Learning latent dynamics for partially observed chaotic systems

Cite as: Chaos **30**, 103121 (2020); https://doi.org/10.1063/5.0019309
Submitted: 24 June 2020 . Accepted: 23 September 2020 . Published Online: 20 October 2020

S. Ouala, D. Nguyen, L. Drumetz, B. Chapron, A. Pascual, F. Collard, L. Gaultier, and R. Fablet

Learning latent dynamics for partially observed chaotic systems

Cite as: Chaos **30**, 103121 (2020); doi: 10.1063/5.0019309
Submitted: 24 June 2020 · Accepted: 23 September 2020 ·
Published Online: 20 October 2020

S. Ouala,[1,a)] D. Nguyen,[1] L. Drumetz,[1] B. Chapron,[2] A. Pascual,[3] F. Collard,[4] L. Gaultier,[4] and R. Fablet[1,b)]

AFFILIATIONS

[1]IMT Atlantique, UMR CNRS Lab-STICC, 29280 Plouzané, France
[2]Ifremer, LOPS, 29280 Plouzané, France
[3]IMEDEA, UIB-CSIC, 07190 Esporles, Spain
[4]OceanDataLab, 29280 Locmaria-Plouzané, France

[a)]**Author to whom correspondence should be addressed:** said.ouala@imt-atlantique.fr
[b)]**Electronic mail:** ronan.fablet@imt-atlantique.fr

ABSTRACT

This paper addresses the data-driven identification of latent representations of partially observed dynamical systems, i.e., dynamical systems for which some components are never observed, with an emphasis on forecasting applications and long-term asymptotic patterns. Whereas state-of-the-art data-driven approaches rely in general on delay embeddings and linear decompositions of the underlying operators, we introduce a framework based on the data-driven identification of an augmented state-space model using a neural-network-based representation. For a given training dataset, it amounts to jointly reconstructing the latent states and learning an ordinary differential equation representation in this space. Through numerical experiments, we demonstrate the relevance of the proposed framework with respect to state-of-the-art approaches in terms of short-term forecasting errors and long-term behavior. We further discuss how the proposed framework relates to the Koopman operator theory and Takens' embedding theorem.

Delay embedding coordinates provide a simple tool for reconstructing a system limit cycle given partial observations of the dynamics. However, finding a parametric approximation of the dynamics based on the delay-embedded states is far from being straightforward since we do not have any relationships between the spanned limit cycle and a particular form of parametric models. In this work, we propose an alternative based on neural networks and automatic differentiation. We learn both the dynamics and the latent states jointly as a solution of an optimization problem.

I. INTRODUCTION

Learning the underlying dynamical representation of observed variables $\mathbf{x}_t \in \mathbb{R}^n$ (where $t \in \{t_0, \ldots, T\}$ is the temporal sampling time and n is the dimension of the observations) is a key challenge in various scientific fields, including control theory, geoscience, fluid dynamics, and economics, and for applications ranging from system identification to forecasting and assimilation issues.[1–4]

For fully observed systems, i.e., when the observed variables $\mathbf{x}_t$ relate to some underlying deterministic states $\mathbf{z}_t$, recent advances[5–10] have shown that one can identify the governing equations of the dynamics of $\mathbf{z}_t$ from a representative dataset of observations $\{\mathbf{x}_{t_i}\}_i$. Unfortunately, when the observed variables $\mathbf{x}_t$ only relate to some but not all the components of underlying states $\mathbf{z}_t$, these approaches cannot apply since no ordinary differential equation (ODE) or, more generally, no one-to-one mapping defined in the observation space can represent the time evolution of the observations. In this context, Takens' theorem states the conditions under which a delay embedding, formed by lagged versions of the observed variables, guarantees the existence of governing equations in the embedded space.[11]

Takens' theorem has motivated a rich literature of machine learning schemes to identify dynamical representations of partially observed systems using a delay embedding. This comprises both non-parametric schemes based on nearest-neighbors or analogs[12] and parametric schemes that include polynomial representations,[13] neural network models,[14] and support vector regression (SVR) models.[15] For all these approaches, the identification of the appropriate delay embedding is a critical issue.[16,17]

From a neural network and machine learning perspective, the inference of a latent space, within a state space model (SSM) framework, for dynamical systems has motivated a broad literature especially for time series forecasting.[18–22] Most of those techniques were introduced in the context of reduced order modeling (ROM) to infer low-dimensional manifolds, where the dynamics of the observations can be represented. When considering partially observed systems, these approaches state this issue as the inference of a (non-linear) projection of an input sequence in a latent space where the observations can be modeled. This projection is usually computed in a probabilistic framework using Bayesian filtering techniques. However, recovering the attractor's dynamics using iterative predictions is still an issue for such models since the explicit modeling of latent space as a delay embedding of the observations may limit the expressiveness of the latent states, especially when considering chaotic dynamics.

In this work, we show that we do not need to rely explicitly on a delay embedding. We address the identification of an augmented space of higher dimension than that of the manifold spanned by the observed variables, where the dynamics of the observations can be fully described by an ODE. Using neural-network representations for the parametrization of the dynamical model, it amounts to jointly learning the governing ODE and reconstructing the augmented latent states for a given observation dataset. We report experiments on linear and chaotic dynamics, which illustrate the relevance of the proposed framework compared to state-of-the-art approaches. We then further discuss the key features of this framework with respect to state-of-the-art dynamical systems identification tools such as the Koopman operator theory.[23]

II. BACKGROUND AND RELATED WORK

This section introduces the learning of dynamical representations for partially observed systems and links this problem to recent advances in machine learning.

Let us consider an **unobserved** state variable $\mathbf{z}$ governed by an autonomous system of s differential equations $\dot{\mathbf{z}}_t = f(\mathbf{z}_t)$. Let us also assume that this system generates a flow $\Phi_{t_i}(\mathbf{z}_{t_0}) = \int_{t_0}^{t_i} f(\mathbf{z}_w)dw \in \mathbb{R}^s$ with trajectories that are asymptotic to a limit-cycle L of dimension d contained in $\mathbb{R}^s$. We further assume that we are provided with a measurement function $\mathscr{H}$ that maps our state variables to our observations $\mathbf{x}_t = \mathscr{H}(\mathbf{z}_t) \in \mathbb{R}^n$.

When considering the data-driven identification of a dynamical mapping that governs some observation data, we first need to evaluate whether the dynamics in the observation space can be described using a smooth[24] ODE. Another way to tackle this question is to find the conditions under which the deterministic properties of the unobserved limit-cycle L are preserved in the observation space in $\mathbb{R}^n$ such that one can reliably perform forecasts in the observation space. The general condition under which a mapping $\mathscr{H}$ preserves the topological properties of the initial limit-cycle involves a differential structure. Assuming that L is a smooth compact differential manifold, the topological properties of L are preserved through a mapping $\mathscr{H}$ in $\mathbb{R}^n$ if $\mathscr{H}$ is one-to-one and is an immersion of L in $\mathbb{R}^n$. Under these conditions, our observation mapping is called an **embedding**.[25]

The simplest example of an embedding involves an identity observation operator $\mathscr{H}$. With this embedding, we have direct access to the state variables, which are governed by a deterministic ODE. This particular case has been widely studied in the literature. Parametric representations have been for decades the most popular models thanks to their simplicity and interpretability.[5,13,26–28] Recently, these approaches have been enriched by neural network and deep learning schemes.[29,30] In particular, the link between residual networks[7,31] and numerical integration schemes has opened new research avenues for learning extremely accurate dynamical models even from irregularly sampled training data. These schemes show greater interpretability and forecasting performance for the data-driven representation of systems governed by an ODE, compared with other state-of-the-art neural network schemes, including recurrent neural networks (RNN) such as LSTM (long-short-term memory). Recent advances in model free representations using, for instance, attention mechanisms as in Ref. 32 and reservoir learning as in Ref. 33 have recently shown meaningful improvements in forecasting applications.

However, for a wide range of real-world systems, we are never provided with an observation operator that forms an embedding of the unobserved dynamical system. In such situations, we do not have any guarantee on the existence of a smooth ODE that governs the temporal evolution of our observations. From this point of view, the question of finding an appropriate dynamical representation of some observed data may not be this straightforward. The fact that our data may come from some unobserved governing equation may restrict the use of the above-mentioned state-of-the-art algorithms. The main difficulty lies in the ability to map observation series to a latent space that provides at least a *one-to-one* mapping between two successive states. From a geometrical point of view, the time delay theorem[11] provides a way to build a latent space that preserves the topological properties of the true (unobserved) dynamics limit-cycle. A generalization of this theorem[25] shows that one can reconstruct topologically similar limit-cycles using any appropriate smooth composition map of the observations. Recent works have also investigated the use of deep learning models to find embedding representations of time series. In the work of Ref. 34, a general embedding technique is proposed based on an autoencoder architecture that successfully enfolded the hidden attractor of several state-of-the-art time series. The derivation of a dynamical system from such representations, however, encounters large disparities since no explicit relationships between the defined phase space and an ODE formulation have been clearly made. Classical state-of-the-art techniques such as polynomial representations[5] and K-Nearest Neighbors (KNN)[35] algorithms were proposed, but they often fail to achieve both accurate short-term forecasting performance and long-term topologically similar reconstructed limit-cycles (see experiments for an illustration). The difficulty in finding such representations remains, in our opinion, in the fact that the embedded attractor is defined independently from the data driven model formulation and learning.

We may also point out that the limitation of ODE-based representation in deep learning architecture has also been pointed out recently in Refs. 36 and 37 for classification issues. As ODE-derived trajectories do not intersect, it may limit the ability of neural ODE representations to reach relevant classification performance

in a given feature space. To address this issue, Dupont *et al.*[36] and Zhang *et al.*[37] propose to consider an augmented state, simply by augmenting the observed state by a number of zeros to create a high-dimensional space in which an ODE representation can be identified. Such a strategy cannot apply to time series modeling as successive augmented states cannot be forced to zero for some dimensions.

Advances in the inference of latent spaces in state space models were introduced essentially, from a dynamical systems perspective, to retrieve low-dimensional manifolds, where the dynamics of the system evolve. When applied to partially observed systems, the latent variables are typically inferred from a sequence of observations through a parametric modeling of the posterior distribution as in Refs. 21, 22, and 7 or through marginalization with model constraints as in Refs. 18 and 19. However, such models often fail in accounting for long term patterns (as shown in the experiments). This is due to the fact that the latent space is constrained to be a non-linear projection of a sequence of observations, which limits the expressiveness of the dynamical model. Interestingly, Mirowski and LeCun[20] do not involve the learning of an inference model as the reconstruction of the latent states is solved as gradient-based minimization of the dynamical prior with respect to an observation series. However, the dynamical prior relies on an explicit delay representation (not necessarily an embedding) as the dynamics of the latent state depend both on the previous latent state and on a delay embedding of the observations.

