Nonlinear Dyn (2022) 108:3783–3805
https://doi.org/10.1007/s11071-022-07365-y

ORIGINAL PAPER

Estimating initial conditions for dynamical systems with incomplete information

Blas Kolic · **Juan Sabuco** · **J. Doyne Farmer**

Received: 12 October 2021 / Accepted: 12 March 2022 / Published online: 1 April 2022

Abstract In this paper, we study the problem of inferring the latent initial conditions of a dynamical system under incomplete information, i.e., we assume we observe aggregate statistics of the system rather than its state variables directly. Studying several model systems, we infer the microstates that best reproduce an observed time series when the observations are sparse, noisy, and aggregated under a (possibly) nonlinear observation operator. This is done by minimizing the least-squares distance between the observed time series and a model-simulated time series using gradient-based methods. We validate this method for the Lorenz and Mackey–Glass systems by making out-of-sample predictions. Finally, we analyze the predicting power of our method as a function of the number of observations available. We find a critical transition for the Mackey–Glass system, beyond which it can be initialized with arbitrary precision.

Keywords Latent state estimation · Chaos · Gradient descent optimization · Data assimilation · Incomplete information · Initialization

B. Kolic (✉) · J. Sabuco · J.D. Farmer
Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford, UK
e-mail: blas.kolic@maths.ox.ac.uk

B. Kolic · J. Sabuco · J.D. Farmer
Mathematical Institute, University of Oxford, Oxford, UK

J. Sabuco
School of Geography and the Environment, University of Oxford, Oxford OX1 3QY, UK

J. Sabuco
Smith School of Enterprise and the Environment, University of Oxford, Oxford, UK

J.D. Farmer
Santa Fe Institute, Santa Fe, USA

1 Introduction

We often model empirical processes as complex systems: they are composed of different agents or *microstates*, that interact through simple rules but evolve in non-trivial ways. A lot of research has taken a complex systems approach, where authors have constructed models from the bottom up, but where empirical observations exhibit incomplete information of the system. Applications range from Earth's global climate [1], to urban dynamics [2], to the human brain [3], to financial markets [4], to transportation networks [5], and marine fisheries [6]. However, data are usually available as an aggregate statistic of the microstates because it is difficult to measure microstates directly. Thus, it is an ongoing challenge to develop methods to recover the latent microstates from aggregate and incomplete observations [7].

Estimating and forecasting complex systems accurately depends on (1) the inherent complexity of the system, which depends on things like the systems' state space dimension, its Lyapunov exponents, and its attractors, (2) the sparsity and quality of the data, and

(3) our ability to model them. In low-dimensional systems with high-quality data, as shown by Packard *et al.* [8] and Takens [9], state-space reconstruction techniques can be applied to partial information to create a representation as a dynamical system. By reconstructing an attractor from data, we can make accurate predictions by choosing its closest points to the current state of the system and extrapolating them [10]. These techniques work well even without any modeling [11].

In high-dimensional systems, it is typically not possible to use time series models. It is nonetheless often possible to use a theoretical model, if one can only measure the initial conditions that the model requires and compare them to the data. The task of estimating initial conditions that match observations is known as *initialization*. Similar to the state-space reconstruction techniques, we require to know the evolution function of the underlying dynamics of the system, or at least some approximation of it [12]. In particular, if the observations available are an aggregate of the dynamical system, then the process is known as *microstate initialization*, or latent state initialization. To do this we need to know how the microstate is aggregated, in addition to having a model of its dynamics.

In the fields of meteorology and numerical weather prediction (NWP), researchers have developed a framework for estimating the latent states of a system [13,14], where a large number of observed states are available. This framework is known as data assimilation, and, although it is well justified from the Bayesian perspective, it incorporates several heuristics pertinent to the weather prediction field. For instance, modelers often incorporate Gaussian priors in the cost functionals involved with known statistics about the latent states [15]. Ideas from data assimilation have already permeated outside the NWP community, such as in urban dynamics [2]. Typically, data assimilation methods operate in a sequential matter: the microstate at time t gets nudged (or corrected) so it optimally approaches the empirical observation at t. Then, the modeler simulates the nudged microstate from time t to time $t+1$, where the microstate is again nudged. Therefore, the resulting sequence of microstates is not a solution of the underlying dynamical system—the microstate was constantly altered throughout its trajectory. We, in contrast, seek to analyze real trajectories of the latent states, so this sequential approach does not suit our goal.

The initialization process is an optimization problem where we minimize a cost function that depends on some notion of distance between the observed and the model-generated data. Hence, in high-dimensional systems, it is essential to develop efficient algorithms to find initial conditions with high precision. Research has been done around the parameter estimation of stochastic dynamical systems in low-dimensional parameter spaces [7,16]. Among other alternatives, gradient descent has proven to be superior in strongly nonlinear models [17,18], but the main drawback is that it can get stuck in local minima. Other methods, like genetic algorithms, simulated annealing, and other meta-heuristics algorithms [19], usually find the global minimum in low dimensions, but they are likely to diverge in high-dimensional systems. We provide a thorough comparison of the state-of-the-art gradient descent methods [20] and discuss their performance in the context of microstate initialization.

Here, we propose a gradient-based method for initializing chaotic systems where only aggregate observations of short observation windows are available. Using a combination of numerical simulations and analytical arguments, we study the conditions in which certain systems may be initialized with arbitrary precision. We explore the performance of several gradient descent algorithms and the effects of observational noise on the accuracy and convergence of the initialization process. Furthermore, we quantify the accuracy of our method numerically with out-of-sample forecasts. Under this framework, we offer a better understanding of what information the observations provide about a system's underlying dynamics, and, additionally, lay out the connections of our method to those in the data assimilation literature.

The remainder of this paper is laid out as follows. In Sect. 2, we describe the initialization problem under the framework of dynamical systems and develop our initialization methodology accordingly. In Sect. 3, we test our method in two systems, namely the Lorenz and Mackey–Glass systems. Finally, in Sect. 4, we discuss our results and suggest further research directions.

2 Methods

2.1 Problem setup

A wide class of dynamical systems can be formulated as follows

$$\boldsymbol{x}(t+\Delta t) = \boldsymbol{f}(\boldsymbol{x}(t);\boldsymbol{\theta}) + \boldsymbol{\xi}(t)\,, \tag{1}$$

where $\boldsymbol{x}(t) \in \mathcal{X} \subset \mathbb{R}^{N_x}$ is the (N_x-dimensional) state of the system at time $t \in \mathbb{R}$ living in some manifold $\mathcal{X}$, $\boldsymbol{\xi}(t) \in \mathbb{R}^{N_x}$ are random variables that model the intrinsic noise of the system and $\boldsymbol{\theta} \in \mathbb{R}^d$ is a vector of parameters. We will hereafter refer to the state $\boldsymbol{x}$ as the *microstate* to emphasize that $\boldsymbol{x}$ is not directly observable and is potentially high dimensional. If $\Delta t > 0$ is finite, the model $\boldsymbol{f} : \mathcal{X} \times \mathbb{R}^d \to \mathcal{X}$ is a discrete mapping that takes the microstate from time t to time $t + \Delta t$. If Δt is infinitesimal, the system (1) defines a continuous-time dynamical system. Here, we assume that the dynamical system $\boldsymbol{f}$ is deterministic and perfectly specified, so that Eq. (1) depends on neither $\boldsymbol{\xi}$ nor $\boldsymbol{\theta}$.

We are interested in using information from the past to estimate the microstate $\boldsymbol{x}(t_0)$ of a dynamical system $\boldsymbol{f}$ at the present time t_0. We assume we know $\boldsymbol{f}$, but cannot directly observe its microstate $\boldsymbol{x}(t)$. Instead, we are only able to observe a sequence of observations $\boldsymbol{y} = (y_{-T}, \ldots, y_0)$ measured at times $\{t_{-T}, \ldots, t_0\}$, where $y_k \in \mathbb{R}^{N_y}$ and $N_y < N_x$, i.e., we are interested in measurements that lose information of the microstate by reducing its dimension from N_x to N_y. Moreover, these observations can be noisy. Thus, we can relate the observations to the dynamical system as

$$y_k = \mathcal{H}\big(\boldsymbol{x}(t_k)\big) + \epsilon_k \,. \tag{2}$$

where $\mathcal{H} : \mathcal{X} \subset \mathbb{R}^{N_x} \to \mathbb{R}^{N_y}$ is a known (possibly) nonlinear *observation operator* and ϵ_k accounts for *observational noise*. Although it is not necessary for the following discussions, we will assume for simplicity that $N_y = 1$ and, that the noise has a known variance σ_y. Our choice of indices from $-T$ to 0 for y_k reflects our interest in the *present time* t_0.

We assume that the observations are sampled at a uniform rate as follows

$$\Delta t_k = t_{k+1} - t_k = m\Delta t \,,$$

where the nonzero integer m is the *sampling interval*[1]. We treat m as a parameter that controls how often we sample observations from the system. Thus, we can recast Eq. (1) as a discrete-time mapping

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}^{\{m\}}(\boldsymbol{x}_k) = \boldsymbol{f}^{\{mk\}}(\boldsymbol{x}_{-T}) \,, \tag{3}$$

where $\boldsymbol{f}^{\{i\}}(\cdot)$ is the composition of $\boldsymbol{f}$ with itself i-times, $\boldsymbol{x}_{-T} = \boldsymbol{x}(t_{-T})$ is the latent microstate at the time of the first observation, and $\boldsymbol{f}^{\{k\}}$ is the self-iteration of f by k times. Note that the whole evolution of the microstates is determined by $\boldsymbol{x}_{-T}$, so the *microstate initialization problem* is that of obtaining the best approximation to $\boldsymbol{x}_0$ given the observations from the *assimilation time* t_{-T} up to present time t_0. These observations, alongside with the model (3) and the observation operator (2), form a nonlinear system of equations with N_x variables and T equations corresponding to the dimension of the microstate space and the number of observations.

[1] The initialization procedure we introduce in the following section works with irregularly sampled observations as well.

2.2 Initialization procedure

As we stated previously, the microstate initialization problem is that of obtaining the best representation of the present-time microstate $\boldsymbol{x}_0$ given the history of observations $\boldsymbol{y}$ under the dynamical system $\boldsymbol{f}$. We define what *best* means in what follows. We make the code of the initialization procedure freely available at https://github.com/blas-ko/LatentStateInitialization.

2.2.1 Cost function

Given an estimate $\boldsymbol{x}$ of the ground-truth microstate $\boldsymbol{x}_{-T}$ at the assimilation time t_{-T}, a natural way to quantify the goodness of $\boldsymbol{x}$ is by measuring the point-to-point discrepancy between the observed and estimated data. We do so with the following mean-squared *cost function*.[2]

$$\mathcal{J}(\boldsymbol{x}) = \frac{1}{T\sigma_y^2} \sum_{k=-T}^{0} \big(y_k - \hat{y}_k\big)^2 \,, \tag{4}$$

where

$$\hat{y}_k(\boldsymbol{x}) = \mathcal{H}\Big(\boldsymbol{f}^{\{mk\}}(\boldsymbol{x})\Big) \tag{5}$$

is our estimate of observation y_k given $\boldsymbol{x}$, and σ_y, the variance of the observed data, is a normalization constant that is arbitrary but will be of practical use later. We call the state $\hat{\boldsymbol{x}}_{-T}$ that minimizes $\mathcal{J}$ the *assimilated microstate* and the present-time state $\hat{\boldsymbol{x}}_0 = \boldsymbol{f}^{\{mT\}}(\hat{\boldsymbol{x}}_{-T})$ the *initialized microstate*. Our goal

[2] We may generalize this cost function for non-scalar observations as

$$\mathcal{J}(\boldsymbol{x}) = \frac{1}{T} \sum_k (\boldsymbol{y}_k - \hat{\boldsymbol{y}}_k)^\dagger \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{y}_k - \hat{\boldsymbol{y}}_k) \,,$$

where $\boldsymbol{\Sigma}_y$ is the covariance matrix of the observations.

is to find an initialized microstate $\hat{\boldsymbol{x}}_0$ that is a good representation of the ground-truth microstate $\boldsymbol{x}_0$.

