Cognitive Computation manuscript No.
(will be inserted by the editor)

Echo State Property of Deep Reservoir Computing Networks

Claudio Gallicchio · Alessio Micheli

Received: date / Accepted: date

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

C. Gallicchio
Department of Computer Science, University of Pisa
Largo B. Pontecorvo, 3, 56127 Pisa. E-mail: gallicch@di.unipi.it

A. Micheli
Department of Computer Science, University of Pisa
Largo B. Pontecorvo, 3, 56127 Pisa. E-mail: micheli@di.unipi.it

Abstract
Background In the last years, the Reservoir Computing (RC) framework has emerged as a state of-the-art approach for efficient learning in temporal domains. Recently, within the RC context, deep Echo State Network (ESN) models have been proposed. Being composed of a stack of multiple non-linear reservoir layers, deep ESNs potentially allow to exploit the advantages of a hierarchical temporal feature representation at different levels of abstraction, at the same time preserving the training efficiency typical of the RC methodology.
Methods In this paper, we generalize to the case of deep architectures the fundamental RC conditions related to the Echo State Property (ESP), based on the study of stability and contractivity of the resulting dynamical system.
Results Besides providing a necessary condition and a sufficient condition for the ESP of layered RC networks, the results of our analysis provide also insights on the nature of the state dynamics in hierarchically organized recurrent models. In particular, we find out that by adding layers to a deep reservoir architecture, the regime of network's dynamics can only be driven towards (equally or) less stable behaviors. Moreover, our investigation shows the intrinsic ability of temporal dynamics differentiation at the different levels in a deep recurrent architecture, with higher layers in the stack characterized by less contractive dynamics. Such theoretical insights are further supported by experimental results that show the effect of layering in terms of a progressively increased short-term memory capacity of the recurrent models.
Conclusions We provide a basic theoretical tool for the construction and further studies of deep RC architectures.

Keywords Reservoir Computing · Deep Learning · Echo State Property · Stability Analysis · Contractivity

1 Introduction

The study of hierarchically organized deep neural network architectures represents a research topic in fast growth [26,3,8,14]. Despite the notable success in real-world applications with significant results in problems such as image processing [25], object detection [6] and handwritten digit recognition [7], there are several open challenges that motivate the research effort in this field [2]. One of the most prominent aspects is related to the huge computational resources typically required by training deep neural networks [3], so that the efficiency of training algorithms is an even more valuable element in this context. Another relevant research topic concerns the extension of the deep learning paradigm to sequential data processing, which opens the possibility to learn temporal representations at different levels of abstraction, thereby allowing to naturally model tasks characterized by a hierarchical organization of the temporal information, for instance in real-world application areas such as text, speech and language processing [37,32, 17,15,9]. Furthermore, hierarchical processing of temporal information has a remarkable biological plausibility as evidenced by studies in the field of neuroscience [23,45,13,33,34] and as also exploited in biologically inspired computational models targeting tasks related to perception and vision (see e.g. [36,1,39]).

In the area of learning in sequential/temporal domains, Reservoir Computing (RC) [27,46] represents a state-of-the-art paradigm for efficient design and training

of Recurrent Neural Networks (RNNs) [24]. Although the principles of RC have been instantiated in different ways in literature (see e.g. [42,28,40]), the reference RC model in the neuro-computing area is represented by the Echo State Network (ESN) [18,21]. An ESN implements a discrete-time dynamical system in which the evolution of state dynamics and the output computation are decoupled both architecturally and under the point of view of training. Specifically, a recurrent non-linear reservoir component is used to model the network dynamics and to provide a temporal context to the external input at each time step. The role of the reservoir is thereby to encode the input history into a (possibly rich and meaningful for the task) state representation that is then used by a linear non-recurrent readout component to compute the output. The extreme training efficiency of the ESN approach derives from the fact that only the readout is trained (typically by direct methods), while the reservoir is initialized under certain conditions and then is left untrained. A pivotal role in this context is played by the *Echo State Property* (ESP) [18], which characterizes valid ESN dynamics and that essentially says that the reservoir state should asymptotically depend only on the driving input signal (the state is an *echo* of the input), while the influence of initial conditions should progressively vanish with time. The ESP has been studied since the seminal works on ESNs, resulting in the definition of two conditions for the ESP to hold, one necessary and one sufficient, which are commonly used in ESN literature for reservoir initialization [18,27]. Moreover, in recent years, the conditions for the ESP have been further investigated and refined in successive contributions [5,49,30,47], witnessing an active and vibrant research interest in the direction of a deeper understanding of the RC networks' dynamics.

Recently, a study of layered deep RC architectures has been proposed in [11,12], with the introduction of the *deepESN* model. The study of deepESNs, comprising a hierarchy of multiple untrained reservoir layers, aims on the one hand to better understand the meaning of stacking RNN layers separately from learning aspects, and on the other hand to represent a starting point for the design of efficiently trained deep learning models for temporal data. The preliminary experimental analysis proposed in [11,12] has evidenced that it is possible to practically exploit the deep architectural design in conjunction with key reservoir hyper-parameters to enhance the time-scale differentiation of layers' dynamics. The results in [11,12] show that higher layers in the hierarchy can effectively develop progressively slower dynamics. Moreover, recent literature studies are addressing the impact of ESNs organized in layered architectures in terms of applications on benchmarks as well as on real world tasks. Specifically, in [20] an ESN-based hierarchical network has been proposed, in which each successive layer is trained to estimate the relevance of the information that is processed at the lower layer, for the final output computation. Being investigated in relation to the problem of temporal feature discovery at different scales, the architecture proposed in [20] showed promising results on a case study with synthetic data. The advantage of multi-layered RC architectures has also been pointed out on time-series benchmarks in the RC area [29] and on real world tasks in the area of speech processing using ad-hoc hierarchical RC settings [43,44]. Overall, such literature works witness the emergence of a growing application interest in this field and further exacerbates the urgency of a timely and theoretically rational support to the set up of hierarchical RC networks.

The goal of this paper is to introduce a theoretical ground to the study of deepESNs under a dynamical system point of view. In particular, we provide an

analysis of the aspects related to stability and contractivity of state dynamics in hierarchical reservoir architectures. Based on such investigations, the main result of this paper is to formulate a necessary condition and a sufficient condition for the ESP to hold in case of deepESNs. Moreover, as a further outcome, our analysis provides useful insights that contribute to explain the diversification of state behavior in recurrent reservoir layers organized in a stack. In line with the theoretical nature of this paper's contribution, the effect of layering on the resulting dynamical regime of deep RC networks is investigated also experimentally by assessing its impact on the short-term memory capacity (MC) task.

The rest of this paper is organized as follows. The deepESN model is described in Section 2, introducing the formalism and the notation adopted throughout the paper. In Section 3 we extend the analysis of shallow reservoir network dynamics to the case of deep reservoir architectures. First, we introduce the ESP in this context, then we study the stability of deepESNs (allowing us to state the necessary condition for the ESP of deepESN) and finally we investigate the conditions under which a deepESN implements contraction mappings (resulting in the sufficient condition for the ESP of deepESN). In Section 4 we provide examples of how the theoretical findings in Section 3 can be used to comprehend the state dynamics diversification at the different layers in stacked recurrent architectures. The experimental results on the MC task are reported in Section 5. Finally, conclusions are drawn in Section 6.

2 Deep Echo State Networks

From an architectural point of view, a deepESN consists in a stack of reservoir layers, where at each time step t the first layer is fed by the external input, while successive layers receive input from the output at step t of the previous layer in the stack. In the following, assuming (for the sake of simplicity) that all the reservoirs in the stack have the same number of units N_R, indicating by N_U the external input dimension and by N_L the number of layers, we denote by:

- $\mathbf{x}^{(i)}(t) \in \mathbb{R}^{N_R}$ the state of the reservoir of layer i at step t
- $\mathbf{u}(t) \in \mathbb{R}^{N_U}$ the external input at step t
- $\mathbf{W}_{in}^{(i)}$ the input-to-reservoir weight matrix for layer i (where $\mathbf{W}_{in}^{(i)} \in \mathbb{R}^{N_R \times N_U}$ for $i = 1$, and $\mathbf{W}_{in}^{(i)} \in \mathbb{R}^{N_R \times N_R}$ for $i > 1$)
- $\boldsymbol{\theta}^{(i)} \in \mathbb{R}^{N_R}$ the weight vector associated to the unitary input bias for layer i
- $\hat{\mathbf{W}}^{(i)} \in \mathbb{R}^{N_R \times N_R}$ the recurrent reservoir weight matrix for layer i.

Throughout the paper we shall assume that the considered input and reservoir state spaces are compact sets, although for the ease of notation we shall keep denoting them as real vector spaces, i.e. as $\mathbb{R}^{N_U}$ and $\mathbb{R}^{N_R}$, respectively. Moreover, we will also take into consideration the global state space of a deepESN, defined as the product of the state spaces of the N_L layers. Accordingly, we shall use the symbol $\mathbf{x}(t)$ to denote the global state of the deepESN at step t, taking into account the states of all the layers, i.e. $\mathbf{x}(t) = (\mathbf{x}^{(1)}(t), \mathbf{x}^{(2)}(t), \ldots, \mathbf{x}^{(N_L)}(t)) \in \mathbb{R}^{N_L N_R}$.

