# Choosing dynamical systems that predict weak input

Sarah E. Marzen*

*W. M. Keck Science Department, Pitzer, Scripps, and Claremont McKenna Colleges, Claremont, California 91711, USA*

(Received 17 November 2020; revised 29 March 2021; accepted 29 June 2021; published 21 July 2021)

Somehow, our brain and other organisms manage to predict their environment. Behind this must be an input-dependent dynamical system, or recurrent neural network, whose present state reflects the history of environmental input. The design principles for prediction—in particular, what kinds of attractors allow for greater predictive capability—are still unknown. We offer some clues to design principles using an attractor picture when the environment perturbs the system's state weakly, motivating and developing some theory for continuous-time time-varying linear reservoirs along the way. Reservoirs that inherently support only stable fixed points are generically good predictors, while reservoirs with limit cycles are good predictors for noisy periodic input.

DOI: 10.1103/PhysRevE.104.014409

## I. INTRODUCTION

In recent years, both the predictive brain and recurrent networks have seen a surge of interest. Those who believe in the predictive brain study how and how well organisms' brains predict their environment, the idea being that prediction is a necessary aspect to survival [1–3]. Those who study recurrent networks engineer systems to predict input, such as written text or audio files. See Refs. [4–7] for examples of advances in designing and training recurrent networks for such applications, and Refs. [8,9] for some applications. But the mechanism behind the predictive brain must be a series of molecular events, describable with nonequilibrium statistical physics. Mathematically, this description is nothing more than a recurrent network. As such, advances in one field can, in principle, advance the other [10].

Many questions still remain about what kinds of recurrent networks are good at predicting their input. We attempt to answer this question in the limit in which environmental input only weakly perturbs the state of the network and in which there is an unspecified sensory readout layer that transforms the system's state into useful predictions of the environmental input (a reservoir [4,5,7,10])—using the tools of dynamical systems theory [11]. Notably, Ref. [12] showed that the vanilla recurrent neural networks that we study here are capable, if trained correctly, of achieving as high a performance as more advanced architectures.

In dynamical systems theory, the attractor-basin portrait reigns supreme. In this viewpoint, one tracks the behavior of the dynamical system (without input) and broadly classifies it as a fixed point, a limit cycle, or a strange attractor. A fixed point means that the dynamical system's state heads towards a particular point and stays there; for a limit cycle, the dynamical system's state oscillates; and strange attractors are famously infinite sets of unstable limit cycles, in which the behavior of the system moves around unpredictably. With input, the attractor-basin portrait might, in principle, jump from one combination of fixed points, limit cycles, and strange attractors to another instantaneously, making it absolutely unsuitable for analyzing the predictive capabilities of recurrent networks. But if the input is weak, there is one attractor-basin portrait that governs the general behavior of the recurrent network for all time. We use this to try and understand which of the three types of attractors—fixed points, limit cycles, or strange attractors—is best for predicting which kinds of input in the weak-input limit. In other words, we attempt to throw the dynamical systems textbook [11] at the still-unanswered question: how can reservoir computers be designed to predict better [10]? The *echo-state property*—that inputs farther in the past have less effect on the present network state—is known to be key, and here we connect this property to attractor types.

In Sec. II, we describe some dynamical systems theory that will aid us in our quest to find design principles for prediction. In Sec. III, we describe the reservoir computing setup. In Sec. IV, we first explain analytically why one can roughly classify the predictive capabilities of a recurrent network with weak input based on attractor type and then run simulations of minimal models to confirm the analytics. Finally, we close in Sec. V with the implications of these findings for the design of reservoir computers and, therefore, potentially, biology.

## II. BACKGROUND

In this paper, we use basic dynamical systems theory to understand reservoir computers. Reservoir computers are actually input-dependent dynamical systems coupled with a linear (or sometimes nonlinear) readout layer. The weights of the readout layer are trained and so we focus on design principles for the underlying input-dependent dynamical system.

Sometimes, one views these input-dependent dynamical systems as random dynamical systems or dynamical systems with noise [13]. However, the so-called noise is actually signal for us and so this viewpoint seems inappropriate. Instead, we turn to Ref. [11]. The setup in that textbook is as follows.

*smarzen@kecksci.claremont.edu

 

There is some dynamical system state $\vec{x}$ that evolves according to the differential equation $\frac{d\vec{x}}{dt} = f(\vec{x})$. Different behaviors result from different choices of $f(\vec{x})$.

It may seem *a priori* that one cannot hope to meaningfully say much about the behavior of $\vec{x}(t)$ without knowing the details of $f$, but, in fact, there is a simple qualitative characterization of such behavior. This is codified in the attractor-basin portrait. Depending on where one starts, $\vec{x}(0)$, one can find different qualitative behaviors. The attractor-basin portrait is simply a codification of which initial conditions lead to which behaviors. Imagine coloring every point in the $d$-dimensional space of possible initial conditions by a color that describes the resulting qualitative behavior. It turns out that the colors form connected regions. Each of these regions is a basin for the "attractor" that describes the qualitative behavior.

There are, surprisingly, only three different qualitative behaviors, though each may be stable, semistable, or unstable: a fixed point, a limit cycle, and a strange attractor. We only care about the stable versions of the attractors in this paper, as the unstable versions will not be chosen by a reservoir computer as described below. If an initial condition belongs to a basin for a stable fixed point, then the state of the system will approach some special point $\vec{x}^*$ such that $f(\vec{x}^*) = \vec{0}$ and stay there. This is a stable fixed point. If an initial condition leads to some sort of oscillation, such that the trajectory is a continuous deformation of a circle, then the attractor is described as a stable limit cycle. And all other stable attractors—those in which one does not leave a certain region but where one does not either stop somewhere or repeat a trajectory—are strange attractors.

For our purposes, a key quantity will be the Jacobian, $J = \nabla_{\vec{x}} f(\vec{x})$. [In the main text, $W(t)$ corresponds to $J(t)$ for a fixed value of the input.] Stable fixed points are rather simple, such that all points in the basin will approach and remain, if $J$ has only eigenvalues with negative real part. Other trajectories have more complicated properties. To understand them somewhat qualitatively, we turn to $e^{\int_0^t J(t')dt'}$, where the Jacobian is measured along the trajectory. For large enough $t$, the logarithm of the eigenvalues of this propagator, scaled by $1/t$, describe the type of trajectory. These quantities are known as Lyapunov exponents. Strange attractors will lead to some eigenvalues that have positive real part.

Strange attractors have many weird properties. The best known of these properties is that if two initial conditions start close together, they diverge in distance exponentially quickly, known as *chaos*. This is a direct result of the positive real part of some of the eigenvalues of the Jacobian. As a point of interest, it is well known that randomly generated dynamical systems are typically strange attractors.

## III. SETUP

Suppose that the system's state $\vec{x}$ evolves according to the equation

$$\frac{d\vec{x}}{dt} = f(\vec{x}, s), \tag{1}$$

where $s(t)$ is the environmental input signal at time $t$ and $f$ is differentiable with respect to $s$. At this point, we assume that $s(t)$ is quite small in its effect on $\frac{d\vec{x}}{dt}$, in that

$$s(t) = s_{\text{carrier}}(t) + \delta s(t), \tag{2}$$

where $s_{\text{carrier}}$ is a carrier signal that changes very slowly and $\delta s(t)$ is a quickly changing signal. For instance, $s_{\text{carrier}}$ can be the mean of the environmental input and thus unchanging. In that situation, it is appropriate to think of $\delta s(t)$, the fluctuations, as being the signal that we desire to predict. We also track only deviations from the natural trajectory, $\delta\vec{x}$, defined more concretely in Results, as this maximizes both memory and predictive power of the reservoir.

