Training a neural network to predict dynamics it has never seen

Anton Pershin *
Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, United Kingdom
and School of Mathematics, University of Leeds, Leeds, OX1 3PU United Kingdom

Cédric Beaume, Kuan Li, and Steven M. Tobias
School of Mathematics, University of Leeds, Leeds, LS2 9JT United Kingdom

(Received 12 January 2022; accepted 15 December 2022; published 23 January 2023)

Neural networks have proven to be remarkably successful for a wide range of complicated tasks, from image recognition and object detection to speech recognition and machine translation. One of their successes lies in their ability to predict future dynamics given a suitable training data set. Previous studies have shown how echo state networks (ESNs), a type of recurrent neural networks, can successfully predict both short-term and long-term dynamics of even chaotic systems. This study shows that, remarkably, ESNs can successfully predict dynamical behavior that is qualitatively different from any behavior contained in their training set. Evidence is provided for a fluid dynamics problem where the flow can transition between laminar (ordered) and turbulent (seemingly disordered) regimes. Despite being trained on the turbulent regime only, ESNs are found to predict the existence of laminar behavior. Moreover, the statistics of turbulent-to-laminar and laminar-to-turbulent transitions are also predicted successfully. The utility of ESNs in acting as early-warning generators for transition is discussed. These results are expected to be widely applicable to data-driven modeling of temporal behavior in a range of physical, climate, biological, ecological, and finance models characterized by the presence of tipping points and sudden transitions between several competing states.

DOI: 10.1103/PhysRevE.107.014304

I. INTRODUCTION

Neural networks are important examples of machine learning techniques that have exhibited tremendous promise in the fields of image recognition, computer vision and speech recognition. Their utility stems from their ability to predict known behavior in new situations but how well they can extend this ability beyond their training set remains an open question [1]. An important aspect of this is related to the prediction of previously unseen temporal behavior. This becomes particularly interesting to explore given that neural networks have recently been introduced to assist physical modeling and to time-dependent partial differential equations [2], where the aim is to predict future dynamics without having to solve a computationally expensive set of equations.

Forecasting in dynamical systems is often achieved using a particular class of neural networks known as recurrent neural networks (RNNs) [3]. These are characterised by the presence of feedback connections within the network to allow it to "remember" the history of the dynamical system and to use it to improve the accuracy of its predictions. Among the many different RNN architectures, we focus on echo state networks (ESNs) [4,5] owing to their relatively low training cost; compared with most other RNNs, only one part of the network is trained while the rest is randomly generated and remains fixed [6]. Echo state networks distinguished themselves by making sound short-term (over typically less than 10 natural periods of the system) and long-term predictions in various low-dimensional chaotic models [4,7–9], in the Kuramoto-Sivashinsky equation (which is a partial differential equation) [7,10,11] and in two-dimensional Rayleigh-Bénard convection [12,13]. Trained using a single time series, ESNs can successfully approximate the statistical and geometric properties of chaotic attractors of a dynamical system [7,14,15] and make short-term predictions with the level of accuracy of state-of-the-art techniques for time-series prediction while significantly outperforming them in terms of memory and CPU usage [6,15].

In this paper, we use ESNs to predict sudden transitions in fluids where a laminar (ordered) flow can undergo an instability and become turbulent (seemingly disordered) and vice versa [16]. Systems exhibiting qualitatively similar, bistable regimes are ubiquitous in both natural and engineering applications. Examples include the transport of liquid and gases through pipelines, bioreactors in biochemical engineering, wind turbines and airfoils, as well as climate [17], ecological [18], Earth's magnetic field and geodynamo models [19,20]. Such transitions are often associated with a change in energy consumption or extreme damage which makes their prediction and, crucially, their control important tasks. We demonstrate that a properly trained ESN is capable of predicting the statistics of both laminar-to-turbulent and turbulent-to-laminar transitions even if it has been trained using a time series containing only turbulent dynamics, i.e., it is able to infer laminar dynamics despite not having seen it during

*anton.pershin@physics.ox.ac.uk

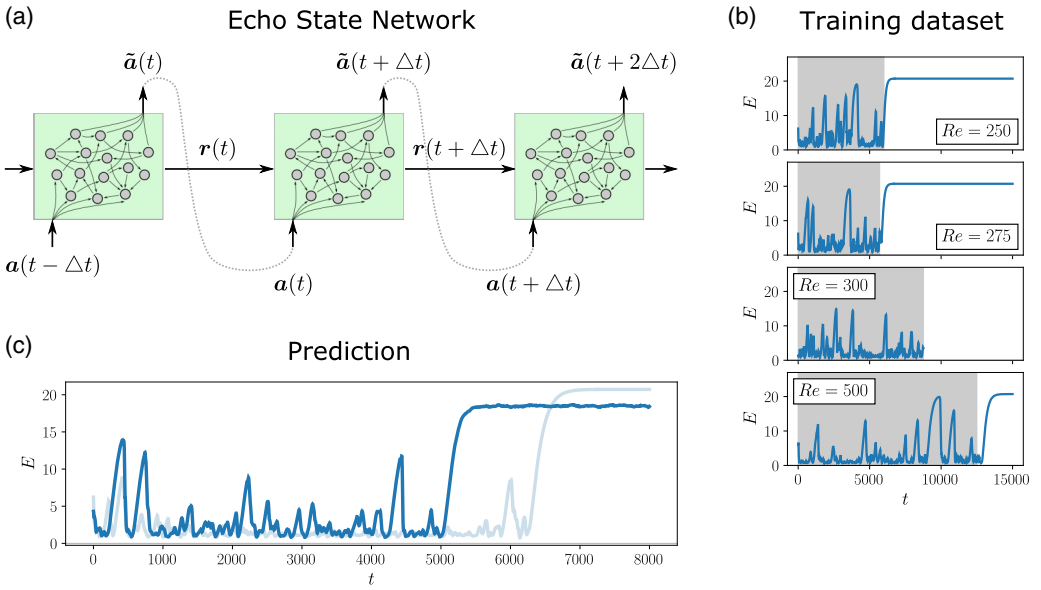

FIG. 1. (a) Schematic of an echo state network. To make a prediction $\tilde{\boldsymbol{u}}(t+\triangle t)$, the flow state $\boldsymbol{u}(t)$ at the previous time step and reservoir state $\boldsymbol{r}(t)$ are passed to the randomly generated reservoir (green box) where they are nonlinearly transformed to yield the prediction. (b) Training time series for Reynolds numbers Re = 250, 275, 300, 500 obtained by time integration of the MFE model and shown in the form of the time evolution of the flow kinetic energy. Only shadowed parts were used for training. (c) Flow prediction made by the echo state network trained at Re = 300 (bright blue curve) and a representative turbulent trajectory of the MFE model computed at the same Reynolds number (light blue curve).

training. As an example of a transitional flow, we consider a paradigm model of plane Couette flow, i.e., the viscous flow between two parallel walls moving in opposite directions at constant and equal velocities. This model is a representative of a wide class of nonlinear dynamical systems exhibiting finite-amplitude instabilities and spontaneous decay of chaotic dynamics. As such, we expect our conclusions to be transferable to a variety of systems with similar dynamical features.

II. MODEL

In plane Couette flow, the velocity field at position $\boldsymbol{x}$ and time t, $\boldsymbol{u}(\boldsymbol{x},t)$, is generally solved for via the integration of the Navier-Stokes equation together with the incompressibility condition, no-slip boundary conditions in the wall-normal direction and spatial periodicity conditions in the streamwise x and spanwise z directions. This set of equations can be reduced to the Moehlis-Faisst-Eckhardt (MFE) model [21] by replacing plane Couette flow with a sinusoidal shear flow, known as the Waleffe flow [22,23], and truncating to nine Fourier-based modes, $\boldsymbol{u}_j(\boldsymbol{x})$, as listed in Appendix A. The fluid velocity can reconstructed via

$$\boldsymbol{u}(\boldsymbol{x},t) = \sum_{j=1}^{9} a_j(t)\boldsymbol{u}_j(\boldsymbol{x}), \tag{1}$$

where the time-dependent amplitudes are $\boldsymbol{a}(t) = [a_1(t),\ldots,a_9(t)]$. The nine-dimensional ordinary differential equation system of coupled amplitude equations obtained by projecting the Waleffe flow equations onto these modes reads

$$\frac{d}{dt}a_j = \delta_{1j}\frac{\pi^2}{4\mathrm{Re}} + \alpha_j(\mathrm{Re})a_j + \sum_{k=1}^{9}\sum_{l=1}^{9}\beta_{jkl}(\mathrm{Re})a_k a_l, \tag{2}$$

where Re is the Reynolds number, δ_{ij} is the Kronecker delta acting on indices i and j, and $\alpha_j(\mathrm{Re})$ and $\beta_{jkl}(\mathrm{Re})$ are Re-dependent coefficients whose full expressions are given in Appendix A. The Reynolds number is the only nondimensional physical parameter in this system. It is a measure of the ratio between inertial and viscous forces. To obtain numerical solutions of this model, we time integrate Eq. (2) using the fourth-order Runge-Kutta scheme with time step $\triangle t_{TI} = 10^{-3}$.