Based on the introduced notation, Figure 1 graphically shows the general architecture of a deepESN. As it can be observed, from an architectural perspective we can view a deep recurrent network as obtained by applying a set of constraints to a shallow fully-connected RNN. Specifically, a hierarchically organized recurrent

model presents connections (without time-delays) only from lower layers to higher layers in the stack. On the other hand, connections from higher to lower layers and connections from the input to layers higher than the first level are avoided.

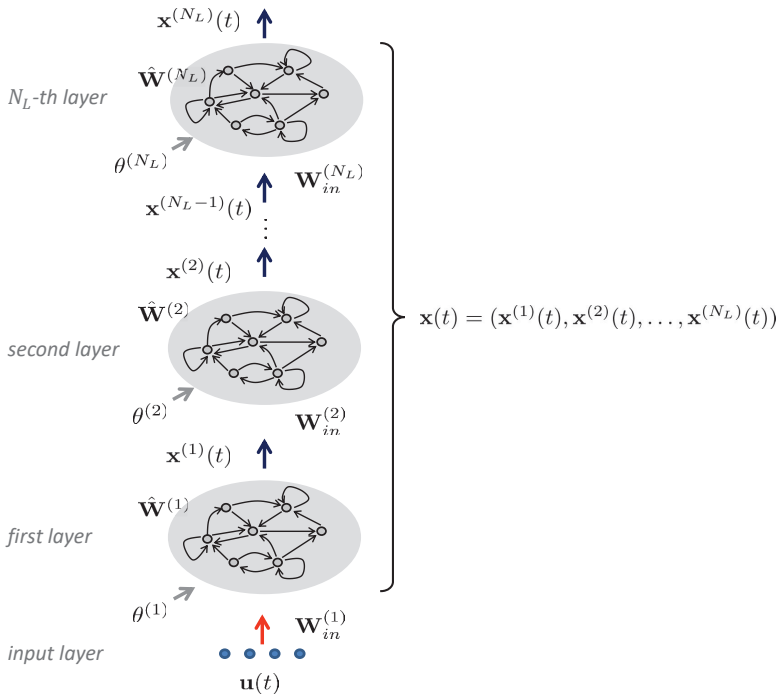

Fig. 1 The reservoir architecture of a deepESN with N_L layers.

Still under an architectural perspective, an interesting case that has been investigated in the RC literature consists in the organization of the reservoir into sub-groups (or ensembles) of recurrent units, which, at a first glance, can be seen also in the architecture in Figure 1. Relevant examples in this concern are represented by the work in [48] in which mechanisms of lateral inhibition among the different sub-groups are implemented to alleviate the problem of coupling among the reservoir units' activations, and by the work in [35] in which the decoupling effect is pursued by applying negatively correlation learning to the readout linked to each of the sub-groups. Differently from such approaches, the architecture of a deepESN is featured by a stacked recurrent architecture, whereas the sub-gropus have the role of layers without mechanisms to direct control their decoupling. The relevance of layering with respect to a sub-groups reservoir organization has been analyzed in terms of the inherent ability to develop richer state dynamics with multiple time-scales in [11, 12].

As in the shallow case, the reservoir component of a deepESN is used to encode the history of the received input into a state space that provides a rich representation of the input signal dynamics, which can then be exploited by a readout tool to learn the desired (temporal) task. As such, the process of training a deepESN is analogous to the case of training a shallow ESN, except for the fact that the input to the readout can take into account the state of the reservoirs at all the layers, e.g. by concatenating them in a predefined order. Besides this consideration, in the rest of this paper we will focus on the study of the state dynamics of a deepESN, generally putting aside the aspects related to the readout training.

From a dynamical systems point of view a deepESN implements an input driven discrete-time non-linear dynamical system, whose dynamics are governed by a function F

$$\begin{aligned} &F : \mathbb{R}^{N_U} \times \underbrace{\mathbb{R}^{N_R} \times \ldots \times \mathbb{R}^{N_R}}_{N_L} \rightarrow \underbrace{\mathbb{R}^{N_R} \times \ldots \times \mathbb{R}^{N_R}}_{N_L} \\ &\mathbf{x}(t) = F(\mathbf{u}(t), \mathbf{x}(t-1)) \end{aligned} \tag{1}$$

that maps the input at step t, i.e. $\mathbf{u}(t)$, and the set of reservoir states in the stack at step $t-1$, i.e. $\mathbf{x}(t-1)$, into the set of reservoir states at step t, i.e. $\mathbf{x}(t)$.

Considering the layer-wise nature of the deepESN dynamics, we can write F also as $F = (F^{(1)}, F^{(2)}, \ldots, F^{(N_L)})$, where for each $i = 1, \ldots, N_L$ the function $F^{(i)}$ describes the evolution of the state of the i-th layer, i.e. how the state of layer i depends on the state of the whole network at the previous time step. In particular, the dynamics of the first layer are the same as in shallow ESNs. By taking into consideration leaky integrator reservoir units [22], the state update equation for the first layer is given by:

$$\begin{aligned} &F^{(1)} : \mathbb{R}^{N_U} \times \mathbb{R}^{N_R} \rightarrow \mathbb{R}^{N_R} \\ &\mathbf{x}^{(1)}(t) = F^{(1)}(\mathbf{u}(t), \mathbf{x}^{(1)}(t-1)) \\ &\qquad = (1 - a^{(1)})\mathbf{x}^{(1)}(t-1) + a^{(1)} \tanh(\mathbf{W}_{in}^{(1)}\mathbf{u}(t) + \boldsymbol{\theta}^{(1)} + \hat{\mathbf{W}}^{(1)}\mathbf{x}^{(1)}(t-1)). \end{aligned} \tag{2}$$

where $a^{(1)} \in [0, 1]$ denotes the leaking rate parameter of the first layer.
For the second layer, the state at step t depends on $\mathbf{x}^{(1)}(t)$, i.e. a function of $\mathbf{u}(t)$ and $\mathbf{x}^{(1)}(t-1)$, and on $\mathbf{x}^{(2)}(t-1)$. Thereby the evolution of the network dynamics at the second layer can be described as $\mathbf{x}^{(2)}(t) = F^{(2)}(\mathbf{u}(t), \mathbf{x}^{(1)}(t-1), \mathbf{x}^{(2)}(t-1))$. In general, for $i > 1$, the state of layer i at step t, i.e. $\mathbf{x}^{(i)}(t)$, depends on the states of the layers $1, 2, \ldots, i$ at step $t-1$, i.e. $\mathbf{x}^{(1)}(t-1), \mathbf{x}^{(2)}(t-1), \ldots, \mathbf{x}^{(i)}(t-1)$,

according to the following state transition function $F^{(i)}$:

$$
\begin{aligned}
&F^{(i)} : \mathbb{R}^{N_U} \times \underbrace{\mathbb{R}^{N_R} \times \ldots \times \mathbb{R}^{N_R}}_{i} \to \mathbb{R}^{N_R} \quad (\text{with } \; i > 1) \\
&\mathbf{x}^{(i)}(t) = F^{(i)}(\mathbf{u}(t), \mathbf{x}^{(1)}(t-1), \mathbf{x}^{(2)}(t-1), \ldots, \mathbf{x}^{(i)}(t-1)) \\
&\quad = (1 - a^{(i)})\mathbf{x}^{(i)}(t-1) + a^{(i)} \tanh(\mathbf{W}_{in}^{(i)} \mathbf{x}^{(i-1)}(t) + \boldsymbol{\theta}^{(i)} + \hat{\mathbf{W}}^{(i)} \mathbf{x}^{(i)}(t-1)) \\
&\quad = (1 - a^{(i)})\mathbf{x}^{(i)}(t-1) + a^{(i)} \tanh(\mathbf{W}_{in}^{(i)} F^{(i-1)}(\mathbf{u}(t), \mathbf{x}^{(1)}(t-1), \\
&\qquad \mathbf{x}^{(2)}(t-1), \ldots, \mathbf{x}^{(i-1)}(t-1)) + \boldsymbol{\theta}^{(i)} + \hat{\mathbf{W}}^{(i)} \mathbf{x}^{(i)}(t-1)).
\end{aligned}
\tag{3}
$$

where $a^{(i)} \in [0, 1]$ is the leaking rate parameter of the i-th layer.

In the following we shall assume that the considered reservoir state spaces, both at individual layers level and at a global level, are endowed with a metric (i.e. a distance) function induced on the corresponding space by a norm, which we indicate by $\| \cdot \|$ in both cases for the ease of notation.

Remark 1 Notice that, without losing generality, the definitions, the derivations and the results presented in this paper are provided for the case of leaky integrator reservoir units. However, all the achieved results are still valid for standard *tanh* reservoir units, corresponding to the case in which the leaking rate parameter is equal to 1 for every layer, i.e. $a^{(i)} = 1$ for $i = 1, 2, \ldots, N_L$.

3 Analysis of Deep Echo State Network Dynamics

In this Section we extend the analysis of standard shallow ESN dynamics to the case of layered architectures, providing a necessary and a sufficient condition for the ESP to hold for a deepESN.

For shallow ESNs, the ESP represents a distinctive characterization of reservoir dynamics. Basically, it states that when the network is driven by a long input sequence, its state will asymptotically depend only on the input itself and the influence of initial conditions will be progressively forgotten. In [18], the ESP has been characterized by three equivalent conditions, namely being uniformly state contracting, state forgetting and input forgetting. Based on such definitions, in [10] it has been adopted a simple and useful definition for the ESP, which is extended here for the case of layered reservoir architectures.