A common way of characterizing the reservoir's predictive capabilities is via the memory and prediction function $m(\tau)$ [5], which is the squared correlation coefficient of a multivariate linear regression of $\delta\vec{x}(t)$ against $s(t+\tau)$. (This is also 1 minus the stimulus-normalized mean-squared error of the optimal linear estimate of an input time $\tau$ in the future based on past inputs.) Here, we take some liberties as was done in Ref. [14] and allow for $\tau$ to be either positive or negative. When $\tau$ is positive, we call $m(\tau)$ the prediction function. One can analytically solve for the memory and prediction function in terms of the covariance matrix,

$$C_{ij} = \langle \delta x_i(t)\delta x_j(t)\rangle - \langle \delta x_i(t)\rangle\langle \delta x_j(t)\rangle \tag{3}$$

and

$$\begin{aligned}(p_\tau)_i &= \langle s(t+\tau)\delta x_i(t)\rangle - \langle s(t+\tau)\rangle\langle \delta x_i(t)\rangle \\ &= \langle \delta s(t+\tau)\delta x_i(t)\rangle.\end{aligned} \tag{4}$$

The formula is simply

$$m(\tau) = \frac{1}{\langle \delta s^2\rangle} p_\tau^\top C^{-1} p_\tau. \tag{5}$$

We will view $\langle \delta s^2\rangle$ as an unimportant constant that says more about the input than the functioning of the system and that cancels a corresponding factor in the covariance matrix $C$. In this paper, we consider the prediction function at times before the system could be considered ergodic and so $m(\tau)$ is also a function of the time since the start of simulation, $t$.

## IV. RESULTS

A straightforward Taylor expansion gives

$$\frac{d\vec{x}}{dt} \approx f(\vec{x}, s_{\text{carrier}}) + f_s(\vec{x}, s_{\text{carrier}})\delta s, \tag{6}$$

where we have assumed that deviations from $s_{\text{carrier}}$ are quite small. For the rest of this paper, we will take this approximation as an equality.

This Taylor expansion is somewhat similar in spirit to Refs. [15,16], though we are dealing with a continuous-time system and (relative to Ref. [15]) considering deviations at all times rather than at just one time in the past. As such, our conclusions about the importance of the Jacobian's eigenvalues in determining memory are somewhat similar.

Though this looks somewhat impossible to solve, we make a further approximation, valid only for a short time for most dynamical systems, that $\vec{x}$ is approximately that which solves

$$\frac{d\vec{x}^*}{dt} = f(\vec{x}^*, s_{\text{carrier}}). \tag{7}$$

In general, $\vec{x}^*$ can be quite complicated. We approximate $\vec{x} = \vec{x}^* + \delta\vec{x}$ and expand to find, to first order in $\delta s$ and $\delta\vec{x}$,

$$\frac{d\delta\vec{x}}{dt} = \nabla_{\vec{x}} f(\vec{x}^*, s_{\text{carrier}})\delta\vec{x} + f_s(\vec{x}^*, s_{\text{carrier}})\delta s. \tag{8}$$

This is exactly a continuous-time linear reservoir,

$$\frac{d\delta\vec{x}}{dt} = W(t)\delta\vec{x} + v(t)\delta s, \tag{9}$$

where the recurrent weights are $W(t) = \nabla_{\vec{x}} f(\vec{x}^*, s_{\text{carrier}})$ and the input is transformed according to the vector $v(t) = f_s(\vec{x}^*, s_{\text{carrier}})$. One can explicitly solve for $\delta\vec{x}$ via

$$\delta\vec{x}(t) = \int_0^t \exp\left[\int_{t'}^t W(t'')dt''\right] v(t')\delta s(t')dt', \tag{10}$$

with $\delta\vec{x}(0) = \vec{0}$.

Notice already that this reservoir will have qualitatively different behavior depending on if $\vec{x}^*$ is a stable fixed point, limit cycle, or strange attractor. If $\vec{x}^*$ is a stable fixed point, all of the eigenvalues of $W(t)$ will be negative and the reservoir will have fading memory, i.e., the echo-state property [5]. If $\vec{x}^*$ is a limit cycle, then all of the eigenvalues of $W(t)$ will have nonpositive real part and imaginary parts, corresponding to infinite memory for some dimensions of the input. And if $\vec{x}^*$ is chaotic, then some of the eigenvalues of $W(t)$ will even be positive, corresponding to stronger memory of the past than present, although exactly which aspects of the stimulus are remembered strongly will change every time the linear approximation fails. In this sense, the typical dynamical systems nomenclature [11] provides a framework for understanding the predictive behavior of dynamical systems in the weak-input limit. One can more easily design dynamical systems for the desired prediction task by examining its behavior without input first.

Note that this will not be a useful picture for understanding prediction when the input is not weak. When the input is strong, the attractor-basin portrait can change considerably as a function of time, destroying our ability to understand the trajectory of the system via recourse to the initial attractor-basin portrait. We do not assume weak input because it is biologically relevant or useful for engineering; we only assume it so that we can make some headway in understanding a special type of recurrent networks.

To elaborate on this intuition, we will pursue a more quantitative characterization of a reservoir's predictive capabilities via the memory and prediction function. Note that in calculating covariance for a time-varying reservoir, we are implicitly taking an ensemble approach, averaging over realizations of the input. The ergodic theorem does not usually apply for the simulations in this paper—e.g., if one is dealing with a strange attractor for relatively small snippets of stimulus—in that the performance of a reservoir stimulated by many realizations of input will be different than the performance of that reservoir stimulated by a single long input.

In the Appendix A, we compute the following formula for the memory and prediction function:

$$p_\tau = \int_0^t \exp\left[\int_{t'}^t W(t'')dt''\right] v(t')R(t+\tau-t')dt', \tag{11}$$

$$C = \int_0^t \int_0^t \exp\left[\int_{t'}^t W(s')ds'\right] v(t')R(t'-t'')v(t'')^\top \exp\left[\int_{t''}^t W(s'')^\top ds''\right] dt'dt''. \tag{12}$$

We could use this to estimate the capability of a variety of systems to remember select time series. An autocorrelation function is selected, the system is simulated, and the integrals above approximated as the squared correlation coefficients from linear regression. In practice, however, these integrals are difficult to compute accurately and so we simulate our reservoirs with one of three kinds of input. Two are governed by the same stochastic differential equations,

$$\frac{ds}{dt} = v_s, \tag{13}$$

$$\frac{dv_s}{dt} = -\gamma v_s - 3s + \eta(t), \tag{14}$$

such that $\eta(t)$ is zero-mean white noise with $\langle\eta(t)\eta(t')\rangle = 0.01^2\delta(t-t')$. Constants were chosen somewhat arbitrarily, though the variance on the white noise was chosen to be small so that the input was "weak." When $\gamma = 0.1$, the stimulus oscillates and we say it is "underdamped"; when $\gamma = 10$, the stimulus has exponentially decaying correlations, and we say it is "overdamped." We also aim to predict more complicated types of input and so use a strange attractor to generate stimuli:

$$\frac{ds_1}{dt} = 10(s_2 - s_1) + \eta(t), \tag{15}$$

$$\frac{ds_2}{dt} = s_1(28 - s_3) - s_2, \tag{16}$$

$$\frac{ds_3}{dt} = s_1 s_2 - \frac{8}{3}s_3, \tag{17}$$

where $\eta$ is unimportant mean-zero white noise. This Lorenz attractor is known to generate chaotic behavior. The first coordinate $s_1$ is used as the stimulus for our reservoirs.

