The Spectral Radius Remains a Valid Indicator of the Echo State Property for Large Reservoirs

Ken Caluwaerts, Francis wyffels, Sander Dieleman and Benjamin Schrauwen

Abstract— In the field of Reservoir Computing, scaling the spectral radius of the weight matrix of a random recurrent neural network to below unity is a commonly used method to ensure the Echo State Property. Recently it has been shown that this condition is too weak. To overcome this problem, other – more involved – sufficient conditions for the Echo State Property have been proposed. In this paper we provide a large-scale experimental verification of the Echo State Property for large recurrent neural networks with zero input and zero bias. Our main conclusion is that the spectral radius method remains a valid indicator of the Echo State Property; the probability that the Echo State Property does not hold, drops for larger networks with spectral radius below unity, which are the ones of practical interest.

I. INTRODUCTION

RESERVOIR COMPUTING is a simple, yet efficient method to train large recurrent networks [1]. It has been successfully applied to a broad range of tasks. The core of the method is the reservoir, a large randomly connected recurrent neural network. Training consists of adjusting the readout layer which is a linear mapping from the neuron states to the output.

To ensure applicability, the reservoir must exhibit the Echo State Property, which prescribes that the system forgets its inputs after a limited amount of time [1]. A commonly used indicator for the Echo State Property is the spectral radius, the supremum among the absolute values of the eigenvalues of the reservoir weight matrix. It is commonly assumed that the spectral radius ought to be smaller than unity in order for a reservoir to exhibit the Echo State Property. Recently, Yildiz et al. (2012) [2] explicitly illustrated – by means of low-dimensional networks – that a spectral radius smaller than unity does in fact not necessarily result in a network for which the Echo State Property holds. While the counterexamples in [2] were small and their extensions to larger networks lead to relatively sparse neural networks, we also found large, densely connected reservoirs which do not exhibit the Echo State Property. Fortunately, our results indicate that such systems occur rarely. We found that both reservoir size and connectivity influence the Echo State Property. Large, densely connected reservoir systems can be used in order to avoid this anomaly.

Ken Caluwaerts, Francis wyffels, Sander Dieleman and Benjamin Schrauwen are with the Reservoir Lab, Electronics and Information Systems departement, Ghent University, Belgium (email: {ken.caluwaerts, francis.wyffels, sander.dieleman, benjamin.schrauwen}@ugent.be).

Despite efforts by many authors (see for example [2]–[5]), the misconception that the spectral radius of *input driven* reservoir systems should always be smaller than unity is still vivid among many researchers.

The remainder of this paper is structured as follows. In Section II we briefly recapitulate the concept of Reservoir Computing. Next, in Section III, we elaborate on the findings of Yildiz et al. (2012) [2]. We will show by means of 2- and 8-dimensional networks that other metrics such as the Schur stability are very restrictive. In Section IV we extend our results to large recurrent neural networks and investigate the influence of reservoir size and connectivity on the Echo State Property. Finally, in Section V, we elaborate on our findings and draw the conclusions.

II. RESERVOIR COMPUTING

Reservoir Computing (RC) is an approach for efficient training of large recurrent neural networks. Typically, a network of randomly connected neurons – the reservoir – is created, excited with one or more inputs and then trained by adjusting the readout weights using linear regression. A schematic overview of an RC system is given in Fig. 1. Formally, for a reservoir with N neurons, the weights of the connections within the reservoir are represented by a matrix $\mathbf{W}_{\mathrm{res}}$ of size $N \times N$. Additionally, matrices $\mathbf{W}_{\mathrm{in}}$ and $\mathbf{W}_{\mathrm{bias}}$ represent the connection weights from the input to the reservoir and from a bias to the reservoir respectively. Typically, $\mathbf{W}_{\mathrm{bias}}$ has dimensions $N \times 1$ and $\mathbf{W}_{\mathrm{in}}$ has dimensions $N \times I$, where I is equal to the number of inputs to the reservoir. After sampling these weights from a random distribution, e.g. a standard normal distribution, the update of the system's state $\mathbf{x}$ at discrete time step k is defined by the following equation:

$$\mathbf{x}[k+1] = \tanh\Big(\mathbf{W}_{\mathrm{res}}\mathbf{x}[k] + \mathbf{W}_{\mathrm{in}}\mathbf{u}[k+1] + \mathbf{W}_{\mathrm{bias}}\Big), \quad (1)$$

where $\mathbf{u}$ is the input of the system. While sometimes other squashing functions are used, we only consider the hyperbolic tangent function in this work.