To this aim, we use the notation $\hat{F}$ to denote the iterated version of the deepESN state transition function in equation 1. Given an input sequence of arbitrary finite length $\mathbf{s} \in (\mathbb{R}^{N_U})^*$, and the global state of the deepESN $\mathbf{x} \in \mathbb{R}^{N_L N_R}$, $\hat{F}(\mathbf{s}, \mathbf{x})$ is the global state of the network which has started in the initial state $\mathbf{x}$ and has

been driven by the sequence $\mathbf{s}$, i.e.

$$\hat{F} : (\mathbb{R}^{N_U})^* \times \underbrace{\mathbb{R}^{N_R} \times \ldots \times \mathbb{R}^{N_R}}_{N_L} \to \underbrace{\mathbb{R}^{N_R} \times \ldots \times \mathbb{R}^{N_R}}_{N_L}$$

$$\hat{F}(\mathbf{s}, \mathbf{x}) = \begin{cases} \mathbf{x} & if\ \mathbf{s} = [\,] \\ F(\mathbf{u}(N), \hat{F}([\mathbf{u}(1), \mathbf{u}(2), \ldots, \mathbf{u}(N-1)], \mathbf{x})) & if\ \mathbf{s} = [\mathbf{u}(1), \mathbf{u}(2), \ldots, \mathbf{u}(N)] \end{cases} \tag{4}$$

where $[\,]$ denotes the null sequence.

Accordingly, the ESP of a deepESN is expressed by the following definition.

Definition 1 (Echo State Property of deepESN)
Assume a deepESN whose global dynamics are ruled by a function F as in equation 1. Then the network has echo states if for each input sequence of length N, $\mathbf{s}_N = [\mathbf{u}(1), \mathbf{u}(2), \ldots, \mathbf{u}(N)]$, and for all couples of deepESN initial states $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{N_L N_R}$, the following condition holds:

$$\|\hat{F}(\mathbf{s}_N, \mathbf{x}) - \hat{F}(\mathbf{s}_N, \mathbf{x}')\| \to 0 \quad as \quad N \to \infty \tag{5}$$

In other words, the distance between the states in which the deepESN is driven after being fed by the same input sequence, but starting from different initial conditions, approaches 0 as the length of the input sequence goes to infinity.

Remark 2 Note that if the RC architecture contains only one layer, i.e. for $N_L = 1$ and with network's dynamics ruled by $F^{(1)}$, the ESP of deepESN in Definition 1 reduces to the literature case of ESP of a standard (shallow) ESN [18,10].

3.1 Stability of deepESN Dynamics

In this Section we provide a necessary condition for the ESP of deep RC networks, in Theorem 1, for the proof of which it is useful to introduce the following definitions and lemmas, which represent interesting results by themselves and that are therefore worth of being presented separately.

We introduce the study of stability of deepESN dynamics by investigating local first-order approximations of the dynamical system in equation 1. Specifically, the linearized version of the system in equation 1 around the state $\mathbf{x}_0 \in \mathbb{R}^{N_L N_R}$ can be described as follows:

$$\mathbf{x}(t) = \mathbf{J}_{F,\mathbf{x}}(\mathbf{u}(t), \mathbf{x}_0)\,(\mathbf{x}(t-1) - \mathbf{x}_0) + F(\mathbf{u}(t), \mathbf{x}_0) \tag{6}$$

where $\mathbf{J}_{F,\mathbf{x}}(\mathbf{u}(t), \mathbf{x}_0)$ is the Jacobian of the deepESN state update equation evaluated in $\mathbf{x}_0$ for external input $\mathbf{u}(t)$. By taking into consideration the layer-wise block organization of F, the Jacobian $\mathbf{J}_{F,\mathbf{x}}(\mathbf{u}(t), \mathbf{x}_0)$ can be written as a block

matrix:

$$\mathbf{J}_{F,\mathbf{x}}(\mathbf{u}(t),\mathbf{x}_0) =$$

$$\begin{pmatrix} \mathbf{J}_{F^{(1)},\mathbf{x}^{(1)}}(\mathbf{u}(t),\mathbf{x}_0) & \mathbf{J}_{F^{(1)},\mathbf{x}^{(2)}}(\mathbf{u}(t),\mathbf{x}_0) & \dots & \mathbf{J}_{F^{(1)},\mathbf{x}^{(N_L)}}(\mathbf{u}(t),\mathbf{x}_0) \\ \mathbf{J}_{F^{(2)},\mathbf{x}^{(1)}}(\mathbf{u}(t),\mathbf{x}_0) & \mathbf{J}_{F^{(2)},\mathbf{x}^{(2)}}(\mathbf{u}(t),\mathbf{x}_0) & \dots & \mathbf{J}_{F^{(2)},\mathbf{x}^{(N_L)}}(\mathbf{u}(t),\mathbf{x}_0) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{J}_{F^{(N_L)},\mathbf{x}^{(1)}}(\mathbf{u}(t),\mathbf{x}_0) & \mathbf{J}_{F^{(N_L)},\mathbf{x}^{(2)}}(\mathbf{u}(t),\mathbf{x}_0) & \dots & \mathbf{J}_{F^{(N_L)},\mathbf{x}^{(N_L)}}(\mathbf{u}(t),\mathbf{x}_0) \end{pmatrix} \tag{7}$$

where for every $i,j = 1,\dots,N_L$, $\mathbf{J}_{F^{(i)},\mathbf{x}^{(j)}}(\mathbf{u}(t),\mathbf{x}_0)$ denotes the partial derivative of $F^{(i)}$ with respect to $\mathbf{x}^{(j)}(t-1)$, evaluated at $\mathbf{x}_0$:

$$\mathbf{J}_{F^{(i)},\mathbf{x}^{(j)}}(\mathbf{u}(t),\mathbf{x}_0) = \frac{\partial F^{(i)}(\mathbf{u}(t),\mathbf{x}^{(1)}(t-1),\dots,\mathbf{x}^{(i)}(t-1))}{\partial \mathbf{x}^{(j)}(t-1)}\Big|_{\mathbf{x}=\mathbf{x}_0} \tag{8}$$

The spectral radius (i.e. the maximum among the eigenvalues in modulus) of the Jacobian in equation 7 has a relevant role in determining the stability/instability behavior of the linearized system in equation 6, as stated by the following Lemma 1. In this regard, we shall use the notation $\rho(\mathbf{J})$ to represent the spectral radius of a matrix $\mathbf{J}$.

Lemma 1 (Necessary condition for the stability of the linearized system around the zero state) *Consider the linearized system in equation 6 and assume a null input sequence as admissible input for the system. Then, a necessary condition for the stability of the system dynamics around the zero state is given by:*

$$\rho(\mathbf{J}_{F,\mathbf{x}}(\mathbf{0}_u,\mathbf{0})) < 1 \tag{9}$$

where $\mathbf{0}_u = [0\dots0]^T \in \mathbb{R}^{N_U}$ denotes the zero input at each pass of the null input sequence, and $\mathbf{0} = [0\dots0]^T \in \mathbb{R}^{N_L N_R}$ is the zero state of the deepESN.

Proof Assuming a constant zero input $\mathbf{0}_u \in \mathbb{R}^{N_U}$ for the system, based on equations 6, the first-order approximation of the deepESN dynamics around the zero state $\mathbf{0} \in \mathbb{R}^{N_L N_R}$ can be expressed as follows:

$$\mathbf{x}(t) = \mathbf{J}_{F,\mathbf{x}}(\mathbf{0}_u,\mathbf{0})\mathbf{x}(t-1), \tag{10}$$

from which it is easy to see that the zero state is a fixed point of the linearized system. This means that the trajectory of the system starting from $\mathbf{0}$ will never move from it. Moreover, the stability of the zero state as fixed point of equation 10 determines the asymptotic behavior of the trajectories starting from a state in a small neighborhood of $\mathbf{0}$. In our case, linear stability analysis tells us that this behavior depends on the spectral radius of $\mathbf{J}_{F,\mathbf{x}}(\mathbf{0}_u,\mathbf{0})$. Specifically, if $\rho(\mathbf{J}_{F,\mathbf{x}}(\mathbf{0}_u,\mathbf{0})) \geq 1$ then we cannot guarantee that the zero state is stable. A necessary condition for the stability of the linearized system around the zero state is therefore given by

$$\rho(\mathbf{J}_{F,\mathbf{x}}(\mathbf{0}_u,\mathbf{0})) < 1 \tag{11}$$

□

Remark 3 Lemma 1 tells us that if the condition in equation 9 is violated, then the linearized system in equation 6 exhibits instability around the zero state when driven by a null input sequence. Such result can be further extended to the non-linear system of equation 1, by observing [18,5] that if the underlying linear system is unstable around the zero state then the non-linear system obtained by passing the state values through a squashing non-linearity (such is tanh) will also show instability. Thereby, the condition in equation 9 can be also interpreted as a necessary condition for the stability of deepESN dynamics around the zero state.

At this point, in order to formulate a condition for the ESP, we need a way to compute the Jacobian $\rho(\mathbf{J}_{F,\mathbf{x}}(\mathbf{0}_u, \mathbf{0}))$, which is provided by the following Lemma 2.