We start by analyzing a reservoir that is simply a stable fixed point. In the Appendix C, we show that for a one-dimensional system with a stable fixed point, the prediction function is simply $m(\tau) = w\frac{(\int \frac{A(\lambda)}{w-\lambda}e^{\lambda\tau}d\lambda)^2}{\int A(\lambda)\frac{\lambda}{w-\lambda}d\lambda}$, where we have represented the input's autocorrelation function as $R(t) = \int A(\lambda)e^{\lambda|t|}d\lambda$. If, for instance, the input comes from an overdamped harmonic oscillator, then $m(\tau) = e^{-2\lambda^*\tau}\frac{w}{w+\lambda^*}$. No matter if the input is an overdamped or underdamped harmonic oscillator, higher $w$ (higher restoring forces) are

preferred for higher predictive power; and longer time horizons are exponentially harder to predict.

These lessons hold for multidimensional stable fixed points. To illustrate, we turn to a simple two-dimensional dynamical system, the damped harmonic oscillator, where the input affects the end position of the spring,

$$\frac{dx_1}{dt} = x_2, \tag{18}$$

$$\frac{dx_2}{dt} = f(x_1 - s) - \gamma x_2. \tag{19}$$

Here, $x_1$ plays the role of position, $x_2$ plays the role of velocity, and $f$ is a nonlinear spring force. There is a stable fixed point at $x_1 = 0,\ x_2 = 0$ with $W = \begin{pmatrix} 0 & 1 \\ -f'(0) & -\gamma \end{pmatrix}$ and $v = \begin{pmatrix} 0 \\ f'(0) \end{pmatrix}$. In the weak-input limit, the predictive properties of this nonlinear reservoir are equivalent to a completely linear reservoir, as analyzed in Refs. [14,17]. One of the lessons from these works was that maximizing predictive power requires, in general, tuning the strength of the restoring force and the damping force to the timescales of the input. However, it seems a highly nontrivial task to predict exactly how the timescales of the system should be optimally matched to the timescales of the input. An additional lesson we learn here from examination of $m(\tau)$ and not just $PC = \int_0^\infty m(\tau)d\tau$ is that, unsurprisingly, if the autocorrelation function of the input decays with time in some fashion, then it is harder to predict longer time horizons.

There is, however, a slight difference between the setup in this article and that of Ref. [17]: namely, here, the state of the system at $t = 0$ is identical for all realizations. Initially, then, any deviations in the system state are directly attributable to differences in the stimulus. Later, past differences in the stimulus effectively amount to additional noise in the system state, so that the system state is less able to correctly track the present stimulus. Since our stimuli are designed so that the present value provides the most information about future values, the predictive power of the reservoir decreases with time $t$. We see an example of this in Figs. 1–6 and in the Appendices B and C.

We now turn to reservoirs that naturally oscillate. The danger with using such systems to predict if input is weak is that the system's internal dynamics will eventually effectively erase any memory it may have of the input in an oscillatory fashion, since the system's past positions will be largely determined by its own internal dynamics. In the Appendix D, we show that this intuition holds—strikingly—for a simple oscillatory dynamical system. For illustration here, we turn to the van der Pol oscillator,

$$\frac{dx_1}{dt} = x_2, \tag{20}$$

$$\frac{dx_2}{dt} = \mu(1 - x_1^2)x_2 - (x_1 - s). \tag{21}$$

The van der Pol oscillator is known to exhibit chaotic behavior when driven strongly enough, but we avoid that possibility by focusing on weak input. We have made a choice to imagine that the input signal primarily affects the restoring force $-x_1$ by jostling the other end of the "spring." However, our qualitative results seem to be somewhat insensitive to this choice. For this oscillator, $W = \begin{pmatrix} 0 & 1 \\ -2\mu x_1 x_2 - 1 & \mu(1 - x_1^2) \end{pmatrix}$ and $v = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. No explicit solution exists for $x_1(t),\ x_2(t)$, although, in the Appendix D, we analyze the forecasting capabilities of a reservoir with an infinite number of limit cycles and find that the prediction function decays roughly inversely with time $t$, with oscillations about this decaying mean. Perhaps surprisingly, we see somewhat similar behavior for the van der Pol oscillator in Figs. 1–6.

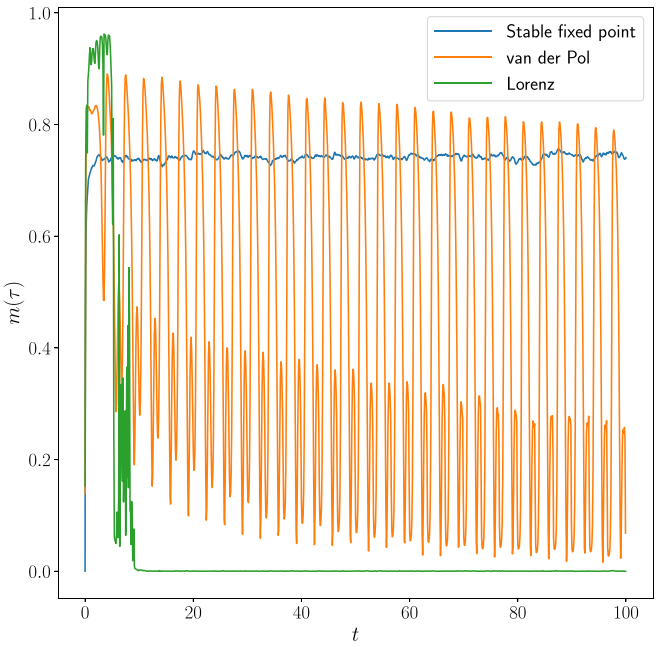


FIG. 1. The prediction function $m(\tau = 0.1)$ as it varies with time since the stimulus was first presented. Shown are three different reservoirs given an overdamped stimulus. The stable fixed point is best equipped to predict this kind of input, as its prediction function levels off at a high level. The van der Pol oscillator's performance oscillates about a decaying mean. The Lorenz attractor starts at a high value, as predicted by short-time analysis in the Appendix, and then decays to minimal predictive performance. All trajectories have the same initial state of the system.

Finally, we examine a prototypical strange attractor, the Lorenz attractor,

$$\frac{dx_1}{dt} = 10(x_2 - x_1) + s, \tag{22}$$

$$\frac{dx_2}{dt} = x_1(28 - x_3) - x_2, \tag{23}$$

$$\frac{dx_3}{dt} = x_1 x_2 - \tfrac{8}{3}x_3. \tag{24}$$

The coupling to the stimulus that we choose is arbitrary, though our qualitative results do not depend so much on this choice. From the prediction functions in Figs. 4–6, we see an extreme version of the behavior shown for stable fixed points and limit cycles. In particular, the prediction function is quite large initially but decays to 0 for the damped harmonic oscillator input considered above. This is strikingly unlike the stable fixed point in that the prediction function does not decay to a nonzero, if small, value, but to zero identically. If the system state is initialized to a value not readily on the attractor, then it takes longer for the prediction function to

