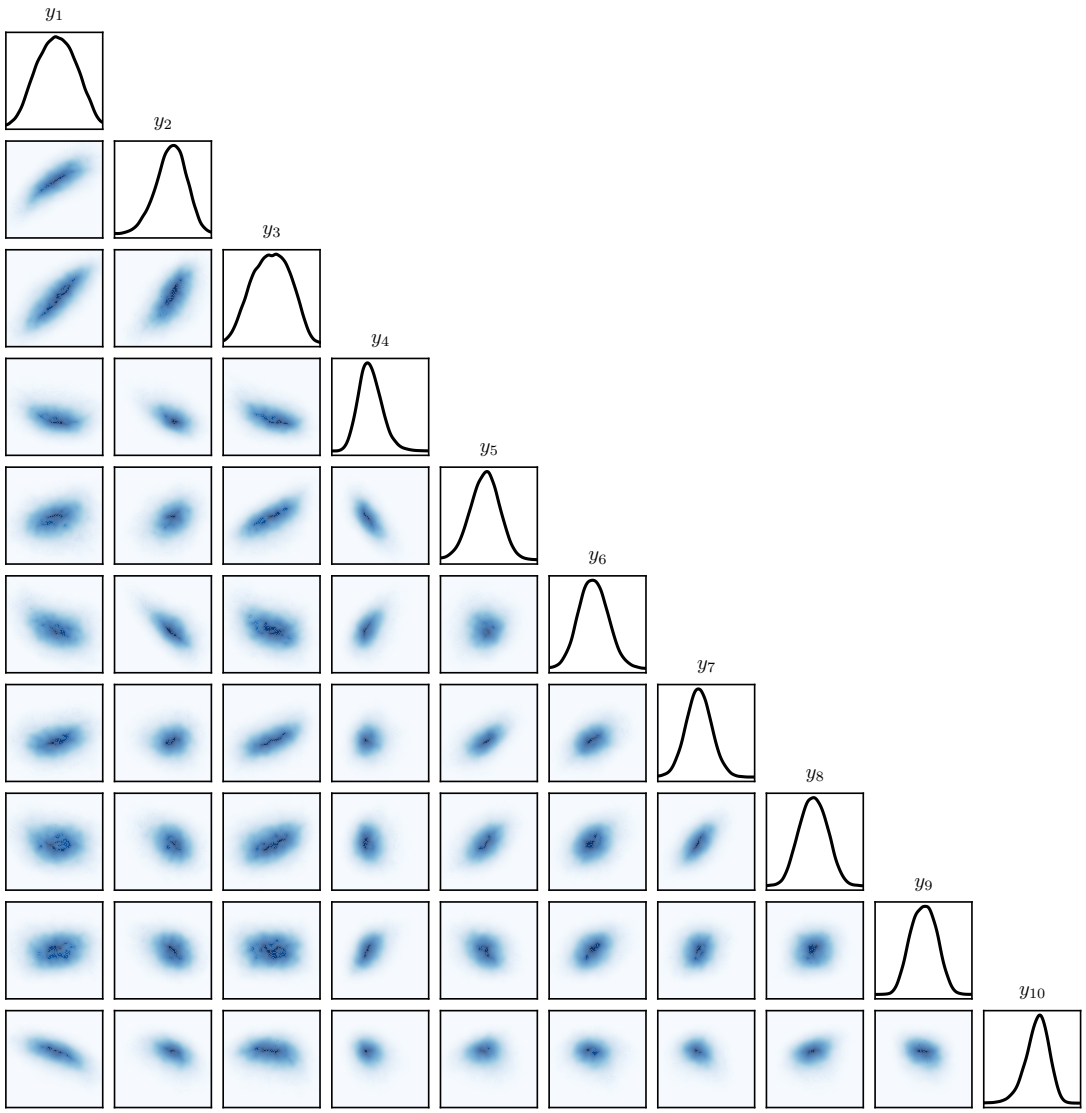


Figure 9. *Marginal distributions of SPOD coordinates in cavity flow from the SDE model. The quantile axes limits are* $(-0.026, 0.026)$. *Compare to the truth in Figure* 8.

