Critical Echo State Networks

Márton Albert Hajnal and András Lőrincz

Eötvös Loránd University, Pázmány P. sétány 1/C, Budapest, Hungary, H-1117,
ouraborous@ludens.elte.hu, andras.lorincz@elte.hu
http://nipg.inf.elte.hu/

Abstract. We are interested in the optimization of the recurrent connection structure of Echo State Networks (ESNs), because their topology can strongly influence performance. We study ESN predictive capacity by numerical simulations on Mackey-Glass time series, and find that a particular small subset of ESNs is much better than ordinary ESNs provided that the topology of the recurrent feedback connections satisfies certain conditions. We argue that the small subset separates two large sets of ESNs and this separation can be characterized in terms of phase transitions. With regard to the criticality of this phase transition, we introduce the notion of Critical Echo State Networks (CESN). We discuss why CESNs perform better than other ESNs.

Keywords: time series, prediction, echo state network, phase transition, critical point.

1 Introduction

Motivation: We are interested in learning the dynamics of deterministic nonlinear systems with artificial neural networks. It is relevant for us that (i) the network captures and represents the dynamical properties, (ii) learning should be fast, and (iii) learning has a neural form.

The Echo State Network (ESN) is an important candidate for such efforts. Despite of its simplicity, it shows immense representation capacity for nonlinear-dynamical systems. Further, the speed of learning is unique amongst Recurrent Neural Networks (RNNs) due to its fast Linear Mean Squared Error (LMSE) tuning algorithm. Finally, the on-line form of any LMSE algorithm corresponds to the well known local Delta-rule that can be implemented in neural networks.

We shall show by numerical experiments that under certain conditions, ESN can gain more than an order of magnitude for Mackey-Glass (MG) time series in terms of prediction length. We provide a set of conditions that achieve this gain. The basic finding is that such ESNs correspond to a critical condition. We describe a framework to measure if an ESN exhibits phase transition and critical behavior. The framework also helps us provide an interpretation.

The paper is built as follows. First, we briefly review background information about ESNs (Section 2.1) and about critical phenomena (Sect. 2.2). We describe our methods in Sect. 3. We study 'macroscopic behavior', optimize the topology,

S. Kollias et al. (Eds.): ICANN 2006, Part I, LNCS 4131, pp. 658–667, 2006.

and test prediction capacities. Section 4 is about our results on the critical point of ESN phase transition and about the predictive potential of some critical ESNs (CESNs). Discussion can be found in Sect. 5. We close with a short summary.

2 Preliminaries

2.1 Echo State Networks

Echo State Network was first introduced by Jaeger [1,2]. We study simple ESNs that contain all necessary components. The ESN has a hidden layer that holds the hidden representation $\mathbf{a} \in \mathbb{R}^l$. It receives input $\mathbf{x} \in \mathbb{R}^k$ and provides output $\mathbf{y} \in \mathbb{R}^m$ (Fig. 1). Network dynamics is governed by the following equations:

$$\mathbf{a}_t = (1-\mu)\mathbf{a}_{t-1} + \sigma(\mathbf{F}\mathbf{a}_{t-1} + \mathbf{W}\mathbf{x}_{t-1}) \tag{1}$$

$$\mathbf{y}_t = \mathbf{H}\mathbf{a}_t \tag{2}$$

where $\mathbf{W}$ and $\mathbf{H}$ are the input and output mappings, respectively, $\mathbf{F}$ represents the recurrent feedback connections of the hidden layer, $\sigma(\cdot)$ is a component-wise non-linearity that we set to $\tanh(\cdot)$, and μ is the parameter of leaky integration. In the ESN approach, a large number of neurons is used with random recurrent connections at the hidden layer ($l \gg k$). They seem to play the role of a 'dynamic reservoir'. We shall consider the configuration when the output of network is an estimation of the next input, that is, $\mathbf{x}_{t+1} = \mathbf{y}_t$ (and $k = m$) at every time step $t > t_0$. In this mode, and upon tuning, the network is capable of approximating the continuation of the experienced time series in the absence of further inputs.