The output $\mathbf{y}$ of a reservoir system is defined by:

$$\mathbf{y}[k] = \mathbf{W}_{\mathrm{out}}^{\mathrm{T}}\mathbf{x}[k], \quad (2)$$

where $\mathbf{W}_{\mathrm{out}}$ are the connection weights from the reservoir to the output. The dimensions of this weight matrix are $N \times O$, where O equals the number of outputs.

After construction of the weight matrices, the reservoir system can be trained by adjusting the readout weights $\mathbf{W}_{\mathrm{out}}$. As reported in [1], the main method for training is linear regression. In practice, training consists of two phases: (1)
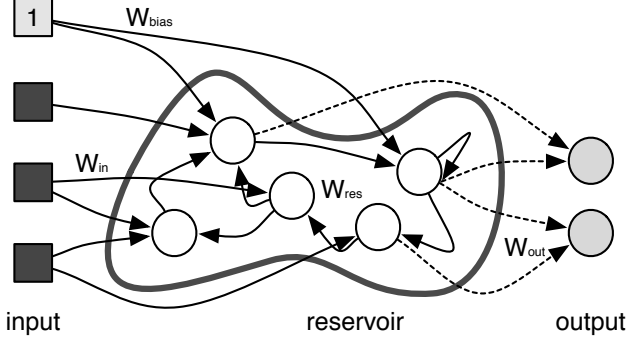
Fig. 1. Schematic overview of an RC system with multiple inputs and outputs. Only the readout weights (dashed connections) are trained.

collecting neuron states during the simulation phase, and (2) adjusting the readout weights by computation of the mean squared error.

During the *simulation phase* the neuron states are collected by stimulating the system using the inputs from the training dataset by applying equation 1. For every time step (from 1 to K) the neuron states (N neurons in total) are collected resulting in a state matrix $\mathbf{X}$ which has size $N \times K$. After collecting the state matrix $\mathbf{X}$, *training the readout weights* is done by linear regression. Let matrix $\mathbf{T}$ be the collection of desired outputs, the readout weights $\mathbf{W}_{\mathrm{out}}$ can then be obtained by minimizing the mean squared error of the linear mapping from the reservoir state matrix $\mathbf{X}$ to the desired output $\mathbf{T}$:

$$\mathbf{W}_{\mathrm{out}} = \arg_{\mathbf{W}} \min \|\mathbf{W}\mathbf{X} - \mathbf{T}\|_2. \qquad (3)$$

This leads to the following matrix solution:

$$\mathbf{W}_{\mathrm{out}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{T}. \qquad (4)$$

One of the key principles behind RC is the *Echo State Property* (ESP) introduced by [1]. A reservoir system exhibits the ESP if it forgets all previous input after a limited time, i.e. it cannot have infinitely long memory (cf. the concept of fading memory introduced in [6]). In other words, without any external input, the system's state should converge to a single fixed point. In order to tune the dynamics of the reservoir, many researchers use the spectral radius ρ, which is defined as the largest absolute eigenvalue of the reservoir weight matrix $\mathbf{W}_{\mathrm{res}}$. The effect of the spectral radius on the dynamics of the reservoir system becomes clear when we plot the bifurcation diagrams of the reservoirs. In Fig. 2 the bifurcation diagram is shown for a 128-dimensional network with neither input, nor bias. The bifurcation diagram shows the different equilibrium points (i.e. local extrema) of three randomly selected neurons of a simulated reservoir after many different initializations. For $\rho < 1.0$, one observes that the system's state converges to a fixed point at the origin. At $\rho = 1$ the system undergoes a bifurcation which makes the reservoir dependent on its initial condition. Consequently, the ESP does not hold anymore. This bifurcation point does not always occur for $\rho = 1$. When the system is fed an
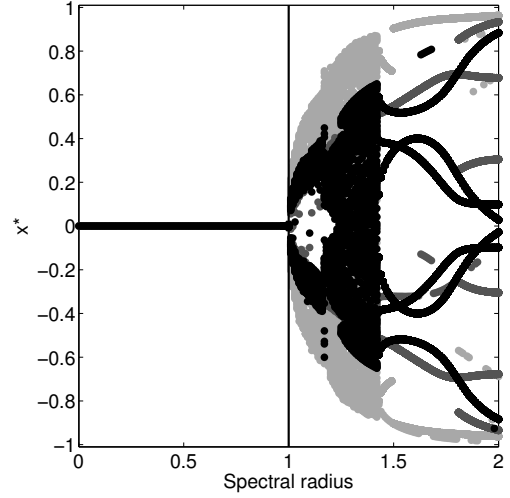