The cost function $\mathcal{J}$ is known as a *filter* in the interpolation assimilation community, a *least-squares optimizer* in the optimization and machine learning communities, and *4D-Var* with infinite state uncertainty in the variational assimilation community. From a Bayesian perspective, the cost function $\mathcal{J}$ emerges naturally when we assume the prior $p(\boldsymbol{x})$ is uniform and observational noise is Gaussian: the posterior $p(\boldsymbol{x}|\boldsymbol{y})$ is maximal when $\mathcal{J}$ is minimal [14].

If the observations in the time series $\boldsymbol{y}$ are noiseless, and given that we assume that the model $\boldsymbol{f}$ perfectly describes the system, then $\mathcal{J}$ has a global minimum $\hat{\boldsymbol{x}}_{-T}$ for which $\mathcal{J}(\hat{\boldsymbol{x}}_{-T}) = 0$. However, even for the noiseless scenario, $\hat{\boldsymbol{x}}_{-T}$ is not necessarily unique. Take, for instance, the Lorenz system [21], which is symmetric around its x-axis, and take an observation operator of the form $\mathcal{H}(\boldsymbol{x}) = \boldsymbol{x}^{\dagger}\boldsymbol{x}$ with $\dagger$ denoting matrix transposition. Note that for any microstate $\boldsymbol{x}$ in the axis of symmetry, the condition $-\boldsymbol{x}$ will produce the same sequence of noiseless observations, so $\boldsymbol{x}$ and $-\boldsymbol{x}$ are indistinguishable in terms of $\mathcal{J}$. We will refer to the set of indistinguishable microstates as the *feasible set* of solutions, and we will denote it as $\boldsymbol{\Omega}$.

An additional problem arises when the observations contain noise. In any finite time series, there might exist microstates with a lower value of the cost function than the ground-truth microstate—i.e., where $\mathcal{J}(\hat{\boldsymbol{x}}_{-T}) < \mathcal{J}(\boldsymbol{x}_{-T})$ for $\hat{\boldsymbol{x}}_{-T} \notin \boldsymbol{\Omega}$—which we will call *dominating microstates*, following [22]. Judd *et al.* [22] suggest that using cost functions that minimize both the variance (as in Eq. (4)) and the kurtosis will do a better job of identifying the true microstate when the observational noise is Gaussian. While this is a good idea when there are many observations, we consider short time series in this work, which calls for other alternatives. Instead of modifying the cost function $\mathcal{J}$, we preprocess the data to reduce the probability of finding dominating trajectories outside of the feasible set $\boldsymbol{\Omega}$.

Following [13], we distinguish between the *error-free* and the *noise* contributions to the cost function $\mathcal{J}$. Isolating the contributions from the discrepancy between ground-truth and assimilated microstate and the observational noise will let us design a well-suited methodology to deal with noise and dominating trajectories in a separate manner. Recall that, given $\boldsymbol{x}_{-T}$, $\mathcal{H}(\boldsymbol{x}_k)$ is the error-free observation at time t_k, ϵ_k its associated noise and $\hat{y}_k$ our estimation of y_k. Thus, we can decompose Eq. (4) into three terms of the form

$$\mathcal{J}(\boldsymbol{x}) = \frac{1}{\sigma_y^2}\Big[\underbrace{\frac{1}{T}\sum_k (\mathcal{H}(\boldsymbol{x}_k) - \hat{y}_k)^2}_{\text{noise-free cost}} + \underbrace{\frac{1}{T}\sum_k \epsilon_k^2}_{\text{obseravation noise}} - \underbrace{\frac{2}{T}\sum_k (\mathcal{H}(\boldsymbol{x}_k) - \hat{y}_k)\epsilon_k}_{\text{error-noise covariates}}\Big]$$

The first term on the RHS is the noise-free cost, $\mathcal{J}_{\text{free}}(\boldsymbol{x})$, that we would obtain in the absence of observational error. The second term is the average contribution of the square of the noise, which converges to the noise variance, σ_n^2, for large T. Finally, the third term captures how the noise and the noise-free cost vary together, which goes to 0 for large T because we assume the observational noise and the system dynamics are uncorrelated. Considering a large number of observations, we can thus approximate $\mathcal{J}$ as

$$\mathcal{J}(\boldsymbol{x}) \approx \mathcal{J}_{\text{free}}(\boldsymbol{x}) + \frac{\sigma_n^2}{\sigma_y^2}\,, \tag{6}$$

which shows the expected behavior of Eq. (4) in the presence of noise.

Now, note that if we take $\hat{y}_k = \mathbb{E}[\boldsymbol{y}]$ for all k [10], which is the best constant predictor for the observed time series, then $\mathcal{J}_{\text{free}} = 1$ and, therefore

$$\mathcal{J}(\boldsymbol{x} : \hat{y}_k = \mathbb{E}[\boldsymbol{y}]\ \forall k) = \mathcal{J}_{\text{const}} := 1 + \sigma_n^2/\sigma_y^2.$$

Naturally, we want to find an assimilated microstate $\hat{\boldsymbol{x}}_{-T}$ that performs better than a constant predictor, so that $\mathcal{J}(\hat{\boldsymbol{x}}_{-T}) \leq \mathcal{J}_{\text{const}}$. This motivates us to consider microstates $\boldsymbol{x}$ such as

$$\{\boldsymbol{x} : \mathcal{J}(\boldsymbol{x}) \leq \alpha + \frac{\sigma_n^2}{\sigma_y^2}\beta\}\,, \tag{7}$$

where the parameter $\alpha \in (0, 1]$ accounts for the error-free tolerance about the global minimum while $\beta \in (0, 1]$ accounts for the noise tolerance. We will explore these parameters in the following sections.

The above discussion suggests that we should pay especial attention on handling dominating trajectories, which might be present by either the presence of observational noise or by microstates that live far away from

the ground-truth but that have low cost function values. Thus, we propose the following three stages to initialize the microstate from aggregate observations: (1) a *preprocessing* stage in which we reduce the noise of the observations, (2) a *bounding* stage in which we limit the region of the microstate space in which we search for an optimal solution, and (3) a *refinement* stage in which we minimize the cost function (4) in a small search space and estimate the optimal microstate given the observations.

2.2.2 Preprocess: noise reduction

First, we preprocess the observed time series to reduce the observational noise and thus lower the probability of obtaining dominating microstates. Casdagli *et al.* [23] showed that in the presence of observational noise, the distribution of local minima gets increasingly complex with increasing levels of noise, especially when the dynamics is chaotic. We handle time series with only a handful of data points (in the order of 100 data points or less), so reducing noise by orbit shadowing [12] or large window impulse response filters [24] are not suitable options. Instead, we find that the best way to reduce the variance of the noise is using a low-pass moving average (LPMA) filter,

$$z_k := \begin{cases} \frac{1}{2}y_k + \frac{1}{2}y_{k+1} & k = -T \\ \frac{1}{2}y_k + \frac{1}{2}y_{k-1} & k = 0 \\ \frac{1}{2}y_k + \frac{1}{4}\left(y_{k-1} + y_{k+1}\right) & \text{otherwise} \end{cases}, \tag{8}$$

where z_k is the filtered data point at time t_k. We control the amount of noise reduction by repeatedly feeding the signal back into the LPMA filter of Eq. (8). Feeding the signal back into the filter q times, hereafter denoted as z_k^q, is equivalent to increasing the filtering window from three to $2q + 1$ points, making the filtered signal smoother.

We expect for the resulting variance, $(\sigma_n^q)^2$, to be lower than the original noise variance σ_n^2. Thus, following [25] and assuming we can rewrite the filtered signal in terms of the microstates as $z_k^q = \mathcal{H}(\boldsymbol{x}_k) + \epsilon_k^q$, we can measure the performance of the LPMA filter with the increase of the signal-to-noise ratio

$$r_0 = \sqrt{\frac{\sum_k (\epsilon_k)^2}{\sum_k (\epsilon_k^q)^2}} \approx \frac{\sigma_n}{\sigma_n^q}, \tag{9}$$

where $r_0 > 1$ whenever $\sigma_n > \sigma_n^q$.

The resulting noise distribution of the filtered signal converges to a zero-mean Gaussian distribution if we have either many data samples or set q big enough. However, if q is too big, we may filter parts of the dynamics and mix them into noise, resulting in exotic noise distributions (see Fig. 12 in Appendix C for examples). What *big enough*, *many samples*, and *too big* mean depend heavily on the dynamical system and the noise distribution, although we stress that the LPMA filter works optimally when the noise distribution has higher frequency spectrum on average than that of the dynamics of the system (see [26], Chapter 6.4.3.1.).

2.2.3 Bound: exploring the attractor

After the preprocessing stage, the next step is to bound the search space. Under no constraints in the cost function, the initialization procedure consists on searching through the whole microstate space $\mathcal{X}$ for a set of microstates that minimize Eq. (4), with no prior preference on where to start the search from. However, many real-world systems are dissipative, meaning that their dynamics relax into an attractor, i.e., a subset manifold $\mathcal{M} \subset \mathcal{X}$ of the microstate space. Thus, it is safe to assume that the observations $\boldsymbol{y}$ derive from a sequence of microstates that live in or near $\mathcal{M}$, and, by the properties of dissipative systems, any microstate $\boldsymbol{x}$ in the basin of attraction will eventually visit every point in $\mathcal{M}$ [27]. This means that, if we wait for long enough, then any point in the basin of attraction will get arbitrarily close to the ground truth microstate.

The bounding stage consists of exploiting the dissipative nature of real-world systems and letting any arbitrary estimate of the microstate explore the basin of attraction until it *roughly* approaches the ground truth microstate. To be more precise, we say that the microstate $\boldsymbol{x}_{-T}^R \in \mathcal{X}$ *roughly approaches* $\boldsymbol{x}_{-T}$ if

$$\mathcal{J}(\boldsymbol{x}_{-T}^R) \leq \delta_R := \alpha_R + \frac{\sigma_n^2}{\sigma_y^2}\beta_R, \tag{10}$$

for some *rough threshold* $0 \ll \delta_R < 1$. Thus, we let an arbitrary microstate evolve according to the model $\boldsymbol{f}$ until Eq. (10) is satisfied.

At this stage, we want to obtain solutions with a cost value of the order of the unfiltered noise level σ_n^2/σ_y^2, so that $\boldsymbol{x}_{-T}^R$ is either near to the feasible set $\boldsymbol{\Omega}$ or to any of the dominating microstates driven by the noise. To achieve this, it suffices to set $\beta_R \sim \mathcal{O}(1)$ and $\alpha_R < \sigma_n^2/\sigma_y^2 \leq 1$.

We note that whenever the time series $\boldsymbol{y}$ is noiseless, the bounding stage is only driven by α_R, the error-free

tolerance of the points situated at the global minima of the cost function, which are exactly those in the feasible set $\mathbf{\Omega}$. Thus, our choice of α_R leverages how closely we approach to $\mathbf{\Omega}$. If we set α_R too close to 0, we would impose for $\hat{\boldsymbol{x}}_{-T}$ to lay near $\mathbf{\Omega}$; however, it would take too long simulation times to satisfy Eq. (10) for this approach to be practical. The idea, ultimately, is to set the lowest δ_R possible such that the time to satisfy Eq. (10) is *short*.

2.2.4 Refine: cost minimization

The final step of the initialization procedure is refining $\boldsymbol{x}^R_{-T}$. By this point, we expect that $\boldsymbol{x}^R_{-T}$, our estimate of $\boldsymbol{x}_{-T}$, has bypassed most of the high-valued local minima of the cost function landscape. Additionally, we preprocessed the observations $\boldsymbol{y}$ to reduce their observational noise, but we have not fully exploited such preprocessing yet. By reducing the variance of the observational noise, we lower the number of dominating trajectories of the cost function—i.e., trajectories for $\boldsymbol{x} \notin \mathbf{\Omega}$ such that $\mathcal{J}(\boldsymbol{x}) < \mathcal{J}(\boldsymbol{x}_{-T})$. Thus, starting from $\boldsymbol{x}^R_{-T}$, we can minimize $\mathcal{J}$ using any optimization scheme until

$$\mathcal{J}(\hat{\boldsymbol{x}}_{-T}) \leq \delta_r := \alpha_r + \frac{\sigma_n^2}{\sigma_y^2}\beta_r \tag{11}$$

for some *refinement threshold* $0 < \delta_r \ll 1$, with $\hat{\boldsymbol{x}}_{-T}$ the assimilated microstate of the system. We then define $\hat{\boldsymbol{x}}_0 = \boldsymbol{f}^{\{mT\}}(\hat{\boldsymbol{x}}_{-T})$ to be the initialized microstate, hoping that $\hat{\boldsymbol{x}}_0$ is a good representation of $\boldsymbol{x}_0$.