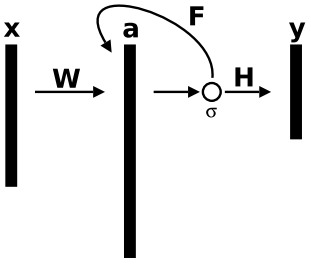


Fig. 1. Structure of the Echo State Network. $\mathbf{x}$: input, $\mathbf{a}$: hidden representation, $\mathbf{y}$: output, $\mathbf{W}$, $\mathbf{F}$, $\mathbf{H}$ linear transformations, σ: nonlinearity.

ESNs are special RNNs: only the hidden-to-output connections (matrix $\mathbf{H}$) are trained. Training is a simple linear regression task that minimizes the time averaged mean squared error between the output and training signal, which is the input itself in our case :

$$J = \frac{1}{2}\sum_t \varepsilon(t) = \frac{1}{2}\sum_t |\mathbf{y}_t - \mathbf{x}_t|^2. \tag{3}$$

Usually, random initialization is used for matrices $\mathbf{W}$ and $\mathbf{F}$. It has been found that the so called echo states may not appear, unless the ESN satisfies the following constraints [3]: Matrix $\mathbf{F}$ should be *sparse*; only a few percent of its elements is non-zero and thus *connectivity* p is low. Also, matrix $\mathbf{F}$ should be *contractive*: the magnitude of singular values should not exceed 1. More details on the operation and tuning of ESNs can be found in the literature, see, e.g. [1,3,4,5]. There are studies about performance and alteration of ESNs [6,7,8]. It has been noted that the nature of the *dynamical reservoir* is not understood yet [9]. Our work aims to shed light on this issue.

2.2 Critical Phenomena

Critical phenomena are notable concepts in physics. The notion refers to many-body interactions, where 'body' is meant in a very general sense. Critical phenomena appear in second order phase transitions and percolation processes, among others. In general, critical phenomena may occur in the transition region that separates 'phases', which may differ in their symmetry properties, in the macroscopic parameters, in their structure, i.e., in their long range order. It is typical to define the *order parameter* of the transition that appears or disappears in one of the phases. The transition between the phases can be a function of the size of the system. The change of the order parameter becomes infinitely sharp in the limit of infinite size. This singular value of the parameter is called the *transition point* of the phase transition. Chaotic behavior is typical for temporal changes at the transition point. These concepts are sufficient for us to proceed. For further details about critical phenomena and for a review of the vast literature of the subject, see, e.g., [10] and references therein.

Below, we define an order parameter for ESNs and present computer simulations. They show that the transition of the network becomes sharp by increasing the size of the network. We also find that at around the transition point predictive capabilities of ESNs can be much better than those of ordinary ESNs.

3 Methods

In this section we describe how the long term behavior of the hidden layer was studied. We establish conditions for finding ESNs with better hidden layer recurrent connections. Our efforts lead to an order parameter and a test that captures the essence of chaotic time series.

3.1 Time Evolving Properties

We are to describe and quantify the special condition mentioned in Sect. 2. Consider the long term behavior of the components of the hidden layer, $a_{i,t}$, $i = 1, \ldots, l$. We would like to eliminate the effects of the input $\mathbf{x}$ and we set $\mathbf{W} \equiv \mathbf{0}$. Under this condition, qualitative description of the time evolution of the components $a_{j,t}$ can be provided, because activity propagation that starts at time 0 and ends at time t is determined solely by $\mathbf{F}^t$ apart from non-linearities.