Fig. 2. Bifurcation diagram of a 128-dimensional reservoir with zero input and zero bias. The equilibrium points for three randomly selected neurons are visualised. By increasing the spectral radius, the system bifurcates from a single fixed point to spontaneous activity.
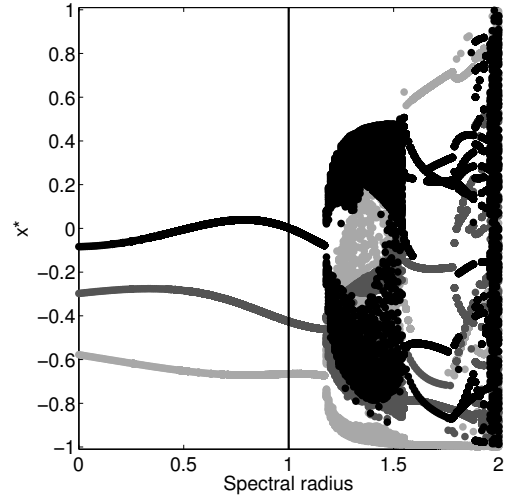

Fig. 3. Bifurcation diagram of 128-dimensional network fed a constant bias. The equilibrium points for three randomly selected neurons are visualised. By increasing the spectral radius, the system bifurcates from a singled fixed point to spontaneous activity. The constant input (e.g. bias) postpones this bifurcation point.

input signal or a constant bias, this bifurcation point can be observed for a spectral radius slightly larger than 1, see Fig. 3. In fact, due to the nonlinearity of the system, in almost all practical situations in which the reservoir is excited with one or more input signals, this will be the case.

III. ECHO STATE PROPERTY REVISITED

The Echo State Property (ESP) is a key concept in Reservoir Computing which can be formally stated as [1]:

Definition 1: A network $F : X \times U \rightarrow X$ (with the compactness condition) has the Echo State Property with respect to U, if for any left infinite input sequence

$u^{-\infty} \in U^{-\infty}$ and any two state vector sequences $x^{-\infty}$, $y^{-\infty} \in X^{-\infty}$ compatible with $u^{-\infty}$, it holds that $x_0 = y_0$.

Consequences of the ESP are that the current network state only depends on a certain number of previous inputs and is not influenced by the initial state after a certain period of time (often called warm-up period).

Reservoirs with the ESP can be used as nonlinear finite impulse response filters. Instead of crafting the neural network such that it performs a certain task (i.e. emulates some desired filter), one typically combines the available nonlinear projections of the reservoir in a linear fashion, assuming that the desired nonlinear computations are available in the system.

Therefore, it is customary to study the global properties of reservoirs with respect to a few parameters, such as the spectral radius, input bias and leak rate. In this spirit, the linear memory capacity [7] has been studied as well as the apparent tradeoff between linear memory and nonlinearity [8]. More recently it has been shown that any dynamical system (under some mild conditions) obeying the fading memory property (which is equivalent to the ESP) essentially has the same amount of computational power with respect to the number of observable variables of the system [9]. Dynamical systems with the ESP do however vary by the precise type of computations they offer, which need to be well-adjusted to the requirements of the application. Rules of thumb are that high bias networks are useful for highly nonlinear computations (because of the nonlinear behavior of the hyperbolic tangent) and a high spectral radius (near 1) results in longer memory [8], [10]. See [5] for a complete overview of design strategies.

One complication of the ESP is that it is input dependent. There has been some research into the input dependent ESP [11], but in practice it is often impossible to know all statistics of the input sequences beforehand. The ESP is thus mostly studied independently of the input sequence.

