On the Interpretation and Characterization of Echo State Networks Dynamics: A Complex Systems Perspective

Filippo Maria Bianchi, Lorenzo Livi and Cesare Alippi

Abstract In this chapter, we discuss recently developed methods for characterizing the dynamics of recurrent neural networks. Such methods rely on theory and concepts coming from the field of complex systems. We focus on a class of recurrent networks called echo state networks. First, we present a method to analyze and characterize the evolution of its internal state. This allows to provide a qualitative interpretation of the network dynamics. In addition, it allows to assess the stability of the system, a necessary requirement in many practical applications. Successively, we focus on the identification of the onset of criticality in such networks. We discuss an unsupervised method based on Fisher information, which can be used to tune the network hyperparameters. With respect to standard supervised techniques, we show that the proposed approach offers several advantages and is effective on a number of tasks.

Keywords Echo state networks · Criticality · Recurrence quantification analysis · Fisher information matrix · Unsupervised learning

F.M. Bianchi
Machine Learning Group, Department of Physics and Technology,
UiT the Arctic University of Norway, Tromsø, Norway
e-mail: filippo.m.bianchi@uit.no

L. Livi · C. Alippi (✉)
Department of Electronics, Information, and Bioengineering,
Politecnico di Milano, Milan, Italy
e-mail: cesare.alippi@polimi.it

L. Livi
e-mail: lorenz.livi@gmail.com

L. Livi · C. Alippi
Faculty of Informatics, Università della Svizzera Italiana, Lugano, Switzerland

A.E. Gawęda et al. (eds.), *Advances in Data Analysis with Computational Intelligence Methods*, Studies in Computational Intelligence 738,
https://doi.org/10.1007/978-3-319-67946-4_5

1 Introduction

Since the very first recurrent neural network (RNN) architectures, several attempts have been made to describe and understand their internal dynamics [64]. Nowadays, such efforts found renewed interest by those researchers trying to "open the black-box" [26, 45, 46, 49]. This is mostly motivated by recent advances in various fields, such as neuroscience [10]. In fact, understanding the inner mechanisms that drive the inductive inference is of utmost importance for deriving novel scientific results [48].

Research on complex dynamical systems is focusing more and more on networks characterized by time-varying properties [2], which can be related to the topology and/or features associated with vertices and edges (e.g., states of networked dynamic systems). Of particular interest are those systems that also perform a computation when driven by an external stimulus. RNNs, initially proposed in the 80s [12, 42, 60], offer an example of those systems. RNNs are universal approximators of Lebesgue measurable dynamical systems [15], with the capability of storing the history of input signals and utilize such information for prediction [8, 23, 40, 50]. While in principle RNNs are characterized by a simple, yet powerful and flexible model, in practice they are hard to train. In fact, in order to learn the internal connection weights, the network designer has to face a series of technical issues [36]. The most important obstacles are due to the vanishing and exploding gradient [3].

In this chapter, we focus on a particular class of RNN, called Echo State Network (ESN). The main peculiarity of ESNs is that the recurrent part, called reservoir, is randomly generated and the connection weights are kept fixed. The only part that is trained is the so-called readout, a memory-less component that combines the neuron activations of the reservoir in order to reproduce a suitable output, according to the specified task at hand. ESNs not only benefit from the presence of feedbacks like any other RNN (the feature which gives to the system the capability to model any complex dynamic behavior) but their sparsely interconnected reservoir of neurons leads to a very fast and simple training procedure. In fact, unlike the complicated and time consuming training process required by standard RNNs, a simple linear readout can be used to solve efficiently a great variety of tasks. On the downside, ESN is characterized by a short-term memory, making it unsuitable for application when long-term correlations must be modeled [37].

Even if ESNs offer an important simplification for what concerns training, they depend on hyperparameters affecting their behavior; additionally, their modus operandi is still not fully understood and it represents an actual object of study [6, 49]. An ESN can generate complex dynamics characterized by sharp transitions between ordered and chaotic regimes. Several experimental results suggest that ESNs achieve the highest information processing capabilities exactly on the edge of this transition, called edge of criticality, resulting in high memory capacity (storage of past events) and good performance on the modeling/prediction task at hand (low prediction errors) [1, 5, 21, 39, 54, 58]. To determine such "critical" network configurations, an ESN requires fine tuning of its controlling hyperparameters. This general behavior is in agreement with the widely-discussed “criticality hypothesis” observed in many biological (complex) systems [14, 16, 41, 43, 51], including the brain [9, 32, 35, 52, 53]. In fact, it was noted [34] that such complex systems tend to self-organize and operate in a critical regime. Investigating weather a given complex system operates more efficiently in the critical regime or not, requires theoretically sound methods for detecting the onset of criticality [44].

Best-performing network configurations are typically identified through supervised methods, such as cross-validation and alike. In this chapter, we present recent research results [6, 22] that focus on unsupervised approaches to characterize ESN dynamics and to identify the edge of criticality. These approaches do not require a validation set, an important limitation in several applications, with scarce amount of data. Another issue of validation procedures is the need to repeat training for each hyperparameter configuration taken into account. Through the proposed unsupervised approaches, hyperparameters are tuned in advance and training is performed just once, at the end. Finally, cross validation considers only the performance obtained on the given task, treating the network as black box. Instead, the presented methods offer insights on the functioning, by modeling dynamics with more easily interpretable tools.

Different unsupervised approaches to identify configurations that maximize ESN computation capability have been proposed in the literature. These, are quickly reviewed in Sect. 2 after an overview on the ESN architecture. In Sect. 3, we address the issue of interpretability of ESN dynamics by relying on recurrence plots and recurrence quantification analysis [6] to characterize the evolution of the internal states. When the network is driven by a specific input signal, these instruments can be used to monitor its degree of stability, for a given configuration of its hyperparameters. In Sect. 4, we define an unsupervised methodology for tuning ESN hyperparameters by means of sensitivity analyses [22]. In particular, we present a theoretical framework based on Fisher information matrix [55, 62] and its related connection with criticality. Conclusions and future research directions are provided in Sect. 5.

2 Echo State Networks

A schematic representation of an ESN is shown in Fig. 1. An ESN consists of a reservoir of N_r nodes characterized by a non-linear transfer function $f(\cdot)$. At time t, the network is driven by the input $\mathbf{x}[t] \in \mathbb{R}^{N_i}$ and produces the output $\mathbf{y}[t] \in \mathbb{R}^{N_o}$, being N_i and N_o the dimensionalities of input and output, respectively. The weight matrices $\mathbf{W}_r^r \in \mathbb{R}^{N_r \times N_r}$ (reservoir internal connections), $\mathbf{W}_i^r \in \mathbb{R}^{N_i \times N_r}$ (input-to-reservoir connections), and $\mathbf{W}_o^r \in \mathbb{R}^{N_o \times N_r}$ (output-to-reservoir feedback connections) contain values in the $[-1, 1]$ interval drawn from a uniform distribution.

ESN is a discrete-time nonlinear system with feedback, whose model reads:

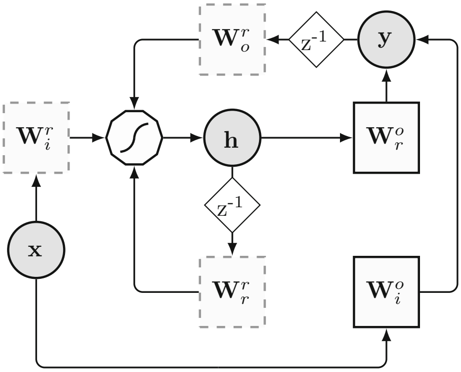


Fig. 1 Schematic depiction of the ESN architecture. The circles represent input **x**, state, **h**, and output, **y**, respectively. Solid squares $\mathbf{W}_r^o$ and $\mathbf{W}_i^o$, are the trainable matrices, respectively, of the readout, while dashed squares, $\mathbf{W}_r^r$, $\mathbf{W}_o^r$, and $\mathbf{W}_i^r$, are randomly initialized matrices. The polygon represents the non-linear transformation performed by neurons and z^{-1} is the unit delay operator

$$\mathbf{h}[t] = f\left(\mathbf{W}_r^r\mathbf{h}[t-1] + \mathbf{W}_i^r\mathbf{x}[k] + \mathbf{W}_o^r\mathbf{y}[t-1]\right); \tag{1}$$

$$\mathbf{y}[t] = g\left(\mathbf{W}_r^o\mathbf{h}[t] + \mathbf{W}_i^o\mathbf{x}[k]\right). \tag{2}$$

Activation functions $f(\cdot)$ and $g(\cdot)$, both applied component-wise, are typically implemented as a sigmoidal (*tanh*) and identity function, respectively. The output weight matrices $\mathbf{W}_r^o \in \mathbb{R}^{N_r \times N_o}$ and $\mathbf{W}_i^o \in \mathbb{R}^{N_i \times N_o}$, which connect, respectively, reservoir and input to the output, represent the readout of the network. The standard training procedure for such matrices requires solving a straightforward regularized least-square problem [18].

Even though the three matrices $\mathbf{W}_r^r$, $\mathbf{W}_o^r$, and $\mathbf{W}_i^r$ are generated randomly, they can be modified in order to obtain desired properties. For instance, $\mathbf{W}_r^o$ is controlled by a multiplicative constant, which in this work is set to 0 to remove the output feedback connection. $\mathbf{W}_i^r$ is controlled by scalar parameter θ_{IS}, which determines the amount of non-linearity introduced by the sigmoid processing units that is largest around the origin. In particular, inputs far from zero tend to drive the activation of the neurons towards saturation where they show more non-linearity. Finally, the parameter θ_{RC} defines the percentage of non-zero connections in $\mathbf{W}_r^r$, while its spectral radius θ_{SR} controls important properties, as discussed in the sequel.