The proportion of non-zero elements in $\mathbf{F}^t$ will be called *time evolving connectivity* and we denote it by p_t. Similarly, let q_t denote the number of non-zero matrix elements of $\mathbf{F}^t$. Thus $q_t = l^2 p_t$, where l is the number of neurons at the hidden layer. Quantities p_t and q_t are *macroscopic* measures of the connectivity structure that – in a broad sense – characterize information transfer from $\mathbf{a}_0$ to $\mathbf{a}_t$ through the non-zero elements of matrix $\mathbf{F}^t$. In the limit $t \to \infty$, p_t may converge. In this case, p_t may increase, decrease, or even vanish. Alternatively, it is easy to find cases, when p_t may keep changing for all times around its average value. In this case, we take this average as p_∞. In line with this note, we shall see that $o = \frac{p_\infty}{p_0}$ is an appropriate order parameter for us.

The activities of the hidden layer are also subject to temporal changes. For $\mu = 1$, Eq. (1) can be rewritten as $\mathbf{a}_t \approx \sigma((\mathbf{F} + \mathbf{WH})\mathbf{a}_{t-1})$. Upon optimizing matrix $\mathbf{H}$ for objective (3), the largest eigenvalue of $\hat{\mathbf{F}} = \mathbf{F} + \mathbf{WH}$ will approximate 1 for non-vanishing deterministic processes.

3.2 Prediction Test

We tested ESNs on Mackey-Glass (MG) [11] time series, derived by means of the delayed parameter differential equation:

$$\dot{x}(t) = -\gamma x(t) + \frac{\alpha x(t-\tau)}{1 + x(t-\tau)^\beta}, \tag{4}$$

where parameter β influences bifurcation, whereas delay parameter τ influences the complexity of the time series. We used $\alpha = 0.2$ and $\gamma = 0.1$, which are widespread in the literature.

Mean squared error is the typical measure of accuracy in the ESN literature. However, if networks are tested on MG time series that may exhibit chaotic patterns depending on the delay parameter, a peculiar effect occurs: prediction estimates usually follow the original trajectory accurately for some time, but – apparently – the network looses the dynamics suddenly. This phenomenon is a general property of chaotic systems, because the divergence of individual trajectories can be exponential. Predictive capacity for chaotic systems is thus better described by the exponent of the divergence of trajectories or by thresholds.

For the comparison of different networks, we introduce a measure of predictive capacity: *successful prediction length*, ζ. Prediction is called 'θ-successful' for time τ with parameters t_p and T, or 'successful', for short, if starting to predict at time t_p and predicting for time durations $t \leq \tau$, the average of the squared prediction error $\boldsymbol{\varepsilon}(t)$ over time interval T does not exceed θ, but it does if $t > \tau$:

$$\zeta(t_\mathrm{p}) = \arg\max_\tau \left(\langle \boldsymbol{\varepsilon}(t_\mathrm{p} + \tau + i) \rangle_{i=1,\ldots,T} < \theta \right), \tag{5}$$

where τ is the growing length of attempted predictions, $\langle \cdot \rangle$ denotes averaging, and i is the running index of averaging.

Should $\langle \boldsymbol{\varepsilon} \rangle$ exceed θ, we consider that the system can not keep the predicted output close to the true input trajectory $\mathbf{x}(t)$ any further. Measure ζ captures the essence of chaotic dynamics [12].

In numerical experiments, different transformations $\mathbf{F}$ and $\mathbf{W}$, starting point t_p, and training lengths were used to learn the distributions of $\zeta(t_\mathrm{p})$ for one-dimensional MG time series. Parameters of this study are provided in Table 1. It may be worth noting that training length and network sizes are much smaller than those of [1]. Now, we describe our experimental findings.