Different methods exist to verify if a given reservoir exhibits the ESP. The most commonly used method (for hyperbolic tangent networks) is to compute the spectral radius of the weight matrix ($\rho = \max_i |\lambda_i|$). If the spectral radius is below unity, one assumes that the ESP is fulfilled. The rationale for this approach is the fact that the hyperbolic tangent has the highest gain at the origin and one thus regards the linear system with the same weight matrix as an upper bound for the stability (eigenvalues within the unit disk).

Unfortunately the spectral radius method is not sufficient for the ESP (e.g. [2]) and it is possible to construct low spectral radius ($\rho << 1$) counterexamples. Consider Fig. 4, which shows the state progression for a 2-dimensional network with $\rho = 0.39$ and neither input, nor bias. We initialized the system in 1×10^5 random states and applied the update equation (equation 1). After a few iterations, all initial states contract into the origin or begin oscillating between two states. One intuitive explanation for such behavior is that the nonlinear network understates negative feedback. It is therefore possible to construct networks with very large weights, which have low spectral radius and do not exhibit the ESP.

Multiple sufficient conditions or tests for the ESP have been proposed. In his original work, Jaeger [1] proved that having the largest singular value of the weight matrix below unity is sufficient for the ESP. It is easy to show that $\max_i(\sigma_i) \geq \rho$, because any consistent matrix norm has a higher value than the spectral radius. Only for normal weight matrices ($\mathbf{W}_{\mathrm{res}}{}^*\mathbf{W}_{\mathrm{res}} = \mathbf{W}_{\mathrm{res}}\mathbf{W}_{\mathrm{res}}{}^*$), both norms coincide and the SVD condition is thus more restrictive than the spectral radius condition.

More recently, the ESP has been studied in terms of Lyapunov exponents [3], operator norms [12] and Schur stability [2]. However, there are multiple reasons to study the usefulness of the spectral radius method. First of all, the proposed methods are generally more complex to verify. Secondly, we shall show that the spectral radius method is often a tight bound for the ESP. Finally, RC is being extended to various domains (e.g. robotics [13], photonics [14] and electronics [15]). Given an equivalent to the weight matrix in another domain and an approximation of the maximum gain of the system, one can define an equivalent to the spectral radius. The performance of a system can thus be quantified in terms of the spectral radius in different domains.

We shall now consider the Schur stability method by Yildiz et al., the largest singular value test and the spectral radius method in more detail for small, zero input and zero bias networks.

The Schur stability method is stated as a linear matrix inequality condition [2]:

Definition 2: A zero bias hyperbolic tangent reservoir has the echo state property for any input if its weight matrix $\mathbf{W}_{\mathrm{res}}$ is diagonally Schur stable, i.e. there exists a diagonal matrix $\mathbf{P} > 0$ such that $\mathbf{W}_{\mathrm{res}}^{\mathrm{T}}\mathbf{P}\mathbf{W}_{\mathrm{res}} - \mathbf{P}$ is negative definite [16].

Fig. 5 and Fig. 6 show the fraction of rejected weight matrices for the different methods for 2 and 8 neuron reservoirs respectively with i.i.d. normally distributed weights as a function of the spectral radius averaged over 1×10^4 random reservoirs. To check if a network has the ESP, we initialized each network in 1×10^3 random (uniform $\in [-1, 1]^2$ or $\in [-1, 1]^8$) states and updated the network for 1×10^3 iterations. If the norm of any final state was above 1×10^{-7}, we considered the network to not have the ESP as not all states contracted to the origin. This is our baseline, indicated as *non-ESP*.

It becomes clear that the singular value test and the Schur method reject many networks for high spectral radii, while the fraction of reservoirs that effectively do not have the ESP for $\rho < 1$ is many times lower. Furthermore, the bounds tend to become weaker for larger networks, while the fraction of non-ESP networks becomes smaller.

