PLOS ONE

RESEARCH ARTICLE

Data-assisted reduced-order modeling of extreme events in complex dynamical systems

Zhong Yi Wan[1], Pantelis Vlachas[2], Petros Koumoutsakos[2], Themistoklis Sapsis[1]*

1 Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States of America, **2** Chair of Computational Science, ETH Zurich, Zurich, Switzerland

* sapsis@mit.edu

OPEN ACCESS

Citation: Wan ZY, Vlachas P, Koumoutsakos P, Sapsis T (2018) Data-assisted reduced-order modeling of extreme events in complex dynamical systems. PLoS ONE 13(5): e0197704. https://doi.org/10.1371/journal.pone.0197704

Editor: Daniel Durstewitz, Heidelberg University, GERMANY

Received: March 8, 2018

Accepted: May 7, 2018

Published: May 24, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: TPS and ZYW have been supported through the Air Force Office of Scientific Research grant FA9550-16-1-0231, the Army Research Office grant W911NF-17-1-0306 and the Office of Naval Research grant N00014-17-1-2676. PK and PV acknowledge support by the Advanced Investigator Award of the European Research Council (Grant No: 341117). The funders had no role in study design, data collection and analysis,

Abstract

The prediction of extreme events, from avalanches and droughts to tsunamis and epidemics, depends on the formulation and analysis of relevant, complex dynamical systems. Such dynamical systems are characterized by high intrinsic dimensionality with extreme events having the form of rare transitions that are several standard deviations away from the mean. Such systems are not amenable to classical order-reduction methods through projection of the governing equations due to the large intrinsic dimensionality of the underlying attractor as well as the complexity of the transient events. Alternatively, data-driven techniques aim to quantify the dynamics of specific, critical modes by utilizing data-streams and by expanding the dimensionality of the reduced-order model using delayed coordinates. In turn, these methods have major limitations in regions of the phase space with sparse data, which is the case for extreme events. In this work, we develop a novel hybrid framework that complements an imperfect reduced order model, with data-streams that are integrated though a recurrent neural network (RNN) architecture. The reduced order model has the form of projected equations into a low-dimensional subspace that still contains important dynamical information about the system and it is expanded by a long short-term memory (LSTM) regularization. The LSTM-RNN is trained by analyzing the mismatch between the imperfect model and the data-streams, projected to the reduced-order space. The data-driven model assists the imperfect model in regions where data is available, while for locations where data is sparse the imperfect model still provides a baseline for the prediction of the system state. We assess the developed framework on two challenging prototype systems exhibiting extreme events. We show that the blended approach has improved performance compared with methods that use either data streams or the imperfect model alone. Notably the improvement is more significant in regions associated with extreme events, where data is sparse.

decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Extreme events are omnipresent in important problems in science and technology such as turbulent and reactive flows [1, 2], Kolmogorov [3] and unstable plane Couette flow [4]), geophysical systems (e.g. climate dynamics [5, 6], cloud formations in tropical atmospheric convection [7, 8]), nonlinear optics [9, 10] or water waves [11–13]), and mechanical systems (e.g. mechanical metamaterials [14, 15]).

The complete description of these system through the governing equations is often challenging either because it is very hard/expensive to solve these equations with an appropriate resolution or due to the magnitude of the model errors. The very large dimensionality of their attractor in combination with the occurrence of important transient, but rare events, makes the application of classical order-reduction methods a challenging task. Indeed, classical Galerkin projection methods encounter problems as the truncated degrees-of-freedom are often essential for the effective description of the system due to high underlying intrinsic dimensionality. On the other hand, purely data-driven, non-parametric methods such as delay embeddings [16–21], equation-free methods [22, 23], Gaussian process regression based methods [24], or recurrent neural networks based approaches [25] may not perform well when it comes to rare events, since the training data-sets typically contain only a small number of the rare transient responses. The same limitations hold for data-driven, parametric methods [26–29], where the assumed analytical representations have parameters that are optimized so that the resulted model best fits the data. Although these methods perform well when the system operates within the main 'core' of the attractor, this may not be the case when rare and/or extreme events occur.

We propose a hybrid method for the formulation of a reduced-order model that combines an imperfect physical model with available data streams. The proposed framework is important for the non-parametric description, prediction and control of complex systems whose response is characterized by both i) high-dimensional attractors with broad energy spectrum distributed across multiple scales, and ii) strongly transient non-linear dynamics such as extreme events.

We focus on data-driven recurrent neural networks (RNN) with a long-short term memory (LSTM) [30] that represents some of the truncated degrees-of-freedom. The key concept of our work is the observation that while the imperfect model alone has limited descriptive and prediction skills (either because it has been obtained by a radical reduction or it is a coarse-grid solution of the original equations), it still contains important information especially for the instabilities of the system, assuming that the relevant modes are included in the truncation. However, these instabilities need to be combined with an accurate description of the nonlinear dynamics within the attractor and this part is captured in the present framework by the recurrent neural network. Note, that embedding theorems [31, 32] make the additional memory of the RNN to represent dimensions of the system that have been truncated, a property that provides an additional advantage in the context of reduced-order modeling [20, 25].

We note that such blended model-data approaches have been proposed previously in other contexts. In [33, 34], a hybrid forecasting scheme based on reservoir computing in conjunction with knowledge-based models are successfully applied to prototype spatiotemporal chaotic systems. In [35–37] the linearized dynamics were projected to low-dimensional subspaces and were combined with additive noise and damping that were rigorously selected to represent the effects nonlinear energy fluxes from the truncated modes. The developed scheme resulted in reduced-order stochastic models that efficiently represented the second order statistics in the presence of arbitrary external excitation. In [38] a deep neural network architecture was

developed to reconstruct the near-wall flow field in a turbulent channel flow using suitable wall only information. These nonlinear near-wall models can be integrated with flow solvers for the parsimonious modeling and control of turbulent flows [39–42]. In [43] a framework was introduced wherein solutions from intermediate models, which capture some physical aspects of the problem, were incorporated as solution representations into machine learning tools to improve the predictions of the latter, minimizing the reliance on costly experimental measurements or high-resolution, high-fidelity numerical solutions. [44] design a stable adaptive control strategy using neural networks for physical systems for which the state dependence of the dynamics is reasonably well understood, but the exact functional form of this dependence, or part thereof, is not, such as underwater robotic vehicles and high performance aircraft. In [29, 45–47] neural nets are developed to simultaneously learn the solution of the model equations using data. In these works that only a small number of scalar parameters is utilized to represent unknown dynamics, while the emphasis is given primarily on the learning of the solution, which is represented through a deep neural network. In other words, it is assumed that a family of models that 'lives' in a low-dimensional parameter space can capture the correct response. Such a representation is not always available though. Here our goal is to apply such a philosophy on the prediction of complex systems characterized by high dimensionality and strongly transient dynamics. We demonstrate the developed strategy in prototype systems exhibiting extreme events and show that the hybrid strategy has important advantages compared with either purely data-driven methods or those relying on reduced-order models alone.

Materials and methods

We consider a nonlinear dynamical system with state variable $\mathbf{u} \in \mathbb{R}^d$ and dynamics given by

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}) = \mathbf{L}\mathbf{u} + \mathbf{h}(\mathbf{u}), \tag{1}$$

where $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^d$ is a deterministic, time-independent operator with linear and nonlinear parts $\mathbf{L}$ and $\mathbf{h}$ respectively. We are specifically interested in systems whose dynamics results in a non-trivial, globally attracting manifold $S \subset \mathbb{R}^d$ to which trajectories quickly decay. The intrinsic dimension of S is presumably much less than d.

In traditional Galerkin-based reduced-order model [48] one typically uses an ansatz of the form

$$\mathbf{u} = \mathbf{Y}\boldsymbol{\xi} + \mathbf{Z}\boldsymbol{\eta} + \mathbf{b}, \tag{2}$$

where the columns of matrix $\mathbf{Y} = [\mathbf{y}_1,\ldots,\mathbf{y}_m]$ form an orthonormal basis of Y, an m-dimensional subspace of $\mathbb{R}^d$, and the columns of $\mathbf{Z} = [\mathbf{z}_1,\ldots,\mathbf{z}_{d-m}]$ make up an orthonormal basis for the orthogonal complement $Z = \mathbb{R}^d \setminus Y$; $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ are the projection coordinates associated with $\mathbf{Y}$ and $\mathbf{Z}$; $\mathbf{b}$ is an offset vector typically made equal to the attractor mean state. This linear expansion allows reduction to take place through special choices of subspace Y and Z, as well as their corresponding basis. For example, the well-known proper orthogonal decomposition (POD) derives the subspace empirically to be such that the manifold S preserves its variance as much as possible when projected to Y (or equivalently, minimizing the variance when projected to Z), given a fixed dimension constraint m.

We show that such a condition enables reduction, by substituting Eq (2) into Eq (1) and projecting onto Y and Z respectively to obtain two coupled systems of differential equations:

$$\begin{aligned}\frac{d\boldsymbol{\xi}}{dt} &= \mathbf{Y}^T\mathbf{L}\mathbf{Y}\boldsymbol{\xi} + \mathbf{Y}^T\mathbf{L}\mathbf{Z}\boldsymbol{\eta} + \mathbf{Y}^T\mathbf{h}(\mathbf{Y}\boldsymbol{\xi} + \mathbf{Z}\boldsymbol{\eta} + \mathbf{b}) + \mathbf{Y}^T\mathbf{L}\mathbf{b} \\ \frac{d\boldsymbol{\eta}}{dt} &= \mathbf{Z}^T\mathbf{L}\mathbf{Y}\boldsymbol{\xi} + \mathbf{Z}^T\mathbf{L}\mathbf{Z}\boldsymbol{\eta} + \mathbf{Z}^T\mathbf{h}(\mathbf{Y}\boldsymbol{\xi} + \mathbf{Z}\boldsymbol{\eta} + \mathbf{b}) + \mathbf{Z}^T\mathbf{L}\mathbf{b}.\end{aligned} \tag{3}$$

If on average $|\boldsymbol{\eta}| \ll |\boldsymbol{\xi}|$, we may make the approximation that $\boldsymbol{\eta} = 0$, leading to a m-dimensional system (ideally $m \ll d$)

$$\frac{d\boldsymbol{\xi}}{dt} = \mathbf{Y}^T\mathbf{L}\mathbf{Y}\boldsymbol{\xi} + \mathbf{Y}^T\mathbf{h}(\mathbf{Y}\boldsymbol{\xi} + \mathbf{b}) + \mathbf{Y}^T\mathbf{L}\mathbf{b} = \mathbf{F}_{\xi}(\boldsymbol{\xi}), \tag{4}$$

which can be integrated in time. This is known as the *flat Galerkin method*. The solution to Eq (1) is approximated by $\mathbf{u} \approx \mathbf{Y}\boldsymbol{\xi} + \mathbf{b}$.

Using (4) as an approximation to Eq (1) is known to suffer from a number of problems. First, the dimension m of the reduction subspace Y may be too large for $|\boldsymbol{\eta}| \approx 0$ to hold true. Second, the subspace Z is derived merely based on statistical properties of the manifold without addressing the dynamics. This implies that even if $\boldsymbol{\eta}$ has small magnitude on average it may play a big role in the dynamics of the high-energy space (e.g. acting as buffers for energy transfer between modes [49]). Neglecting such dimensions in the description of the system may alter its dynamical behaviors and compromise the ability of the model to generate reliable forecasts.