2.1 ESN Dynamics and Stability Measures

An ESN is typically designed so that the influence of past inputs gradually fades away and the initial state of the reservoir is eventually washed out. This is formalized by the Echo State Property (ESP), which ensures that, given any input sequence taken from a compact set, trajectories of any two different initial states become eventually indistinguishable. ESP was originally investigated in [18] and successively in [61]; we refer the interested reader to [25] for a more recent definition, where also the influence of input is explicitly accounted for. In ESNs with no output feedback, as in our case, the state update of Eq. (1) reduces to:

$$\mathbf{h}[t] = f(\mathbf{W}_r^r \mathbf{h}[t-1] + \mathbf{W}_i^r \mathbf{x}[k]). \tag{3}$$

In order to study the stability of the network, we compute the maximal local Lyapunov exponent (λ) from the Jacobian of the state update (3) of the reservoir. This quantity is used to approximate (for an autonomous system) the separation rate in phase space of trajectories having very similar initial conditions. λ is derived from the Jacobian at time t, which can be conveniently expressed if neurons are implemented with a *tanh* activation function as

$$\mathbf{J}(\mathbf{h}[t]) = \mathbb{I}_{N_r} \cdot \left[1 - (h_1[t])^2, 1 - (h_2[t])^2, \ldots, 1 - (h_{N_r}[t])^2\right]^T. \tag{4}$$

where $h_l[t]$ is the activation of the l-th neuron, with $l = 1, 2, \ldots, N_r$. λ is then computed as

$$\lambda = \max_{n=1,\ldots,N_r} \frac{1}{t_{\max}} \sum_{t=1}^{t_{\max}} \log\left(r_n[t]\right), \tag{5}$$

being $r_n[t]$ the module of n-th eigenvalue of $\mathbf{J}(h[t])$ and $t_{\max}$ the total number of time-steps in the considered trajectory.

Local, first-order approximations provided by Eq. 4 are useful to study the stability of a (simplified) reservoir operating around the zero state, $\mathbf{0}$. In fact, implementing $f(\cdot)$ as a *tanh* assures $f(\mathbf{0}) = \mathbf{0}$, i.e., $\mathbf{0}$ is a fixed point of the ESN dynamics. Therefore, by linearizing (3) around $\mathbf{0}$ and assuming a zero-input, we obtain from (4)

$$\mathbf{h}[t] = \mathbf{J}(\mathbf{0})\mathbf{h}[t-1] = \mathbf{W}_r^r \mathbf{h}[t-1]. \tag{6}$$

Linear stability analysis of (6) suggests that, if $\theta_{\mathrm{SR}} < 1$, the dynamic around $\mathbf{0}$ is stable. In the more general case, the non-linearity of the sigmoid functions in (3) forces the norm of the state vector of the reservoir to remain bounded. Therefore, the condition $\theta_{\mathrm{SR}} < 1$ looses its significance and does not guarantee stability when the system deviates from a small region around $\mathbf{0}$ [57]. Notably, it is possible to find reservoirs (3) having $\theta_{\mathrm{SR}} > 1$, which still possess the ESP. In fact, the effective local gain decreases when the operating point of the neurons shifts toward the positive/negative branch of the sigmoid, where stabilizing saturation effects start to influence the excitability of reservoir dynamics [61]. In the more realistic and useful scenario where the input driving the network is a generic (non-zero) signal, a sufficient condition for the ESP is met if $\mathbf{W}_r^r$ is diagonally Schur-stable, i.e., if there exists a positive definite diagonal matrix, $\mathbf{P}$, such that $(\mathbf{W}_r^r)^T \mathbf{P} \mathbf{W}_r^r - \mathbf{P}$ is negative definite [61]. However, this recipe is fairly restrictive in practice as this condition might generate reservoirs that are not rich enough in terms of provided dynamics, since the use of a conservative scaling factor might compromise the amount of memory in the network and thus the ability to accurately model a given problem. Therefore, for most practical purposes, the necessary condition $\theta_{\text{SR}} < 1$ is considered "sufficient in practice", since the state update map is contractive with high probability, regardless of the input and given a sufficiently large reservoir [63].

2.2 Edge of Criticality

The number of reservoir neurons and the bounds on θ_{SR} can be used for a naïve quantification of the computational capability of a reservoir [61]. However, those are static measures that only consider the algebraic properties of $\mathbf{W}_r^r$, without taking into account other factors, such as the input scaling θ_{IS} and the particular properties of the given input signals. Moreover, it is still not clear how, in a mathematical sense, these stability bounds relate to the actual ESN dynamics when processing non-trivial input signals [25]. In this context, the idea of pushing the system toward the edge of criticality has been explored. In [5, 20, 21] it is shown that several dynamical systems, among which randomly connected RNNs, achieved the highest computational capabilities when moving toward the unstable (sometime even chaotic) regime, where the ESP is lost and the system enters into an oscillatory behavior. This justifies the use of spectral radii above the unity in some practical applications.

The stable–unstable transition can be detected numerically by considering the sign of λ (5). In fact, in autonomous systems, $\lambda > 0$ indicates that the dynamics is chaotic. Relative to ESNs, λ was proposed to characterize reservoir dynamics and it demonstrated its efficacy in designing a suitable network configuration in several applications [56, 57]. Further descriptors used for characterizing the dynamics of a reservoir are based on information-theoretic quantities, such as (average) transfer entropy and active information storage [7]. The authors have shown that such quantities peak right when $\lambda > 0$. In addition, the minimal singular value of the Jacobian (4), denoted as η, was demonstrated to be an accurate predictor of ESN performance, providing more accurate information regarding the ESN dynamics than both λ and θ_{SR} [56]. Hyperparameters that maximize η generate a dynamical system that is far from singularity, it has many degrees of freedom, a good excitability, and it separates well the input signals in phase space [56].

3 Interpreting and Tuning ESN Through Recurrence Quantification Analysis

Poincaré recurrence provides fundamental information for the analysis of dynamical systems [29]. This follows from Poincaré's theorem, which guarantees that the states of a dynamic system must recur during its evolution. Recurrences contain all relevant information regarding a system behavior in phase space and can be linked also with dynamical invariants (e.g., metric entropy) and features related to stability. However, especially for high-dimensional complex systems, the recurrence time elapsed between recurring states is difficult to calculate, even when assuming full analytical knowledge of the system.

Recurrence Plots (RPs) [11, 27, 29, 30], together with the computation of dynamical invariants and heuristic complexity measures called Recurrence Quantification Analysis (RQA), offer a simple yet effective tool to analyze such recurrences starting from a time-series derived from the system under analysis. RP provides a visual representation of recurrence time and its line patterns contain information about the duration of the recurrence [28]. RPs are constructed by considering a suitable distance in the phase space and a threshold τ_{RP} is used to determine the recurrence/similarity of states during the evolution of the system.

In the following, we address the interpretability issue of ESNs by analyzing the dynamics of the reservoir neuron activations with RPs and RQA complexity measures. Techniques based on RPs and RQA allow the designer to visualize and characterize (high-dimensional) dynamical systems starting from a matrix encoding the recurrences of the system states over time.

3.1 Representing ESN Dynamics with RP

The sequence of ESN states can be seen as a multivariate time-series $\mathbf{h}$, relative to the N_r neuron activations. An RP is constructed by calculating a $t_{\max} \times t_{\max}$ binary matrix $\mathbf{R}$. The generic element R_{ij} is defined as

$$R_{ij} = \Theta(\tau_{\mathrm{RP}} - d(\mathbf{h}[i], \mathbf{h}[j])), \quad 1 \leq i, j \leq t_{\max}, \tag{7}$$

where $d(\cdot, \cdot)$ is a dissimilarity measure operating in phase space (e.g., Euclidean, Manhattan, or max-norm distance), $\Theta(\cdot)$ is the Heaviside function and $\tau_{\mathrm{RP}} > 0$ is a user-defined threshold used to identify recurrences. τ_{RP} can be defined in different ways, but typically chosen to be proportional to a percentage of the average or the maximum phase space distance between the states. Figure 2 depicts the algorithmic steps required to generate an RP on ESN states.

Depending on the properties of the analyzed time-series, different line patterns emerge in a RP [28]. Besides providing an immediate visualization of the system properties, from $\mathbf{R}$ it is possible to derive several complexity measures, those associated with an RQA. Such measures are defined by the distribution of both vertical/horizontal and diagonal line structures present in the RP and provide a numerical characterization of the underlying dynamics. Several RQA measures are based on the histograms $P(l)$ and $P(v)$, counting, respectively, the diagonal and vertical lines having lengths l and v,

$$P(l) = \sum_{i,j=1}^{t_{\max}-l} (1 - R_{i-1,j-1})(1 - R_{i+l,j+l}) \prod_{k=0}^{l-1} R_{i+k,j+k};$$
$$P(v) = \sum_{i,j=1}^{t_{\max}-v} (1 - R_{i,j})(1 - R_{i,j+v}) \prod_{k=0}^{v-1} R_{i,j+k}.$$

The RQA measures considered here are summarized in Table 1; abbreviations and notation are kept consistent with [29].