We want for $\hat{\boldsymbol{x}}_{-T}$ to have *the lowest cost possible* at this stage. In terms of the error-free tolerance, we look for $\alpha_r \ll \alpha_R < \sigma_n^2/\sigma_y^2 \leq 1$, but the actual magnitude of α_r is left to the modeler to choose. Regarding the contribution of the noise, recall from Eq. (9) that $r_0 \approx \sigma_n/\sigma_n^q > 1$, so the lowest expected cost we can get is $\sigma_n^2/\sigma_y^2 r_0^{-2}$ (see Eq. (6)). Thus, we define the *refinement bound* of our initialization procedure as that of setting $\alpha_r \ll \alpha_R$ and $\beta_r \sim \mathcal{O}(r_0^{-2})$.

For our optimization scheme, we explore a plethora of the most successful *gradient-based algorithms* in the literature [20]. These algorithms include stochastic gradient descent [28], momentum descent [29], Nesterov [30], Adagrad [31], Adadelta [32], Rmsprop [33], Adam [34], and AdamX [35] and YamAdam [36].

In all cases, we set the hyper-parameters to be those given in the literature. We then compute the gradient of $\mathcal{J}$ using centered finite differences of step size $\sqrt{\varepsilon_M} \approx 1.5 \times 10^{-8}$, where ε_M corresponded to double precision arithmetic on our machine. In the absence of observational noise, provided that the dynamics is not degenerate and that we have sufficient observations, we expect that the feasible set $\mathbf{\Omega}$ collapses to the ground-truth microstate only, so at this stage we expect to infer it with high precision from the data.

2.3 Validation

We validate the initialized microstate $\hat{\boldsymbol{x}}_0$ by comparing them with the present-time microstate $\boldsymbol{x}_0$ and making out-of-sample predictions of the observed time series. Recall that we take the convention in which the observations, $\boldsymbol{y} = (y_{-T}, \ldots, y_0)$, have non-positive time indexes, so we refer to the observation times t_k for $k < 0$ as *assimilative* while we refer to times for $k \geq 0$ as *predictive*. We measure the *discrepancy* between the real and the simulated observations using both the normalized squared error in the observation space and the normalized error in the model space, i.e.,

$$\mathrm{NSE}_k^{obs} = \frac{(y_k - \hat{y}_k)^2}{\sigma_y^2}, \tag{12}$$

$$\mathrm{NSE}_k^{mod} = \frac{1}{N_x}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k)^\dagger \Sigma_{\boldsymbol{x}}^{-1}(\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k), \tag{13}$$

where σ_y^2 is the variance of the data, $\Sigma_{\boldsymbol{x}}$ the covariance matrix of the microstates and $(\cdot)^\dagger$ denotes matrix transposition. In general, $\boldsymbol{x}_k$ and $\Sigma_{\boldsymbol{x}}$ are unknown to the modeler, but we use them to measure the performance of our initializations in the latent space of the microstates.

Note that if we let $T \to \infty$, then $\mathcal{J}(\boldsymbol{x})$ converges to $\mathbb{E}[NSE_0^{obs}]$. When k grows, the trajectories y_k and $\hat{y}_k$ diverge exponentially until they lose all memory about their initial conditions ($\boldsymbol{x}_0$ and $\hat{\boldsymbol{x}}_0$, respectively). Given that such divergence is exponential, we therefore take the *median* of the NSE when comparing the performance over an ensemble of experiments as a better alternative to the mean.

On the same note, another way to validate the inferred microstate is by looking for how long our predictions accurately describe the system. Chaotic systems are by definition sensitive to initial conditions, meaning that microstates that are close to each other diverge exponentially over time. If these microstates live in a chaotic attractor, the distance between them is

bounded by the size of the attractor [13]. Thus, we can assess the quality of our predictions by measuring how long the real and simulated observations retain memory about each other [10]. We refer to this limiting time as the *predictability horizon* of the system k_{max}, and we define it as the average number of steps before the separation between y_k and $\hat{y}_k$ is greater than the distance between two random points in the attractor of the system. Mathematically,

$$k_{max} := \mathbb{E}\Big[\arg\min_{k\geq 0}\left\{(y_k - \hat{y}_k)^2 \geq \mathcal{D}_{\mathcal{M}}\right\}\Big],$$

where $\mathcal{D}_{\mathcal{M}}$ is average squared distance between two random points in the attractor $\mathcal{M}$ normalized by the variance of the attractor. More specifically, if X, Y are two i.i.d. random variables such that $X \sim \mu(\mathcal{M})$ with $\mu(\cdot)$ denoting the natural measure, then

$$\mathcal{D}_{\mathcal{M}} := \frac{\mathbb{E}[\|X - Y\|^2]}{Var(X)} = 2.$$

Thus, given that $\mathbb{E}[\sigma_y] = Var(X)$, we can simplify k_{max} into an expression that only depends on Eq. (12) so that

$$k_{max} = \mathbb{E}\Big[\arg\min_{k\geq 0}\left\{\mathrm{NSE}_k \geq 2\right\}\Big], \tag{14}$$

which is the formula we use to compute k_{max}.

Finally, we benchmark the inferred microstate by comparing k_{max} with a measure of the natural rate of divergence of the dynamics. As a measure of this, we use the *Lyapunov tenfold time* t_λ, which indicates the average time for two neighboring microstates to diverge from each other by one order of magnitude [37]. This is defined in terms of the inverse of the maximum Lyapunov exponent λ as

$$t_\lambda = \frac{\ln 10}{m\Delta t}\lambda^{-1}, \tag{15}$$

where the factor $\ln 10/(m\Delta t)$ lets us interpret t_λ in the units of the number of observations after which on average the dynamics causes the loss of an order of magnitude of precision. We obtain λ numerically using the two-particle method from Benettin *et al.* [38].

2.4 Initialization procedure summary

We summarize our *microstate initialization procedure* in the following steps:

1. **Preprocess:** Smooth the observed time series using the LMPA filter (see Eq. (8)) or any other suitable noise reduction technique. The smoothed signal will have fewer dominating trajectories [22] and a simpler distribution of local minima than the full noisy signal [23].
2. **Bound:** Make an arbitrary guess $\boldsymbol{x} \in \mathcal{X}$ of the microstate, and let it evolve under the model $\boldsymbol{f}$ until the microstate *roughly approaches* the smoothed observations, i.e., until $\mathcal{J}(\boldsymbol{x}^R_{-T}) \leq \delta_R$ for $\boldsymbol{x}^R_{-T} = \boldsymbol{f}^{\{mR\}}(\boldsymbol{x})$ for some $R \geq 0$. (see Eq. (10)). If several attractors exist, make one arbitrary guess for each of the different basins of attraction in the system.
3. **Refine:** Minimize the cost function $\mathcal{J}$ starting from $\mathbf{x}^R_{-T}$ using Adam gradient descent or any other suitable optimization scheme until $\mathcal{J}(\hat{\boldsymbol{x}}_{-T}) \leq \delta_r$ (see Eq. (11)) and call $\hat{\boldsymbol{x}}_{-T}$ the *assimilated microstate* and $\hat{\boldsymbol{x}}_0 = \boldsymbol{f}^{\{mT\}}(\hat{\boldsymbol{x}}_{-T})$ the *initialized microstate*.
4. **Validate:** Compute the discrepancy between the real and simulated observations (see Eq. (12)) and the predictability horizon k_{max} (see Eq. (14)) on out-of-sample predictions of the system to evaluate the quality of the initialized microstate $\hat{\boldsymbol{x}}_0$. If possible, benchmark the predictability horizon with the Lyapunov tenfold time (see Eq. (15)) of the system considered.

3 Results

In this section, we test the microstate initialization procedure on two paradigmatic chaotic systems: the well-known Lorenz system [21] and the high-dimensional Mackey–Glass system [39]. We approximate both systems using numerical integrators (described in each section that follows), and we take the approximated system as the real dynamical system.

In all cases, we sample the ground-truth microstate $\boldsymbol{x}_{-T}$ from the attractor of the system considered. We generate observations using the following nonlinear observation operator

$$\mathcal{H}(\boldsymbol{x}) = \sqrt[3]{\sum_{i=1}^{N_x}(\boldsymbol{x})_i^3}, \tag{16}$$

where $(\boldsymbol{x})_i$ is the i-th component of $\boldsymbol{x}$. Note that, while $\mathcal{H}$ is nonlinear, the mappings $x \to x^3$ and $x \to \sqrt[3]{x}$ are bijective, so there exists a diffeomorphism between $\mathcal{H}$ and any non-degenerate linear operator $\boldsymbol{H} : \mathcal{X} \subset \mathbb{R}^{N_x} \to \mathbb{R}$. Additionally to this operator, in Appendix B we study the quality of our ini-

Table 1 Parameters and other quantities

	α_R	α_r	β_R	β_r	T	N_x	m	σ_n/σ_y	q	r_0	t_λ
Lorenz	0.05	10^{-4}	0.5	$0.8r_0^{-2}$	50	3	2	0.3	4	2.02	127
Mackey–Glass	0.05	10^{-5}	0.5	$0.2r_0^{-2}$	25	50	2	0.3	5	2.41	230

α_R, α_r, β_R, and β_r are the parameters of our procedure. T is the number of data points in the time series $\boldsymbol{y}$, N_x is the dimension of the microstate space, m is the sampling interval between observations, q is the number of times we feed the signal back into the LPMA filter, r_0 is the increase in the signal-to-noise ratio, σ_n/σ_y is the noise level, and t_λ is the Lyapunov tenfold time of the system

tialization procedure with several observations operators each with different levels of coupling between the microstate components. The predictions in the observation space are almost identical regardless of what observation operator we use. However, given the symmetry about the x-axis of the Lorenz system, the operator with the maximum coupling recovers the right x component, but a reflection of the ground truth of the y and z components.

We test our initialization procedure on both noiseless time series and noisy time series. For the noisy series, we take a zero-mean Gaussian noise distribution $\epsilon_k \sim \mathcal{N}(0, \sigma_n^2)$ for all k. Here, σ_n represents the noise level of the observations, which we take to be 30% of the standard deviation of the observed data; i.e., $\sigma_n = 0.3\sigma_y$. We always construct the noiseless and noisy time series from the same ground-truth microstate so that all our results are comparable.

Although our choice for the initial guess is arbitrary, we initialize our method with a microstate that matches the first observation of $\boldsymbol{y}$ exactly. This is, we take our initial guess at random from the set $\{\boldsymbol{x} \in \mathcal{X} : \mathcal{H}(\boldsymbol{x}) = y_{-T}\}$. This is straightforward to do for any homogeneous function, such as the one of Eq. (16).

Throughout the results, we use the parameters and system features described in Table 1 unless otherwise stated. However, we evaluate the performance of our method for various choices of the rough parameter δ_R (see Fig. 16 in Appendix C) and the optimizers described in Sect. 2.2.4 (See Figs. 10 and 11 in Appendix C). Varying δ_R determines the effect of bounding the search space into finer-grained regions of the attractor, and we find that when observations are noiseless, the lower δ_R the better the initialization but the longer it takes to meet condition (10). For noisy observations, we find that if we set δ_R very small, the refinement stage yields no improvement over the initialized microstates obtained. In terms of the optimization schemes of the refinement stage, we find that Adadelta [32] and the various flavors of Adam [34–36] outperform all other alternatives, with significantly better results than vanilla gradient descent.

