Randomly Coupled Oscillators for Time Series Processing

Andrea Ceni [*1] **Andrea Cossu** [*1] **Jingyue Liu** [2] **Maximilian Stölzle** [2] **Cosimo Della Santina** [2]
Claudio Gallicchio [1] **Davide Bacciu** [1]

Abstract

We investigate a physically-inspired recurrent neural network derived from a continuous-time ODE modelling a network of coupled oscillators. Enthralled by the Reservoir Computing paradigm, we introduce the Randomly Coupled Oscillators (RCO) model, which leverages an untrained recurrent component with a smart random initialization. We analyse the architectural bias of RCO and its neural dynamics. We derive sufficient conditions for the model to have a unique asymptotically uniformly stable input-driven solution. We also derive necessary conditions for stability, that permit to push the system of oscillators slightly beyond the edge of stability. We empirically assess the effectiveness of RCO in terms of its stability and its long-term memory properties. We compare its performance against both fully-trained and randomized recurrent models in a number of time series processing tasks. We find that RCO provides an excellent trade-off between robust long-term memory properties and ability to predict the behavior of non-linear, chaotic systems.

1. Introduction

Machine learning (ML) is nowadays ubiquitous in our society covering applications from healthcare (De Fauw et al., 2018) to chatbots (Lund & Wang, 2023). Major breakthroughs have been made especially on feed-forward neural network architectures leveraging on convolutions, e.g. AlexNet, VGG (Krizhevsky et al., 2017; Simonyan & Zisserman, 2014), residual connections, e.g. ResNet, EfficientNet (He et al., 2016a; Tan & Le, 2019), and on attention mechanisms, e.g. Transformers (Vaswani et al., 2017). Contrariwise, Recurrent Neural Network (RNN) architectures witnessed a slight disinterest in the last years. The soaring computational power provided by modern hardware accelerators made less appealing RNN architectures, due to their intrinsic difficulty to be implemented in parallel. Nevertheless, RNNs represent more plausibly the way microcircuits in the brain perform computations (Mante et al., 2013). Unlike feed-forward neural networks, the key characteristic of RNNs is to spatially encode into the RNN parameters and its hidden state the temporal information present in the input. In theory, the current recurrent hidden state conveys the context of the input signal from the infinite past to present time.

In this paper, we investigate a physically-inspired RNN derived from a continuous-time ODE describing a network of oscillators. We cast this model into the Reservoir Computing (RC) framework and we dub it Randomly Coupled Oscillators (RCO). Oscillators represent an archetypal dynamical behaviour ubiquitous in nature that can be found in chemical reactions such as Belousov–Zhabotinsky, electronic circuits like amplifiers, business cycles, central nervous system diseases like Parkinson's, pendulum clocks, fireflies' light pulses, and many others (Pikovsky et al., 2002). This motivates us to inspect to what extent an *untrained* ensemble of randomly coupled heterogeneous oscillators can be exploited for time series processing. We take inspiration from a recently proposed deep learning model based on oscillators (Rusch & Mishra, 2023), called coupled oscillatory RNN (coRNN), and extend this model to the case of heterogeneous oscillators coupled only in the position variable, but being uncoupled in the damping terms. Moreover, our RCO follows the principles of RC by leveraging untrained hidden parameters.
The contribution of our paper can be summarised as follows:

- We introduce the RCO model and, in Section 3.4, we demonstrate that it can be interpreted as a generalisation of a popular RC model called Leaky-ESN. As such, the RCO model has the potential to be implemented on a variety of tasks involving time series, among which classification (Jaeger et al., 2007), and chaotic attractor reconstruction (Lu et al., 2018).

[*] Equal contribution [1] Department of Computer Science, University of Pisa, Pisa, Italy [2] CoR Department, TU Delft, Delft, Netherlands. Correspondence to: Andrea Ceni <andrea.ceni@di.unipi.it>, Andrea Cossu <andrea.cossu@di.unipi.it>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023. Copyright 2023 by the author(s).

- In Section 4, we provide a theoretical analysis of RCO linear stability. We discover that RCO exhibits an architectural bias towards the identity (see in particular Theorem 4.1), a widely recognised important feature in the context of deep learning (He et al., 2016b). Moreover, we provide a characterisation of the eigenspectrum of the linearised RCO model in Theorem 4.4.

- In Proposition 4.2, we derive sufficient conditions to impose on the hyperparameters of the RCO model to ensure the existence of a unique asymptotically uniformly stable input-driven solution. In Proposition 4.5, we also derive necessary conditions for the same property, that are less restrictive than the sufficient ones.

- In Sections 5-6, we test our RCO model on popular time series benchmarks. In time series classification tasks, RCO outperforms randomized models and suffers only a slight degradation in performance with respect to fully-trained models, while being two orders of magnitude more efficient. On chaotic systems forecasting, RCO clearly outperforms fully-trained recurrent models and performs similarly to randomized recurrent models.

2. Reservoir Computing background

From a purely mathematical perspective, universal approximation theorems prove that feed-forward neural networks can learn any function within a given precision (Hornik et al., 1990), as RNNs can learn any dynamical system (Funahashi & Nakamura, 1993), provided a sufficient number of parameters. Dynamical systems and RNNs are intimately linked together by the internal state which traces an input-driven trajectory in phase space. Adopting this perspective allows to use standard tools from dynamical systems theory, like bifurcation theory, Lyauponov stability analysis, and control theory, to get insights on the inner functioning of RNNs. The key idea of Reservoir Computing (RC) (Lukoševičius & Jaeger, 2009) is to treat the internal dynamics as fixed, i.e. untrained. RC leverages on smart random instantiations of the recurrent part of the model, called in this context the reservoir, and only train an output layer, usually with simple linear regression techniques, to decode the internal dynamics into an output signal. A good reservoir must balance between two contrasting properties: fading memory, and input separation. The former forces the reservoir to weight more the present information at the expenses of far-in-the-past information. This property of gradually fading the memory of the past is closely related to the concept of asymptotic convergence in dynamical systems, thus a property of stability. The latter equips the reservoir with the ability to map different inputs into internal states that separate enough one other. This is particularly important in classification tasks involving long time series where apparently slight differences in the input might be crucial for the assignment to the appropriate class. In a sense, mildly unstable reservoir dynamics can be beneficial to the ability to separate inputs. However, these instabilities should not be systematic of the model, but rather temporary, allowing to exploit transient dynamics for the purpose of the classification. In summary, it is desirable to design reservoirs whose input-driven dynamics occur at the edge of the stability region, which has been recognised to be useful for time series processing (Bertschinger & Natschläger, 2004; Legenstein & Maass, 2007).

We take inspiration from a particular class of RC models called Echo State Networks (ESNs) (Jaeger, 2002). An ESN is a discrete-time recurrent model whose reservoir state update is computed as follows:

$$\mathbf{y}_k = \tanh(\mathbf{W}\mathbf{y}_{k-1} + \mathbf{V}\mathbf{u}_k + \mathbf{b}), \tag{1}$$

where $\mathbf{y}_k \in \mathbb{R}^N$, $\mathbf{u}_k \in \mathbb{R}^I$ are the internal state and input trajectories at time-step k, respectively and $\mathbf{W} \in \mathbb{R}^{N\times N}$, $\mathbf{V} \in \mathbb{R}^{N\times I}$, $\mathbf{b} \in \mathbb{R}^N$ are fixed, randomly-initialised parameters. ESNs work under the fundamental assumption of the Echo State Property (ESP), which is defined as follows (Yildiz et al., 2012).

Definition 2.1. An input-driven system $\mathbf{y}_{k+1} = G(\mathbf{y}_k, \mathbf{u}_{k+1})$ has the ESP if there exists a sequence δ_k converging to zero such that, for all input sequences $\mathbf{u}_k$, and all pairs of initial states $\mathbf{y}_0, \overline{\mathbf{y}}_0$, it holds that

$$||\mathbf{y}_k - \overline{\mathbf{y}}_k|| \leq \delta_k \tag{2}$$

where $\mathbf{y}_k$, and $\overline{\mathbf{y}}_k$, are the $\mathbf{u}_k$-driven trajectories starting from $\mathbf{y}_0$, and $\overline{\mathbf{y}}_0$, respectively, and $||\cdot||$ denotes the Euclidean norm.

Remark 2.2. The ESP of definition 2.1 is a property of asymptotic uniform contraction in phase space for an input-driven system. Note in fact that, although $\delta_k \to 0$, it is not necessary for δ_k to be monotonically decreasing. The ESP implies that, in the long term, eventually all trajectories starting from all initial conditions will synchronise to a unique input-driven solution.