3.2 Visualize and Classify Reservoir Dynamics

In the following, we show how RPs permit to visualize, and hence classify, reservoir dynamics when ESN is fed with inputs possessing well-known characteristics. We consider a stable ESN described by (3); RPs are constructed following the procedure depicted in Fig. 2. Although many classes of signals/systems exist (with related subclasses) [29], here we focus on the ability to discriminate between important classes for the input signals: (i) with/without time-dependence, (ii) periodic/non-periodic

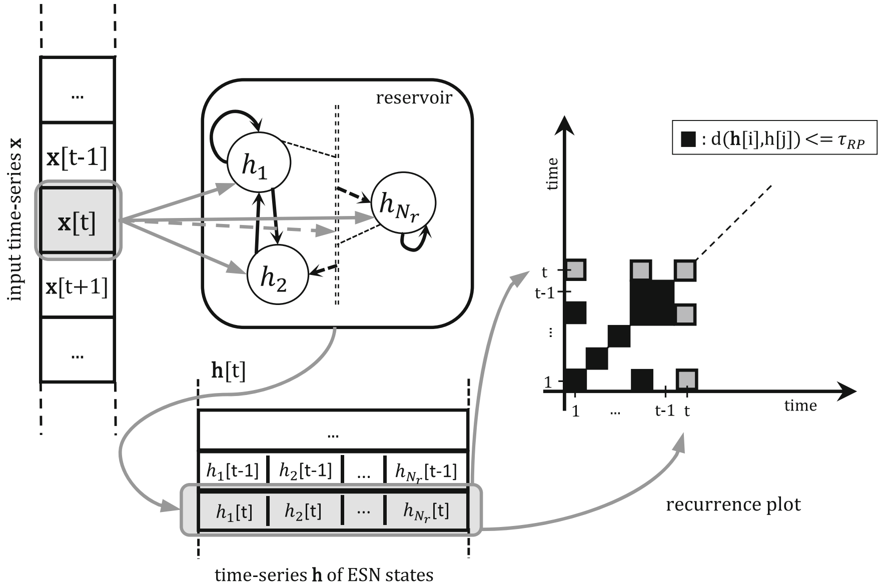


Fig. 2 When $\mathbf{x}[t]$ is fed as input to the N_r neurons of the ESN reservoir, the internal state is updated to $\mathbf{h}[t] = [h_1[t], h_2[t], \ldots, h_{N_r}[t]]^T$, where $h_n[t]$ is the output of the n-th neuron. Once the time-series $\mathbf{h}$ is generated, the RP is constructed by using a threshold τ_{RP} and a dissimilarity measure $d(\cdot, \cdot)$. If $d(\mathbf{h}[t], \mathbf{h}[i]) \leq \tau_{\mathrm{RP}}$, the cell of the RP in position (t, i) is colored in black, otherwise it is left white. The elements in gray highlight the operations performed at time-step t. Taken from [6]

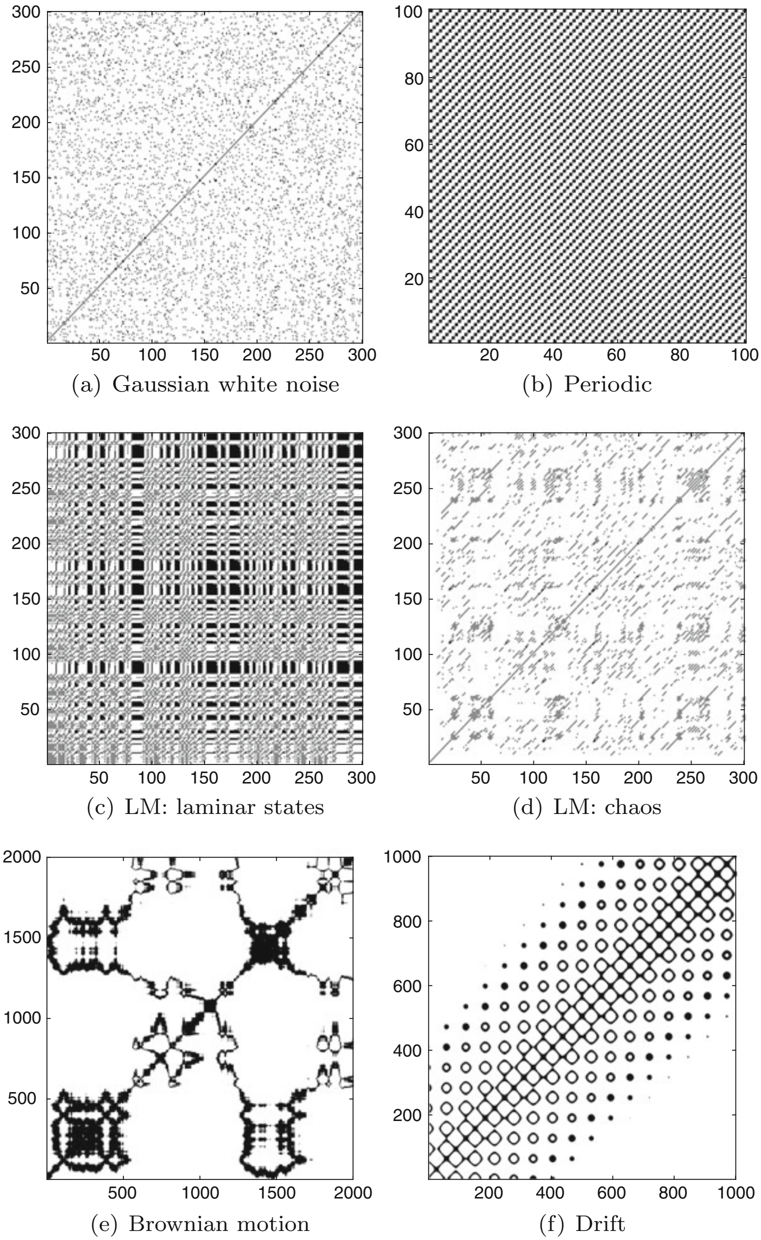


Fig. 3 RPs generated by state sequences **h** of ESNs fed with input signals taken into account. Both axes represent time. Taken from [6]

Table 1 Definition of RQA measures

$\mathrm{RR} = \frac{1}{t_{\max}^2} \sum_{i,j=1}^{t_{\max}} R_{ij}$	Recurrence rate, a measure of density of recurrences in **R**. It corresponds to the correlation sum, an important concept used in chaos theory. RR can help to select τ_{RP} when performing multiple tests on different conditions, e.g., by preserving the rate
$\mathrm{DET} = \frac{\sum_{l=l_{\min}}^{t_{\max}} lP(l)}{\sum_{l=1}^{t_{\max}} lP(l)}$	Determinism level of the system, based on the percentage of diagonal lines of minimum length $l_{\min}$. A periodic system would have DET close to unity and close to zero for a signal with no time-dependency
$\mathrm{L}_{\max} = \max\{l_i\}_{i=1}^{N_l}$	Maximum diagonal line length, with $1 \leq \mathrm{L}_{\max} \leq \sqrt{2} t_{\max}$. l_i is ith diagonal line length and N_l is the total number of diagonal lines, defined as $N_l = \sum_{l \geq l_{\min}} P(l)$
$\mathrm{DIV} = 1/\mathrm{L}_{\max}$	Mean exponential divergence in phase space, related to correlation entropy of the system. Notably, chaotic systems do not present long diagonal lines, as trajectories diverge exponentially fast
$\mathrm{LAM} = \frac{\sum_{v=v_{\min}}^{t_{\max}} vP(v)}{\sum_{v=1}^{t_{\max}} vP(v)}$	Presence of laminar phases, which denote states of the system that do not change or change very slowly for a number of consecutive time-steps. $v_{\min}$ is the minimal vertical line length considered
$\mathrm{ENTR} = -\sum_{l=1}^{t_{\max}} p(l) \ln(p(l))$	Diagonal lines distribution, with $p(l) = P(l)/N_l$. In absence of time-dependence, ENTR $\simeq 0$, i.e., the diagonal lines distribution is fully concentrated on very short lines. Conversely, ENTR increases when the diagonal lines distribution become heterogeneous

motions, (iii) laminar behaviours, (iv) chaotic dynamics, and finally (v) non-stationary processes. We refer to the examples depicted in Fig. 3 to discuss the RP relative to each class.

Time-dependency: a uniformly distributed RP denotes absence of time-dependence in the time-series. Specific RQA measures, such as DET and ENTR (Table 1), can be used to numerically investigate the presence of time-dependency, as their values is very low if the signal is uncorrelated. For periodic signal with a strong time-dependency, DET would be very high, but ENTR would still be low. In fact, ENTR measures the complexity of the signal, which is low if there is no temporal structure. Figure 3a depicts the RP generated by feeding the ESN with Gaussian white noise, a typical example of signal with no time-dependency. Reservoir states generates a uniform RP, which is peculiar of signals composed by realizations of statistically independent variables.

Periodicity: every periodic system would induce long diagonal lines and the vertical spacing provides the period of the oscillation. A periodic system is typically accompanied by high values for DET and $\mathrm{L}_{\max}$, while its low complexity is expressed by ENTR. In Fig. 3b, we show an example of periodic motion generated by reservoir neurons, when ESN is fed with a sinusoid having a single dominating frequency. The regularity of the diagonal lines can be immediately recognized from the figure.

Laminarity: a system presents laminar phases if its state does not change or change very slowly over a number of successive time-steps. Laminar phases can be visually recognized in an RP by the presence of black rectangles. Every system possessing laminar phases is characterized by high values for LAM. To provide an example, we consider the logistic map (LM), defined by the differential equation $\mathbf{x}[t+1] = \tau_{\mathrm{LM}}\mathbf{x}[t](1-\mathbf{x}[t])$, where usually $\tau_{\mathrm{LM}} \in (0, 4]$; here we set the initial condition $\mathbf{x}[1] = 0.5$. Figure 3c depicts RP obtained for $\tau_{\mathrm{LM}} = 3.679$, where the system exhibits chaos-chaos transitions. In fact, such a RP is compatible with the one of a (mildly) chaotic system, showing the presence of laminar phases (large black rectangles).

Chaoticity: RPs offer a particularly useful visual tool in the case of chaotic dynamics, which are characterized by the presence of erratic and very short diagonal lines. As a consequence, RR would be very low. ENTR is also useful to determine the degree of chaoticity: the higher its value, the more chaotic/complex the system. Chaos is characterized by trajectories diverging exponentially fast. This can be quantified with $\mathrm{L}_{\max}$ and DIV, whose values would be respectively very low and close to one for systems with a high degree of chaoticity. As an example, we consider a chaotic system obtained through LM set with $\tau_{\mathrm{LM}} = 4$. The reservoir dynamics, as shown in the RP in Fig. 3d, denotes fully developed chaos, as indicated by the presence of short and erratic diagonal lines.