3.1 Low-dimensional example: Lorenz system

We first test our microstate initialization procedure on the Lorenz system [21], described by the following differential equations

$$\begin{aligned} \dot{x} &= \sigma(y - x)\,, \\ \dot{y} &= x(\rho - z) - y\,, \\ \dot{z} &= xy - \beta z\,, \end{aligned} \tag{17}$$

where $\boldsymbol{x}_k = \big(x(t_k), y(t_k), z(t_k)\big)$ is the microstate of the system at time t_k. The dynamics exhibit chaotic behavior for the parameters $\sigma = 28$, $\rho = 10$, and $\beta = 8/3$. We solve the system using a 4-th order Runge–Kutta integrator with a fixed step size of 0.01 units. Under these settings, the system's Lyapunov tenfold time is $t_\lambda = 127$ samples while its attractor has a Kaplan–Yorke dimension of $N_{\mathcal{M}} = 2.06$.

We perform 1000 independent experiments. For each experiment, we sample ground-truth microstates at random from the attractor of the system and generate time series of $T = 50$ observations with a sampling interval of $m = 2$ time steps per observation. We initialize the microstate for each time series following the steps presented in Sect. 2.4 using the parameters summarized in Table 1. We present our results in Fig. 1, where we show the median assimilation ($k < 0$) and prediction errors ($k \geq 0$) of the noiseless and noisy time series for both the model and observation spaces (see Eqs. (12)-(13)).

From the assimilation side, our estimations get progressively better the closer they are to the present time at $k = 0$. This means that the longer the time series—and

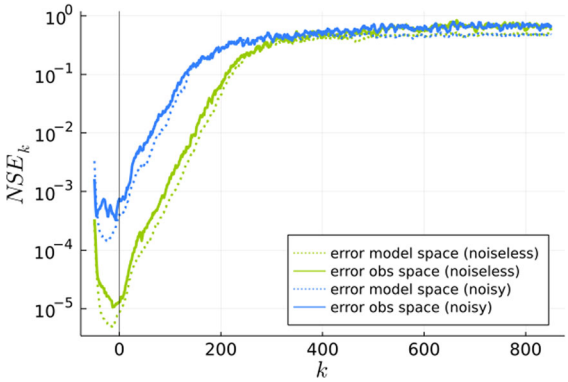


Fig. 1 Prediction error for the Lorenz system, showing the median normalized squared error over 1000 experiments for the observation space (solid lines) and the model space (dotted lines) for the case of noiseless (green) and noisy (blue) observations. The solid vertical line separates the assimilative regime ($k < 0$) from the predictive regime ($k \geq 0$)

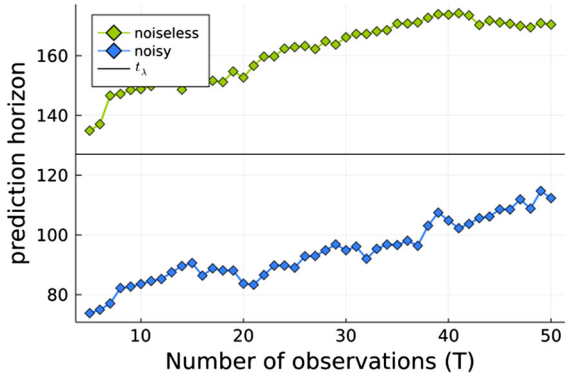


Fig. 2 Predictability vs. number of observations. We show how the predictability horizon k_{max} for the Lorenz system changes with the number of observations T for noiseless (green) and noisy (blue) ensembles of time series. The horizontal black solid line indicates the Lyapunov tenfold time t_λ

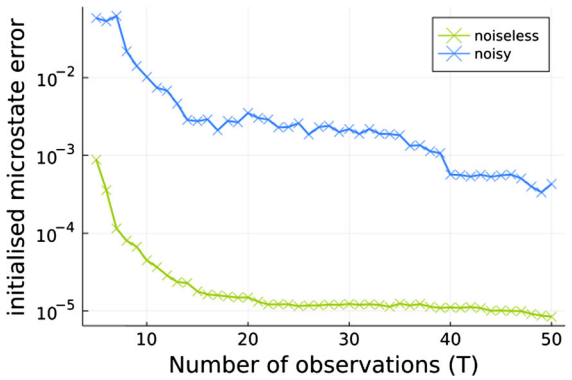


Fig. 3 Initialized microstate error for the Lorenz system. We show how the average discrepancy NSE_0^{model} between the true present-time microstate $\boldsymbol{x}_0$ and the initialized microstate $\hat{\boldsymbol{x}}_0$ changes with the number of observations T for noiseless (green) and noisy (blue) ensembles of time series

the more information we have from the *past*—the better the quality of the initialized microstate. Note that in the noisy case, the assimilative error plateaus near 10^{-3} in the observation space, marking the noise level of the observations. In contrast, the estimations keep getting better in the model space, indicating that even in the presence of noise, having more observations mitigates the probability of having dominating trajectories.

From the prediction side, we observe that the error diverges at essentially the same rate in both the noiseless and noisy cases. The main difference is that the error intercept at $k = 0$ is higher in the noisy case, thus making the predictions saturate earlier than when the observations are noiseless. Specifically, we find prediction horizons of $k_{max} = 171$ and $k_{max}^{noisy} = 113$ steps, which correspond to $1.35t_\lambda$ vs. $0.89t_\lambda$ for the noiseless and noisy time series (see Fig. 13 in Appendix C for an alternative approach on the prediction horizons). Moreover, we find that the prediction errors on the model and observation spaces are almost identical throughout the whole prediction window. Thus, measuring how the errors diverge in the observation space gives us a good proxy of the out-of-sample behavior of the latent microstates of the system.

In Figs. 2 and 3, we analyze how the performance of our method depends on the length of the observation window, showing how the prediction horizons and the discrepancies between $\boldsymbol{x}_0$ and $\hat{\boldsymbol{x}}_0$ change with the number of observations.

In Fig. 2, we find that the prediction horizon increases linearly with the number of observations available, with a similar slope for both the noiseless and noisy cases. Not surprisingly, the noise affects the time horizon over which one can make an effective prediction.

Additionally, we observe in Fig. 3 that the discrepancy decreases monotonically in both the noiseless and noisy cases. For the noiseless case, we observe a higher than exponential decrease in the discrepancy that ranges from $\text{NSE}_0^{model} \sim 10^{-3}$ for $T = 5$ to $\text{NSE}_0^{model} \sim 10^{-5}$ for $T = 50$ observations. While the change in discrepancy is less pronounced for the noisy time series, it decreases 2 orders of magnitude with

$\text{NSE}_0^{model} \sim 10^{-1.5}$ for $T = 5$ to $\text{NSE}_0^{model} \sim 10^{-3.5}$ for $T = 50$ observations.

The Lorenz equations are a low-dimensional system ($N_x = 3$) with a low dimensional attractor of dimension $N_{\mathcal{M}} = 2.06$. The number of observations in our experiments can be much larger than the dimension of the system. When combined with the fact that this system does not have any severe degeneracies, (see Appendix A), we recover the ground-truth microstate precisely with only a handful of noiseless observations. Every additional observation is, in theory, redundant for finding $\boldsymbol{x}_{-T}$ but, in practice, measurements can have several sources of error such as observational noise or the finite precision of numerical integration methods. Each additional observation thus further averages out these errors, which probably why k_{max} gets better proportionally to T.

3.2 High-dimensional system: Mackey–Glass

The Mackey–Glass system [39] describes the dynamics of the density of cells in the blood with the following delayed differential equation

$$\dot{x} = \mathscr{F}\left(x, x_{t_d}\right) = \frac{a x_{t_d}}{1 + x_{t_d}^c} - bx. \tag{18}$$

The state $x_{t_d} = x(t - t_d)$ is the density of cells delayed by t_d time units and a, b, and c are parameters. It exhibits chaotic dynamics for $t_d > 16.8$ with $a = 0.2$, $b = 0.1$ and $c = 10$ [40]. In terms of blood cell density, the chaotic regime represents a pathological behavior.

The evolution of Eq. (18) relies on knowing the state of x in the continuous interval $[t - t_d, t]$, making its state space infinite-dimensional. However, we can approximate such state by taking N_x samples at intervals of length $\Delta t = t_d/N_x$ and constructing the N_x-dimensional microstate vector

$$\begin{aligned}\boldsymbol{x}_k &= \left((\boldsymbol{x}_k)_1, \cdots, (\boldsymbol{x}_k)_{N_x}\right) \\ &= \Big(x(t_k - \underbrace{N_x \Delta t}_{t_d}), \ldots, x(t_k - \Delta t), x(t_k)\Big),\end{aligned} \tag{19}$$

where $(\boldsymbol{x}_i)_k = x(t - (N_x - i)\Delta t)$.

Using this vector, we can obtain trajectories of the Mackey–Glass system with any numerical integrator, for which we use the Euler method with a fixed step size of Δt for simplicity. We can thus recast this approximate system with the following N_x-dimensional deterministic mapping

$$\begin{aligned}\boldsymbol{x}_{k+1} &= \boldsymbol{f}(\boldsymbol{x}_k) \\ &= \begin{cases} (\boldsymbol{x}_t)_{N_x} + \Delta t \mathscr{F}\left((\boldsymbol{x}_t)_{N_x}, (\boldsymbol{x}_t)_1\right) \\ (\boldsymbol{x}_{t+1})_1 + \Delta t \mathscr{F}\left((\boldsymbol{x}_{t+1})_1, (\boldsymbol{x}_t)_2\right) \\ \vdots \\ (\boldsymbol{x}_{t+1})_{N_x - 1} + \Delta t \mathscr{F}\left((\boldsymbol{x}_{t+1})_{N_x - 1}, (\boldsymbol{x}_t)_{N_x}\right), \end{cases}\end{aligned} \tag{20}$$

using $\mathscr{F}$ as defined in Eq. (18). The microstate-space dimension N_x of this mapping is determined by $t_d/\Delta t$, for which we take $N_x = 50$ and $t_d = 25$ so that the system exhibits chaotic dynamics. Under these settings, the system's Lyapunovs tenfold time is $t_\lambda = 230$ samples while its attractor has a Kaplan–Yorke dimension of $N_{\mathcal{M}} = 2.34$.

As before, we perform 1000 independent experiments in which, for each experiment, we sample ground-truth microstates at random from the attractor of the system and generate time series of $T = 25$ observations with a sampling interval of $m = 2$ time steps per observation (see Table 1 for details). In contrast to the Lorenz system, we consider time series containing fewer data points than the dimension of the microstate space (in this case $T = 25$ and $N_x = 50$, respectively), making the problem under-determined. We present our results in Fig. 4, where we show the median assimilation ($k < 0$) and prediction errors ($k \geq 0$) for the noiseless and noisy time series for both the model and observation spaces.

Unexpectedly, our method yields more accurate initializations for the Mackey–Glass system than the Lorenz system, even though both the attractor and the microstate space of the former have a higher dimension than the latter. However, if we compare the power spectra of the two systems (see Fig. 14 in Appendix C), we find that the Mackey–Glass system has a faster frequency decay and more frequency peaks than the Lorenz system, suggesting that the former is easier to initialize than the latter. For instance, the power amplitude for frequency 1/6 is more than 1000 higher in the Lorenz system than in the Mackey–Glass system. This further suggests that looking at the power spectra of the system is a better indicator of the initializability of a system than the dimension of its chaotic attractor.