IV. ECHO STATE PROPERTY IN LARGE RESERVOIRS

To study the ESP in large reservoirs, we performed an extensive numerical verification. These experiments are designed to show the influence of the reservoir size and connectivity on the ESP.
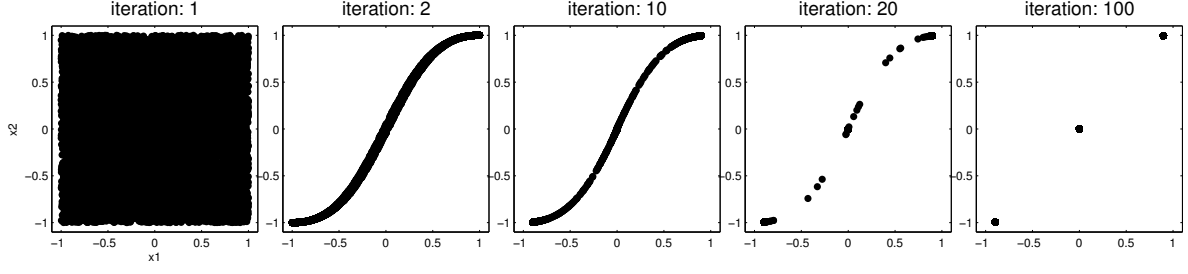
Fig. 4. A low spectral radius ($\rho = 0.39$) two neuron network without the Echo State Property. 1×10^5 initial states were sampled (uniform $\in [-1, 1]^2$) and we recorded the states for multiple time(SVD) steps (increasing from left to right). All initial states converge to the origin or continue to oscillate between $\pm[0.8975\,0.9946]^{\mathrm{T}}$. The weight matrix is given by $\mathbf{W}_{\mathrm{res}} = [-3\,1.24; -5.968\,2.416]$. This is not a degenerate case, as small variations of the weights also result in a non-ESP network (e.g. $A = [-3\,1.2; -6\,2.4]$, $\rho = 0.6$).
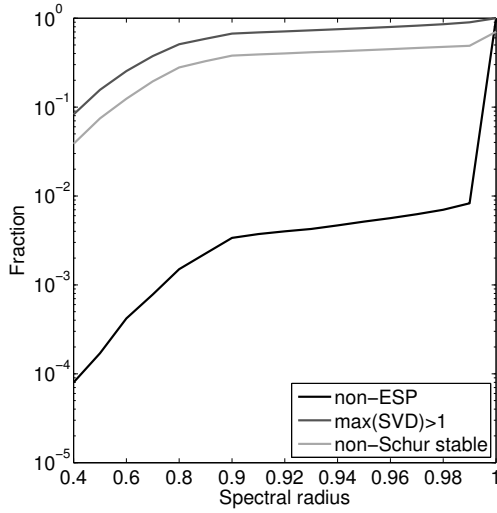

Fig. 5. Fraction of the 2-dimensional networks (zero input, zero bias) that do not have the experimentally verified ESP in function of ρ, fraction of the 2-dimensional networks with $\max(\mathrm{SVD}) > 1$ and fraction of the 2-dimensional networks which are not Schur stable.

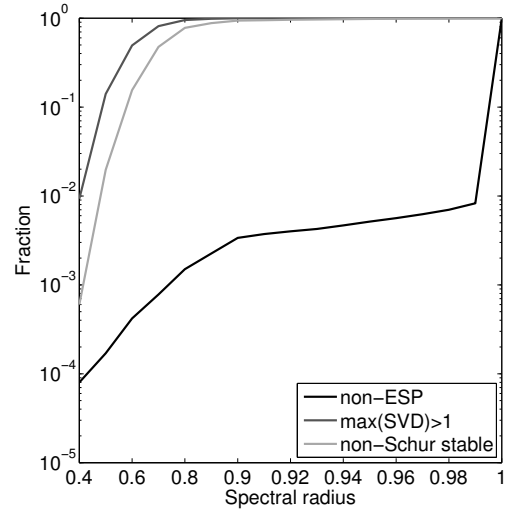Fig. 6. Fraction of the 8-dimensional networks (zero input, zero bias) that do not have the experimentally verified ESP in function of ρ, fraction of the 8-dimensional networks with $\max(\mathrm{SVD}) > 1$ and fraction of the 8-dimensional networks which are not Schur stable.