An existing method that attempts to address the truncation effect of the $\boldsymbol{\eta}$ terms is the *nonlinear Galerkin projection* [48, 50], which expresses $\boldsymbol{\eta}$ as a function of $\boldsymbol{\xi}$:

$$\boldsymbol{\eta} = \boldsymbol{\Phi}(\boldsymbol{\xi}), \tag{5}$$

yielding a reduced system

$$\frac{d\boldsymbol{\xi}}{dt} = \mathbf{Y}^T\mathbf{L}\mathbf{Y}\boldsymbol{\xi} + \mathbf{Y}^T\mathbf{L}\mathbf{Z}\boldsymbol{\Phi}(\boldsymbol{\xi}) + \mathbf{Y}^T\mathbf{h}(\mathbf{Y}\boldsymbol{\xi} + \mathbf{Z}\boldsymbol{\Phi}(\boldsymbol{\xi}) + \mathbf{b}) + \mathbf{Y}^T\mathbf{L}\mathbf{b}. \tag{6}$$

The problems boils down to finding Φ, often empirically. Unfortunately, Φ is well-defined only when the inertial manifold S is fully parametrized by dimensions of Y (see Fig 1), which is a difficult condition to achieve for most systems under a reasonable m. Even if the condition is met, how to systematically find Φ remains a big challenge.

Data-assisted reduced-order modeling

In this section we introduce a new framework for improving the reduced-space model that assists, with data streams, the nonlinear Galerkin method. Our main idea relies on building an additional data-driven model from data series observed in the reduction space to assist the equation-based model Eq (4).

We note that the *exact* dynamics of $\boldsymbol{\xi}$ can be written as

$$\frac{d\boldsymbol{\xi}}{dt} = \mathbf{F}_{\xi}(\boldsymbol{\xi}) + \mathbf{G}(\boldsymbol{\xi}, \boldsymbol{\eta}), \tag{7}$$

where $\mathbf{F}_{\xi}$ is defined in Eq (4) and $\mathbf{G} : \mathbb{R}^m \times \mathbb{R}^{d-m} \to \mathbb{R}^m$ encompasses the coupling between $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$. We will refer to $\boldsymbol{\psi} = \mathbf{G}(\boldsymbol{\xi}, \boldsymbol{\eta})$ as the *complementary dynamics* since it can be thought of as a correction that complements the flat Galerkin dynamics $\mathbf{F}_{\xi}$.

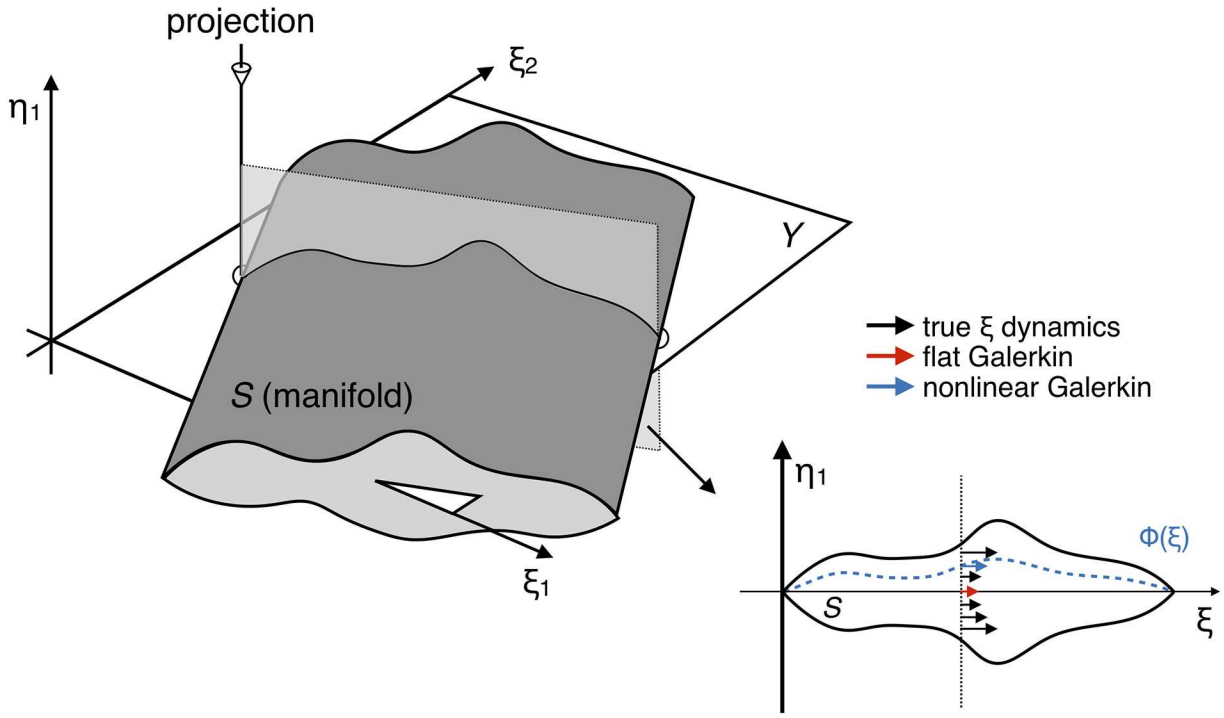

Fig 1. Geometric illustration of flat and nonlinear Galerkin projected dynamics in $\mathbb{R}^3$. 3D manifold S living in (ξ_1, ξ_2, η_1) is projected to 2D plane parametrized by (ξ_1, ξ_2). Parametrization is assumed to be imperfect, i.e. out-of-plane coordinate η_1 cannot be uniquely determined from (ξ_1, ξ_2). Flat Galerkin method always uses the dynamics corresponding to $\eta_1 = 0$. Nonlinear Galerkin method uses the dynamics corresponding to $\eta_1 = \Phi(\xi_1, \xi_2)$ where Φ is determined by some prescribed criterion (e.g. minimization of $L2$ error).

https://doi.org/10.1371/journal.pone.0197704.g001

The key step of our framework is to establish a data-driven model $\hat{\mathbf{G}}$ to approximate $\mathbf{G}$:

$$\boldsymbol{\psi}(t) = \mathbf{G}(\boldsymbol{\xi}(t), \boldsymbol{\eta}(t)) \approx \hat{\mathbf{G}}(\boldsymbol{\xi}(t), \boldsymbol{\xi}(t-\tau), \boldsymbol{\xi}(t-2\tau), \ldots) \tag{8}$$

where $\boldsymbol{\xi}(t)$, $\boldsymbol{\xi}(t-\tau)$,... are uniformly time-lagged states in $\boldsymbol{\xi}$ up to a reference initial condition. The use of delayed $\boldsymbol{\xi}$ states makes up for the fact that Y may not be a perfect parametrization subspace for S. The missing state information not directly accessible from within Y is instead inferred from these delayed $\boldsymbol{\xi}$ states and then used to compute $\boldsymbol{\psi}$. This model form is motivated by the embedding theorems developed by Whitney [31] and Takens [32], who showed that the attractor of a deterministic, chaotic dynamical system can be fully embedded using delayed coordinates.

We use the long short-term memory (LSTM) [30], a regularization of recurrent neural network (RNN), as the fundamental building block for constructing $\hat{G}$. The LSTM has been recently deployed successfully for the formulation of fully data-driven models for the prediction of complex dynamical systems [25]. Here we employ the same strategy to model the complementary dynamics while we preserve the structure of the projected equations. LSTM takes advantage of the sequential nature of the time-delayed reduced space coordinates by processing the input in chronological order and keeping memory of the useful state information that complements $\boldsymbol{\xi}$ at each time step. An overview of the RNN model and the LSTM is given in S1 Appendix.

Building from LSTM units, we use two different architectures to learn the complementary dynamics from data. The first architecture reads a sequence of $\boldsymbol{\xi}$ states, i.e. states projected to the d–dimensional subspace and outputs the corresponding sequence of complementary

dynamics. The second architecture reads an input sequence and integrate the output dynamics to predict future. The details of both architectures are described below.

Data series. Both architectures are trained and tested on the same data set consisting of N data series, where N is assumed to be large enough such that the low-order statistics of S are accurately represented. Each data series is a sequence of observed values in reduced space Y, with strictly increasing and evenly spaced observation times. Without loss of generality, we assume that all data series have the same length. Moreover, the observation time spacing τ is assumed to be small so that the true dynamics at each step of the series can be accurately estimated with finite difference. We remark that for single-step prediction (architecture I below) increasing τ (while keeping the number of steps constant) is beneficial for training as it reduces the correlation between successive inputs. However, for multi-step prediction (architecture II), large τ incurs integration errors which quickly outweigh the benefit of having decorrelated inputs. Hence, we require small τ in data.

Architecture I. We denote an input sequence of length-p as $\{\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_p\}$ and the corresponding finite-difference interpolated dynamics as $\{\dot{\boldsymbol{\xi}}_1,\ldots,\dot{\boldsymbol{\xi}}_p\}$. A forward pass in the first architecture works as follows (illustrated in Fig 2I). At time step i, input $\boldsymbol{\xi}_i$ is fed into a LSTM cell with n_{LSTM} hidden states, which computes its output $\mathbf{h}_i$ based on the received input and its previous memory states (initialized to zero). The LSTM output is then passed through an intermediary fully-connected (FC) layer with n_{FC} hidden states and rectified linear unit (ReLU) activations to the output layer at desired dimension m. Here $\mathbf{h}_i$ is expected to contain state information of the *unobserved* $\boldsymbol{\eta}$ at time step i, reconstructed effectively as a function of all previous *observed* states $\{\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{i-1}\}$. The model output is a predicted sequence of complementary dynamics $\{\hat{\boldsymbol{\psi}}_1,\ldots,\hat{\boldsymbol{\psi}}_p\}$.

Optionally, $\mathbf{h}_i$ can be concatenated with LSTM input $\boldsymbol{\xi}_i$ to make up the input to the FC layer. The concatenation is necessary when n_{LSTM} is small relative to m. Under such conditions, the LSTM hidden states $\mathbf{h}_i$ are more likely trained to represent $\boldsymbol{\eta}_i$ alone, as opposed to $(\boldsymbol{\xi}_i,\boldsymbol{\eta}_i)$ combined. $\boldsymbol{\xi}_i$ thus needs to be seen by the FC layer in order to have all elements necessary in order to estimate $\boldsymbol{\psi}$. If the LSTM cell has sufficient room to integrate incoming input with memory (i.e. number of hidden units larger than the intrinsic dimensionality of the attractor), the concatenation may be safely ignored.

In the case of models with lower complexity for faster learning a small n_{LSTM} is preferred. However, finding the minimum working n_{LSTM}, which is expected to approach the intrinsic attractor dimension, is a non-trivial problem. Therefore, it is sometimes desirable to conservatively choose n_{LSTM}. In this case the LSTM unit is likely to have sufficient cell capacity to integrate incoming input with memory, rendering the input concatenation step unnecessary.

