




Predicting extreme events in a data-driven model of turbulent shear flow using an atlas of charts

Andrew J. Fox , C. Ricardo Constante-Amores , and Michael D. Graham ^{*}

*Department of Chemical and Biological Engineering,
University of Wisconsin-Madison, Madison, Wisconsin 53706, USA*



(Received 30 January 2023; accepted 14 August 2023; published 5 September 2023)

Dynamical systems with extreme events are difficult to capture with data-driven modeling due to the relative scarcity of data within extreme events compared to the typical dynamics of the system and the strong dependence of the long-time occurrence of extreme events on short-time conditions. A recently developed technique [D. Floryan and M. D. Graham, *Nat. Mach. Intell.* **4**, 1113 (2022)], here denoted as *Charts and Atlases for Non-linear Data-Driven Dynamics on Manifolds*, or CANDyMan, overcomes these difficulties by decomposing the time series into separate charts based on data similarity, learning dynamical models on each chart via individual time-mapping neural networks, then stitching the charts together to create a single atlas to yield a global dynamical model. We apply CANDyMan to a nine-dimensional model of turbulent shear flow between infinite parallel free-slip walls under a sinusoidal body force [J. Moehlis, H. Faisst, and B. Eckhardt, *New J. Phys.* **6**, 56 (2004)], which undergoes extreme events in the form of intermittent quasi-laminarization and long-time full laminarization. The multichart model created by the CANDyMan technique is compared with both a standard data-driven model (i.e., the “single-chart” limit of the CANDyMan method) and a Koopman-based model created through extended dynamic mode decomposition-dictionary learning. We demonstrate that the CANDyMan method allows the trained dynamical models to more accurately forecast the evolution of the model coefficients than both a single-chart model and a Koopman model, reducing the error in the predictions as the model evolves forward in time. The technique exhibits more accurate predictions of extreme events than either a single-chart model or Koopman model, capturing the frequency of quasi-laminarization events and predicting the time until full laminarization more accurately than a single neural network.

DOI: [10.1103/PhysRevFluids.8.094401](https://doi.org/10.1103/PhysRevFluids.8.094401)

I. INTRODUCTION

Real-world dynamical systems often produce unusual behaviors in the form of extreme events. These extreme events are characterized by a dissimilarity to the typical dynamics of the system, usually greater in scope or scale, that occur relatively infrequently compared to the typical dynamics. Common examples include rogue waves in the ocean [1], extreme weather patterns such as hurricanes and tornadoes [2,3], and intermittency in turbulent flows [4]. While extreme events are a consequence of the same dynamical system that governs the nonextreme state, they are often difficult to forecast using data-driven modeling. The relative scarcity of data within extreme events both limits the overall observations of the extreme events on which to train the model and reduces the relative influence of extreme event behavior on data-driven model training. Thus, creating a data-driven model that can accurately capture extreme events remains an active challenge.

*mdgraham@wisc.edu

Recent studies proposed various techniques for analyzing and forecasting the occurrence of extreme events. Guth and Sapsis [5] developed a probabilistic framework for the use of indicator observables as predictors of the extreme events. Ragone and Bouchet [6] supplemented climate model simulations with a rare-event algorithm to examine and more accurately capture the increasing frequency of extreme heatwaves in Europe. Blanchard *et al.* [7] built a machine learning framework to correct a biased climate model to produce better forecasts of extreme events. Mendez and Farazmand [8] applied probabilistic models toward predicting indirect spreading of wildfires by wind to improve forecasts of new wildfire locations. Gomé *et al.* [9] applied a rare-event algorithm to analyze the transition between states in turbulent pressure-driven flow and more efficiently predict passage time between states. While these studies improved predictions of extreme events, they primarily corrected and supplemented the forecasts of existing models; we will instead aim to develop an improved model.

One attractive test case of a dynamical system with extreme events is the nine-dimensional model for turbulent flow developed by Moehlist, Faisst, and Eckhardt (MFE) [10]. The MFE model, an extension of a model by Waleffe [11], governs the evolution of nine amplitudes of combinations of spatial Fourier modes describing an incompressible turbulent shear flow between infinite parallel free-slip walls under a sinusoidal body force. These nine modes provide a minimal description of the mechanisms for self-sustenance in turbulence, allowing the resulting flow field to display realistic turbulent dynamics. In particular, the model displays features consistent with turbulence in the transition region, namely, long periods of turbulent behavior with infrequent quasi-laminarization events (also called quiescent [12] or hibernating [13] intervals) and ultimately full laminarization [12–14]. These quasi- and complete relaminarizations will be the extreme events considered in the present work, in which we use time series from the MFE model as “data” with which to develop a data-driven model.

In recent years, several attempts were made to reproduce the dynamics of the MFE model (and other flow systems) through data-driven techniques based on neural networks (NNs). Neural networks are a powerful data-driven modeling technique that were shown to accurately recreate the dynamics of systems such as the viscous Burgers equation [15], the Kuramoto-Sivashinsky equation [16,17], and Kolmogorov flow [18]. Srinivasan *et al.* [19] developed both feedforward neural networks (FNNs) and long short-term memory (LSTM) networks to recreate the MFE model as discrete-time maps. While the FNNs were unable to reproduce the model, LSTMs were able to accurately reconstruct long-time behaviors of the full-field velocity statistics. This problem was revisited by Eivazi *et al.* [20], where the reconstruction via a LSTM network was compared to predictions generated via a Koopman-operator-inspired framework with nonlinear forcing. In their approach, the observables incorporated time-delay embeddings, and they imposed a nonlinear forcing [21,22]. Their work demonstrated that this framework could reproduce short-time and long-time statistics as well or better than the LSTM networks. (We further discuss the Koopman operator approach to dynamics below, for the moment simply noting that the original Koopman operator formalism is linear and Markovian, neither of which property is exhibited by the methodology of the authors of [20].) Pandey *et al.* [23] introduced the use of reservoir computing in the form of an echo-state network (ESN) to reproduce the MFE model as a discrete-time map, and provided comparisons to both a FNN and a LSTM network. The LSTM network and the ESN were shown to perform similarly, with both adequately capturing the full-field velocity statistics, while again the FNN was shown to perform appreciably worse. Racca and Magri [24] specifically examined the ability of an ESN to forecast the occurrence of an extreme event within a future time window. They determined that their data-driven model could accurately forecast extreme event episodes far into the future without incorrectly predicting false quasi-laminarization events. Pershin *et al.* [25] assessed the ability of an ESN to forecast time until full laminarization. They showed that their model could adequately reproduce the lifetime distribution of the MFE data, correctly predicting the probability of an arbitrary MFE time series remaining in the turbulent state some time in the future. These studies only successfully modeled the MFE equations through the use of non-Markovian models, which forecasted the future state through input of the current and past states. As the MFE model

is itself Markovian, we will instead endeavor to model the MFE data with a Markovian dynamical system.