Lemma 2 (Jacobian of deepESN state transition function) *Consider a deepESN whose dynamics are defined by equations 1, 2 and 3. Then the Jacobian of the deepESN state transition function for null input and zero state, i.e.* $\rho(\mathbf{J}_{F,\mathbf{x}}(\mathbf{0}_u, \mathbf{0}))$*, can be computed as follows:*

$$\rho(\mathbf{J}_{F,\mathbf{x}}(\mathbf{0}_u, \mathbf{0})) = \max_{k=1,2,\dots,N_L} \rho\left((1 - a^{(k)})\mathbf{I} + a^{(k)}\hat{\mathbf{W}}^{(k)}\right), \tag{12}$$

where $\mathbf{I} \in \mathbb{R}^{N_R \times N_R}$ *denotes the identity matrix.*

The proof is given in Appendix A.

Based on the results of Lemmas 1 and 2, we can state the following theorem, which provides a necessary condition for the ESP to hold in the case of a deepESN.

Theorem 1 (Necessary condition for the ESP of a deepESN) *Consider a deepESN whose dynamics is defined by equations 1, 2 and 3, and assume a null sequence as admissible input for the system. Then, a necessary condition for the ESP of the deepESN dynamics around the zero state is given by:*

$$\rho_g = \max_{k=1,2,\dots,N_L} \left(\rho((1 - a^{(k)})\mathbf{I} + a^{(k)}\hat{\mathbf{W}}^{(k)})\right) < 1, \tag{13}$$

where we call ρ_g *the global spectral radius of the deepESN.*

Proof As a consequence of Lemma 1 (see Remark 3), and by using the result of Lemma 2, we can see that a necessary condition for the stability of the deepESN dynamics around the zero state is given by

$$\rho(\mathbf{J}_{F,\mathbf{x}}(\mathbf{0}_u, \mathbf{0})) = \max_{k=1,2,\dots,N_L} \left(\rho((1 - a^{(k)})\mathbf{I} + a^{(k)}\hat{\mathbf{W}}^{(k)})\right) < 1. \tag{14}$$

By following a similar line of reasoning as in [18], we observe that if the zero state is not a stable fixed point of the deepESN state update equation 1, then there exist a state $\mathbf{x}'_0$ in the neighborhood of $\mathbf{0}$ such that, given the same null input sequence, the network starting from $\mathbf{x}'_0$ will not converge to $\mathbf{0}$ (i.e. it will follow a different trajectory than the one of the system starting in $\mathbf{0}$), and therefore the ESP condition in equation 5 is violated. This means that the equation 14 provides a necessary condition for the ESP of a deepESN.

□

Remark 4 Note that if the RC architecture contains only one layer, i.e. if $N_L = 1$, the necessary condition for the ESP of a deepESN in Theorem 1 reduces to the necessary condition for the ESP (around the zero state and for null input sequence) of standard (shallow) ESNs commonly reported in literature [18,27].

As in typical standard ESN applications, the necessary condition for the ESP provides a mean for deepESN initialization, consisting in setting the values of the leaking rate parameters and scaling the recurrent reservoir weight matrices in the layers of the network hierarchy in accordance to equation 13. However, a note of caution is due in interpreting the meaning of the condition for the ESP in Theorem 1, as it is actually valid only in a neighborhood of the zero state, while in more general cases the stability of network's dynamics depends also on the properties of the driving input signals. This aspect has been analyzed in recent literature on standard RC networks [30,49,47,4], with the aim of characterizing the state dynamics of (shallow) ESNs also as a function of the input. Although investigations of this kind are out of the scopes of this paper, we point out here that depending on the driving input signals, it could be possible that a deepESN exhibits stable dynamics even if the global spectral radius ρ_g violates the necessary condition in equation 13. Thereby, for practical applications of deepESNs, values of $\rho_g \geq 1$ should also be explored.

Remark 5 Finally, it is worth noticing that, as a consequence of Lemma 2, when we incrementally add new layers to a deepESN architecture, the Jacobian of the state update in equation 14 can never decrease. As such, the global spectral radius ρ_g of a deepESN is a monotonically non-decreasing function of the number of layers. Therefore, when we add new layers to a deepESN, the resulting network is characterized by an identical or less stable dynamical regime.

3.2 Contractivity of deepESN Dynamics

In standard RC networks, the notion of *contraction* mapping plays a fundamental role in the analysis of reservoir dynamics in relation to the sufficient condition for the ESP [10,18]. In order to apply such notion to the state evolution of a deepESN, we need to pursue the concept of metrics on the deepESN reservoir spaces mentioned in Section 2. Specifically, assuming that the reservoir spaces of the layers of a deepESN are equipped with a distance function, the metric of the global deepESN state space can be defined by resorting to the notion of metric on the product [31] of the layers' state spaces. In particular, in the following, we take into consideration the distance induced by the L_2-norm (i.e. the Euclidean distance) as metric for the layers' state spaces, and the maximum product metric as metric for the global deepESN state space. Accordingly, denoting the distance between two states at any layer k, $\mathbf{x}^{(k)}, \mathbf{x}'^{(k)} \in \mathbb{R}^{N_R}$ by $\|\mathbf{x}^{(k)} - \mathbf{x}'^{(k)}\|$, the distance between two global states $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N_L)}) \in \mathbb{R}^{N_L N_R}$ and $\mathbf{x}' = (\mathbf{x}'^{(1)}, \mathbf{x}'^{(2)}, \ldots, \mathbf{x}'^{(N_L)}) \in \mathbb{R}^{N_L N_R}$ is defined by [31]:

$$\|\mathbf{x} - \mathbf{x}'\| = \max_{k=1,2,\ldots,N_L} \|\mathbf{x}^{(k)} - \mathbf{x}'^{(k)}\|. \tag{15}$$

Furthermore, note that the same distance defined in equation 15 can be used as metric for the space obtained by the product of the reservoir state spaces

for any number of layers. Namely, for every $i > 0$, and for every $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}$, $\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)} \in \mathbb{R}^{N_R}$, the distance between $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)})$ and $(\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})$ is computed as $\|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}) - (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})\| = \max_{k=1,2,\ldots,i} \|\mathbf{x}^{(k)} - \mathbf{x}'^{(k)}\|$.

Based on such considerations, we can provide definitions of contractivity for the state transition functions of a deepESN at a layer level, in Definition 2, and at a global level, in Definition 3.

Definition 2 (Contractivity of layers' state transition function)
Assume a deepESN whose state dynamics at layer i are described by the state transition function $F^{(i)}$, defined as in equation 2 for $i = 1$, and as in equation 3 for $i > 1$. The function $F^{(i)}$ implements a contraction with respect to the state space if there exist a coefficient $C^{(i)} \in \mathbb{R}$, with $0 \leq C^{(i)} < 1$, such that:
(a) if $i = 1$
$\forall \mathbf{u} \in \mathbb{R}^{N_U}$, $\forall \mathbf{x}^{(1)}, \mathbf{x}'^{(1)} \in \mathbb{R}^{N_R}$:

$$\|F^{(1)}(\mathbf{u}, \mathbf{x}^{(1)}) - F^{(1)}(\mathbf{u}, \mathbf{x}'^{(1)})\| \leq C^{(1)} \|\mathbf{x}^{(1)} - \mathbf{x}'^{(1)}\| \tag{16}$$

(b) if $i > 1$
$\forall \mathbf{u} \in \mathbb{R}^{N_U}$, $\forall \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)} \in \mathbb{R}^{N_R}$:

$$\begin{aligned} &\|F^{(i)}(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}) - F^{(i)}(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})\| \leq \\ &C^{(i)} \|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}) - (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})\| \end{aligned} \tag{17}$$

In other words, $F^{(i)}$ is Lipschitz continuous (with respect to the state space) with a Lipschitz constant $C^{(i)} < 1$, which we call *contraction coefficient* of $F^{(i)}$. In this case we also say that the dynamics of the deepESN at layer i is contractive. Moreover, note that higher values of the contraction coefficients $C^{(i)}$ lead to less contractive dynamics.

Definition 3 (Contractivity of global state transition function)
Assume a deepESN whose state dynamics are described by the state transition function F in equation 1. The function F implements a contraction with respect to the state space if there exists a coefficient $C \in \mathbb{R}$, with $0 \leq C < 1$, such that $\forall \mathbf{u} \in \mathbb{R}^{N_U}$, $\forall \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_L)}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(N_L)} \in \mathbb{R}^{N_R}$:

$$\begin{aligned} &\|F(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_L)}) - F(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(N_L)})\| \leq \\ &C \|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_L)}) - (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(N_L)})\| \end{aligned} \tag{18}$$

In other words, F is Lipschitz continuous (with respect to the state space) with a Lipschitz constant $C < 1$, which we also call *contraction coefficient* of F. In this case we also say that the global deepESN reservoir dynamics is contractive. Note that higher values of the contraction coefficient C lead to less contractive network dynamics.

Grounded in Definitions 2 and 3, the following lemmas provide sufficient conditions for obtaining deepESN state transition functions implementing contraction mappings.

Lemma 3 (Sufficient condition for contractivity of layer's dynamics) *Consider a deepESN in which the state dynamics at layer i are described by the state transition function $F^{(i)}$, defined as in equation 2 for $i = 1$, and as in equation 3 for $i > 1$. Then a sufficient condition for the contractivity of the dynamics of the reservoir at layer i is given by:*
(a) if $i = 1$

$$C^{(1)} = (1 - a^{(1)}) + a^{(1)} \|\hat{\mathbf{W}}^{(1)}\| < 1 \tag{19}$$

(b) if $i > 1$, and assuming $F^{(i-1)}$ is contractive with coefficient $C^{(i-1)} < 1$

$$C^{(i)} = (1 - a^{(i)}) + a^{(i)} \left(C^{(i-1)} \|\mathbf{W}_{in}^{(i)}\| + \|\hat{\mathbf{W}}^{(i)}\| \right) < 1 \tag{20}$$

The proof is provided in Appendix B.