Non-stationarity: Peculiar line patterns observed for all nonstationary signals include large white areas with irregular patterns denoting abrupt changes in the dynamics. Drift is a typical form of nonstationarity, which is visually recognized in an RP by the fading of recurrences in the upper-left and lower-right corners. In Fig. 3e, we show an example by feeding the ESN with a well-known nonstationary signal: Brownian motion, a random walk resulting in a nonstationary stochastic process; whose increments correspond to Gaussian white noise, a stationary process. In Fig. 3f we show an example of drift, obtained by adding a linear trend to a sinusoid. Nonstationarity can be numerically detected by considering an RQA measure called TREND (not used in our study) and by analyzing the variation of RQA measures when time-delay is applied to the signal (see [29] for technical details).

3.3 Recurrence Analysis to Determine ESN Stability

In this section, we show how recurrence analysis can be used to assess stability for a given configuration. We perform two experiments: in the first one, we use RPs to visualize reservoir dynamics when driven by a given input signal. When the reservoir operates in a stable regime, RPs of reservoir and input show similar line patterns. In a second experiment, We show that $\mathrm{L}_{\max}$ is anticorrelated with λ and hence it can be considered as a reliable indicator for the (input-dependent) degree of network stability.

To test our methodology, we consider two time-series generated respectively by an oscillatory and by the Mackey-Glass (MG) dynamical system [47]. We chose these two signals since both of them are often considered as benchmarks for prediction in the ESN literature [18, 57] and they exemplify a very regular and a mildly chaotic system, respectively. In both experiments, we consider an ESN with no output feedback, configured with a standard setting: uniformly distributed weights in $[-1, 1]$ for W_i^r and W_r^r, percentage of non-zero reservoir connections $\theta_{\mathrm{RC}} = 25\%$. The readout is trained by setting the regularization parameter in the ridge regression to 0.1. According to the standard drop-out procedure, we discarded the first 100 elements of $\mathbf{h}$ in order to get rid of the ESN transient states. The number of reservoir neurons is set to $N_r = 75$. We used the Manhattan distance for evaluating the dissimilarity in the phase space. The threshold τ_{RP} has been calculated by using a percentage of the average dissimilarity value between the states in $\mathbf{h}$. Our results are easily reproducible by using the ESN[1] and RP[2] toolboxes available online.

The first experiment consists in generating the RP relative to the input sequence $\{\mathbf{x}[t]\}_{t=1}^{t_{\max}}$ (sinusoid or MG time-series) and the ones relative to neuron activations $\{\mathbf{h}[t]\}_{t=1}^{t_{\max}}$, when the reservoir is configured with a spectral radius θ_{SR} that determines a ordered or a chaotic dynamics.

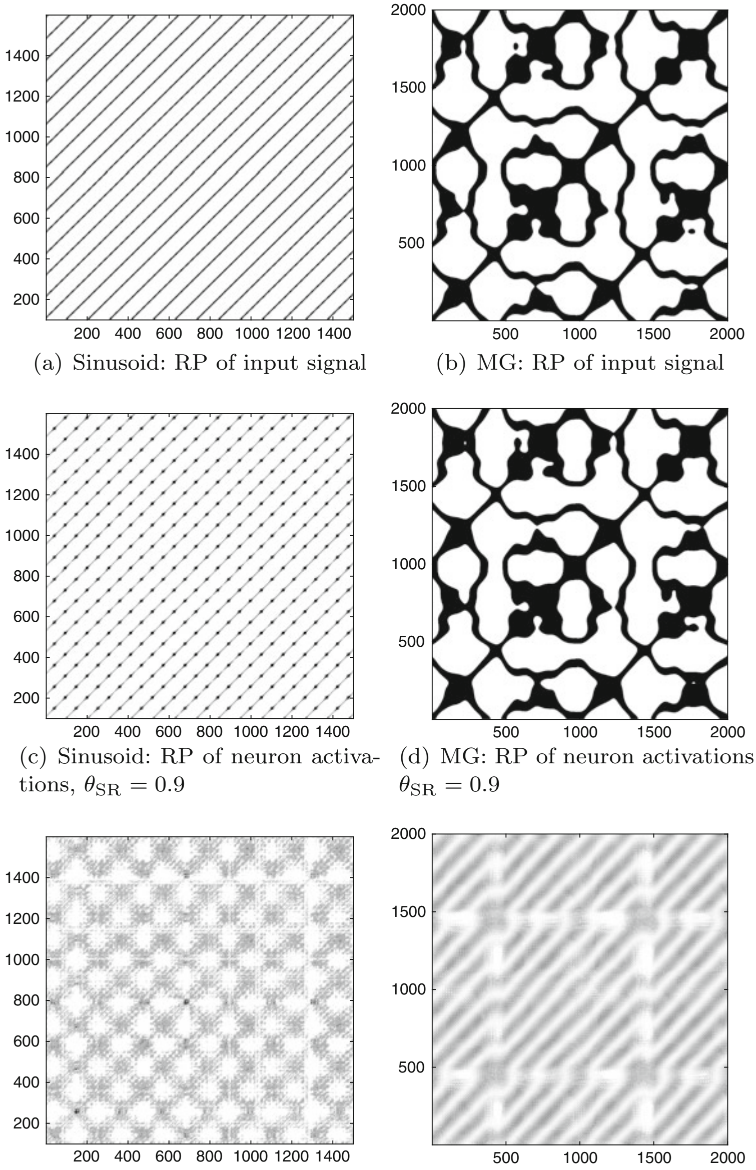
In Fig. 4, we report the RPs relative to the input signal and the reservoir states, generated for two different values of θ_{SR}. The left column is relative to the ESN fed with a sinusoid and the right column to the ESN fed with the MG time-series. As we can see, when $\theta_{\mathrm{SR}} = 0.9$ the ESN is stable and the dynamics of the input, represented by the RPs in Fig. 4a and b produce very similar line patterns in the RPs of the reservoirs, reported in Fig. 4c and d. Instead, when the spectral radius is pushed far beyond unity, the ESN dynamics become unstable and the similarity in the reservoir RPs is lost, as we can see from Fig. 4e and f.

In the second experiment, we evaluate the effectiveness of $\mathrm{L}_{\max}$ and DIV in determining the degree of stability in the ESN. Specifically, the higher the value of $\mathrm{L}_{\max}$, the more stable the system. The opposite holds for DIV, which is computed as the reciprocal of $\mathrm{L}_{\max}$ (see Table 1). Our evaluation consists in comparing, λ, a global indicator of stability (see Eq. 5), with $\mathrm{L}_{\max}$, the value of the longest diagonal line in an RP, and with DIV. As before, we consider two ESN fed with the sinusoid and the MG time-series. The correlations of these measures are reported in Table 2.

To visually assess the agreement of λ with $\mathrm{L}_{\max}$ and DIV, in Fig. 5 we show a 2D depiction obtained by selecting a specific input scaling $\theta_{\mathrm{IS}} = 0.8$ and by varying θ_{SR} in the interval $[0.1, 2]$. For the sinusoidal input, λ and $\mathrm{L}_{\max}$ are anticorrelated with (Pearson) correlation equal to -0.74: the value of $\mathrm{L}_{\max}$ decreases as θ_{SR} increases, while λ, as expected, increases with θ_{SR}. Additionally, it is possible to observe that there exists a positive correlation (0.53) between λ and DIV. Also for the MG time-series, λ and $\mathrm{L}_{\max}$ show a good anticorrelation, with a value of -0.65. Analogously, λ and DIV are correlated with a slightly lower value of 0.57.

[1] http://www.reservoir-computing.org/node/129.

[2] http://www.recurrence-plot.tk/.



(a) Sinusoid: RP of input signal

(b) MG: RP of input signal

(c) Sinusoid: RP of neuron activations, $\theta_{\mathrm{SR}} = 0.9$

(d) MG: RP of neuron activations, $\theta_{\mathrm{SR}} = 0.9$

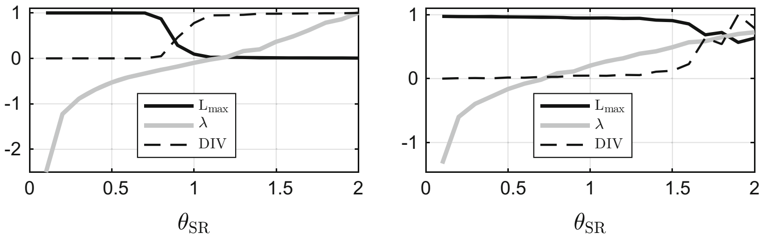
(e) Sinusoid: RP of neuron activations, $\theta_{\mathrm{SR}} = 1.5$

(f) MG: RP of neuron activations, $\theta_{\mathrm{SR}} = 1.5$

Fig. 4 RPs of input signal and sequence of states of the reservoir. When $\theta_{\mathrm{SR}} = 0.9$, the ESN is stable and the activations are compatible with the input dynamics. When θ_{SR} exceeds one, the activations denote instability. Taken from [6]

Table 2 Correlations between λ, DIV, and $\mathrm{L}_{\max}$ for sinusoid input and MG time-series

	$\mathrm{corr}(\lambda, \mathrm{L}_{\max})$	$\mathrm{corr}(\lambda, \mathrm{DIV})$
Sin	−0.74	0.53
MG	−0.65	0.57



(a) Sinusoid: $\mathrm{L}_{\max}$, λ and DIV (b) MG: $\mathrm{L}_{\max}$, λ and DIV

Fig. 5 Value of λ (gray solid line), value of $\mathrm{L}_{\max}$ (solid black line), and the value of DIV (dashed black line) for the ESN fed with sinusoid input (left) and MG time-series (right). Taken from [6]

Even if in Fig. 5 we provide a visualization only for a specific value of input scaling, it is important to remark that the agreement between λ and $\mathrm{L}_{\max}$ is consistent for the entire range of θ_{IS}, confirming that statistics of the RP diagonal lines offer consistent and solid complexity measures that are able to characterize the network stability.