The model is trained by minimizing a loss function with respect to the weights of the LSTM cell and FC layer. The loss function is defined as a weighted sum of the mean squared error (MSE) of the complementary dynamics:

$$L=\sum_{i=1}^{p} w_i\left|\left|\hat{\boldsymbol{\psi}}_i-\boldsymbol{\psi}_i\right|\right|^2 \tag{9}$$

where $\boldsymbol{\psi}_i=\dot{\boldsymbol{\xi}}_i-\mathbf{F}_{\tilde{\xi}}(\boldsymbol{\xi})$ is the true complementary dynamics at step i. Note that for this architecture it is equivalent to defining the loss based on MSE of the *total* dynamics. For weights w_i we use a step profile:

$$w_i=\begin{cases} w_0 & 0<i\le p_t \\ 1 & p_t<i\le p, \end{cases} \tag{10}$$

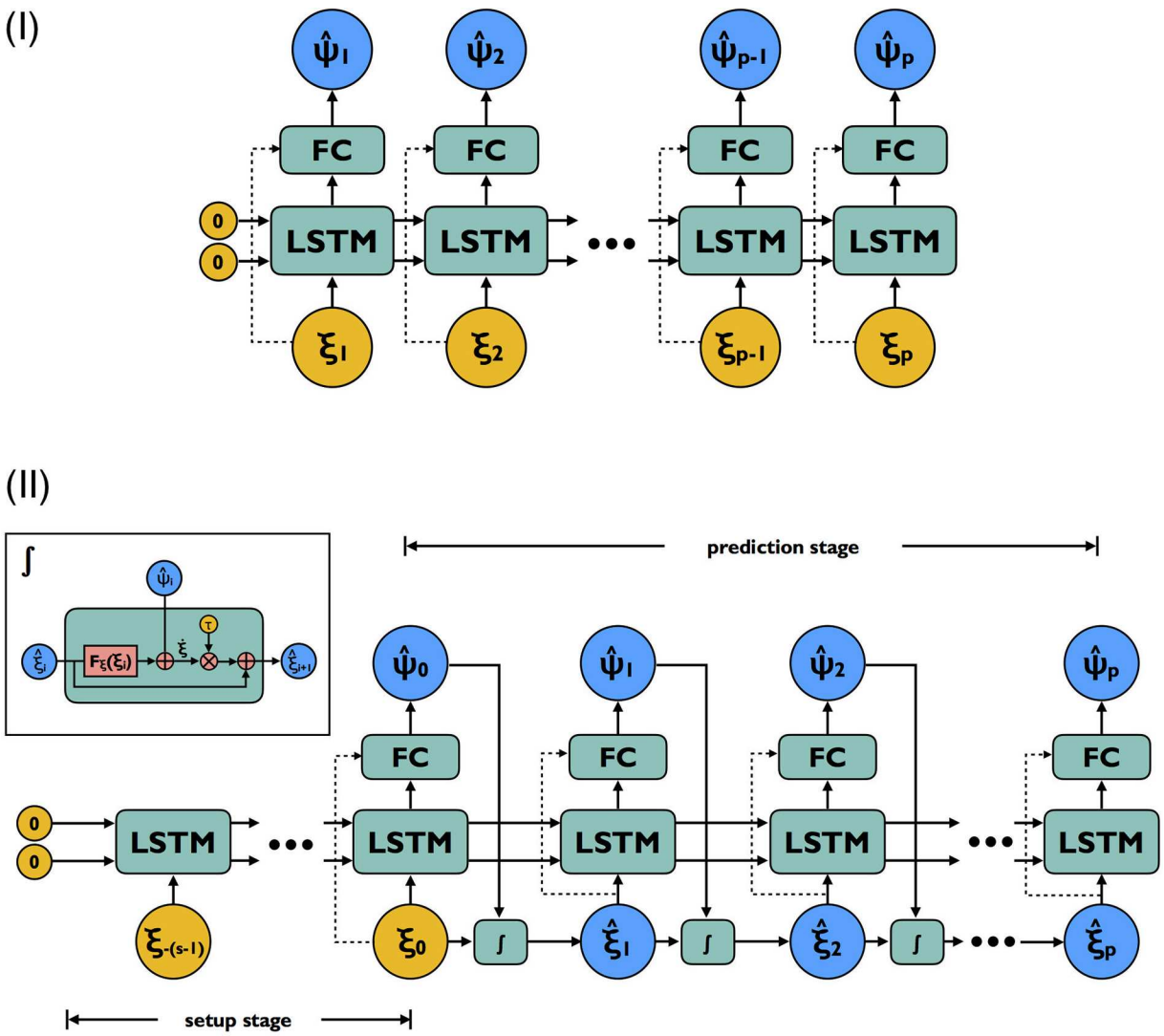

Fig 2. Computational graph for model architecture I and II. Yellow nodes are input provided to the network corresponding to sequence of states and blue nodes are prediction targets corresponding to the complementary dynamics (plus states for architecture II). Blocks labeled 'FC' are fully-connected layers with ReLU activations. Dashed arrows represent optional connections depending on the capacity of LSTM relative to the dimension of ξ. Both architectures share the same set of trainable weights. For architecture I, predictions are made as input is read; input is always accurate regardless of any prediction errors made in previous steps. This architecture is used *only for training*. Architecture II makes prediction in a sequence-to-sequence (setup sequence to prediction sequence) fashion. Errors made early do impact all predictions that follow. This architecture is used for *fine-tuning weights* and *multi-step-ahead prediction*.

https://doi.org/10.1371/journal.pone.0197704.g002

where $w_0 \ll 1$ is used to weight the first p_t steps when the LSTM unit is still under the transient effects of the cell states being initialized to zero. Predictions made during this period is therefore valued much less. In practice, p_t is usually negatively correlated with the parametrization power of the reduction subspace Y and can be determined empirically. For optimization we use the gradient-based Adam optimizer [51] (also described in S1 Appendix) with early stopping. The gradient is calculated for small batches of data series (batch size n_{batch}) and across the entire training data for n_{ep} epochs.

A notable property of this model architecture is that input representing the reduced state is always accurate regardless of any errors made in predicting the dynamics previously. This is undesirable especially for chaotic systems where errors tend to grow exponentially. Ideally, the

model should be optimized with respect to the cumulative effects of the prediction errors. To this end, this architecture is primarily used for pre-training and a second architecture is utilized for fine-tuning and multi-step-ahead prediction.

Architecture II. The second architecture bears resemblance to the sequence-to-sequence (seq2seq) models which have been widely employed for natural language processing tasks [52, 53]. It consists of two stages (illustrated in Fig 2II): a set-up stage and a prediction stage. The set-up stage has the same structure as architecture I, taking as input a uniformly spaced sequence of s reduced-space states which we call $\{\boldsymbol{\xi}_{-s+1}, \boldsymbol{\xi}_{-s+2}\ldots, \boldsymbol{\xi}_0\}$. No output, however, is produced until the very last step. This stage acts as a spin-up such that zero initializations to the LSTM memory no longer affects prediction of dynamics at the beginning of the next stage. The output of the set-up stage is a single prediction of the complementary dynamics $\hat{\boldsymbol{\psi}}_0$ corresponding to the last state of the input sequence and the ending LSTM memory states. This dynamics is combined with $\mathbf{F}_{\xi}(\boldsymbol{\xi}_0)$ to give the total dynamics at $\boldsymbol{\xi}_0$. The final state and dynamics are passed to an integrator to obtain the first input state of the prediction stage $\hat{\boldsymbol{\xi}}_1$. During the prediction stage, complementary dynamics is predicted iteratively based on the newest state prediction and the LSTM memory content before combined with $\mathbf{F}_{\xi}$ dynamics to generate the total dynamics and subsequently the next state. After p prediction steps, the output of the model is obtained as a sequence of predicted states $\{\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_p\}$ and a sequence of complementary dynamics $\{\hat{\boldsymbol{\psi}}_1, \ldots, \hat{\boldsymbol{\psi}}_p\}$.

For this architecture we define the loss function as

$$L = \sum_{i=1}^{p} w_i ||\hat{\boldsymbol{\psi}}_i + \mathbf{F}_{\xi}(\hat{\boldsymbol{\xi}}_i) - \dot{\hat{\boldsymbol{\xi}}}_i||^2. \tag{11}$$

This definition is based on MSE of the *total* dynamics so that the model learns to 'cooperate' with the projected dynamics $\mathbf{F}_{\xi}$. For weights we use an exponential profile:

$$w_i = \gamma^{i-1}, \ 0 < i \leq p \tag{12}$$

where $0 < \gamma < 1$ is a pre-defined ratio of decay. This profile is designed to counteract the exponentially growing nature of the errors in a chaotic system and prevent exploding gradients. Similar to architecture I, training is performed in batches using the Adam algorithm.

Architecture II, in contrast with the architecture I, finishes reading the entire input sequence before producing the prediction sequence. For this reason it is suitable for running multi-step-ahead predictions. Both architectures, however, share the same set of trainable weights used to estimate the complementary dynamics. Hence, we can utilize architecture I as a pre-training facility for architecture II because it tends to have smaller gradients (as errors do not accumulate over time steps) and thus faster convergence. This idea is very similar to teacher forcing method used to accelerate training (see [54]). On the other hand, architecture II is much more sensitive to the weights. Gradients tend to be large and only small learning rates can be afforded. For more efficient training, it is therefore beneficial to use architecture I to find a set of weights that already work with reasonable precision and perform fine-tuning with architecture II. In addition, the p_t parameter for architecture I also provides a baseline for the set-up stage length s to be used for architecture II.

Another feature of architecture II is that the length of its prediction stage can be arbitrary. Shorter length limits the extent to which errors can grow and renders the model easier to train. In practice we make sequential improvements to the model weights by progressively increasing the length p of the prediction stage.

Table 1. Summary of hyperparameters for data-driven model architectures.

Category	Symbol	Hyperparameter	Architecture
Layers	n_{LSTM}	number of hidden units, LSTM layer	I & II
	n_{FC}	number of hidden units, fully connected layer	I & II
Series	s	number of time steps, set-up stage	II
	p	number of time steps, prediction stage	I & II
	τ	time step	I & II
Loss	p_t	length of transient (low-weight) period	I
	w_0	transient weight	I
	γ	weight decay	II
Training	n_{batch}	batch size	I & II
	n_{ep}	number of epochs	I & II
	η, β_1, β_2	learning rate and momentum control	I & II

https://doi.org/10.1371/journal.pone.0197704.t001

For convenience, the hyperparameters involved in each architecture are summarized in Table 1.

Fully data-driven modeling

Both of the proposed architectures can be easily adapted for a fully data-driven modeling approach (see [25]): for architecture I the sequence of total dynamics $\{\dot{\boldsymbol{\xi}}_1, \ldots, \dot{\boldsymbol{\xi}}_p\}$ is used as the training target in place of the complementary dynamics and for architecture II the FC layer output is directly integrated to generate the next state. Doing so changes the distribution of model targets and implicitly forces the model to learn more. For comparison, we examine the performance of this fully data-driven approach through the example applications in the following section.

Results and discussion

A chaotic intermittent low-order atmospheric model

We consider a chaotic intermittent low-order atmospheric model, the truncated Charney-DeVore (CDV) equations, developed to model barotropic flow in a β-plane channel with orography. The model formulation used herein is attributed to [55, 56], and employs a slightly different scaling and a more general zonal forcing profile than the original CDV. Systems dynamics are governed by the following ordinary differential equations:

$$\begin{aligned}
\dot{x}_1 &= \gamma_1^* x_3 - C(x_1 - x_1^*), & \dot{x}_2 &= -(\alpha_1 x_1 - \beta_1)x_3 - Cx_2 - \delta_1 x_4 x_6, \\
\dot{x}_3 &= (\alpha_1 x_1 - \beta_1)x_2 - \gamma_1 x_1 - Cx_3 + \delta_1 x_4 x_5, & \dot{x}_4 &= \gamma_2^* x_6 - C(x_4 - x_4^*) + \varepsilon(x_2 x_6 - x_3 x_5), \\
\dot{x}_5 &= -(\alpha_2 x_1 - \beta_2)x_6 - Cx_5 - \delta_2 x_4 x_3, & \dot{x}_6 &= (\alpha_2 x_1 - \beta_2)x_5 - \gamma_2 x_4 - Cx_6 + \delta_2 x_4 x_2,
\end{aligned} \tag{13}$$

where the model coefficients are given by

$$\begin{aligned}
\alpha_m &= \frac{8\sqrt{2}m^2(b^2+m^2-1)}{\pi(4m^2-1)(b^2+m^2)}, & \beta_m &= \frac{\beta b^2}{b^2+m^2}, \\
\delta_m &= \frac{64\sqrt{2}}{15\pi}\frac{b^2-m^2+1}{b^2+m^2}, & \gamma_m^* &= \gamma\frac{4\sqrt{2}mb}{\pi(4m^2-1)}, \\
\varepsilon &= \frac{16\sqrt{2}}{5\pi}, & \gamma_m &= \gamma\frac{4\sqrt{2}m^3 b}{\pi(4m^2-1)(b^2+m^2)},
\end{aligned} \tag{14}$$