Throughout the paper, we will use the notation $||\cdot||$ to denote the Euclidean norm of a vector, or the matrix norm induced by it, if the argument is a matrix. One sufficient condition to ensure the ESP for the input-driven system $\mathbf{y}_{k+1} = G(\mathbf{y}_k, \mathbf{u}_{k+1})$ is to impose that (Ceni et al., 2020)

$$\sup_{\mathbf{y},\mathbf{u}} \left\| \frac{\partial G}{\partial \mathbf{y}}(\mathbf{y}, \mathbf{u}) \right\| < 1. \tag{3}$$

However, eq. (3) is a much stronger condition than asymptotic uniform contraction, since it implies contraction at each time step. As such, this strong contraction property might harm the expressiveness of an RC model. Notably, the same sufficient condition of eq. (3) can be derived applying known results from the control theory literature, for example see section 4.2 of the book (Slotine et al., 1991).

It is common practice in RC with ESN to initialise the hidden matrix $\mathbf{W}$ such that its spectral radius ρ is smaller than 1. This is not a sufficient condition for the ESP, but in practice it has been shown to work well. Moreover, the matrix $\mathbf{V}$ is scaled by a scalar value v. Both ρ and v are task-specific and are usually determined by a model selection phase (e.g., with random/grid search).

3. From harmonic oscillator to Randomly Coupled Oscillators

In this section we trace a path from the classical harmonic oscillator to our proposed recurrent neural network models. In Section 3.1 we provide a gentle introduction to the damped harmonic oscillator, framing it in the context of fading memory systems. Section 3.2 introduces the coRNN model, while in Section 3.3 we describe our proposed models, hcoRNN and RCO. Finally, in Section 3.4 we link our RCO model to a well-known RC model called Leaky-ESN.

3.1. Damped harmonic oscillator

Harmonic oscillators are at the core of classical mechanics. They describe simple oscillatory motions around an equilibrium point without experiencing any dissipation of energy. The equation of an harmonic oscillator reads as follows:

$$\ddot{y} = -\gamma y. \tag{4}$$

The general solution of eq. (4) is given by

$$y(t) = A\cos(\sqrt{\gamma}t + \psi), \tag{5}$$

where the amplitude A, and phase ψ, are uniquely determined by the initial conditions. Here different initial conditions give rise to different solutions.

A key feature required for a dynamical system to be exploited for computational purposes is the *fading memory*. In rough terms, we aim for a stable dynamical system which is able to wash out in the long term any dependencies from the initial conditions. Therefore, exploiting a network of randomised oscillators of the form of eq. (5) for computational purposes is impractical. Fading memory can be brought by a damping term into eq. (4), thus introducing a source of energy dissipation. The equation of a damped harmonic oscillator reads as follows:

$$\ddot{y} = -\gamma y - \varepsilon \dot{y}. \tag{6}$$

In eq. (6), the γ scalar term relates with the intrinsic frequency of the underlying harmonic oscillator, while the ε scalar term refers to the strength of the damping force (also called friction) exerting against the harmonic oscillator. Any value of $\varepsilon > 0$ induces the system of eq. (6) to converge towards the resting state of $y = 0$. According to the value of ε, two main behaviours can be observed: overdamped dynamics (when $\varepsilon > 2\sqrt{\gamma}$) characterised by exponential decay towards $y = 0$ without any oscillations, or underdamped dynamics (when $0 < \varepsilon < 2\sqrt{\gamma}$) characterised by an oscillatory behaviour with decreasing amplitude in time. In this sense, the damped harmonic oscillator possesses a fading memory property. Often, an external time-varying force $f(t)$ drives the damped oscillator giving rise to the following equation:

$$\ddot{y} = f(t) - \gamma y - \varepsilon \dot{y}. \tag{7}$$

The driven damped harmonic oscillators of eq. (7) emerges in many physical, engineering and biological systems.

3.2. Coupled oscillatory RNN

The next step is to build a network of input-driven damped harmonic oscillators, and use this physically-inspired neural network model to perform computations. Let us denote with vectors $\boldsymbol{\gamma}, \boldsymbol{\varepsilon} \in \mathbb{R}^N$, the characteristic frequencies and damping ratios of each oscillator in the network. Then, the following equation describes a network of heterogeneous driven damped harmonic oscillators:

$$\ddot{\mathbf{y}} = \mathbf{f}(t) - \boldsymbol{\gamma} \odot \mathbf{y} - \boldsymbol{\varepsilon} \odot \dot{\mathbf{y}}, \tag{8}$$

where $\odot$ denotes the point-wise multiplication of vectors.

In Rusch & Mishra (2023), the authors introduce a parametrisation of the function $\mathbf{f}(t)$ as an input-driven nonlinear layer as follows:

$$\mathbf{f}(t) = \tanh(\mathbf{W}\mathbf{y} + \mathcal{W}\dot{\mathbf{y}} + \mathbf{V}\mathbf{u}(t) + \mathbf{b}), \tag{9}$$

where $\mathbf{u}(t)$ is the external input driving the network. Note that in eqs. (8)-(9) the coupling between neurons is nonlinear, due to $\tanh$, and defined by the matrices $\mathbf{W}$ and $\mathcal{W}$. Eqs. (8)-(9) describe an RNN model with hidden state $\mathbf{y} \in \mathbb{R}^N$, with N being the number of neurons. $\mathbf{W} \in \mathbb{R}^{N\times N}$ and $\mathcal{W} \in \mathbb{R}^{N\times N}$ are the hidden-to-hidden connections, $\mathbf{V} \in \mathbb{R}^{N\times I}$ are the input-to-hidden connections, and $\mathbf{b} \in \mathbb{R}^N$ the bias vector of the RNN. The hyperbolic tangent mediates a nonlinear bounded response in the oscillators. This approach has the advantage to architecturally constraint the driving force to be bounded in $(-1, 1)$ for each neuron, regardless of the other neurons and the external input.

Introducing the variable $\mathbf{z} = \dot{\mathbf{y}}$, we get the following first order system of ODEs

$$\begin{aligned}\dot{\mathbf{y}} &= \mathbf{z},\\ \dot{\mathbf{z}} &= \tanh(\mathbf{W}\mathbf{y} + \mathcal{W}\mathbf{z} + \mathbf{V}\mathbf{u}(t) + \mathbf{b}) - \boldsymbol{\gamma} \odot \mathbf{y} - \boldsymbol{\varepsilon} \odot \mathbf{z},\end{aligned} \tag{10, 11}$$

that we discretise with an implicit (the $\dot{\mathbf{y}}$ equation), and an explicit (the $\dot{\mathbf{z}}$ equation) Euler numerical scheme. The result is the coRNN model.

Definition 3.1. coupled oscillatory RNN (coRNN), from Rusch & Mishra (2023). A coRNN with unique scalar values $\boldsymbol{\gamma} \equiv \gamma, \boldsymbol{\varepsilon} \equiv \varepsilon$ is defined by the following equation:

$$\begin{aligned} \mathbf{y}_{k+1} &= \mathbf{y}_k + \tau \mathbf{z}_{k+1}, \\ \mathbf{z}_{k+1} &= \mathbf{z}_k + \tau \big(\tanh(\mathbf{W}\mathbf{y}_k + \mathcal{W}\mathbf{z}_k + \mathbf{V}\mathbf{u}_{k+1} + \mathbf{b}) \\ &\qquad - \gamma \mathbf{y}_k - \varepsilon \mathbf{z}_k \big). \end{aligned} \tag{12}$$

3.3. Randomly Coupled Oscillators

We will consider the particular case of $\mathcal{W} = \mathbf{0}$, with heterogeneous oscillators (i.e. $\boldsymbol{\gamma}, \boldsymbol{\varepsilon}$, vectors), and we will use the name hcoRNN (heterogeneous coRNN) to refer to this model.

Definition 3.2. Heterogeneous coRNN (hcoRNN). We introduce the hcoRNN model, whose update reads as follows:

$$\begin{aligned} \mathbf{y}_{k+1} &= \mathbf{y}_k + \tau \mathbf{z}_{k+1}, \\ \mathbf{z}_{k+1} &= \mathbf{z}_k + \tau \big(\tanh(\mathbf{W}\mathbf{y}_k + \mathbf{V}\mathbf{u}_{k+1} + \mathbf{b}) \\ &\qquad - \boldsymbol{\gamma} \odot \mathbf{y}_k - \boldsymbol{\varepsilon} \odot \mathbf{z}_k \big). \end{aligned} \tag{13}$$

In practice, in an hcoRNN the coupling occurs only by means of the activations $\mathbf{y}$, while the damping terms remains decoupled, i.e. characteristic to each neuron regardless of the behaviour of the other neurons.