4 Detection of Critical Dynamics with Fisher Information

In the last part of this chapter, we present a theoretically motivated, unsupervised method based on Fisher information for determining the edge of criticality in ESNs (see [22] for details). It is proven that Fisher information is maximized for (finite-size) systems operating near or on the edge of criticality [38]. Accordingly, the hyperparameters, which indirectly affect ESN performance, are suitably controlled to identify a collection of network configurations that maximize Fisher information and computational performance. Since no assumption regarding the mathematical model of the (input-driven) dynamic system is made, the method can handle any type of applications. Additionally, it is independent of the particular reservoir topology, since it operates in the hyperparameter space. This allows the network designer to instantiate a specific architecture based on problem-dependent design choices. However, Fisher information is notoriously difficult to compute and either requires the probability density function or the conditional dependence of the system states with respect to the model parameters. In the proposed framework, we take advantage of a recently-developed non-parametric estimator of the Fisher information matrix [4].

4.1 Fisher Information Matrix and the Non-parametric Estimator

Fisher information matrix (FIM) [62] is a symmetric positive semi-definite (PD) matrix, whose elements are defined as follows:

$$F_{ij}(p_{\boldsymbol{\theta}}(\cdot)) = \int_{\mathcal{D}} p_{\boldsymbol{\theta}}(\mathbf{u}) \left(\frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{u})}{\partial \theta_i} \right) \left(\frac{\partial \ln p_{\boldsymbol{\theta}}(\mathbf{u})}{\partial \theta_j} \right) d\mathbf{u}, \tag{8}$$

where $p_{\boldsymbol{\theta}}(\cdot)$ is a parametric probability density function (PDF), which depends on d parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_d]^T \in \Theta \subset \mathbb{R}^d$; Θ is the parameter space. In the ESN framework, $\boldsymbol{\theta}$ contains the hyperparameters under consideration. In (8), $\ln p_{\boldsymbol{\theta}}(\cdot)$ is the log-likelihood function and $\mathcal{D} \subseteq \mathbb{R}^D$ denotes the domain of the PDF. To simplify notation, we denote $\mathbf{F}(p_{\boldsymbol{\theta}}(\cdot))$ as $\mathbf{F}(\boldsymbol{\theta})$. The FIM contains $d(d+1)/2$ distinct entries encoding the sensitivity of the PDF with respect to the parameters $\boldsymbol{\theta}$.

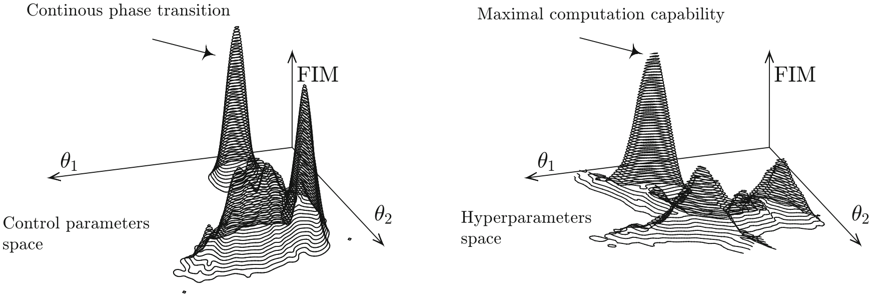
Fisher information is tightly linked with statistical mechanics and, in particular, with the field of (continuous) phase transitions. In fact, it is possible to provide a thermodynamic interpretation of Fisher information in terms of rate of change of the order parameter [38], quantities used to discriminate the different phases of a system. This fact provides an important link between the concept of criticality and statistical modeling of complex systems. It emerges that the critical phase of a thermodynamic system can be mathematically described as that region of the phase space where the order parameters vanish and their derivatives diverge. This implies that, on the critical region, FIM diverges as well, hence providing a quantitative, well-justified tool for detecting the onset of criticality in both theoretical models and computational simulations [59]. In the ESN framework considered here, we identify the edge of criticality as the region in parameter space where the Fisher information is maximized. Figure 6 provides an intuitive illustration, linking criticality and ESNs.

Calculation of the FIM (8) requires full analytical knowledge of the PDF. However, in many experimental settings either the PDF underlying the observed data is unknown or the relation linking the variation of the control parameters $\boldsymbol{\theta}$ and the resulting $p_{\boldsymbol{\theta}}(\cdot)$ depends on an unknown function. Recently, a non-parametric estimator of the FIM based on divergence measure

$$D_\alpha(p, q) = \frac{1}{4\alpha(1-\alpha)} \int_{\mathcal{D}} \frac{(\alpha p(\mathbf{u})(1-\alpha)q(\mathbf{u}))^2}{\alpha p(\mathbf{u})(1-\alpha)q(\mathbf{u})} d\mathbf{u} - (2\alpha - 1)^2, \tag{9}$$

was proposed [4], with $\alpha \in (0, 1)$; $p(\cdot)$ and $q(\cdot)$ are PDFs both supported on $\mathcal{D}$. D_α belongs to the family of f-divergences and it can be computed directly by means of an extension of the Friedman-Rafsky multi-variate two-sample test statistic [13].

FIM can be approximated by using a proper f-divergence measure computed between the parametric PDF of interest and a perturbed version of it [17]. Notably, by expanding Eq. 9 up to the second order we obtain:



(a) Thermodynamic systems (b) Echo State Networks

Fig. 6 The approach based on FIM maximization used to identify a continuous phase transition can be adopted also to characterize dynamics in ESNs. In this context, ESN hyperparameters (e.g., spectral radius, input scaling) play the same role of the control parameters in a thermodynamic system. FIM can be used to identify the critical region in the ESN hyperparameter space, where the computational capability is maximized. Taken from [22]

$$D_\alpha(p_{\boldsymbol{\theta}}, p_{\hat{\boldsymbol{\theta}}}) \simeq \frac{1}{2}\mathbf{r}^T\mathbf{F}(\boldsymbol{\theta})\mathbf{r}, \tag{10}$$

where $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \mathbf{r}$, being $\mathbf{r} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{d\times d})$ a small normally distributed perturbation vector with standard deviation σ.

In the following, we omit $\boldsymbol{\theta}$ and we refer to the estimated FIM as $\hat{\mathbf{F}}$. According to [4], FIM can be estimated through least-square optimization:

$$\hat{\mathbf{F}}_{\text{hvec}} = (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{v}_{\boldsymbol{\theta}}, \tag{11}$$

where $\mathbf{v}_{\boldsymbol{\theta}} = [v_{\boldsymbol{\theta}}(\mathbf{r}_1), \ldots, v_{\boldsymbol{\theta}}(\mathbf{r}_M)]^T$, with $v_{\boldsymbol{\theta}}(\mathbf{r}_i) = 2D_\alpha(p_{\boldsymbol{\theta}}, p_{\hat{\boldsymbol{\theta}}_i}), i = 1, \ldots, M$, and $D_\alpha(\cdot, \cdot)$ is computed by means of the Friedman-Rafsky test. $\mathbf{R}$ is a matrix containing all M perturbation vectors $\mathbf{r}_i$ arranged as column vectors, and $\hat{\mathbf{F}}_{\text{hvec}}$ is the half-vector representation of $\hat{\mathbf{F}}$. Note that a vector representation $\hat{\mathbf{F}}_{\text{vec}}$ of $\hat{\mathbf{F}}$ reads as $\left[f_{11}, \ldots, f_{m1}, f_{12}, \ldots, f_{mn}\right]^T$. Since $\hat{\mathbf{F}}$ is symmetric, it can be represented through the half-vector representation, $\hat{\mathbf{F}}_{\text{hvec}}$, which is obtained by eliminating all superdiagonal elements of $\hat{\mathbf{F}}$ from $\hat{\mathbf{F}}_{\text{vec}}$ [24]. $\hat{\mathbf{F}}_{\text{hvec}}$ in Eq. 11 is hence defined as $\left[\hat{f}_{11}, \ldots, \hat{f}_{dd}, \hat{f}_{12}, \ldots, \hat{f}_{d(d-1)}\right]^T$, where the diagonal elements are located in the first components of the vector.

4.2 Tuning ESN by Exploiting FIM Properties

In the following, we define the procedure to identify the edge of criticality, here defined as parameter configurations $\mathcal{K} \subset \Theta$ that maximize the ESN computational

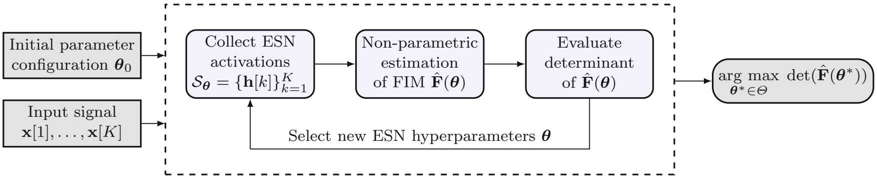

Initial parameter configuration $\boldsymbol{\theta}_0$
Input signal $\mathbf{x}[1], \ldots, \mathbf{x}[K]$
Collect ESN activations $\mathcal{S}_{\boldsymbol{\theta}} = \{\mathbf{h}[k]\}_{k=1}^{K}$
Non-parametric estimation of FIM $\hat{\mathbf{F}}(\boldsymbol{\theta})$
Evaluate determinant of $\hat{\mathbf{F}}(\boldsymbol{\theta})$
Select new ESN hyperparameters $\boldsymbol{\theta}$
$\arg\max_{\boldsymbol{\theta}^* \in \Theta} \det(\hat{\mathbf{F}}(\boldsymbol{\theta}^*))$

Fig. 7 Schematic, high-level description of the proposed procedure. Taken from [22]

capability. Figure 7 shows a schematic description of the main phases involved in the proposed method.