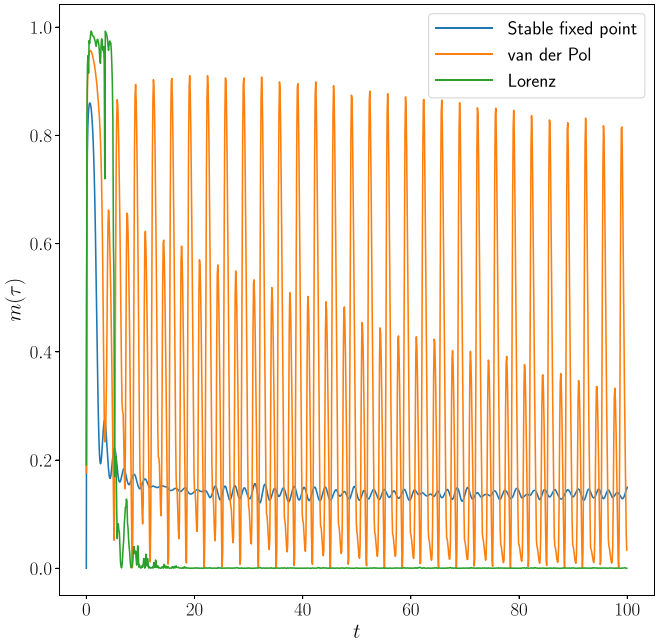


FIG. 2. The prediction function $m(\tau = 0.1)$ as it varies with time since the stimulus was first presented. Shown are three different reservoirs given a chaotic stimulus. The stable fixed point is, surprisingly, best equipped to predict this kind of input, with a prediction function that levels off at a nonzero level. The van der Pol oscillator's performance oscillates about a decaying mean, attaining nearly zero predictive power at the simulation's end. The Lorenz attractor starts at a high value and stays at high values for longer than that in other reservoirs, but then decays to minimal predictive performance. All trajectories have the same initial state of the system.

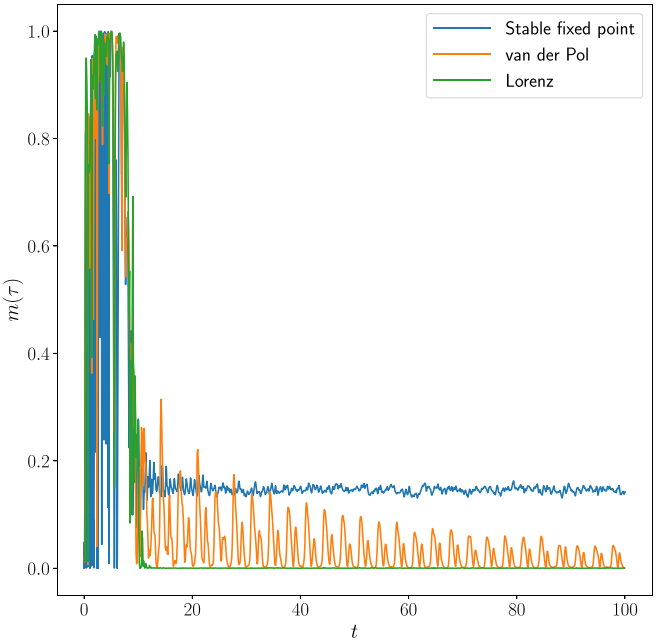


FIG. 3. The prediction function $m(\tau = 0.1)$ as it varies with time since the stimulus was first presented. Shown are three different reservoirs given an underdamped stimulus. The stable fixed point has a prediction function that levels off at a small but nonzero value. The van der Pol oscillator's performance oscillates about a slowly decaying mean. The Lorenz attractor starts at a high value, as predicted by short-time analysis in the Appendix B, and then decays to minimal predictive performance. All trajectories have the same initial state of the system.

decay to 0, as the path to the attractor is stereotyped enough that the "noise" from past stimulus does not greatly affect the system's ability to sense the most recent stimulus.

Initial conditions were set so that there was no jitter in the initial state of the system, but when this condition is removed, the behavior of the prediction function is similar, as these are stable fixed points, stable limit cycles, or strange attractors. The effect of noise has been discussed at greater length in Ref. [18]; its main effect is to add to the covariance matrix and thus degrade both the memory and prediction function.

From Figs. 1–6, we see that it takes some time for the Lyapunov exponents to govern the ability of the reservoir to predict. Exactly how much time this takes depends on exactly the trajectory. The prediction functions shown here for $\tau = 0.1$ and $\tau = 1$ with Gaussian noise determining the jitter in initial conditions shown here for $\tau = 0.1$ is qualitatively typical. An alternative explanation that only appears to work in certain special cases, however, comes from the Appendix D, in which the covariance matrix has eigenvalues that increase linearly with time for limit cycles. This might be expected from the analysis of Ref. [18] if one is to think of the instantaneous filters as having exponential amplification for strange attractors or no exponential decrease for limit cycles. In fact, one might expect the bounds of Ref. [19] to be weaker for the limit cycles and strange attractors.

Why are strange attractors such comparatively bad dynamical systems for use as predictive reservoirs? Along the strange attractor trajectory, the eigenvalues of the Jacobian will contain some positive values. These positive eigenvalues essentially introduce noise into the system's state, so that the incoming stimulus value has to compete with a large amount of past stimulus-induced noise. In other reservoirs, this stimulus-induced noise is more or less useful for understanding future input (see the Appendix B). The dependence on stimulus "noise" could in theory be a boon for specially designed input, such that the right aspect of the past is amplified by the natural dynamics of the system. In practice, we conjecture that these types of input are nearly impossible to design. This does not mean that strange attractors are useless as predictive recurrent networks for input that is not weak, so that this attractor-basin portrait does not hold.

## V. CONCLUSION

Overall, we have found that the characteristics of both the memory and prediction function are qualitatively different for different types of attractors in the limit of weak input, almost regardless of the type of weak input. When the attractor is a stable fixed point, the prediction function decays from a large initial value to a smaller nonzero value; when the attractor is a stable limit cycle, the prediction function decays from its large initial value and oscillates; and when the attractor is a strange attractor, the prediction function decays from its large initial value to a value of 0. These attributes are understandable after an appeal to the eigenvalues of the Jacobian of the natural

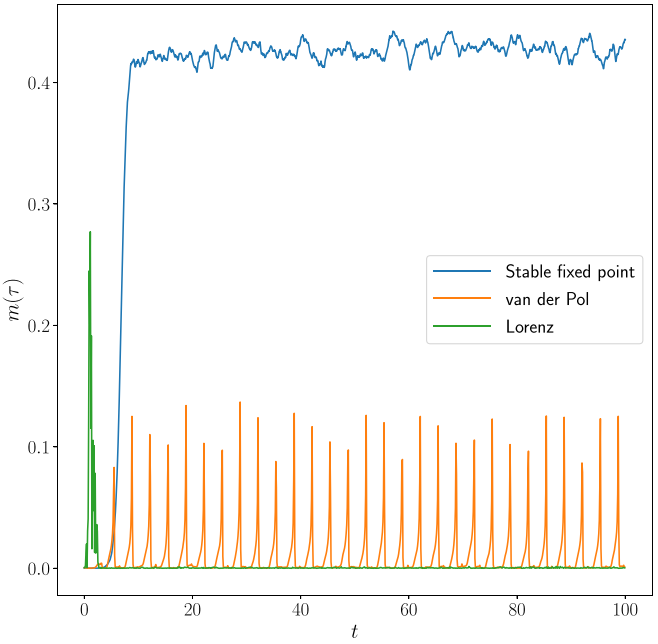


FIG. 4. The prediction function $m(\tau = 1)$ as it varies with time since the stimulus was first presented. Shown are three different reservoirs given an overdamped stimulus. The stable fixed point is best equipped to predict this kind of input, as its prediction function levels off at a high level. The van der Pol oscillator's performance oscillates about a decaying mean. The Lorenz attractor starts at a high value, as predicted by short-time analysis in the Appendix B, and then decays to minimal predictive performance. All trajectories have the same initial state of the system.