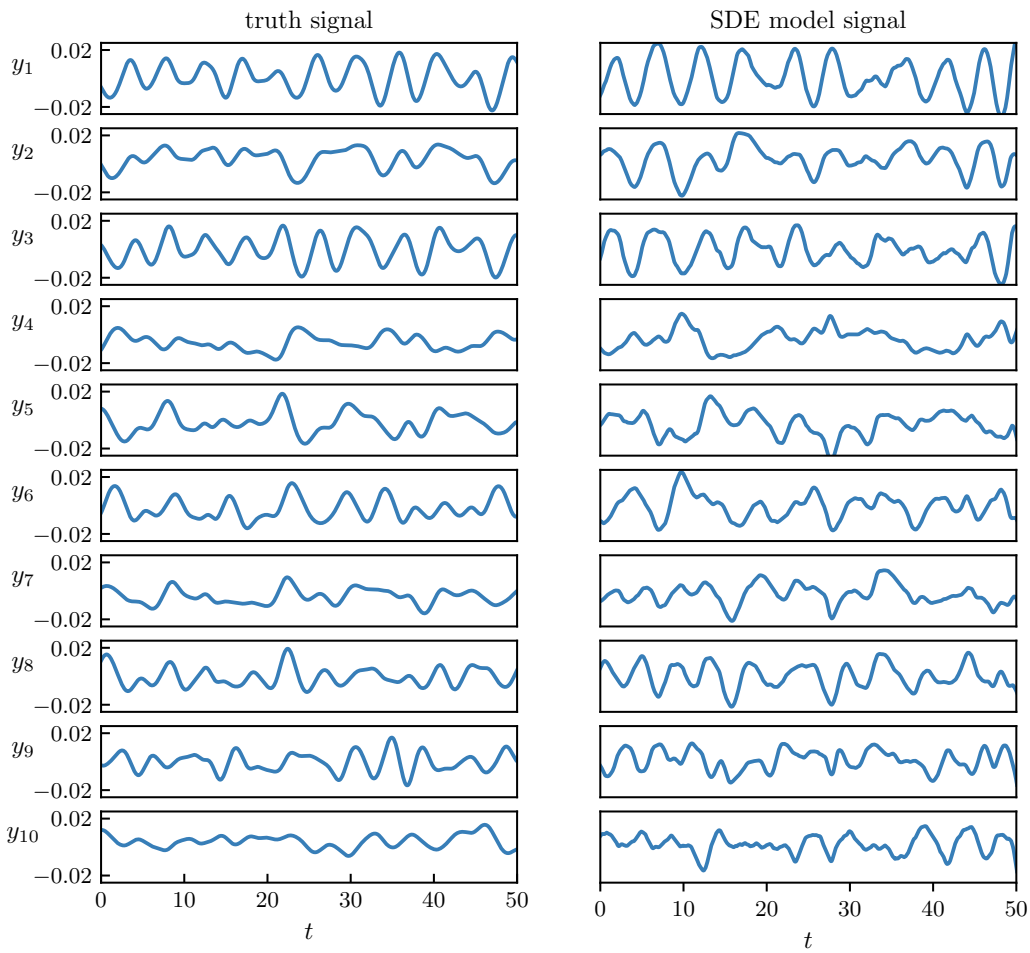


Figure 10. *Time evolution of SPOD coordinates for cavity flow with samples from the truth data (numerical simulation of the flow, shown on the left) and a realization from the SDE model (right).*