Table 1. Experimental parameters

size of hidden layer	l in the range $20 - 400$
value of elements of $\mathbf{W}$	randomly chosen; $\approx \pm 0.07$
value of non-zero elements of $\mathbf{F}$	*equal* and positive
max. eigenvalue of $\mathbf{F}$ (scale factor)	0.9
value of leaky integrator, μ	0.7
Mackey-Glass parameters as in [1]	$\alpha = 0.2, \gamma = 0.1, \beta = 10$
Mackey-Glass delay parameters	$\tau = 17$ & 30
training length	1500 (with sub sampling 10)
threshold and averaging window in Eq. (5)	$\theta = 0.2$, $T = 10$

4 Results

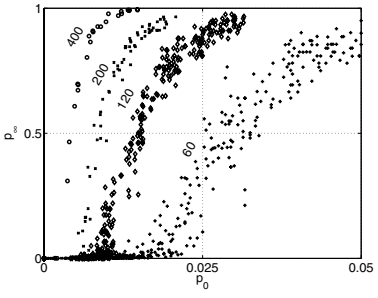
First, we shall show that a sharp transition appears in time evolving connectivity and describe how the final phase depends on the initialization. After measuring the value of the critical point we shall conclude that permutation matrices, or orthogonal matrices in general, satisfy the critical condition. We shall demonstrate the superior performance of permutation matrices.

4.1 ESN Phase Transition and the Critical Point

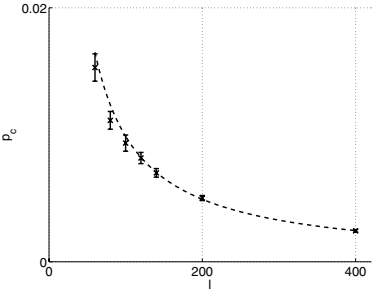
We have created a large number of networks of various sizes. The connectivity structure of the hidden layer was set randomly. We have determined p_∞ for all of them. We have plotted p_∞ against p_0 (Fig. 2a) for different inner layer sizes. Two phases emerged with a transition interval between them. By increasing the size of the network, the position of the interval underwent a monotone shift towards lower values and the width of the interval became narrower (Fig. 2b).

According to our original *critical point conjecture* sharp phase transition emerges with a critical point that separates the two phases at around $p_t \approx p_0$. We define the critical point of ESNs as $p_c = p_0 = p_\infty$ (but see also Sect. 5). Figure 2b shows that a $p_c \approx 1/l$ relation is apparent for larger network sizes with a few percent relative standard error. Thus, according to Section 3.1, we have $q_c \approx l$, because $q_t = l^2 p_t$ and $p_c \approx 1/l$. For a *critical network* subject to our choices detailed in Table 1, the number of equal and non-zero elements in $\mathbf{F}$ is equal to the dimension of the hidden layer.

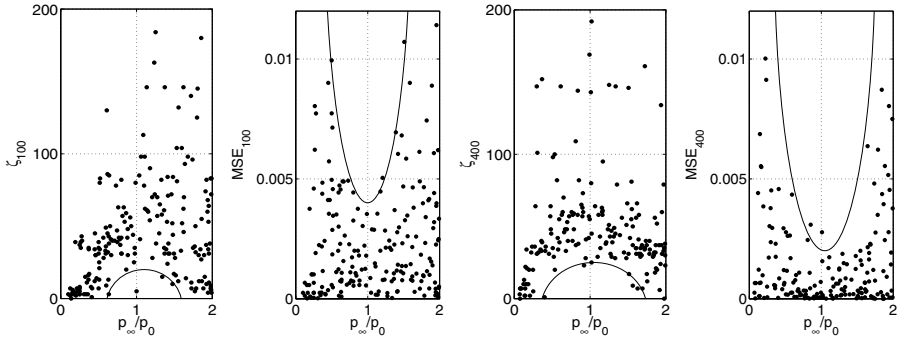
In the next section we introduce *exact critical structures* for the hidden matrix. We shall see that critical structure often exhibits superior performance.