A. Experimental setup

In large-scale numerical experiments we varied the number of neurons N, the connectivity (the fraction of weights that is non-zero) c of the reservoir weight matrix and the spectral radius ρ:

$$N \in \{2, 4, 8, 16, 32, 64, 128, 256\}$$
$$c \in \{0.01, 0.0167, 0.0278,$$
$$0.0464, 0.0774, 0.1292,$$
$$0.2154, 0.3594, 0.5995, 1.0\}$$
$$\rho \in \{0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9,$$
$$0.91, 0.92, 0.93, 0.94, 0.95, 0.96,$$
$$0.97, 0.98, 0.99, 1, 1.01, 1.05, 1.1, 1.2\}.$$

Each parameter combination was tested for 1×10^5 randomly generated networks with weights $\mathbf{W}_{\mathrm{res}}$ sampled from a standard normal distribution and 1×10^3 random (uniform $\in [-1, 1]^N$) initial states. To test the ESP for a (zero-input)

network, each network was updated by applying equation 2 with zero bias and zero input ($\mathbf{u}$ and $\mathbf{W}_{\mathrm{bias}}$ equal to $\mathbf{0}$) for $1,000$ iterations for each initial state. We then stored the largest norm of the final states ($\|\mathbf{x}[1000]\|$). We present here the results for fully dense weight matrices with varying network size and for networks with 128 neurons with varying connectivity.

B. Results

Figs. 7 and 8 show the fraction of networks for which the ESP does not hold ($\|\mathbf{x}[1000]\| > 10^{-7}$) as a function of the spectral radius with respect to network size and connectivity respectively. One can observe in Fig. 7 that for relatively large (fully connected) networks ($N > 32$) the probability of finding a network without the ESP is below 1×10^{-3}. This is interesting because most reservoir systems of practical use are quite large ($N > 50$) and, consequently they are not affected by the fact that the spectral radius method is not a sufficient condition. Additionally, from Fig. 8 one learns that
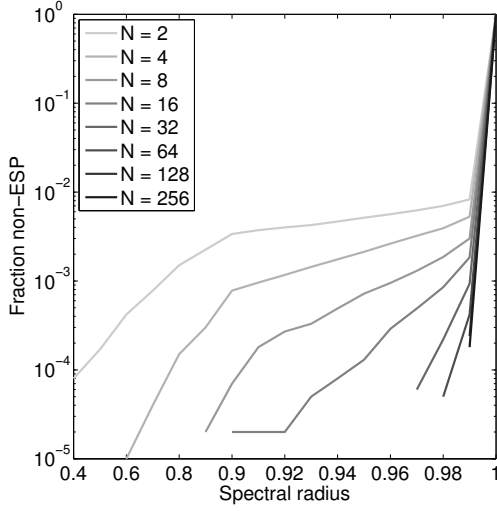
Fig. 7. Fraction of the fully connected networks with varying size for which the ESP does not hold in function of the spectral radius. The larger the network, the less likely that it does not exhibit the ESP for $\rho < 1.0$.

Fig. 9. Bifurcation diagram for 3 neurons of a fully connected 128-dimensional network. As can be observed in the bifurcation diagram, the ESP does *also* not hold for $\rho < 1.0$.
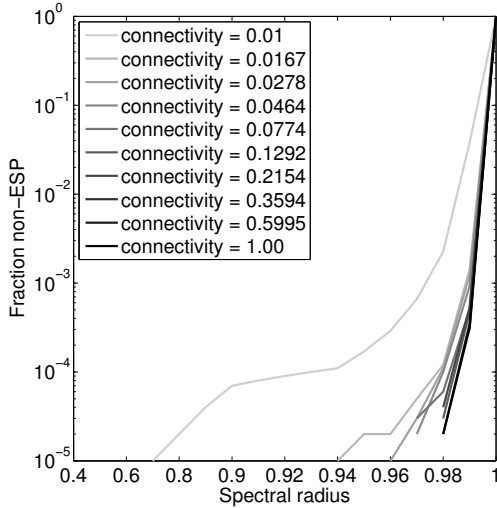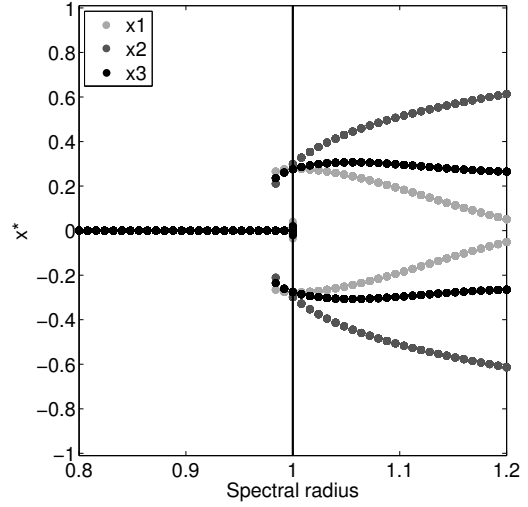