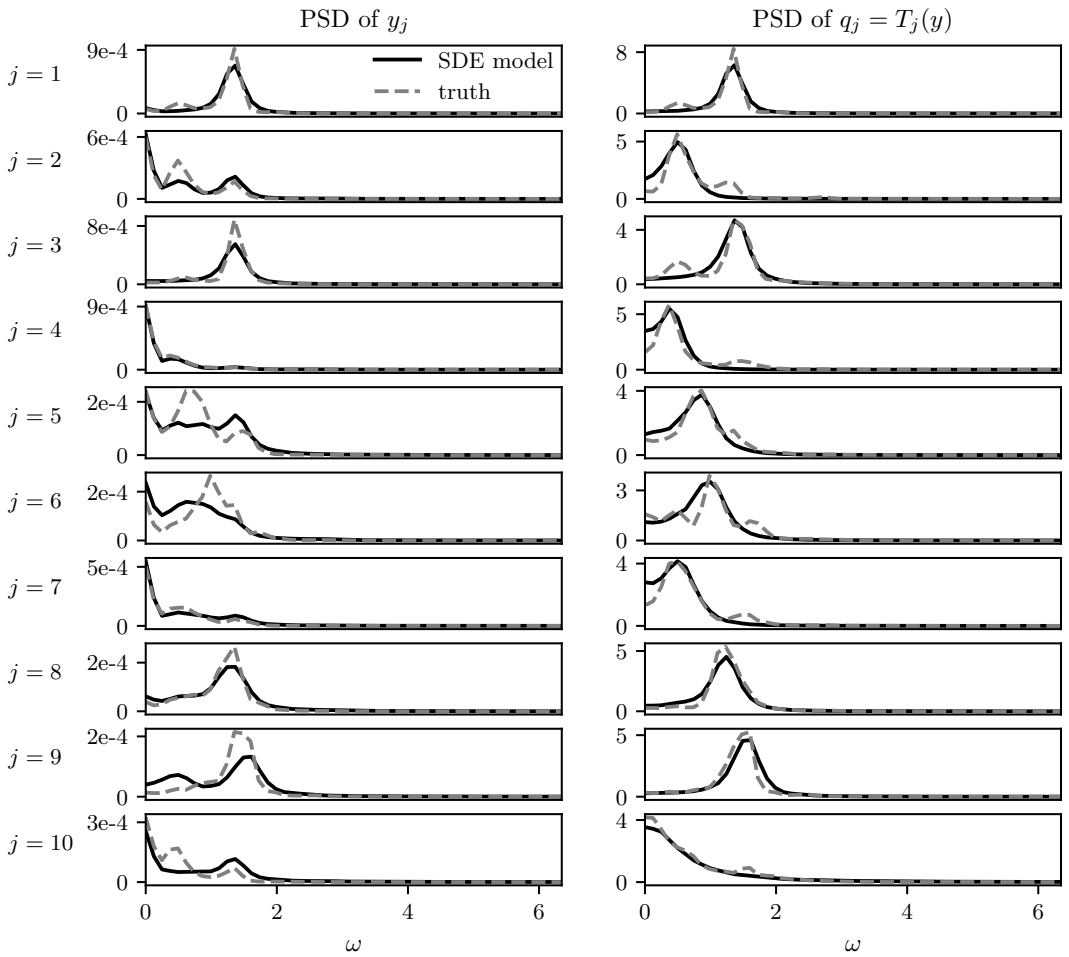


Figure 11. *Power spectral density (PSD) of SPOD coordinates from the flow simulation and the SDE model (left) and the transported variable (right).*

Acknowledgments. We are grateful to Dr. Boyko Dodov and AIR Worldwide for providing the reanalysis climate data, as well as Prof. Christian Lessig, who performed the projection of the data to a spherical wavelet basis. We also thank Profs. Youssef Marzouk, Igor Mezić, and Yannis Kevrekidis for instructive discussions and pointing out related references. H.A. is grateful to Dr. Antoine Blanchard for notes on usage of NEK5000 and to Dr. Daniele Bigoni on usage of the computational package for transport.

Data and source code. A Python implementation of our framework and the data for producing the figures in this paper is available at https://github.com/arbabiha/StochasticModelingwData.

REFERENCES

[1] H. Arbabi and I. Mezićc, *Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator*, SIAM J. Appl. Dyn. Syst., 16 (2017), pp. 2096–2126, https://doi.org/10.1137/17M1125236.

[2] H. Arbabi and I. Mezić, *Study of dynamics in post-transient flows using Koopman mode decomposition*, Phys. Rev. Fluids, 2 (2017), 124402, https://doi.org/10.1103/PhysRevFluids.2.124402.

[3] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein GAN*, preprint, https://arxiv.org/abs/1701.07875, 2017.