(a) Connectivity in the limit p_∞ versus initial connectivity p_0



(b) Critical 'point' p_c versus network size l



(c) **(d)** **(e)** **(f)**

Fig. 2. Phase transition and improved performance around the critical point. **(a):** Phase transition in time evolving connectivity. *Zero phase*: connections of the hidden layer disappear for sufficiently large, but finite times. *Saturated phase*: (almost) all connections contribute after sufficiently large times. Transition between the phases becomes sharp for larger hidden layers. **(b):** Position of p_c shifts to lower values as hidden layer size l increases. Dashed line: fit by assuming $p_c = 1/l$. **(c)** and **(e):** estimated successful prediction length ζ, **(d)** and **(f):** MSE of ζ, **(c)** and **(d):** size of hidden network is 100, **(e)** and **(f):** size of hidden network is 400. Solid lines: approximate (indicative) 'boundaries' that show improvements around the critical point $p_\infty/p_0 = 1$.

4.2 Critical Echo State Networks

Condition $q_c = l$ for matrix $\mathbf{F}^t$ is satisfied e.g., if every row and every column of $\mathbf{F}^t$ contains one non-zero element in the limit. Such structure will be called *exact critical structure*. For example, ESNs with permutation matrices in the hidden layer (PESNs) have exact critical structure.

Before proceeding, we conclude for the general case: according to our numerical studies, there is a critical region for networks. In this region, the time evolving hidden layer connectivity may not loose all connections (may not enter the *zero phase*) or may not get close to full connectivity (the *saturation phase*). See also

Fig. 2a and the caption of Fig. 2. Such networks will be called *Critical Echo State Networks* (CESNs).

There are special cases that belong to CESNs. For example, if the eigenvalues of matrix $\mathbf{F}$ are bounded by the unit sphere, two of them are on this sphere, and these two do not form a diagonal sub-matrix, that is they mix elements of the internal representation, then $\mathbf{F}^t$ will not belong to the zero phase nor to the saturation phase. Also, ESNs with hidden orthogonal matrices are CESNs, because their connectivity structure neither vanishes nor saturates in the limit.

In our investigations we shall turn to hidden permutation matrices, because otherwise the relative number of critical structures generated randomly may be very low, especially for large hidden layers. A particular $l \times l$ permutation matrix contains $1 \leq l_\nu \leq l$ number of cycles. A cycle of length l_ν exchanges the corresponding elements of a vector in l_ν steps. Similarly, orthogonal matrices mix subspaces.

Now, we present results for hidden permutation matrices, i.e., for PESNs and we set

$$\mathbf{F} = \mathbf{P} \ , \tag{6}$$

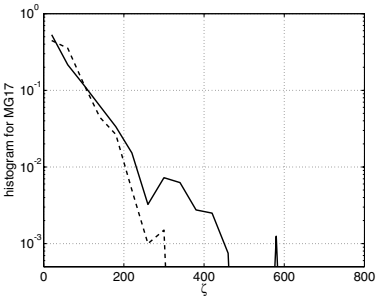
where $\mathbf{P}$ is a permutation matrix.

4.3 Prediction Gain over Ordinary ESNs

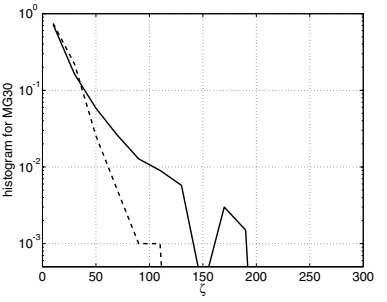
In this section we compare the performance of ESNs with PESNs on Mackey-Glass time series with delay parameters 17 (MG17) and 30 (MG30).

Figures 3a and 3b show distributions of successful prediction length ζ for 4,000 ordinary ESNs. The distributions are compact. The same figures depict the distributions for 4,000 PESNs with randomly generated input matrix, hidden permutation matrix, and optimized output matrix. PESNs show more asymmetric distributions for ζs. The average and the median are about the same for the two distributions, but the ESN distributions are much narrower. A large proportion of PESNs are very successful, whereas we have barely encountered significantly better than average randomly initialized ESNs, in agreement with the results reported in the literature. In Figs. 3a and 3b, the decrease of the PESN distribution is slower than that of the ESN distribution; the PESN distribution seems to have a long tail. For the more difficult MG30 time series prediction length is shorter for both networks.