Specifically, we will use a recently developed method that will be denoted here as *Charts and Atlases for Nonlinear Data-Driven Dynamics on Manifolds* (CANDyMan) [26,27]. CANDyMan operates by decomposing the data distribution in state space into separate regions called charts with a clustering algorithm, learning local dynamical models in each chart using FNNs, then stitching together the charts to create a single atlas containing the global dynamical model. This approach is quite distinct from cluster-based network modeling [28], where clustering is used to construct Markov chains modeling transitions between cluster centroids. Here we are constructing deterministic dynamical systems. This technique was previously applied to dimension reduction problems, accurately learning reduced-order dynamical models whose dimension is equal to the intrinsic dimensionality of the system [26]. The use of multiple charts allows low-dimensional manifolds embedded in high-dimensional space to be broken down into locally low-dimensional structures, capturing the dynamics of a system with the minimal number of dimensions, in a way that single chart methods cannot. Here, we do not perform dimension reduction, but rather utilize the clustering of data to break down the dynamical system into separate regions representing extreme and nonextreme states. By learning the dynamics in the extreme region separately and independently from the nonextreme regions, CANDyMan inherently overcomes the imbalance of extreme versus nonextreme information and thus the limited influence of extreme events in data-driven model training.

Here, we will use CANDyMan to reconstruct the dynamics of the MFE model. A data set containing time series of the MFE amplitudes will be decomposed using k -means clustering into atlases containing between one and five charts. We will train deep neural networks to reconstruct the time evolution of the MFE amplitudes within each of the charts, then stitch them together to create five global models. To assess the accuracy of the models, we will first consider their ability to reconstruct the turbulent flow field. Next, we will analyze their performance in reproducing short-time and long-time statistics, and we will compare our framework with a Koopman framework in which the observable evolves under linear dynamics. Finally, we will assess the extreme event forecasting of the data-driven models by determining the statistical accuracy of forecasting extreme event occurrences and comparing the predicted laminarization lifetime distribution to the true data.

II. FORMULATION

The MFE model is a severely truncated Fourier-Galerkin approximation to the Navier-Stokes equations (NSE) for incompressible flow between two free-slip walls and driven by a spatially sinusoidal body force. The flow is composed of nine combinations of spatial Fourier modes $\mathbf{u}_i(\mathbf{x})$, describing the basic profile, streaks, and vortices, as well as interactions between them. The velocity field at position \mathbf{x} and time t is given by a superposition of the nine modes as $\mathbf{u}(\mathbf{x}, t) = \sum_{i=1}^9 a_i(t) \mathbf{u}_i(\mathbf{x})$. The mode amplitudes $a_i(t)$ satisfy a system of nine ordinary differential equations (ODEs), generated through Galerkin projection, whose explicit form is given in Moehlis *et al.* [10]. Our study considers a domain of size $L_x \times L_y \times L_z$, with infinite, parallel walls at $y = -L_y/2$ and $y = L_y/2$ and periodic boundaries $x = 0$, $x = L_x$, $z = 0$, and $z = L_z$; x , y , and z are the streamwise, wall-normal, and spanwise coordinates, respectively. The domain size of $L_x = 4\pi$, $L_y = 2$, $L_z = 2\pi$ was used, with a channel Reynolds number of 400; these parameters produce turbulent behavior of suitable length for data-driven model development [19].

As training data, we generated 100 unique time series from a fourth-order Runge-Kutta integration of the MFE equation, with a time step of 0.5. Each time series encompasses the transient turbulent state, consisting of turbulent intervals interspersed with quasi-laminarization events, with terminal laminarization occurring at long times. We will often characterize the flow using the *total* kinetic energy (KE), given by $\text{KE} = \frac{1}{2} \sum_{i=1}^9 a_i^2$. Therefore, the turbulent state is *low* energy while the laminar is *high* energy. Every time series collapses to the known laminar fixed point $a_i = \delta_{i1}$. To generate the time series, initial conditions of eight of the amplitudes are given as follows:

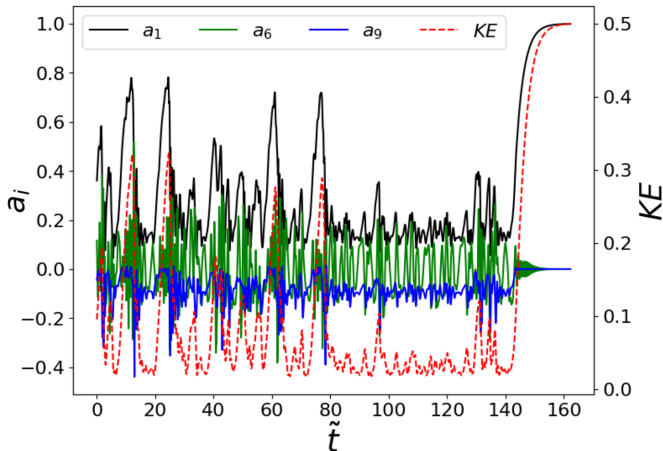


FIG. 1. Evolution of three amplitudes, a_1 , a_6 , a_9 and corresponding kinetic energy from one time series of the MFE data set.