[4] P. Berrisford, D. Dee, P. Poli, R. Brugge, M. Fielding, M. Fuentes, P. Kallberg, S. Kobayashi, S. Uppala, and A. Simmons, *The ERA-interim archive version* 2.0, ECMWF Report, 2011.

[5] D. Bigoni, A. Spantini, R. Morrison, R. M. Baptista, and Y. Marzouk, *Transport Maps Software Documentation*, 2015–2020, http://transportmaps.mit.edu/docs/

[6] S. Boyd and L. Chua, *Fading memory and the problem of approximating nonlinear operators with Volterra series*, IEEE Trans. Circuits Syst., 32 (1985), pp. 1150–1161.

[7] H. Broer and F. Takens, *Mixed spectra and rotational symmetry*, Arch. Ration. Mech. Anal., 124 (1993), pp. 13–42.

[8] B. W. Brunton, L. A. Johnson, J. G. Ojemann, and J. N. Kutz, *Extracting spatial-temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition*, J. Neurosci. Methods, 258 (2016), pp. 1–15.

[9] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz, *Chaos as an intermittently forced linear system*, Nat. Commun., 8 (2017), 19.

[10] A. Chattopadhyay, P. Hassanzadeh, and D. Subramanian, *Data-driven predictions of a multiscale Lorenz* 96 *chaotic system using machine-learning methods: Reservoir computing, artificial neural network, and long short-term memory network*, Nonlinear Proc. Geophys., 27 (2020), pp. 373–389.

[11] N. Chen, A. J. Majda, and D. Giannakis, *Predicting the cloud patterns of the Madden-Julian oscillation through a low-order nonlinear stochastic model*, Geophys. Res. Lett., 41 (2014), pp. 5612–5619.

[12] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.

[13] M. J. Colbrook and A. Townsend, *Rigorous Data-Driven Computation of Spectral Properties of Koopman Operators for Dynamical Systems*, preprint, https://arxiv.org/abs/2111.14889, 2021.

[14] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An Introduction to Statistical Modeling of Extreme Values*, Springer Ser. Statist. 208, Springer, 2001.

[15] P. Cvitanović, R. Artuso, R. Mainieri, G. Tanner, and G. Vattay, *Chaos: Classical and Quantum*, ChaosBook.org, Niels Bohr Institute, Copenhagen, 2005.

[16] A. de la Torre and J. Burguete, *Slow dynamics in a turbulent von kármán swirling flow*, Phys. Rev. Lett., 99 (2007), pp. 054101.

[17] M. Dellnitz and O. Junge, *On the approximation of complicated dynamical behavior*, SIAM J. Numer. Anal., 36 (1999), pp. 491–515, https://doi.org/10.1137/S0036142996313002.

[18] M. Dellnitz and O. Junge, *Set oriented numerical methods for dynamical systems*, in Handbook of Dynamical Systems, Vol. 2, North-Holland, Amsterdam, 2002, pp. 221–264.