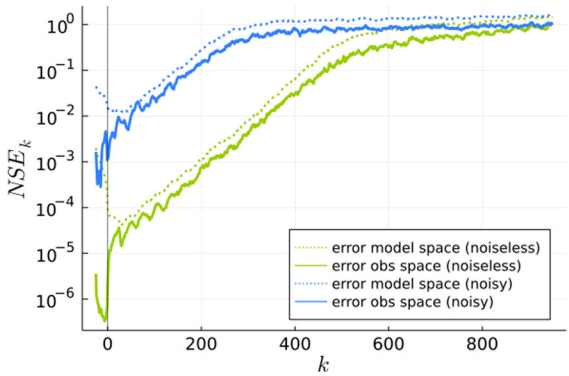


Fig. 4 Prediction error for the Mackey–Glass system. We show the median normalized squared error over 1000 experiments for the observation space (solid lines) and the model space (dotted lines) for the case of noiseless (green) and noisy (blue) observations. The solid vertical line separates the assimilative regime ($k < 0$) from the predictive regime ($k \geq 0$)

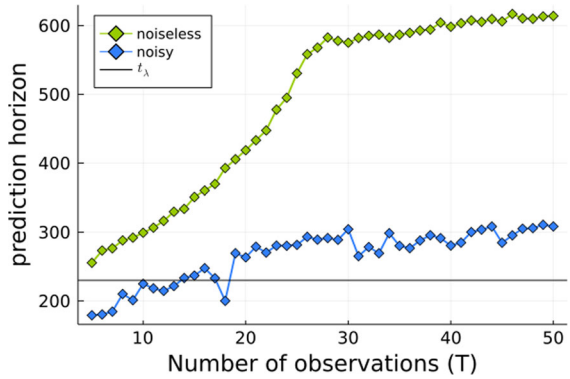


Fig. 5 Predictability horizon of the Mackey–Glass system. We show how the predictability horizon k_{max} changes with the number of observations T for noiseless (green) and noisy (blue) ensembles of time series. The horizontal black solid line indicates the Lyapunov tenfold time t_λ. For the noiseless case, we observe a critical transition on the behavior of k_{max} for $T_c = 25$

From the prediction side, our results are qualitatively similar to what we saw for the Lorenz system. The error diverges at roughly the same rate in both the noiseless and noisy time series, with an error intercept at $k = 0$ that is higher for the noisy experiments. Additionally, we find a very close correspondence between the errors in the model and observation spaces, which supports our claim that measuring the error in the observation space gives us a good proxy of the out-of-sample behavior of the latent microstate dynamics.

In terms of their predictability horizons, we find that $k_{max} = 556$ and $k_{max}^{noisy} = 285$, corresponding to $2.4t_\lambda$ and $1.2t_\lambda$ for the noiseless and noisy ensembles, respectively, (see Fig. 13 in Appendix C for an alternative approach on the prediction horizons). We find it remarkable that with only 25 observations of the system, we obtain predictions that stay accurate for significantly longer than the Lyapunov tenfold time of the system.

From the assimilation side, the microstate estimations get progressively better the closer they are to the present time, similar to what we observed in the Lorenz system. However, the assimilation error is significantly lower in the observation space than in the model space, suggesting, misleadingly, that the initialized microstate is much more accurate than the error we observe in the model space. Nonetheless, the errors in model and observation spaces converge to each other as soon as the prediction window starts, meaning that the error in the observation space is still an accurate proxy of the out-of-sample behavior in the model space.

Interestingly, the first few out-of-sample predictions in the model space have a lower discrepancy than in the observation space in both the noiseless and noisy cases. This happens, we believe, because the time span of the observations $\boldsymbol{y}$ is not long enough for the initialized microstate to converge onto the attractor. With only $T = 25$ data points of a 50-dimensional chaotic system, we do not possess enough information to recover the present-time microstate precisely.

The previous discussion suggests that we need more observations to better initialize the system. To investigate this we perform a series of experiments in which we vary the number of data points of the observed time series and assess the quality of the predictions. Similar to what we did for the Lorenz system, we focus on the prediction horizon (see Fig. 5) and the discrepancy between the present-time and the initialized microstate (see Fig. 6).

We find a contrasting behavior regarding the experiments between noisy and noiseless observations. When the observations are noisy, the prediction horizon increases linearly with the number of observations (see Fig. 5 blue), ranging from $k_{max} = 0.78t_\lambda$ when $T = 5$ to $k_{max} = 1.34t_\lambda$ when $T = 50$. We also find that, in general, the discrepancy between the initialized and ground-truth present microstate decreases monotonically with the number of observations (see Fig. 6 bottom). These results are qualitatively similar to what we

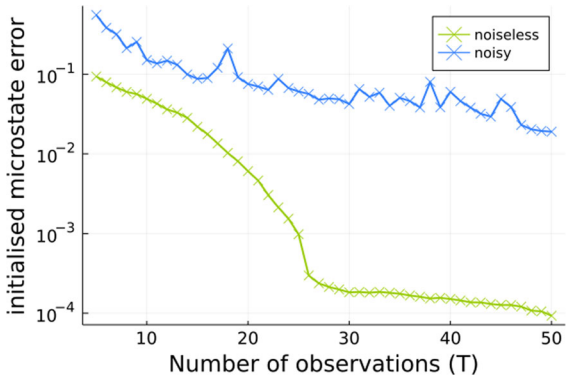


Fig. 6 Initialized model error for the Mackey–Glass system. We show how NSE_0^{model}, the average discrepancy between the true present-time microstate $\boldsymbol{x}_0$ and the initialized microstate $\hat{\boldsymbol{x}}_0$, changes with the number of observations T for noiseless (green) and noisy (blue) ensembles of time series. For the noiseless case, we observe a critical transition in the behavior of NSE_0^{model} for $T_c = 25$

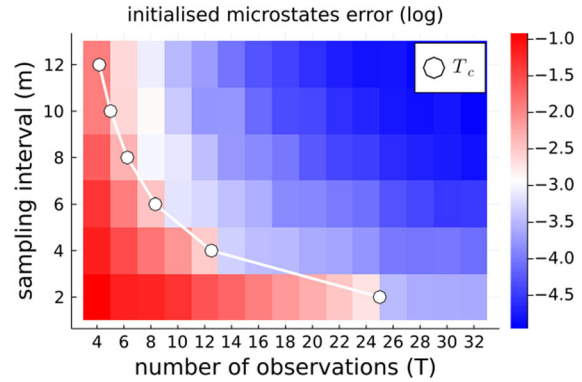


Fig. 7 Critical transition heatmap of the Mackey–Glass system: In the z-axis, we show the (base-10 logarithm of the) initialized microstate discrepancy, NSE_0^{model}, as a function of the number of observations T and the sampling interval m for ensembles of noiseless time series. In white, we plot the $m = N_x/T$ curve for fixed $N_x = 50$. We find that the microstate discrepancies decrease abruptly before and after this curve

found for the Lorenz system: the more observations the better.

When the observations are noiseless, we find a *critical change of behavior* at roughly $T = 25$ observations. In Fig. 5 (green), we find that the prediction horizon rises superlinearly for $T < 25$, with $k_{max} = 1.11t_\lambda$ for $T = 5$ to $k_{max} = 2.31t_\lambda$ for $T = 25$. Afterward, the prediction horizon grows linearly and with a marginal increase, getting to $k_{max} = 2.67t_\lambda$ for $T = 50$. We note, however, that the increase in this linear regime is almost double of what we find in the noisy counterpart, with a slope of $\Delta k_{max} = 15.2$ steps per observation against $\Delta k_{max}^{noisy} = 8.3$, respectively. In parallel, we find that the discrepancy of the initialized microstate decreases abruptly for the time series of $T \geq 25$ observations, as we show in Fig. 6 (green). This sharp transition reflects that time series with more than 25 observations have enough information to pin down the present-time latent microstate precisely, meaning that we possess enough data points to uniquely separate the mixing of the microstate generated by the observation operator $\mathcal{H}$ into its individual components.

In short, having 25 (or more) noiseless measurements of the system gives us enough information to precisely recover the present-time microstate, which is 50-dimensional. Recall that we are considering the discrete map (20) as the real system, so the only information we lose comes from either measuring the system with $\mathcal{H}$ or taking time series with a coarse-grained sampling frequency. Thus, for a fixed $\mathcal{H}$ and noiseless observations, recovering the initial microstates precisely should depend solely on how well the samples describe the latent trajectory of the system. Inspired by the Nyquist–Shannon sampling theorem [41], if we can establish a clear cutoff frequency on the power spectrum of the system, we could argue that if we observe the system with twice the frequency as the systems's cutoff frequency, then the observed signal would not lose any information with respect to the signal sampled for every update of map (20).

We thus claim that, if we can establish a clear cutoff frequency f_c and the dynamical system does not suffer from severe degeneracies, we can precisely obtain the initial conditions of an N_x-dimensional (possibly) nonlinear system observed every m updates with a scalar (possibly) nonlinear observation operator $\mathcal{H}$ if 1) $m^{-1} \geq 2f_c$ and 2) $T_c \geq N_x/m$. Having these conditions satisfied is equivalent to having an invertible observation-matrix $\boldsymbol{M}$ that determines the solution of the system $\boldsymbol{y} = \boldsymbol{M}\boldsymbol{x}_0$ exactly (see Appendix A for a deeper development of this discussion).

We make further experiments to check the validity of our claim, where we measure the initialized microstate discrepancy when varying both the number of observations T and the sampling interval m while leaving the dimension of the system fixed to $N_x = 50$. If our claim about the Nyquist–Shannon theorem is a good approximation, we expect to find a $T_c \sim 1/m$ relation

where, for $T \geq T_c$, the error in the initialized microstate becomes significantly lower than for $T < T_c$. In Fig. 7, we show the results of these experiments, in which we observe such a change of regimes before and after the $T = N_x/m$ line, thus supporting the Nyquist–Shannon hypothesis.

4 Conclusions

Many natural and social processes can be accurately described by how their microstates interact and evolve into rich non-trivial dynamics with emerging properties. We often only possess aggregate noisy measurements of such processes, so it is of great interest to develop methods that let us extract information about the latent microstate dynamics from a given dataset.

In this paper, we tackled the problem of initializing the latent microstate of a known (possibly) nonlinear dynamical system from aggregate (possibly) noisy and nonlinear observations of the system. We propose a three-step method to obtain such latent microstate that consists of (1) reducing the observational noise to mitigate possible dominating trajectories, (2) letting the system explore its attractor(s) and thus limiting the region in which we search for an optimal solution, and (3) minimizing the discrepancy between the simulated and real observations to obtain a refined estimation of the ground-truth microstate. We quantified the discrepancy between observations and simulations using a least-squares cost function in the observation space, similar to [13,14]. We minimized the cost function using a plethora of gradient-based algorithms, for which we find that Adadelta [32] and Adam-oriented schemes [34–36] perform the best.

We tested our method on two chaotic paradigmatic examples: the Lorenz and a high-dimensional approximation of the Mackey–Glass systems. We obtained initialized microstates that accurate fit the data, with out-of-sample predictions that outperformed the systems' Lyapunov tenfold times, even when the observed time series were very short. We found that good predictions in the observations space always implied good predictions in the space of the microstates. We considered nonlinear observation operators that aggregate all the microstate component into a real number in all cases, with robust result with all the operators considered. Surprisingly, we obtained better results for the Mackey–Glass system, which has a higher-dimensional model space and higher-dimensional attractor but faster-decaying frequency spectrum than the Lorenz system, suggesting that the frequency spectrum gives us a better proxy of the initiability of the system than the observations to dimension of the model ratio.

In most experiments, the quality of the initialized microstate was proportional to the number of data points of the observed time series. However, when the dimension of the system was higher than the number of observations and these observations were noiseless, the quality of the initialized microstate grew superlinearly. This superlinear regime transitions into the more common linear regime in a non-trivial manner, and we explored the conditions for such a transition. We claim that as long as we can establish a clear cutoff frequency of the observed data and this data meets the Nyquist–Shannon sampling theorem conditions with respect to the cutoff frequency, we can recover the ground-truth microstate precisely with fewer observations than the dimension of the system, thus marking the transition between regimes. This implies that if we possess a dataset where observations are sampled at an optimal rate such that we lose the least possible information of the underlying system, we can obtain high-quality initializations with just a handful of samples.

How well we can initialize a system depends on the amount of information the observed data contains and on the intrinsic features of the system. On the one hand, the amount of observational noise, the mixing that results from aggregating the system, the number of observations, and the data sampling rate contribute significantly in estimating microstates that may fit the data well but extrapolate poorly into the future. On the other hand, the dimension of the dynamical system, the frequency spectrum of its attractor(s), and how chaotic the system is, determines the window in which the predictions stay accurate.

This work gives us a conceptual framework to understand the interface between aggregate data and microscopic interactions. However, it is limited the case in which we know the model that perfectly specifies the system, as well as the observation operator and the characteristics of the noise in the observations.