In this work, we address the identification of a latent embedding, associated with an ODE representation, for partially observed systems. The core idea of this work is to infer an augmented latent space, governed by an ODE, which fully explain the observed time series and their dynamics. In contrast to previous work,[7,18,19,21,22] we do not exploit either a delay embedding or an explicit modeling of the inference model (i.e., the reconstruction of the latent states given the observed time series). As such, our scheme only involves the selection of the class of ODEs of interest. The expected benefits are as follows: (1) our model ensures the existence of a latent embedding associated with an ODE, which may not be guaranteed when considering a parametric inference model and/or a delay embedding, (2) our model reduces the complexity of the overall scheme to the complexity of the ODE representation, and (3) our model guarantees the consistency of the reconstructed latent states with respect to the learned ODE.

III. LEARNING LATENT REPRESENTATIONS OF PARTIALLY OBSERVED DYNAMICS

A. Augmented latent dynamics

Let us assume a continuous s-dimensional dynamical system $\mathbf{z}_t$ governed by an autonomous ODE $\dot{\mathbf{z}}_t = f(\mathbf{z}_t)$, with Φ_t being the corresponding flow $\Phi_t(\mathbf{z}_{t_0}) = \int_{t_0}^{t} f(\mathbf{z}_w)dw$ with trajectories that are asymptotic to a limit-cycle L of dimension d contained in $\mathbb{R}^s$.

In many applications, one cannot fully access the state z and the observations only relate to some components of this state. Formally, we can define an observation function $\mathscr{H}: \mathbb{R}^s \longrightarrow \mathbb{R}^n$ such that the observations $\mathbf{x}_t$ follow $\mathbf{x}_t = \mathscr{H}(\mathbf{z}_t)$. We can also define a bijective map $\mathscr{M}$ that maps our observations $\mathbf{x}_t$ to some low dimensional manifold $\mathbf{r}_t = \mathscr{M}(\mathbf{x}_t) \in \mathbb{R}^k$. The definition of this operator is crucial in the data driven identification of ROMs[38] of real data since in this case, the provided data are usually mapped through $\mathscr{H}$ in a higher dimensional space. Besides, $\mathscr{M}$ is supposed to be bijective so that the dynamics in $\mathbb{R}^n$ are fully determined by the dynamics in $\mathbb{R}^k$. From now on, and for the sake of simplicity, we will refer to both $\mathbf{r}_t \in \mathbb{R}^k$ and $\mathbf{x}_t \in \mathbb{R}^n$ as observations since they are equivalent up to a bijective map $\mathscr{M}$.

We aim to derive an ODE representation of $\mathbf{x}_t \in \mathbb{R}^n$. However, the key question arising here is the extent to which the dynamics expressed in the observations space reflect the true underlying dynamics in $\mathbb{R}^s$, and consequently, the conditions on $\mathscr{H}$ under which the predictable deterministic dynamical behavior of the hidden states is still predictable in the observations space. To illustrate this issue, we may consider a linear dynamical system in the complex domain governed by the following linear ODE:

$$\begin{cases} \dot{\mathbf{z}}_t = \alpha \mathbf{z}_t, \\ \mathbf{z}_{t_0} = \mathbf{z}_0, \end{cases} \tag{1}$$

with $\mathbf{z} \in \mathbb{C}$ being a state variable and $\alpha \in \mathbb{C}$ being a complex imaginary number. The solution of this problem is

$$\mathbf{z}_t = \mathbf{z}_0 e^{\alpha t}. \tag{2}$$

Let us assume now that we are only provided with the real part as direct measurements of the unobserved state, i.e., $\mathscr{H}(.) = Real(.)$: $\mathbf{x}_t = Real(\mathbf{z}_t)$ so in this case $\mathscr{M} = I_1$ and $k = n$.

Proposition 1: *The flow of an ODE cannot represent the time evolution of* $\boldsymbol{x}_t$.

The proof of the proposition is given in the Appendix A and the intuition behind it is as follows. Assuming that we are only provided with the real part as direct measurements $\mathbf{x}_t \in \mathbb{R}$ of the true states $\mathbf{z}_t$, no smooth autonomous ODE model in the scalar observation space can describe the trajectories of the observations as the mapping between two observations is not one-to-one. For example, assuming that $\mathbf{z}_{t_0}$ and $\mathbf{z}_{t_1}$ correspond to two states that have the same real part but distinct imaginary parts, the associated observed states are equal $\mathbf{x}_{t_0} = \mathbf{x}_{t_1}$. However, the time evolution of the states $\mathbf{z}_{t_0}$ and $\mathbf{z}_{t_1}$ differ if they have different imaginary parts such that the observed states $\mathbf{x}_{t_0+\delta}$ and $\mathbf{x}_{t_1+\delta}$ after any time increment δ are no longer equal. As a consequence, a given observation may have more than one future state and this behavior cannot be represented by a smooth ODE in the observation space. And the application of an ODE mapping[6,7] for such observations will lead to poor forecasting performance. From a Naïve neural networks point of view, fitting such a model will most likely force the forecasting into an equilibrium point since we are iteratively matching the same inputs with different output predictions. For a given observation operator $\mathscr{H}$ of a deterministic underlying dynamical system that governs $\mathbf{z}_t$, Takens' theorem guarantees the existence of an augmented space, defined as a delay embedding of the observations, in which a one-to-one mapping exists between successive time steps of the observation series.[11] Rather than exploring such delay embedding, we aim to identify an augmented latent space, where the latent dynamics are governed by a smooth ODE and can be mapped to the observations. Let us define

$\mathbf{u}_t \in \mathbb{R}^{d_E}$ a d_E-dimensional augmented latent state as follows:

$$\mathbf{u}_t{}^T = [\mathscr{M}(\mathbf{x}_t)^T, \mathbf{y}_t^T], \tag{3}$$

with $\mathbf{y}_t \in \mathbb{R}^l$ being the unobserved component of latent state $\mathbf{u}_t$. The augmented latent space evolves in time according to the following state space model:

$$\begin{cases} \dot{\mathbf{u}}_t = f_\theta(\mathbf{u}_t), \\ \mathbf{x}_t = \mathscr{M}^{-1}(G(\mathbf{u}_t)), \end{cases} \tag{4}$$

where the dynamical operator f_θ belongs to a family of smooth operators (in order to guarantee uniqueness[39]) parametrized by θ. We typically consider a neural-network representation with Lipschitz nonlinearities and finite weights. G is a projection matrix that satisfies $\mathscr{M}(\mathbf{x}_t) = G(\mathbf{u}_t)$. As detailed in Secs. III B–III D, we address the identification of the operator f_θ and of the associated latent space $\mathbf{u}$ from a dataset of observations $\{\mathbf{x}_0, \ldots, \mathbf{x}_T\}$ as well as the exploitation of the identified latent dynamics for the forecasting of the time evolution of the observed states, for instance, unobserved future states $\{\mathbf{x}_{T+1}, \ldots, \mathbf{x}_{T+N}\}$.

B. Learning scheme

Given an observation time series $\{\mathbf{x}_0, \ldots, \mathbf{x}_T\}$ and the bijective map $\mathscr{M}$, we aim to identify the state-space model defined by (4), which amounts to learning the parameters θ of the dynamical operator f_θ. However, as the component $\mathbf{y}_t$ of the latent state $\mathbf{u}_t$ is never observed, this identification requires the joint optimization of the model parameters θ as well as of the hidden component $\mathbf{y}_t$. Formally, this problem is stated as the following minimization of the forecasting error on the observed variables:

$$\hat{\theta} = \arg\min_\theta \min_{\{\mathbf{y}_t\}_t} \sum_{t=1}^{T} \|\mathbf{x}_t - \mathscr{M}^{-1}(G\left(\Phi_{\theta,t}\left(\mathbf{u}_{t-1}\right)\right))\|^2, \tag{5}$$

$$\text{subject to} \begin{cases} \mathbf{u}_t & = \Phi_{\theta,t}(\mathbf{u}_{t-1}), \\ \mathscr{M}(G(\mathbf{u}_t)) & = \mathbf{x}_t, \end{cases}$$

with $\Phi_{\theta,t}$ being the one-step-ahead diffeomorphic mapping associated with operator f_θ such that

$$\Phi_{\theta,t}(\mathbf{u}_{t-1}) = \mathbf{u}_{t-1} + \int_{t-1}^{t} f_\theta(\mathbf{u}_w) dw.$$

In (5), the loss to be minimized involves the one-step-ahead forecasting error for the observed variable $\mathbf{x}_t$. The constraints state that the augmented state $\mathbf{u}_t$ is composed of observed component and $G(\mathbf{u}_t)$ should be a solution of the ODE (4). Here, we numerically minimize the equivalent formulation

$$\min_\theta \min_{\{\mathbf{y}_t\}_t} \sum_{t=1}^{T} \|\mathbf{x}_t - \mathscr{M}^{-1}(G\left(\Phi_{\theta,t}\left(\mathbf{u}_{t-1}\right)\right))\|^2 + \lambda \|\mathbf{u}_t - \Phi_{\theta,t}(\mathbf{u}_{t-1})\|^2 \tag{6}$$

where $\mathbf{u}_t^T = [\mathscr{M}(\mathbf{x}_t^T), \mathbf{y}_t^T]$ and λ is a weighting parameter. The term $\|\mathbf{u}_t - \Phi_{\theta,t}(\mathbf{u}_{t-1})\|^2$ may be regarded as a regularization term such that the inference of the unobserved component $\mathbf{y}_{t-1}$ of the augmented state $\mathbf{u}_{t-1}$ is not solved independently for each time step.

Using a neural-network parametrization for the ODE operator f_θ, the corresponding forecasting operator $\Phi_{\theta,t}$ is also stated as a neural network based on a numerical integration scheme formulation (typically a 4th-order Runge–Kutta scheme). This architecture, quite similar to a ResNet,[21] allows very accurate identification of ODE models.[6,31] Hence, for a given observed state series $\{\mathbf{x}_0, \ldots, \mathbf{x}_T\}$, we minimize (6) jointly with respect to θ and unobserved variables $\{\mathbf{y}_0, \ldots, \mathbf{y}_T\}$. In the experiments reported in Sec. IV, we consider bilinear architectures.[6] However, the proposed framework applies to any neural-network architecture.