[19] T. DelSole, *Stochastic models of quasigeostrophic turbulence*, Surveys Geophys., 25 (2004), pp. 107–149.
[20] J. L. Doob, *Stochastic Processes*, Wiley, New York, 1953.
[21] T. A. El Moselhy and Y. M. Marzouk, *Bayesian inference with optimal maps*, J. Comput. Phys., 231 (2012), pp. 7815–7850.
[22] N. B. Erichson, S. L. Brunton, and J. N. Kutz, *Compressed dynamic mode decomposition for background modeling*, J Real-Time Image Process., 16 (2019), pp. 1–14.
[23] B. F. Farrell and P. J. Ioannou, *A theory for the statistical equilibrium energy spectrum and heat flux produced by transient baroclinic waves*, J. Atmos. Sci., 51 (1994), pp. 2685–2698.
[24] A. C. M. Freitas and J. M. Freitas, *On the link between dependence and independence in extreme value theory for dynamical systems*, Stat. Probab. Lett., 78 (2008), pp. 1088–1093.
[25] D. Giannakis, *Delay-coordinate maps, coherence, and approximate spectra of evolution operators*, Res. Math. Sci., 8 (2021), pp. 1–33.
[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in Neural Information Processing Systems, Curran Associates, 2014, pp. 2672–2680.
[27] N. Govindarajan, R. Mohr, S. Chandrasekaran, and I. Mezić, *On the Approximation of Koopman Spectra for Measure Preserving Transformations*, preprint, https://aps.arxiv.org/abs/1803.03920, 2018.
[28] J. Harlim and A. Majda, *Filtering nonlinear dynamical systems with linear stochastic models*, Nonlinearity, 21 (2008), pp. 1281.
[29] K. Hasselmann, *Stochastic climate models part* I. *Theory*, Tellus, 28 (1976), pp. 473–485.
[30] E. Kaiser, B. R. Noack, L. Cordier, A. Spohn, M. Segond, M. Abel, G. Daviller, J. Östh, S. Krajnović, and R. K. Niven, *Cluster-based reduced-order modelling of a mixing layer*, J. Fluid Mech., 754 (2014), pp. 365–414.
[31] M. Khodkar and P. Hassanzadeh, *Data-driven reduced modelling of turbulent Rayleigh-Bénard convection using DMD-enhanced fluctuation-dissipation theorem*, J. Fluid Mech., 852 (2018), R3.
[32] D. P. Kingma and M. Welling, *Auto-encoding Variational Bayes*, preprint, https://arxiv.org/abs/1312.6114, 2013.
[33] B. Kinsman, *Wind Waves: Their Generation and Propagation on the Ocean Surface*, Courier Corporation, 1984.
[34] S. Klus, P. Koltai, and C. Schütte, *On the Numerical Approximation of the Perron-Frobenius and Koopman Operator*, preprint, https://arxiv.org/abs/1512.05997, 2015.
[35] B. O. Koopman, *Hamiltonian systems and transformation in Hilbert space*, Proc. Natl. Acad. Sci. USA, 17 (1931), pp. 315–318.
[36] M. Korda and I. Mezić, *On convergence of extended dynamic mode decomposition to the Koopman operator*, J. Nonlinear Sci., 28 (2018), pp. 687–710.
[37] M. Korda, M. Putinar, and I. Mezić, *Data-driven spectral analysis of the Koopman operator*, Appl. Comput. Harmon. Anal., 48 (2020), pp. 599–629.
[38] S. Kravtsov, D. Kondrashov, and M. Ghil, *Multilevel regression modeling of nonlinear processes: Derivation and applications to climatic variability*, J. Climate, 18 (2005), pp. 4404–4424.
[39] A. Lasota and M. C. Mackey, *Chaos, Fractals and Noise*, Springer-Verlag, New York, 1994.
[40] C. Leith, *Climate response and fluctuation dissipation*, J. Atmos. Sci., 32 (1975), pp. 2022–2026.
[41] C. Lessig, *Divergence free polar wavelets for the analysis and representation of fluid flows*, J. Math. Fluid Mech., 21 (2019), 18.
[42] Q. Li, F. Dietrich, E. M. Bollt, and I. G. Kevrekidis, *Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the Koopman operator*, Chaos, 27 (2017), 103111.
[43] V. Lucarini, D. Faranda, A. C. G. M. M. de Freitas, J. M. M. de Freitas, M. Holland, T. Kuna, M. Nicol, M. Todd, and S. Vaienti, *Extremes and Recurrence in Dynamical Systems*, John Wiley & Sons, 2016.
[44] J. L. Lumley, *Stochastic Tools in Turbulence*, Academic Press, 1970.
[45] B. Lusch, J. N. Kutz, and S. L. Brunton, *Deep learning for universal linear embeddings of nonlinear dynamics*, Nature Commun., 9 (2018), 4950.