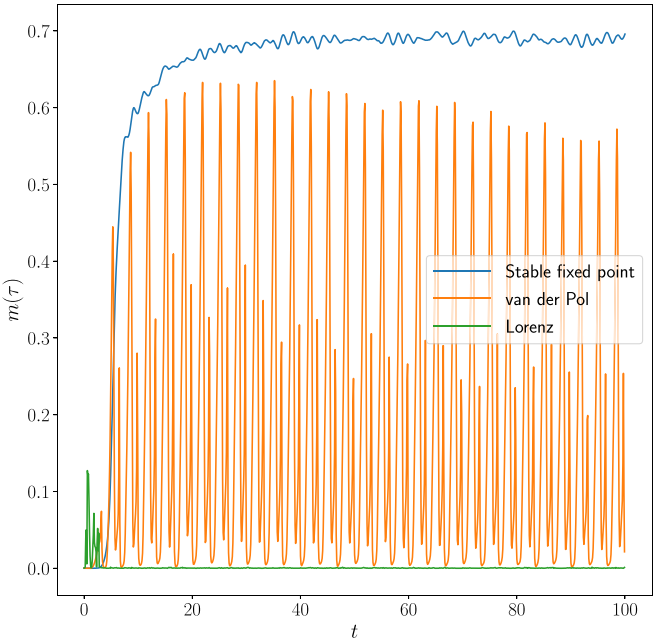


FIG. 5. The prediction function $m(\tau = 1)$ as it varies with time since the stimulus was first presented. Shown are three different reservoirs given an underdamped stimulus. The stable fixed point has a prediction function that levels off at a small but nonzero value. The van der Pol oscillator's performance oscillates about a slowly decaying mean. The Lorenz attractor starts at a high value, as predicted by short-time analysis in the Appendix, and then decays to minimal predictive performance. All trajectories have the same initial state of the system.

system dynamics. Negative eigenvalues lead to exponential decay of the prediction function to a nonzero value; purely imaginary eigenvalues usually lead to some sort of cycling; and positive eigenvalues eventually destroy any and all predictive capability.

Again, this does not necessarily mean that using a strange attractor as a reservoir is always a bad idea, and the analysis in this paper only applies if input is weak. However, if one desires higher predictive capacity or forecasting capacity, the analysis here suggests that one is better served by stable fixed points or limit cycles, unless one takes care to only use short snippets of input time series. This classification is mostly in line with current thinking on the types of reservoirs that predict well, as practitioners prefer reservoirs to have the echo-state property [10]. Reservoirs that are inherently chaotic, while internally rich [16], will likely not predict well, unless the input itself has unusual long-range memory. For specially designed networks, this may not be true [20], though see Ref. [21]. On the other hand, perhaps surprisingly, reservoirs that inherently support stable fixed points or even limit cycles can have rich enough dynamics for prediction while satisfying this echo-state property.

This analysis also complements a previous understanding of chaotic synchronization and other types of synchronization. Suppose that one has access to the dynamical system producing the stimulus. One can replicate this dynamical system, knock out one of the nodes, and feed in that aspect of the stimulus instead. For instance, one can generate chaotic trajectories from the Lorenz equation and feed $s$ into the reservoir given by

$$\frac{dx_2}{dt} = s(28 - x_2) - x_1, \tag{25}$$

$$\frac{dx_3}{dt} = sx_1 - \tfrac{8}{3}x_2. \tag{26}$$

The trajectory of the reservoir will converge to the trajectory of the hidden inputs in the stimulus [22]. If we analyze this reservoir using our methods, we would classify this reservoir as having a stable fixed point attractor and a specially designed $v$, a classification that is borne out by simulations. The presence of chaotic synchronization therefore proves that it is possible to match the reservoir to the input by specially designing not just $W$, but also $v$. Surprisingly, the existence of chaotic synchronization is not at odds with our analysis, but complementary, as our analysis has found that stable fixed points are well suited to the prediction of nearly all signals, and that by designing $W$ and $v$, one can get larger and larger predictive capabilities.

It is unlikely that these insights will hold when the input is not weak, in which case the attractor-basin portrait of dynamical systems will likely not be useful in trying to understand the predictive capabilities of nonlinear reservoirs. However, when the input is weak, the analyses in this paper suggest that one should use a recurrent network whose attractor is a stable fixed point. The design of the stable fixed point is nontrivial

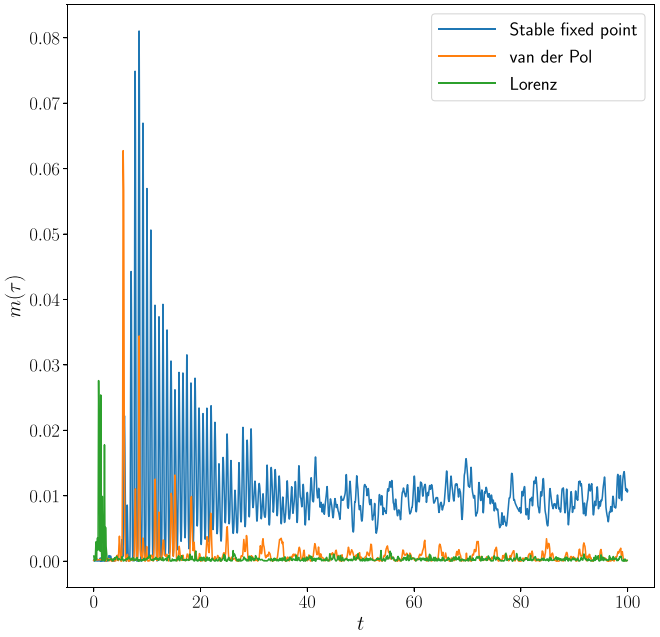


FIG. 6. The prediction function $m(\tau = 1)$ as it varies with time since the stimulus was first presented. Shown are three different reservoirs given a chaotic stimulus. The stable fixed point is, surprisingly, best equipped to predict this kind of input, with a prediction function that levels off at a nonzero level. The van der Pol oscillator's performance oscillates about a decaying mean, attaining nearly zero predictive power at the simulation's end. The Lorenz attractor starts at a high value and stays at high values for longer than that in other reservoirs, but then decays to minimal predictive performance. All trajectories have the same initial state of the system.

[14,17]. It may be useful to use a stable limit cycle when the input is noisy and periodic. Strange attractors seem less well performing, but they are generic, and so one must take care when designing reservoir computers with weak input as a result.

## ACKNOWLEDGMENTS

S.E.M. thanks James P. Crutchfield for useful comments on an earlier draft. S.E.M. was funded by the Air Force Office of Scientific Research under Award No. FA9550-19-1-0411.