$(a_1, a_2, a_3, a_5, a_6, a_7, a_8, a_9) = (1, 0.07066, -0.07076, 0, 0, 0, 0, 0)$. The initial value of a_4 is arbitrarily generated in the range $[-0.1, 0.1]$. These initial conditions were previously demonstrated to generate chaotic dynamical data with quasi-laminarization events [19]. We will report all results in units $\tilde{t} = t/\tau_L$, where τ_L is the Lyapunov time for the system; in the original nondimensionalization $\tau_L \approx 61$ [24]. The first 1000 time steps of each time series are discarded to eliminate the dependence on initial conditions. Each time series is evolved until the laminar state is reached, taking a varying number of time steps depending on the initial conditions; the resulting training data set consist of approximately two million snapshots of MFE amplitudes and have a mean lifetime of $164\tau_L$. Amplitudes and KE from a randomly chosen time series are shown in Fig. 1.

In this study, we examine the behavior of multichart models with between two and five charts, as well as a standard approach with one global model—the “one-chart” limit of CANDyMan. Here, the one chart refers to a global state-space representation by a single dynamical model, whereas in the multichart models the separate charts create independent local representations of the state space which are then combined to form a single global model. The dynamical system data are first clustered into k charts via k -means clustering, which partitions a data set into k clusters, minimizing the within-cluster variance [29,30]. Other clustering techniques, such as k nearest neighbors [31] or single-linkage clustering [32], could be used, provided the clustering technique produces charts that encompass contiguous regions of the state space. Furthermore, machine learning techniques involving clustering have recently appeared, including chart autoencoders [33], where clustering of the data is learned simultaneously with local dimension reduction, and mixture models of variational autoencoders [34], where clustering is performed with mixture models. We selected k -means clustering for this study due to its simplicity for implementation and its previous success in modeling dynamical systems with the CANDyMan method [26]. The clusters are then augmented so that they overlap by locating the k_{NN} nearest neighbors to each data point in a cluster by Euclidean distance and adding these to the original cluster. This creates overlap regions between neighboring clusters, providing transition regions in which the dynamics are described in multiple charts and allowing for the movement into and out of the region to be handled by the separate local models.

Then, in each augmented chart, we generated discrete-time models of the form $a^{(j)}(t + \tau) = F^{(j)}(a^{(j)}(t); \theta^{(j)})$, where $a^{(j)}(t) \in \mathbb{R}^9$ is the representation of the state in chart j , the discrete time step is $\tau = 0.5$, and $F^{(j)}$ is the corresponding discrete-time map, which takes the form of a FNN. The quantities $\theta^{(j)}$ are the neural network weights for $F^{(j)}$, which are learned from the data using a standard stochastic gradient descent method and trained to minimize the loss function $L^{(j)} = \langle \|a^{(j)}(t) - \tilde{a}^{(j)}(t)\|_2 \rangle$, where $\langle \cdot \rangle$ is the average over the training data. To ensure that the

comparison between different numbers of charts was standardized, each global model contains the same number of total neurons $N_T = 1800$; a system of k charts would then use $N_N = N_T/k$ neurons in each local model, each containing four fully connected hidden layers of $N_N/6$, $N_N/3$, $N_N/3$, and $N_N/6$ of the total number of local neurons, respectively. Increasing the total number of neurons beyond $N_T = 1800$ was not found to significantly impact the performance of the data-driven models. Each neural network was trained using a learning-rate scheduler with an initial learning rate of 0.01, decaying at a rate of 0.9 every 2000 steps. Each model was then trained for 100 epochs, which was found to accurately reproduce the training data while avoiding overfitting. The computational cost of training global models containing multiple local models is no greater than that of training a single global model, as the total trainable parameters, total amount of training data, and training procedure is held constant between the different models; additionally, the multichart models could potentially be trained in less real time by training models in each chart in parallel.

Finally, we briefly describe the methodology used for the Koopman predictions. For a Markovian autonomous deterministic dynamical system with state $\mathbf{a}(t)$, the Koopman operator \mathcal{K}_τ describes the evolution of an arbitrary observable $G(\mathbf{a})$ from time t to time $t + \tau$: $G[\mathbf{a}(t + \tau)] = \mathcal{K}_\tau G[\mathbf{a}(t)]$ [35,36]. The Koopman operator is linear and time independent, so evolution of the observables of the state can be expressed as a sum (or integral, if \mathcal{K}_τ has a continuous spectrum) of ‘‘Koopman modes’’ with complex-exponential time dependence. The tradeoff for gaining linearity is that \mathcal{K}_τ is also infinite-dimensional, requiring for implementation some finite-dimensional truncation of the space of observables. Here we use the ‘‘extended dynamic mode decomposition-dictionary learning’’ (EDMD-DL) approach [37]. Given a vector of observables $\Psi[\mathbf{a}(t)]$, now there is an approximate matrix-valued Koopman operator \mathbf{K} such that the evolution of observables is approximated by $\Psi[\mathbf{a}(t + \tau)] = \mathbf{K}\Psi[\mathbf{a}(t)]$. The EDMD-DL approach aims to simultaneously learn the operator \mathbf{K} and the best set of observables $\Psi[\mathbf{a}(t)]$, represented as neural networks, to accurately approximate the evolution of the system. The key idea behind finding \mathbf{K} is to determine the linear operator which best maps between corresponding pairs of observables (in a least-squares sense). Given a matrix of observables whose columns are the vector of observables at different times, $\psi(t) = \{ \Psi[\mathbf{a}(t_1)] \ \Psi[\mathbf{a}(t_2)] \ \cdots \}$ and its corresponding matrix at $t + \tau$, $\psi(t + \tau) = \{ \Psi[\mathbf{a}(t_1 + \tau)] \ \Psi[\mathbf{a}(t_2 + \tau)] \ \cdots \}$, the approximate matrix-valued Koopman operator is defined as $\mathbf{K} = \psi(t + \tau)\psi(t)^+$, where the $+$ superscript denotes the pseudo (Moore-Penrose) inverse. Our method relies on automatic differentiation to find \mathbf{K} and the set of observables simultaneously. In this study, we create a set of observables of 100 elements in addition to the state. We select three hidden layers with 100 neurons, each hidden layer is followed by an activation function with ELU. We thoroughly vary the numbers of observables and the neural network architecture finding no difference in the short- and long-time tracking of the results [38].