C. Links to manifold embedding theorems

Whitney's embedding theorem guarantees that a generic map $\mathscr{H}: \mathbb{R}^s \longrightarrow \mathbb{R}^{d_E}$ is an embedding of the manifold in $\mathbb{R}^{d_E}$ as long as $d_E > 2d + 1$. However, from an experimentalist perspective, being able to observe a large number of independent quantities (typically $2d + 1$) is usually impossible. The Takens delay embedding theorem overcomes this issue by using time delay coordinates of a single generic variable (under some technical assumptions) as an embedding of the manifold in $\mathbb{R}^{d_E}$. However, and as stated above, modeling the delay embedding attractor is not straightforward. In the proposed framework, the embedded attractor is learned jointly with the data driven dynamical model, which makes us to find the most appropriate embedding for a given architecture of the data driven model. Furthermore, and supposing that the model architecture is representative enough (typically nonlinear), the learned latent space can be considered a generic observation basis of the underlying dynamics which, corresponding to Whitney's theorem, and similarly to Takens' embedding theorem, forms an embedding of the unseen attractor.

D. Application to forecasting

We also apply the proposed framework to the forecasting of the observed states $\mathbf{x}_t$. Given a trained latent dynamical model (4), forecasting future states for $\mathbf{x}_t$ relies on the forecasting of the entire augmented latent state $\mathbf{u}_t$. The latter amounts to determining an initial condition of the unobserved component $\mathbf{y}_t$ and performing a numerical integration of the trained ODE (4).

Let us denote by $\mathbf{x}_t^n$, $t \in \{t_0, \ldots, T\}$ a new series of observed states. We aim to forecast future states $\mathbf{x}_t^n$, $t \in \{T+1, \ldots, T+\delta T\}$. Following (6), we infer the unobserved component $\widehat{\mathbf{y}}_T$ of latent state X_T^n at time T from the following minimization:

$$\hat{\mathbf{y}}_T^n = \arg\min_{\mathbf{y}_T^n} \min_{\{\mathbf{y}_t^n\}_{t<T}} \sum_{t=T+1}^{T+\delta T} \|\mathbf{x}_t^n - \mathscr{M}^{-1}(G\left(\Phi_{\theta,t}\left(\mathbf{u}_{t-1}^n\right)\right))\|^2 + \lambda \|\mathbf{u}_t^n - \Phi_{\theta,t}(\mathbf{u}_{t-1}^n)\|^2. \tag{7}$$

Here, we minimize only with respect to latent variables $\{\mathbf{y}_t^n\}$ given the trained forecasting operator $\Phi_{\theta,t}$. This minimization relates to a variational assimilation issue with partially observed states and known dynamical and observation operators.[40] Similarly to the learning step, we benefit from the neural-network parameterization of operator $\Phi_{\theta,t}$ and from the associated automatic differentiation tool to compute the solution of the above minimization using a gradient descent.

We may consider different initialization strategies for this minimization problem. Besides a simple random initialization, we may benefit from the information gained on the manifold spanned by the unobserved components during the training stage. The basic idea comes to assume that the training dataset is likely to comprise state trajectories which are similar to the new one. As the training step embeds the inference of the whole latent state sequence, we may pick as initialization for minimization (7) the inferred augmented latent state in the training dataset which leads to the observed state trajectory that is the most similar (in the sense of the L2 norm) to the new observed sequence $\mathbf{x}_t^n$. The interest of this initialization scheme is twofold: (1) speeding-up the convergence of minimization (7) as we expect to be closer to the minimum and (2) considering an initial condition which is in the basin of attraction of the reconstructed limit-cycle. The latter may be critical as we cannot guarantee that the learned model does not involve other limit-cycles than the ones truly revealed by the training dataset, which may lead to a convergence to a local and poorly relevant minimum. Reaching the global minimum of the optimization problem of Eq. (6) (which is the actual governing equations and attractor of the system) would cancel this issue. However, reaching the global minimum only knowing partial observations of the system is almost deterministically impossible since it depends on the parametrization of the approximate dynamical model and the initialization of the latent states. In this context, we may also argue that given partial observations of the system, several models can reflect the variability of the observed variables while being diffeomorphic to the actual governing dynamics in the attractor spanned by the observations (not necessarily away from the attractor as the approximate model may involve several limit-cycles other than the one spanned by the observations). Given these considerations, we can retrieve most of the time a relevant local minimum, which reflects the topological properties of the initial model and attractor.

IV. NUMERICAL EXPERIMENTS

In this section, we report numerical experiments to illustrate the key features of proposed framework. We consider three case-studies: a linear ODE case-study; a chaotic system, namely, Lorenz-63 dynamics, and real upper ocean data.

A. Application to a linear ODE

In order to illustrate the key principles of the proposed framework, we consider the following linear ODE in the complex domain:

$$\begin{cases} \dot{\mathbf{z}}_t = \alpha \mathbf{z}_t, \\ \mathbf{z}_{t_0} = \mathbf{z}_0, \end{cases} \tag{8}$$

with $\alpha = -0.1 - 0.5j$, $j^2 = -1$, and $\mathbf{z}_0 = 0.5$. As $\alpha \in \mathbb{C}$ with $Real(\alpha) < 0$ and $\mathbf{z}_0 \neq 0$, the solution of this ODE is an ellipse in the complex plane (Fig. 1).

As observation, we consider the real part of the underlying state, i.e., the observation function $\mathscr{H}: \mathbb{C} \longrightarrow \mathbb{R}$ is given by $\mathbf{x}_t = Real(\mathbf{z}_t)$. This is a typical example, where the mapping between two successive observations is not a one-to-one mapping since all the states that have the same real part lead to the same observation. As explained in Sec. III, one cannot identify an autonomous ODE model that will reproduce the dynamical behavior of the observations in the observations space.

We apply the proposed framework to this toy example. We consider a two-dimensional augmented state $\mathbf{u}_t = [\mathbf{x}_t, \mathbf{y}_t^1]$ with $\mathscr{M} = I_1$. As neural-network parametrization for operator f_θ, we consider a neural network with a single linear fully connected layer. We use an observation series of 10 000 time steps as training data. As illustrated in Figs. 1 and 2, given the same initial condition over the observable state, the inferred latent state dynamics, though different from the true ones, depict a similar spiral pattern. This result is in agreement with geometrical reconstruction techniques[11] of the latent dynamics up to a diffeomorphic mapping. Overall, our model learns a dynamical behavior similar to the true model represented by an elliptic transient and an equilibrium point limit-set. Furthermore, the projection of the augmented latent space and the true solution of Eq. (8) in the real axis illustrate the relevance of the proposed framework in forecasting the observations dynamics (mean square error $<1E-6$).

B. Lorenz-63 dynamics

Lorenz-63 dynamical system is a three-dimensional model that involves, under some specific parametrizations,[41] chaotic dynamics with a strange attractor. We simulate chaotic Lorenz-63 state

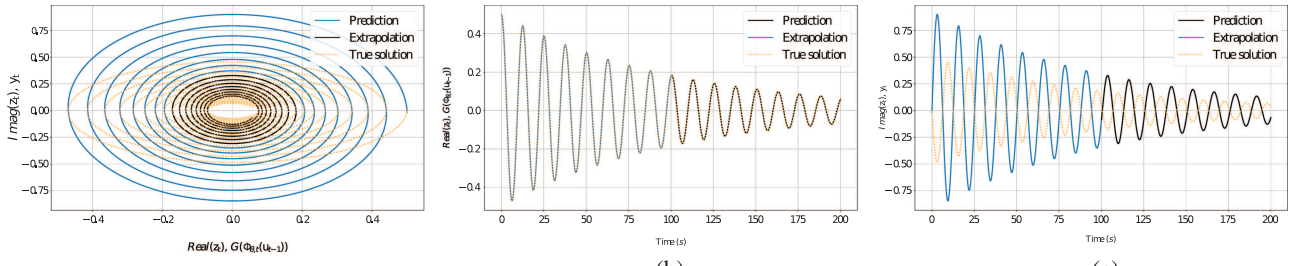

FIG. 1. Illustration for a two-dimensional linear ODE. Forecasted augmented latent space (a) with respect to the true states given the same initial condition as the training sequence. We illustrate both the prediction (forecast up to the end of the training time) of the trained model and the extrapolation (forecast beyond the training time) performances with respect to the true trajectory. (b) [respectively, (c)] illustrates the forecasting of the observations (respectively, the inference of the imaginary part).

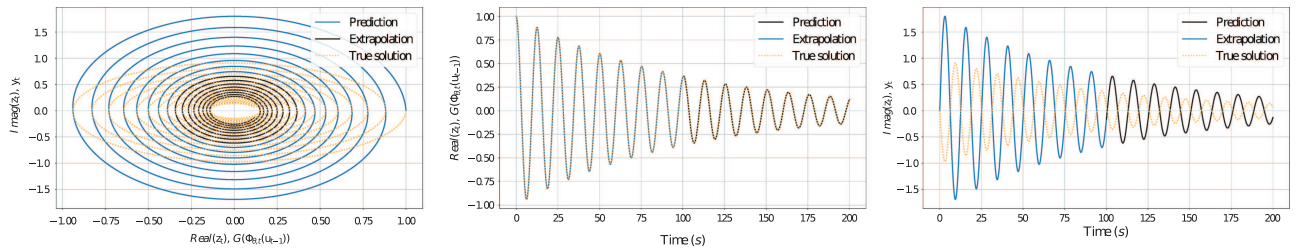

FIG. 2. Illustration for a two-dimensional linear ODE. Forecasted augmented latent space (a) with respect to the true states given a new initial condition. Similarly to Fig. 1 given the initial condition we illustrate both the prediction and the extrapolation performances with respect to the true trajectory. (b) [respectively, (c)] illustrates the forecasting of the observations (respectively, the inference of the imaginary part).

sequences with the same model parameters as proposed in Ref. 41 using the LOSDA ODE solver[42] with a sampling time step of 0.01. We assume that only the first Lorenz-63 variable is observed $x_t = z_{t,1}$ and we set $\mathscr{M} = I_1$. We apply the proposed framework to this experimental setting using a training sequence of 4000 time steps.

For benchmarking purposes, we perform a quantitative comparison with state-of-the-art approaches using delay embedding representations.[11] The parameters of the delay embedding representation, namely, the lag τ and the dimension d_E of the augmented space were computed using state-of-the-art techniques. Specifically, the lag parameter was computed using both the mutual information and correlation techniques,[16] respectively, denoted as τ_{MI} and τ_{Corr}. Regarding the dimension of the embedding representation, we used the Whitney's embedding condition $d_E = 2d + 1$ with d being the dimension of the hidden limit-cycle. The delay embedding dimension was also computed using the false nearest neighbors (FNN) method.[17] We also tested arbitrary parameters for the delay embedding dimension. Given the delay embedding representation, we tested two state-of-the-art data-driven representations of the dynamics: the analog forecasting (AF) technique which is based on the nearest neighbors algorithm[35] and the sparse regression (SR) method on a second order polynomial representation of the delay embedding states.