In future developments, we will explore how to deal with misspecified models and systems with stochastic behavior. We should include these new sources of uncertainty in the cost functions involved. With stochastic behavior alone, there are several new considerations for inferring individual agents' evolution in

a system [42]. Once we better understand what information low-dimensional observations gives us about a high-dimensional dynamical system, the natural next step is test our initialization method on real-world datasets.

Acknowledgements The authors are grateful with Maria del Rio-Chanona, Marco Pangallo and Sylvain Barde for stimulating discussions, as well as the Oxford INET Complexity Economics group for valuable feedback.

Funding BK has received research support from the Conacyt-SENER: Sustentabilidad Energética scholarship.

Code availability All the code described in the methodology is coded in Julia and is freely available at https://github.com/blas-ko/LatentStateInitialization.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendices

A noiseless non-autonomous linear systems

In this section, we assume that the dynamical system (1), the observations (2), and the corresponding dynamical mapping (3) are given by time-varying linear functions. We further assume that the dynamics are deterministic and the observations are noiseless so that $\boldsymbol{\xi}_k = 0$ and $\epsilon_k = 0$ for all k. Making these assumptions provides us with two advantages over arbitrary nonlinear dynamical systems: (1) linear systems are much easier to handle than nonlinear systems, and (2) nonlinear systems can be approximated by time-varying systems arbitrarily well if they are locally Lipchitz [43].

Additionally, we find it convenient to slightly change the notation introduced in the main text. In the main text, we indexed the observations of the system so that the time stamps describe the observations $\boldsymbol{y}$ in the most natural way. Thus, we defined t_k such that $y_k = y(t_k)$ is the k-th data point of the series. Consequently, we indexed the evolution of the microstates as $\boldsymbol{x}(t_{k+1}) = \boldsymbol{x}(t_k + m\Delta t) = \boldsymbol{f}^{\{m\}}(\boldsymbol{x}(t_k))$, i.e., we needed to update the system m times before sampling the next observation. Here, we index the passing of time in the time scale of the microstates, so that $\boldsymbol{x}(t_{k+1}) = \boldsymbol{x}(t_k + \Delta t) = \boldsymbol{f}(\boldsymbol{x}(t_k))$. Thus, we label the observed time series as $\boldsymbol{y} = (y_0, y_{m-1}, \ldots, y_{mT-1})$ so that y_{mk-1} is the k-th data point of a time series of T observations. This approach emphasizes that $\boldsymbol{y}$ is a coarse-grained sample of the underlying dynamics of the microstates.

Under the above considerations, we can recast Eqs. (1–2), respectively, as follows

$$\boldsymbol{x}_{k+1} = \boldsymbol{F}_k \boldsymbol{x}_k = (\boldsymbol{F}_k \boldsymbol{F}_{k-1} \cdots \boldsymbol{F}_0)\boldsymbol{x}_0 \,, \tag{21}$$

$$y_k = \boldsymbol{H}_k \boldsymbol{x}_k \,, \tag{22}$$

where, at every time t_k, $\boldsymbol{F}_k$ is an $N_x \times N_x$ matrix representing a linear dynamical process and $\boldsymbol{H}_k$ is an $1 \times N_x$ matrix representing a linear observation operator. We assume that every element of $\boldsymbol{H}_k$ is nonzero for every k. Note that under the current notation, the sequence $y_0, y_1, \ldots, y_k, \ldots$ represents the ground-truth dynamics under the (time-varying) observation operator $\boldsymbol{H}_k$, and only those indexes k that are a multiple of m are included in the observations $\boldsymbol{y}$.

From the RHS of Eq. (21), we see the explicit dependence of any observation y_k from the initial conditions $\boldsymbol{x}_0$, so we can define an $1 \times N_x$ matrix[3]

$$\boldsymbol{M}_k := \boldsymbol{H}_k \boldsymbol{F}_{k-1} \cdots \boldsymbol{F}_0$$

that takes us from $\boldsymbol{x}_0$ to y_k for any k. Thus, if we possess a time series of T observations such that the last observations happen at time t_{mT-1}, then we can recast the microstate initialization problem as the following linear

[3] We define matrix $\boldsymbol{M}$ for time t_0 as $\boldsymbol{M}_0 := \boldsymbol{H}_0$.

system of equations of N_x variables and mT equations

$$\begin{aligned} y_0 &= \boldsymbol{M}_0 \boldsymbol{x}_0 \\ y_1 &= \boldsymbol{M}_1 \boldsymbol{x}_0 \\ &\vdots \\ y_{mT-1} &= \boldsymbol{M}_{mT-1} \boldsymbol{x}_0 \ , \end{aligned}$$

or, more compactly,

$$\boldsymbol{y}^* = \boldsymbol{M}^* \boldsymbol{x}_0, \tag{23}$$

where $\boldsymbol{y}^* = (y_0, y_1, \ldots, y_{mT-1}) \in \mathbb{R}^{mT}$ is the *extended* sequence of observations (it is extended in that it includes all the observations in $\boldsymbol{y} \in \mathbb{R}^T$ plus its intermediate, unobserved samples) and $\boldsymbol{M}^* = [\boldsymbol{M}_0 | \boldsymbol{M}_1 | \ldots | \boldsymbol{M}_{mT-1}]$, the *extended observation matrix* of size $mT \times N_x$.

We may solve system (23) exactly whenever $\boldsymbol{M}$ is invertible. If the matrices $\boldsymbol{M}_k$ are non-singular, then $\boldsymbol{M}^*$ becomes invertible when $mT = N_x$.

We do not possess $\boldsymbol{y}^*$ but the coarse-grained time series $\boldsymbol{y}$ of T observations, where two consecutive samples are m time steps apart. Thus, we may only express a reduced form of system (23) with T equations and N_x variables as

$$\boldsymbol{y} = \boldsymbol{M} \boldsymbol{x}_0, \tag{24}$$

with $\boldsymbol{M} = [\boldsymbol{M}_0 | \boldsymbol{M}_{m-1} | \ldots | \boldsymbol{M}_{mT-1}]$ of size $T \times N_x$. The system (24) spans the same time interval as system (23), but it has m less equations, so it is underdetermined and might have several solutions.

Under the conditions we describe in what follows, we may obtain a solution $\boldsymbol{x}$ of the reduced system (24) that is also a solution of the extended system (23). If $\boldsymbol{y}$ consists of $T = N_x/m$ (or more) data points, the solution $\boldsymbol{x}$ is unique and equal to the ground-truth microstate $\boldsymbol{x}_0$. More specifically, if the sampling frequency $(m\Delta t)^{-1}$ is higher than twice the cutoff frequency of the spectrum of the system and the matrices $\boldsymbol{M}_k$ are non-singular for $k \in \{0, m-1, \ldots, mT-1\}$, then the time series $\boldsymbol{y}$ determines the extended series $\boldsymbol{y}^*$ uniquely, and therefore any solution $\boldsymbol{x}$ that solves (24) also solves (23). The above conditions establish the necessary and sufficient conditions for the Nyquist–Shannon sampling theorem [41] to be true, so our result is a direct application of the theorem.

Note that the power spectra of the systems considered in this work (and most chaotic systems) exhibit a power-law decay on their power spectrum, so there is not a well-defined cutoff frequency on neither the Mackey–Glass nor the Lorenz systems. Nevertheless, if their power spectrum decays fast enough, we should be able to define an effective cutoff frequency for the previous arguments to be a good approximation. We will investigate the validity of our arguments in future iterations of this work.

In this paper, we chose non-degenerate dynamical systems that are Lipschitz continuous and chose an observation operator (see Eq. (16)) that is bijective to the linear operator $\boldsymbol{H} = [1, \ldots, 1]^\dagger$. Therefore, the results of this Appendix apply in all our systems whenever the observations are noiseless. In particular, the results of this Appendix explain the critical transition we show in Figs. 5 and 6 for $T = N_x/m = 25$.

B Comparison of several nonlinear observation operators

We assess the robustness of our initialization method using different (nonlinear) observation operators. The way we aggregate the microstates affects how much information we retain about the latent dynamics, so we consider the three following operators each with different levels of coupling between the microstate components

$$\mathcal{H}_1(\boldsymbol{x}) = \sqrt[3]{S_{\boldsymbol{x}}}, \tag{25}$$

$$\mathcal{H}_2(\boldsymbol{x}) = sign\,(P_{\boldsymbol{x}})\ \sqrt[N_x]{|P_{\boldsymbol{x}}|}, \tag{26}$$

$$\mathcal{H}_3(\boldsymbol{x}) = sign\,(C_{\boldsymbol{x}})\ \sqrt{|C_{\boldsymbol{x}}|}, \tag{27}$$

where

$$S_{\boldsymbol{x}} = \sum_{i=1}^{N_x} \boldsymbol{x}_i^3, \tag{28}$$

$$P_{\boldsymbol{x}} = \prod_{i=1}^{N_x} \boldsymbol{x}_i, \tag{29}$$

$$C_{\boldsymbol{x}} = \sum_{i<j} \boldsymbol{x}_i \boldsymbol{x}_j. \tag{30}$$

Our rationale for choosing these operators goes as follows. We point out that Eq. (25) is the same as Eq. (16): it is the cubic root of sum of cubes of the components of the microstates. This operator has no coupling between the microstate variables. As we mentioned before, Eq(25) is bijective to any non-degenerate

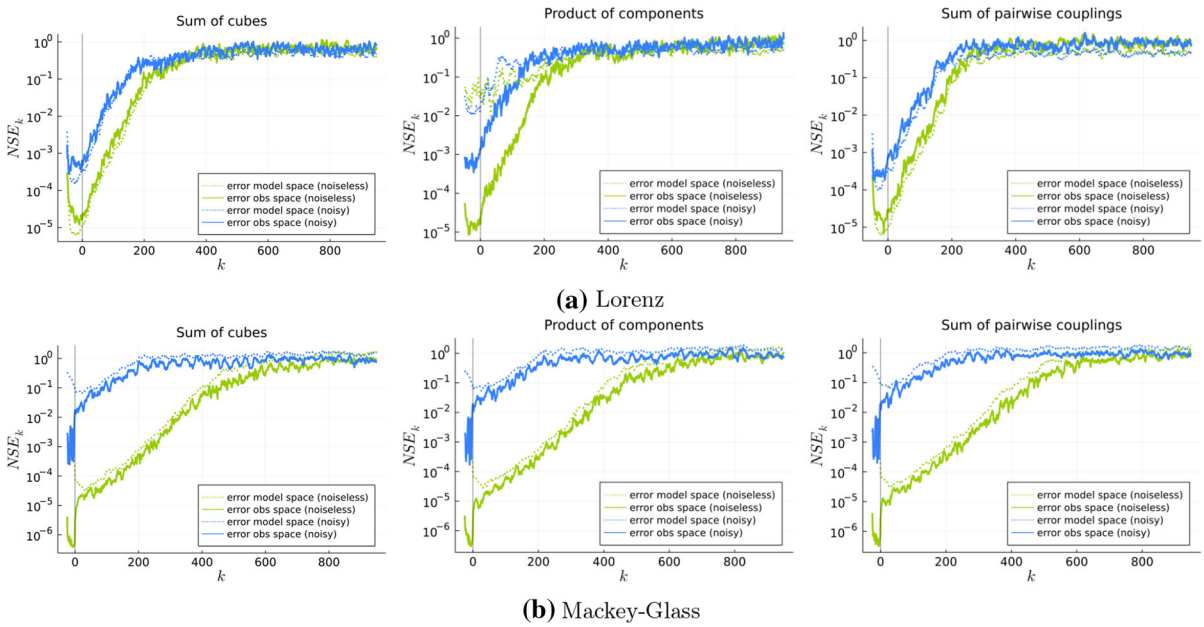


Fig. 8 Prediction error for different observation operators. We show the median normalized squared error over 200 experiments for the observation space (solid lines) and the model space (dotted lines) for the case of noiseless (green) and noisy (blue) observations for (**a**) the Lorenz system and (**b**) the Mackey–Glass system. From left to right, we show the behavior for operators representing the sum of cubes (see Eq. (25)), the product between microstate components (see Eq. (26)), and the sum of pairwise couplings (see Eq. (27)). We take all parameters as in Table 1

linear operator. For Eq. (26), we take an operator that couples all the microstate components by multiplying them. If all the microstate components are positive—i.e., if $\boldsymbol{x}_i > 0$ for all i—then we can take the logarithm of $\mathcal{H}_2$ and recover a non-degenerate linear operator. If any of the components is non-positive, then we cannot transform $\mathcal{H}_2$ to a linear operator, making the decoupling impossible. Finally, Eq. (27) is the sum of pairwise couplings between the microstate components. In this case, there is no smooth transformation between $\mathcal{H}_3$ and a linear operator, so we can make no further simplification of $\mathcal{H}_3$. We consider the cubic-, N_xth-, and square-root of $S_{\boldsymbol{x}}$, $P_{\boldsymbol{x}}$, and $C_{\boldsymbol{x}}$, respectively, so that the physical units of the observations are the same as the units of the microstates.