III. RESULTS AND DISCUSSION

A. Distribution of data into clusters

Insight into the number of charts necessary for properly reconstructing the MFE data can be gained by observing the clustering of the training data set. Figure 2 shows how one trajectory from the data set is partitioned when we use different numbers of charts, in terms of Figs. 2(a) to 2(e) the time series of KE and Figs. 2(f) to 2(j) state-space projections onto amplitudes a_1 , a_6 , a_9 . With two charts, the data partitions into one cluster covering the low-energy (turbulent) nonextreme states and the second containing the high-energy extreme (quasi-laminar, laminarizing) states. When three charts were used, the clusters are further segmented, with one covering the low-energy turbulent state, the second primarily consisting of the transition into quasi-laminarization and laminarization events, and the third consisting mainly of the high-energy components of these events. Clustering into four charts breaks down the low-energy region into two separate clusters that remain relatively distinct. When the data are clustered into five charts, the distinction between the charts in the low-energy turbulent regime decreases and the charts containing the turbulent states are described by increasingly similar centroids.

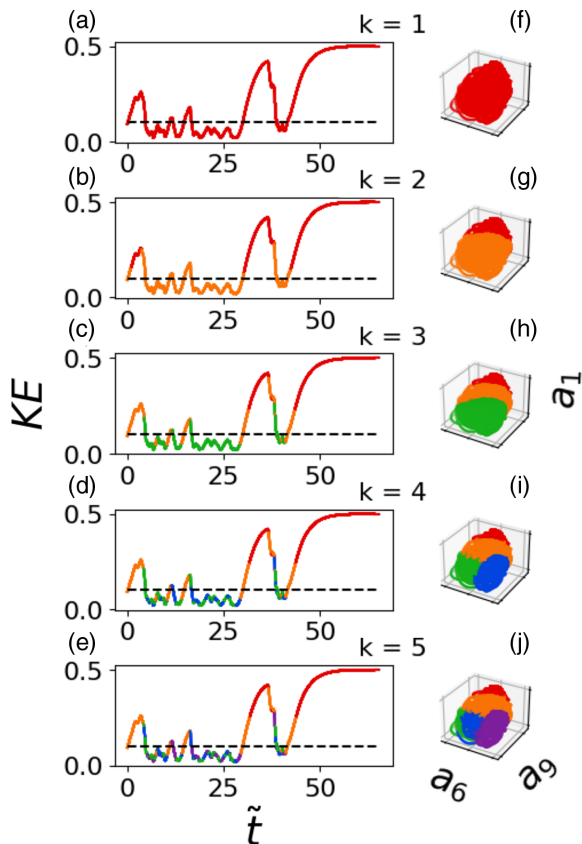


FIG. 2. Clustering of a randomly selected trajectory of (a)–(e) kinetic energy and (f)–(k) the projection of the clustering of the first, sixth, and ninth MFE amplitude for one to five charts, color coded by cluster.

B. Trajectory predictions and time-averaged statistics

The performance of the data-driven models is evaluated on their ability to reconstruct the evolution of the MFE model amplitudes. Two test data sets were generated for comparison between the MFE dynamics and the single- and multichart data-driven models, each with separate and unique initial conditions from the training data. For trajectory predictions, 1000 trajectories of MFE amplitudes were generated from arbitrary initial conditions and time-integrated for 10 Lyapunov times, with the same initial conditions separately evolved forward using the generated data-driven models for the same length of time; this will henceforth be denoted as data set A. The purpose of this data set is to determine the short-time precision of the predictions generated by the single- and multichart models, regardless of any observed or predicted laminarization. With this data set, we will evaluate the similarity of trajectories produced by both the MFE model and the data-driven models from identical initial conditions. For time-averaged statistics, 100 trajectories of MFE amplitudes were generated from random initial conditions and time-integrated for 100 Lyapunov times or until a laminarization event occurred, with the initial conditions separately evolved forward using the generated data-driven models and the same ending criteria; this will henceforth be denoted as data set B. The purpose of this data set is to assess the accuracy of the predicted long-time turbulent state statistics, and as such removes any observed or predicted laminarization. With this data set, we will evaluate the ability of the data-driven models to produce trajectories whose dynamics reside on the same manifold as those produced by the MFE model. As with the creation of the training data, the

first 1000 time steps of each time series following initialization were discarded for both test data sets.

The data-driven models are first evaluated on their ability to reconstruct the velocity statistics of the turbulent regime, the essential function of the MFE model. Using data set B, we project the amplitudes on to the spatial Fourier modes of the MFE model and compare the accuracy of the predicted velocity statistics in the turbulent state to the exact solution. The mean streamwise velocity and Reynolds shear stress were calculated for each data set, as shown in Fig. 3. The mean streamwise velocity and Reynolds shear stress profiles are practically identical to those created by the turbulent portion of the training data. As the figure shows, the single-chart model and the Koopman operator capture well the form of the velocity statistics, but fail to accurately capture the exact values. The three-chart model creates much better predictions, quantitatively capturing the flow profile.

Now we turn to the prediction of trajectories. To quantify the performance of the trajectory predictions, we analyze the data-driven models' ability to accurately forecast the evolution of MFE amplitudes. Using data set A, the error in the predictions $E(t)$ is then calculated for each time series, averaged, and normalized, such that $E(t) = \frac{\|a(t) - \hat{a}(t)\|_2}{D}$. Here, D is the average L_2 -norm between randomly chosen time instants in the turbulent state. Figure 4 shows $E(t)$ for the single- and multichart models and for the Koopman model as a function of time. All FNN models create accurate predictions for $\sim 0.5\tau_L$, with the error remaining close to 0, while the error in the predictions by the Koopman model increases rapidly even at short times. The rapid growth in error is caused by a small number of predictions that evolve rapidly away from the true solution whose large deviations from the true values dominate the error calculation. The error in the predictions of the single-chart model also grows much more rapidly than the multichart models, indicating that the forecasting ability is much stronger in the multichart models. Furthermore, the performance of the multichart models improves as the number of charts increases from two to three, plateaus from three and four, and diminishes from four to five. This indicates that the three charts are sufficient to improve reconstruction via the CANDyMan technique, while further increases in the number of charts can, in fact, impair performance. Possible reasons for this are discussed below. As such, for the remainder of the paper, we will focus our comparison between the single- and three-chart models.