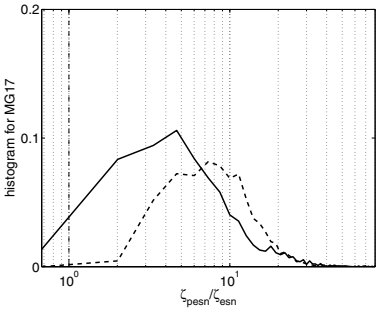
Figures 3c and 3d compare the number of occasions that a particular successful prediction length was achieved by ESNs and PESNs for MG17 (Fig. 3c) and for MG30 (Fig. 3d). Performances were evaluated over 2500 different starting points and two comparisons were made. The best PESN out of 4000 randomly chosen PESN networks was compared to (a) the average ESN out of 4000 randomly chosen ESN networks and (b) the best ESN out of the same 4000 randomly chosen ESN networks. For high ratios, i.e., when the performance of the PESN is much better than that of the ESN, the curves become similar for both MG17 and MG30. Results indicate that for large successful prediction lengths, performances of the average and the best ESNs out of 4000 randomly generated networks are poor and are very similar. Thus, high performance ESNs are rare compared to



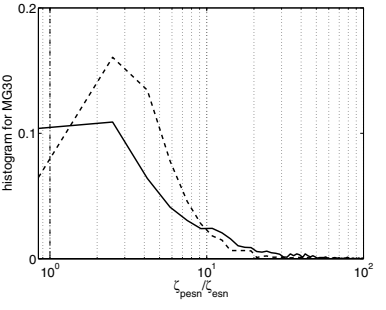
(a) Performance distributions over 2500 starting points for randomly generated ESN and PESN networks for MG17.



(b) Performance distributions over 2500 starting points for randomly generated ESN and PESN networks for MG30.



(c) Ratios of numbers of occasions of successful predictions length of PESNs and ESNs for 2500 starting points for MG17.



(d) Ratios of numbers of occasions of successful predictions length of PESNs and ESNs for 2500 starting points for MG30.

Fig. 3. Comparisons of ESNs and PESNs for MG17 and MG30 time series. **(a)** and **(b)**: Dashed lines: ESN, solid lines: PESN, network size: $l = 60$, **(c)** and **(d):** Dashed lines: average ESN (out of 4000) and best PESN (out of 4000), solid lines: best ESN (out of 4000) and best PESN (out of 4000), network size: $l = 60$, vertical line at value 1 (at 10^0): 'curve' for identical distributions.

high performance PESNs: PESNs form a highly efficient subgroup within ESN networks – at least for MG chaotic time series.

We found that matrix $\mathbf{W}$ had an effect on the performance of PESNs. For example, $\mathbf{W}$ with similar elements had a negative effect. Uneven averages and variances for elements of $\mathbf{W}$ belonging to *different* cycles improved performance.

5 Discussion

We studied critical ESNs with single inputs. PESN performances have broader distributions than ESN ones. For PESNs, the *probability* that extremely good

ESN is found is dramatically increased. One can quickly find extremely good PESNs, whereas good ESNs are rare amongst ordinarily initialized ESNs.

Why do we find high performance PESNs significantly more often? Consider the permutation matrix in the hidden layer of the PESN. In general, it connects disjoint sets of elements, that is, we have disjoint cycles. We found that neither the single cycle case, nor the case of a large number but small cycles exhibited good performances. This was expected because of the following reason. For a single cycle of size n, identical representation arises after n steps. However, if there are more cycles, the identical representation appears after m_{LCM} steps, where m_{LCM} is the least common multiple of the sizes of the cycles. LCM is small if all cycles are equal, if cycles are small, or if there are single cycles. Such PESNs, show poor performances, but form only a small subset of randomly generated PESNs.