[46] B. MacCluer, *Elementary Functional Analysis*, Grad. Texts in Math. 253, Springer Science & Business Media, 2008.
[47] A. Majda, R. V. Abramov, and M. J. Grote, *Information Theory and Stochastics for Multiscale Nonlinear Systems*, CRM Monogr. Ser. 25. American Mathematical Society, 2005.
[48] A. Majda, I. Timofeyev, and E. Vanden-Eijnden, *Stochastic models for selected slow variables in large deterministic systems*, Nonlinearity, 19 (2006), pp. 769–794.
[49] A. J. Majda and J. Harlim, *Physics constrained nonlinear regression models for time series*, Nonlinearity, 26 (2012), pp. 201–217.
[50] A. J. Majda, I. Timofeyev, and E. V. Eijnden, *Models for stochastic climate prediction*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 14687–14691.
[51] A. J. Majda, I. Timofeyev, and E. Vanden Eijnden, *A mathematical framework for stochastic climate models*, Comm. Pure Appl. Math., 54 (2001), pp. 891–974.
[52] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini, *Sampling via measure transport: An introduction*, in Handbook of Uncertainty Quantification, Springer, Cham, 2017, pp. 785–825.
[53] I. Mezić, *Spectral properties of dynamical systems, model reduction and decompositions*, Nonlinear Dynam., 41 (2005), pp. 309–325.
[54] I. Mezić, *Analysis of fluid flows via spectral properties of the Koopman operator*, Annu. Rev. Fluid Mech., 45 (2013), pp. 357–378.
[55] I. Mezić, *Koopman operator, geometry, and learning of dynamical systems*, Notices Amer. Math. Soc., 68 (2021), pp. 1087–1105.
[56] I. Mezić, *Spectrum of the Koopman operator, spectral expansions in functional spaces, and state-space geometry*, J. Nonlinear Sci., 30 (2020), pp. 2091–2145.
[57] M. A. Mohamad, W. Cousins, and T. P. Sapsis, *A probabilistic decomposition-synthesis method for the quantification of rare events due to internal instabilities*, J. Comput. Phys., 322 (2016), pp. 288–308.
[58] M. A. Mohamad and T. P. Sapsis, *Sequential sampling strategy for extreme event statistics in nonlinear dynamical systems*, Proc. Natl. Acad. Sci. USA, 115 (2018), pp. 11138–11143.
[59] R. Morrison, R. Baptista, and Y. Marzouk, *Beyond normality: Learning sparse probabilistic graphical models in the non-Gaussian setting*, in Advances in Neural Information Processing Systems, Curran Associates, 2017, pp. 2359–2369.
[60] F. Noé and F. Nüske, *A variational approach to modeling slow processes in stochastic dynamical systems*, Multiscale Model. Simul., 11 (2013), pp. 635–655, https://doi.org/10.1137/110858616.
[61] S. E. Otto and C. W. Rowley, *Linearly recurrent autoencoder networks for learning dynamics*, SIAM J. Appl. Dyn. Syst., 18 (2019), pp. 558–593, https://doi.org/10.1137/18M1177846.
[62] M. D. Parno and Y. M. Marzouk, *Transport map accelerated Markov chain Monte Carlo*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 645–682, https://doi.org/10.1137/17M1134640.
[63] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, *Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach*, Phys. Rev. Lett., 120 (2018), 024102.
[64] B. Peherstorfer and Y. Marzouk, *A transport-based multifidelity preconditioner for Markov chain Monte Carlo*, Adv. Comput. Math., 45 (2019), pp. 2321–2348.
[65] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, *Identification of slow molecular order parameters for Markov model construction*, J. Chem. Phys., 139 (2013), 07B604˙1.
[66] G. Peyré and M. Cuturi, *Computational optimal transport*, Found. Trends Mach. Learn., 11 (2019), pp. 355–607.
[67] F. Raak, Y. Susuki and T. Hikihara, *Data-driven partitioning of power networks via Koopman mode analysis*, IEEE Trans. Power Syst., 31 (2015), pp. 2799–2808.
[68] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Numerical Gaussian processes for time-dependent and nonlinear partial differential equations*, SIAM J. Sci. Comput., 40 (2018), pp. A172–A198, https://doi.org/10.1137/17M1120762.
[69] R. Rico-Martinez, I. Kevrekidis, and K. Krischer, *Nonlinear system identification using neural networks: Dynamics and instabilities,* in Neural Networks for Chemical Engineers, Elsevier, 1995, pp. 409–442.
[70] R. Rico-Martinez, K. Krischer, I. Kevrekidis, M. Kube, and J. Hudson, *Discrete- vs. continuous-time nonlinear signal processing of cu electrodissolution data*, Chem. Engrg. Commun., 118 (1992), pp. 25–48.