The coRNN and hcoRNN models are fully trained with the backpropagation through-time algorithm, which realises a stochastic gradient descent. In contrast, we propose an alternative variant of hcoRNN where we leverage fixed, untrained hidden parameters $(\mathbf{W}, \mathbf{V}, \mathbf{b})$.

Definition 3.3. Randomly Coupled Oscillators (RCO). An RCO is an hcoRNN (Definition 3.2) with fixed, random parameters. In line with the RC framework, the matrix $\mathbf{W}$ is tuned by simply rescaling its spectral radius ρ, and the matrix $\mathbf{V}$ is scaled with a scalar value ν.

In eqs. (13) of hcoRNN and RCO, the scalar value τ is linked to the step size of the numerical integration. Therefore, if one wants to discretise the continuous-time model for the sake of merely reproducing the continuous-time dynamics, then an opportunely small value of the step size is required, i.e. $\tau \ll 1$. However, here we are not interested in reliably simulating trajectories of the continuous-time dynamical system defined by eqs. (10)-(11), but rather to investigate the expressiveness of the physically-inspired discrete-time RNN model of eqs. (13) that we derived. As a consequence, in the remainder of this paper we will treat $\tau > 0$ as an hyperparameter of the RNN model.

From the point of view of ML applications, the hidden states computed by eqs. (13) are exploited as features encoding crucial temporal information for the processing of the input time series $\mathbf{u}_k$. To solve time series tasks, we consider a linear transformation of the hidden state $\mathbf{y}$ to an output state $\mathbf{r}$ as follows:

$$\mathbf{r}_{k+1} = \mathbf{W}_o \mathbf{y}_{k+1} + \mathbf{b}_o \tag{14}$$

where $\mathbf{W}_o, \mathbf{b}_o$, are weights and biases of the output layer. Eqs. (13)-(14) describe an RNN model mapping input sequences $\mathbf{u}_k$ into output sequences $\mathbf{r}_k$. The parameters $\mathbf{W}_o, \mathbf{b}_o$ are the only trainable parameters of an RCO.

3.4. RCO as Leaky-ESNs

In the particular case of eqs. (13) with $\boldsymbol{\varepsilon} \equiv \dfrac{1}{\tau}$, the $\mathbf{z}$-dynamics become completely determined by the $\mathbf{y}$-dynamics. Therefore, the hcoRNN equation becomes

$$\mathbf{y}_{k+1} = \mathbf{y}_k + \tau^2 \tanh(\mathbf{W}\mathbf{y}_k + \mathbf{V}\mathbf{u}_{k+1} + \mathbf{b}) - \tau^2 \boldsymbol{\gamma} \odot \mathbf{y}_k.$$

Interestingly, setting further $\boldsymbol{\gamma} \equiv 1$, we recover a popular RC model named Leaky-ESN (Jaeger et al., 2007), whose state-update equation reads as follows:

$$\mathbf{y}_{k+1} = \tau^2 \tanh(\mathbf{W}\mathbf{y}_k + \mathbf{V}\mathbf{u}_{k+1} + \mathbf{b}) + (1 - \tau^2)\mathbf{y}_k. \tag{15}$$

In the context of eq. (15), the hyperparameter τ is interpreted as the squared root of the leak rate of the model. The sufficient condition of eq. (3) is ensured for eq. (15) whenever $||\mathbf{W}|| < 1$. The Leaky-ESN model has been successfully used in many ML tasks involving time series processing. Remarkably, the Leaky-ESN with linear output layer as in eq. (14) can accurately learn the climate of chaotic attractors (Lu et al., 2018).

The above remarks established the discrete-time model of eqs. (13) as a bridge between the coRNN model and the Leaky-ESN model. From this perspective, the hcoRNN and RCO stand out as powerful models able to describe both stable complex oscillatory dynamics and chaotic dynamics, provided with an opportune choice of hyperparameters.

4. Linear stability analysis

In this section we perform a linear stability analysis of the hcoRNN model (Definition 3.2) and RCO model (Definition 3.3). Let us denote $\mathbf{X}_k = \begin{pmatrix} \mathbf{y}_k \\ \mathbf{z}_k \end{pmatrix}$. Then, the hcoRNN and RCO models can be defined by the input-driven state-update equation $\mathbf{X}_{k+1} = G(\mathbf{X}_k, \mathbf{u}_{k+1})$, where $G : \mathbb{R}^{2N} \times \mathbb{R}^{I} \to \mathbb{R}^{2N}$ is defined by eqs. (13). The Jacobian of the G map computed on $(\mathbf{X}_k, \mathbf{u}_{k+1})$, denoted with $\mathbf{J}_k$, reads:

$$\mathbf{J}_k = \begin{bmatrix} \frac{\partial \mathbf{y}_{k+1}}{\partial \mathbf{y}_k} & \frac{\partial \mathbf{y}_{k+1}}{\partial \mathbf{z}_k} \\ \frac{\partial \mathbf{z}_{k+1}}{\partial \mathbf{y}_k} & \frac{\partial \mathbf{z}_{k+1}}{\partial \mathbf{z}_k} \end{bmatrix} = \begin{bmatrix} \mathbf{I} + \tau^2 \mathbf{A}_k & \tau\big(\mathbf{I} - \tau \mathrm{diag}(\boldsymbol{\varepsilon})\big) \\ \tau \mathbf{A}_k & \mathbf{I} - \tau \mathrm{diag}(\boldsymbol{\varepsilon}) \end{bmatrix}, \tag{16}$$

where

$$\begin{aligned} \mathbf{A}_k =& \mathbf{S}_k \mathbf{W} - \operatorname{diag}(\boldsymbol{\gamma}), & (17)\\ \mathbf{S}_k =& \operatorname{diag}\big(1 - \tanh^2(\mathbf{W}\mathbf{y}_k + \mathbf{V}\mathbf{u}_{k+1} + \mathbf{b})\big). & (18) \end{aligned}$$

In the following theorem we provide an upper bound for the Euclidean norm of the Jacobian. Let us define the following quantities

$$\begin{aligned} \xi =& \max_j |1 - \tau \varepsilon_j|, & (19)\\ \eta =& \max_j |1 - \tau^2 \gamma_j|, & (20)\\ \sigma =& ||\mathbf{W}||. & (21) \end{aligned}$$

Theorem 4.1. *The norm of the Jacobian matrix of the hcoRNN and RCO models admits the following upper bound*

$$||\mathbf{J}_k|| \leq \max(\eta + \tau^2 \sigma, \xi) + \tau \max(\xi, \gamma_{max} + \sigma). \quad (22)$$

In particular, for $\tau \ll 1$, and assuming $\varepsilon_{min} > 0$, and $\gamma_{max} \geq 1$, the bound reads

$$1 + \tau(\gamma_{max} + \sigma) + O(\tau^2). \quad (23)$$

The proof of Theorem 4.1 can be found in appendix B. A widely known stability condition, sufficient for the ESP to hold, is given by imposing that the Jacobian of eq. (16) is a contraction. One way to ensure this condition is to impose the Euclidean norm of the Jacobian to be uniformly less than 1, i.e. eq. (3). As can be deduced from Theorem 4.1, the RCO model results disinclined to this strong condition of stability, even for very small values of τ. The linearised RCO model is indeed strongly biased towards the identity for small values of τ. Although, the entire spectrum can be uniformly bounded around a neighbourhood of the identity by means of τ, it is an hard task to find combinations of hyperparameters ensuring that $||\mathbf{J}_k|| < 1$, and so ensuring the ESP for the RCO model. One interesting example is given by the particular case of $\boldsymbol{\varepsilon} \equiv \frac{1}{\tau}$, where the Jacobian of eq. (16) becomes $\mathbf{J}_k = \begin{bmatrix} \mathbf{I} + \tau^2 \mathbf{A}_k & \mathbf{0} \\ \tau \mathbf{A}_k & \mathbf{0} \end{bmatrix}$, reflecting the decoupling of the variable $\mathbf{y}$ from the variable $\mathbf{z}$, as discussed in section 3.4. In such a case, the $\mathbf{y}$-dynamics (and so the $\mathbf{z}$-dynamics) are contracting whenever $||\mathbf{I} + \tau^2 \mathbf{A}_k|| < 1$, i.e. whenever $||\mathbf{W}|| < \gamma_{\text{min}}$, for $\tau^2(\gamma_{\text{min}} + \gamma_{\text{max}}) \leq 2$. We prove this fact within Appendix C, which is dedicated to the particular case of $\boldsymbol{\varepsilon} \equiv \frac{1}{\tau}$. In the general case, imposing the upper bound of Theorem 4.1 to be less than 1, we can obtain sufficient conditions for the asymptotic uniform stability of the RCO model. We summarise in the proposition below a scheme of sufficient conditions to impose to have a contractive RCO, thus an uniformly asymptotically stable RCO in particular.