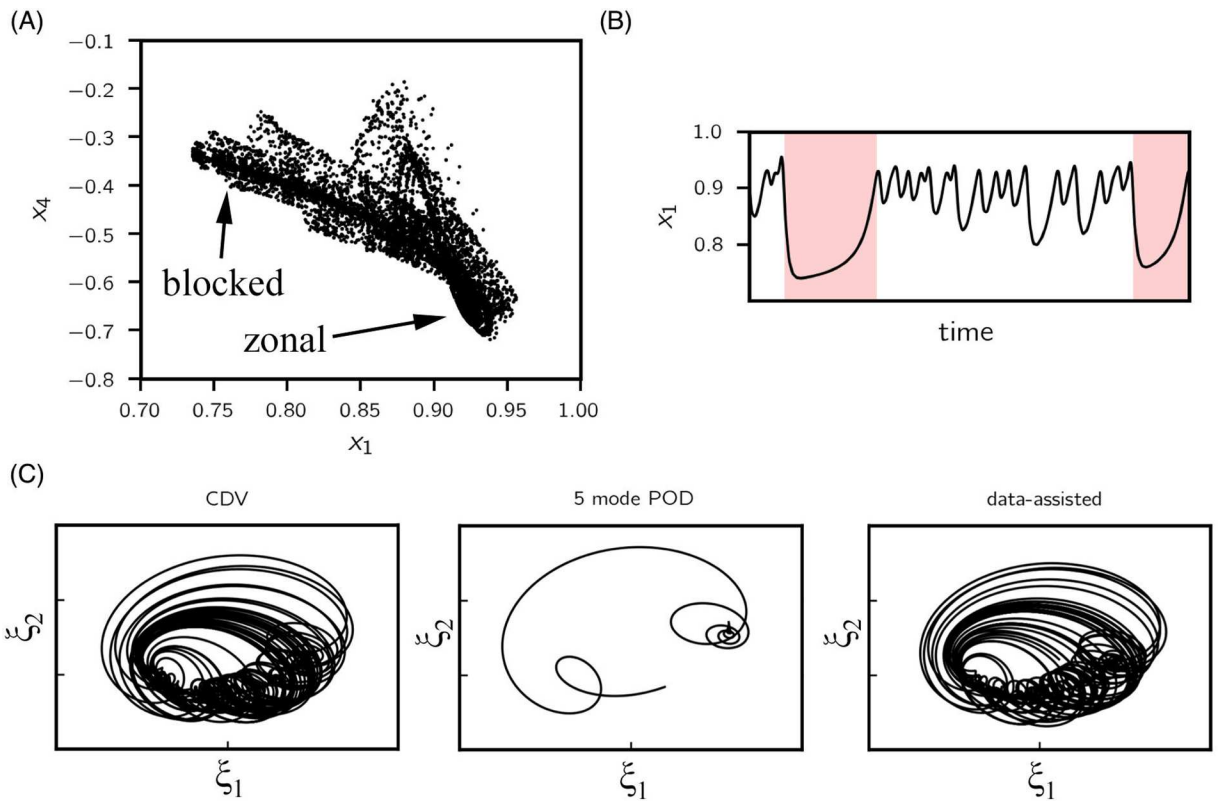

Fig 3. CDV system. (A) 10^4 points sampled from the CDV attractor, projected to (x_1, x_4) plane. (B) Example time series for x_1; *blocked* flow regime is shaded in red. (C) Length-2000 trajectory projected to the first two POD modes (normalized) integrated using the CDV model (left), 5-mode POD projected model (middle) and data-assisted model (right). Despite preserving 99.6% of the total variance, the 5-mode projected model has a single fixed point as opposed to a chaotic attractor. Data-assisted model, however, is able to preserve the geometric features of the original attractor.

https://doi.org/10.1371/journal.pone.0197704.g003

for $m = 1, 2$. Here we examine the system at a fixed set of parameters $(x_1^*, x_4^*, C, \beta, \gamma, b) = (0.95, -0.76095, 0.1, 1.25, 0.2, 0.5)$, which is found to demonstrate chaotic intermittent transitions between *zonal* and *blocked* flow regime, caused by the combination of topographic and barotropic instabilities [55, 56]. These highly transient instabilities render this model an appropriate test case for evaluating the developed methodology. The two distinct regimes are manifested through x_1 and x_4 (Fig 3A and 3B).

For reduction of the system we attempt the classic proper orthogonal decomposition (POD) whose details are described in S1 Appendix. The basis vectors of the projection subspace are calculated using the method of snapshots on a uniformly sampled time series of length 10,000 obtained by integrating Eq (13). The first five POD modes collectively account for 99.6% of the total energy. However, despite providing respectable short-term prediction accuracy, projecting the CDV system to its most energetic five modes completely changes the dynamical behavior and results in a single globally attracting fixed point instead of a strange attractor. The difference between exact and projected dynamics can be seen in terms of the two most energetic POD coefficients, ξ_1, ξ_2, in Fig 3C (left and middle subplots).

In the context of our framework, we construct a data-assisted reduced-order model that includes the dynamics given by the 5-mode POD projection. We set $n_{\mathrm{LSTM}} = 1$ (because one dimension is truncated) and $n_{\mathrm{FC}} = 16$. Input to the FC layer is a concatenation of LSTM output and reduced state because $n_{\mathrm{LSTM}} = 1$ is sufficient to represent the truncated mode. Data is

obtained as 10,000 trajectories, each with $p = 200$ and $\tau = 0.01$. We use 80%, 10%, 10% for training, validation and testing respectively. For this setup it proves sufficient, based on empirical evidence, to train the assisting data-driven model with Architecture I for 1000 epochs, using a batch size of 250. The trained weights are plugged in architecture II to generate sequential predictions. As we quantify next, it is observed that (a) the trajectories behave much like the 6-dimensional CDV system in the long term by forming a similar attractor, as shown in Fig 3C, and (b) the short-term prediction skill is boosted significantly.

We quantify the improvement in prediction performance by using two error metrics—root mean squared error (RMSE) and correlation coefficient. For comparison we also include prediction errors when using a purely data-driven model based on LSTM. RMSE in ith reduced dimension is computed as

$$\mathrm{RMSE}_i(t_l) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\xi_i^{(n)}(t_l) - \hat{\xi}_i^{(n)}(t_l))^2}, \quad i = 1, \ldots, m, \tag{15}$$

where $\xi_i^{(n)}(t_l)$ and $\hat{\xi}_i^{(n)}(t_l)$ represent the truth and prediction for the nth *test* trajectory at prediction lead time t_l respectively. The results are plotted in Fig 4. We end remark that the

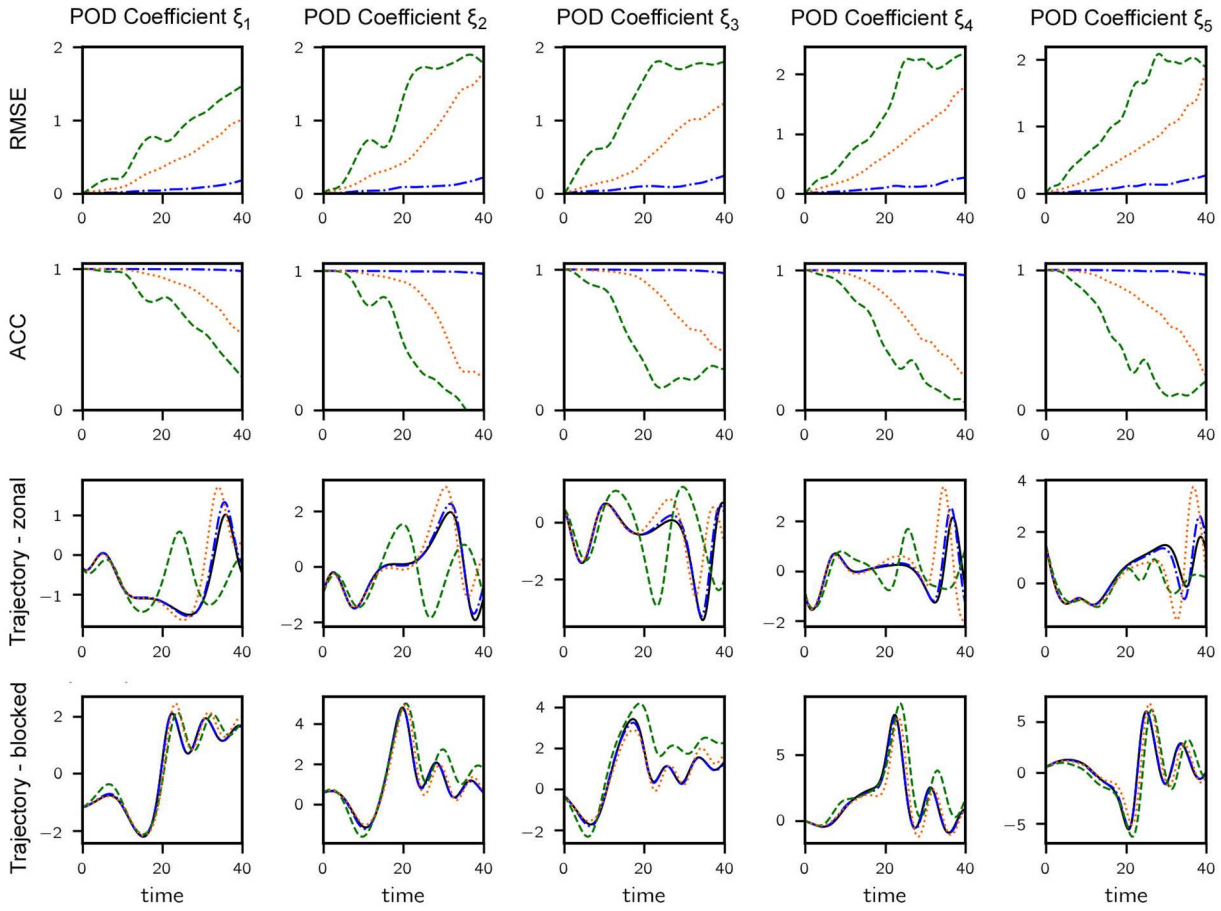

Fig 4. Results for CDV system. (Row 1) RMSE vs. lead time for 5-mode POD projected model (orange dotted), data-assisted model (blue dashdotted) and purely data-driven model (green dashed). (Row 2) ACC vs. lead time. (Row 3) A sample trajectory corresponding to zonal flow—true trajectory is shown (black solid). (Row 4) A sample trajectory involving regime transition (happening around $t = 20$). For rows 1, 3 and 4, plotted values are normalized by the standard deviation of each dimension.

https://doi.org/10.1371/journal.pone.0197704.g004

predictions obtained by the proposed data-assisted model are significantly better than the projected model, as well as than the purely data-driven approach. Low error levels are maintained by the present approach even when the other methods under consideration exhibit significant errors.

The anomaly correlation coefficient (ACC) [57] measures the correlation between anomalies of forecasts and those of the truth with respect to a reference level and is defined as

$$\mathrm{ACC}_i(t_l) = \frac{\sum_{n=1}^{N}(\xi_i^{(n)}(t_l) - \bar{\xi}_i)(\hat{\xi}_i^{(n)}(t_l) - \bar{\xi}_i)}{\sqrt{\sum_{n=1}^{N}(\xi_i^{(n)}(t_l) - \bar{\xi}_i)^2 \sum_{n=1}^{N}(\hat{\xi}_i^{(n)}(t_l) - \bar{\xi}_i)^2}} \tag{16}$$

where $\bar{\xi}_i$ is the reference level set to the observation average by default. ACC takes a maximum value of 1 if the variation pattern of the anomalies of forecast is perfectly coincident with that of truth and a minimum value of -1 if the pattern is completely reversed. Again, the proposed method is able to predict anomaly variation patterns which are almost perfectly correlated with the truth at very large lead times when the predictions made by the compared methods are mostly uncorrelated (Fig 4—second row).