The only known stable solution of (2) is the steady laminar flow: $\boldsymbol{a}_{\mathrm{lam}} = [1,0,\ldots,0]^T$, which is equivalent to $\boldsymbol{u}_{\mathrm{lam}} = \sqrt{2}\sin(\pi y/2)\boldsymbol{e}_x$ in physical space; here $\boldsymbol{e}_x$ is the unit vector in the x direction. Despite the linear stability of the laminar flow, we can observe long-lived turbulence for $\mathrm{Re} \gtrsim 150$ [21]. Examples of turbulent flows at different values of the Reynolds numbers are shown in Fig. 1(b) through time series of the kinetic energy:

$$E = \frac{1}{2}||\boldsymbol{u}||_2^2 = \Gamma_x\Gamma_z\sum_{j=1}^{9}a_j^2, \tag{3}$$

where $\Gamma_x = 1.75\pi$ (resp. $\Gamma_z = 1.2\pi$) is the imposed solution wavelength in the x (resp. z) direction. All our simulations display chaotic dynamics over thousands of time units but eventually relax to the laminar flow, which is expected to be the global attractor at least for $\mathrm{Re} \lesssim 335$ [24]. This phenomenon, called hereafter turbulent-to-laminar transition,

is a prominent feature of transitional shear flows [16]. The opposite process of laminar-to-turbulent transition is equally important both from a theoretical and a practical viewpoint. In this study, we show that statistical features associated with both laminar-to-turbulent and turbulent-to-laminar transitions can be successfully predicted by an ESN trained solely on a transient segment of a turbulent trajectory, i.e., with no experience of laminarization.

III. METHOD

Echo state networks (ESNs) belong to a class of artificial recurrent neural networks (RNNs) that is characterized by the presence of internal feedback connections in their architecture allowing the network to have its own "memory" and, thereby, generate time series with a greater accuracy compared with its nonrecurrent companions. Figure 1(a) shows a schematic representation of a typical RNN architecture that takes the flow state $\boldsymbol{a}(t) \in \mathbb{R}^{N_a}$ at time t as an input, where $N_a = 9$ for the MFE model, and outputs the prediction of the flow state at time $t + \triangle t$, $\tilde{\boldsymbol{a}}(t + \triangle t)$, where we used $\triangle t = 1$ throughout this study. It should be noted that, in our case, the training data and the ESN prediction time steps are significantly larger than the time-integration time step $\triangle t \gg \triangle t_{TI} = 10^{-3}$. In addition to the input flow state, the RNN uses its own internal state $\boldsymbol{r}(t) \in \mathbb{R}^{N_r}$, which is also updated. In the context of ESNs, $\boldsymbol{r}(t)$ is called the reservoir state. The ESN prediction is done in two stages. First, the reservoir state $\boldsymbol{r}(t)$ and the input flow state $\boldsymbol{a}(t)$ are nonlinearly transformed to get the reservoir state at time $t + \triangle t$:

$$\boldsymbol{r}(t + \triangle t) = \tanh\left[\boldsymbol{b} + \boldsymbol{W}\boldsymbol{r}(t) + \boldsymbol{W}_{\text{in}}\boldsymbol{a}(t)\right] + \xi \boldsymbol{Z}, \tag{4}$$

where $\boldsymbol{W}$ and $\boldsymbol{W}_{\text{in}}$ are fixed $N_r \times N_r$ and $N_r \times N_a$ weight matrices, $\boldsymbol{b}$ is a fixed N_r-dimensional bias vector, $\boldsymbol{Z}$ is a random vector uniformly distributed between -0.5 and 0.5 and ξ is a hyperparameter controlling the amplitude of the additive noise. Second, the reservoir state is mapped back into the flow space via the linear transformation:

$$\tilde{\boldsymbol{a}}(t + \triangle t) = \boldsymbol{W}_{\text{out}} \begin{bmatrix} \boldsymbol{r}(t + \triangle t) \\ 1 \end{bmatrix}, \tag{5}$$

where $\boldsymbol{W}_{\text{out}}$ is an $N_a \times (N_r + 1)$ weight matrix. The result, $\tilde{\boldsymbol{a}}(t + \triangle t)$, is called the prediction. The network is trained so that the prediction approximates the true flow state $\boldsymbol{a}(t + \triangle t)$ as accurately as possible. Note that the addition of the unit component in the right-hand-side vector of Eq. (5) is intended to create a bias and improve the performance of the ESN.

The characteristics that distinguish ESNs from the vast majority of other RNN architectures are that the weight matrices $\boldsymbol{W}$ and $\boldsymbol{W}_{\text{in}}$ and the bias term $\boldsymbol{b}$ are initialized randomly and remain fixed, i.e., they are not trained, and that the weight matrices are often chosen to be sparse, resulting in a sparsely connected network (see Appendix B for details). This greatly simplifies the training process, which becomes equivalent to solving the linear regression problem:

$$\min_{\boldsymbol{W}_{\text{out}}} \sum_{k=1}^{N_t} ||\boldsymbol{W}_{\text{out}}\boldsymbol{r}(k\triangle t) - \boldsymbol{a}(k\triangle t)||_2^2, \tag{6}$$

where it is assumed that the training dataset is composed of $N_t + 1$ flow states $\boldsymbol{a}(t)$ known at times $t = 0, \triangle t, 2\triangle t, \ldots, N_t \triangle t$. The flow state at $t = 0$ is used as an initial condition only to compute the first prediction $\tilde{\boldsymbol{a}}(\triangle t)$. This minimization problem possesses a closed-form solution given by the normal equation:

$$\boldsymbol{W}_{\text{out}}^T = (\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{A}, \tag{7}$$

where matrix $\boldsymbol{R} \in \mathbb{R}^{N_t \times (N_r+1)}$ is made of vectors $\boldsymbol{r}(\triangle t), \boldsymbol{r}(2\triangle t), \ldots, \boldsymbol{r}(N_t\triangle t)$ and an all-ones vector and $\boldsymbol{A} \in \mathbb{R}^{N_t \times N_a}$ is made of vectors $\boldsymbol{a}(\triangle t), \boldsymbol{a}(2\triangle t), \ldots, \boldsymbol{a}(N_t\triangle t)$. We wish to emphasize two modifications which differentiate our ESN architecture from more standard alternatives found in the literature. The first one is the presence of a random bias term $\boldsymbol{b}$ in Eq. (4), which significantly improves the accuracy of predictions in our case. The second one is the presence of additive noise in the same equation which is introduced to regularize the regression problem and, at the same time, improve the stability of our ESN [5].

The aforementioned architecture involves several hyperparameters: the reservoir state dimension N_r, the spectral radius $\rho(\boldsymbol{W})$ of matrix $\boldsymbol{W}$, its sparsity s, and the noise amplitude ξ. Though we did not perform an exhaustive search of optimal hyperparameter values, several points need to be highlighted. The success of ESNs relies on a high-dimensional reservoir space whose dimension N_r is expected to be much higher than that of the flow state. Consequently, we chose $N_r = 1500$. We also fixed the noise amplitude $\xi = 10^{-3}$ and the spectral radius $\rho(\boldsymbol{W}) = 0.5$, values that allowed us to minimize the expression (B3). Surprisingly, we found a weak dependence of the quality of prediction on s and used $s = 0.5$ for $\text{Re} = 250$ and $s = 0.9$ for other Reynolds numbers. See Appendix B for further details of the hyperparameter search.