Proposition 4.2. ***Sufficient conditions.*** *If $\frac{\xi - \eta}{\tau^2} \leq \xi - \gamma_{max}$ then the hcoRNN and RCO models are asymptotically uniformly stable whenever one of the following three conditions hold:*

- $\sigma \leq \frac{\xi - \eta}{\tau^2}$, *and* $\xi < \frac{1}{1 + \tau}$,
- $\frac{\xi - \eta}{\tau^2} < \sigma \leq \xi - \gamma_{max}$, *and* $\sigma < \frac{1 - \tau\xi - \eta}{\tau^2}$,
- $\sigma \geq \xi - \gamma_{max}$, *and* $\sigma < \frac{1 - \eta - \tau\gamma_{max}}{\tau(1 + \tau)}$.

If $\xi - \gamma_{max} < \frac{\xi - \eta}{\tau^2}$ then the hcoRNN and RCO models are stable whenever one of the following three conditions hold:

- $\sigma \leq \xi - \gamma_{max}$, *and* $\xi < \frac{1}{1 + \tau}$,
- $\xi - \gamma_{max} < \sigma \leq \frac{\xi - \eta}{\tau^2}$, *and* $\sigma < \frac{1 - \xi}{\tau} - \gamma_{max}$,
- $\sigma \geq \frac{\xi - \eta}{\tau^2}$, *and* $\sigma < \frac{1 - \eta - \tau\gamma_{max}}{\tau(1 + \tau)}$.

The proof of Proposition 4.2 can be found in appendix D. The eigenspectrum for a combination of hyperparameters satisfying Proposition 4.2 is plotted in the upper left plot of Figure 2. The sufficient conditions that we found in Proposition 4.2 define a very narrow region of hyperparameters. The difficulty to satisfy these bounds reflects how closely the RCO model is biased around the identity. We might relax the request of contracting at each time step in favour of the less stringent requirement of having a spectral radius less than 1. Note however that, for a generic linear non-autonomous system, having a spectral radius less than 1 is not sufficient to imply asymptotic stability (Slotine et al., 1991; Kozachkov et al., 2022). In fact, there might be strong asymmetries promoting expanding dynamics in phase space that eventually lead to unstable behaviour. Here below, we introduce a variation of the RCO model that promotes fading memory via slightly pushing the eigenvalues towards the inside of the unit circle.

Definition 4.3. Fading RCO (F-RCO). An F-RCO is an RCO where stability is *promoted*, but not guaranteed, by an additional term. The state update of an F-RCO reads as follows:

$$\begin{aligned} \mathbf{y}_{k+1} =& \mathbf{y}_k + \tau \mathbf{z}_{k+1} - \tau \mathbf{y}_k, \\ \mathbf{z}_{k+1} =& \mathbf{z}_k + \tau \big(\tanh(\mathbf{W}\mathbf{y}_k + \mathbf{V}\mathbf{u}_{k+1} + \mathbf{b}) \quad (24)\\ & - \boldsymbol{\gamma} \odot \mathbf{y}_k - \boldsymbol{\varepsilon} \odot \mathbf{z}_k\big) - \tau \mathbf{z}_k. \end{aligned}$$

The last term of each equation ($-\tau\mathbf{y}_k, -\tau\mathbf{z}_k$) has the effect to push the eigenvalues slightly more towards the inside of the unitary circle (but not ensuring them to be inside). In this sense, we are enforcing a fading memory property without constraining the hyperparameters $\tau, \boldsymbol{\varepsilon}, \boldsymbol{\gamma}, \|\mathbf{W}\|$.

We provide a more precise picture of the eigenvalues distribution of the RCO model in the following theorem.

Theorem 4.4. *For all μ eigenvalues of the Jacobian of the hcoRNN and RCO models there exists a point $\lambda \in \{ 1 - \tau^2 \gamma_j \, , \, 1 - \tau \, \varepsilon_j \}_{j=1}^{N}$ such that*

$$|\mu - \lambda| \leq C, \tag{25}$$

where $C = \tau^2 \sigma + \tau \max\big(\xi, \gamma_{max} + \sigma\big)$.

The proof of Theorem 4.4 can be found in appendix E. According to Theorem 4.4, the eigenspectrum of the Jacobian of an RCO model is contained inside the union of disks of radius C centered on the points $1 - \tau\varepsilon_i, 1 - \tau^2\gamma_i$, see Figure 1 for a visual representation of this fact.

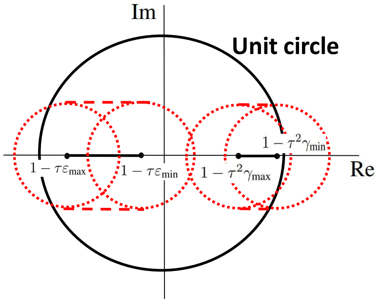

Figure 1. Depiction of the eigenspectrum's bound given by Theorem 4.4 for the Jacobian of an RCO model.

Theorem 4.4 allows us to impose conditions for the entire eigenspectrum of the RCO to be inside the unit circle via the inequality $\max(\xi, \eta) + C \leq 1$. However, it turns out that these conditions are almost equivalent to those derived in Proposition 4.2. Imposing such strict conditions on the RCO model might harm its expressiveness. For this reason, we rather consider simpler necessary conditions to impose to the RCO model for asymptotic stability. Specifically, we exploit the characterisation of the eigenspectrum given by Theorem 4.4 to derive more loose necessary conditions for linear asymptotic stability, by simply imposing that all the points $1 - \tau\varepsilon_i, 1 - \tau^2\gamma_i$, lie inside the unit circle. More precisely, we refer to the necessary conditions as stated in the following proposition.

Proposition 4.5. ***Necessary conditions.*** *If the hcoRNN and RCO models are asymptotically uniformly stable, then all the following hold true*

$$\varepsilon_{min} \geq 0, \tag{26}$$

$$\gamma_{min} \geq 0, \tag{27}$$

$$\tau\varepsilon_{max} \leq 2, \tag{28}$$

$$\tau^2\gamma_{max} \leq 2. \tag{29}$$

The proof of Proposition 4.5 can be found in appendix F. Although, the inequalities of eqs. (26)-(29) do not ensure linear asymptotic stability of an RCO model, following them as guidelines make sure to exclude RCO models that are undoubtedly linearly unstable. Moreover, selecting τ values small enough while satisfying eqs. (26)-(29) will generate typically RCO models with an underlying Jacobian just marginally unstable with eigenvalues at most slightly beyond the unitary circle in a neighbourhood of the value of 1, see Figure 2. Therefore, promoting the computation at the edge of stability.

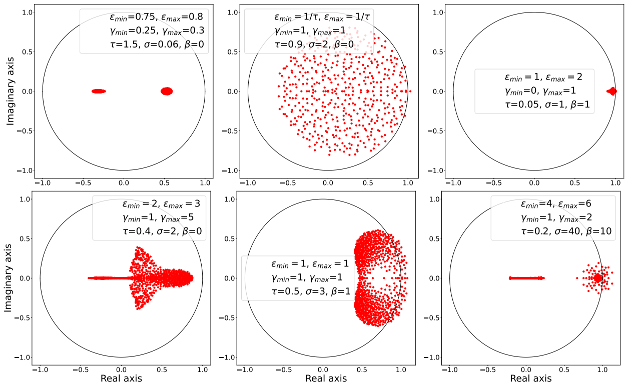

Figure 2. Plots of the eigenspectrum of the Jacobian eq. (16) for various combinations of hyperparameters. The bias vector inside the $\tanh$ has ben set to the value of β in all its components, while the input-to-hidden matrix $\mathbf{V}$ has been set to zero.