It is worth to observe that the sufficient conditions in Lemma 3, besides representing the ground for the following results reported in this Section, provide as corollary interesting insights on the effective diversification of the temporal dynamics emerging in the layers of a hierarchically organized RC network. This aspect is addressed and discussed through an illustrative example in Section 4.

Lemma 4 (Sufficient condition for contractivity of global dynamics) *Consider a deepESN in which the state dynamics is defined by means of equations 1, 2 and 3. Assume that for every $i = 1, \ldots, N_L$ the dynamics at layer i is contractive (according to Definition 2) with a Lipschitz constant $C^{(i)}$ identified in Lemma 3, then a sufficient condition for the contractivity of the global deepESN reservoir dynamics is given by:*

$$C = \max_{k=1,2,\ldots,N_L} \left(C^{(k)} \right) < 1 \tag{21}$$

Proof $\forall \mathbf{u} \in \mathbb{R}^{N_U}$ and $\forall \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_L)}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(N_L)} \in \mathbb{R}^{N_R}$ it results that:

$$\begin{aligned}
&\|F(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_L)}) - F(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(N_L)})\| = \\
&\|(F^{(1)}(\mathbf{u}, \mathbf{x}^{(1)}), F^{(2)}(\mathbf{u}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}), \ldots, F^{(N_L)}(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_L)})) - \\
&\quad (F^{(1)}(\mathbf{u}, \mathbf{x}'^{(1)}), F^{(2)}(\mathbf{u}, \mathbf{x}'^{(1)}, \mathbf{x}'^{(2)}), \ldots, F^{(N_L)}(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(N_L)}))\| = \\
&\max_{k=1,2,\ldots,N_L} \left(\|F^{(k)}(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}) - F^{(k)}(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(k)})\| \right) \leq \\
&\max_{k=1,2,\ldots,N_L} \left(C^{(k)} \|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}) - (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(k)})\| \right) \leq \\
&\max_{k=1,2,\ldots,N_L} \left(C^{(k)} \right) \ \|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_L)}) - (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(N_L)})\|
\end{aligned} \tag{22}$$

from which it follows that $C = \max_{k=1,2,\ldots,N_L} \left(C^{(k)} \right)$ is a Lipschitz constant of F, thus if $C < 1$, F is a contraction according to Definition 3.
□

The condition of contractivity of global dynamics in Lemma 4 allows to ensure the ESP of deepESNs, as stated in the following Theorem 2.

Theorem 2 (Sufficient condition for the ESP of a deepESN) *Consider a deepESN whose dynamics is defined by equations 1, 2 and 3. Assume that the deepESN is characterized by globally contractive dynamics according to the conditions in Lemma 4, with a Lipschitz constant $C < 1$, and that the reservoir state space at every layer is bounded with diameter D. Then the deepESN satisfies the ESP.*

The proof is given in Appendix C.

Remark 6 Note that if the considered RC architecture consists of only one layer, i.e. if $N_L = 1$, then the sufficient condition for the ESP of a deepESN in Theorem 2 reduces to the sufficient condition reported in literature for the ESP of a standard (shallow) ESN [18,27], i.e. ultimately to the condition in equation 19.

4 On the Diversification of Layers' Dynamics

In this Section we show how the results of our analysis on the contractivity of deepESN state transition functions can be used to investigate the intrinsic differentiation among the temporal dynamics developed by the layers of a stacked recurrent network.

Specifically, we consider a base deepESN architecture with N_L layers, in which the hyper-parameters of the model take the same values in all the layers, i.e. for every $i = 1, \ldots, N_L$, we have that $a^{(i)} = \alpha$, $\|\hat{\mathbf{W}}^{(i)}\| = \omega$ and $\|\mathbf{W}_{in}^{(i)}\|$ is fixed to 1. In this setting, referring to the outcomes of Lemma 3, we compute the values of the contraction coefficients $C^{(i)}$ of the different layers in the network hierarchy.

By solving the recurrence in equations 19 and 20 we obtain:

$$\begin{aligned} C^{(i)} &= 1 - \alpha^i + \omega \sum_{k=1}^{i} \alpha^k \\ &= \begin{cases} \omega\, i & if \quad \alpha = 1 \\ (1-\alpha^i)(1+\frac{\alpha\,\omega}{1-\alpha}) & if \quad \alpha < 1 \end{cases} \end{aligned} \tag{23}$$

from which we can see that $C^{(i)}$ is an increasing function of the layer number i, indicating that the contraction coefficient gradually increases in higher layers.

Here it is worth noticing that, in general, we cannot say that for each state transition function $F^{(i)}$ the value of $C^{(i)}$ found out in Lemma 3 is the smallest among its possible Lipschitz constants, and in this sense the contraction coefficients computed by applying equation 19, 20, and 23 actually represent upper bounds to the effective contractivity of the state dynamics. As such, beyond the numerical values of $C^{(i)}$, what is really interesting for the scopes of our analysis is the qualitative trend of differentiation among the layers. Furthermore, note that the contraction coefficient of a contractive dynamical system is related to its memory length, as pointed out in the context of RNNs and ESNs in [41,16,10]. Thereby, the differentiation of the values of the contraction coefficient across the layered deepESN architecture, as evidenced in a base deepESN setting by equation 23, implies a different extent of the memory length among the different layers, with higher layers in the stack being characterized by less contractive dynamics and longer memory.

A graphical representation of this characterization is reported in Figure 2. In particular, Figure 2(a) shows the values of the contraction coefficients computed by equation 23 for a 5-layers deepESN with $\alpha = 0.5$ and different values of ω in the range $0.1 - 0.9$, while Figure 2(b) corresponds to the case in which $\omega = 0.9$ and α varies in the range $0.1 - 1.0$. From Figure 2(a) it is possible to observe, with respect to the number of the layers, the clear increasing trend of the contraction coefficients, whose values scale with the value of ω. For what concerns the α parameter, in Figure 2(b) we can see how the increasing trend of the contraction coefficients, scaling with the value of α, has a different behavior depending on the particular value of α. In particular, while the growth is linear for $\alpha = 1$, a progressively stronger saturation effect is evident for values of α closer to 0.

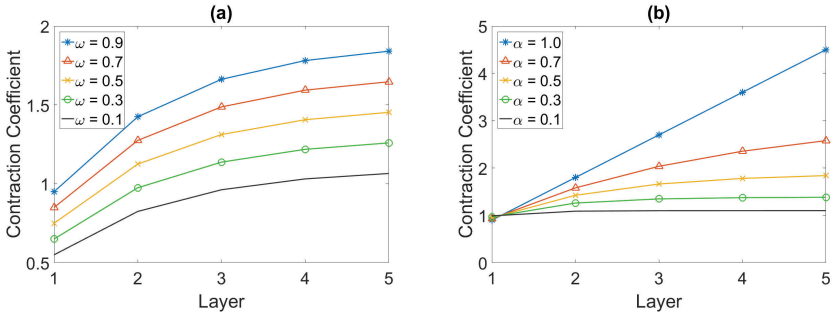

Fig. 2 Contraction coefficient of deepESN layers. **(a)**: $\alpha = 0.5$ and ω varying in $0.1 - 0.9$, **(b)**: $\omega = 0.9$ and α varying in $0.1 - 1.0$.

Finally, we can draw two main lines of observations. First, we can qualitatively characterize the inherent differentiation among the layers of a hierarchically organized recurrent network in relation to the occurrence of temporal dynamics with progressively longer memory length, with a saturation effect ruled by the values of network's hyper-parameters. Second, the growth of the contraction coefficient, obtained by increasing the number of layers in the architecture, explains the intrinsic increase of the memory length in deep recurrent networks as an architectural effect of the layering. In this concern, it is useful to recall that, on the basis of the result of Lemma 4, the maximum value of the contraction coefficient among the network's layers characterizes the contractivity of the global network dynamics, and thereby the extent of its memory length.

5 Short-term Memory Capacity

The relation between the memory length of recurrent networks' dynamics and the effect of layering is further investigated through the experimental assessment of the short-term memory of the hierarchically organized RC networks. To this end we considered the MC task [19], which consists in recovering delayed versions of a temporal input signal whose elements $u(t)$ are randomly drawn from a uniform distribution in $[-0.8, 0.8]$. The dataset contained a total number of 6000 time

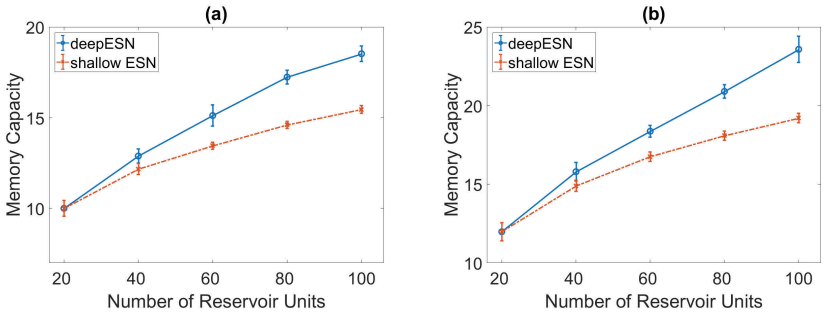

Fig. 3 Averaged MC values (and standard deviations) on the test set achieved by deepESN for increasing number of layers, plotted with respect to the total number of reservoir units (with 20 units per layer). For comparison, results achieved by shallow ESN under the same hyper-parametrization settings are reported as well. **(a)**: $\alpha = 0.5$ and $\omega = 0.9$, **(b)**: $\alpha = 1.0$ and $\omega = 0.9$.

steps, of which the first 5000 have been used for training and the remaining 1000 for test. Denoting by $y_d(t)$ the output of the readout unit trained to reconstruct the input signal with delay d, i.e. $u(t-d)$, the MC of the network is defined as

$$MC = \sum_{d=0}^{\infty} r^2(u(t-d), y_d(t)), \tag{24}$$

where $r^2(u(t-d), y_d(t))$ represents the squared correlation coefficient between $u(t-k)$ and $y_d(t)$.