To make predictions of the flow state, we simply replace $\boldsymbol{a}$ in (4) with $\tilde{\boldsymbol{a}}$, which is equivalent to activating one more feedback connection [gray dotted line in Fig. 1(a)]. This makes (4) and (5) a closed system of recurrent equations which only requires initial conditions $\tilde{\boldsymbol{a}}(0)$ and $\boldsymbol{r}(0)$. Since the initial reservoir state $\boldsymbol{r}(0)$ is not known in advance, we must determine it through a process termed synchronization. We take a small number of states from the recent flow history, $\boldsymbol{a}(-9\triangle t), \boldsymbol{a}(-8\triangle t), \ldots, \boldsymbol{a}(-\triangle t)$, and subsequently generate reservoir states $\boldsymbol{r}(-8\triangle t), \boldsymbol{r}(-7\triangle t), \ldots, \boldsymbol{r}(0)$ using Eq. (4) and a trivial initial condition: $\boldsymbol{r}(-9\triangle t) = \boldsymbol{0}$. At the end of this synchronization process, we obtain the required initial reservoir state $\boldsymbol{r}(0)$ to predict the flow dynamics following the procedure described above.

IV. RESULTS

Here we provide evidence that ESNs are able to predict laminar dynamics without having observed it before. We first generate transient turbulent trajectories by time integrating random initial conditions using a fourth-order Runge-Kutta scheme applied to (2) with time step 10^{-3}. One such trajectory is generated for each of the following Reynolds numbers: $\text{Re} = 250, 275, 300$, and 500 [see Fig. 1(b)]. As these simulations eventually relax to the laminar flow ($E \approx 20.7$), we selected the training set to comprise only turbulent dynamics,

as shown by the shaded regions in Fig. 1(b). For each of these training sets, we trained one ESN, which we then identify using their Reynolds number. For ease of reference, we call *truth* the results produced by the MFE model and *prediction* those computed by the ESN.

Each of the trained ESNs is able to generate turbulent trajectories whose statistical properties are similar to those of the original model. This fact has already been established in [9], where ESNs were used to predict statistical properties and extreme events associated with the dynamics of the MFE model. In this paper, we show that ESNs are able to perform the more difficult task of predicting laminarization, i.e., the decay process of a turbulent trajectory towards a laminar state, despite having been solely trained on turbulent trajectories. One such prediction is shown in Fig. 1(c) as an example. At $t \approx 5000$, the predicted flow (dark blue curve) terminates its low-energy chaotic oscillations to relax to a higher energy behavior with only weak temporal variations attributed to the presence of small-amplitude noise in Eq. (4). This is similar to the true laminar state, located at $E \approx 20.7$.

Despite the difference between the true and the predicted laminar flow, it is particularly noteworthy that the ESN is able to predict laminarization, a transition to which was not exposed during training. We obtained a similar prediction, but without temporal oscillations, if we turned off the noise in Eq. (4) while proceeding to the prediction step. However, we found that the presence of noise leads to better prediction of the transition statistics, so we kept using noise throughout this study. It is also important to emphasize that the transition occurring in our prediction at $t \approx 5000$ is different from the "amplitude death" phenomenon, i.e., a sudden collapse of oscillatory or chaotic dynamics caused by a parameter shift, since our main parameter Re is not changed dynamically. The amplitude death phenomenon has recently been shown to be well replicated by ESNs [25].

As we shall see in the next sections, ESNs are capable of more surprising predictions. First, we demonstrate their ability to learn turbulent-to-laminar transition by showing that ESNs can successfully recover the distribution of lifetimes of turbulent trajectories [26]. Additionally, we provide evidence of their ability to make short-term probabilistic predictions of transitional events. This paves the way for their use as generators of early-warning signals of critical transitions [27]. Finally, we examine the opposite kind of transition, laminar-to-turbulent transition, and show that ESNs can be used to approximate the transition probability, one of the key statistics associated with this type of instability [28].

A. Turbulent-to-laminar transition

Turbulent-to-laminar transitions are often characterized using statistical tools similar to the survival function $S(t) = P(T \geqslant t)$, which represents the probability that turbulent behavior remains observed for a duration t or, equivalently, that the time T at which the laminarization event eventually takes place is larger than t [16,21,26]. Within the context of the MFE model, for $\mathrm{Re} \lesssim 300$, this

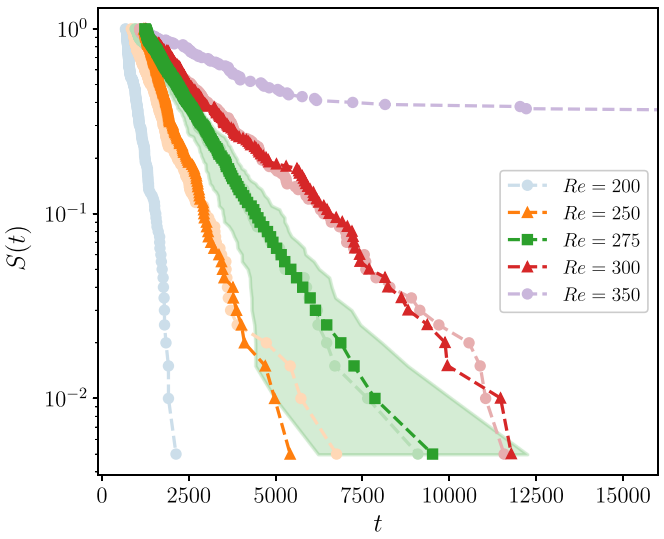

FIG. 2. Lifetime distributions for Reynolds numbers from $\mathrm{Re} = 200$ to $\mathrm{Re} = 350$ shown in the form of survival functions. Dark (resp. light) colors correspond to the distributions generated by echo state networks (resp. MFE model). At $\mathrm{Re} = 275$, the ESN result was constructed based on a distribution of survival functions computed for 100 ESNs: the squares and the dashed curve denote the median values, and the shaded area denotes the 80% confidence band.

distribution takes the form [21]:

$$S(t;\mathrm{Re}) = \exp\left[\frac{t - t_0}{\tau(\mathrm{Re})}\right], \tag{8}$$

where t_0 is the time taken for the initial condition to approach the turbulent saddle and $1/\tau(\mathrm{Re})$ is the Re-dependent escape rate. To build the lifetime distribution for the original MFE model at a fixed value of the Reynolds number, we time-integrate 200 random initial conditions generated by drawing initial amplitudes $a_j(0)$ from the uniform distribution with support $[-1; 1]$ such that the kinetic energy of any initial condition is equal to $E = 0.3\Gamma_x\Gamma_z$, as described in [21]. The lifetime T is measured for each of these initial conditions in the following way: we assume that a laminarization event has taken place and, thus, record T if the total kinetic energy of the flow $E > 15$ from time $T - 1000$ to time T.

This procedure is used for $\mathrm{Re} = 200, 250, 275, 300, 350$ and the resulting survival functions are shown in Fig. 2 (light colors). For $\mathrm{Re} = 350$ and beyond, the lifetime distribution does not follow law (8), as was already observed in [21], so we did not investigate such values of the Reynolds number. To provide matching ESN predictions, we used the same initial conditions, augmented by the first nine time steps of the associated time-integration that we use for synchronization. The resulting predictions are shown by the dark color curves in Fig. 2. We did not obtain results for $\mathrm{Re} = 200$ owing to the fact that laminarization occurs too soon to generate a sufficiently long laminarization-free time series for training. To demonstrate the level of sensitivity of our results to the randomness inherent to the ESN generation process, we additionally computed an empirical distribution of survival functions at $\mathrm{Re} = 275$ by sampling 100 ESNs and building a survival function for each of them. The median

TABLE I. Maximum likelihood estimates of parameters t_0 and $\tau(\mathrm{Re})$ of the exponential distribution (8) approximating lifetime distributions computed for both the MFE model and ESNs.

	t_0		$\tau(\mathrm{Re})$		
Re	Truth	Prediction	Truth	Prediction	Relative error in $\tau(\mathrm{Re})$
250	845	1207	835	734	0.121
275	956	1222	1202	1249	0.169
300	1086	1248	2161	2089	0.033

and 80% confidence band of the empirical distribution are shown in Fig. 2 with green squares and green shadowed area, respectively.