## APPENDIX A: COMPUTATION OF MEMORY AND PREDICTION FUNCTION

We have

$$p_\tau = \langle \delta s(t+\tau)\delta x(t)\rangle \tag{A1}$$

$$= \langle \delta s(t+\tau) \int_0^t \exp\left[\int_{t'}^t W(t'')dt''\right] v(t')\delta s(t')dt'\rangle \tag{A2}$$

$$= \int_0^t \exp\left[\int_{t'}^t W(t'')dt''\right] v(t')\langle \delta s(t+\tau)\delta s(t')\rangle dt' \tag{A3}$$

$$= \int_0^t \exp\left[\int_{t'}^t W(t'')dt''\right] v(t')R(t+\tau-t')dt', \tag{A4}$$

and since $\langle \delta s\rangle = 0$, $\langle \delta x\rangle = \vec{0}$, we have

$$C = \langle \delta\vec{x}\delta\vec{x}^\top\rangle \tag{A5}$$

$$= \left\langle \left\{\int_0^t \exp\left[\int_{t'}^t W(t'')dt''\right] v(t')\delta s(t')dt'\right\}\left\{\int_0^t \exp\left[\int_{t'}^t W(t'')dt''\right] v(t')\delta s(t')dt'\right\}^\top\right\rangle \tag{A6}$$

$$= \left\langle \left\{\int_0^t \exp\left[\int_{t'}^t W(t'')dt''\right] v(t')\delta s(t')dt'\right\}\left\{\int_0^t \delta s(t')v(t')^\top \exp\left[\int_{t'}^t W^\top(t'')dt''\right] dt'\right\}\right\rangle \tag{A7}$$

$$= \int_0^t\int_0^t \exp\left[\int_{t'}^t W(s')ds'\right] v(t')\langle \delta s(t')\delta s(t'')\rangle v(t'')^\top \exp\left[\int_{t''}^t W(s'')^\top ds''\right] dt'dt'' \tag{A8}$$

$$= \int_0^t\int_0^t \exp\left[\int_{t'}^t W(s')ds'\right] v(t')R(t'-t'')v(t'')^\top \exp\left[\int_{t''}^t W(s'')^\top ds''\right] dt'dt''. \tag{A9}$$

So only the autocorrelation function matters, as usual, for the memory and prediction function.

## APPENDIX B: MEMORY AND PREDICTION FUNCTION NEAR $t = 0$ AND AT LARGE TIMES

Initially, the reservoirs are initialized so that there is no variability in where they start. As the reservoirs progress, variability in the stimuli being presented translates into effective variability in their system states, degrading their ability to sense new stimuli values for all the stimuli considered here. (It should be said that the stimuli considered here are all representatives of a special class of stimuli—a quite common one—for which the most recent value is typically the most informative.) But initially, all reservoirs show strikingly similar performance.

We can Taylor expand $p_\tau$ and $C$ to understand this phenomenon. Let $v_0$ be the initial value of $v(t)$. For $p_\tau$, we find

$$p_\tau \approx R(\tau)t v_0, \tag{B1}$$

where corrections are of the order of $t^2$ and up. For the covariance matrix, we find

$$C \approx \tfrac{1}{2}t^2 R(0)v_0 v_0^\top, \tag{B2}$$

although this expression is delicate, as this covariance matrix approximation is singular. Higher-order terms—of the order of $t^3$ and up—turn the covariance matrix nonsingular and depend on not only $v_0$, but also the initial value of $W$, which we call $W_0$. We obtain those expressions as follows:

$$C = \left[\tfrac{1}{2}t^2R(0) + \tfrac{1}{3}t^3R'(0)\right]v_0v_0^\top + \tfrac{1}{2}R(0)t^3(W_0v_0v_0^\top + v_0v_0^\top W_0^\top). \tag{B3}$$

Usually, the right-hand terms will ensure nonsingularity of the approximate covariance matrix, though approximate inversion is delicate because the higher-order terms are ensuring that the matrix actually can be inverted. The memory and prediction function is therefore

$$m(\tau) = \frac{1}{\sigma_y^2}p_\tau^\top C^{-1}p_\tau \tag{B4}$$

$$= \frac{1}{\sigma_y^2}R(\tau)^2t^2v_0^\top C^{-1}v_0. \tag{B5}$$

Essentially, one of the eigenvalues of $C$ will be of the order of $t^2$, and the remaining eigenvalues will be of the order of $t^3$. (Imagine rotating into a basis in which $v_0$ is a unit vector.) The one eigenvalue of the order of $t^2$ will be $\frac{1}{2}t^2R(0)$. Then,

$$m(\tau) \approx \frac{1}{\sigma_y^2}\frac{2R(\tau)^2t^2}{R(0)t^2} = \frac{1}{\sigma_y^2}\frac{2R(\tau)^2}{R(0)}. \tag{B6}$$

Note that this is independent of $v_0$ and so independent of exactly what the reservoir is, or even where we start the reservoir. The factor $\frac{2R(\tau)^2}{R(0)\sigma_y^2}$ is only dependent on the type of stimulus. This explains why, in simulation, all reservoirs seemed to do roughly the same near $t = 0$.

At larger times, we can roughly think of the reservoir as having turned past stimulus values into effective noise. In other words, past stimulus values (with some stochasticity) will affect the exact system state. In the worst-case scenario, the variations in the system state will be essentially uncorrelated with the future stimulus value because the reservoir will have amplified the *wrong* aspects of the past stimulus. Then, we can understand the memory and prediction function by returning to a discrete-time linear reservoir,

$$x(1) = Wx(0) + vs(0). \tag{B7}$$

Unlike the setup in the main text, we think of $x(0)$ as being a random variable due to the stochasticity of past inputs that is completely uncorrelated with the present and future stimulus, so that our ability to predict the future essentially relies on our ability to record information about $s(0)$. The memory and prediction function can be obtained from

$$p_\tau = \langle x(1)s(1+\tau)\rangle \tag{B8}$$

$$= W\langle x(0)s(1+\tau)\rangle + v\langle s(0)s(1+\tau)\rangle \tag{B9}$$

$$= vR(\tau+1) \tag{B10}$$

and

$$C = \langle x(1)^2\rangle \tag{B11}$$

$$= W^2\langle x(0)^2\rangle + v^2R(0), \tag{B12}$$

and so, putting these together,

$$m(\tau) = \frac{p_\tau^2}{C} \tag{B13}$$

$$= \frac{R(\tau+1)^2v^2}{W^2\langle x(0)^2\rangle + v^2R(0)} \tag{B14}$$

$$= \frac{R(\tau+1)^2}{1 + \frac{W^2}{v^2}\langle x(0)^2\rangle}. \tag{B15}$$

As $\langle x(0)^2\rangle$ grows, $m(\tau)$ decreases. The longer the past of the stimulus that is presented, the larger $\langle x(0)^2\rangle$ is expected to become, and so, the smaller the memory and prediction function. This problem is especially pernicious for the reservoir based on strange attractors as the positive eigenvalues of the Jacobian amplify past stimulus values in a typically useless way. Past values are preferred over more recent values, but for most stimuli, the most recent value contains more information about future values.

In between the small and large time limits, we expect intermediate behavior. $x(t)$ will be somewhat correlated with $s(t)$ and thus $s(t+\tau)$, but the strength of this correlation will decrease as the attractor stores unnecessary information about the growing past stimulus. We will therefore go from the small-$t$ limit above, in which the memory and prediction function attains a maximal, reservoir-independent value, to something closer to the large-$t$ limit, in which the memory and prediction function attains a small, reservoir-dependent value. For the reservoir that is the Lorenz attractor, $\langle x(t)s(t)\rangle$ seems to tend to values that are of the order of $10^{-3}$, while the eigenvalues of the covariance matrix $\langle x(t)x(t)^\top\rangle$ tend to larger values in the hundreds no matter the type of input. This is not the case for the reservoir whose attractor is a stable fixed point or for the stable limit cycle, for which the eigenvalues of $\langle x(t)x(t)^\top\rangle$ and the values of $\langle s(t)x(t)\rangle$ are comparable. This makes sense in that the positive eigenvalues of the Jacobian of the Lorenz attractor are liable to amplify aspects of the stimulus that are essentially noise, while the nonpositive real parts of the eigenvalues of the Jacobian for stable fixed points and stable limit cycles will reward more salient aspects of the past stimulus.