In order to determine $\mathcal{K}$, we introduce an algorithm that take advantage of the FIM properties on a system undergoing a continuous phase transition. FIM defines a metric tensor for the smooth manifold of parametric PDFs embedded in Θ [38], providing thus a geometric characterization of the system under analysis. It is possible to prove [33] that $\mathcal{K}$ corresponds to a region in Θ characterized by the largest volume (high concentration of parametric PDFs). This geometric result is reflected in the determinant $\det(\mathbf{F}(\boldsymbol{\theta}))$, which is monotonically related to the aforementioned volume in parameter space. Therefore, considering that the FIM is a PD matrix, and hence its determinant is always non-negative, we identify $\mathcal{K}$ with all those hyperparameters $\boldsymbol{\theta}^*$ for which:

$$\boldsymbol{\theta}^* = \arg\max_{\theta \in \Theta} \det(\mathbf{F}(\boldsymbol{\theta})). \tag{12}$$

Algorithm 1 delivers the pseudo-code of the proposed procedure. The impact provided by the variation of the control parameters $\boldsymbol{\theta}$ on the resulting ESN state cannot be described analytically without making further assumptions [31]: the (unknown) input signal driving the network plays an important role in the resulting ESN dynamics. Therefore, in order to calculate $\mathbf{F}(\boldsymbol{\theta})$, in Algorithm 1 we rely on the non-parametric FIM estimator described in Sect. 4.1. The estimation of the FIM for a given $\boldsymbol{\theta}$ is performed by analyzing the sequence $\mathcal{S}_{\theta} = \{\mathbf{h}[t]\}_{t=1}^{t_{\max}}$ of reservoir neuron activations. Since $\mathbf{h}[t] \in [-1, 1]^{N_r}$, the domain of the PDF in (8) is defined as $\mathcal{D} = [-1, 1]^{N_r}$. Additional sequences of activations, $\mathcal{S}_{\hat{\theta}_j}$, are considered (see line 7), which are obtained by perturbing M times the current network configuration $\boldsymbol{\theta}$ under analysis, and processing the same input $\mathbf{x}$. Perturbations are introduced by means of a small zero-mean noise with spherical covariance matrix, thus characterized by a single scalar parameter σ controlling the magnitude of the perturbation. FIM is estimated according to Eq. 11. In order to make the estimation more robust, we follow an ensemble approach and perform a number of trials (see line 3). The determinant is computed only once on the resulting average FIM, which is obtained by using T different (and independent) random realizations of the ESN architecture (see line 16).

Algorithm 1 Procedure for determining an ESN configuration on the edge of criticality.

```
Input: An ESN architecture, input $\mathbf{x} = \{x[i]\}_{i=1}^{t_{\max}}$, quantized parameter space $\Theta$, standard deviation
    $\sigma$ for the perturbations, number of trials $T$ and perturbations $M$.
Output: A configuration $\boldsymbol{\theta}^* \in \mathcal{K}$
Select an initial parameter configuration, $\boldsymbol{\theta} \in \Theta$; maximum $\eta = 0$
loop
    for $t = 1$ to $T$ do
        Randomly initialize the ESN weight matrices
        Configure ESN with $\boldsymbol{\theta}$ and process input $\mathbf{x}$
        Collect the related activations $\mathcal{S}_{\boldsymbol{\theta}} = \{\mathbf{h}[i]\}_{i=1}^{t_{\max}}$
        for $j = 1$ to $M$ do
            Generate a perturbation vector $\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{d \times d})$
            Randomly initialize the ESN weight matrices
            Configure ESN with perturbed version $\hat{\boldsymbol{\theta}}_j = \boldsymbol{\theta} + \mathbf{r}_j$ and process input $\mathbf{x}$
            Collect the related activations $\mathcal{S}_{\hat{\boldsymbol{\theta}}_j} = \{\mathbf{h}[i]\}_{i=1}^{t_{\max}}$
        end for
        Define $\mathcal{S}_{\hat{\boldsymbol{\theta}}} = \cup_{j=1}^{M} \mathcal{S}_{\hat{\boldsymbol{\theta}}_j}$
        Estimate the FIM $\mathbf{F}^{(t)}(\boldsymbol{\theta})$ of trial $t$ using $\mathcal{S}_{\boldsymbol{\theta}}$ and $\mathcal{S}_{\hat{\boldsymbol{\theta}}}$ with the non-parametric estimator
        introduced in Sect. 4.1
    end for
    Compute the average FIM, $\mathbf{F}(\boldsymbol{\theta})$, using all $\mathbf{F}^{(t)}(\boldsymbol{\theta}), t = 1, \ldots, T$
    if $\det(\mathbf{F}(\boldsymbol{\theta})) > \eta$ then
        Update $\eta = \det(\mathbf{F}(\boldsymbol{\theta}))$ and $\boldsymbol{\theta}^* = \boldsymbol{\theta}$
    end if
    if Stop criterion is met then
        return $\boldsymbol{\theta}^*$
    else
        Select a new $\boldsymbol{\theta} \in \Theta$ based on a suitable search scheme
    end if
end loop
```

4.3 Results

In the following, we compare the agreement between the hyperparameter configurations identified by the unsupervised FIM-based approach as the edge of criticality, with the configurations where supervised performance measures are maximized. Specifically, we consider the prediction accuracy, defined as $\gamma = \max\{1 - \mathrm{NRMSE}, 0\}$, where NRMSE is the Normalized Root Mean Squared Error of the ESN. Then, we account the memory capacity (MC), which quantifies the capability of ESN to remember previous inputs, relative to an i.i.d. signal. MC is measured as the squared correlation coefficient between the desired output, which is the input signal delayed by different delays $\delta > 0$, and the observed network output $\mathbf{y}[t]$:

$$\mathrm{MC} = \sum_{\delta=1}^{\delta_{\max}} \frac{\mathrm{cov}^2\left(\mathbf{x}[t-\delta], \mathbf{y}[t]\right)}{\mathrm{var}\left(\mathbf{x}[t-\delta]\right) \mathrm{var}\left(\mathbf{y}[t]\right)}. \tag{13}$$

MC is computed by training several readout layers, one for each delay $\delta \in \{1, 10, \ldots, 100\}$, while keeping fixed input and reservoir layers.

To test the effectiveness of the identified edge of criticality in terms of forecast accuracy, we consider the prediction of the sinusoid and the MG time-series. We also take into account the NARMA task,

$$\mathbf{y}[t+1] = 0.3\mathbf{y}[t] + 0.05\mathbf{y}[t]\left(\sum_{i=0}^{r-1} \mathbf{y}[t-i]\right) + 1.5\mathbf{x}[t-r]\mathbf{x}[t] + 0.1, \tag{14}$$

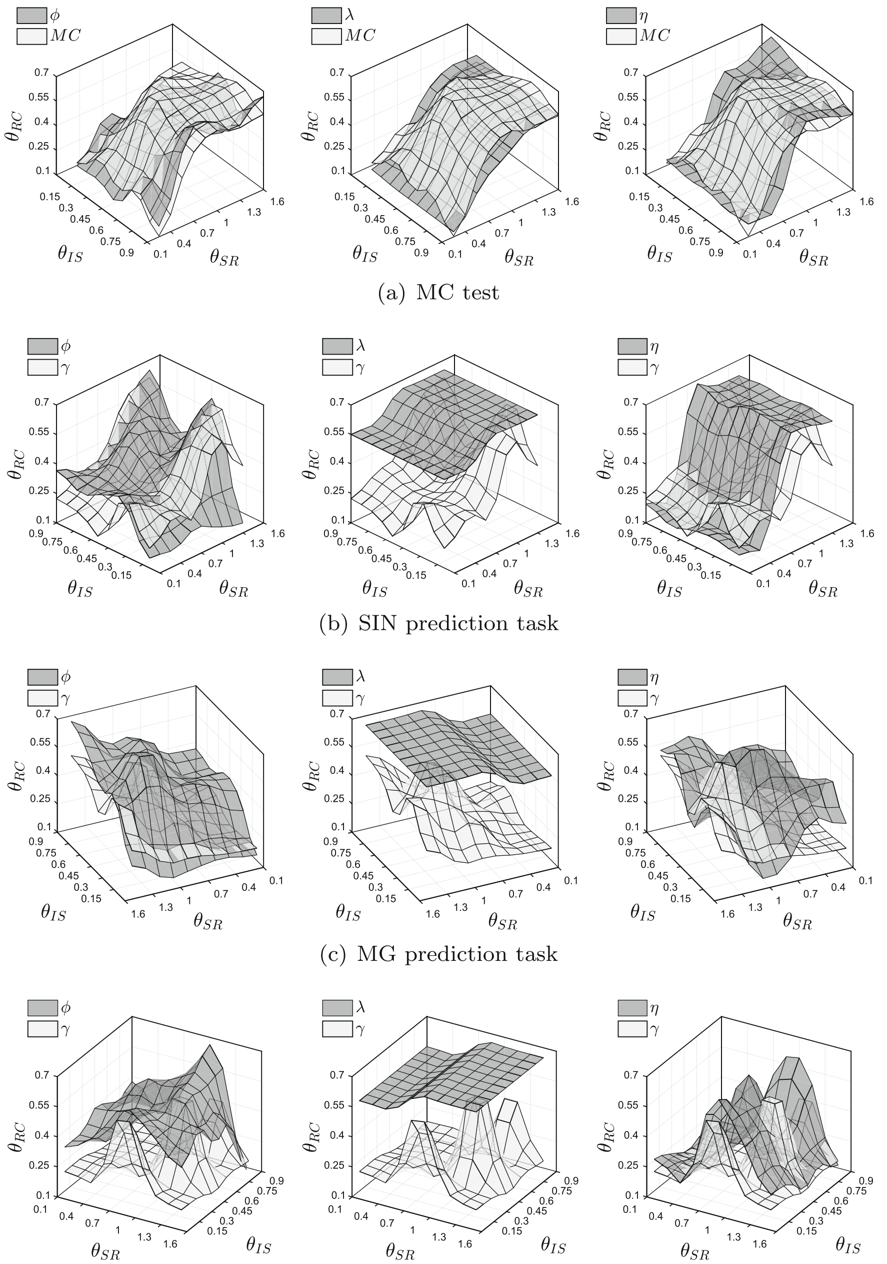
being $\mathbf{x}[t]$ an i.i.d. uniform noise in [0, 1].