The ESN predictions are excellent: they preserve the main qualitative feature of the true distributions, their exponential structure, implying that the memoryless nature of the laminarization process has been adequately learned. Furthermore, the escape rate of these survival laws, $1/\tau(\mathrm{Re})$, is also well predicted as one can observe in Table I. This is a surprising result given that our ESNs had not seen any laminarization event during training. Moreover, the ESN predictions are fairly robust to random choices of reservoirs which can be concluded from the small spread of the empirical distribution at $\mathrm{Re} = 275$ and the median overlapping the true survival function. A relatively large spread of the empirical distribution for $S(t) \lesssim 4 \times 10^{-1}$ does not change this conclusion since it is mainly explained by the increasing statistical uncertainty inevitably taking place when estimating the distribution tail with modest-size samples.

Interestingly, ESNs have recently been shown to successfully replicate a similar type of distributions which statistically describe random transitions between laminar and chaotic states in the Navier-Stokes equation [29]. However, in contrast to our work, time series used for training there contained a relatively small number of transitions.

B. Early warning of turbulent-to-laminar transition

Lifetime distributions, such as those considered above, are used to predict statistics about the long-term behavior of the system. In many cases, however, it is a short-term prediction that is of interest, like that of critical transitions in, for example, climate [30,31], geophysical [32], ecological [33], and many other complex nonlinear systems [27,34].

In the transitional flow problem considered here, we may want to determine whether a given turbulent flow will laminarize within a relatively short time window, e.g., $T = 2000$. In the case of a deterministic system, such as the MFE model, it is sufficient to time integrate a given initial condition to learn whether the laminarization event occurs. Our ESN, in contrast, is a stochastic model by design owing to the presence of the noise term in (4) and, thus, does not need any alteration or the creation of any additional perturbation to assess the probability of turbulent-to-laminar transition within a given time window. To compute this probability, we perform ensemble predictions, where each prediction within the ensemble starts from the same initial condition but, due to the noise, evolves differently from other predictions. The probability of turbulent-to-laminar transition is then computed as the fraction of these predictions leading to laminarization.

We start by exploring the average level of "leakiness" of the turbulent saddle generated by the ESN at $\mathrm{Re} = 500$, i.e., the average probability that a typical turbulent flow suddenly laminarizes given a short sequence of its previous states. This value will act as a reference for future predictions. To make such a measurement using the ESN, we pick $N = 100$ random states $\boldsymbol{a}(t_j)$, $j = 1, \ldots, N$, from a time series, which was not previously used. We make ensemble predictions for each of these random states after synchronizing the ESN using the previous nine time steps $\boldsymbol{a}(t_j - 9\triangle t), \boldsymbol{a}(t_j - 8\triangle t), \ldots, \boldsymbol{a}(t_j)$. For each ensemble member, we predict the next 2000 time units using the ESN. Each of the predictions is then classified as either exhibiting turbulent-to-laminar transition or not. The probability of turbulent-to-laminar transition, denoted as $P_{\mathrm{T}\to\mathrm{L}}(t_j)$, is then estimated as a fraction of laminarizing trajectories. The average probability of turbulent-to-laminar transition, which we will refer to as the reference probability P_{ref}, is obtained by averaging $P_{\mathrm{T}\to\mathrm{L}}(t_j)$ with respect to t_j: $P_{\mathrm{ref}} \approx 0.11$. We can then generate an early warning of turbulent-to-laminar transition whenever the probability $P_{\mathrm{T}\to\mathrm{L}}(t)$ takes significantly larger values than P_{ref}.

We demonstrate that ESNs are able to act as generators of early warning signals by making probabilistic predictions of the $\mathrm{Re} = 500$ transition shown in Fig. 1(b), starting approximately at $t \approx 14\,000$. To that aim, we use the part of the time series preceding the transition but not included in the training set (small unshaded part in the figure). We expect that the probabilities of turbulent-to-laminar transition $P_{\mathrm{T}\to\mathrm{L}}(t)$ estimated by the ESN become higher as t approaches the transition point. To verify this, we pick five initial states at times $t_j = 13\,840 + 100(j-1)$, where $j = 1, \ldots, 5$, and compute the corresponding probabilities of turbulent-to-laminar transition $P_{\mathrm{T}\to\mathrm{L}}(t_j)$ using the ESN trained at $\mathrm{Re} = 500$ and $N = 100$ ensemble members for each initial state. As required by the synchronization procedure, we also use nine flow states prior to each given initial state. The probability of turbulent-to-laminar transition is then computed using exactly the same ensemble-based algorithm as we used to compute the reference probability P_{ref}. The resulting probabilities together with a small selection of predictions generated by ensemble members are shown in Fig. 3. These results are in line with our expectations. The initial probability prediction is 0.13 at $t_1 = 13\,840$ (top panel in Fig. 3), a value comparable to the reference value $P_{\mathrm{ref}} \approx 0.11$ and thereby implying a low likelihood of transition to the laminar state. At slightly later initial time $t_2 = 13\,940$, the probability of turbulent-to-laminar transition jumps up to a significantly larger value, 0.38, which

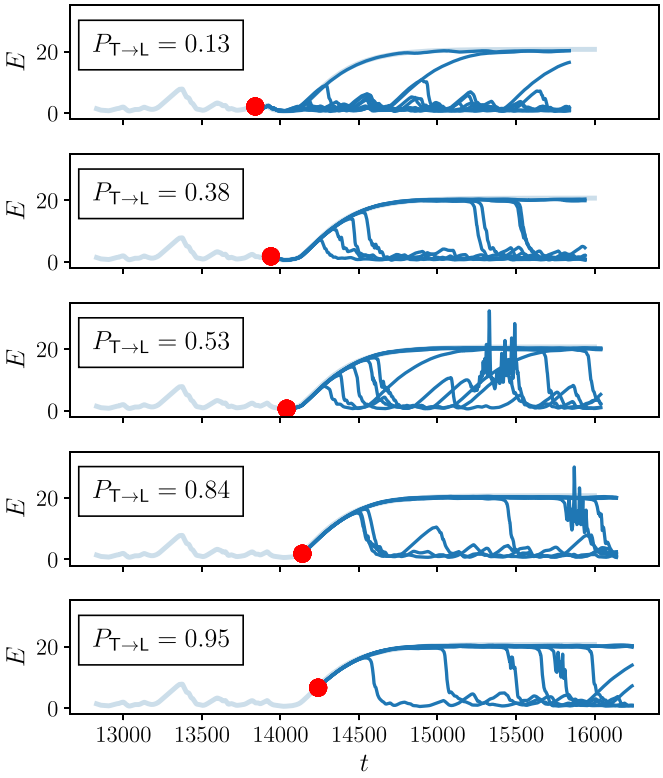

FIG. 3. Prediction of turbulent-to-laminar transition at Re = 500 based on the calculation of the probability of turbulent-to-laminar transition $P_{T\to L}$ which is estimated using an ensemble approach for five initial states (red dots) at times $t_j = 13\,840 + 100(j-1)$, where $j = 1, \ldots, 5$. Every 10th ensemble member of the prediction generated by the echo state network is plotted in bright blue. The true flow trajectory obtained by time integration of the MFE model is plotted in light blue.