5. Experiments

Our empirical evaluation focuses on two key RCO properties, which have been the subject of the theoretical analysis of Section 3[1]:

1. We study the impact of weight *randomization* by comparing the performance of an RCO against fully-trained hcoRNN, coRNN and LSTM. We used both sequence classification and time series forecasting benchmarks.
2. We study the role played by the dynamical system *stability* in an RCO. To this end, we leverage the F-RCO model, which enforces stability, and we compare

[1] The code to reproduce the experiments can be found at https://github.com/AndreaCossu/RandomizedCoupledOscillators.

Table 1. Test Accuracy for sMNIST, psMNIST and npCIFAR-10 and test NRMSE for Lorenz96. Hidden size of LSTM and coRNN is 256, while hidden size for the other models is 362 (accounting for the lack of friction in the non-linearity). We did not experiment with hcoRNN on Lorenz96, since coRNN is already surpassed by an ESN.

Model	sMNIST ↑	psMNIST ↑	npCIFAR-10 ↑	Lorenz96 ↓
LSTM (Rusch & Mishra, 2023)	0.99	0.93	0.12	6.8×10^{-2}
coRNN (Rusch & Mishra, 2023)	0.99	0.97	0.59	9.8×10^{-2}
hcoRNN (our)	0.99	0.95	0.55	—
ESN	0.85	0.75	0.22	3.8×10^{-2}
RCO (our)	0.94	0.90	0.36	4.9×10^{-2}

Table 2. Number of adaptive parameters and total training time (in minutes) for each benchmark and model. Number of parameters include linear classifier/predictor. hcoRNN is trained for 120 epochs, as the original coRNN from Rusch & Mishra (2023).

Model	sMNIST	psMNIST	npCIFAR-10	Lorenz96
hcoRNN (our)	135,388 / 230M	135,388 / 230M	52,128 / 360M	—
ESN	3,620 / 2M	3,620 / 2M	1,810 / 1M	300 / 0.16M
RCO (our)	3,620 / 3M	3,620 / 3M	1,810 / 2M	300 / 0.17M

its performance against an RCO (possibly unstable) on the same benchmarks used in the previous point.

To guarantee a fair comparison, we adopt the experimental setup of Rusch & Mishra (2023), where the coRNN model was first introduced. We use the sequential MNIST (sMNIST), permuted sequential MNIST (psMNIST) and the noise padded CIFAR-10 (npCIFAR-10) as our sequence classification benchmarks. In such tasks, the recurrent model is required to show robust long-term memory capabilities. In sMNIST, the model observes one pixel at a time and it is required to predict the digit class after having processed all the 784 pixels. The psMNIST benchmark is the same as sMNIST, except that the pixels in an image are shuffled according to a fixed, random permutation. The npCIFAR-10 benchmark presents each RGB image of CIFAR-10 in a row-wise fashion (flattening the RGB channels into a single vector), leading to sequences of 32 elements. A randomly generated suffix is added to each sequence, reaching a final sequence length of 1,000 time steps. In npCIFAR-10, the information required to classify the image is contained in the very beginning of the sequence. Therefore, the model needs to extend its memory over hundreds of steps.

In their paper, Rusch & Mishra (2023) discussed how coRNN models are not tailored to time series forecasting for chaotic systems, due to their inability (by design) to generate chaotic dynamics. We know that randomly initialised models like ESN are usually very effective in predicting chaotic systems. Following Rusch & Mishra (2023), we ran experiments on the Lorenz96 system, in order to assess the effectiveness of RCO in chaotic systems forecasting.
The Lorenz96 system is defined by the following differential equation:

$$\dot{x}_i = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F, \tag{30}$$

with $i = 1, \ldots, 5$ and F an external driving force. We replicated the experiments in Rusch & Mishra (2023) with the Lorenz96 dynamical system, by choosing the chaotic regime with $F = 8$. The Lorenz96 task consists in predicting the next 25-th state of the system. The training, validation and test sets are composed by 128 trajectories of length 2000. Each trajectory is independently generated by solving eq. (30) with a random initial condition sampled uniformly from $[F - 0.5, F + 0.5]$ and a discretization time-step of 0.01. As commonly done for time series forecasting, we used an initial washout of length 200 (the first 200 steps are used to warm-up the model, but are not used when evaluating its performance). The performance of the models is measured by the Normalized Root Mean Squared Error (NRMSE), where normalization is performed by diving the RMSE by the root mean square of the target trajectory.

For the hcoRNN, we ran a model selection around the optimal hyper-parameters found by Rusch & Mishra (2023). Since we removed the adaptive parameters contained in $\mathcal{W}$, we increased the hidden size of the models accordingly, to match the total number of parameters of the original coRNN. For RCO, we ran a wider model selection, since we could not leverage any available results. We report all the details related to model selection for all models in Appendix A.

6. Results

Table 1 reports the results on the sMNIST, psMNIST, npCIFAR-10 and Lorenz96 benchmarks for LSTM, coRNN, hcoRNN and RCO. The results for the first two models are taken from Rusch & Mishra (2023).

Impact of randomization. Our RCO model surpasses the performance of an ESN in sMNIST, psMNIST and npCIFAR-10. As expected, fully-trained models like coRNN and hcoRNN achieve better results than RCO. The difference is larger for npCIFAR-10, which is the most difficult benchmark in terms of long-term memory requirements. However, RCO is able to outperform LSTM and ESN by a large margin in npCIFAR-10, showing that randomization does not completely ruin the excellent long-term memory capabilities present in coRNN models. In fact, on sMNIST and psMNIST the RCO model is able to achieve very good performance, reaching competitive accuracies with respect to fully-trained models. As showed by the Lorenz96 results in Table 1, RCO obtains a strong performance when modelling chaotic dynamical systems. While ESNs are still able to achieve a lower NRMSE, RCO clearly surpasses both (h)coRNN and LSTM. Hence the intrinsic inability of the coRNN family models to generate chaotic dynamics is not as detrimental as in Rusch & Mishra (2023) for chaotic systems forecasting. The randomization of the recurrent component seems in fact beneficial. Moreover, we observed that increasing the number of units in the ESN and RCO quickly reduces the NRMSE. Even 500 units are sufficient to get a NRMSE of 2.5×10^{-2} for ESN and 3.3×10^{-2} for RCO.

RCO stability. We verified whether or not the best RCO configurations (Appendix A) satisfy the *necessary* conditions of Proposition 4.5. Although we cannot state that the RCO is truly stable (Proposition 4.5 provides necessary, but not sufficient conditions for stability), we can observe from Table 3 that the best RCO in our experiments behave similarly to a stable RCO, since the conditions are satisfied in most cases. We then compared the performance of our best RCO with the Stable-RCO, that is an RCO that satisfies the *sufficient* conditions of Proposition 4.2. We also leveraged the F-RCO model, where stability is enforced but not guaranteed. Table 4 shows that the Stable-RCO is overly restrictive and does not allow to learn properly any of the time series tasks. The F-RCO model performs better than Stable-RCO, but it still does not match the best, unconstrained RCO. This supports that RCO needs to possess some degree of instability in order to tackle our time series tasks.

Computational efficiency. The number of adaptive parameters is two orders of magnitude smaller in RCO than in fully-trained models (Table 2). We already noticed that this comes at the cost of a compromise, which is however often favorable: an RCO exhibits less robust long-term memory capabilities, but it is able to outperform (h)coRNN in time series forecasting. Moreover, the small number of adaptive parameters makes the RCO a much more efficient model, both in terms of memory and computational time (Table 2). In fact, training time for (h)coRNNs scales as $O(N^2L)$ (back-propagation through time), where N is the recurrent matrix dimension and L is the length of the input sequence. On the contrary, RCO can be trained very quickly in $O(NL)$ with closed-form solutions that do not require back-propagation through time (e.g., least-mean squares). The lower training time for RCO allow to explore more configuration than in fully-trained models. This appears to be crucial, since the (h)coRNN is quite sensitive to the choice of its hyper-parameters, in particular τ.

Table 3. Validity of conditions of Proposition 4.5 for best RCO configurations from Table 1. The value of the best hyper-parameters can be found in Appendix A.

	sMNIST	psMNIST	npCIFAR-10	Lorenz96
$\epsilon_{\text{MIN}} \geq 0$	×	×	✓	✓
$\gamma_{\text{MIN}} \geq 0$	✓	✓	✓	✓
$\tau\epsilon_{\text{MAX}} \leq 2$	✓	✓	✓	✓
$\tau^2\gamma_{\text{MAX}} \leq 2$	✓	✓	✓	✓

Table 4. Test Accuracy for sMNIST, psMNIST and npCIFAR-10 and test NRMSE for Lorenz96. Stable-RCO refers to an RCO configuration that satisfies the *sufficient* conditions of Proposition 4.2.