Regarding deep learning models, we compare our method to a stacked bidirectional LSTM (RNN) and to the latent-ODE model.[7] The proposed framework, referred to as Neural embedding for Dynamical Systems (NbedDyn), was tested for different dimensions of the augmented state space, namely, from 3 to 6 (please refer to the Appendix for details on the considered neural network architectures).

Figure 3 illustrates the learning process for the latent space from the initialization to the last training epoch. The optimization of the training criterion with respect to both the model parameters and the latent states leads to a topologically similar spanned manifold with respect to the truly unobserved high dimensional one. We also illustrate the convergence of the training procedure in terms of short term forecast and topological invariants of the learned embedding and model as shown in Fig. 4. Our method is able to get similar results as classical attractor dimension unfolding algorithms such as FNN using both short and long term criterions since we show that three dimensions of the latent state are enough to get a converged architecture. Regarding the quantitative analysis, we report both the analysis of short-term forecasting performance and the long-term asymptotic behavior characterized by the largest Lyapunov exponent of the benchmarked models in Table I. The proposed model leads to significant improvements in terms of short term forecasting performance with respect to the other approaches. Surprisingly, the latent-ODE and RNN models lead to the poorest performance in terms of both forecasting error and asymptotic behavior. This is mainly due, in the latent-ODE case, to the fact that the latent space is seen as a non-linear projection of the observed variables through the optimization of the Evidence Lower Bound (ELBO) loss.[22] By contrast, our latent embedding formulation optimizes the latent states to forecast the observed variables which explicitly constrain the latent space to be an embedding of the true underlying dynamics. The RNN model in the other hand converges to a periodic solution (please refer to the Appendix for forecasting figures) with still a poor short term forecasting performance. Overall, these results suggest that one should use such deep learning models with care to reach satisfying performance. The SR model seems to lead to better short term forecast using *ad hoc* parameters ($\tau = 6$, $d_E = 3$); however, it does not capture well the chaotic patterns, which are associated with a positive largest Lyapunov exponent. This may suggest that the combination of the SR model and a delay embedding may require additional investigation as a good geometrical reconstruction of the phase space as stated in Takens' theorem does not guarantee the existence of a parametric ODE model based on the corresponding delay embedding variables. Better performance is reported using an analog forecasting approach. The performance, however, greatly varies depending on the considered definition of the delay embedding. Using *ad hoc* parameters ($\tau = 6$, $d_E = 3$), one may retrieve the expected long-term chaotic behavior ($\lambda_1 = 0.87$) with a relatively low short-term forecasting error ($8.0E-4$ for a one-step-ahead forecast). When considering the proposed model, we report for all the parametrizations of the dimension of the augmented space from 3 to 6, performance at least in the same range as the best analog forecasting setting. Besides, when increasing the dimension of the augmented space, we significantly decrease the short-term forecasting errors ($<1.E-4$ for a one-step-ahead prediction when considering the best fit for $d_E = 6$, i.e., one order of magnitude compared to the best benchmark model) while keeping

FIG. 3. Evolution of the learned latent space. Starting from a random initialization of the augmented states $\mathbf{y}_i$, the latent space is optimized according to the minimization of Eq. (6) to form a limit-cycle similar to the true Lorenz 63 attractor. We depict three-dimensional projections of the learned latent space for the proposed model with different embedding dimensions from $d_E = 3$ to $d_E = 6$.

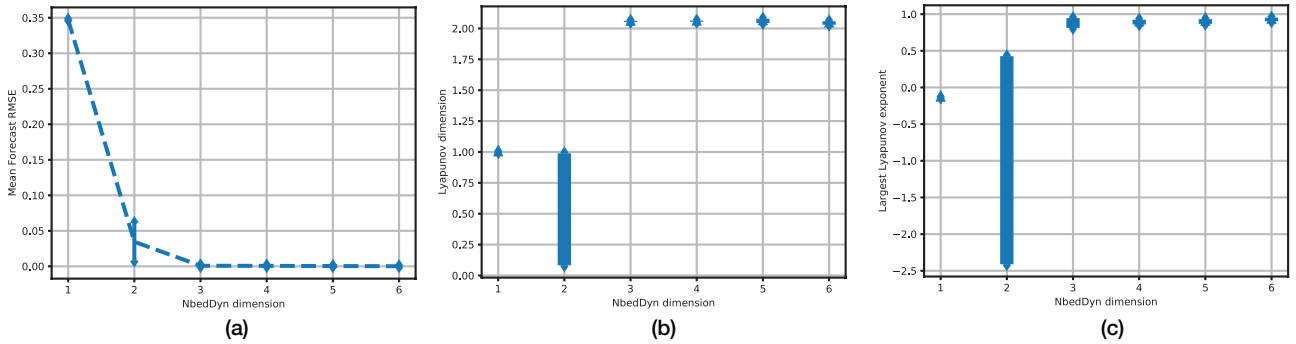

FIG. 4. Convergence of the proposed NbedDyn architecture as a function of the dimension of the augmented space. Evolution of the short term forecast performances (a) as well as the largest Lyapunov exponent (c) and Lyapunov dimension (b) of the NbedDyn attractor as a function of the dimension of the embedding. Our architecture is able to unfold the underlying Lorenz dynamics given a sufficient dimension of the augmented state. We may note that the topological properties illustrated here were estimated using iterative forecast of the trained model, only given an initial condition inside the basin of attraction of the spanned manifold. (The true Lyapunov dimension of the Lorenz 63 model is 2.06.[43])

an appropriate chaotic long-term pattern ($\lambda_1 = 0.92$). Finally, since all the learned attractors (as long as $d_E > 2$) are diffeomorphic to the actual Lorenz 63 model, we show in Fig. 5 that we can map theme to the actual Lorenz 63 attractor only using an affine transformation (statistically, since some runs fail to be mapped to the true Lorenz using an affine transformation. However, the proportion of these runs is around 20% and they can be mapped to the true Lorenz using a linear quadratic model instead). This result can be interpreted as follows. Given a single generic observation, we need only three variables to model the Lorenz attractor (this result is shown in the learning convergence figure above and can be easily verified using state of the art techniques such as the FNN), one may expect a strong relationship between the latent variables of different NbedDyn architectures with $d_E > 2$. We show that this relationship is linear (up to modeling errors) and more importantly is also linear with respect to the true unseen underlying dynamics.

C. Modeling sea level anomaly (SLA)

The data driven identification of dynamical representations of real data is an extremely difficult task especially when the underlying processes involve non-stable behaviors such as chaotic attractors.

TABLE I. Forecasting performance on the test set of data-driven models for Lorenz-63 dynamics where only the first variable is observed. First two columns: mean RMSE for different forecasting time steps, third column: the largest Lyapunov exponent of a predicted series of length of 10 000 time steps. (The true largest Lyapunov exponent of the Lorenz 63 model is 0.91.[43])

Model			$t_0 + h$	$t_0 + 4h$	λ_1
AF	$\tau_{MI} = 16$	d_E(FNN) = 3	$5.6E-3$	$1.3E-2$	0.85
	$\tau_{MI} = 16$	d_E(Takens) = 6	$9.9E-3$	$2.4E-2$	*NaN*
	$\tau_{Corr} = 27$	d_E(FNN) = 3	$8.9E-3$	$2.3E-2$	12.35
	$\tau_{Corr} = 27$	d_E(Takens) = 6	$8.5E-3$	$1.9E-2$	*NaN*
	$\tau = 6$	$d_E = 3$	$8.0E-4$	$9.0E-4$	0.87
	$\tau = 10$	$d_E = 3$	$2.1E-3$	$4.9E-3$	0.60
SR	$\tau_{MI} = 16$	d_E(FNN) = 3	$7.8E-2$	$2.5E-1$	0.12
	$\tau_{MI} = 16$	d_E(Takens) = 6	$4.5E-2$	$1.7E-1$	*NaN*
	$\tau_{Corr} = 27$	d_E(FNN) = 3	$1.4E-1$	$4.6E-1$	*NaN*
	$\tau_{Corr} = 27$	d_E(Takens) = 6	$2.1E-1$	$8.4E-1$	*NaN*
	$\tau = 6$	$d_E = 3$	$7.6E-3$	$7.4E-3$	*NaN*
	$\tau = 10$	$d_E = 3$	$2.5E-2$	$5.7E-2$	0.2535
	Latent-ODE RNN		$6.9E-2 \pm 2.9E-2$	$1.5E-1 \pm 3E-2$	*NaN*
			$6.9E-2 \pm 4.6E-2$	$1.5E-1 \pm 1.1E-1$	-6.79 ± 0.0
NbedDyn		$d_E = 3$	$3.2E-4 \pm 1.3E-4$	$1.7E-3 \pm 7.5E-4$	0.81 ± 0.09
		$d_E = 4$	$1.3E-4 \pm 5.2E-5$	$7.3E-4 \pm 2.2E-4$	0.82 ± 0.06
		$d_E = 5$	$3.8E-4 \pm 7.4E-4$	$2.0E-3 \pm 3.4E-4$	0.80 ± 0.02
		$d_E = 6$	$3.7E-4 \pm 2.8E-4$	$2.0E-3 \pm 1.7E-3$	0.92 ± 0.02
		$d_E = 6$ (Best)	$9.1E-5$	$4.7E-4$	0.92