In the third and fourth rows of Fig 4 we illustrate the improvement that we obtain with the data-assisted approach throughout the systems attractor, i.e. in both zonal and blocked regimes. In the third row of Fig 4 the flow in the zonal regime is shown and in the fourth row we demonstrate the flow transitions into the blocked regime around $t = 20$. In both cases, the data-assisted version clearly improves the prediction accuracy.

We emphasize that the presence of the equation-driven part contributes largely to the long-term stability (vs. purely data driven models) while the data-driven part serves to improve the short-term prediction accuracy. These two ingredients of the dynamics complement each other favorably in achieving great prediction performance. In addition, the data-assisted approach successfully produces a chaotic structure that is similar to the one observed in the full-dimensional system, a feat that cannot be replicated by either methods using equation or data alone.

Intermittent bursts of dissipation in Kolmogorov flow

We consider the two-dimensional incompressible Navier-Stokes equations

$$\begin{aligned} \partial_t \mathbf{u} &= -\mathbf{u}\cdot\nabla\mathbf{u} - \nabla p + \nu\Delta\mathbf{u} + \mathbf{f} \\ \nabla\cdot\mathbf{u} &= 0 \end{aligned} \tag{17}$$

where $\mathbf{u} = (u_x, u_y)$ is the fluid velocity defined over the domain $(x, y) \in \Omega = [0, 2\pi] \times [0, 2\pi]$ with periodic boundary conditions, $\nu = 1/Re$ is the non-dimensional viscosity equal to reciprocal of the Reynolds number and p denotes the pressure field over Ω. We consider the flow driven by the monochromatic Kolmogorov forcing $\mathbf{f}(\mathbf{x}) = (f_x, f_y)$ with $f_x = \sin(k_f y)$ and $f_y = 0$. $\mathbf{k}_f = (0, k_f)$ is the forcing wavenumber.

Following [58], the kinetic energy E, dissipation D and energy input I are defined as

$$E(\mathbf{u}) = \frac{1}{|\Omega|}\int_\Omega \frac{1}{2}|\mathbf{u}|^2 \; d\Omega,$$
$$D(\mathbf{u}) = \frac{\nu}{|\Omega|}\int_\Omega |\nabla u|^2 \; d\Omega, \tag{18}$$
$$I(\mathbf{u}) = \frac{1}{|\Omega|}\int_\Omega \mathbf{u}\cdot\mathbf{f} \; d\Omega$$

satisfying the relationship $\dot{E} = I - D$. Here $|\Omega| = (2\pi)^2$ denotes the area of the domain.

The Kolmogorov flow admits a laminar solution $u_x = (Re/k_f^2)\sin(k_f y), u_y = 0$. For sufficiently large k_f and Re, this laminar solution is unstable, chaotic and exhibiting intermittent surges in energy input I and dissipation D. Here we study the flow under a particular set of parameters $Re = 40$ and $k_f = 4$ for which we have the occurrence of extreme events. Fig 5A shows the bursting time series of the dissipation D along a sample trajectory.

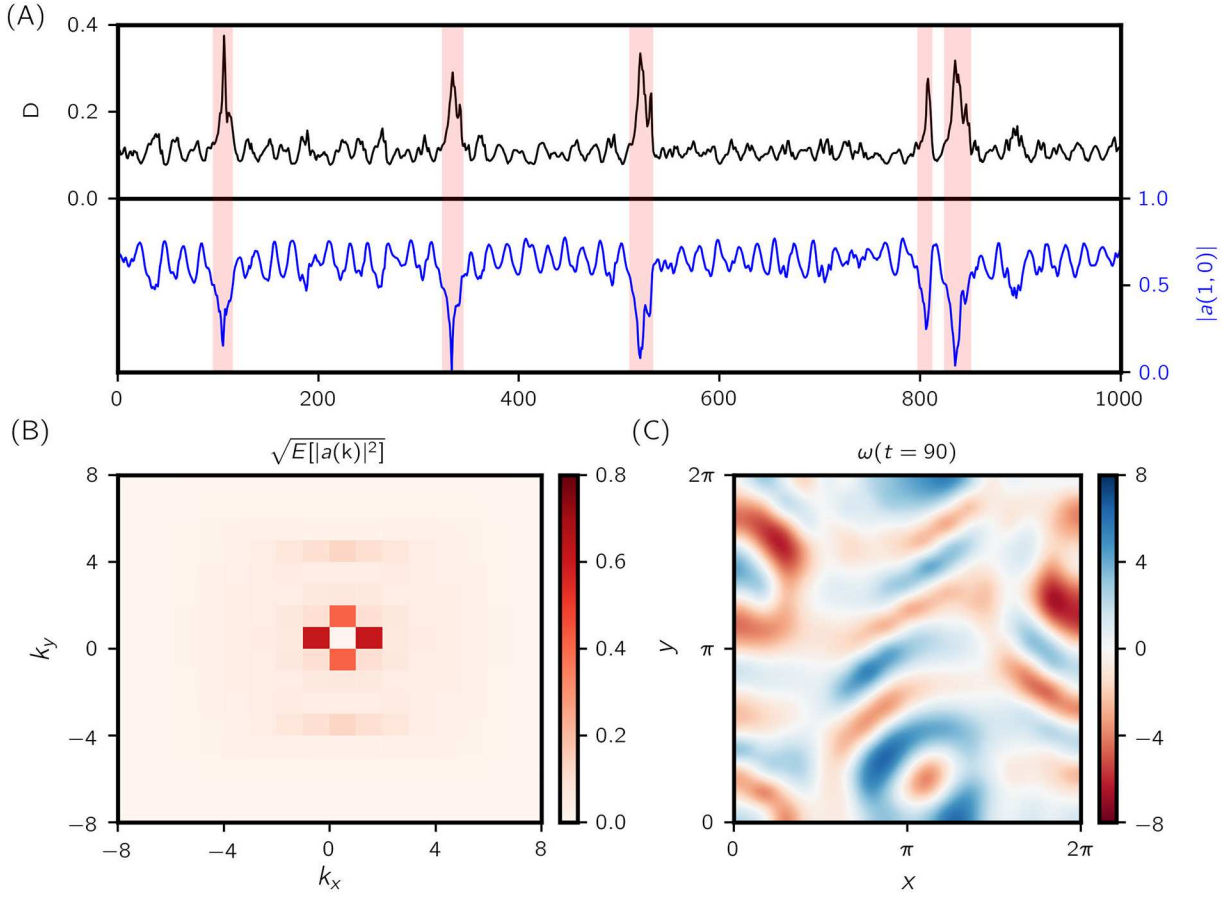

Fig 5. Kolmogorov flow. (A) Time series of energy dissipation rate D and Fourier coefficient modulus $|a(1, 0)|$—rare events are signaled by burst in D and sudden dip in $|a(1, 0)|$. (B) Root mean squared (RMS) modulus for wavenumbers $-8 \le k_1, k_2 \le 8$. (C) Vorticity field $\nabla \times \mathbf{u} = \omega$ at time $t = 90$ over the domain $\mathbf{x} \in [0, 2\pi] \times [0, 2\pi]$.

https://doi.org/10.1371/journal.pone.0197704.g005

Due to spatial periodicity, it is natural to examine the velocity field in Fourier space. The divergence-free velocity field $\mathbf{u}$ admits the following Fourier series expansion:

$$\mathbf{u}(\mathbf{x},t)=\sum_{\mathbf{k}}\frac{a(\mathbf{k},t)}{|\mathbf{k}|}\begin{pmatrix}k_2\\-k_1\end{pmatrix}e^{i\mathbf{k}\cdot\mathbf{x}} \tag{19}$$

where $\mathbf{k}=(k_1,k_2)$ is the wavenumber and $a(\mathbf{k},t)=-\overline{a(-\mathbf{k},t)}$ for $\mathbf{u}$ to be real-valued. For notation clarity, we will not explicitly write out the dependence on t from here on. Substituting Eq (19) into the governing equations Eq (17) we obtain the evolution equations for a as (more details are presented in S1 Appendix)

$$\dot{a}(\mathbf{k})=\sum_{\mathbf{p}+\mathbf{q}=\mathbf{k}}i\frac{(p_1q_2-p_2q_1)(k_1q_1+k_2q_2)}{|\mathbf{p}||\mathbf{q}||\mathbf{k}|}a(\mathbf{p})a(\mathbf{q})-\nu|\mathbf{k}|^2a(\mathbf{k})-\frac{1}{2}i(\delta_{\mathbf{k},\mathbf{k}_f}+\delta_{\mathbf{k},-\mathbf{k}_f}) \tag{20}$$

The first term suggests that any mode with wavenumber $\mathbf{k}$ is directly affected, in a nonlinear fashion, by pairs of modes with wavenumbers $\mathbf{p}$ and $\mathbf{q}$ such that $\mathbf{k}=\mathbf{p}+\mathbf{q}$. A triplet of modes $\{\mathbf{p},\mathbf{q},\mathbf{k}\}$ satisfying this condition is referred to as a *triad*. It is worth noting that a mode which does not form a triad with mode $\mathbf{k}$ can still have an indirect effect on dynamics $\dot{a}(\mathbf{k})$ through interacting with modes that do form a triad with $\mathbf{k}$.

In [58], it is found that the most revealing triad interaction to observe, in the interest of predicting intermittent bursts in the energy input/dissipation, is amongst modes $(0,k_f)$, $(1,0)$ and $(1,k_f)$. Shortly prior to an intermittent event, mode $(1,0)$ transfers a large amount of energy to mode $(0,k_f)$, leading to rapid growth in the energy input rate I and subsequently the dissipation rate D (see Fig 5A). However, projecting the velocity field and dynamics to this triad of modes and their complex conjugates fails to faithfully replicate the dynamical behaviors of the full system (the triad only accounts for 59% of the total energy; Fourier energy spectrum is shown in Fig 5B). We use the present framework to complement the projected triad dynamics.

Quantities included in the model are $a(1,0)$, $a(0,k_f)$ and $a(1,k_f)$ and their conjugate pairs, which amount to a total of six independent dimensions. Data is generated as a single time series of length 10^5 at $\Delta t=1$ intervals, by integrating the full model equations Eq (17) using a spectral grid of size 32 (wavenumbers truncated to $-16\leq k_1,k_2\leq 16$) [59]. Each data point is then taken as an initial condition from which a trajectory of length 1 (200 steps of 0.005) is obtained. The sequence of states along the trajectory is projected to make up the 6-dimensional input to the LSTM model. The ground truth total dynamics is again approximated with first-order finite differences. The first 80% of the data is used for training, 5% for validation and the remaining 15% for testing.

For this problem it is difficult to compute the true minimum parametrizing dimension so we conservatively choose $n_{\mathrm{LSTM}}=70$ and $n_{\mathrm{FC}}=38$. It is found that the models do not tend to overfit, nor is their performance sensitive to these hyper-parameters around the chosen value. Since the number of hidden units used in the LSTM is large relative to the input dimension, they are not concatenated with the input before entering the output layer. We first perform pre-training with architecture I for 1000 epochs and fine-tune the weights with architecture II. Due to the low-energy nature of the reduction space, transient effects are prominent (see S1 Appendix) and thus a sizable set-up stage is needed for training and prediction with architecture II. Using a sequential training strategy, we keep $s=100$ fixed and progressively increase prediction length at $p=\{10,30,50,100\}$ (see S1 Appendix). At each step, weights are optimized for 1000 epochs using a batch size of 250. The hyperparameters defining the loss functions are $p_t=60$, $w_0=0.01$ and $\gamma=0.98$, which are found empirically to result in favorable weight convergence.