In Fig. 8, we show the median assimilation ($k < 0$) and prediction ($k \geq 0$) errors over 200 runs for the Lorenz and the Mackey–Glass systems for each of the observation operators described above. We took the same set of 200 initial conditions and noise seeds for each operator to make the results comparable.

In almost every case, we observe that the results for each system are almost identical regardless of the operator we use, both in the observation and the model spaces. This behavior suggests that our method is robust

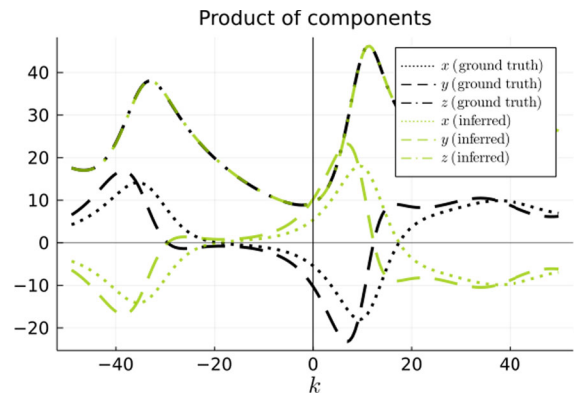


Fig. 9 Dynamics of the individual microstate components of the Lorenz system for an example run of the ground truth (black) and the initialized (green) microstate when the observation are taken using the operator of Eq. (26)—i.e., the operator that multiplies all the microstate components— in a noiseless scenario. The y and z components are symmetric with respect to the 0 line. Similar to Fig. 8, $k < 0$ denotes assimilation times and $k \geq 0$ prediction times

to the observation operator: if the observation convey enough information about the latent dynamics of the system, our method will initialize a microstate with good prediction power.

However, we observe an anomaly on the model space error for the Lorenz system: the model space error corresponding to $\mathcal{H}_2$—the product of components—is significantly higher than i) its observation space error, and ii) the model error corresponding to $\mathcal{H}_1$ and $\mathcal{H}_3$. Recall that the Lorenz system is symmetric around its x-axis, so the operator $\mathcal{H}_2$ cannot resolve if the product $\prod_i \boldsymbol{x}_i$ corresponds to (x, y, z) or to $(x, -y, -z)$, regardless of the number of measurements we have about the system. Thus, our method sometimes initializes the ground truth microstate and the other times it initializes its reflection about the x-axis, as we show in Fig. 9.

Summarizing, the feasible set of solutions for the Lorenz system observed through $\mathcal{H}_2$ include both (x, y, z) and $(x, -y, -z)$ which, incidentally, result in the same observation-space dynamics for time windows of arbitrary size. In this case, the initialization is more difficult because some microstates are indistinguishable. For all the other observation operators, the only feasible solution is the ground-truth microstate, hence their results are almost identical in 8. These results suggest, that it is possible to improve the initialization procedure in systems with symmetries. It would only be necessary to add an extra step to the initialization procedure. This extra step would involve the rotation of the microstate found across all the symmetries of the system. We would then continue refining in parallel all those different symmetric microstates. The global solution would be then the one with the lowest cost function. We leave a deep analysis of this scenario for a future study.

C Initialization method performance & system features

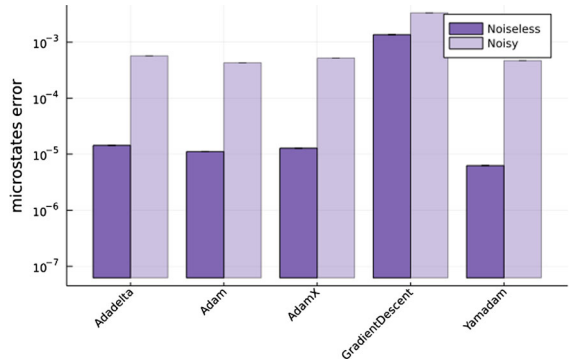


Fig. 10 Gradient-based optimizers performance on the Lorenz system. We measure the performance of the gradient-based optimizers (described in Sect. 2.2.4) with NSE_0^{model}, the average discrepancy between the present-time microstate $\boldsymbol{x}_0$ and the initialized microstate $\hat{\boldsymbol{x}}_0$, for noiseless (dark purple) and the noisy (light purple) time series. We show only the four best performing optimizers—namely Adadelta, Adam, AdamX, and YamAdam—as well as stochastic gradient descent, which serves as our benchmark

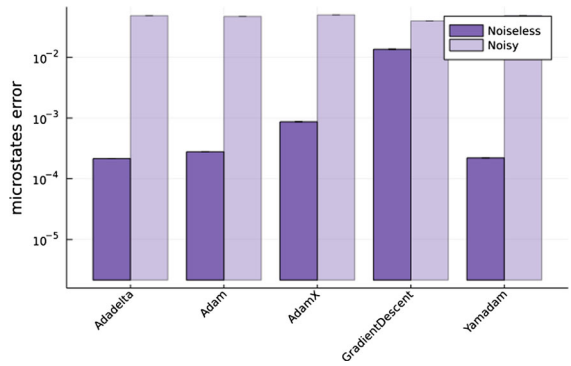
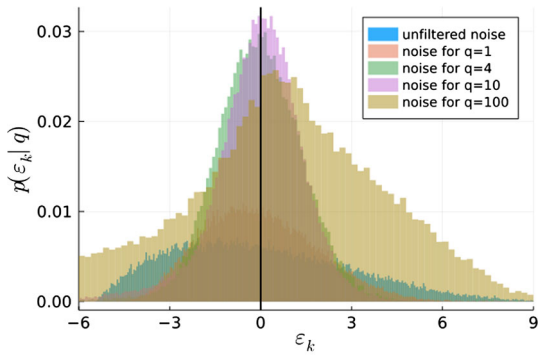
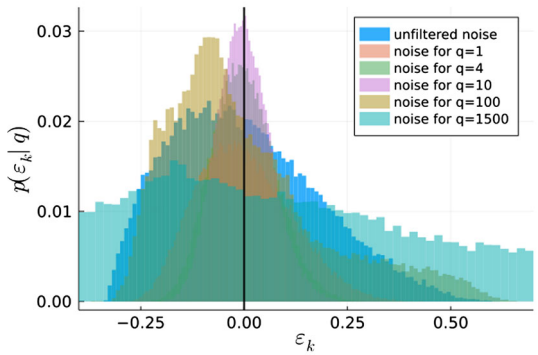


Fig. 11 Gradient-based optimizers performance on the Mackey–Glass system. See description in Fig. 10 for details. The best performing optimizers were the same as with the Lorenz system

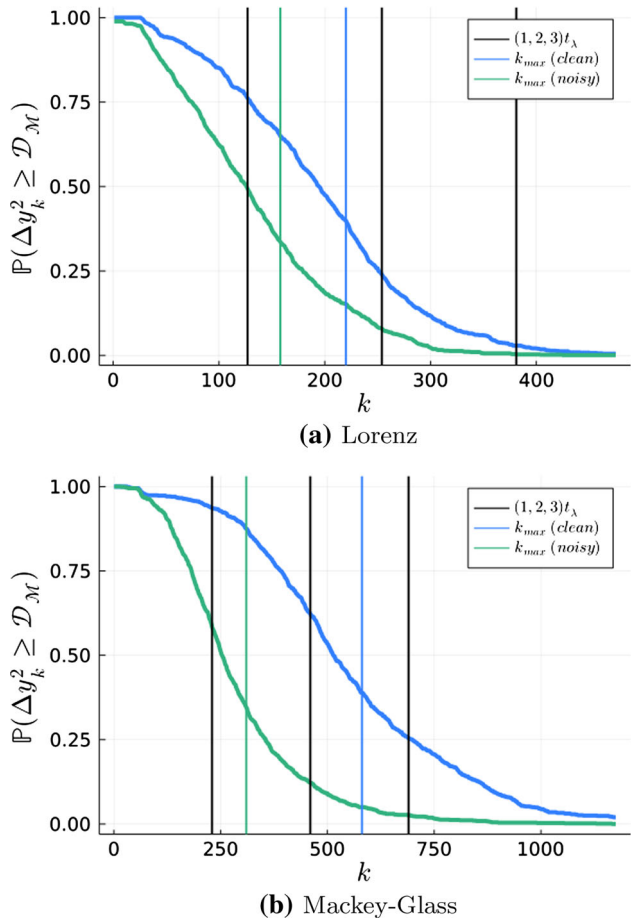


(a) Lorenz



(b) Mackey-Glass

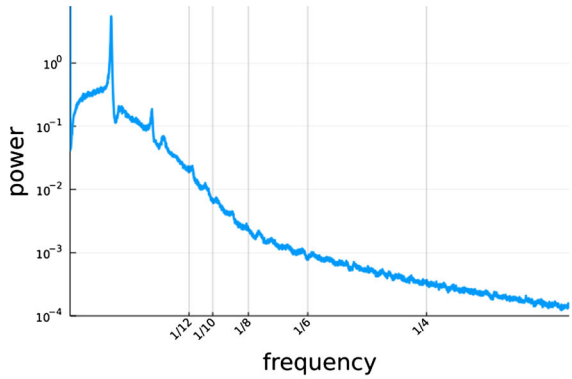
Fig. 12 Filtered noise distributions Starting from a simulated observational noise described by a left-skewed Beta distribution (see "unfiltered noise" in the legend) of very long time series ($T = 50000$ samples) of the *a)* Lorenz and *b)* Mackey–Glass systems, we plot the noise distribution of the unfiltered noise (solid blue) as well as the noise distribution we obtain after smoothing the corrupted signal with the LMPA filter for filters of increasing strength q. See Sect. 2.2.2 for details on the filter. For small q, the resulting distribution looks like a Gaussian distribution while for $q \gg 1$, the filter takes a significant of the signal, resulting in exhotic-shaped distributions



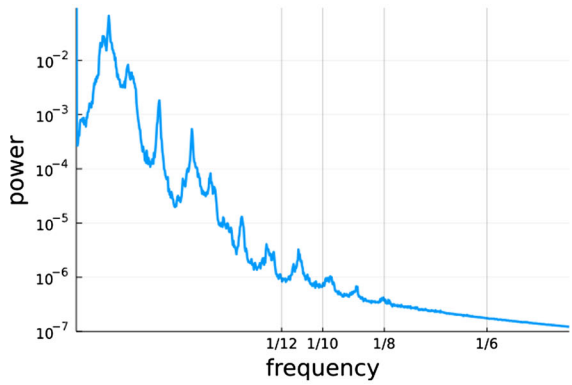
(a) Lorenz

(b) Mackey-Glass

Fig. 13 Cumulative probability of divergence for *a)* the Lorenz system and *b)* the Mackey–Glass system. We show the fraction of trajectories—out of 1000—for which $NSE_k^{obs} \geq 2$ as a function of the prediction step k. This approach generalizes our definition of prediction horizon of Eq. (14) into a distribution-like quantity. In solid vertical lines, we show the Lyapunov tenfold time of the system as well as the prediction horizons k_{max} for the noiseless and noisy cases