Now, consider general orthogonal matrices in the hidden layer. They belong to the class of CESNs, because their time evolving connectivity can neither saturate nor disappear for large times. It is possible that connectivity structure does not converge: periodic or never repeating structures may occur. We have studied CESNs starting from good PESNs. For example, we changed the sign of one or more non-zero elements of a permutation matrix. In all cases, the good predictive performance dropped to average. We also tried to modify different non-diagonal 2×2 sub-matrices defined by two non-zero elements of the permutation matrix to a rotation matrix. Performance decreased in most cases unless the angle of rotation was small. Note that permutation corresponds to rotation by 90^0 *and* a reflection, whereas 180^0 rotation corresponds to the change of the sign of one of the components. Combinations of these changes also spoiled performance in an overwhelming majority of the experiments.

The hidden permutation matrix is able to approximate non-periodic dynamical systems, because the hidden layer is embedded by the input matrix $\mathbf{W}$ and the output matrix $\mathbf{H}$, and they can modify finite cycles; matrix $\hat{\mathbf{F}} = \mathbf{F} + \mathbf{WH}$ counts in this respect. Note however, that matrix $\mathbf{WH}$, which can modify the permutation matrix, has limited capabilities, because the rank of this matrix is 1. Chances are high that changes of permutation matrices of good PESNs destroy performance, thus such changes seem to be out of reach for the optimization procedure of matrix $\mathbf{H}$ of the PESN.

Identification capabilities of *general* CESNs for dynamical systems *beyond* MG time series deserve further studies. A rich repertoire of phenomena may appear for input dimensions larger than 1.

6 Summary

We have shown that ESNs undergo sharp phase transition depending on the connectivity properties at the hidden layer. We have introduced and studied *critical echo state networks*. We found – by means of a large number of numerical simulations – that a large proportion of exact critical structures exhibit highly superior performance as opposed to ordinary ESNs, at least for MG time series.

We have argued that CESNs with permutation matrices in the hidden layer can identify both periodic and aperiodic time series, because the permutation matrix is complemented by other structures of the ESN.

References

1. Jaeger, H.: The Echo State Approach to Analysing and Training Recurrent Neural Networks. Technical Report 148, Fraunhofer Institute for Autonomous Intelligent Systems (2001)
2. Maas, W., Natschläger, T., Markram, H.: Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. Neural Computation **14** (2002) 2531–2560
3. Jaeger, H.: Short Term Memory in Echo State Networks. Technical report, German National Research Center for Information Technology (2002)
4. Jaeger, H., Haas, H.: Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. Science **304** (2004) 78–80
5. Ishii, K., Zant, T., Becanovic, V., Plöger, P.: Identification of Motion with Echo State Network. In: Proc. Oceans. (2004) 1205–1230
6. Baier, N., De Feo, O.: Chaotic Model Identification Using a Biologically Inspired Algorithm. Aperest, Universidad Complutense de Madrid (2004)
7. Mayer, N., Browne, M.: Echo State Networks and Self-Prediction. In: Lecture Notes in Computer Science. Volume 3141., Springer Berlin / Heidelberg (2004) 40
8. Fette, G., Eggert, J.: Short Term Memory and Pattern Matching with Simple Echo State Networks. In: Lecture Notes in Computer Science. Volume 3696., Springer Berlin / Heidelberg (2005) 13
9. Jaeger, H.: Reservoir Riddles: Suggestions for Echo State Network Research (Extended Abstract). In: Proceedings of International Joint Conference on Neural Networks, Montreal, Canada. (2005)
10. Sornette, D.: Critical Phenomena in Natural Sciences. Springer Series in Synergetics. Springer, Berlin, Germany (2003)
11. Mackey, M., Glass, L.: Oscillation and chaos in physiological control systems. Science **197** (1977) 287–289
12. Cvitanovic, P., Artuso, R., Mainieri, R., Tanner, G., Vattay, G., Whelan, N., Wirzba, A.: Chaos: Classical and Quantum. Niels Bohr Institue, Copenhagen, chaosbook.org version 11 (2004)