C. Prediction of extreme events

Now we examine the ability of the data-driven model to correctly capture the structure of the extreme events. An extreme event can be identified by a growth in the first MFE amplitude, which represents the mean shear, with a corresponding decrease in the remaining eight amplitudes, which capture the turbulent fluctuations. In Fig. 5(a), we show the joint probability density function (jPDF) of a_1 and a_3 for the reconstruction of the ensemble of trajectories from data set B, which is similar in distribution to that created by the turbulent portion of the training data. The extreme events can be seen as the long tail extending to the right toward the laminar state $a_1 = 1, a_3 = 0$. The prediction of the single-chart model, shown in Fig. 5(b) fails to accurately capture the structure of the extreme events, with the tail almost entirely absent. By contrast, the three-chart model, shown in Fig. 5(c), captures the structure of the extreme events well, accurately reproducing the shape of the joint probability density function. The Koopman model, shown in Fig. 5(d), entirely fails to accurately capture the structure of the joint probability density function. The accuracy of the reconstruction of the jPDFs is quantified by calculating the relative mean squared error (MSE) of the reconstructed jPDF, which is defined as the MSE between the jPDF of test data set and the reconstruction normalized by the MSE between two jPDFs produced by the MFE model from two sets of unique initial conditions. The relative MSE of the reconstruction by the three-chart model is only 1.5, indicating excellent reconstruction of the jPDF. The relative MSEs of the reconstructions by the single-chart model and the Koopman operator, however, are much larger at 24.5 and 463.6, respectively, reflecting the much poorer reconstructions.

We now examine the ability of the single- and multichart models to forecast an extreme event, defined by the kinetic energy of the time series increasing to $KE > 0.1$. To analyze the ability to

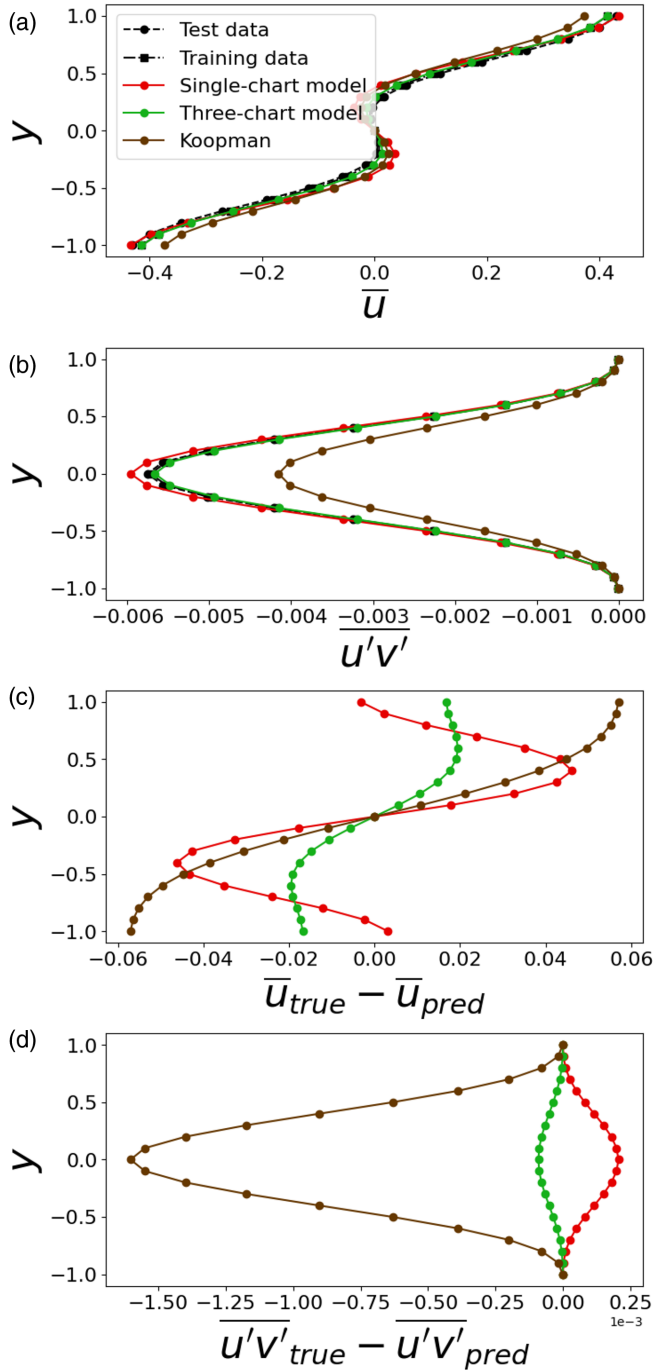


FIG. 3. (a) Mean streamwise velocity and (b) Reynolds shear stress of the full field of the test and training data and of the reconstruction of the MFE model by the single- and three-chart model and by the Koopman model, with (c), (d) the difference between the true values and the predicted values.

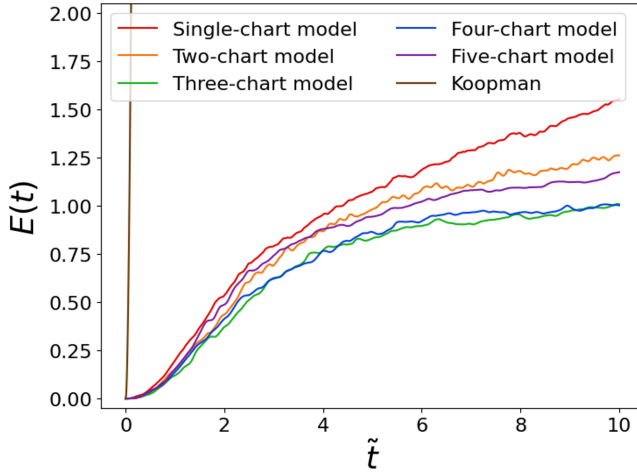


FIG. 4. Ensemble-averaged short-time error tracking of the reconstruction of the MFE model by the single- and multichart models, as well as the reconstruction by the Koopman model.