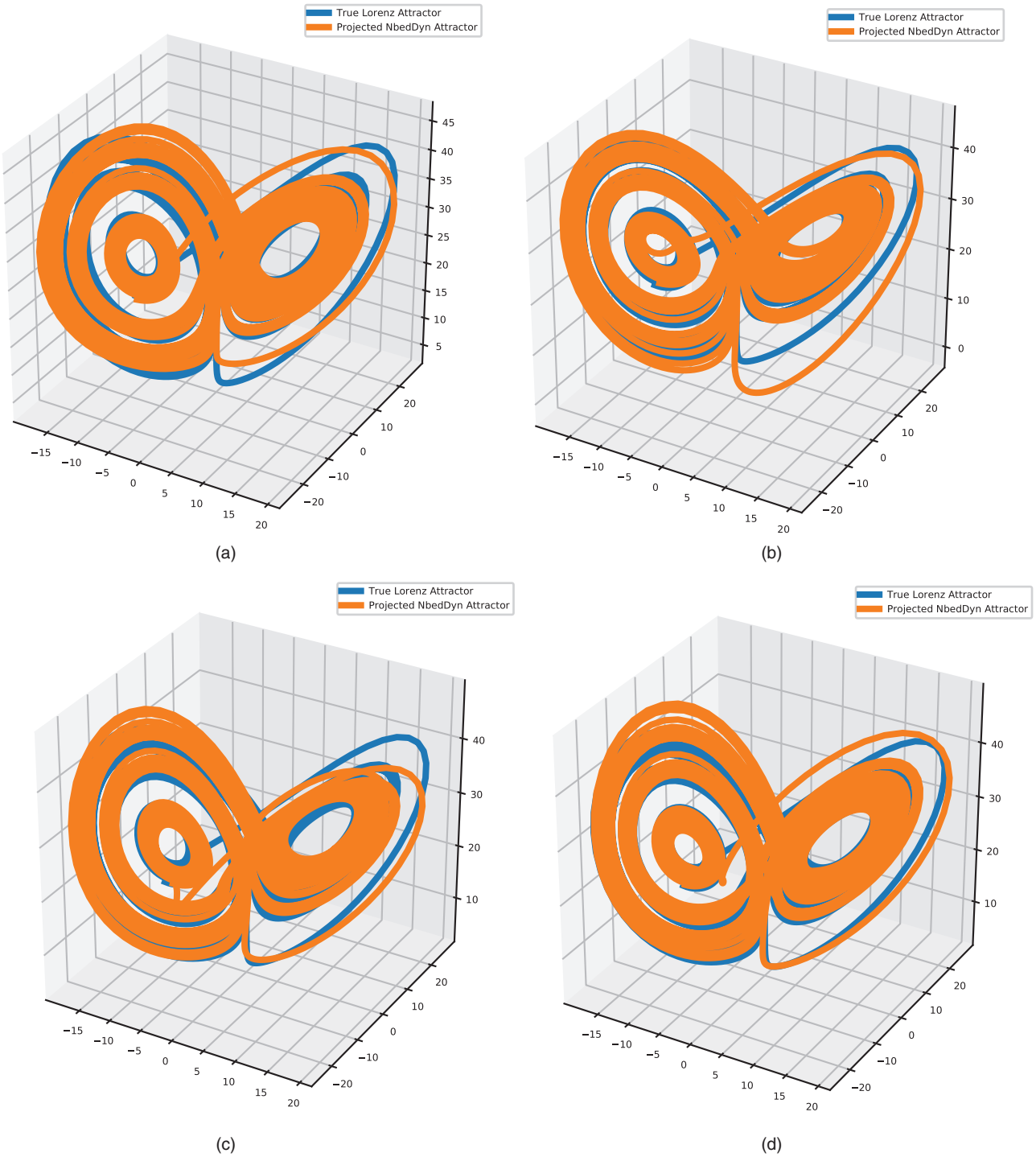

FIG. 5. Mapping the NbedDyn attractors to the true Lorenz attractor. An affine transformation is trained to map the NbedDyn attractor to the true Lorenz attractor. (a)–(d) highlight these projections given an NbedDyn attractor of dimensions 3, 4, 5, and 6, respectively. We show that the relationship between the embeddings unfolded by our architecture for different dimensions of the augmented space is linear (up to modeling errors) and more importantly is also linear to the true unseen dynamics.

This is mainly due to the fact that we do not have any exact knowledge of the closed form of the equations governing the temporal evolution of our variables. Furthermore, the measured quantity may depend on other unobserved variables which makes the exploitation of data-driven techniques highly challenging.

In this context, we report an application to SLA (sea level anomaly) dynamics, which relate to upper ocean dynamics and are monitored by satellite altimeters.[44] Sea surface dynamics are chaotic and clearly involve latent processes, typically subsurface and atmospheric processes. The dataset used in our experiments is a SLA time series obtained using the WMOP product.[45] The spatial resolution of our data is 0.05°, and the temporal resolution $h = 1$ day. We use the data from January 2009 to December 2014 as training data, and we tested our approach on the last month of the year 2014. The considered region is located on south Mallorca (2.5°E–4.25°E, 37.25°N–39.5°N). Finally, and in order to identify a ROM, we mapped our data through a projection defined offline using a PCA as follows: $\mathbf{r}_t = \mathscr{M}(\mathbf{x}_t) \in \mathbb{R}^k$ with $k = 15$ which amounts to capture 92% of the total variance (here $\mathscr{M}$ is simply a linear PCA projection).

We report forecasting performance for our model and include a comparison with analog methods (AF), sparse regression (SR), LSTM (RNN), and a neural ODE setting (latent-ODE) in Table II. (The results of the neural network based models were averaged over five runs.) Regarding the proposed NbedDyn model, we consider an augmented latent space with $d_E = 60$. Our model clearly outperforms the three benchmarked schemes with a very significant gain for the forecasting performance at 1 day (relative gain greater than 90%) and 2 days (relative gain greater than 90%). For a 4-day-ahead forecasting, our model still outperforms the other ones though the gain is lower (relative gain of 40%). In order to illustrate the influence of adding extra dimensions to define an augmented latent space on real data, we show in Fig. 6 the convergence of the solution in terms of forecasting performances as a function of the dimension of the embedding. We also tested the proposed NbedDyn model directly on the PCA space ($d_E = k = 15$); this model is referred to as NbedDynZERO and the influence of the latent components is clear from the results given in Table II. We report a relative gain up to 90% with respect to the same model directly applied onto the PCA space. We let the reader refer to Appendixes A–F for a more detailed analysis of these experiments, including visual comparisons of the forecasts.

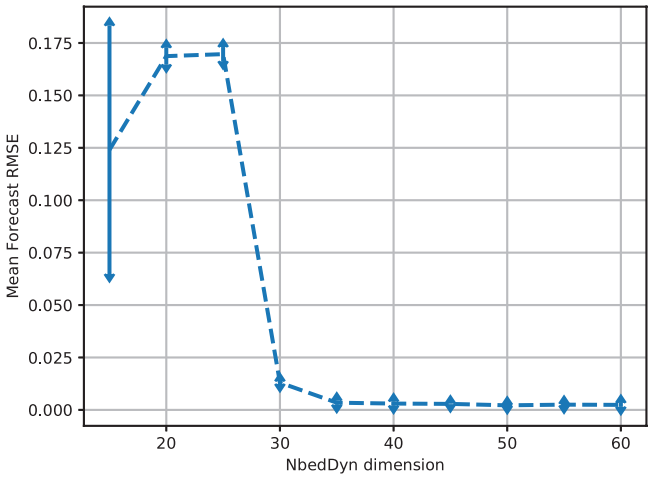

FIG. 6. Convergence of the proposed NbedDyn architecture in the SLA application as a function of the dimension of the augmented space.

V. DISCUSSION

In this work, we address the data-driven identification of latent dynamics for systems which are only partially observed, i.e., when some components of the system of interest are never observed. The reported forecasting performance for Lorenz-63 dynamics is in line with the forecasting performance of state-of-the-art learning-based approaches for a noise-free and fully observed setting. This is of key interest for real-world applications, where observing systems most often monitor only some components of the underlying systems. As a typical example, the SLA forecasting experiment clearly motivates the proposed framework in the context of ocean dynamics for which neither *in situ* nor satellite observing systems can provide direct observations for all state variables (e.g., subsurface velocities and fine-scale sea surface currents[46]).

TABLE II. SLA forecasting performance on the test set of data-driven models. RMSE and correlation coefficients for different forecasting time steps.

Model		$t_0 + h$	$t_0 + 2h$	$t_0 + 4h$
AF	RMSE	0.036	0.049	0.067
	Corr	98.93%	96.97%	93.99%
SR	RMSE	0.014	0.021	0.037
	Corr	99.42%	97.63%	90.91%
Latent-ODE	RMSE	0.030 ± 0.05	0.031 ± 0.031	0.040 ± 0.040
	Corr	$98.20\% \pm 0.39\%$	$97.39\% \pm 0.36\%$	$93.42\% \pm 0.55\%$
RNN	RMSE	0.026 ± 0.003	0.038 ± 0.007	0.053 ± 0.016
	Corr	$98.36\% \pm 0.40\%$	$95.29\% \pm 1.73\%$	$74.97\% \pm 5.75\%$
NbedDynZERO	RMSE	0.016 ± 0.0	0.023 ± 0.0	0.038 ± 0.0
	Corr	$99.44\% \pm 0.0\%$	$97.71\% \pm 0.0\%$	$91.18\% \pm 0.0\%$
NbedDyn	RMSE	0.002 ± 0.0003	0.006 ± 0.001	0.020 ± 0.004
	Corr	$99.99\% \pm 0.0017\%$ %	$99.91\% \pm 0.01\%$	$99.01\% \pm 0.04\%$

We may also further discuss how the proposed framework relates to state-of-the-art dynamical system theory approaches. Most of these approaches rely on delay embedding, as Takens' theorem states the existence of a delay embedding in which the topological properties of the hidden dynamical system are equivalent to those of the true systems up to a diffeomorphic mapping. Hence, state-of-the-art approaches typically combine the selection of a delay embedding representation within classic regression models to represent the one-step-ahead mapping in the considered embedding. Here, we consider latent dynamics governed by an unknown ODE (4) but we do not explicitly state the latent space. This is, however, implicit in our forecasting framework. By construction, the considered forecasting model relies on the integration of the learned ODE (4) from an initial condition given as the solution of minimization (7). Let us consider the following embedding ψ:

$$\psi\left(\{\mathbf{x}_t\}_{t_0:T}\right) = \arg\min_{\mathbf{u}_T} \min_{\{\mathbf{u}_t\}_{t<T}} \sum_{t=1}^{T} \|\mathbf{x}_t - \mathscr{M}^{-1}(G\left(\Phi_{\theta,t}\left(\mathbf{u}_{t-1}\right)\right)\|^2 + \lambda\|\mathbf{u}_t - \Phi_{\theta,t}(\mathbf{u}_{t-1})\|^2. \tag{9}$$

Given this embedding, the resulting one-step-ahead forecasting for the observed variable may be written as

$$\mathbf{x}_{T+1} = \mathscr{M}(G\left(\Phi_{\theta,t}\left(\psi\left(\{\mathbf{x}_t\}_{t=t_0:T}\right)\right)\right)). \tag{10}$$

Hence, ψ defines a delay embedding representation implicitly stated through minimization (7). In this embedding, the dynamics of the observed system $\mathbf{x}$ is governed by the composition of observation operator G and forecasting operator $\Phi_{\theta,t}$. Regarding the literature on the Koopman operator theory, most approaches rely on the explicit identification of eigenfunctions and eigenvalues of the Koopman operator.[23,47,48] Our framework relates to the identification of the infinitesimal generator f_θ of the one-parameter subgroup defined by the Koopman operator through the ODE representation (4). By construction, the Koopman operator associated with the identified operator $f_{\hat{\theta}}$ is also diagonalizable such that the identification of infinitesimal generator $f_{\hat{\theta}}$ provides an implicit decomposition of the Koopman operator of the underlying and unknown dynamical system onto the eigenbasis of the learned latent dynamics governed by ODE (4).