## APPENDIX C: SEMIQUANTITATIVE RESULTS FOR STABLE FIXED POINTS

To get some intuition for what memory and prediction functions look like for two types of attractors, i.e., that of stable fixed points and that of limit cycles, for both cases, we imagine that we start *on* the attractor, avoiding a discussion of transients that take us closer to the stable fixed point or limit cycle.

We can get slightly more illuminating expressions by considering the integral

$$R(\tau) = \int A(\lambda)e^{\lambda|\tau|}d\lambda, \tag{C1}$$

integrated over the complex plane. Unlike in previous attempts, integration over the complex plane is necessary, as the input that is best matched to a limit cycle might be something that has oscillations in its autocorrelation function.

In the case of a stable fixed point—and we consider a one-dimensional example here so that we can get a little more intuition, though the intuition carries over to the multidimensional case based on the simulations in the main text—we have $W(t) = -w$ and $v(t) = v$ with $w$ positive, and so

$$p_\tau = \int_0^t e^{-w(t-t')} v R(t+\tau-t') dt' \tag{C2}$$

$$= \int_0^t e^{-w(t-t')} v \int A(\lambda) e^{\lambda|t+\tau-t'|} d\lambda \tag{C3}$$

$$= \int A(\lambda) \int_0^t e^{-w(t-t')} v e^{\lambda(t+\tau-t')} dt' d\lambda \tag{C4}$$

$$= \int A(\lambda) v e^{\lambda\tau} \int_0^t e^{(-w+\lambda)(t-t')} dt' d\lambda \tag{C5}$$

$$= \int A(\lambda) v e^{\lambda\tau} \frac{e^{(-w+\lambda)t} - 1}{-w+\lambda} d\lambda, \tag{C6}$$

which in the long-time limit (equivalent to nonequilibrium steady state) converges to

$$p_\tau = \int A(\lambda) e^{\lambda\tau} \frac{v}{w-\lambda} d\lambda. \tag{C7}$$

Similarly, we can treat the covariance matrix,

$$C = \int_0^t \int_0^t e^{-w(t-t')} v R(t'-t'') v e^{-w(t-t'')} dt' dt'' \tag{C8}$$

$$= \int A(\lambda) v^2 \int_0^t \int_0^t e^{-w(2t-t'-t'')} e^{\lambda|t'-t''|} dt' dt'' \tag{C9}$$

$$= \int A(\lambda) v^2 \int_0^t \left( \int_0^{t'} e^{-w(2t-t'-t'')} e^{\lambda(t'-t'')} dt'' + \int_{t'}^t e^{-w(2t-t'-t'')} e^{\lambda(t''-t')} dt'' \right) dt' \tag{C10}$$

$$\approx \int A(\lambda) v^2 \left[ -\frac{1}{w(\lambda-w)} \right] d\lambda, \tag{C11}$$

where we have again taken the long-time limit. Then we find

$$m(\tau) = w \frac{\left( \int \frac{A(\lambda)}{w-\lambda} e^{\lambda\tau} d\lambda \right)^2}{\int A(\lambda) \frac{\lambda}{w-\lambda} d\lambda}. \tag{C12}$$

Note that the expression for the case of higher-dimensional stable fixed points is far more complicated, e.g., see Ref. [17].

To see what kind of input the stable fixed point is best tuned for, we stimulate with an input with exponentially decaying autocorrelation function, $A(\lambda) = \sigma_s^2 \delta(\lambda + \lambda^*)$, and an input that oscillates and decoheres, $A(\lambda) = \frac{\sigma_s^2}{2} \delta(\lambda + \gamma + i\omega) + \frac{\sigma_s^2}{2} \delta(\lambda + \gamma - i\omega)$.

For the first, we find that

$$m(\tau) = e^{-2\lambda^*\tau} \frac{w}{w+\lambda^*}. \tag{C13}$$

In this case, the stronger the restoring force towards the stable fixed point, the higher the memory and prediction function and, as expected, memory strictly decays as the time horizon increases. One can also decrease memory by having the input have more strongly decaying correlations. A system with a stable fixed point that is well suited for this kind of input will try, paradoxically, to get the input back to the stable fixed point as soon as possible. In that way, only the most recent information about $\delta s$ matters for determining the system state.

For the second kind of input, we find

$$p_\tau = \frac{e^{-\gamma\tau}}{(w+\gamma)^2 + \omega^2} [(w+\gamma)\cos(\omega\tau) + \omega \sin(\omega\tau)], \tag{C14}$$

$$C = \frac{w+\gamma}{w} \frac{1}{(w+\gamma)^2 + \omega^2}, \tag{C15}$$

$$m(\tau) = e^{-2\gamma\tau} \frac{w}{w+\gamma} \frac{[(w+\gamma)\cos(\omega\tau) + \omega \sin(\omega\tau)]^2}{(w+\gamma)^2 + \omega^2}. \tag{C16}$$

Note that this memory and prediction function is essentially the memory and prediction function of the input with an exponentially decaying autocorrelation function, but augmented with an oscillatory component that oscillates at the same frequency as the oscillations in the input.

## APPENDIX D: SEMIQUANTITATIVE RESULTS FOR LIMIT CYCLES

To get some intuition for what memory and prediction functions look like for two types of attractors, i.e., that of stable fixed points and that of limit cycles, for both cases, we imagine that we start *on* the attractor, avoiding a discussion of transients that take us closer to the stable fixed point or limit cycle.

We can get slightly more illuminating expressions by considering the integral

$$R(\tau) = \int A(\lambda) e^{\lambda|\tau|} d\lambda, \tag{D1}$$

integrated over the complex plane. Unlike in previous attempts, integration over the complex plane is necessary, as the input that is best matched to a limit cycle might be something that has oscillations in its autocorrelation function.

At this point, we have to choose the limit cycle that we wish to study. For analytic ease, we focus on the system

$$\frac{dx_1}{dt} = x_2 + x_2 \delta s, \tag{D2}$$

$$\frac{dx_2}{dt} = -x_1 - x_1 \delta s, \tag{D3}$$

to which the solution (choosing the time origin properly) is always $x_1(t) = a\cos t, x_2(t) = -a \sin t$. This dynamical system does not actually have stable limit cycles, in that you are not attracted to this limit cycle in particular; depending on one's initial conditions, you travel on one of an infinite number of possible limit cycles. But this dynamical system will suffice for our purposes, as all we need is some sense of $W$ and $v$ for a system with stable limit cycles.

Note that the choice of $v$ here is somewhat ad hoc and does affect the memory and prediction function—but the results that we find for the $v$ chosen here seem to hold true for some other choices of $v$ as well.