predict quasi-laminarization events, each time series in data set A, which contains 90 extreme events, is segmented into time windows of duration $0.5\tau_L$ and analyzed for the presence of an extreme event (i.e., KE exceeding 0.1 in the window), where the window at \tilde{t} refers to the time between \tilde{t} and $\tilde{t} + 0.5\tau_L$. The exact solution and data-driven models are then compared to determine if each predict whether an extreme event occurred. If an extreme event occurred in both the exact solution and the model predictions, this is labeled as a *true positive* (*TP*). If the exact solution exhibited an

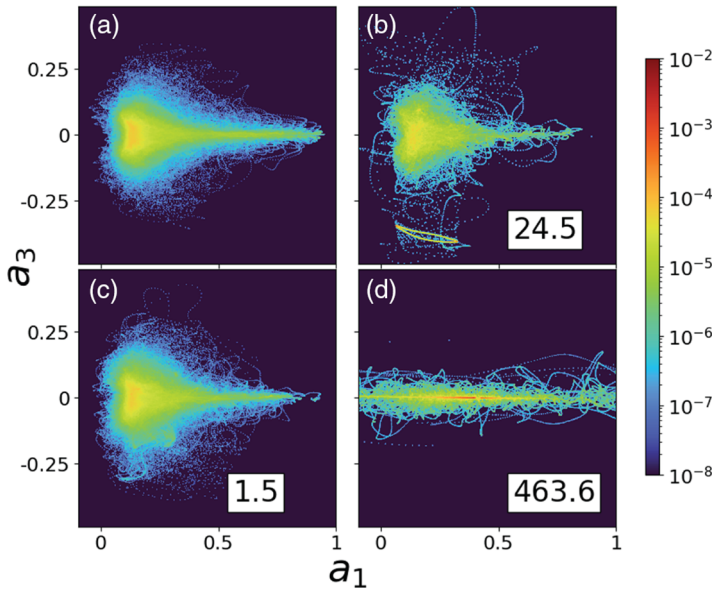


FIG. 5. (a) Joint probability density function (jPDF) of a_1 and a_3 ; (b)–(d) predictions of the MFE model by the single- and three-chart models and by the Koopman model, respectively, with the mean squared error (MSE) of the reconstruction, normalized by the MSE between two jPDFs produced by the MFE model with unique initial conditions, inset. Note the logarithmic scale.

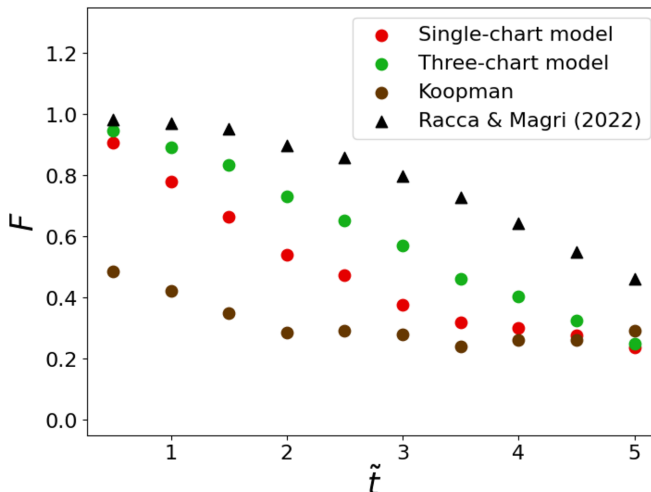


FIG. 6. F-score of extreme event forecasting of the MFE model by the single- and three-chart models and by the Koopman model, with comparison to prior study by Racca and Magri [24].

extreme event, but the data-driven model failed to forecast one, this is labeled as a *false negative* (FN). If the model predicted an extreme event when the exact solution showed none, it is identified as a *false positive* (FP) [24]. The total number of each identification type in each window was tabulated and the F -score, F , was calculated in each window, where $F = (1 + \frac{FP+FN}{2TP})^{-1}$.

Figure 6 shows the F -score as a function of prediction time for the single- and multichart models, as well as a comparison to results from Racca and Magri [24] using an echo-state network; this study did not perform quantitative trajectory comparisons, and as such a similar comparison can not be made for Fig. 4. Both the single- and multichart models produce more accurate forecasts of extreme event occurrences than the Koopman model, with the F -score of the Koopman model falling below 0.5 at all time windows. The multichart model outperforms the single-chart model, more accurately forecasting extreme events at all prediction times. Our multichart model performs similarly to the (non-Markovian) echo-state network developed by Racca and Magri [24] at short times, while the accuracy falls below that of the ESN at longer prediction times.

Finally, we determine the ability of the data-driven models to forecast the lifetime of the turbulence before permanent laminarization. At long times, all time series generated by the MFE model at the given parameters collapse to the laminar fixed point; the lifetime of each time series is dependent on the initial condition, with the probability of remaining in the turbulent state approaching zero at long times. At $Re \lesssim 320$, the probability that a given time series remains in the turbulent state for a duration t , known as the survival function $S(t)$, takes the form [10,25] $S(t; Re) = \exp[-\frac{t-t_0}{\tau_S(Re)}]$, where t_0 is the time delay caused by the approach to the attractor and $1/\tau_S(Re)$ is the Re -dependent decay rate. At $Re \gtrsim 320$, the distribution, particularly at long lifetimes, is known to deviate from an exponential decay, requiring increased time to laminarize.