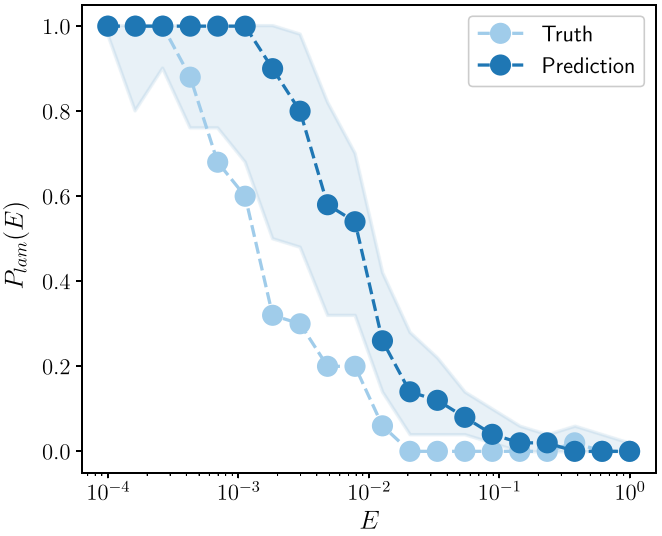

FIG. 4. Laminarization probability as a function of the kinetic energy of random perturbations plotted for the MFE model (light blue curve) and the ensemble of 100 echo state networks (bright blue curve denotes the median, light blue shadowed area denotes the interdecile range) at Re = 500.

can already be considered as an early warning. The predicted probability keeps increasing as we get closer to the actual transition, thereby confirming that this measure can indeed act as an early warning.

C. Laminar-to-turbulent transition

The transition from turbulence to laminar flow does not follow similar dynamical processes to its reciprocal laminar-to-turbulent transition. While the former is a sudden escape from a turbulent saddle (i.e., not an attractor), transition to turbulence is a finite-amplitude instability: the laminar flow is linearly stable, so a sufficiently large perturbation is necessary to trigger transition to turbulence. In this section, we also show that ESNs can be used to predict this transition.

To characterize the transition from laminar flow to turbulence statistically, it is convenient to introduce the laminarization probability $P_{\text{lam}}(E)$, which is the probability that a random perturbation to the laminar flow decays as a function of its kinetic energy E [28]. The laminarization probability is related to the relative volume of the basin of attraction of the laminar flow and, therefore, to the notion of basin stability [18].

We compute the laminarization probability for 20 different values of the kinetic energy of perturbations evenly spaced in the logarithmic scale: $E_1 = 10^{-4}, \ldots, E_{20} = 1$. For each energy level E_j, we generate 50 random perturbations by drawing $a_k(0)$, $k = 1, \ldots, 9$, from the uniform distribution with support $[-1; 1]$ and scale them such that the kinetic energy of perturbations is equal to E_j. We then time integrate the MFE model starting from each of the generated perturbations for 300 time units and record transition to turbulence if E reaches values lower than 10 within this time window. The laminarization probability $P_{\text{lam}}(E_j)$ is then approximated as the fraction of random perturbations which do not lead to transition to turbulence. We used the same procedure to estimate the laminarization probability using the ESN trained on the turbulent state, except that each perturbation is time advanced for 10 time steps using the MFE model to provide sufficient data for synchronization. To provide a statistical confirmation of the ESN ability to learn the laminarization probability, statistically, we estimated $P_{\text{lam}}(E_j)$ for 100 randomly generated ESNs thereby generating an empirical distribution of laminarization probability curves.

The resulting dependence of $P_{\text{lam}}(E)$ on the perturbation kinetic energy E for Re = 500 is shown in Fig. 4 for both the truth and prediction. The laminarization probability of the original model almost monotonically decreases with E. It tends to 1 for small perturbation energies (the laminar flow is linearly stable) and we found that $P_{\text{lam}}(E) = 0$ for $E \gtrsim 2 \times 10^{-2}$, indicating that all the perturbations beyond this energy trigger transition to turbulence. The ESN prediction exhibits the same trend and compares qualitatively well with the truth, showing that ESNs are also capable of learning the statistical boundaries of the basin of attraction of the laminar flow. It is, in fact, remarkable that the ESN can successfully estimate the threshold for laminar-to-turbulent transition *despite having only been trained on fully turbulent time series.* Despite these qualitatively striking predictions, the ESN does somewhat overestimate the laminarization probability for

intermediate perturbation energies, thereby overestimating the nonlinear stability of the laminar flow. This is likely related to the fact that the ESN generates the laminar state in the presence of $O(10^{-3})$ noise making it more stable to perturbations of very small amplitudes. As a result, we observe a systematically increased laminarization probability in the interval $5 \times 10^{-3} \lesssim E \lesssim 10^{-2}$ in Fig. 4.

V. DISCUSSION

In this work, we have shown that echo state networks, a class of recurrent neural networks, are able to capture dynamical behavior qualitatively different from anything included in their training data set. We demonstrated this on the Moehlis-Faisst-Eckhardt (MFE) model, a classical example of fluid dynamics where the flow can display two distinct types of behavior, laminar flow and turbulence. In this problem, the transition from laminar flow to turbulence is a finite-amplitude instability, while the reverse transition is a spontaneous escape from a chaotic saddle. We computed predictions of these transitions using echo state networks trained solely on turbulent dynamics and compared them to the "truth," which we determined by directly time integrating the MFE model.

Remarkably, our echo state networks were able to learn laminar dynamics despite not having seen it during training. In addition, they were capable of successfully reproducing the statistical properties of both types of transition. Finally, we demonstrated that echo state networks can successfully act as generators of early warning signals of transition by tracking the predicted probability of turbulent-to-laminar transition in time. In our study, each echo state network was trained at a specific value of Re separately from other echo state networks. However, we believe that transfer learning already adapted for echo state networks [35] can be used to train only one echo state network using a long training time series and then adjust it to a new value of Re based on a small-size time series. Similarly, transfer learning could help improve the prediction accuracy by adding a small sample of laminar dynamics to the training set.

This success may be related to the echo state network approximation theorem recently proved for a one-dimensional observable of the true dynamical system and in the absence of noise [36]. It states that, under some mild conditions, for a sufficiently large reservoir and structurally stable true dynamical system, there exists such a matrix $\boldsymbol{W}_{\text{out}}$ that the dynamical system defined by the resulting echo state network is topologically conjugate to the true dynamical system. A crucial consequence of this theorem is that an echo state network is also expected to embed attractors of the true system. However, this theorem does not provide us with particular rules for building the matrices $\boldsymbol{W}$, $\boldsymbol{W}_{\text{in}}$ and $\boldsymbol{W}_{\text{out}}$ guaranteeing that a given attractor will be embedded into the manifold generated by an echo state network. We found that the training time series plays a crucial role in this process. In particular, echo state networks failed to learn the precise laminar dynamics and, as a consequence, were not able to produce any transitions, when the training time series did not include at least one large-amplitude excursion pulling the flow relatively close to the laminar state. This fact is illustrated in Fig. 5, where we show predictions made by echo state networks trained on four time series. The first one (top plots) does not cover the large excursion at $t \approx 3500$ whatsoever and results in an echo state network that is unable to predict laminar dynamics and turbulent-to-laminar transition. The second one covers only a small piece of this excursion which is enough for the echo state network to reproduce laminar dynamics, but not its stability. Finally, the third and fourth time series cover a sufficient part of the excursion in order for echo state networks to generate stable laminar flows. To predict new dynamics, echo state networks therefore require training time series that include some indication that the turbulent dynamics may not be ubiquitously stable.

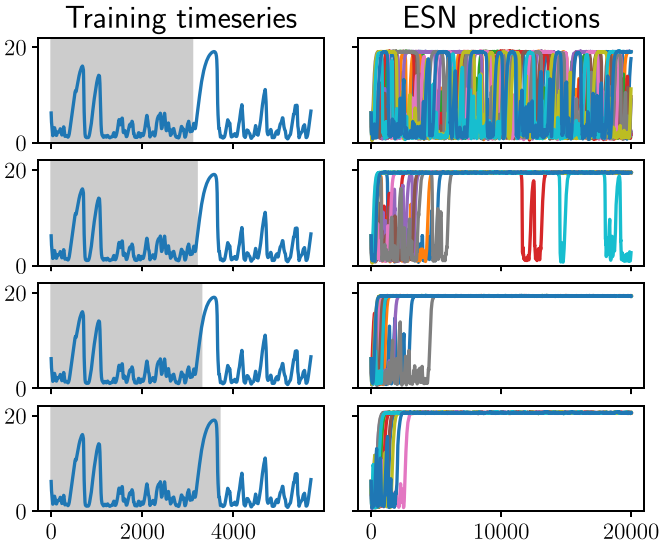

FIG. 5. Predictions made by echo state networks (right) each of which was trained using the corresponding shadowed part of the time series shown on the left. Trajectories of different colors on the right were obtained using random initial conditions.