In addition to the spectral radius θ_{SR} and the input scaling θ_{IS}, we consider also the effect of the density of the reservoir connections θ_{RC} as a core hyperparameter. The hyperparameters are searched in a discretized space through a grid search, which considers 10 different configurations for each parameter. Specifically, we search for the spectral radius θ_{SR} in [0.4, 1.6], input scaling θ_{IS} in [0.3, 0.8], and reservoir connectivity θ_{RC} in [0.1, 0.7], evaluating a total of 1000 hyperparameter configurations. Since we considered a parameter space with three dimensions, the related edge of criticality $\mathcal{K}$ is a two-dimensional manifold embedded in such a three-dimensional space. For each hyperparameter configuration, in Algorithm 1 we perform $T = 10$ independent trials and $M = 80$ perturbations to compute the ensemble average of the FIM; the variance for the perturbations is set to $\sigma^2 = 0.25$. In each trial, we sample new (and independent) input and reservoir connection weights (W_i^r and W_r^r).

In Fig. 8, we report the critical regions of the parameter space identified in each test by: maximization of FIM determinant, denoted by ϕ, zero-crossing of MLLE (λ), and maximization of minimum singular value of the Jacobian (η). The light gray manifold corresponds to the regions in parameter space where the performance of the network is maximized and the dark gray manifolds represent ϕ, λ, and η. In Table 3, we report the numerical values of the correlations between the light gray manifold and the dark gray ones.

The numerical values of the correlations are reported in Table 3. As it is possible to notice in Fig. 8a, the critical regions identified by each one of the three methods follow with good accuracy the region of the hyperparameter space where MC is maximized. The degrees of correlation for the MC task are described in Table 3. It is interesting to note that λ shows a very high correlation (81%) preforming better than η for this task. The correlation between ϕ and the region with maximum MC is also very high (75%), showing that both ϕ and λ can be used as reliable indicators to identify the optimal configurations that enhance the short-term memory capacity of ESNs. The p-values for each correlation measure are lower than 0.05, indicating statistical significance of the results.

Relative to the prediction of the sinusoid, as it is possible to observe in Fig. 8b, both ϕ and η are consistent with γ, while λ shows a lower agreement. From Table 3, we see that ϕ achieves the best results, all the measures have positive degrees of correlation with γ and small p-values (hence statistical significance).



(a) MC test

(b) SIN prediction task

(c) MG prediction task

(d) NARMA prediction task

Fig. 8 In each figure, the light gray manifold represents configurations of spectral radius (θ_{SR}), input scaling (θ_{IS}), and reservoir connectivity (θ_{RC}) that maximize Memory Capacity (MC) or prediction accuracy (γ). The dark gray manifolds represent (from left to right): configurations where the FIM determinant is maximized (ϕ); configurations where MLLE crosses zero (λ); configurations where mSVJ is maximized (η). Taken from [22]

Table 3 Correlations between the regions where FIM determinant is maximized (ϕ), MLLE crosses zero (λ), minimum singular value of the Jacobian is maximized (η) and performances are maximized (γ/MC). Best results are shown in bold, p-values are reported in brackets

Test	Corr (ϕ, γ/MC)	(λ, γ/MC)	Corr (η, γ/MC)
MC	0.75 (1e-5)	**0.81** (1e-8)	0.65 (1e-4)
Predict—SIN	**0.58** (0.02)	0.52 (1e-3)	0.56 (1e-3)
Predict—MG	**0.71** (1e-5)	0.66 (1e-4)	0.38 (0.06)
Predict—NARMA	**0.52** (0.01)	0.25 (0.22)	0.48 (0.02)

In MG test, both ϕ and λ provide better results than η to identify the optimal configuration, as we can see from Fig. 8c and the results in the table. Notably, the correlation between γ and η has a p-value beyond the confidence level 0.05, suggesting that correlations are not different from zero.

According to the results shown in Fig. 8d and Table 3, in the NARMA task ϕ and η perform significantly better than λ for identifying the critical region. If fact, the correlation between γ and λ is low and not statistically significant. Even in this case, the best results in terms of correlation are achieved by ϕ.

5 Concluding Remarks and Future Research Perspectives

In this chapter, we presented recent research developments for the characterization and tuning of echo state networks. We have shown how recurrence plots can be generated from reservoir neurons activations and exploited by the designer as visual tools to analyze the response of the network to a specific input. Recurrence plots provide an immediate visual interpretation of network stability: short and erratic diagonal lines denote instability/chaoticity, while long diagonal lines denote regularity (e.g., a periodic motion). Through the recurrence quantification analysis, the designer can deduce important and consistent conclusions about the behavior of the network, depending on the actual input driving the system and the current configurations of the hyperparameters.

Successively, we discussed a method that establishes a connection between the notion of continuous phase transition, echo state networks, and Fisher information. Based on this interplay, we have developed a principled approach to configure ESNs on the edge of criticality, where computational capability (defined in terms of prediction performance and short-term memory capacity) is maximized. The proposed methodology is completely unsupervised and it opens new perspectives for analyzing the dynamics of driven recurrent neural networks. Fisher information requires analytic knowledge of the distribution ruling the system. To address this issue, we have followed an ensemble estimation approach based on a recently proposed nonparametric FIM estimator, which, thanks to a graph-based representation of the data, is also applicable to high-dimensional densities. This last aspect plays a fundamental role in our domain of application, since we analyze the network through a multivariate sequence of reservoir neuron activations; hence the number of dimensions is determined by the number of reservoir neurons. We evaluated the proposed method on benchmarks of short-term memory capacity and prediction accuracy, to identify the ESN hyperparameters maximizing the computational capability. We compared our method with established criteria based on the sign of the maximum local Lyapunov exponent and the minimum singular value of the Jacobian. Our experiments demonstrated that the FIM-based approach achieves comparable or even better accuracy than the two other indicators in identifying the onset of criticality.

The methodologies discussed here are independent of the particular task at hand and offer an insight on the dynamics and actual functioning of the network. In this sense, the proposed framework of analysis represents a step forward to the understanding of these systems that, even if are capable of solving efficiently a variety of tasks, are often treated as black boxes. We believe that, the linkage of methods from the complex systems field with recurrent neural networks offers the potential to disclose a whole new set of opportunities for further studies and applications. Our future directions point toward graph-based approaches, which demonstrated to be powerful tools to represent complex systems and to model their dynamics when observed through time-series [19].

References

1. Aljadeff, J., Stern, M., Sharpee, T.: Transition to chaos in random networks with cell-type-specific connectivity. Phys. Rev. Lett. **114**, 088101 (2015). doi:10.1103/PhysRevLett.114.088101
2. Barzel, B., Barabási, A.-L.: Universality in network dynamics. Nat. Phys. **9**(10), 673–681 (2013). doi:10.1038/nphys2741
3. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994). ISSN 1045-9227. doi:10.1109/72.279181
4. Berisha, V., Hero, A.Q. III.: Empirical non-parametric estimation of the Fisher information. IEEE Signal Process. Lett. **22**(7), 988–992 (2015). ISSN 1070-9908. doi:10.1109/LSP.2014.2378514
5. Bertschinger, N., Natschläger, T.: Real-time computation at the edge of chaos in recurrent neural networks. Neural Comput. **16**(7), 1413–1436 (2004). doi:10.1162/089976604323057443
6. Bianchi, F.M., Livi, L., Alippi, C.: Investigating echo state networks dynamics by means of recurrence analysis. IEEE Trans. Neural Netw. Learn. Syst. 1–13 (2016). doi:10.1109/TNNLS.2016.2630802
7. Boedecker, J., Obst, O., Lizier, J.T., Mayer, N.M., Asada, M.: Information processing in echo state networks at the edge of chaos. Theory Biosci. **131**(3), 205–213 (2012). doi:10.1007/s12064-011-0146-8
8. Charles, A., Yin, D., Rozell, C.: Distributed sequence memory of multidimensional inputs in recurrent networks. arXiv:1605.08346 (2016)
9. De Arcangelis, L., Lombardi, F., Herrmann, H.J.: Criticality in the brain. J. Stat. Mech. Theory Exp. **2014**(3), P03026 (2014). doi:10.1088/1742-5468/2014/03/P03026