Here, we define a laminarization event as a high-energy state ($KE > 0.1$) for which the kinetic energy over 1 τ_L levels off. The survival function $S(t)$ is shown in Fig. 7 for the test data set and the one- and three-chart models. The test data set consists of 100 time series of varying lengths and has a mean lifetime of 169 τ_L , within 3% of the mean lifetime of the training data set (164 τ_L); this will henceforth be denoted as data set C. The one-chart model and Koopman model produce poor predictions of the lifetime distribution, vastly underestimating the lifetimes of the turbulent state, with mean lifetimes of 19 τ_L and 30 τ_L , respectively. The three-chart model produces a much more accurate representation of the lifetime distribution. The predicted distribution closely matches the exact solution for \tilde{t} up to about 150, while overestimating the lifetimes at longer times, and

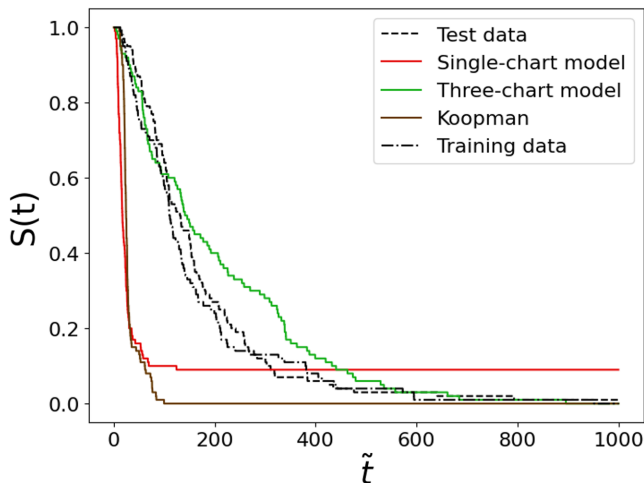


FIG. 7. Lifetime distribution of test and training data and the reconstruction from the MFE model by the single- and three-chart models and by the Koopman model.

predicts an average lifetime of $200 \tau_L$, overestimating the true result by less than 20%. It should be emphasized that we are measuring time here in units of Lyapunov time, so the inaccuracy of $S(t)$ in the three-chart model only arises at extremely long times.

IV. CONCLUSION

In this paper, we applied the CANDyMan [26] technique towards data-driven modeling of a dynamical system with extreme events: the MFE model [10] for turbulent shear flow. We showed that clustering data sets and training multiple local data-driven models allows unique features of distinct data regimes (e.g., extreme events) to be separately and more accurately captured by a multichart global model than in a conventional data-driven model. Thus, multichart models were able to more accurately reproduce the evolution of this system, reducing forecasting error and improving reconstruction of the structure and frequency of extreme events. Importantly, multichart models dramatically improved predictions of extreme event occurrences compared to the single-chart models used previously. While we were not able to fully match the predictive capabilities of the non-Markovian echo state networks in this regard, our Markovian models more accurately complied with the true nature of the underlying system being modeled. Finally, we demonstrated the ability of multichart models to reconstruct the lifetime distribution of turbulent states, accurately predicting the distribution of survival times hundreds of Lyapunov times in the future.

In all cases, we showed that our model outperforms a Koopman approach (EDMD-DL) in which all the observables are evolved linearly, in terms of short- and long-time tracking of the dynamics. While the approach captures key time-averaged quantities such as the mean velocity profile and Reynolds shear stress, its predictions of dynamics are less robust. The performance of the Koopman approach appears consistent with other studies (e.g., [22]), where nonlinear modifications seem to be required to effectively model systems with chaotic dynamics.

An open question in the use of multichart models generated through the CANDyMan technique concerns the number of charts necessary for adequately capturing a dynamical system with extreme events. In this study, we observed that multichart data-driven models, specifically a three-chart model, can forecast one such system more accurately than a single-chart model. Given the clustering of the training data into three charts, shown in Fig. 2, the desired number of charts should be sufficient to separate the extreme from the nonextreme states with a transition region between them. This suggests that a dynamical system with one extreme state, such as the MFE model, would

require at least three charts for optimal forecasting, whereas a dynamical system with multiple types of extreme events, such as weather systems, would necessitate more. We have, however also observed that the improvement in predictive capabilities decreases above a certain number of charts, suggesting that the advantage gained through multichart models is limited and does not increase indefinitely with additional charts. In fact, the performance of our data-driven models decrease above three charts, suggesting that an unnecessarily large number of charts may hinder forecasting. Observing the clustering shown in Fig. 2, it can be seen that, between three and five clusters, the additional clustering does not further segment the extreme and nonextreme states, but rather separates only the nonextreme state into additional regions. Further clustering of the data reduces the total data used to train any individual local data-driven model, potentially decreasing the performance of the local model. In this case, the additional local models describe similar regions of data already well described by a single local model, and as such the decreased training data worsens the individual local models and thereby the global multichart model. Accordingly, the optimal number of charts used to train a dynamical system with extreme events should allow clustering to separate out distinct features, while not further segmenting already distinct regions. Further work is necessary to systematically understand how to best choose the number of charts for a given data set.

Now that we have seen that CANDyMan improved the performance of data-driven models forecasting a low-dimensional dynamical system with extreme events, future investigations should determine its applicability to higher-dimensional systems. As has been previously shown, the use of a charting technique such as CANDyMan allows improved dimension reduction through the use of autoencoder neural networks, capturing the intrinsic dimensionality of dynamical systems [26]. For high-dimensional dynamical systems with intermittency, such as turbulent fluid flows, the application of CANDyMan could not only aid in improved dimension reduction, but also produce more accurate forecasting than conventional single-chart techniques.

A repository containing the PYTHON code used to generate the training data and data-driven models through the CANDyMan technique can be found at [38].

ACKNOWLEDGMENTS

This work was supported by Office of Naval Research N00014-18-1-2865 (Vannevar Bush Faculty Fellowship). We gratefully acknowledge Daniel Floryan for helpful discussions.

-
- [1] K. Dysthe, H. E. Krogstad, and P. Muller, Oceanic rogue waves, *Annu. Rev. Fluid Mech.* **40**, 287 (2008).
 - [2] D. R. Easterling, J. L. Evans, P. Y. Groisman, T. R. Karl, K. E. Kunkel, and P. Ambenje, Observed variability and trends in extreme climate events: A brief review, *Bull. Am. Meteorol. Soc.* **81**, 417 (2000).
 - [3] A. J. Majda, Challenges in climate science and contemporary applied mathematics, *Commun. Pure Appl. Math.* **65**, 920 (2012).
 - [4] N. Platt, L. Sirovich, and N. Fitzmaurice, An investigation of chaotic Kolmogorov flows, *Phys. Fluids* **3**, 681 (1991).
 - [5] S. Guth and T. P. Sapsis, Machine learning predictors of extreme events occurring in complex dynamical systems, *Entropy* **21**, 925 (2019).
 - [6] F. Ragone and F. Bouchet, Rare event algorithm study of extreme warm summers and heatwaves over europe, *Geophys. Res. Lett.* **48**, e2020GL091197 (2021).
 - [7] A. Blanchard, N. Parashar, B. Dodov, C. Lessig, and T. Sapsis, A multi-scale deep learning framework for projecting weather extremes, *arXiv:2210.12137* (2022).
 - [8] A. Mendez and M. Farazmand, Quantifying rare events in spotting: How far do wildfires spread? *Fire Safety J.* **132**, 103630 (2022).