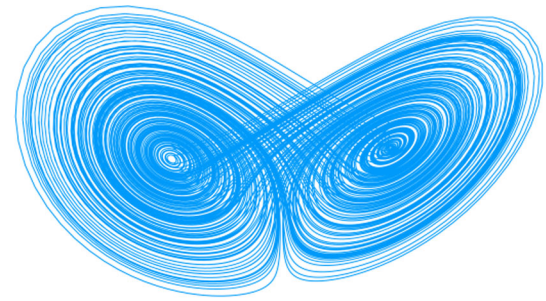


(a) Lorenz

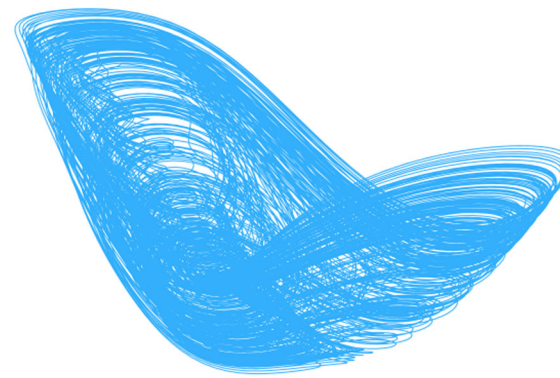


(b) Mackey-Glass

Fig. 14 Normalized power spectra for **a** the Lorenz system and **b** the Mackey–Glass system. The plot shows the average power spectra over 100 random in-attractor initial conditions with trajectories of 2^{12} points. We note that the Lorenz system has a clear power-law frequency decay with only a few low frequency peaks. While the Mackey–Glass system exhibits a nonvanishing spectrum characteristic of chaotic systems, it has more defined frequency peaks and decays much faster than the Lorenz system. Thus, we expect for the Mackey–Glass system to be easier to initialize in general

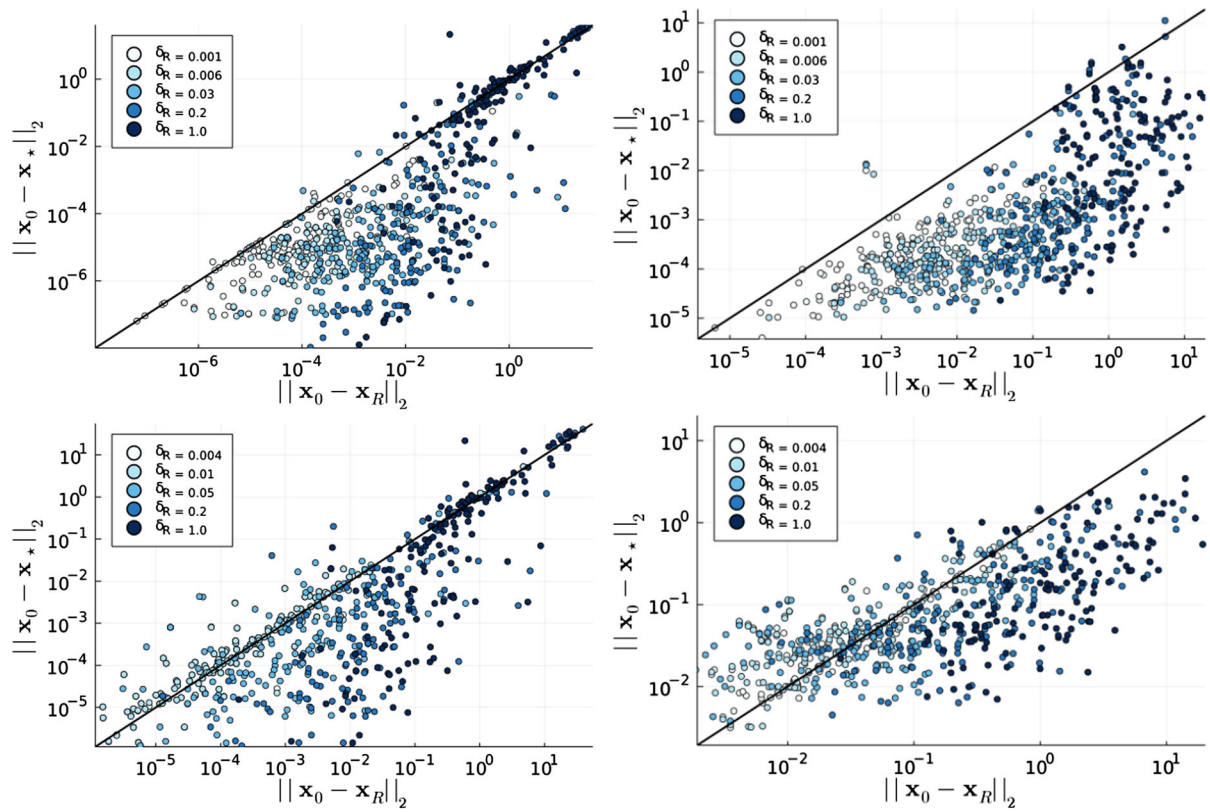


(a) Lorenz



(b) Mackey-Glass

Fig. 15 Chaotic attractors for **a** the state space portrait of the Lorenz system, where the axis show x, y and z, respectively, and **b** the state space portrait of the Mackey–Glass system, where we show $x(t)$ against $x(t - t_d)$. Each attractor consists of a *long* trajectory of 15000 points coming from a random in-attractor initial condition



(a) Lorenz **(b)** Mackey-Glass

Fig. 16 Initialization performance for different bounding-stage parameters. We present, on the x axis, the discrepancy of $\hat{\boldsymbol{x}}_0^R$ against, on the y axis, the discrepancy of the initialized microstates $\hat{\boldsymbol{x}}_0$ for increasing levels of δ_R (see legend) on **a** the Lorenz system and **b** the Mackey–Glass system. The microstate $\hat{\boldsymbol{x}}_0^R$ is the *rough* estimation of $\boldsymbol{x}_0$ after Eq. (10) is zatisfied. Values below the identity mean that $\hat{\boldsymbol{x}}_0^R$ improves refining it as described in Sect. 2.2.4. Values on the diagonal mean that the initialized microstates do not get any better by applying refinement methods

References

1. Rind, D.: Complexity and climate. Science **284**, 105–107 (1999)
2. Ward, J.A., Evans, A.J., Malleson, N.S.: Dynamic calibration of agent-based models using data assimilation. R. Soc. Open Sci. **3**(4), 150703 (2016)
3. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional system. Nature **10**, 186–198 (2009)
4. Farmer, J.D., Gallegati, M., Hommes, C., Kirman, A., Ormerod, P., Cincotti, S., Sanchez, A., Helbing, D.: A complex systems approach to constructing better models for managing financial markets and the economy. Eur. Phys. J. **214**, 295–324 (2012)
5. Manfredi, S., Di Tucci, E., Latora, V.: Mobility and congestion in dynamical multilayer networks with finite storage capacity. Phys. Rev. Lett. **120**(6), 068301 (2018)
6. Glaser, S.M., Fogarty, M.J., Liu, H., Altman, I., Hsieh, C.H., Kaufman, L., Sugihara, G.: Complex dynamics may limit prediction in marine fisheries. Fish Fisheries **15**(4), 616–633 (2014)
7. Grazzini, J., Richiardi, M.G., Tsionas, M.: Bayesian estimation of agent-based models. J. Econ. Dyn. Control **77**, 26–47 (2017)
8. Packard, N.H., Crutchfield, J.P., Farmer, J.D., Shaw, R.S.: Geometry from a time series. Phys. Rev. Lett. **45**(9), 712–716 (1980)
9. Takens, F.: Detecting strange attractors in turbulence. In Dynamical systems and turbulence, Warwick 1980, pages 366–381. Springer, (1981)
10. Farmer, J.D., Sidorowich, J.J.: Predicting chaotic time series. Phys. Rev. Lett. **59**(8), 845 (1987)
11. Ye, H., Beamish, R.J., Glaser, S.M., Grant, S.C.H., Hsieh, C.-H., Richards, L.J., Schnute, J.T., Sugihara, G.: Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. Proceed. Natl. Acade. Sci. **116**(41), E1569–E1576 (2015)
12. Farmer, J.D., Sidorowich, J.J.: Optimal shadowing and noise reduction. Physica D: Nonlinear Phenomena **47**(3), 373–392 (1991)
13. Pires, C., Vautard, R., Talagrand, O.: On extending the limits of variational assimilation in nonlinear chaotic systems. Tellus A **48**(1), 96–121 (1996)
14. Carrassi, A., Bocquet, M., Bertino, L., Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives. Wiley Interdiscip. Rev.: Clim. Change **9**(5), 1–50 (2018)
15. Navon, I. M.: Data assimilation for numerical weather prediction: a review. In Data assimilation for atmospheric, oceanic and hydrologic applications, pages 21–65. Springer (2009)
16. Platt, D.: A comparison of economic agent-based model calibration methods. J. Econ. Dyn. Control **113**, 103859 (2020)
17. Evensen, G.: Advanced data assimilation for strongly nonlinear dynamics. Mon. Weather Rev. **125**(6), 1342–1354 (1997)
18. Judd, K.: Forecasting with imperfect models, dynamically constrained inverse problems, and gradient descent algorithms. Physica D **237**(2), 216–232 (2008)
19. Yang, X.-S.: Nature-inspired Metaheuristic Algorithms. Luniver Press, (2010)
20. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
21. Lorenz, E.N.: Deterministic nonperiodic flow. J. Atmos. Sci. **20**(2), 130–141 (1963)
22. Judd, K.: Failure of maximum likelihood methods for chaotic dynamical systems. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **75**(3), 1–7 (2007)
23. Casdagli, M., Eubank, S., Farmer, J.D., Gibson, J.: State space reconstruction in the presence of noise. Physica D: Nonlinear Phenomena **51**(1–3), 52–98 (1991)
24. Guiñón, J. L., Ortega, E., García-Antón, J., Pérez-Herranz, V.: Moving average and savitzki-golay smoothing filters using mathcad. Papers ICEE, 2007 (2007)
25. Grassberger, P., Hegger, R., Kantz, H., Schaffrath, C., Schreiber, T.: On noise reduction methods for chaotic data. Chaos: An Interdisciplinary Journal of Nonlinear Science, **3**(2):127–141 (1993)
26. Mary, N. et al.: E-handbook of statistical methods. NIST/SEMATECH, 49, (2010)
27. Anishchenko, V.S., Vadivasova, T.E., Kopeikin, A.S., Kurths, J., Strelkova, G.I.: Peculiarities of the relaxation to an invariant probability measure of nonhyperbolic chaotic attractors in the presence of noise. Phys. Rev. E **65**(3), 036206 (2002)
28. Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics, pages 400–407, (1951)
29. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. Ussr Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
30. Nesterov, Y. E.: A method for solving the convex programming problem with convergence rate o $(1/k^2)$. In Dokl. akad. nauk Sssr, vol. 269, pp. 543–547, (1983)
31. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
32. Zeiler, M. D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701, (2012)
33. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. University of Toronto, Technical Report, (2012)
34. Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR), (2014)
35. Reddi, S. J., Kale, S., Kumar, S.: On the convergence of adam and beyond. Proceedings of the 6th International Conference on Learning Representations (ICLR), (2019)
36. Yamada, K. D.: Yamadam: a hyperparameter-free gradient descent optimizer that incorporates unit correction and moment estimation. BioRxiv, page 348557, (2018)
37. Aurell, E., Boffetta, G., Crisanti, A., Paladin, G., Vulpiani, A.: Growth of noninfinitesimal perturbations in turbulence. Phys. Rev. Lett. **77**(7), 1262 (1996)
38. Benettin, G., Galgani, L., Strelcyn, J.-M.: Kolmogorov entropy and numerical experiments. Phys. Rev. A **14**(6), 2338 (1976)
39. Mackey, M.C., Glass, L.: Oscillation and chaos in physiological control systems. Science **197**(4300), 287–289 (1977)

40. Farmer, J.D.: Chaotic attractors of an infinite-dimensional dynamical system. Physica D: Nonlinear Phenomena **4**(3), 366–393 (1982)
41. Shannon, C.E.: Communication in the presence of noise. Proc. IRE **37**(1), 10–21 (1949)
42. Barde, S.: Back to the future: economic rationality and maximum entropy prediction. Technical report, School of Economics Discussion Papers (2012)
43. Tomás-Rodríguez, M., Banks, S. P.: Linear, time-varying approximations to nonlinear dynamical systems: with applications in control and optimization, volume 400. Springer Science & Business Media, (2010)