Future work will further explore methodological aspects, especially the application to high-dimensional and stochastic systems. In the considered framework, the operator $\mathscr{M}$ is stated as an identity operator on the observed component of state $\mathbf{u}_t$ or as a simple PCA projection. Although for the geosciences community, using PCA to reduce the dimensionality is motivated by the Galerkin derivation of reduced order models from complex high dimensional governing partial differential equations,[49] using auto-encoders has shown promising results in discovering optimal coordinates when trained jointly with a dynamical system. The combination of the proposed framework with the variational setting considered in the latent-ODE model[7] also appears as an interesting direction for future work.

The extension to stochastic systems through the identification of a stochastic ODE is also of key interest, for instance, for future applications of the proposed framework to geophysical random flows, especially to the simulation and forecasting of ocean-atmosphere dynamics in which stochastic components naturally arise.[50]

ACKNOWLEDGMENTS

This work was supported by Labex Cominlabs (grant SEACS), Region Bretagne, CNES (grant OSTST-MANATEE), and Microsoft (AI EU Ocean awards) and by MESR, FEDER, Région Bretagne, Conseil Général du Finistère, Brest Métropole, and Institut Mines Télécom in the framework of the VIGISAT program managed by "Groupement Bretagne Télédétection," CNES (grant OSTST-MANATEE), and ANR (Melody and OceaniX grants). It benefited from HPC and GPU resources from Azure (Microsoft EU Ocean awards) and from GENCI-IDRIS (Grant No. 2020-101030).

APPENDIX A: PROOF OF PROPOSITION 1

This proposition can be easily extended to any observation function that does not form an embedding of the initial unobserved ODE. However, for the sake of simplicity, we will consider the example given in Eq. (1).

Let us suppose a smooth ODE in the observation space that governs the time evolution of $\mathbf{x}$ from Eq. (1),

$$\begin{cases} \dot{\mathbf{x}}_t = f(\mathbf{x}_t), \\ \mathbf{x}_{t_0} = \mathbf{x}_0. \end{cases} \tag{A1}$$

This ODE generates a flow $\mathbf{x}_t = \Psi_t(\mathbf{x}_0)$.

Since our observation operator is not one-to-one, we can assume the existence of some $\hat{t}, t_1, t_2$, where $Real(\Phi_{\hat{t}}(\mathbf{z}_{t_1})) = Real(\Phi_{\hat{t}}(\mathbf{z}_{t_2}))$ with $Real(\mathrm{z}_{t_1}) \neq Real(\mathrm{z}_{t_2})$ [Φ is the flow generated by the unobserved ODE illustrated in Eq. (1)]. Projecting this equality to the observation space leads to $\Psi_{\hat{t}}(\mathbf{x}_{t_1}) = \Psi_{\hat{t}}(\mathbf{x}_{t_2})$ with $\mathbf{x}_{t_1} \neq \mathbf{x}_{t_2}$.

Since the above ODE is smooth (or continuously differentiable), we can show that f is locally Lipschitz on any interval containing t_0,[51] which guarantees by Picard's existence theorem the existence of a unique solution.[39] Formally, for the times $\hat{t}, t_1, t_2$, $\Psi_{\hat{t}}(\mathbf{x}_{t_1}) = \Psi_{\hat{t}}(\mathbf{x}_{t_2})$ if and only if $\mathbf{x}_{t_1} = \mathbf{x}_{t_2}$. This contradicts the assumption that $\mathbf{x}_{t_1} \neq \mathbf{x}_{t_2}$, and thus there is no existence of $\hat{t}$ such that $Real(\Phi_{\hat{t}}(\mathbf{z}_{t_1})) = Real(\Phi_{\hat{t}}(\mathbf{z}_{t_2}))$ with $Real(\mathrm{z}_{t_1}) \neq Real(\mathrm{z}_{t_2})$.

APPENDIX B: DIMENSIONALITY ANALYSIS OF THE NbedDyn MODEL

One of the key parameters of the proposed model is the dimension of the latent space. Despite the fact that it is extremely challenging to get a prior idea of the dimension of the model in the case of real data experiments, and similarly to the performance analysis of the NbedDyn model illustrated, for instance, in Figs. 4 and 6, one can analyze the spanned manifold of the learned latent states to get an idea of the true dimension of the underlying model (true here stands for a sufficient dimension of the latent space). The idea here is to compute the modulus of the eigenvalues of the Jacobian matrix for each input of the training data. An eigenvalue does not influence the temporal evolution of the latent state if it has a modulus that tends to zero. The number of non-zero eigenvalues can then be seen as a sufficient dimension of the latent space.

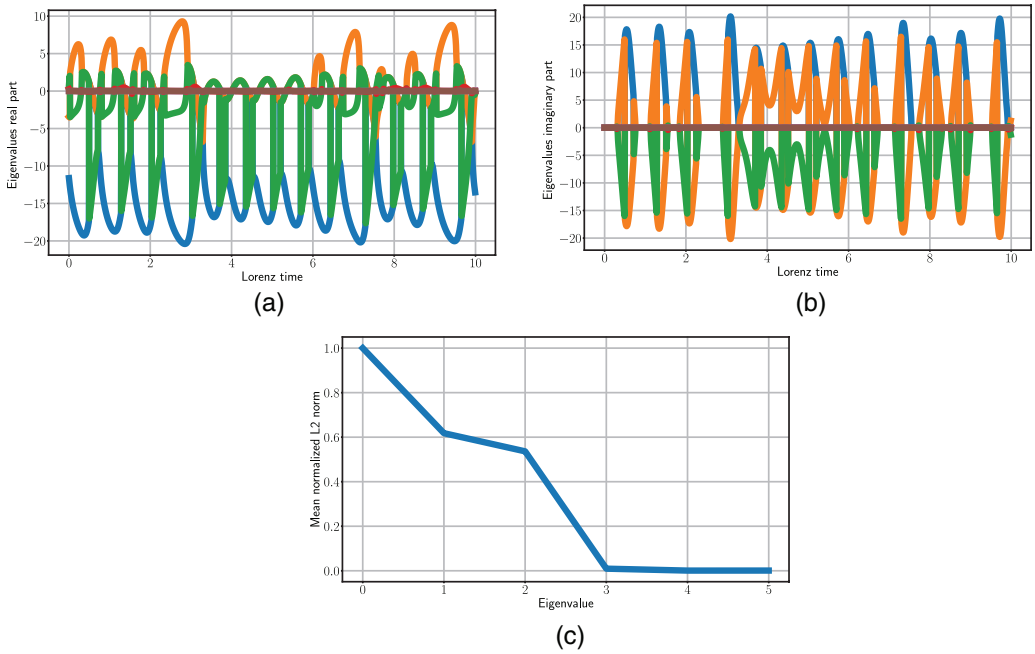

FIG. 7. Analysis of the eigenvalues of the NbedDyn model Jacobian matrix. Lorenz-63 case-study with $d_E = 6$. We illustrate the real part in (a), the imaginary part in (b), and the modulus in (c) of the eigenvalues of the Jacobian matrix.

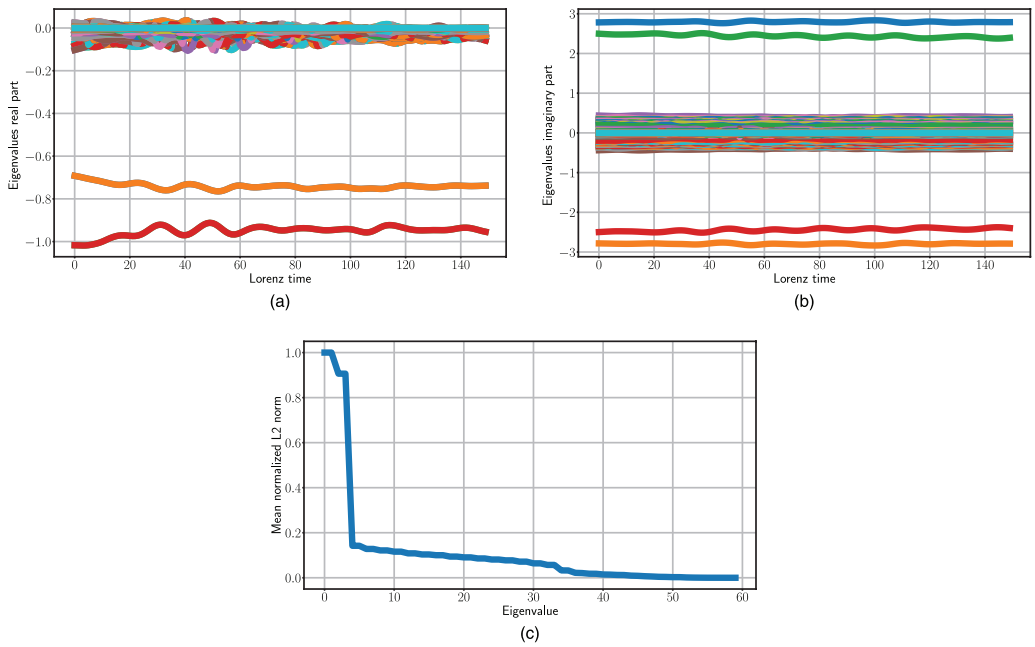

FIG. 8. Analysis of the eigenvalues of the NbedDyn model Jacobian matrix. Sea level anomaly case-study with $d_E = 60$. We illustrate the real part in (a), the imaginary part in (b), and the modulus in (c) of the eigenvalues of the Jacobian matrix.