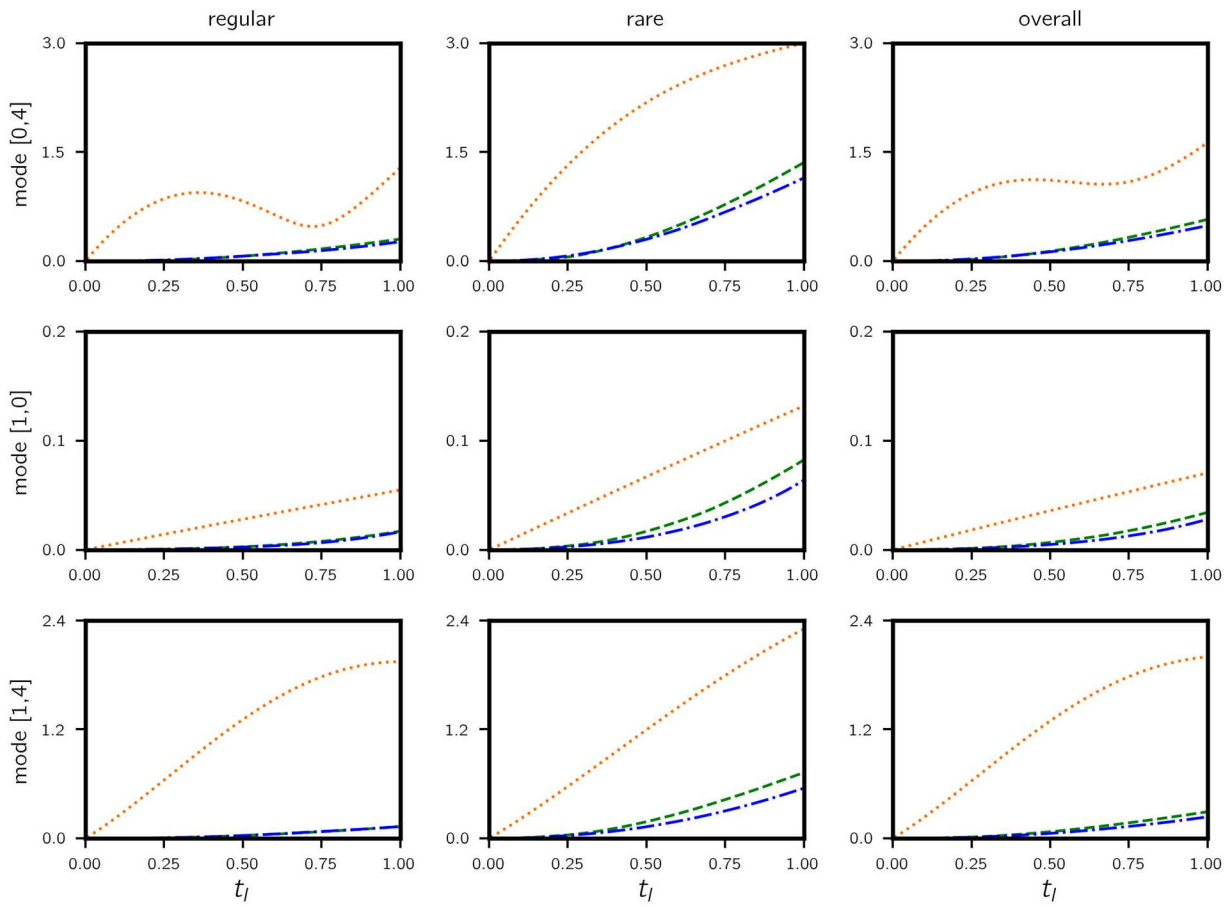

Fig 6. Kolmogorov flow—RMSE vs. time. Errors are computed for 10^4 test trajectories (legend: fully data-driven—green dashed; data-assisted—blue dashdotted; triad—orange dotted). The RMSE in each mode is normalized by the corresponding amplitude $E(\mathbf{k}) = \sqrt{\mathbb{E}[|a(\mathbf{k})|^2]}$. A test trajectory is classified as *regular* if $|a(1,0)| > 0.4$ at $t = 0$ and *rare* otherwise. Performance for regular, rare and all trajectories are shown in three columns. Data-assisted model has very similar errors to those of purely data-driven models for regular trajectories, but the performance is visibly improved for rare events.

https://doi.org/10.1371/journal.pone.0197704.g006

Similar to the CDV system, we measure the prediction performance using RMSE and correlation coefficient. Since the modeled Fourier coefficients are complex-valued, the sum in Eq (15) is performed on the squared complex magnitude of the absolute error.

The resulting normalized test error curves are shown in Figs 6 and 7 respectively, comparing the proposed data-assisted framework with the original projected model and the fully data-driven approach as the prediction lead time increases. At 0.5 lead time (approximately 1 eddy turn-over time t_e), the data-assisted approach achieves 0.13, 0.005, 0.058 RMSE in mode [0, 4], [1, 0] and [1, 4] respectively. Predictions along a sample trajectory is shown in Fig 8.

Overall, the data-assisted approach produces the lowest error, albeit narrowly beating the fully data-driven model but significantly outperforming the projected model (88%, 86% and 95% reduction in error for the three modes). This is because data is used to assist a projected model that ignores a considerable amount of state information which contribute heavily to the dynamics. It is therefore all up to the data-driven model to learn this missing information. For this reason we observe similar performance between data-assisted and fully data-driven models. However, when we classify the test cases into regular and rare events based on the value of $|a(1, 0)|$ and examine the error performance separately, the advantage of the data-assisted

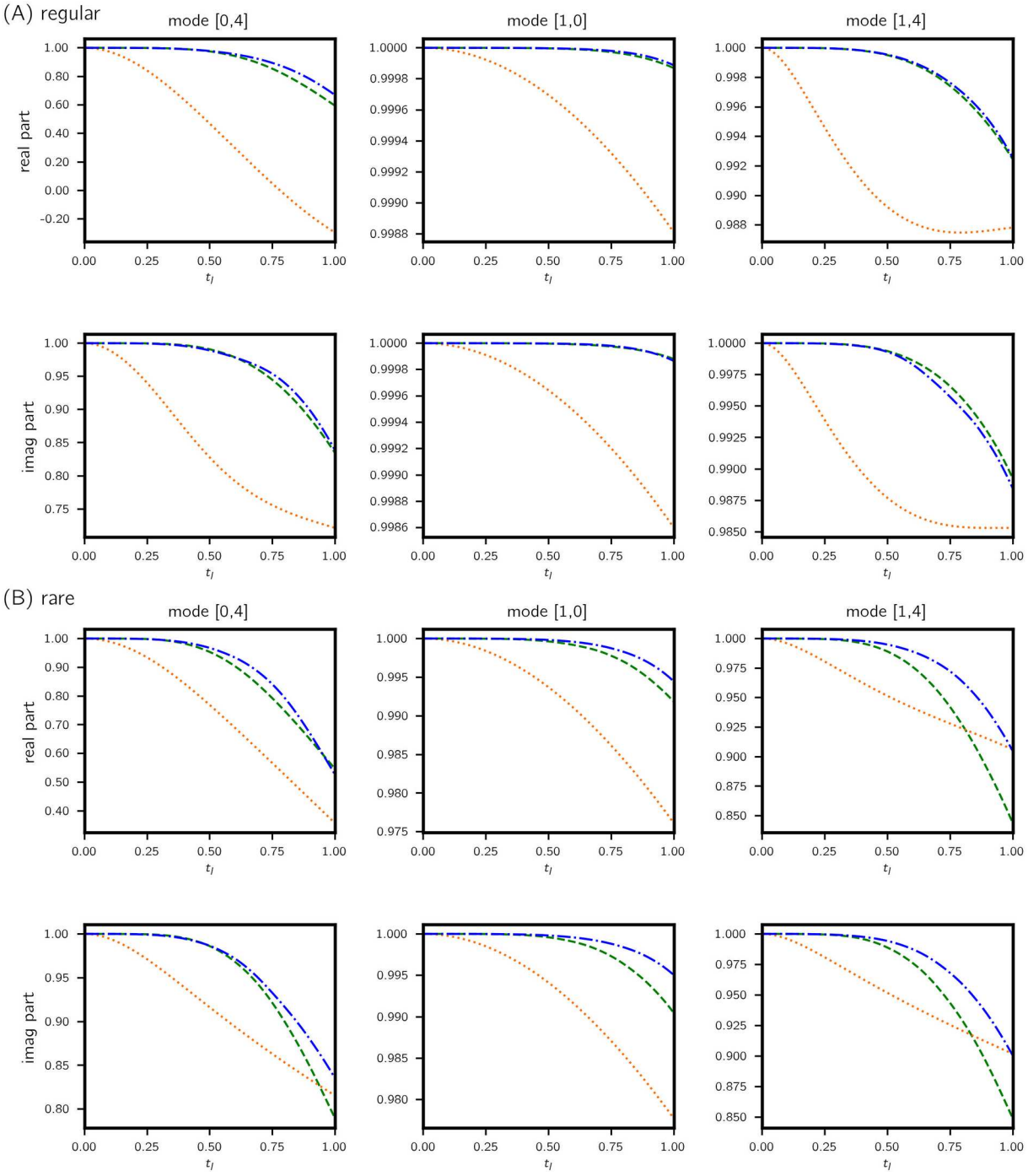

Fig 7. Kolmogorov flow: ACC vs. time. Values are computed for (A) regular and (B) rare trajectories classified from 10^4 test cases. Legend: fully data-driven—green dashed; data-assisted—blue dashdotted; triad dynamics—orange dotted. Real and imaginary parts are treated independently. Similarly to RMSE in Fig 6, improvements in predictions made by the data-assisted model are more prominent for rare events.

https://doi.org/10.1371/journal.pone.0197704.g007

approach is evident in the latter category, especially for mode (1, 4). The is mainly due to (a) rare events appear less frequently in data such that the corresponding dynamics is not learned well compared to regular events and (b) the triad of Fourier modes selected play more prominent role in rare events and therefore the projected dynamics contain relevant dynamical