- [9] S. Gomé, L. S. Tuckerman, and D. Barkley, Extreme events in transitional turbulence, *Phil. Trans. R. Soc. A* **380**, 20210036 (2022).
- [10] J. Moehlis, H. Faisst, and B. Eckhardt, A low-dimensional model for turbulent shear flows, *New J. Phys.* **6**, 56 (2004).
- [11] F. Waleffe, On a self-sustaining process in shear flows, *Phys. Fluids* **9**, 883 (1997).
- [12] J. M. Hamilton, J. Kim, and F. Waleffe, Regeneration mechanisms of near-wall turbulence structures, *J. Fluid Mech.* **287**, 317 (1995).
- [13] L. Xi and M. D. Graham, Active and Hibernating Turbulence in Minimal Channel Flow of Newtonian and Polymeric Fluids, *Phys. Rev. Lett.* **104**, 218301 (2010).
- [14] B. Hof, J. Westerweel, T. M. Schneider, and B. Eckhardt, Finite lifetime of turbulence in shear flows, *Nature (London)* **443**, 59 (2006).
- [15] A. J. Linot, J. W. Burby, Q. Tang, P. Balaprakash, M. D. Graham, and R. Maulik, Stabilized neural ordinary differential equations for long-time forecasting of dynamical systems, *J. Comput. Phys.* **474**, 111838 (2023).
- [16] A. J. Linot and M. D. Graham, Deep learning to discover and predict dynamics on an inertial manifold, *Phys. Rev. E* **101**, 062209 (2020).
- [17] A. J. Linot and M. D. Graham, Data-driven reduced-order modeling of spatiotemporal chaos with neural ordinary differential equations, *Chaos* **32**, 073110 (2022).
- [18] C. E. Perez De Jesus and M. D. Graham, Data-driven low-dimensional dynamic model of Kolmogorov flow, *Phys. Rev. Fluids* **8**, 044402 (2023).
- [19] P. A. Srinivasan, L. Guastoni, H. Azizpour, P. Schlatter, and R. Vinuesa, Predictions of turbulent shear flows using deep neural networks, *Phys. Rev. Fluids* **4**, 054603 (2019).
- [20] H. Eivazi, L. Guastoni, P. Schlatter, H. Azizpour, and R. Vinuesa, Recurrent neural networks and Koopman-based frameworks for temporal predictions in a low-order model of turbulence, *Int. J. Heat Fluid Flow* **90**, 108816 (2021).
- [21] M. A. Khodkar, and P. Hassanzadeh, A data-driven, physics-informed framework for forecasting the spatiotemporal evolution of chaotic dynamics with nonlinearities modeled as exogenous forcings, *J. Comp. Phys.* **440**, 110412 (2021).
- [22] 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**, 19 (2017).
- [23] S. Pandey, J. Schumacher, and K. R. Sreenivasan, A perspective on machine learning in turbulent flows, *J. Turbul.* **21**, 567 (2020).
- [24] A. Racca and L. Magri, Data-driven prediction and control of extreme events in a chaotic flow, *Phys. Rev. Fluids* **7**, 104402 (2022).
- [25] A. Pershin, C. Beaume, K. Li, and S. M. Tobias, Training a neural network to predict dynamics it has never seen, *Phys. Rev. E* **107**, 014304 (2023).
- [26] D. Floryan and M. D. Graham, Data-driven discovery of intrinsic dynamics, *Nat. Mach. Intell.* **4**, 1113 (2022).
- [27] D. Floryan and M. D. Graham, “Dfloryan/neural-manifold-dynamics: V1.0 (2022)”, doi: [10.5281/zenodo.7219159](https://doi.org/10.5281/zenodo.7219159).
- [28] D. Fernex, B. R. Noack, and R. Semaan, Cluster-based network modeling-from snapshots to complex dynamical systems, *Sci. Adv.* **7**, eabf5006 (2021).
- [29] J. MacQueen, Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 5.1.*, edited by L. M. L. Cam and J. Neyman (Statistical Laboratory of the University of California, Berkeley, CA, 1967), pp. 281–297.
- [30] E. W. Forgy, Cluster analysis of multivariate data: efficiency versus interpretability of classifications, *Biometrics* **21**, 768 (1965).
- [31] E. Fix and J. L. Hodges, Jr., Discriminatory analysis. nonparametric discrimination: Consistency properties, *Int. Stat. Rev./Revue Int. Stat.* **57**, 238 (1989).
- [32] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, 5th ed., Wiley Series in Probability and Statistics (Wiley-Blackwell, Hoboken, NJ, 2011).

- [33] S. C. Schonscheck, J. Chen, and R. Lai, Chart auto-encoders for manifold structured data, [arXiv:1912.10094](#) (2019).
- [34] G. S. Alberti, J. Hertrich, M. Santacesaria, and S. Sciutto, Manifold learning by mixture models of VAEs for inverse problems, [arXiv:2303.15244](#) (2023).
- [35] B. O. Koopman, Hamiltonian systems and transformation in Hilbert space, [Proc. Natl. Acad. Sci. **17**, 315 \(1931\)](#).
- [36] A. Lasota and M. C. Mackey, *Chaos, Fractals and Noise: Stochastic Aspects of Dynamics*, 2nd ed. (Springer, New York, 1994).
- [37] 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**, 103111 \(2017\)](#).
- [38] A. Fox and C. R. Constante-Amores, “MFE data generation and CANDyMan data-driven model training (Zenodo, 2022)”, doi: [10.5281/zenodo.8092386](#).