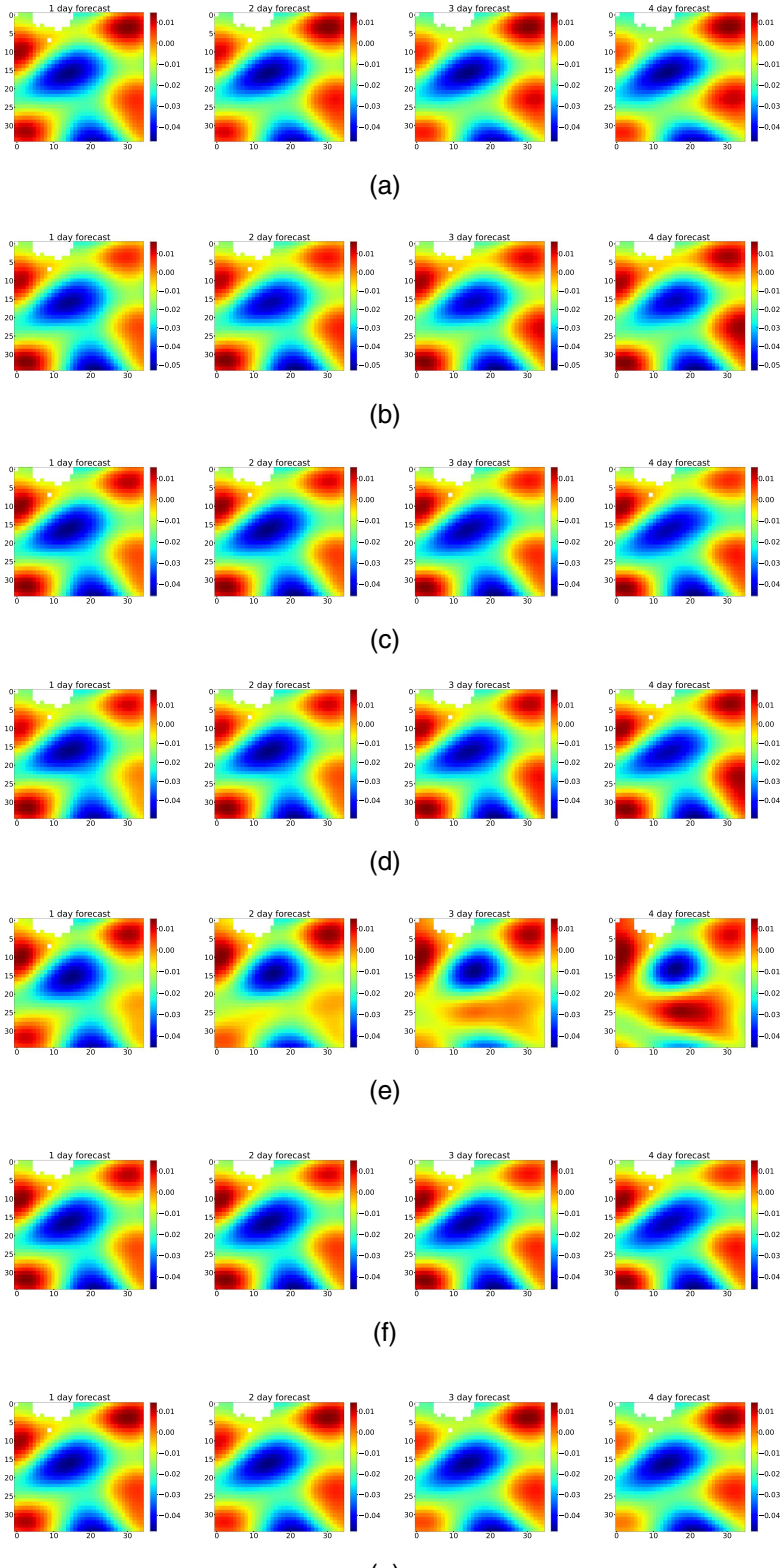

FIG. 9. Forecasted SLA states of the proposed models. We illustrate the forecasted SLA fields using analog forecasting in (b), sparse regression in (c), latent-ODE model in (d), RNN in (e), NbedDynZERO in (f), and the proposed architecture in (g) with respect to the true field illustrated in (a).

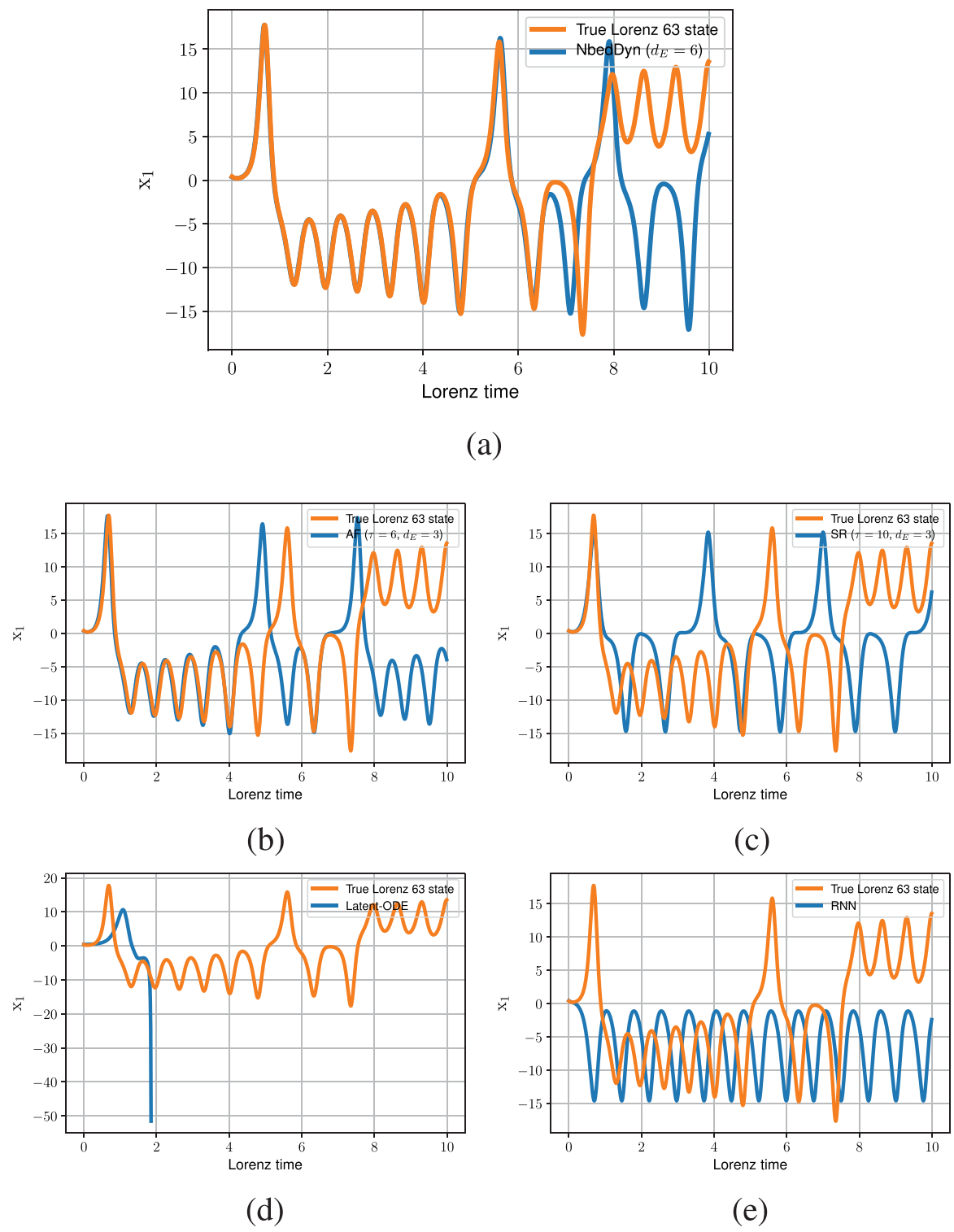

FIG. 10. Generated time series of the proposed models. Given an initial condition, we generated a time series of 1000 time steps using the proposed architecture in (a). We also show the forecasting performances, given the same initial condition, of the state-of-the-art models [(b)–(e)].

Regarding the identification of an ODE model governing the first state variable of the Lorenz 63 model, Fig. 7 illustrates the eigenvalues of the Jacobian matrix and their modulus for a dimension of the latent space $d_E = 6$. Interestingly, only three eigenvalues have non-zero modulus and are effectively influencing the underlying dynamics. This result shows that one can use a three-dimensional latent-space as a sufficient dimension to identify an ODE model governing the first state of the Lorenz 63 system, which is the same dimension as the true Lorenz 63 model.

The analysis of the eigenvalues of the sea level anomaly model in the other hand is not as straightforward as in the case of the Lorenz model since we do not have any idea on the analytical form of the underlying dynamical model. Figure 8 illustrates that using a 60-dimensional latent space for the NbedDyn model, only 50 eigenvalues have non-zero modulus and thus are effectively influencing the underlying dynamics. The conclusion in this case is that the observed SLA data evolve in a 50-dimensional latent space parametrized by the dynamical model f_θ.

APPENDIX C: ADDITIONAL FIGURES OF THE SEA LEVEL ANOMALY EXPERIMENT

Forecasted states of the sea level anomaly are illustrated in Fig. 9. The visual analysis of the forecasted SLA states emphasizes the relevance of the proposed NbedDyn model. While state of the art approaches generally overestimate the time evolution of some structures such as eddies, our model is the only one to give near perfect forecasting up to 4 days.

APPENDIX D: ADDITIONAL FIGURES OF THE LORENZ 63 EXPERIMENT

We illustrate the forecasting performance of the tested models for the Lorenz-63 experiment through an example of forecasted trajectories in Fig. 10. Our model with $d_E = 6$ leads to a trajectory similar to the true one up to seven Lyapunov times, when the best alternative approach diverge from the true trajectory beyond four Lyapunov times.

An other interesting experiment is to find the initial condition for new observation data. This issue is addressed as presented in Sec. III as follows. Given a new noisy and partial observation sequence (Fig. 11), we first look for a potential initial condition in the inferred training latent state sequence. This initial condition is then optimized using the cost function described by Eq. (7) to minimize the forecasting error of the new observation sequence.

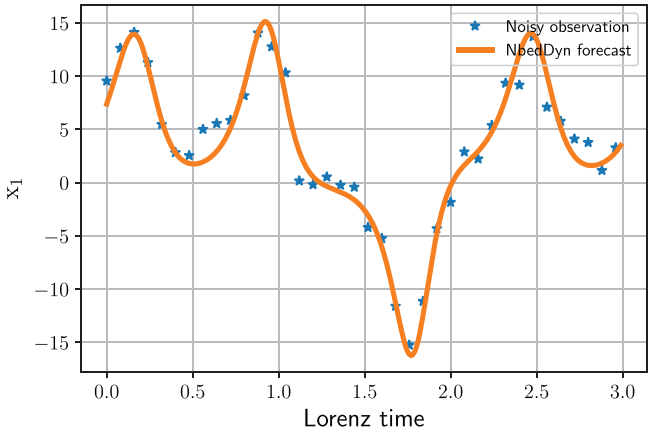

FIG. 11. Forecasted Lorenz 63 state sequence given noisy and partial observations. Given noisy and partial observations, our model optimizes Eq. (7) to infer an initial condition that minimizes the forecasting of the observations.