Fig. 8. Fraction of the 128-dimensional networks with varying connectivity for which the ESP does not hold in function of the spectral radius. Larger or denser networks with $\rho < 1.0$ are less likely not to exhibit the ESP.

the connectivity greatly influences the ESP. The sparser the network, the less likely it is to exhibit the ESP. For very sparse networks, with a connectivity of 1%, the fraction increases to 3.8% of the networks. In comparison, only 0.031% of the fully connected 128-dimensional networks did not exhibit the ESP. This observation corresponds with the intuition that it is very likely to find oscillating sub-networks, e.g. the small networks given in [2], in sparse networks.

In [2] a method is given for constructing large reservoir systems for which the ESP does not hold for spectral radii below 1. By following that procedure, the obtained networks will be sparse. A logical question is whether large dense networks can be found for which the ESP does not hold

with $\rho < 1$. The answer is positive, as we learned from Figs. 7 and 8. A bifurcation plot of such a system is given in Fig. 9. For $\rho < 0.98$ this system exhibits the ESP. At $\rho \approx 0.98$ the system bifurcates and starts to oscillate for some of the initial conditions. This contrasts with the behavior of a *normal* reservoir as depicted in Fig. 2.

V. DISCUSSION

The spectral radius is the most commonly used indicator for the dynamics of a reservoir. As a rule of thumb, it is assumed that a reservoir will exhibit the Echo State Property for spectral radii below unity. The ESP indicates that a reservoir has fading memory and thus that the network state will eventually become independent of the initial state and past inputs. One consequence of this is that a zero input, zero bias network has to converge to the origin (cf. Fig. 2).

However, as has been indicated in the past by a number of researchers (e.g. [5]) and pointed out explicitly in [2], this simple rule does not always hold. Low-dimensional examples in [2] illustrated that in some cases a network oscillates despite the fact that $\rho < 1$. These low-dimensional examples can be extended to higher-dimensional networks, however these are always sparse by construction. By large-scale numerical experimentation we also found explicit examples of large networks ($N \geq 128$) for which the ESP does not hold for $\rho < 1$ (see for example Fig. 9).

Fortunately, we found that the fraction of such networks rapidly decreases with increasing network size. Apart from reservoir size, the connectivity of the network also influences the ESP. In particular, we showed that sparser networks with $\rho < 1$ are more likely not to exhibit the ESP compared to dense reservoirs with equal spectral radius.

These findings do not suggest that the spectral radius should always be below 1. As the bifurcation plot in Fig. 3 indicates, inputs also greatly influence the ESP. We thus
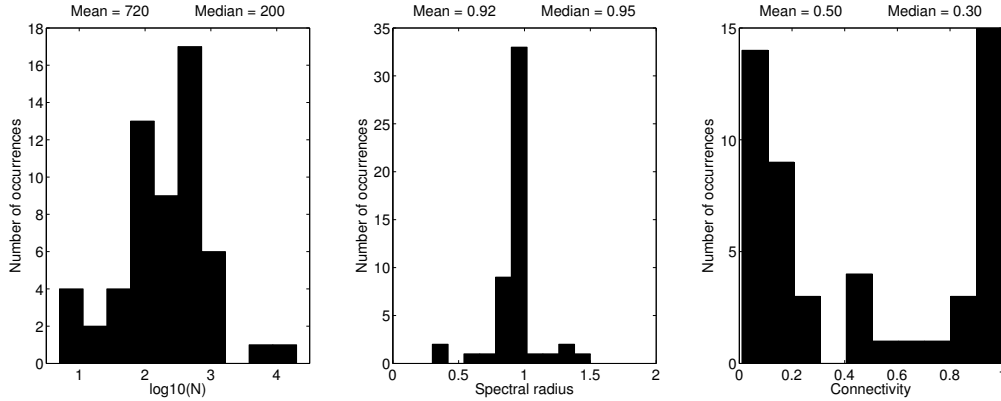
Fig. 10. Commonly used parameters for reservoir computing: the reservoir size N, spectral radius ρ and the connectivity c. All accessible studies citing [3] were consulted. Researchers tend to use relatively large networks. The majority of the researchers use a spectral radius slightly below unity. Typically very sparsely or very densely connected networks are preferred. We conclude that most studies used parameter ranges for which the spectral radius is a good indicator for the ESP.

advocate the exploration of larger ρ, like [2]. One should consider ρ as a task dependent global parameter for optimization.