	Stable-RCO	F-RCO	RCO
sMNIST ↑	0.15	0.88	0.94
psMNIST ↑	0.22	0.85	0.90
npCIFAR-10 ↑	0.10	0.26	0.36
Lorenz96 ↓	3.9×10^{-1}	6.1×10^{-2}	4.9×10^{-2}

7. Conclusion and future works

We developed a theoretical and empirical analysis of recurrent dynamical systems based on randomly coupled oscillators. We introduced the RCO model, which endows the coRNN model (Rusch & Mishra, 2023) with randomization properties. We provided a theoretical analysis of RCO and derived both necessary and sufficient conditions for its linear stability. We empirically evaluated RCO on a set of sequence classification and time series forecasting benchmarks. Our results show that RCO exhibits an effective long-term memory while greatly improving the performance in chaotic systems prediction. The computational efficiency of RCO in terms of training time and number of parameters is orders of magnitude better than the coRNN model.

Following the principles of deep RC, in the future we plan to study deep versions of RCO, where multiple layers of randomized oscillators are stacked together. This would lead to a more expressive model, able to learn richer latent representations. Due to its physically-inspired design, RCO could also be implemented in analogical and neuromorphic devices, which can mimic the behavior of coupled oscillators. We believe RCO to be a promising model for time series processing which provides an excellent trade-off between long-term memory capabilities and the ability to model nonlinear relationships.

Acknowledgments

This work is supported by the EU EIC project EMERGE, grant number 101070918.

References

Bauer, F. L. and Fike, C. T. Norms and exclusion theorems. *Numerische Mathematik*, 2:137–141, 1960.

Bertschinger, N. and Natschläger, T. Real-time computation at the edge of chaos in recurrent neural networks. *Neural computation*, 16(7):1413–1436, 2004.

Ceni, A., Ashwin, P., Livi, L., and Postlethwaite, C. The echo index and multistability in input-driven recurrent neural networks. *Physica D: Nonlinear Phenomena*, 412: 132609, 2020.

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Funahashi, K.-i. and Nakamura, Y. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural networks*, 6(6):801–806, 1993.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645. Springer, 2016b.

Hornik, K., Stinchcombe, M., and White, H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural networks*, 3(5):551–560, 1990.

Jaeger, H. Adaptive nonlinear system identification with echo state networks. *Advances in neural information processing systems*, 15, 2002.

Jaeger, H., Lukoševičius, M., Popovici, D., and Siewert, U. Optimization and applications of echo state networks with leaky-integrator neurons. *Neural networks*, 20(3): 335–352, 2007.

Kozachkov, L., Ennis, M., and Slotine, J.-J. Rnns of rnns: Recursive construction of stable assemblies of recurrent neural networks. *Advances in Neural Information Processing Systems*, 35:30512–30527, 2022.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Legenstein, R. and Maass, W. What makes a dynamical system computationally powerful. *New directions in statistical signal processing: From systems to brain*, pp. 127–154, 2007.

Lu, Z., Hunt, B. R., and Ott, E. Attractor reconstruction by machine learning. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(6):061104, 2018.

Lukoševičius, M. and Jaeger, H. Reservoir computing approaches to recurrent neural network training. *Computer science review*, 3(3):127–149, 2009.

Lund, B. D. and Wang, T. Chatting about chatgpt: how may ai and gpt impact academia and libraries? *Library Hi Tech News*, 40(3):26–29, 2023.

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

Pikovsky, A., Rosenblum, M., and Kurths, J. Synchronization: a universal concept in nonlinear science, 2002.

Rusch, T. K. and Mishra, S. Coupled Oscillatory Recurrent Neural Network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=F3s69XzWOia`.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Slotine, J.-J. E., Li, W., et al. *Applied nonlinear control*, volume 199. Prentice hall Englewood Cliffs, NJ, 1991.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yildiz, I. B., Jaeger, H., and Kiebel, S. J. Re-visiting the echo state property. *Neural networks*, 35:1–9, 2012.

Table 5. Model selection configurations for sMNIST. α is the ESN leaky rate, ν is the input scaling, ρ is the spectral radius.

Model	Configuration
hcoRNN	$\epsilon = 4.7 \pm \{2, \mathbf{1}\}, \gamma = 2.7 \pm \{2, \mathbf{1}\}, \tau = \{0.42, \mathbf{0.042}\}$
ESN	$\alpha = \{1, 0.5, 0.1, 0.01, \mathbf{0.001}\}, \rho = \{900, 90, 9, \mathbf{0.999}, 0.99, 0.9\}, \nu = \{10, \mathbf{1}, 0.1\}$
RCO	$\tau = \{0.42, \mathbf{0.042}\}, \rho = \{900, 90, \mathbf{9}, 0.9\}, \nu = \{10, \mathbf{1}, 0.1\}, \epsilon = \{4.7, \mathbf{0.47}\} \pm \{2, \mathbf{1}\}, \gamma = \{\mathbf{2.7}, 0.27\} \pm \{\mathbf{2}, 1\}$

Table 6. Model selection configurations for psMNIST. α is the ESN leaky rate, ν is the input scaling, ρ is the spectral radius.

Model	Configuration
hcoRNN	$\epsilon = 8.0 \pm \{2, \mathbf{1}\}, \gamma = 0.4 \pm \{\mathbf{2}, 1\}, \tau = \{0.76, \mathbf{0.076}\}$
ESN	$\alpha = \{1, 0.5, 0.1, \mathbf{0.01}, 0.001\}, \rho = \{900, 90, 9, \mathbf{0.999}, 0.99, 0.9\}, \nu = \{10, 1, \mathbf{0.1}\}$
RCO	$\tau = \{0.76, \mathbf{0.076}\}, \rho = \{900, 90, 9, \mathbf{0.9}\}, \nu = \{10, \mathbf{1}, 0.1\}, \epsilon = \{8, \mathbf{0.8}\} \pm \{\mathbf{2}, 1\}, \gamma = \{\mathbf{4}, 0.4\} \pm \{\mathbf{2}, 1\}$

A. Model selection

We provide the complete experimental setup used for model selection. Table 5, 6, 7, 8 reports the grid search performed during model selection, with the best value in bold.
The number of units is 362 for sMNIST and psMNIST, 181 for npCIFAR-10 for hcoRNN, RCO and ESN. On Lorenz96, RCO and ESN use 300 units. The hcoRNN model has been trained for the same number of epochs (120) as the original coRNN model from Rusch & Mishra (2023). For hcoRNN the grid search was performed around the best hyper-parameters found by the original paper (Rusch & Mishra, 2023).

For the Stable-RCO model in Table 4, we used the following configuration: $\tau = 1.1, \gamma = 0.58 \pm 0.03, \epsilon = 0.77 \pm 0.10, \rho = 0.01$. These hyper-parameters satisfy the sufficient conditions of Proposition 4.2 but are overly restrictive for the model.

B. Proof of Theorem 4.1

We will make use of the following lemma.

Lemma B.1. *Let be given two square matrices* $\mathbf{M}, \mathbf{N}$ *of the same dimension. Then it holds that*

$$\textit{(i)} \quad \left|\left| \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{bmatrix} \right|\right| \leq \max(||\mathbf{M}||, ||\mathbf{N}||),$$

$$\textit{(ii)} \quad \left|\left| \begin{bmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{N} & \mathbf{0} \end{bmatrix} \right|\right| \leq \max(||\mathbf{M}||, ||\mathbf{N}||).$$

Proof. We notice that for any unitary vector $\mathbf{X} = \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix}$ it holds

$$\left|\left| \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{bmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} \right|\right|^2 = \|\mathbf{My}\|^2 + \|\mathbf{Nz}\|^2 \leq \max(\|\mathbf{My}\|, \|\mathbf{Nz}\|)^2,$$

from which it follows that $\left|\left| \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{bmatrix} \right|\right| \leq \max(||\mathbf{M}||, ||\mathbf{N}||)$. For the antidiagonal case, note that $\begin{bmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{N} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$, hence we have for any unitary vector $\mathbf{X} = \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix}$ that

$$\left|\left| \begin{bmatrix} \mathbf{0} & \mathbf{M} \\ \mathbf{N} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} \right|\right|^2 = \left|\left| \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} \right|\right|^2 \leq \max(\|\mathbf{M}\|, \|\mathbf{N}\|)^2 \left(\left|\left| \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} \right|\right|^2 \right) = \max(\|\mathbf{M}\|, \|\mathbf{N}\|)^2.$$

Table 7. Model selection configurations for npCIFAR-10. α is the ESN leaky rate, ν is the input scaling, ρ is the spectral radius.