In line with the analysis reported in Section 4, we adopted a base deepESN architecture in which the reservoir hyper-parameters have the same value in all the layers, i.e. $a^{(i)} = \alpha$, $\|\hat{\mathbf{W}}^{(i)}\| = \omega$ and $\|\mathbf{W}_{in}^{(i)}\| = \omega_{in}$ for every $i = 1, \ldots, N_L$. We considered deepESNs in which each layer contained $N_R = 20$ fully connected reservoir units, varying the number of layers N_L from 1 to 5, thereby varying the total number of reservoir units in the range $20 - 100$. By following a similar approach to e.g. [38], we practically implemented the MC task by using a finite number of delayed signals equal to 200, i.e. twice the maximum number of total reservoir units, and we used a value of $\omega_{in} = 0.1$. With the only scope of comparative analysis (and without aiming at reaching the best performance on the task), we conducted experiments with $\omega = 0.9$ and $\alpha \in \{0.5, 1\}$, and for each reservoir hyper-parametrization we independently instantiated a number of 10 reservoir guesses (for random initialization), over which the results have been averaged. For comparison, we also ran experiments with standard (shallow) ESNs, using the same hyper-parametrizations considered for the experiments with deepESNs, and organizing in this case the total number of reservoir units (variable in the range 20 - 100) into a single fully connected reservoir layer. Hence, the experiments considered five steps with a growing number of units/layers for the standard/deep cases. Note that, since the total number of reservoir units is the same for deepESN and standard ESN in each of the five steps, the same number of free parameters for the readout, and hence the same time complexity, are considered in the two cases.

The results achieved on the test set of the MC task are illustrated in Figure 3, respectively in correspondence of the setting with $\alpha = 0.5$ and $\omega = 0.9$ in

Figure 3(a), and with $\alpha = 1$ and $\omega = 0.9$ in Figure 3(b). Results clearly show that the MC of deepESN networks increases as the number of layer is increased. Remarkably, this characterization does not simply depend on the increasing number of reservoir units, but it is actually due to the layered organization of the recurrent architecture, as testified by the fact that the MC value of deepESN systematically outperforms the result achieved by shallow ESN with the same hyper-parametrization and total number of reservoir units. Moreover, the gap between the MC of deepESN and of shallow ESN tends to increase when a larger number of layers are considered. We can also note that results obtained in correspondence of the setting with $\alpha = 1$ are consistently better than those achieved with $\alpha = 0.5$. In particular, for the 5-layered deepESN setting the achieved MC value is 18.52 ± 0.43, for $\alpha = 0.5$, and 23.58 ± 0.83, for $\alpha = 1$.

Overall, the results presented in this Section represent an experimental evidence that confirms the observations made in Section 4, and practically show that the short-term memory of RC networks is actually amplified as an effect of layering, where the extent of such increase might be modulated by the values of the network's hyper-parameters (i.e. by the value of α in our case).

6 Conclusions

The study of deep recurrent models is still in its initial phases. In this paper we have proposed an analysis of the state dynamics in deep recurrent architectures, establishing fundamental conditions for the ESP of deepESNs that extend and generalize known results in standard RC literature to the case of layered RC networks. As such, the work in this paper yields a basic tool for concretely creating and studying deepESNs. In particular, through the study of the stability of deepESN dynamics we have provided a necessary condition for the ESP that is related to a global spectral radius of the system, which turned out to be a non-decreasing function of the number of layers in the architecture. Moreover, the analysis of the contractivity of deepESN dynamics allowed us to formulate a sufficient condition for the ESP, related to the value of key RC hyper-parameters.

As a further result of our investigation, we have provided insights on the inherent differentiation among the temporal dynamics developed in the layers of a hierarchically organized recurrent neural architecture. Already in a base deepESN setting, in which hyper-parameters of the network take the same values in every layer, the state dynamics are characterized by progressively increasing contraction coefficients. This aspect contributes to explain the intrinsic gain of the memory length in deeper recurrent networks, as also empirically shown through numerical simulations on the MC task.

A Proof of Lemma 2

Let us consider the Jacobian of deepESN state transition function in equation 1 evaluated at $\mathbf{u}(t)$ and $\mathbf{x}(t-1)$. From equation 8 we can easily see that for every $j > i$, we have that $\mathbf{J}_{F^{(i)},\mathbf{x}^{(j)}}(\mathbf{u}(t), \mathbf{x}(t-1))$ is a zero matrix, as the hierarchical structure of the deepESN architecture (see Figure 1 and equations 2 and 3) tells us that state update at layer i does not depend on the previous state of the system at higher layers in the stack (i.e. at layers j, with $j > i$). Thereby we can notice that $\mathbf{J}_{F,\mathbf{x}}(\mathbf{u}(t), \mathbf{x}(t-1))$ has the structure of a lower-triangular block

matrix. As such, the eigenvalues of $\mathbf{J}_{F,\mathbf{x}}(\mathbf{u}(t),\mathbf{x}(t-1))$ are the eigenvalues of the matrices on its block diagonal, i.e. the eigenvalues of $\mathbf{J}_{F^{(i)},\mathbf{x}^{(i)}}(\mathbf{u}(t),\mathbf{x}(t-1))$ for every $i = 1,2,\ldots,N_L$. Accordingly, we have that the spectral radius of $\mathbf{J}_{F,\mathbf{x}}(\mathbf{u}(t),\mathbf{x}(t-1))$ is the maximum among the spectral radii of its diagonal blocks, i.e.

$$\rho(\mathbf{J}_{F,\mathbf{x}}(\mathbf{u}(t),\mathbf{x}(t-1))) = \max_{k=1,2,\ldots,N_L} \left(\rho(\mathbf{J}_{F^{(k)},\mathbf{x}^{(k)}}(\mathbf{u}(t),\mathbf{x}(t-1))\right). \tag{25}$$

With the aim of computing the diagonal block matrices $\mathbf{J}_{F^{(k)},\mathbf{x}^{(k)}}(\mathbf{u}(t),\mathbf{x}(t-1))$, we observe that from equation 8 we have that

$$\mathbf{J}_{F^{(k)},\mathbf{x}^{(k)}}(\mathbf{u}(t),\mathbf{x}(t-1)) = \frac{\partial F^{(k)}(\mathbf{u}(t),\mathbf{x}^{(1)}(t-1),\ldots,\mathbf{x}^{(k)}(t-1))}{\partial \mathbf{x}^{(k)}(t-1)} =$$

$$\frac{\partial}{\partial \mathbf{x}^{(k)}(t-1)} \Big((1-a^{(k)})\mathbf{x}^{(k)}(t-1) + a^{(k)} \tanh(\mathbf{W}_{in}^{(k)} F^{(k-1)}(\mathbf{u}(t),\mathbf{x}^{(1)}(t-1),\ldots,\mathbf{x}^{(k-1)}(t-1)) +$$

$$\boldsymbol{\theta}^{(k)} + \hat{\mathbf{W}}^{(k)}\mathbf{x}^{(k)}(t-1))\Big) =$$

$$(1-a^{(k)})\mathbf{I} + a^{(k)} \begin{pmatrix} 1-(\tilde{x}_1^{(k)}(t))^2 & 0 & \ldots & 0 \\ 0 & 1-(\tilde{x}_2^{(k)}(t))^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1-(\tilde{x}_{N_R}^{(k)}(t))^2 \end{pmatrix} \hat{\mathbf{W}}^{(k)} \tag{26}$$

where for $j = 1,2,\ldots,N_R$, $\tilde{x}_j^{(k)}(t)$ are the elements of $\tilde{\mathbf{x}}^{(k)}(t) = \tanh(\mathbf{W}_{in}^{(k)} F^{(k-1)}(\mathbf{u}(t),\mathbf{x}^{(1)}(t-1),\mathbf{x}^{(2)}(t-1),\ldots,\mathbf{x}^{(k-1)}(t-1)) + \boldsymbol{\theta}^{(k)} + \hat{\mathbf{W}}^{(k)}\mathbf{x}^{(k)}(t-1))$.