Our results provide strong evidence that echo state networks can be used for data-driven discovery of new dynamical regimes, early warning of transitions between different dynamical modes and prediction of reversals from transitions. While these results were obtained for a low-dimensional turbulence model, they could be extended to higher-dimensional systems given that echo state networks have already been shown to successfully replicate the Kuramoto-Sivashinsky equation and 2D turbulence [7,10–13] where echo state networks can be complemented by autoencoders for the sake of dimensionality reduction [37]. We believe that our findings have a potential to be useful in a range of applications involving complex nonlinear systems characterized by abrupt transitions between dynamical regimes [27,30–34].

ACKNOWLEDGMENTS

A.P. is grateful for the support of Professor Timothy Palmer and the European Research Council grant ITHACA (Grant Agreement No. 741112) at the University of Oxford. C.B. acknowledges support from the Leverhulme Trust under Research Project Grant No. RPG-2018-311. K.L. and S.M.T. would like to acknowledge support of funding from the

Research Council under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. D5S-DLV-786780).

APPENDIX A: MOEHLIS-FAISST-ECKHARDT MODEL

The Moehlis-Faisst-Eckhardt model is an extension of Waleffe's eight-dimensional model which is believed to capture many features of shear flow turbulence including transient chaos [21]. In this model, a computational domain $\Omega = [0;\Gamma_x]\times[-1;1]\times[0;\Gamma_z]$, periodic in the x and z directions, is assumed. The velocity field is represented by the following decomposition:

$$\boldsymbol{u}(\boldsymbol{x},t) = \sum_{j=1}^{9} a_j(t)\boldsymbol{u}_j(\boldsymbol{x}), \tag{A1}$$

where the modes $\boldsymbol{u}_j(\boldsymbol{x})$ are defined as follows:

$$\boldsymbol{u}_1 = \begin{bmatrix} \sqrt{2}\sin(\pi y/2) \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{u}_2 = \begin{bmatrix} \frac{4}{\sqrt{3}}\cos^2(\pi y/2)\cos(\gamma z) \\ 0 \\ 0 \end{bmatrix},$$

$$\boldsymbol{u}_3 = \frac{2}{\sqrt{4\gamma^2+\pi^2}}\begin{bmatrix} 0 \\ 2\gamma\cos(\pi y/2)\cos(\gamma z) \\ \pi\sin(\pi y/2)\sin(\gamma z) \end{bmatrix}, \quad \boldsymbol{u}_4 = \begin{bmatrix} 0 \\ 0 \\ \frac{4}{\sqrt{3}}\cos(\alpha x)\cos^2(\pi y/2) \end{bmatrix},$$

$$\boldsymbol{u}_5 = \begin{bmatrix} 0 \\ 0 \\ 2\sin(\alpha x)\sin(\pi y/2) \end{bmatrix}, \quad \boldsymbol{u}_6 = \frac{4\sqrt{2}}{\sqrt{3(\alpha^2+\gamma^2)}}\begin{bmatrix} -\gamma\cos(\alpha x)\cos^2(\pi y/2)\sin(\gamma z) \\ 0 \\ \alpha\sin(\alpha x)\cos^2(\pi y/2)\cos(\gamma z) \end{bmatrix},$$

$$\boldsymbol{u}_7 = \frac{2\sqrt{2}}{\sqrt{\alpha^2+\gamma^2}}\begin{bmatrix} \gamma\sin(\alpha x)\sin(\pi y/2)\sin(\gamma z) \\ 0 \\ \alpha\cos(\alpha x)\sin(\pi y/2)\cos(\gamma z) \end{bmatrix}, \quad \boldsymbol{u}_8 = N_8\begin{bmatrix} \pi\alpha\sin(\alpha x)\sin(\pi y/2)\sin(\gamma z) \\ 2(\alpha^2+\gamma^2)\cos(\alpha x)\cos(\pi y/2)\sin(\gamma z) \\ -\pi\gamma\cos(\alpha x)\sin(\pi y/2)\cos(\gamma z) \end{bmatrix},$$

$$\boldsymbol{u}_9 = \begin{bmatrix} \sqrt{2}\sin(3\pi y/2) \\ 0 \\ 0 \end{bmatrix}.$$

The model itself is defined by a system of nine ordinary differential equations:

$$\frac{da_1}{dt} = \frac{\beta^2}{\mathrm{Re}} - \frac{\beta^2}{\mathrm{Re}}a_1 - \sqrt{\frac{3}{2}}\frac{\beta\gamma}{k_{\alpha\beta\gamma}}a_6a_8 + \sqrt{\frac{3}{2}}\frac{\beta\gamma}{k_{\beta\gamma}}a_2a_3,$$

$$\frac{da_2}{dt} = -\left(\frac{4\beta^2}{3}+\gamma^2\right)\frac{a_2}{\mathrm{Re}} + \frac{5\sqrt{2}}{3\sqrt{3}}\frac{\gamma^2}{k_{\alpha\gamma}}a_4a_6 - \frac{\gamma^2}{\sqrt{6}k_{\alpha\gamma}}a_5a_7 - \frac{\alpha\beta\gamma}{\sqrt{6}k_{\alpha\gamma}k_{\alpha\beta\gamma}}a_5a_8 - \sqrt{\frac{3}{2}}\frac{\beta\gamma}{k_{\beta\gamma}}a_1a_3 - \sqrt{\frac{3}{2}}\frac{\beta\gamma}{k_{\beta\gamma}}a_3a_9,$$

$$\frac{da_3}{dt} = -\frac{\beta^2+\gamma^2}{\mathrm{Re}}a_3 + \frac{2}{\sqrt{6}}\frac{\alpha\beta\gamma}{k_{\alpha\gamma}k_{\beta\gamma}}(a_4a_7+a_5a_6) + \frac{\beta^2(3\alpha^2+\gamma^2)-3\gamma^2(\alpha^2+\gamma^2)}{\sqrt{6}k_{\alpha\gamma}k_{\beta\gamma}k_{\alpha\beta\gamma}}a_4a_8,$$

$$\frac{da_4}{dt} = -\frac{3\alpha^2+4\beta^2}{3\mathrm{Re}}a_4 - \frac{\alpha}{\sqrt{6}}a_1a_5 - \frac{10}{3\sqrt{6}}\frac{\alpha^2}{k_{\alpha\gamma}}a_2a_6 - \sqrt{\frac{3}{2}}\frac{\alpha\beta\gamma}{k_{\alpha\gamma}k_{\beta\gamma}}a_3a_7 - \sqrt{\frac{3}{2}}\frac{\alpha^2\beta^2}{k_{\alpha\gamma}k_{\beta\gamma}k_{\alpha\beta\gamma}}a_3a_8 - \frac{\alpha}{\sqrt{6}}a_5a_9,$$

$$\frac{da_5}{dt} = -\frac{\alpha^2+\beta^2}{\mathrm{Re}}a_5 + \frac{\alpha}{\sqrt{6}}a_1a_4 + \frac{\alpha^2}{\sqrt{6}k_{\alpha\gamma}}a_2a_7 - \frac{\alpha\beta\gamma}{\sqrt{6}k_{\alpha\gamma}k_{\alpha\beta\gamma}}a_2a_8 + \frac{\alpha}{\sqrt{6}}a_4a_9 + \frac{2}{\sqrt{6}}\frac{\alpha\beta\gamma}{k_{\alpha\gamma}k_{\beta\gamma}}a_3a_6,$$

$$\frac{da_6}{dt} = -\frac{3\alpha^2+4\beta^2+3\gamma^2}{3\mathrm{Re}}a_6 + \frac{\alpha}{\sqrt{6}}a_1a_7 + \sqrt{\frac{3}{2}}\frac{\beta\gamma}{k_{\alpha\beta\gamma}}a_1a_8 + \frac{10}{3\sqrt{6}}\frac{\alpha^2-\gamma^2}{k_{\alpha\gamma}}a_2a_4 - 2\sqrt{\frac{2}{3}}\frac{\alpha\beta\gamma}{k_{\alpha\gamma}k_{\beta\gamma}}a_3a_5 + \frac{\alpha}{\sqrt{6}}a_7a_9 + \sqrt{\frac{3}{2}}\frac{\beta\gamma}{k_{\alpha\beta\gamma}}a_8a_9,$$