Model	Configuration
hcoRNN	$\epsilon = 12.7 \pm \{2, \mathbf{1}\}, \gamma = 1.3 \pm \{2, \mathbf{1}\}, \tau = \{0.76, \mathbf{0.076}\}$
ESN	$\alpha = \{1, 0.5, 0.1, 0.01, \mathbf{0.001}\}, \rho = \{\mathbf{900}, 90, 9, 0.9\}, \nu = \{10, \mathbf{1}, 0.1\}$
RCO	$\tau = \{0.34, \mathbf{0.034}\}, \rho = \{900, 90, \mathbf{9}, 0.9\}, \nu = \{10, 1, \mathbf{0.1}\}, \epsilon = \{\mathbf{12.7}, 1.27\} \pm \{2, \mathbf{1}\}, \gamma = \{13, \mathbf{1.3}\} \pm \{2, \mathbf{1}\}$

Table 8. Model selection configurations for Lorenz96. α is the ESN leaky rate, ν is the input scaling, ρ is the spectral radius.

Model	Configuration
ESN	$\alpha = \{1, \mathbf{0.5}, 0.1\}, \rho = \{900, 90, 9, \mathbf{0.9}\}, \nu = \{10, 1, \mathbf{0.1}\}$
RCO	$\tau = \{1, 0.7, 0.5, \mathbf{0.17}, 0.1, 0.05, 0.01, 0.001\}, \rho = \{90, 9, 0.999, \mathbf{0.99}, 0.9\}, \nu = \{10, 1, \mathbf{0.1}\}$ $\epsilon = \{\mathbf{10}, 5, 2, 1\} \pm \{2, \mathbf{1}\}, \gamma = \{\mathbf{10}, 5, 2, 1\} \pm \{\mathbf{2}, 1\}$

□

Lemma B.1 allows us to prove Theorem 4.1 as follows.

Proof. We decompose the Jacobian of eq. (16), with $\mathcal{W} = \mathbf{0}$, in the sum of two matrices, one diagonal one anti-diagonal, as follows

$$\mathbf{J}_k = \left[\begin{array}{cc} \mathbf{I} - \tau^2 \text{diag}(\boldsymbol{\gamma}) + \tau^2 \mathbf{S}_k \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \tau \text{diag}(\boldsymbol{\varepsilon}) \end{array} \right] + \left[\begin{array}{cc} \mathbf{0} & \tau\big(\mathbf{I} - \tau \text{diag}(\boldsymbol{\varepsilon})\big) \\ \tau \mathbf{A}_k & \mathbf{0} \end{array} \right]. \tag{31}$$

The upper left block of the diagonal matrix in eq. (31) admits the following bound.

$$||\mathbf{I} - \tau^2 \text{diag}(\boldsymbol{\gamma}) + \tau^2 \mathbf{S}_k \mathbf{W}|| \leq ||\mathbf{I} - \tau^2 \text{diag}(\boldsymbol{\gamma})|| + \tau^2 ||\mathbf{S}_k \mathbf{W}|| \leq \max_j |1 - \tau^2 \gamma_j| + \tau^2 ||\mathbf{W}|| = \eta + \tau^2 \sigma, \tag{32}$$

where we used the triangle inequality, and the fact that $||\mathbf{S}_k \mathbf{W}|| \leq ||\mathbf{W}||$, see definition of $\mathbf{S}_k$ in eq. (18).

The bottom right block of the diagonal matrix in eq. (31), is itself diagonal, and admits the following exact estimation.

$$||\mathbf{I} - \tau \text{diag}(\boldsymbol{\varepsilon})|| = \max_j |1 - \tau \varepsilon_j| = \xi. \tag{33}$$

The upper right block of the diagonal matrix in eq. (31), is itself diagonal, and admits the following exact estimation.

$$||\tau(\mathbf{I} - \tau \text{diag}(\boldsymbol{\varepsilon}))|| = \tau \max_j |1 - \tau \varepsilon_j| = \tau \xi. \tag{34}$$

The bottom left block of the diagonal matrix in eq. (31) admits the following bound.

$$||\tau(\mathbf{S}_k \mathbf{W} - \text{diag}(\boldsymbol{\gamma}))|| \leq \tau(||\mathbf{S}_k \mathbf{W}|| + ||\text{diag}(\boldsymbol{\gamma})||) \leq \tau(||\mathbf{W}|| + \gamma_{\text{max}}) = \tau(\sigma + \gamma_{\text{max}}). \tag{35}$$

Putting together eqs. (32)-(35), and Lemma B.1, we obtain

$$||\mathbf{J}_k|| \leq \max(\eta + \tau^2 \sigma, \xi) + \tau \max(\xi, \sigma + \gamma_{\text{max}}),$$

which is the thesis.

In particular, for small enough values of τ we have that $\tau \varepsilon_{\text{max}} \leq 1$, and $\tau^2 \gamma_{\text{max}} \leq 1$, which in turns imply that $\xi = 1 - \tau \varepsilon_{\text{min}}$, and that $\eta = 1 - \tau^2 \gamma_{\text{min}}$, respectively. Furthermore, assuming that $\varepsilon_{\text{min}} > 0$, and $\gamma_{\text{max}} \geq 1$, we have that $\xi < 1 \leq \sigma + \gamma_{\text{max}}$. Therefore, the bound reads

$$\max(1 - \tau^2 \gamma_{\text{min}} + \tau^2 \sigma, 1 - \tau \varepsilon_{\text{min}}) + \tau(\sigma + \gamma_{\text{max}}).$$

Finally note that for $\tau \ll 1$, and $\varepsilon_{\text{min}} > 0$, we have that $1 - \tau \varepsilon_{\text{min}} \leq 1 - \tau^2 \gamma_{\text{min}} + \tau^2 \sigma$. Hence, the bound has the following expansion for small values of τ

$$1 + \tau(\gamma_{\text{max}} + \sigma) + O(\tau^2). \tag{36}$$

C. Contractivity for the particular case of $\varepsilon \equiv \frac{1}{\tau}$

We already noticed that for the particular case of $\varepsilon \equiv \frac{1}{\tau}$ the $\mathbf{z}$-dynamics in eqs. (13) become unidirectionally driven by the $\mathbf{y}$-dynamics. In such a case, we can focus only on the $\mathbf{y}$-dynamics which reads

$$\mathbf{y}_{k+1} = (\mathbf{I} - \tau^2 \mathrm{diag}(\boldsymbol{\gamma}))\mathbf{y}_k + \tau^2 \tanh(\mathbf{W}\mathbf{y}_k + \mathbf{V}\mathbf{u}_{k+1} + \mathbf{b}). \tag{37}$$

We provide the following sufficient conditions for contraction in the particular case of (37).

Proposition C.1. *In the particular case of $\varepsilon \equiv \frac{1}{\tau}$, the hcoRNN and RCO models are contractive whenever*

$$\textit{(i)}\ \ \sigma < \gamma_{min},\ \textit{if}\ \tau^2(\gamma_{min} + \gamma_{max}) \leq 2;$$

$$\textit{(ii)}\ \ \sigma < \frac{2 - \tau^2\gamma_{max}}{\tau^2},\ \textit{if}\ \tau^2(\gamma_{min} + \gamma_{max}) > 2.$$

Proof. The Jacobian of eq. (37) reads $\mathbf{J}_k = \mathbf{I} + \tau^2\mathbf{A}_k = (\mathbf{I} - \tau^2\mathrm{diag}(\boldsymbol{\gamma})) + \tau^2\mathbf{S}_k\mathbf{W}$. Therefore, it holds $||\mathbf{J}_k|| \leq ||\mathbf{I} - \tau^2\mathrm{diag}(\boldsymbol{\gamma})|| + ||\tau^2\mathbf{S}_k\mathbf{W}|| \leq \eta + \tau^2\sigma$. Thus, $||\mathbf{J}_k|| < 1$ holds whenever $\sigma < \frac{1-\eta}{\tau^2}$. Finally note that, due to Definition 20, there are two possibilities for η, either $\eta = 1 - \tau^2\gamma_{\mathrm{min}}$, if $\tau^2(\gamma_{\mathrm{min}} + \gamma_{\mathrm{max}}) \leq 2$, or $\eta = \tau^2\gamma_{\mathrm{max}} - 1$, if $\tau^2(\gamma_{\mathrm{min}} + \gamma_{\mathrm{max}}) > 2$. The first case implies the thesis of (i), while the second case implies the thesis of (ii). □

Note that, in order for (i) and (ii) to hold in Proposition C.1, two necessary conditions must hold, namely $\gamma_{\mathrm{min}} \geq 0$, and $\tau^2\gamma_{\mathrm{max}} \leq 2$.