10. Enel, P., Procyk, E., Quilodran, R., Dominey, P.F.: Reservoir computing properties of neural dynamics in prefrontal cortex. PLoS Comput. Biol. **12**(6), e1004967 (2016). doi:10.1371/journal.pcbi.1004967
11. Eroglu, D., Peron, T.K.D.M., Marwan, N., Rodrigues, F.A., da Costa, L.F., Sebek, M., Kiss, I.Z., Kurths, J.: Entropy of weighted recurrence plots. Phys. Rev. E **90**(4), 042919 (2014). doi:10.1103/PhysRevE.90.042919
12. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**(2), 179–211 (1990). ISSN 0364-0213. doi:10.1016/0364-0213(90)90002-E
13. Friedman, J.H., Rafsky, L.C.: Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. Ann. Stat. **7**(4), 697–717 (1979)
14. Grigolini, P.: Emergence of biological complexity: criticality, renewal and memory. Chaos, Solitons Fractals (2015). doi:10.1016/j.chaos.2015.07.025
15. Hammer, B., Micheli, A., Sperduti, A., Strickert, M.: Recursive self-organizing network models. Neural Netw. **17**(8–9), 1061–1085 (2004). ISSN 0893-6080. doi:10.1016/j.neunet.2004.06.009
16. Hidalgo, J., Grilli, J., Suweis, S., Muñoz, M.A., Banavar, J.R., Maritan, A.: Information-based fitness and the emergence of criticality in living systems. Proc. Natl. Acad. Sci. **111**(28), 10095–10100 (2014). doi:10.1073/pnas.1319166111
17. Hidalgo, J., Grilli, J., Suweis, S., Maritan, A., Muñoz, M.A.: Cooperation, competition and the emergence of criticality in communities of adaptive systems. J. Stat. Mech. Theory Exp. **2016**(3), 033203 (2016). doi:10.1088/1742-5468/2016/03/033203
18. Jaeger, H.: The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, vol. 148, p. 34 (2001)
19. Lacasa, L., Nicosia, V., Latora, V.: Network structure of multivariate time series. Sci. Rep. **5**, (2015). doi:10.1038/srep15508
20. Langton, C.G.: Computation at the edge of chaos: Phase transitions and emergent computation. Phys. D Nonlinear Phenom. **42**(1), 12–37 (1990). doi:10.1016/0167-2789(90)90064-V
21. Legenstein, R., Maass, W.: Edge of chaos and prediction of computational performance for neural circuit models. Neural Netw. **20**(3), 323–334 (2007). doi:10.1016/j.neunet.2007.04.017
22. Livi, L., Bianchi, F.M., Alippi, C.: Determination of the edge of criticality in echo state networks through Fisher information maximization. IEEE Trans. Neural Netw. Learn. Syst. 1–12 (2017). doi:10.1109/TNNLS.2016.2644268
23. Maass, W., Joshi, P., Sontag, E.D.: Computational aspects of feedback in neural circuits. PLoS Comput. Biol. **3**(1), e165 (2007). doi:10.1371/journal.pcbi.0020165.eor
24. Magnus, J.R., Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, New York (1995)
25. Manjunath, G., Jaeger, H.: Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks. Neural Comput. **25**(3), 671–696 (2013). doi:10.1162/NECO_a_00411
26. Marichal, R.L., Piñeiro, J.D.: Analysis of multiple quasi-periodic orbits in recurrent neural networks. Neurocomputing **162**, 85–95 (2015). doi:10.1016/j.neucom.2015.04.001
27. Marwan, N.: How to avoid potential pitfalls in recurrence plot based data analysis. Int. J. Bifurcat. Chaos **21**(04), 1003–1017 (2011). doi:10.1142/S0218127411029008
28. Marwan, N., Kurths, J.: Line structures in recurrence plots. Phys. Lett. A **336**(4), 349–357 (2005). doi:10.1016/j.physleta.2004.12.056
29. Marwan, N., Carmen, M., Thiel, R.M., Kurths, J.: Recurrence plots for the analysis of complex systems. Phys. Rep. **438**(5), 237–329 (2007). doi:10.1016/j.physrep.2006.11.001
30. Marwan, N., Schinkel, S., Kurths, J.: Recurrence plots 25 years later-Gaining confidence in dynamical transitions. EPL (Europhys. Lett.) **101**(2), 20007 (2013). doi:10.1209/0295-5075/101/20007
31. Massar, M., Massar, S.: Mean-field theory of echo state networks. Phys. Rev. E **87**(4), 042809 (2013). doi:10.1103/PhysRevE.87.042809

32. Massobrio, P., de Arcangelis, L., Pasquale, V., Jensen, H.J., Plenz, D.: Criticality as a signature of healthy neural systems. Front. Syst. Neurosci. **9**, 22 (2015). doi:10.3389/fnsys.2015.00022
33. Mastromatteo, I., Marsili, M.: On the criticality of inferred models. J. Stat. Mech. Theory Exp. **2011**(10), P10012 (2011). doi:10.1088/1742-5468/2011/10/P10012
34. Mora, T., Bialek, W.: Are biological systems poised at criticality? J. Stat. Phys. **144**(2), 268–302 (2011). doi:10.1007/s10955-011-0229-4
35. Mora, T., Deny, S., Marre, O.: Dynamical criticality in the collective activity of a population of retinal neurons. Phys. Rev. Lett. **114**(7), 078105 (2015). doi:10.1103/PhysRevLett.114.078105
36. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. arXiv:1211.5063 (2012)
37. Peng, Y., Lei, M., Li, J.-B., Peng, X.-Y.: A novel hybridization of echo state networks and multiplicative seasonal ARIMA model for mobile communication traffic series forecasting. Neural Comput. Appl. **24**(3–4), 883–890 (2014)
38. Prokopenko, M., Lizier, J.T., Obst, O., Wang, X.R.: Relating Fisher information to order parameters. Phys. Rev. E **84**(4), 041116 (2011). doi:10.1103/PhysRevE.84.041116
39. Rajan, K., Abbott, L.F., Sompolinsky, H.: Stimulus-dependent suppression of chaos in recurrent neural networks. Phys. Rev. E **82**(1), 011903 (2010). doi:10.1103/PhysRevE.82.011903
40. Reinhart, R.F., Steil, J.J.: Regularization and stability in reservoir networks with output feedback. Neurocomputing **90**, 96–105 (2012). doi:10.1016/j.neucom.2012.01.032
41. Roli, A., Villani, M., Filisetti, A., Serra, R.: Dynamical criticality: overview and open questions. arXiv:1512.05259 (2015)
42. Rumelhart, D.E., Smolensky, P., McClelland, J.L., Hinton, G.: Sequential thought processes in pdp models. V **2**, 3–57 (1986)
43. Scheffer, M., Bascompte, J., Brock, W.A., Brovkin, V., Carpenter, S.R., Dakos, V., Held, H., Van Nes, E.H., Rietkerk, M., Sugihara, G.: Early-warning signals for critical transitions. Nature **461**(7260), 53–59 (2009). doi:10.1038/nature08227
44. Scheffer, M., Carpenter, S.R., Lenton, T.M., Bascompte, J., Brock, W., Dakos, V., van De Koppel, J., van De Leemput, I.A., Levin, S.A., van Nes, E.H., Pascual, M., Vandermeer, J.: Anticipating critical transitions. Science **338**(6105), 344–348 (2012). doi:10.1126/science.1225244
45. Schiller, U.D., Steil, J.J.: Analyzing the weight dynamics of recurrent learning algorithms. Neurocomputing **63**, 5–23 (2005). doi:10.1016/j.neucom.2004.04.006
46. Shen, Y., Wang, J.: An improved algebraic criterion for global exponential stability of recurrent neural networks with time-varying delays. IEEE Trans. Neural Netw. **19**(3), 528–531 (2008). ISSN 1045-9227. doi:10.1109/TNN.2007.911751
47. Steil, J.J.: Memory in backpropagation-decorrelation o(n) efficient online recurrent learning. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds.) Artificial Neural Networks: Formal Models and Their Applications-ICANN 2005, pp. 649–654. Springer, Berlin, Heidelberg (2005)
48. Sussillo, D.: Neural circuits as computational dynamical systems. Curr. Opin. Neurobiol. **25**, 156–163 (2014). doi:10.1016/j.conb.2014.01.008
49. Sussillo, D., Barak, O.: Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. Neural Comput. **25**(3), 626–649 (2013). doi:10.1162/NECO_a_00409
50. Tiňo, P., Rodan, A.: Short term memory in input-driven linear dynamical systems. Neurocomputing **112**, 58–63 (2013). doi:10.1016/j.neucom.2012.12.041
51. Tkačik, G., Bialek, W.: Information processing in living systems. Ann. Rev. Condens. Matter Phys. **7**(1), 89–117 (2016). doi:10.1146/annurev-conmatphys-031214-014803
52. Tkačik, G., Mora, T., Marre, O., Amodei, D., Palmer, S.E., Berry, M.J., Bialek, W.: Thermodynamics and signatures of criticality in a network of neurons. Proc. Natl. Acad. Sci. **112**(37), 11508–11513 (2015). doi:10.1073/pnas.1514188112
53. Torres, J.J., Marro, J.: Brain performance versus phase transitions. Sci. Rep. **5** (2015). doi:10.1038/srep12216
54. Toyoizumi, T., Abbott, L.F.: Beyond the edge of chaos: amplification and temporal integration by recurrent networks in the chaotic regime. Phys. Rev. E **84**(5), 051908 (2011). doi:10.1103/PhysRevE.84.051908

55. Toyoizumi, T., Aihara, K., Amari, S.-I.: Fisher information for spike-based population decoding. Phys. Rev. Lett. **97**(9), 098102 (2006). doi:10.1103/PhysRevLett.97.098102
56. Verstraeten, D., Schrauwen, B.: On the quantification of dynamics in reservoir computing. In: Artificial Neural Networks–ICANN 2009, pp. 985–994. Springer, Berlin (2009). doi:10.1007/978-3-642-04274-4_101
57. Verstraeten, D., Schrauwen, B., D'Haene, M., Stroobandt, D.: An experimental unification of reservoir computing methods. Neural Netw. **20**(3), 391–403 (2007). ISSN 0893-6080. doi:10.1016/j.neunet.2007.04.003. Echo State Networks and Liquid State Machines
58. Wainrib, G., Touboul, J.: Topological and dynamical complexity of random neural networks. Phys. Rev. Lett. **110**, 118101 (2013). doi:10.1103/PhysRevLett.110.118101
59. Wang, X., Lizier, J., Prokopenko, M.: Fisher information at the edge of chaos in random boolean networks. Artif. Life **17**(4), 315–329 (2011). ISSN 1064-5462. doi:10.1162/artl_a_00041
60. Werbos, P.J.: Backpropagation: past and future. Proc. IEEE Int. Conf. Neural Netw. **1**, 343–353 (1988). doi:10.1109/ICNN.1988.23866
61. Yildiz, I.B., Jaeger, H., Kiebel, S.J.: Re-visiting the echo state property. Neural Netw. **35**, 1–9 (2012). doi:10.1016/j.neunet.2012.07.005
62. Zegers, P.: Fisher information properties. Entropy **17**(7), 4918–4939 (2015). doi:10.3390/e17074918
63. Zhang, B., Miller, D.J., Wang, Y.: Nonlinear system modeling with random matrices: echo state networks revisited. IEEE Trans. Neural Netw. Learn. Syst. **23**(1), 175–182 (2012). ISSN 2162-237X. doi:10.1109/TNNLS.2011.2178562
64. Zhang, Y., Wang, J.: Global exponential stability of recurrent neural networks for synthesizing linear feedback control systems via pole assignment. IEEE Trans. Neural Netw. **13**(3), 633–644 (2002). ISSN 1045-9227. doi:10.1109/TNN.2002.1000129