To get an overview of how reservoirs are typically tuned, we analyzed all accessible papers citing [3] that used hyperbolic tangent neurons. We learned that the majority of the researchers use $0.9 < \rho < 1.0$, see Fig. 10. Based on practical experience, it is our belief that in many applications, the spectral radius should be much larger or much smaller and consequently that this distribution should be more bell-shaped.

Our meta-analysis also shows that most researchers are using large networks. This is positive, since the ESP is more likely to hold for $\rho < 1$ in such networks. More surprising is the connectivity used in many reservoirs. There seem to be two factions; one preferring fully dense networks for which the spectral radius seems to be a valid indicator for the ESP in general and the other preferring very sparse networks. Although the latter group comprises a significant portion of the studies, this does not necessarily indicate a problem as the networks were typically large.

In [1] and [2] different metrics for the ESP were given. By large-scale experimentation on 2- and 8-dimensional networks, we showed that both the largest singular value method and the Schur stability are too restrictive conditions for practical use. Further experiments on larger networks indicated that the probability of a network with spectral radius below unity not exhibiting the ESP quickly drops as a function of the network size for zero input, zero bias networks. Therefore, we conclude that the spectral radius remains a good indicator of the ESP, especially in large reservoir systems.

REFERENCES

[1] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks," German National Research Center for Information Technology, Tech. Rep. GMD Report 148, 2001.

[2] I. Yildiz, H. Jaeger, and S. Kiebel, "Re-visting the echo state property," Neural Networks, vol. 35, pp. 1–9, 2012.

[3] D. Verstraeten, B. Schrauwen, M. D'Haene, and D. Stroobandt, "An experimental unification of reservoir computing methods," Neural Networks, vol. 20, pp. 391–403, 2007.

[4] D. Verstraeten and B. Schrauwen, "On the quantification of dynamics in reservoir computing," in Lecture Notes in Computer Science, vol. 5768, 2009, pp. 985–994.

[5] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," Computer Science Review, vol. 3, pp. 127–149, 2009.

[6] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," Neural Computation, vol. 14, no. 11, pp. 2531–2560, 2002.

[7] M. Hermans and B. Schrauwen, "Memory in reservoirs for high dimensional input," in Proceedings of the International Joint Conference on Neural Networks, 2010.

[8] D. Verstraeten, J. Dambre, X. Dutoit, and B. Schrauwen, "Memory versus non-linearity in reservoirs," in Proceedings of the International Joint Conference on Neural Networks, 2010.

[9] J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar, "Information Processing Capacity of Dynamical Systems," Scientific Reports, vol. 2, 2012.

[10] S. Ganguli, D. Huh, and H. Sompolinsky, "Memory traces in dynamical systems," Proceedings of the National Academy of Sciences, vol. 105, no. 48, pp. 18 970–18 975, 2008.

[11] G. Manjunath and H. Jaeger, "Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks," Neural Computation, vol. 25, no. 3, pp. 671–696, 2013.

[12] M. Buehner and P. Young, "A tighter bound for the echo state property," IEEE Transactions on Neural Networks, vol. 17, pp. 820–4, May 2006.

[13] K. Caluwaerts, M. D'Haene, D. Verstraeten, and B. Schrauwen, "Locomotion Without a Brain: Physical reservoir computing in tensegrity structures," Artificial Life, vol. 19, pp. 35–66, 2013.

[14] K. Vandoorne, W. Dierckx, B. Schrauwen, D. Verstraeten, R. Baets, P. Bienstman, and J. Van Campenhout, "Toward optical signal processing using Photonic Reservoir Computing," Optics Express, vol. 16, no. 15, pp. 11 182–11 192, 2008.

[15] L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer, "Information processing using a single dynamical node as complex system," Nature Communications, vol. 2, 2011.

[16] E. Kaszkurewicz and A. Bhaya, "Matrix Diagonal and D-Stability," in Matrix Diagonal Stability in Systems and Computation. Birkhuser Boston, 2000, pp. 25–89.