[71] G. Rigas, A. Morgans, R. Brackston, and J. Morrison, *Diffusive dynamics and stochastic models of turbulent axisymmetric wakes*, J. Fluid Mech., 778 (2015), R2.
[72] C. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. Henningson, *Spectral analysis of nonlinear flows*, J. Fluid Mech., 641 (2009), pp. 115–127.
[73] W. J. Rugh, *Nonlinear System Theory*, Johns Hopkins University Press, Baltimore, MD, 1981.
[74] P. J. Schmid, *Dynamic mode decomposition of numerical and experimental data*, J. Fluid Mech., 656 (2010), pp. 5–28.
[75] P. J. Schmid, L. Li, M. Juniper, and O. Pust, *Applications of the dynamic mode decomposition*, Theoret. Comput. Fluid Dyn., 25 (2011), pp. 249–259.
[76] A. Singer and R. R. Coifman, *Non-linear independent component analysis with diffusion maps*, Appl. Comput. Harmon. Anal., 25 (2008), pp. 226–239.
[77] K. Sobczyk, *Stochastic Differential Equations: With Applications to Physics and Engineering*, Math. Appl. (East European Ser.) 40, Kluwer, 1991.
[78] A. Spantini, D. Bigoni, and Y. Marzouk, *Inference via low-dimensional couplings*, J. Mach. Learn. Res., 19 (2018), pp. 2639–2709.
[79] K. R. Sreenivasan, A. Bershadskii, and J. Niemela, *Mean wind and its reversal in thermal convection*, Phys. Rev. E, 65 (2002), 056306.
[80] N. Takeishi, Y. Kawahara, and T. Yairi, *Learning Koopman invariant subspaces for dynamic mode decomposition*, in Advances in Neural Information Processing Systems, Curran Associates, 2017, pp. 1130–1140.
[81] A. Towne, O. T. Schmidt, and T. Colonius, *Spectral proper orthogonal decomposition and its relationship to dynamic mode decomposition and resolvent analysis*, J. Fluid Mech., 847 (2018), pp. 821–867.
[82] C. Villani, *Optimal Transport: Old and New*, Grundlehren Math. Wiss. 338, Springer Science & Business Media, 2008.
[83] Z. Y. Wan and T. P. Sapsis, *Machine learning the kinematics of spherical particles in fluid flows*, J. Fluid Mech., 857 (2018), R2.
[84] Z. Y. Wan, P. Vlachas, P. Koumoutsakos, and T. Sapsis, *Data-assisted reduced-order modeling of extreme events in complex dynamical systems*, PloS One, 13 (2018), e0197704.
[85] J.-X. Wang, J.-L. Wu, and H. Xiao, *Physics-informed machine learning approach for reconstructing Reynolds stress modeling discrepancies based on DNS data*, Phys. Rev. Fluids, 2 (2017), 034603.
[86] P. Welch, *The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms*, IEEE Trans. Acoust. Speech, 15 (1967), pp. 70–73.
[87] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Vol. 2. Springer Science & Business Media, 2003.
[88] M. O. Williams, C. W. Rowley, and I. Kevrekidis, *A kernel-based method for data-driven Koopman spectral analysis*, J. Comput. Dynam., 2 (2015), pp. 247–265.
[89] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, *A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition*, J. Nonlinear Sci., 25 (2015), pp. 1307–1346.
[90] E. Yeung, S. Kundu, and N. Hodas, *Learning Deep Neural Network Representations for Koopman Operators of Nonlinear Dynamical Systems*, preprint, https://arxiv.org/abs/1708.06850, 2017.
[91] A. Zare, M. R. Jovanović, and T. T. Georgiou, *Colour of turbulence*, J. Fluid Mech., 812 (2017), pp. 636–680.
[92] J. Zech and Y. Marzouk, *Sparse Approximation of Triangular Transports on Bounded Domains*, preprint, https://arxiv.org/abs/2006.06994, 2020.
[93] Z. Zhao and D. Giannakis, *Analog forecasting with dynamics-adapted kernels*, Nonlinearity, 29 (2016), pp. 2888–2939.