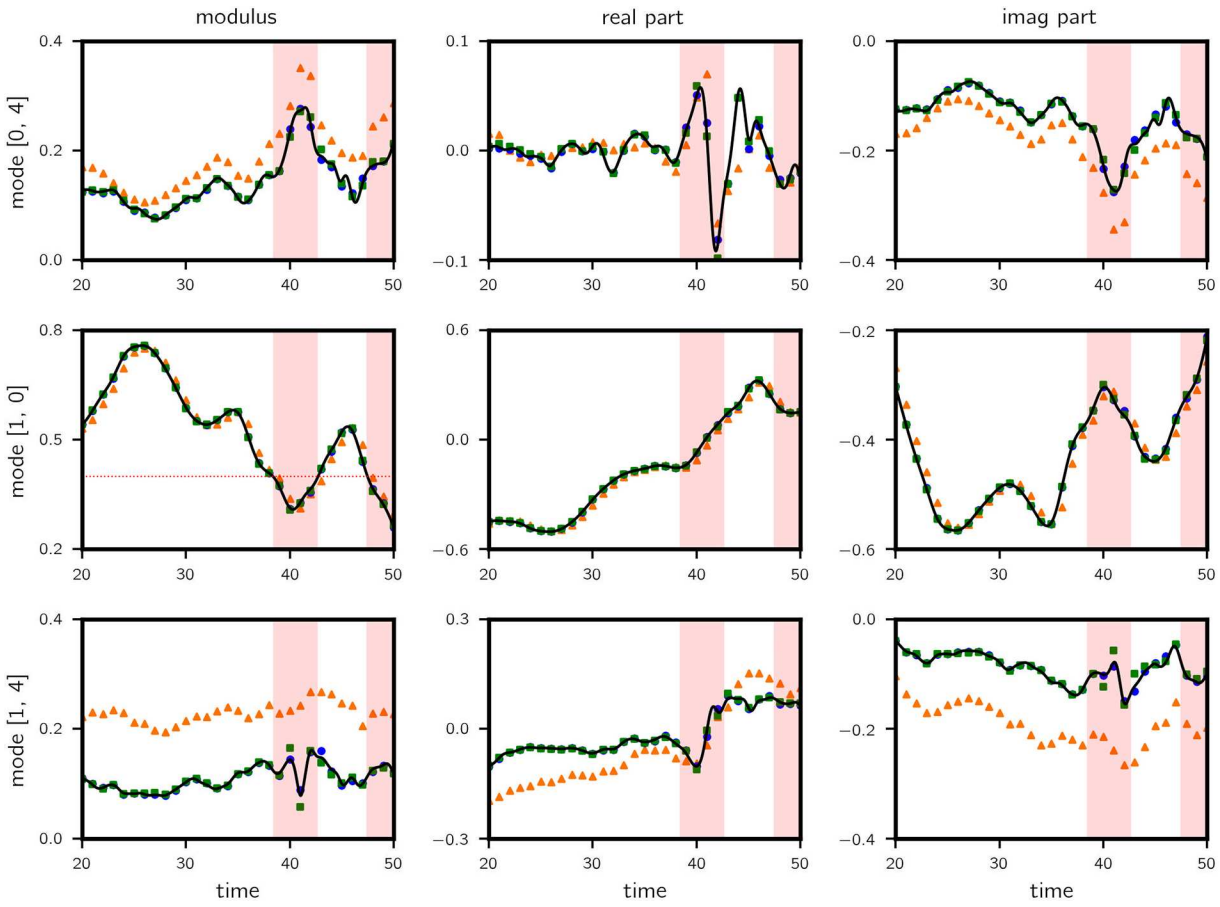

Fig 8. Kolmogorov flow: Predictions along a sample trajectory with lead time = 0.5. Results for the complex modulus (left column), real part (middle column) and imaginary part (right column) of the wavenumber triad are shown. Legend: truth—black solid line; data-assisted—blue circle; triad dynamics—orange triangle; purely data-driven—green square. Rare events are recorded when $|a(1, 0)|$ (left column, mid row) falls below 0.4 (shaded in red). Significant improvements are observed for wavenumbers (0, 4) and (1, 4).

https://doi.org/10.1371/journal.pone.0197704.g008

information. Nevertheless, errors for rare events are visibly higher (about 5 times), attesting to their unpredictable nature in general.

To better understand the favorable properties of the hybrid scheme when it comes to the prediction of extreme events we plot the probability density function (pdf) of complementary and total dynamics (see Eq (7)), calculated from the 10^5-point training data set with a kernel density estimator (Fig 9). The dynamics values are standardized so that 1 unit in horizontal axis represent 1 standard deviation. We immediately notice that total dynamics in every dimension have a fat-tailed distribution. This signifies that the data set contains several *extreme* observations, more than 10 standard deviations away from the mean. For data-driven models these dynamics are difficult to learn due to their sporadic occurrence in sample data and low density in phase space.

In contrast, the marginal pdf of the complementary dynamics have noticeably different characteristics, especially in both the real and imaginary parts of mode (1, 4) (and to a smaller degree for the real part of mode (0, 4)). The distribution is bimodal-like; more importantly, density falls below 10^{-4} level within 3 standard deviations. Because of its concentrated character, this is a much better conditioned target distribution, as the data-driven scheme would never have to learn extreme event dynamics; these are captured by the projected equations. As

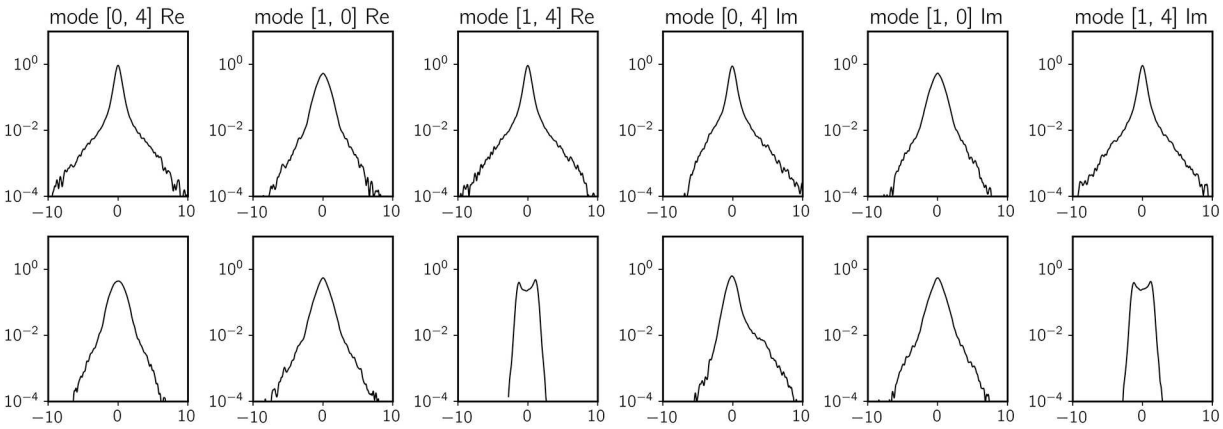

Fig 9. Marginal probability density function of total dynamics (top row) and complementary dynamics (bottom row). Horizontal axes are scaled by standard deviations of each quantity. For the real and imaginary parts of mode (1, 4) (and real part of mode (0, 4) to a smaller degree) triad dynamics help remove large-deviation observations in the complementary dynamics.

https://doi.org/10.1371/journal.pone.0197704.g009

expected, the error plot in Fig 7 suggests that the biggest improvement from a purely data-driven approach to a data-assisted approach is indeed for mode (1, 4).

Conclusion

We introduce a data assisted framework for reduced-order modeling and prediction of extreme transient events in complex dynamical systems with high-dimensional attractors. The framework utilizes a data-driven approach to complement the dynamics given by imperfect models obtained through projection, i.e. in cases when the projection subspace does not perfectly parametrize the inertial manifold of the system. Information which is invisible to the subspace but important to the dynamics is extracted by analyzing the time history of trajectories (data-streams) projected in the subspace, using a RNN strategy. The LSTM based architecture of the employed RNN allows for the modeling of the dynamics using delayed coordinates, a feature that significantly improves the performance of the scheme, complementing observation in fully data-driven schemes.

We showcase the capabilities of the present approach through two illustrative examples exhibiting intermittent bursts: a low dimensional atmospheric system, the Charney-DeVore model and a high-dimensional system, the Kolmogorov flow described by Navier-Stokes equations. For the former the data-driven model helps to improve significantly the short-term prediction skill in a high-energy reduction subspace, while faithfully replicating the chaotic attractor of the original system. In the infinite-dimensional example it is clearly demonstrated that in regions characterized by extreme events the data-assisted strategy is more effective than the fully data-driven prediction or the projected equations. On the other hand, when we consider the performance close to the main attractor of the dynamical system the purely data-driven approach and the data-assisted scheme exhibit comparable accuracy.

The present approach provides a non-parametric framework for the improvement of imperfect models through data-streams. For regions where data is available we obtain corrections for the model, while for regions where no data is available the underlying model still provides a baseline for prediction. The results in this work emphasize the value of this hybrid strategy for the prediction of extreme transient responses for which data-streams may not contain enough information. In the examples considered the imperfect models where obtained through projection to low-dimensional subspaces. It is important to emphasize that such

imperfect models should contain relevant dynamical information for the modes associated to extreme events. These modes are not always the most energetic modes (as illustrated in the fluids example) and numerous efforts have been devoted for their characterization [58, 60, 61].

Apart of the modeling of extreme events, the developed blended strategy should be of interest for data-driven modeling of systems exhibiting singularities or singular perturbation problems. In this case the governing equations have one component that is particularly challenging to model with data, due to its singular nature. For such systems it is beneficial to combine the singular part of the equation with a data-driven scheme that will incorporate information from data-streams. Future work will focus on the application of the formulated method in the context of predictive control [62–64] for turbulent fluid flows and in particular for the suppression of extreme events.

Supporting information

S1 Appendix. Supplementary notes. In the notes we provide some background theory on recurrent neural networks, LSTM and momentum based optimization methods. Additional computation results for CDV and Kolmogorov flow are also included.
(PDF)

S1 Code. Python source code. All code used in this study is available at: https://github.com/zhong1wan/data-assisted. In addition, all training and testing data files are available from the authors upon request.
(ZIP)

Acknowledgments

TPS and ZYW have been supported through the AFOSR-YIA FA9550-16-1-0231, the ARO MURI W911NF-17-1-0306 and the ONR MURI N00014-17-1-2676. PK and PV acknowledge support by the Advanced Investigator Award of the European Research Council (Grant No: 341117).

Author Contributions

Conceptualization: Zhong Yi Wan, Petros Koumoutsakos, Themistoklis Sapsis.

Formal analysis: Zhong Yi Wan.

Funding acquisition: Petros Koumoutsakos, Themistoklis Sapsis.

Investigation: Zhong Yi Wan.

Methodology: Zhong Yi Wan, Themistoklis Sapsis.

Software: Zhong Yi Wan.

Supervision: Petros Koumoutsakos, Themistoklis Sapsis.

Validation: Zhong Yi Wan.

Visualization: Zhong Yi Wan.

Writing – original draft: Zhong Yi Wan, Themistoklis Sapsis.

Writing – review & editing: Zhong Yi Wan, Pantelis Vlachas, Petros Koumoutsakos, Themistoklis Sapsis.