Considering zero input and state, from equation 26, we can derive that for every $k = 1,2,\ldots,N_L$

$$\rho(\mathbf{J}_{F^{(k)},\mathbf{x}^{(k)}}(\mathbf{0}_u,\mathbf{0})) = \rho\left((1-a^{(k)})\mathbf{I} + a^{(k)}\hat{\mathbf{W}}^{(k)}\right) \tag{27}$$

and therefore equation 25 becomes

$$\rho(\mathbf{J}_{F,\mathbf{x}}(\mathbf{0}_u,\mathbf{0})) = \max_{k=1,2,\ldots,N_L} \rho\left((1-a^{(k)})\mathbf{I} + a^{(k)}\hat{\mathbf{W}}^{(k)}\right). \tag{28}$$

□

B Proof of Lemma 3

case (a): This case follows from the case of contractivity in standard shallow ESNs [10]. Indeed, $\forall \mathbf{u} \in \mathbb{R}^{N_U}$ and $\forall \mathbf{x}^{(1)}, \mathbf{x}'^{(1)} \in \mathbb{R}^{N_R}$

$$\begin{aligned}
&\|F^{(1)}(\mathbf{u},\mathbf{x}^{(1)}) - F^{(1)}(\mathbf{u},\mathbf{x}'^{(1)})\| = \\
&\|(1-a^{(1)})\mathbf{x}^{(1)} + a^{(1)}\tanh(\mathbf{W}_{in}^{(1)}\mathbf{u} + \boldsymbol{\theta}^{(1)} + \hat{\mathbf{W}}^{(1)}\mathbf{x}^{(1)}) - \\
&\quad (1-a^{(1)})\mathbf{x}'^{(1)} - a^{(1)}\tanh(\mathbf{W}_{in}^{(1)}\mathbf{u} + \boldsymbol{\theta}^{(1)} + \hat{\mathbf{W}}^{(1)}\mathbf{x}'^{(1)})\| = \\
&\|(1-a^{(1)})(\mathbf{x}^{(1)} - \mathbf{x}'^{(1)}) + a^{(1)}(\tanh(\mathbf{W}_{in}^{(1)}\mathbf{u} + \boldsymbol{\theta}^{(1)} + \hat{\mathbf{W}}^{(1)}\mathbf{x}^{(1)}) - \\
&\quad \tanh(\mathbf{W}_{in}^{(1)}\mathbf{u} + \boldsymbol{\theta}^{(1)} + \hat{\mathbf{W}}^{(1)}\mathbf{x}'^{(1)}))\| \leq \\
&(1-a^{(1)})\|\mathbf{x}^{(1)} - \mathbf{x}'^{(1)}\| + a^{(1)}\|\mathbf{W}_{in}^{(1)}\mathbf{u} + \boldsymbol{\theta}^{(1)} + \hat{\mathbf{W}}^{(1)}\mathbf{x}^{(1)} - \\
&\quad \mathbf{W}_{in}^{(1)}\mathbf{u} - \boldsymbol{\theta}^{(1)} - \hat{\mathbf{W}}^{(1)}\mathbf{x}'^{(1)}\| \leq \\
&(1-a^{(1)})\|\mathbf{x}^{(1)} - \mathbf{x}'^{(1)}\| + a^{(1)}\|\hat{\mathbf{W}}^{(1)}\|\|\mathbf{x}^{(1)} - \mathbf{x}'^{(1)}\| = \\
&\left((1-a^{(1)}) + a^{(1)}\|\hat{\mathbf{W}}^{(1)}\|\right)\|\mathbf{x}^{(1)} - \mathbf{x}'^{(1)}\|
\end{aligned} \tag{29}$$

from which it follows that $C^{(1)} = (1 - a^{(1)}) + a^{(1)}\|\hat{\mathbf{W}}^{(1)}\|$ is a Lipschitz constant for $F^{(1)}$. Thus if $C^{(1)} < 1$ then $F^{(1)}$ is a contraction (see Definition 2).

case (b): In this case, assuming $F^{(i-1)}$ is a contraction with a Lipschitz constant $C^{(i-1)} < 1$, $\forall \mathbf{u} \in \mathbb{R}^{N_U}$ and $\forall \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)} \in \mathbb{R}^{N_R}$

$$
\begin{aligned}
&\|F^{(i)}(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}) - F^{(i)}(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})\| = \\
&\|(1 - a^{(i)})\mathbf{x}^{(i)} + a^{(i)} \tanh(\mathbf{W}_{in}^{(i)} F^{(i-1)}(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i-1)}) + \boldsymbol{\theta}^{(i)} + \hat{\mathbf{W}}^{(i)}\mathbf{x}^{(i)}) - \\
&\quad (1 - a^{(i)})\mathbf{x}'^{(i)} - a^{(i)} \tanh(\mathbf{W}_{in}^{(i)} F^{(i-1)}(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i-1)}) + \boldsymbol{\theta}^{(i)} + \hat{\mathbf{W}}^{(i)}\mathbf{x}'^{(i)})\| = \\
&\|(1 - a^{(i)})(\mathbf{x}^{(i)} - \mathbf{x}'^{(i)}) + a^{(i)}(\tanh(\mathbf{W}_{in}^{(i)} F^{(i-1)}(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i-1)}) + \boldsymbol{\theta}^{(i)} + \hat{\mathbf{W}}^{(i)}\mathbf{x}^{(i)}) - \\
&\quad \tanh(\mathbf{W}_{in}^{(i)} F^{(i-1)}(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i-1)}) + \boldsymbol{\theta}^{(i)} + \hat{\mathbf{W}}^{(i)}\mathbf{x}'^{(i)}))\| \leq \\
&(1 - a^{(i)})\|\mathbf{x}^{(i)} - \mathbf{x}'^{(i)}\| + a^{(i)}\|\mathbf{W}_{in}^{(i)} F^{(i-1)}(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i-1)}) + \boldsymbol{\theta}^{(i)} + \hat{\mathbf{W}}^{(i)}\mathbf{x}^{(i)} - \\
&\mathbf{W}_{in}^{(i)} F^{(i-1)}(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i-1)}) - \boldsymbol{\theta}^{(i)} - \hat{\mathbf{W}}^{(i)}\mathbf{x}'^{(i)}\| = \\
&(1 - a^{(i)})\|\mathbf{x}^{(i)} - \mathbf{x}'^{(i)}\| + a^{(i)}\|\mathbf{W}_{in}^{(i)}(F^{(i-1)}(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i-1)}) - \\
&F^{(i-1)}(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i-1)})) + \hat{\mathbf{W}}^{(i)}(\mathbf{x}^{(i)} - \mathbf{x}'^{(i)})\| \leq \\
&(1 - a^{(i)})\|\mathbf{x}^{(i)} - \mathbf{x}'^{(i)}\| + a^{(i)} \left(\|\mathbf{W}_{in}^{(i)}\| \|F^{(i-1)}(\mathbf{u}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i-1)}) - \right. \\
&\left. F^{(i-1)}(\mathbf{u}, \mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i-1)})\| + \|\hat{\mathbf{W}}^{(i)}\| \|\mathbf{x}^{(i)} - \mathbf{x}'^{(i)}\| \right) \leq \\
&(1 - a^{(i)})\|\mathbf{x}^{(i)} - \mathbf{x}'^{(i)}\| + a^{(i)} \left(\|\mathbf{W}_{in}^{(i)}\| \, C^{(i-1)} \, \|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i-1)}) - \right. \\
&\left. (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i-1)})\| + \|\hat{\mathbf{W}}^{(i)}\| \|\mathbf{x}^{(i)} - \mathbf{x}'^{(i)}\| \right) \leq \\
&(1 - a^{(i)})\|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}) - (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})\| + a^{(i)} \left(C^{(i-1)} \|\mathbf{W}_{in}^{(i)}\| \|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}) - \right. \\
&\left. (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})\| + \|\hat{\mathbf{W}}^{(i)}\| \|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}) - (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})\| \right) = \\
&(1 - a^{(i)})\|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}) - (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})\| + \\
&a^{(i)} \left(C^{(i-1)} \|\mathbf{W}_{in}^{(i)}\| + \|\hat{\mathbf{W}}^{(i)}\| \right) \|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}) - (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})\| = \\
&\left[(1 - a^{(i)}) + a^{(i)} \left(C^{(i-1)} \|\mathbf{W}_{in}^{(i)}\| + \|\hat{\mathbf{W}}^{(i)}\| \right) \right] \|(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}) - (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(i)})\|
\end{aligned}
\tag{30}
$$

from which we can see that $C^{(i)} = (1 - a^{(i)}) + a^{(i)} \left(C^{(i-1)} \|\mathbf{W}_{in}^{(i)}\| + \|\hat{\mathbf{W}}^{(i)}\| \right)$ is a Lipschitz constant for $F^{(i)}$, and thereby whenever $C^{(i)} < 1$ it results that $F^{(i)}$ is a contraction (see Definition 2).