APPENDIX E: SCOPE AND LIMITATIONS

1. Constraining limit cycles

The proposed augmented ODE formulation does not suppose any prior knowledge on the underlying dynamics responsible for the temporal evolution of the observations. This can lead in some cases (especially when working on chaotic dynamics) to output a dynamical representation that has several attracting regions in addition to the one leading to the observations limit cycle. This can lead to inappropriate results when trying to find an initial condition that forecasts a given observation sequence. The idea of using the manifold spanned by the augmented training data allows to bypass this issue but we believe that adding additional constraints (energy preserving constraints, known symmetries in the models, etc.) can significantly improve the quality of the learned dynamical models.

APPENDIX F: NEURAL NETWORKS' HYPERPARAMETERS

1. Lorenz 63 experiments' hyperparameters

Tables III–V show the RNN, Latent-ODE and the NbedDyn parameters in the Lorenz 63 experiment.

TABLE III. RNN parameters in the Lorenz 63 experiment.

Parameter	Value
Number of LSTM layers	10
Hidden size	10
Sequence length	30
Learning rate	0.001
Optimizer	Adam
Training data	4000

TABLE IV. Latent-ODE parameters in the Lorenz 63 experiment, please refer to Ref. 7 for more details.

Parameter	Value
Latent dimension	4
Hidden size	15
RNN hidden size	100
Learning rate	0.01
Optimizer	Adam
Training data	4000

TABLE V. NbedDyn parameters in the Lorenz 63 experiment, please refer to Ref. 6 for more details.

Parameter	Value
Augmented latent dimension	6
Number of bilinear layers	6
Number of linear layers	6
Integration scheme	Runge–Kutta 4
Learning rate	0.001
Optimizer	Adam
Training data	4000

2. SLA experiments' hyperparameters

Tables VI–VIII show RNN parameters in the SLA experiment, latent-ODE parameters in the SLA experiment, and NbedDyn parameters in the SLA experiment.

TABLE VI. RNN parameters in the SLA experiment.

Parameter	Value
Number of LSTM layers	5
Hidden size	20
Sequence length	40
Learning rate	0.001
Optimizer	Adam
Training data	2000

TABLE VII. Latent-ODE parameters in the SLA experiment, please refer to Ref. 7 for more details.

Parameter	Value
Latent dimension	60
Hidden size	70
RNN hidden size	200
Learning rate	0.01
Optimizer	Adam
Training data	2000

TABLE VIII. NbedDyn parameters in the SLA experiment, please refer to Ref. 6 for more details.

Parameter	Value
Augmented latent dimension	60
Number of bilinear layers	60
Number of linear layers	60
Integration scheme	Runge–Kutta 4
Learning rate	0.001
Optimizer	Adam
Training data	2000

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," Ann. Stat. **10**, 154–166 (1982).
[2]H. D. I. Abarbanel and U. Lall, "Nonlinear dynamics of the great salt lake: System identification and prediction," Clim. Dyn. **12**, 287–297 (1996).
[3]J. Jeong and F. Hussain, "On the identification of a vortex," J. Fluid Mech. **285**, 69–94 (1995).
[4]T. C. Koopmans, "Identification problems in economic model construction," Econometrica **17**, 125–144 (1949).
[5]S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," Proc. Nat. Acad. Sci. **113**, 3932–3937 (2016).
[6]R. Fablet, S. Ouala, and C. Herzet, "Bilinear residual neural network for the identification and forecasting of geophysical dynamics," in *2018 26th European Signal Processing Conference (EUSIPCO)* (IEEE, 2018), pp. 1477–1481.
[7]T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems* (NeurIPS, 2018), pp. 6571–6583.
[8]D. Nguyen, S. Ouala, L. Drumetz, and R. Fablet, "Em-like learning chaotic dynamics from noisy and partial observations," arXiv:1903.10335[cs.LG] (2019).
[9]M. Bocquet, J. Brajard, A. Carrassi, and L. Bertino, "Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models," Nonlinear Process. Geophys. **26**, 143–162 (2019).
[10]J. Brajard, A. Carassi, M. Bocquet, and L. Bertino, "Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model," arXiv:2001.01520 (2020).
[11]F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980*, edited by D. Rand and L.-S. Young (Springer, Berlin, 1981), pp. 366–381.
[12]H. D. I. Abarbanel, "Modeling chaos," in *Analysis of Observed Chaotic Data* (Springer, New York, NY, 1996), pp. 95–114.
[13]J. Paduart, L. Lauwers, J. Swevers, K. Smolders, J. Schoukens, and R. Pintelon, "Identification of nonlinear systems using polynomial nonlinear state space models," Automatica **46**, 647–656 (2010).
[14]J. Frank, S. Mannor, and D. Precup, "Activity and gait recognition with time-delay embeddings," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI'10* (AAAI Press, 2010), pp. 1581–1586.
[15]A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, and O. K. Hussain, "Support vector regression with chaos-based firefly algorithm for stock market price forecasting," Appl. Soft Comput. **13**, 947–958 (2013).
[16]H. D. I. Abarbanel, "Choosing time delays," in *Analysis of Observed Chaotic Data* (Springer, New York, NY, 1996), pp. 25–37.
[17]H. D. I. Abarbanel, "Choosing the dimension of reconstructed phase space," in *Analysis of Observed Chaotic Data* (Springer, New York, NY, 1996), pp. 39–67.
[18]Z. Ghahramani and S. T. Roweis, "Learning nonlinear dynamical systems using an EM algorithm," in *Advances in Neural Information Processing Systems* (NeurIPS, 1999), pp. 431–437.
[19]J. Wang, A. Hertzmann, and D. J. Fleet, "Gaussian process dynamical models," in *Advances in Neural Information Processing Systems* (NeurIPS, 2006), pp. 1441–1448.
[20]P. Mirowski and Y. LeCun, "Dynamic factor graphs for time series modeling," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Springer, 2009), pp. 128–143.
[21]K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," arXiv:1512.03385[cs] (2015).
[22]R. G. Krishnan, U. Shalit, and D. Sontag, "Structured inference networks for nonlinear state space models," arXiv:1609.09869[cs,stat] (2016).

[23]B. O. Koopman, "Hamiltonian systems and transformations in hilbert space," Proc. Nat. Acad. Sci. U.S.A. **17**, 315–318 (1931).
[24]The word smooth here stands for continuously differentiable or $\mathscr{C}^1$.
[25]T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," J. Stat. Phys. **65**, 579–616 (1991).
[26]M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," Science **324**, 81–85 (2009).
[27]Y. Yuan, X. Tang, W. Zhou, W. Pan, X. Li, H.-T. Zhang, H. Ding, and J. Goncalves, "Data driven discovery of cyber physical systems," Nat. Commun. **10**, 1–9 (2019).
[28]W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and C. Grebogi, "Predicting catastrophes in nonlinear dynamical systems by compressive sensing," Phys. Rev. Lett. **106**, 154101 (2011).
[29]S. Wiewel, M. Becher, and N. Thuerey, "Latent-space physics: Towards learning the temporal evolution of fluid flow," arXiv:1802.10123[cs.LG] (2018).
[30]M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Multistep neural networks for data-driven discovery of nonlinear dynamical systems," arXiv:1801.01236 (2018).
[31]S. Ouala, A. Pascual, and R. Fablet, "Residual integration neural network," in *ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2019), pp. 3622–3626.
[32]G. Shen, J. Kurths, and Y. Yuan, "Sequence-to-sequence prediction of spatiotemporal systems," Chaos **30**, 023102 (2020).
[33]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).
[34]W. Gilpin, "Deep reconstruction of strange attractors from time series," arXiv:2002.05909[cs.LG] (2020).
[35]R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, "The analog data assimilation," Mon. Weather Rev. **145**, 4093 (2017).
[36]E. Dupont, A. Doucet, and Y. W. Teh, "Augmented neural ODEs," arXiv:1904.01681 (2019).
[37]H. Zhang, X. Gao, J. Unterman, and T. Arodz, "Approximation capabilities of neural ordinary differential equations," arXiv:1907.12998 (2019).
[38]K. Champion, B. Lusch, J. N. Kutz, and S. L. Brunton, "Data-driven discovery of coordinates and governing equations," arXiv:1904.02107 (2019).
[39]E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations* (Tata McGraw-Hill Education, 1955).
[40]P. Lynch and X.-Y. Huang, "Initialization," in *Data Assimilation: Making Sense of Observations*, edited by W. Lahoz, B. Khattatov, and R. Menard (Springer, Berlin, 2010), pp. 241–260.
[41]E. N. Lorenz, "Deterministic nonperiodic flow," J. Atmos. Sci. **20**, 130–141 (1963).
[42]A. C. Hindmarsh, "ODEPACK, a systematized collection of ODE solvers," IMACS Trans. Sci. Comput. **1**, 55–64 (1983).
[43]J. C. Sprott, *Chaos and Time-Series Analysis* (Oxford University Press, Inc., New York, NY, 2003).
[44]S. Calmant, F. Seyler, and J. F. Cretaux, "Monitoring continental surface waters by satellite altimetry," Surv. Geophys. **29**, 247–269 (2008).
[45]M. Juza, B. Mourre, L. Renault, S. Gómara, K. Sebastián, S. Lora, J. P. Beltran, B. Frontera, B. Garau, C. Troupin, M. Torner, E. Heslop, B. Casas, R. Escudier, G. Vizoso, and J. Tintoré, "Socib operational ocean forecasting system and multi-platform validation in the western mediterranean sea," J. Oper. Oceanogr. **9**, s155–s166 (2016).
[46]F. d'Ovidio, A. Pascual, J. Wang, A. M. Doglioli, Z. Jing, S. Moreau, G. Grégori, S. Swart, S. Speich, F. Cyr, B. Legresy, Y. Chao, L. Fu, and R. A. Morrow, "Frontiers in fine-scale in situ studies: Opportunities during the swot fast sampling phase," Front. Mar. Sci. **6**, 168 (2019).
[47]S. L. Brunton, B. W. Brunton, J. L. Proctor, and J. N. Kutz, "Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control," PLoS ONE **11**, e0150171 (2016).
[48]J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, "On dynamic mode decomposition: Theory and applications," J. Comput. Dyn. **1**, 391–421 (2014).
[49]P. Holmes, J. L. Lumley, G. Berkooz, and C. W. Rowley, "Galerkin projection," in *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge Monographs on Mechanics, 2nd ed. (Cambridge University Press, 2012), pp. 106–129.
[50]B. Chapron, P. Dérian, E. Mémin, and V. Resseguier, "Large-scale flows under location uncertainty: A consistent stochastic framework," Q. J. R. Meteorol. Soc. **144**, 251–260 (2018).
[51]H. H. Sohrab, *Basic Real Analysis* (Springer, 2003), Vol. 231.