For this dynamical system and the trajectory that are given, we find that

$$W = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \; v(t) = \begin{pmatrix} -a\sin t \\ -a\cos t \end{pmatrix}. \tag{D4}$$

From there, we find that

$$p_\tau = \int_0^t \exp\left[W(t-t')\right]\begin{pmatrix} -a\sin t' \\ -a\cos t' \end{pmatrix} R(t+\tau-t')dt' \tag{D5}$$

$$= \int A(\lambda)\int_0^t \exp\left[W(t-t')\right]\begin{pmatrix} -a\sin t' \\ -a\cos t' \end{pmatrix} e^{\lambda(t+\tau-t')}dt' \tag{D6}$$

$$= \int A(\lambda)e^{\lambda\tau}\int_0^t \begin{pmatrix} a\sin(t-2t') \\ a\cos(t-2t') \end{pmatrix} e^{\lambda(t-t')}dt'd\lambda \tag{D7}$$

$$\approx \int aA(\lambda)\frac{e^{\lambda\tau}}{4+\lambda^2}\begin{pmatrix} 2\cos t + \lambda\sin t \\ -\lambda\cos t + 2\sin t \end{pmatrix} d\lambda, \tag{D8}$$

again taking the long-time limit, and that

$$C = \int_0^t\int_0^t e^{W(t-t')}v(t')R(t'-t'')v(t'')^\top e^{W^\top(t-t'')}dt'dt'' \tag{D9}$$

$$= a^2\int_0^t\int_0^t \begin{pmatrix} \sin(t-2t') \\ \cos(t-2t') \end{pmatrix} R(t'-t'')(\sin(t-2t'') \quad \cos(t-2t''))dt'dt'' \tag{D10}$$

$$= a^2\int A(\lambda)\int_0^t\int_0^t e^{\lambda|t'-t''|}\begin{pmatrix} \sin(t-2t')\sin(t-2t'') & \sin(t-2t')\cos(t-2t'') \\ \cos(t-2t')\sin(t-2t'') & \cos(t-2t')\cos(t-2t'') \end{pmatrix} dt'dt''d\lambda \tag{D11}$$

$$\approx a^2\int A(\lambda)\frac{\lambda}{4+\lambda^2}\begin{pmatrix} t & 0 \\ 0 & -t \end{pmatrix} d\lambda, \tag{D12}$$

again taking the long-time limit, where we have ignored many oscillatory components. We already see strikingly different behavior for the covariance matrix than with the stable fixed point, in that the covariance elements decay linearly with time rather than approaching a finite value. This then leads to a huge degradation in predictive capacity via a decrease in the memory and prediction function,

$$m(\tau) = \frac{1}{\sigma_s^2}p_\tau^\top C^{-1}p_\tau. \tag{D13}$$

We again specialize to the two kinds of input considered in the last section of the Appendix. First, in the case of $R(\tau) = e^{-\lambda^*|\tau|}$, we have that

$$p_\tau = a\frac{e^{-\lambda^*\tau}}{4+(\lambda^*)^2}\begin{pmatrix} 2\cos t - \lambda^*\sin t \\ \lambda^*\cos t + 2\sin t \end{pmatrix}, \tag{D14}$$

$$C = -a^2\frac{\lambda^*}{4+(\lambda^*)^2}\begin{pmatrix} t & 0 \\ 0 & -t \end{pmatrix}, \tag{D15}$$

$$m(\tau) = \frac{1}{t}\frac{\lambda^* e^{-2\lambda^*\tau}}{4+(\lambda^*)^2}\left\{4\lambda^*\sin(2t) + [4+(\lambda^*)^2]\cos(2t)\right\}. \tag{D16}$$

The $1/t$ factor implies that the limit cycle is incredibly bad at forming predictive features of the input with an exponentially decaying autocorrelation function. This holds true not just for the likely Markovian input, but also for the oscillatory input.

But to see if the oscillatory input is also badly predicted by the limit cycle, we turn to

$$A(\lambda) = \tfrac{1}{2}\delta(\lambda+\gamma+i\omega) + \tfrac{1}{2}\delta(\lambda+\gamma-i\omega), \tag{D17}$$

so that

$$p_\tau = \frac{a}{2}\frac{e^{(-\gamma+i\omega)\tau}}{4+(-\gamma+i\omega)^2}\begin{pmatrix} 2\cos t + (-\gamma+i\omega)\sin t \\ -(-\gamma+i\omega)\cos t + 2\sin t \end{pmatrix} + \frac{a}{2}\frac{e^{(-\gamma-i\omega)\tau}}{4+(-\gamma-i\omega)^2}\begin{pmatrix} 2\cos t + (-\gamma-i\omega)\sin t \\ -(-\gamma-i\omega)\cos t + 2\sin t \end{pmatrix}, \tag{D18}$$

which is $o(t)$ and

$$C \approx \frac{a^2}{2}\left[\frac{-\gamma+i\omega}{4+(-\gamma+i\omega)^2} + \frac{-\gamma-i\omega}{4+(-\gamma-i\omega)^2}\right]\begin{pmatrix} t & 0 \\ 0 & -t \end{pmatrix} \tag{D19}$$

$$= -a^2\frac{\gamma(4+\gamma^2-\omega^2)+2\omega^2\gamma}{(4+\gamma^2-\omega^2)^2+(2\omega\gamma)^2}\begin{pmatrix} t & 0 \\ 0 & -t \end{pmatrix}, \tag{D20}$$

which is $O(t)$ in terms of its eigenvalues. We can conclude that for oscillatory input as well, the memory and prediction function decays as $1/t$.

One might think that a naturally oscillatory dynamical system would be a well-designed predictive dynamical sensor of oscillatory input, but, in fact, the dynamics of the system interfere with its sensory capabilities. Its motion is determined not by the sensor input, but by its own internal dynamics, and so the system becomes solipsistic, losing sight of what it is trying to sense.

## APPENDIX E: EFFECT OF $\tau$ AND NOISE IN THE INITIAL CONDITION

Figures 4–6 show the prediction functions for when $\tau = 1$ and when there is a small amount of Gaussian noise added to the initial state of the system.


[1] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, Predictive information in a sensory population, Proc. Natl. Acad. Sci. USA **112**, 6908 (2015).

[2] R. P. N. Rao and D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects, Nat. Neurosci. **2**, 79 (1999).

[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2018).

[4] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Comput. **9**, 1735 (1997).

[5] H. Jaeger, *Short Term Memory in Echo State Networks*, Vol. 5 (GMD-Forschungszentrum Informationstechnik, Germany, 2001).

[6] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv:1406.1078.

[7] B. Schrauwen, D. Verstraeten, and J. V. Campenhout, An overview of reservoir computing: Theory, applications and implementations, in *Proceedings of the 15th European Symposium on Artificial Neural Networks* (ESANN, Bruges, Belgium, 2007), pp. 471–482.

[8] 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).

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, Language models are few-shot learners, arXiv:2005.14165.

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

[11] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (Addison-Wesley, Reading, MA, 1994).

[12] J. Collins, J. Sohl-Dickstein, and D. Sussillo, Capacity and trainability in recurrent neural networks, arXiv:1611.09913.

[13] L. Arnold, Random dynamical systems, in *Dynamical Systems* (Springer, New York, 1995), pp. 1–43.

[14] S. Marzen, Difference between memory and prediction in linear recurrent networks, Phys. Rev. E **96**, 032308 (2017).

[15] M. Inubushi and K. Yoshimura, Reservoir computing beyond memory-nonlinearity trade-off, Sci. Rep. **7**, 10199 (2017).

[16] M. C. Ozturk, D. Xu, and J. C. Príncipe, Analysis and design of echo state networks, Neural Comput. **19**, 111 (2007).

[17] A. Hsu and S. E. Marzen, Time cells might be optimized for predictive capacity, not redundancy reduction or memory capacity, Phys. Rev. E **102**, 062404 (2020).

[18] M. Hermans and B. Schrauwen, Memory in linear recurrent neural networks in continuous time, Neural Networks **23**, 341 (2010).

[19] L. Gonon, L. Grigoryeva, and J.-P. Ortega, Memory and forecasting capacities of nonlinear recurrent networks, Physica D: Nonlin. Phenom. **414**, 132721 (2020).

[20] T. Toyoizumi and L. F. Abbott, Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime, Phys. Rev. E **84**, 051908 (2011).

[21] J. Schuecker, S. Goedeke, and M. Helias, Optimal Sequence Memory in Driven Random Networks, Phys. Rev. X **8**, 041029 (2018).

[22] L. M. Pecora and T. L. Carroll, Synchronization in Chaotic Systems, Phys. Rev. Lett. **64**, 821 (1990).