References

1. Smooke MD, Mitchell RE, Keyes DE. Numerical Solution of Two-Dimensional Axisymmetric Laminar Diffusion Flames. Combustion Science and Technology. 1986; 67(4-6):85–122. https://doi.org/10.1080/00102208908924063
2. Pope SB. Computationally efficient implementation of combustion chemistry using'in situ' adaptive tabulation. Combustion Theory and Modelling. 1997; 1(1):41–63. https://doi.org/10.1080/713665229
3. Chandler GJ, Kerswell RR. Invariant recurrent solutions embedded in a turbulent two-dimensional Kolmogorov flow. J Fluid Mech. 2013; 722:554–595. https://doi.org/10.1017/jfm.2013.122
4. Hamilton JM, Kim J, Waleffe F. Regeneration mechanisms of near-wall turbulence structures. J Fluid Mech. 1995; 287:243. https://doi.org/10.1017/S0022112095000978
5. Majda AJ. Challenges in Climate Science and Contemporary Applied Mathematics. Communications on Pure and Applied Mathematics. 2012; 65:920. https://doi.org/10.1002/cpa.21401
6. Majda AJ. Real world turbulence and modern applied mathematics. In: Mathematics: Frontiers and Perspectives, International Mathematical Union. American Mathematical Society; 2000. p. 137–151.
7. Grabowski WW, Smolarkiewicz PK. CRCP: a Cloud Resolving Convection Parameterization for modeling the tropical convecting atmosphere. Physica D: Nonlinear Phenomena. 1999; 133(1-4):171–178. https://doi.org/10.1016/S0167-2789(99)00104-9
8. Grabowski WW. Coupling Cloud Processes with the Large-Scale Dynamics Using the Cloud-Resolving Convection Parameterization (CRCP). Journal of the Atmospheric Sciences. 2001; 58(9):978–997. https://doi.org/10.1175/1520-0469(2001)058%3C0978:CCPWTL%3E2.0.CO;2
9. Akhmediev N, Dudley JM, Solli DR, Turitsyn SK. Recent progress in investigating optical rogue waves. Journal of Optics. 2013; 15(6):60201. https://doi.org/10.1088/2040-8978/15/6/060201
10. Arecchi FT, Bortolozzo U, Montina A, Residori S. Granularity and inhomogeneity are the joint generators of optical rogue waves. Physical Review Letters. 2011; 106(15):2–5. https://doi.org/10.1103/PhysRevLett.106.153901
11. Onorato M, Residori S, Bortolozzo U, Montina A, Arecchi FT. Rogue waves and their generating mechanisms in different physical contexts. Physics Reports. 2013; 528(2):47–89. https://doi.org/10.1016/j.physrep.2013.03.001
12. Muller P, Garrett C, Osborne A. Rogue Waves. Oceanography. 2005; 18(3):66–75.
13. Fedele F. Rogue waves in oceanic turbulence. Physica D: Nonlinear Phenomena. 2008; 237(14):2127–2131. https://doi.org/10.1016/j.physd.2008.01.022
14. Kim E, Li F, Chong C, Theocharis G, Yang J, Kevrekidis PG. Highly Nonlinear Wave Propagation in Elastic Woodpile Periodic Structures. Phys Rev Lett. 2015; 114:118002. https://doi.org/10.1103/PhysRevLett.114.118002 PMID: 25839312
15. Li F, Anzel P, Yang J, Kevrekidis PG, Daraio C. Granular acoustic switches and logic gates. Nature Communications. 2014; 5:5311.
16. Farmer JD, Sidorowich JJ. Predicting chaotic time series. Physical Review Letters. 1987; 59(8):845–848. https://doi.org/10.1103/PhysRevLett.59.845 PMID: 10035887
17. Crutchfield JP, McNamara BS. Equations of motion from a data series. Comp Sys. 1987; 1:417–452.
18. Sugihara G, May RM. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature. 1990; 344(6268):734–741. https://doi.org/10.1038/344734a0 PMID: 2330029
19. Rowlands G, Sprott JC. Extraction of dynamical equations from chaotic data. Physica D. 1992; 58:251–259. https://doi.org/10.1016/0167-2789(92)90113-2
20. Berry T, Giannakis D, Harlim J. Nonparametric forecasting of low-dimensional dynamical systems. Physical Review E. 2015; 91(3):032915. https://doi.org/10.1103/PhysRevE.91.032915
21. Berry T, Harlim J. Forecasting Turbulent Modes with Nonparametric Diffusion Models: Learning from Noisy Data. Physica D. 2016; 320:57–76. https://doi.org/10.1016/j.physd.2016.01.012
22. Kevrekidis IG, Gear CW, Hyman JM, Kevrekidis PG, Runborg O, Theodoropoulos C. Equation-Free, Coarse-Grained Multiscale Computation: Enabling Mocroscopic Simulators to Perform System-Level Analysis. Comm Math Sci. 2003; 1:715–762. https://doi.org/10.4310/CMS.2003.v1.n4.a5
23. Sirisup S, Xiu D, Karniadakis GE, Kevrekidis IG. Equation-free/Galerkin-free POD-assisted computation of incompressible flows. J Comput Phys. 2005; 207:568–587. https://doi.org/10.1016/j.jcp.2005.01.024
24. Wan ZY, Sapsis TP. Reduced-space Gaussian Process Regression for data-driven probabilistic forecast of chaotic dynamical systems. Physica D: Nonlinear Phenomena. 2017; 345. https://doi.org/10.1016/j.physd.2016.12.005

25. Vlachas PR, Byeon W, Wan ZY, Sapsis TP, Koumoutsakos P. Data-Driven Forecasting of High Dimensional Chaotic Systems with Long-Short Term Memory Networks. Submitted. 2018;.
26. Bongard J, Lipson H. Automated reverse engineering of nonlinear dynamical systems. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(24):9943–9948. https://doi.org/10.1073/pnas.0609476104 PMID: 17553966
27. Schmidt M, Lipson H. Distilling Free-Form Natural Laws from Experimental Data. Science. 2009; 324 (5923). https://doi.org/10.1126/science.1165893
28. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences of the United States of America. 2016; 113(15):3932–7. https://doi.org/10.1073/pnas.1517384113 PMID: 27035946
29. Raissi M, Perdikaris P, Karniadakis GE. Machine learning of linear differential equations using Gaussian processes. J Comp Phys. 2017; 348:683–693. https://doi.org/10.1016/j.jcp.2017.07.050
30. Hochreiter Sepp and Schmidhueber Juergen. Long short-term memory. Neural Computation. 1997; 9 (8):1–32.
31. Whitney H. Differentiable manifolds. Ann Math. 1936; 37:645–680. https://doi.org/10.2307/1968482
32. Takens F. Detecting strange attractors in turbulence. Lecture notes in mathematics 898. 1981; p. 366–381. https://doi.org/10.1007/BFb0091924
33. Pathak J, Hunt BR, Girvan M, Lu Z, Ott E. Model-free prediction of large spatiotemporally chaotic systems from data: a reservoir computing approach. Physical Review Letters. 2018; 120(024102).
34. Pathak J, Wikner A, Fussell R, Chandra S, Hunt BR, Girvan M, et al. Hybrid forecasting of chaotic processes: using machine learning in conjunction with a knowledge-based model. ArXiv e-prints. 2018;.
35. Sapsis TP, Majda AJ. Statistically Accurate Low Order Models for Uncertainty Quantification in Turbulent Statistically Accurate Low Order Models for Uncertainty Quantification in Turbulent Dynamical Systems. Proceedings of the National Academy of Sciences. 2013; 110:13705–13710. https://doi.org/10.1073/pnas.1313065110
36. Sapsis TP, Majda AJ. A statistically accurate modified quasilinear Gaussian closure for uncertainty quantification in turbulent dynamical systems. Physica D. 2013; 252:34–45. https://doi.org/10.1016/j.physd.2013.02.009
37. Sapsis TP, Majda AJ. Blending Modified Gaussian Closure and Non-Gaussian Reduced Subspace methods for Turbulent Dynamical Systems. Journal of Nonlinear Science (In Press). 2013;. https://doi.org/10.1007/s00332-013-9178-1
38. Milano M, Koumoutsakos P. Neural network modeling for near wall turbulent flow. Journal of Computational Physics. 2002; 182(1):1–26. https://doi.org/10.1006/jcph.2002.7146
39. Bright I, Lin G, Kutz JN. Compressive sensing based machine learning strategy for characterizing the flow around a cylinder with limited pressure measurements. Physics of Fluids. 2013; 25(12):127102. https://doi.org/10.1063/1.4836815
40. Ling J, Jones R, Templeton J. Machine learning strategies for systems with invariance properties. Journal of Computational Physics. 2016; 318:22–35. https://doi.org/10.1016/j.jcp.2016.05.003
41. Guniat F, Mathelin L, Hussaini MY. A statistical learning strategy for closed-loop control of fluid flows. Theoretical and Computational Fluid Dynamics. 2016; 30(6):497–510. https://doi.org/10.1007/s00162-016-0392-y
42. Wang JX, Wu JL, Xiao H. Physics-informed machine learning approach for reconstructing Reynolds stress modeling discrepancies based on DNS data. Phys Rev Fluids. 2017; 2:034603. https://doi.org/10.1103/PhysRevFluids.2.034603
43. Weymouth GD, Yue DKP. Physics-Based Learning Models for Ship Hydrodynamics. Journal of Ship Research. 2014; 57(1):1–12. https://doi.org/10.5957/JOSR.57.1.120005
44. Sanner RM, Slotine JJE. Stable Adaptive Control of Robot Manipulators Using "Neural" Networks. Neural Comput. 1995; 7(4):753–790. https://doi.org/10.1162/neco.1995.7.4.753
45. Kevrekidis RRMJSAIG. Continuous-time nonlinear signal processing: a neural network based approach for gray box identification. In: Neural Networks for Signal Processing [1994] IV. Proceedings of the 1994 IEEE Workshop; 1994.
46. Lagaris IE, Likas A, Fotiadis DI. Artificial neural networks for solving ordinary and partial differential equations. IEEE Transactions on Neural Networks. 1998; 9(5):987–1000. https://doi.org/10.1109/72.712178 PMID: 18255782
47. Raissi M, Perdikaris P, Karniadakis GE. Physics Informed Deep Learning (Part I): Data-driven solutions of nonlinear partial differential equations. SIAM J Sci Comput. 2018;In press.
48. Matthies HG, Meyer M. Nonlinear Galerkin methods for the model reduction of nonlinear dynamical systems. Comput Struct. 2003; 81:1277–1286. https://doi.org/10.1016/S0045-7949(03)00042-7

49. Sapsis TP, Majda AJ. Blended reduced subspace algorithms for uncertainty quantification of quadratic systems with a stable mean state. Physica D. 2013; 258:61. https://doi.org/10.1016/j.physd.2013.05.004

50. Foias C, Jolly MS, Kevrekidis IG, Sell GR, Titi ES. On the computation of inertial manifolds. Phys Lett A. 1988; 131(7-8):433–436. https://doi.org/10.1016/0375-9601(88)90295-2

51. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. CoRR. 2014;abs/1412.6980.

52. Sutskever I, Vinyals O, Le QV. Sequence to seqence learning with neural networks. In: Proc. NIPS. Montreal, CA; 2014.

53. Britz D, Goldie A, Luong T, Le Q. Massive Exploration of Neural Machine Translation Architectures. ArXiv e-prints. 2017;.

54. Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. IEEE Neural Computation. 1989; 1:270–280. https://doi.org/10.1162/neco.1989.1.2.270

55. Crommelin DT, Opsteegh JD, Verhulst F. A mechanism for atmospheric regime behavior. J Atmos Sci. 2004; 61(12):1406–1419. https://doi.org/10.1175/1520-0469(2004)061%3C1406:AMFARB%3E2.0.CO;2

56. Crommelin DT, Majda AJ. Strategies for model reduction: comparing different optimal bases. J Atmos Sci. 2001; 61(17):2206–2217. https://doi.org/10.1175/1520-0469(2004)061%3C2206:SFMRCD%3E2.0.CO;2

57. Jolliffe IT, Stephensen DB, editors. Forecast verification: a practitioner's guide in atmospheric science. Wiley and Sons; 2012.

58. Farazmand M, Sapsis TP. A variational approach to probing extreme events in turbulent dynamical systems. Sci Adv. 2017; 3(9). https://doi.org/10.1126/sciadv.1701533 PMID: 28948226

59. Uecker H. A short ad hoc introduction to spectral methods for parabolic pde and the Navier-Stokes equations. Summer School Modern Computational Science. 2009; p. 169–209.

60. Cousins W, Sapsis TP. Reduced order prediction of rare events in unidirectional nonlinear water waves. Submitted to Journal of Fluid Mechanics. 2015;.

61. Farazmand M, Sapsis T. Dynamical indicators for the prediction of bursting phenomena in high-dimensional systems. Physical Review E. 2016;032212:1–31.

62. Sanner RM, Slotine JJE. Gaussian networks for direct adaptive control. IEEE Transactions on Neural Networks. 1992; 3:837–863. https://doi.org/10.1109/72.165588 PMID: 18276483

63. Proctor JL, Brunton SL, Kutz JN. Dynamic mode decomposition with control. SIAM J Appl Dyn Syst. 2016; 15:142–161. https://doi.org/10.1137/15M1013857

64. Korda M, Mezic I. Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. ArXiv e-prints. 2016;.