D. Proof of Proposition 4.2

Recall the upper bound of the Jacobian found in Theorem 4.1, that we denote for the purpose of the proof as

$$c = \max(\eta + \tau^2\sigma, \xi) + \tau\max(\xi, \sigma + \gamma_{\mathrm{max}}). \tag{38}$$

The proof is divided in 4 cases.

CASE 1.
Assume that $\eta + \tau^2\sigma \leq \xi$ and $\gamma_{\mathrm{max}} + \sigma \leq \xi$. These assumptions hold if and only if $\sigma \leq \min(\frac{\xi - \eta}{\tau^2}, \xi - \gamma_{\mathrm{max}})$. If such assumptions are true, then the constant (38) reads $c = \xi + \tau\xi$. Therefore, by Theorem 4.1, the Jacobian has norm less than 1 whenever $\xi < \frac{1}{1+\tau}$.

CASE 2.
Assume that $\eta + \tau^2\sigma \geq \xi$ and $\gamma_{\mathrm{max}} + \sigma \leq \xi$. These assumptions hold if and only if $\frac{\xi - \eta}{\tau^2} \leq \sigma \leq \xi - \gamma_{\mathrm{max}}$. If such assumptions are true, then the constant (38) reads $c = \eta + \tau^2\sigma + \tau\xi$. Therefore, by Theorem 4.1, the Jacobian has norm less than 1 whenever $\sigma < \frac{1 - \tau\xi - \eta}{\tau^2}$.

CASE 3.
Assume that $\eta + \tau^2\sigma \leq \xi$ and $\gamma_{\mathrm{max}} + \sigma \geq \xi$. These assumptions hold if and only if $\xi - \gamma_{\mathrm{max}} \leq \sigma \leq \frac{\xi - \eta}{\tau^2}$. If such assumptions are true, then the constant (38) reads $c = \xi + \tau(\sigma + \gamma_{\mathrm{max}})$. Therefore, by Theorem 4.1, the Jacobian has norm less than 1 whenever $\sigma < \frac{1-\xi}{\tau} - \gamma_{\mathrm{max}}$.

CASE 4.
Assume that $\eta + \tau^2\sigma \geq \xi$ and $\gamma_{\mathrm{max}} + \sigma \geq \xi$. These assumptions hold if and only if $\sigma \geq \max(\xi - \gamma_{\mathrm{max}}, \frac{\xi - \eta}{\tau^2})$. If such

assumptions are true, then the constant (38) reads $c = \eta + \tau^2\sigma + \tau(\sigma + \gamma_{\text{max}})$. Therefore, by Theorem 4.1, the Jacobian has norm less than 1 whenever $\sigma < \dfrac{1 - \xi - \tau\gamma_{\text{max}}}{\tau(1+\tau)}$.

The statement of Proposition 4.2 organises these results depending on whether $\dfrac{\xi - \eta}{\tau^2} \leq \xi - \gamma_{\text{max}}$, or vice versa.

E. Proof of Theorem 4.4

The proof is a straightforward application of the Bauer-Fike's theorem (Bauer & Fike, 1960) that we report here for ease of comprehension.

Theorem E.1 (Bauer-Fike)**.** *Let* $\mathbf{D}$ *be a diagonalisable matrix, and let* $\mathbf{H}$ *be the eigenvector matrix such that* $\mathbf{D} = \mathbf{H}\mathbf{\Lambda}\mathbf{H}^{-1}$ *where* $\mathbf{\Lambda}$ *is the diagonal matrix of the eigenvalues of* $\mathbf{D}$. *Let* $\mathbf{E}$ *be an arbitrary matrix of the same dimension of* $\mathbf{D}$. *Then, for all* μ *eigenvalues of* $\mathbf{D} + \mathbf{E}$, *there exists an eigenvalue* λ *of* $\mathbf{D}$ *such that*

$$|\mu - \lambda| \leq \|\mathbf{H}\|\|\mathbf{H}^{-1}\|\|\mathbf{E}\|. \tag{39}$$

Let us denote

$$\mathbf{E}_k = \begin{bmatrix} \tau^2\mathbf{S}_k\mathbf{W} & \tau\big(\mathbf{I} - \tau\text{diag}(\boldsymbol{\varepsilon})\big) \\ \tau\mathbf{A}_k & \mathbf{0} \end{bmatrix}. \tag{40}$$

The norm of the matrix $\mathbf{E}_k$ can be bounded as stated in the following lemma.

Lemma E.2. *The matrix* $\mathbf{E}_k$ *admits the following upper bound*

$$||\mathbf{E}_k|| \leq C, \tag{41}$$

where C *is defined as follows*

$$C = \tau^2\sigma + \tau\max\big(\xi, \gamma_{max} + \sigma\big), \tag{42}$$

where ξ *is defined in eq.* (19)*, and* σ *is defined in eq.* (21)*.*

Proof. We decompose the matrix $\mathbf{E}_k$ in its diagonal and antidiagonal parts, and apply Lemma B.1 obtaining the thesis. □

Then, Theorem E.1 in combination with Lemma E.2 give us all the ingredients to prove Theorem 4.4, as follows.

Proof. We decompose the Jacobian of eq. (16) in the sum of two matrices as follows

$$\mathbf{J}_k = \begin{bmatrix} \mathbf{I} - \tau^2\text{diag}(\boldsymbol{\gamma}) & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \tau\text{diag}(\boldsymbol{\varepsilon}) \end{bmatrix} + \begin{bmatrix} \tau^2\mathbf{S}_k\mathbf{W} & \tau\big(\mathbf{I} - \tau\text{diag}(\boldsymbol{\varepsilon})\big) \\ \tau\mathbf{A}_k & \mathbf{0} \end{bmatrix}, \tag{43}$$

and apply the Bauer-Fike's theorem E.1, choosing $\mathbf{D} = \begin{bmatrix} \mathbf{I} - \tau^2\text{diag}(\boldsymbol{\gamma}) & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \tau\text{diag}(\boldsymbol{\varepsilon}) \end{bmatrix}$, and $\mathbf{E} = \mathbf{E}_k$ as defined in eq. (40). Noticing that $\mathbf{D}$ is already diagonalized, i.e. $\mathbf{D} = \mathbf{\Lambda}$, thus $\mathbf{H}$ is the identity matrix, and the eigenspectrum of $\mathbf{D}$ is the set of all the points $\{1 - \tau^2\gamma_j, 1 - \tau\varepsilon_j\}_j$. The norm of the matrix $\mathbf{E}_k$ is bounded with C as stated in Lemma E.2. □

F. Proof of Proposition 4.5

If the RCO is asymptotically stable, then there exists at least one point of the input-driven solution $(\mathbf{y}_k, \mathbf{z}_k)$ for which the linearised RCO evaluated on it has all the eigenvalues inside the unit circle. Otherwise, there is at each time step at least one expanding direction, which contradicts the asymptotic uniform stability hypothesis. Now, due to Theorem 4.4 it must hold, in particular, that $1 - \tau^2\gamma_i \leq 1$, for all i, which translates in the condition $\gamma_{\text{min}} \geq 0$. In fact, if the point $1 - \tau^2\gamma_i$ is outside the unit circle, then there exists a smaller $0 < \tilde{\tau} < \tau$ such that the same RCO with $\tilde{\tau}$ is not asymptotically stable.
A similar argument implies that $1 - \tau\varepsilon_i \leq 1$ must hold for all i, which translates in the condition $\varepsilon_{\text{min}} \geq 0$. On the other hand, it must also hold that $1 - \tau^2\gamma_i \geq -1$, for all i, which translates in the condition $\tau^2\gamma_{\text{max}} \leq 2$, and analogously, that $1 - \tau\varepsilon_i \geq -1$, for all i, which translates in the condition $\tau\varepsilon_{\text{max}} \leq 2$. □