$$\frac{da_7}{dt} = -\frac{\alpha^2+\beta^2+\gamma^2}{\mathrm{Re}}a_7 - \frac{\alpha}{\sqrt{6}}(a_1a_6+a_6a_9) + \frac{1}{\sqrt{6}}\frac{\gamma^2-\alpha^2}{k_{\alpha\gamma}}a_2a_5 + \frac{1}{\sqrt{6}}\frac{\alpha\beta\gamma}{k_{\alpha\gamma}k_{\beta\gamma}}a_3a_4,$$

$$\frac{da_8}{dt} = -\frac{\alpha^2+\beta^2+\gamma^2}{\mathrm{Re}}a_8 + \frac{2}{\sqrt{6}}\frac{\alpha\beta\gamma}{k_{\alpha\gamma}k_{\alpha\beta\gamma}}a_2a_5 + \frac{\gamma^2(3\alpha^2-\beta^2+3\gamma^2)}{\sqrt{6}k_{\alpha\gamma}k_{\beta\gamma}k_{\alpha\beta\gamma}}a_3a_4,$$

$$\frac{da_9}{dt} = -\frac{9\beta^2}{\mathrm{Re}}a_9 + \sqrt{\frac{3}{2}}\frac{\beta\gamma}{k_{\beta\gamma}}a_2a_3 - \sqrt{\frac{3}{2}}\frac{\beta\gamma}{k_{\alpha\beta\gamma}}a_6a_8,$$

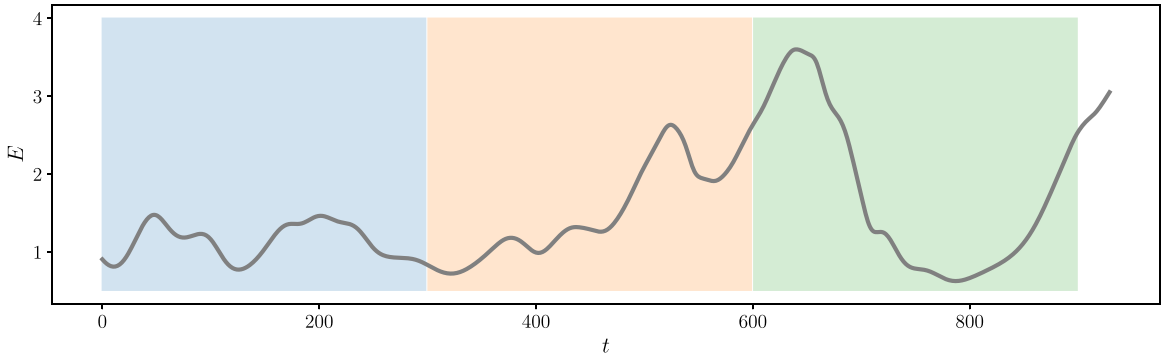

FIG. 6. An example of turbulent trajectory used for finding an optimal combination of hyperparameter values. Short-term predictions are made within each colored block.

where Re is the Reynolds number and the remaining coefficients are defined as follows:

$$\alpha = 2\pi/\Gamma_x, \quad \beta = \pi/2, \quad \gamma = 2\pi/\Gamma_z,$$
$$N_8 = \frac{2\sqrt{2}}{\sqrt{(\alpha^2+\gamma^2)(4\alpha^2+4\gamma^2+\pi^2)}},$$
$$k_{\alpha\gamma} = \sqrt{\alpha^2+\gamma^2}, \quad k_{\beta\gamma} = \sqrt{\beta^2+\gamma^2},$$
$$k_{\alpha\beta\gamma} = \sqrt{\alpha^2+\beta^2+\gamma^2}.$$

In this study, we consider fixed wavelengths values $\Gamma_x = 1.75\pi$ and $\Gamma_z = 1.2\pi$ corresponding to the minimal flow unit in plane Couette flow allowing for sustained turbulence [38].

APPENDIX B: ECHO STATE NETWORK GENERATION AND TRAINING

The echo state network (ESN) architecture we use in this work can be fully described by a two-equation system:

$$\boldsymbol{r}(t+\triangle t) = \tanh\left[\boldsymbol{b} + \boldsymbol{W}\boldsymbol{r}(t) + \boldsymbol{W}_{\text{in}}\boldsymbol{a}(t)\right] + \xi\boldsymbol{Z}, \tag{B1}$$

$$\tilde{\boldsymbol{a}}(t+\triangle t) = \boldsymbol{W}_{\text{out}}\begin{bmatrix}\boldsymbol{r}(t+\triangle t)\\ 1\end{bmatrix}, \tag{B2}$$

where $\tilde{\boldsymbol{a}}(t+\triangle t) \in \mathbb{R}^{N_a}$ is the prediction of the flow state at time $t+\triangle t$ based on its state $\boldsymbol{a}(t) \in \mathbb{R}^{N_a}$ at time t and the reservoir state $\boldsymbol{r}(t) \in \mathbb{R}^{N_r}$. The key feature of ESNs making them different from many other examples of recurrent neural networks is that the matrices $\boldsymbol{W}$, $\boldsymbol{W}_{\text{in}}$ and vector $\boldsymbol{b}$ are generated randomly. Moreover, the weight matrices are often assumed to be sparse which is akin to using pruning, a technique successfully employed in neural networks [39]. This allows us to avoid complicated and computationally demanding backpropagation-based algorithms for training and formulate training as a linear-regression problem while keeping a high accuracy of the final prediction.

Matrix $\boldsymbol{W} \in \mathbb{R}^{N_r \times N_r}$ is generated in three steps. First, we generate a random matrix $\widetilde{\boldsymbol{W}}$ by drawing all its coefficients from uniform distribution with support $(-0.5; 0.5)$. Second, we impose required sparsity s by setting to zero sN_r^2 randomly chosen matrix elements. Finally, we rescale matrix $\widetilde{\boldsymbol{W}}$ to ensure that the resulting matrix $\boldsymbol{W}$ has a prescribed spectral radius $\rho = \rho(\boldsymbol{W})$:

$$\boldsymbol{W} = \widetilde{\boldsymbol{W}}\frac{\rho}{\rho(\widetilde{\boldsymbol{W}})}.$$

Matrix $\boldsymbol{W}_{\text{in}} \in \mathbb{R}^{N_r \times N_u}$ and vector $\boldsymbol{b} \in \mathbb{R}^{N_r}$ are generated by drawing their elements from uniform distribution with support $(-1; 1)$ without imposing any constraints on sparsity.

The least-squares optimization problem is then formulated to minimize the sum of squares of deviations of one-step predictions with respect to matrix $\boldsymbol{W}_{\text{out}}$:

$$\min_{\boldsymbol{W}_{\text{out}}} \sum_{k=1}^{N_t} ||\boldsymbol{W}_{\text{out}}\boldsymbol{r}(k\triangle t) - \boldsymbol{a}(k\triangle t)||_2^2, \tag{B3}$$

where we assume that $N_t + 1$ flow states $\boldsymbol{a}(t)$ are known at times $t = 0, \triangle t, 2\triangle t, \ldots, N_t\triangle t$ and, thus, constitute our training data set. The flow state at $t = 0$ is only needed to compute the first prediction $\tilde{\boldsymbol{a}}(\triangle t)$. Instead of directly solving the normal equation, we find the solution by taking the Moore-Penrose pseudoinverse $\boldsymbol{R}^+$ of matrix $\boldsymbol{R}$:

$$\boldsymbol{W}_{\text{out}}^T = \boldsymbol{R}^+\boldsymbol{A}, \tag{B4}$$

where matrices $\boldsymbol{R}$ and $\boldsymbol{A}$ are defined as follows:

$$\boldsymbol{R} = \begin{bmatrix} & \boldsymbol{r}(\triangle t) & & 1\\ & \boldsymbol{r}(2\triangle t) & & 1\\ & \vdots & & 1\\ & \boldsymbol{r}(N_t\triangle t) & & 1\end{bmatrix}, \tag{B5}$$

$$\boldsymbol{A} = \begin{bmatrix} & \boldsymbol{a}(\triangle t) & \\ & \boldsymbol{a}(2\triangle t) & \\ & \vdots & \\ & \boldsymbol{a}(N_t\triangle t) & \end{bmatrix}. \tag{B6}$$

The pseudoinverse is computed using the singular value decomposition of $\boldsymbol{R}$. We pursue this approach in our work owing to the relative low dimensionality of our problem. It must be emphasized however that for large-scale problems, one may want to turn to semidirect or purely iterative methods for solving the normal equation.