□

C Proof of Theorem 2

Proof Given any input string of length N, denoted by $\mathbf{s}_N = [\mathbf{u}(1), \ldots, \mathbf{u}(N)]$, and for every couple of deepESN global states $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N_L)}) \in \mathbb{R}^{N_L N_R}$ and $\mathbf{x}' = (\mathbf{x}'^{(1)}, \ldots, \mathbf{x}'^{(N_L)}) \in \mathbb{R}^{N_L N_R}$, we have that:

$$
\begin{aligned}
&\|\hat{F}(\mathbf{s}_N, \mathbf{x}) - \hat{F}(\mathbf{s}_N, \mathbf{x}')\| = \\
&\|\hat{F}([\mathbf{u}(1), \dots, \mathbf{u}(N)], \mathbf{x}) - \hat{F}([\mathbf{u}(1), \dots, \mathbf{u}(N)], \mathbf{x}')\| = \\
&\|F(\mathbf{u}(N), \hat{F}([\mathbf{u}(1), \dots, \mathbf{u}(N-1)], \mathbf{x})) - \\
&\quad F(\mathbf{u}(N), \hat{F}([\mathbf{u}(1), \dots, \mathbf{u}(N-1)], \mathbf{x}'))\| \leq \\
&C\|\hat{F}([\mathbf{u}(1), \dots, \mathbf{u}(N-1)], \mathbf{x}) - \hat{F}([\mathbf{u}(1), \dots, \mathbf{u}(N-1)], \mathbf{x}')\| = \\
&C\|F(\mathbf{u}(N-1), \hat{F}([\mathbf{u}(1), \dots, \mathbf{u}(N-2)], \mathbf{x})) - \\
&\quad F(\mathbf{u}(N-1), \hat{F}([\mathbf{u}(1), \dots, \mathbf{u}(N-2)], \mathbf{x}'))\| \leq \\
&C^2\|\hat{F}([\mathbf{u}(1), \dots, \mathbf{u}(N-2)], \mathbf{x}) - \hat{F}([\mathbf{u}(1), \dots, \mathbf{u}(N-2)], \mathbf{x}')\| \leq \\
&\dots \\
&C^{N-1}\|\hat{F}([\mathbf{u}(1)], \mathbf{x}) - \hat{F}([\mathbf{u}(1)], \mathbf{x}')\| = \\
&C^{N-1}\|F(\mathbf{u}(1), \hat{F}([\,], \mathbf{x})) - F(\mathbf{u}(1), \hat{F}([\,], \mathbf{x}'))\| = \\
&C^{N-1}\|F(\mathbf{u}(1), \mathbf{x}) - F(\mathbf{u}(1), \mathbf{x}')\| \leq \\
&C^N \|\mathbf{x} - \mathbf{x}'\| = \\
&C^N \max_{k=1,2,\dots,N_L} \|\mathbf{x}^{(k)} - \mathbf{x}'^{(k)}\| \leq \\
&C^N D
\end{aligned}
\tag{31}
$$

from which it follows that $\|\hat{F}(\mathbf{s}_N, \mathbf{x}) - \hat{F}(\mathbf{s}_N, \mathbf{x}')\|$ is upper bounded by a term that approaches 0 as $N \to \infty$. Thereby the ESP condition in Definition 1 holds.
□

Compliance with Ethical Standards

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

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Aboudib, A., Gripon, V., Coppin, G.: A biologically inspired framework for visual information processing and an application on modeling bottom-up visual attention. Cognitive Computation **8**(6), 1007–1026 (2016)
2. Angelov, P., Sperduti, A.: Challenges in deep learning. In: Proceedings of the 24th European Symposium on Artificial Neural Networks (ESANN), pp. 489–495. i6doc.com (2016)
3. Bengio, Y.: Learning deep architectures for ai. Foundations and trends® in Machine Learning **2**(1), 1–127 (2009)
4. Bianchi, F., Livi, L., Alippi, C.: Investigating echo state networks dynamics by means of recurrence analysis. arXiv preprint arXiv:1601.07381 pp. 1–25 (2016)
5. Buehner, M., Young, P.: A tighter bound for the echo state property. IEEE Transactions on Neural Networks **17**(3), 820–824 (2006)
6. Cireşan, D., Giusti, A., Gambardella, L., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 411–418. Springer (2013)

7. Cireşan, D., Meier, U., Gambardella, L., Schmidhuber, J.: Deep, big, simple neural nets for handwritten digit recognition. Neural computation **22**(12), 3207–3220 (2010)
8. Deng, L., Yu, D.: Deep learning. Signal Processing **7**, 3–4 (2014)
9. El Hihi, S., Bengio, Y.: Hierarchical recurrent neural networks for long-term dependencies. In: NIPS, pp. 493–499 (1995)
10. Gallicchio, C., Micheli, A.: Architectural and markovian factors of echo state networks. Neural Networks **24**(5), 440–456 (2011)
11. Gallicchio, C., Micheli, A.: Deep reservoir computing: A critical analysis. In: Proceedings of the 24th European Symposium on Artificial Neural Networks (ESANN), pp. 497–502. i6doc.com (2016)
12. Gallicchio, C., Micheli, A., Pedrelli, L.: Deep reservoir computing: A critical experimental analysis. Neurocomputing (2016). (Accepted)
13. Gerstner, W., Kistler, W.: Spiking neuron models: Single neurons, populations, plasticity. Cambridge university press (2002)
14. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning (2016). Book in preparation for MIT Press, http://www.deeplearningbook.org
15. Graves, A., Mohamed, A.R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 ieee international conference on Acoustics, speech and signal processing (ICASSP),, pp. 6645–6649. IEEE (2013)
16. Hammer, B., Tiňo, P.: Recurrent neural networks with small weights implement definite memory machines. Neural Computation **15**(8), 1897–1929 (2003)
17. Hermans, M., Schrauwen, B.: Training and analysing deep recurrent neural networks. In: NIPS, pp. 190–198 (2013)
18. Jaeger, H.: The echo state approach to analysing and training recurrent neural networks - with an erratum note. Tech. rep., GMD - German National Research Institute for Computer Science, Tech. Rep. (2001)
19. Jaeger, H.: Short term memory in echo state networks. Tech. rep., German National Research Center for Information Technology (2001)
20. Jaeger, H.: Discovering multiscale dynamical features with hierarchical echo state networks. Tech. rep., Jacobs University Bremen (2007)
21. Jaeger, H., Haas, H.: Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. Science **304**(5667), 78–80 (2004)
22. Jaeger, H., Lukoševičius, M., Popovici, D., Siewert, U.: Optimization and applications of echo state networks with leaky-integrator neurons. Neural Networks **20**(3), 335–352 (2007)
23. Klopf, A., Weaver, S., Morgan, J.: A hierarchical network of control systems that learn: Modeling nervous system function during classical and instrumental conditioning. Adaptive behavior **1**(3), 263–319 (1993)
24. Kolen, J.F., Kremer, S.C.: A field guide to dynamical recurrent networks. IEEE Press (2001)
25. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105 (2012)
26. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
27. Lukoševičius, M., Jaeger, H.: Reservoir computing approaches to recurrent neural network training. Computer Science Review **3**(3), 127–149 (2009)
28. Maass, W., Natschläger, T., Markram, H.: Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural computation **14**(11), 2531–2560 (2002)
29. Malik, Z.K., Hussain, A., Wu, Q.J.: Multilayered echo state machine: A novel architecture and algorithm. IEEE Transactions on Cybernetics (2016). (In Press)
30. Manjunath, G., Jaeger, H.: Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks. Neural computation **25**(3), 671–696 (2013)
31. O'Searcoid, M.: Metric spaces. Springer Science & Business Media (2006)
32. Pascanu, R., Gulcehre, C., Cho, K., Bengio, Y.: How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026v5 (2014)
33. Rabinovich, M., Huerta, R., Varona, P., Afraimovich, V.: Generation and reshaping of sequences in neural systems. Biological cybernetics **95**(6), 519–536 (2006)
34. Rabinovich, M., Varona, P., Selverston, A., Abarbanel, H.: Dynamical principles in neuroscience. Reviews of modern physics **78**(4), 1213 (2006)

35. Rodan, A., Tiňo, P.: Negatively correlated echo state networks. In: Proceedings of the 19th European Symposium on Artificial Neural Networks (ESANN), pp. 53–58. i6doc.com (2011)
36. Sato, Y., Nagatomi, T., Horio, K., Miyamoto, H.: The cognitive mechanisms of multi-scale perception for the recognition of extremely similar faces. Cognitive Computation **7**(5), 501–508 (2015)
37. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural Networks **61**, 85–117 (2015)
38. Schrauwen, B., Wardermann, M., Verstraeten, D., Steil, J., Stroobandt, D.: Improving reservoirs using intrinsic plasticity. Neurocomputing **71**(7), 1159–1171 (2008)
39. Spratling, M.: A hierarchical predictive coding model of object recognition in natural images. Cognitive Computation pp. 1–17 (2016)
40. Steil, J.: Backpropagation-decorrelation: online recurrent learning with o (n) complexity. In: Proceedings of the 2004 IEEE International Joint Conference on Neural Networks (IJCNN), vol. 2, pp. 843–848. IEEE (2004)
41. Tiňo, P., Hammer, B., Bodén, M.: Markovian bias of neural-based architectures with feedback connections. In: Perspectives of neural-symbolic integration, pp. 95–133. Springer (2007)
42. Tiňo, P., Dorffner, G.: Predicting the future of discrete sequences from fractal representations of the past. Machine Learning **45**(2), 187–217 (2001)
43. Triefenbach, F., Jalalvand, A., Demuynck, K., Martens, J.P.: Acoustic modeling with hierarchical reservoirs. IEEE Transactions on Audio, Speech, and Language Processing **21**(11), 2439–2450 (2013)
44. Triefenbach, F., Jalalvand, A., Schrauwen, B., Martens, J.P.: Phoneme recognition with large hierarchical reservoirs. In: Advances in neural information processing systems, pp. 2307–2315 (2010)
45. Tyrrell, T.: The use of hierarchies for action selection. Adaptive Behavior **1**(4), 387–420 (1993)
46. Verstraeten, D., Schrauwen, B., d'Haene, M., Stroobandt, D.: An experimental unification of reservoir computing methods. Neural networks **20**(3), 391–403 (2007)
47. Wainrib, G., Galtier, M.: A local echo state property through the largest lyapunov exponent. Neural Networks **76**, 39–45 (2016)
48. Xue, Y., Yang, L., Haykin, S.: Decoupled echo state networks with lateral inhibition. Neural Networks **20**(3), 365–376 (2007)
49. Yildiz, I., Jaeger, H., Kiebel, S.: Re-visiting the echo state property. Neural networks **35**, 1–9 (2012)