For each Re, we train a separate ESN. For training purposes, we use a single turbulent trajectory without laminarization events. Our networks have four hyperparameters: reservoir state dimension N_r, spectral radius ρ, sparsity s

and noise strength ξ. To find an optimal combination of hyperparameter values, we use another turbulent trajectory simulated at the same Reynolds number and divide it into a set of blocks of equal length $t = 300$ (see Fig. 6 for an example). Then, short-term predictions are made by an ESN within each block to estimate its performance as a residual sum of squares. The optimal configuration is then found by train 10 ESNs per combination of hyperparameter values, ranking them according to their performance and selecting the best one. To reduce the dimensionality of the hyperparameter space, we fix $N_r = 1500$ and $\xi = 10^{-3}$. As a result, values $\rho = 0.5$ and $s = 0.9$ have been found to be optimal for all the Reynolds numbers except for $\mathrm{Re} = 250$ for which $s = 0.5$ has been used. The number of blocks used for the hyperparameter search depends on available turbulent trajectories and is typically equal to 10.

[1] K. Xu, M. Zhang, J. Li, S. S. Du, K.-I. Kawarabayashi, and S. Jegelka, How neural networks extrapolate: From feedforward to graph neural networks, in *ICLR 2021: The Ninth International Conference on Learning Representations* (2021).

[2] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, Machine learning for fluid mechanics, Annu. Rev. Fluid Mech. **52**, 477 (2020).

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).

[4] H. Jaeger and H. Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, Science **304**, 78 (2004).

[5] M. Lukoševičius and H. Jaeger, Reservoir computing approaches to recurrent neural network training, Comput. Sci. Rev. **3**, 127 (2009).

[6] P. R. Vlachas, J. Pathak, B. R. Hunt, T. P. Sapsis, M. Girvan, E. Ott, and P. Koumoutsakos, Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics, Neural Netw. **126**, 191 (2020).

[7] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott, Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data, Chaos **27**, 121102 (2017).

[8] X. Chen, T. Weng, H. Yang, C. Gu, J. Zhang, and M. Small, Mapping topological characteristics of dynamical systems into neural networks: A reservoir computing approach, Phys. Rev. E **102**, 033314 (2020).

[9] N. A. K. Doan, W. Polifke, and L. Magri, Short- and long-term predictions of chaotic flows and extreme events: A physics-constrained reservoir computing approach, Proc. R. Soc. A **477**, 20210135 (2021).

[10] J. Pathak, A. Wikner, R. Fussell, S. Chandra, B. R. Hunt, M. Girvan, and E. Ott, Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model, Chaos **28**, 041101 (2018).

[11] 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**, 024102 (2018).

[12] S. Pandey and J. Schumacher, Reservoir computing model of two-dimensional turbulent convection, Phys. Rev. Fluids **5**, 113506 (2020).

[13] F. Heyder and J. Schumacher, Echo state network for two-dimensional turbulent moist Rayleigh-Bénard convection, Phys. Rev. E **103**, 053107 (2021).

[14] A. Haluszczynski and C. Räth, Good and bad predictions: Assessing and improving the replication of chaotic attractors by means of reservoir computing, Chaos **29**, 103143 (2019).

[15] 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, Nonlin. Proc. Geophys. **27**, 373 (2020).

[16] D. Barkley, Theoretical perspective on the route to turbulence in a pipe, J. Fluid Mech. **803**, P1 (2016).

[17] V. Lucarini and T. Bódai, Transitions across Melancholia States in a Climate Model: Reconciling the Deterministic and Stochastic Points of View, Phys. Rev. Lett. **122**, 158701 (2019).

[18] P. J. Menck, J. Heitzig, N. Marwan, and J. Kurths, How basin stability complements the linear-stability paradigm, Nat. Phys. **9**, 89 (2013).

[19] F. Pétrélis, S. Fauve, E. Dormy, and J.-P. Valet, Simple Mechanism for Reversals of Earth's Magnetic Field, Phys. Rev. Lett. **102**, 144503 (2009).

[20] S. Tobias, The turbulent dynamo, J. Fluid Mech. **912**, P1 (2021).

[21] J. Moehlis, H. Faisst, and B. Eckhardt, A low-dimensional model for turbulent shear flows, New J. Phys. **6**, 56 (2004).

[22] F. Waleffe, On a self-sustaining process in shear flows, Phys. Fluids **9**, 883 (1997).

[23] C. Beaume, G. P. Chini, K. Julien, and E. Knobloch, Reduced description of exact coherent states in parallel shear flows, Phys. Rev. E **91**, 043010 (2015).

[24] J. Moehlis, H. Faisst, and B. Eckhardt, Periodic orbits and chaotic sets in a low-dimensional model for shear flows, SIAM J. Appl. Dyn. Syst. **4**, 352 (2005).

[25] R. Xiao, L.-W. Kong, Z.-K. Sun, and Y.-C. Lai, Predicting amplitude death with machine learning, Phys. Rev. E **104**, 014205 (2021).

[26] K. Avila, D. Moxey, A. de Lozar, M. Avila, D. Barkley, and B. Hof, The onset of turbulence in pipe flow, Science **333**, 192 (2011).

[27] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. Van Nes, M. Rietkerk, and G. Sugihara, Early-warning signals for critical transitions, Nature (London) **461**, 53 (2009).

[28] A. Pershin, C. Beaume, and S. M. Tobias, A probabilistic protocol for the assessment of transition and control, J. Fluid Mech. **895**, A16(2020).

[29] M. U. Kobayashi, K. Nakai, Y. Saiki, and N. Tsutsumi, Dynamical system analysis of a data-driven model constructed by reservoir computing, Phys. Rev. E **104**, 044215 (2021).

[30] T. M. Lenton, Early warning of climate tipping points, Nat. Clim. Change **1**, 201 (2011).

[31] P. Ashwin, S. Wieczorek, R. Vitolo, and P. Cox, Tipping points in open systems: Bifurcation, noise-induced and rate-dependent examples in the climate system, Philos. Trans. R. Soc. A **370**, 1166 (2012).

[32] P. D. Ditlevsen and S. J. Johnsen, Tipping points: Early warning and wishful thinking, Geophys. Res. Lett. **37**, L19703 (2010).

[33] S. Kefi, V. Guttal, W. A. Brock, S. R. Carpenter, A. M. Ellison, V. N. Livina, D. A. Seekell, M. Scheffer, E. H. van Nes, and V. Dakos, Early warning signals of ecological transitions: Methods for spatial patterns, PLoS ONE **9**, e92097 (2014).

[34] M. Scheffer, S. R. Carpenter, T. M. Lenton, J. Bascompte, W. Brock, V. Dakos, J. Van de Koppel, I. A. Van de Leemput, S. A. Levin, E. H. Van Nes *et al.*, Anticipating critical transitions, Science **338**, 344 (2012).

[35] M. Inubushi and S. Goto, Transfer learning for nonlinear dynamics and its application to fluid turbulence, Phys. Rev. E **102**, 043301 (2020).

[36] A. Hart, J. Hook, and J. Dawes, Embedding and approximation theorems for echo state networks, Neural Netw. **128**, 234 (2020).

[37] N. A. K. Doan, W. Polifke, and L. Magri, Auto-encoded reservoir computing for turbulence learning, in *International Conference on Computational Science*, edited by M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. Sloot (Springer, Krakow, Poland, 2021), pp. 344–351.

[38] J. M. Hamilton, J. Kim, and F. Waleffe, Regeneration mechanisms of near-wall turbulence structures, J. Fluid Mech. **287**, 317 (1995).

[39] Y. LeCun, J. S. Denker, and S. A. Solla, Optimal brain damage, in *Advances in Neural Information Processing Systems* edited by D. S. Touretzky (Morgan Kaufmann, San Francisco CA, 1990), pp. 598–605.