

*Re: Submission to Science, "Harnessing nonlinearity: predicting chaotic systems and boosting wireless communication." (Ref: 1091277)*

## Refutation of Second Reviewer's Objections

*Herbert Jaeger, Dec. 23, 2003*

For convenient reference, the second reviewer's objections are appended at the end of this refutation.

Among the three objections raised by the second reviewer, the second is the most serious. Therefore I treat it first. The objection has several aspects that I will address separately.

### Objection 2, aspect 1

*Allow me to re-phrase: in computing the training and testing data for the chaotic attractor example, I used a large stepsize for simulating the attractor trajectory. Therefore, it might not be an accurate version of the attractor; it might even be non-chaotic. To remove such suspicion, I should report the Lyapunov exponent.*

#### Refutation:

- **The stepsize that I use is actually not large.** The stepsize 0.1 that I used must be seen in relation to the attractor's inherent timescale, which is very slow. With this stepsize, one loop around the attractor takes about 500 simulation steps, which cannot be considered a coarse approximation. When the Mackey-Glass system is simulated with a standard commercial solver for differential delay equations (namely, Matlab's dde23 function), a stepsize of about 2.0 is automatically selected, 20 times the stepsize I used.
- **The stepsize that I use yields an accurate simulation of the attractor.** I computed estimates of the first Lyapunov exponent  $\lambda_1$  for both my simulation (stepsize 0.1) and the Matlab simulation function dde23. I obtained  $\lambda_1 \approx 0.0064$  for my simulation and  $\lambda_1 \approx 0.0059$  for the commercial simulation tool. Both values are in agreement with estimates found in the literature.

Taken both points together, my simulation method clearly yielded an accurate model of the MG attractor. All computations relating to the calculation of Lyapunov exponents are online in the form of commented executable Matlab files at

<http://www.ais.fraunhofer.de/INDY/herbert/1091277/MGLyapunov.zip> .

**Suggested improvement to the manuscript:** if the occasion arises, I would include a statement about the Lyapunov exponent into the SOM, as suggested by the reviewer.

## Objection 2, aspect 2

First I rephrase this objection: *the chaotic system I investigated is not strongly chaotic and therefore easy to predict.*

### Refutation:

- **The Mackey-Glass system is indeed not strongly chaotic, but that is not the point.** A first Lyapunov exponent of about 0.006 implies weak chaos. But this fact is common knowledge in the field of nonlinear dynamics; in fact it is one of the reasons why the Mackey-Glass system is arguably the most popular benchmark for chaotic time series prediction methods. I chose it because of this benchmark character which allowed me to put my approach into perspective. The MG system (with delay 17) is uniquely suited as a demonstrator because it is the only system where, for historical reasons, virtually all authors use the same criterium for measuring model precision, namely, the 84 step prediction. This allowed me to formulate my claim that the ESN method outperforms existing techniques by 2.7 orders of magnitude.
- **On more chaotic or more complex systems the ESN method works superbly as well.** Because of limited space, I did not include other case studies in the submitted manuscript. In a separate document attached to this refutation<sup>1</sup>, I report ESN training experiments on various chaotic attractors that each pose distinctive difficulties:
  - The Lorenz attractor, which with a largest Lyapunov exponent of about 0.9 exhibits significant chaos. Here, ESNs appear to perform better than previous techniques by 4-5 orders of magnitude.
  - The MG system with a delay of 30, whose chaos is as mild as in attractor investigated in the submitted manuscript, but which has a higher embedding dimension and exhibits a very rich phenomenology and is therefore difficult to model. Again, ESNs advance the state of the art dramatically.
  - The empirical Laser time series from the 1994 Santa Fe time series prediction competition, which contains noise and in which a crucial "breakdown" event that has to be predicted occurs only once in the training sequence. I include this example because the second reviewer suggested that I treat time series competition data. ESNs perform as well as the best other techniques; however, since 1994 it has been found out that this time series is actually easy to learn optimally by a simple pattern matching. Thus, ESNs and the other best approaches simply reach the optimal possible performance.

### Suggested improvements to the manuscript:

- If recommended by the editors, I would include prediction studies of other well-known chaotic attractors into the SOM.
- In the meantime I was able to further increase the prediction accuracy by a factor of about 4, achieving in the MG (delay 17) task an NRMSE for the standard 84-step prediction of  $1e-4.75$ , a little more than 3.0 orders of magnitude better than previously possible. The idea of the method refinement is sketched in the supplementary manuscript and can be inspected in detail from the Matlab files that I made accessible online. If suggested by the editors, I would use the improved results and methods for a revised version of the manuscript.

---

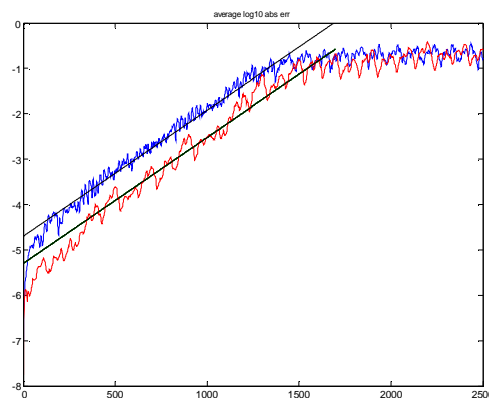
<sup>1</sup> *Various chaotic attractors predicted by ESN models*, manuscript specially prepared for this refutation.

### Objection 2, aspect 3

First I rephrase this objection: *I should compare my prediction results with the theoretical optimum.*

#### Refutation:

- **It is currently not possible to define the theoretical optimum.** Such an optimum should indicate the maximal achievable prediction accuracy using any prediction method, given a set of training data. In computing such an optimum, the mathematics of a general learning theory (independent of learning method) must be combined with the mathematics of chaotic attractors. But even a general learning theory is not available. The best that we have is known as *statistical learning theory*<sup>2</sup>, but the error bounds yielded by that theory are far from tight, and the theory does not yet cover models of dynamical systems.
- **The optimality criterium suggested by the reviewer is closely met by the ESN method, but this is not revealing.** The reviewer does not state which optimality criterium s/he has in mind, but I take it that s/he refers to the criterium used in the supplied reference. According to this criterium, a prediction is optimal if it diverges from the true continuation with the rate of the true system's maximal Lyapunov exponent. In this sense, the predictions yielded by the ESN method in the original submission are almost optimal (Fig. 1). But, the refined ESN method developed in the meantime is 4 times more precise and meets the criterium just as well. Therefore, the criterium is not informative about how close a method comes to the theoretical optimum.



**Figure 1:** Illustrating the optimality criterium suggested by the second reviewer. The  $x$ -axis shows network updates which correspond in this case to MG time units, the  $y$ -axis shows  $\log_{10}$  of absolute error. Blue line: absolute error of MG predictions averaged over 20 prediction trials, all using the same ESN trained as described in the submitted manuscript. Red line: same for an ESN model trained with the improved method developed in the meantime (averaged over 2 prediction trials). Black lines: rate of divergence according to a Lyapunov exponent of 0.0064. Even without a rigorous statistical analysis it is clear that the slope of the model divergence comes close to the Lyapunov rate for both ESN models; in this sense, both model predictions would qualify as "optimal". However, the second model is 10 times as precise as the first.

---

<sup>2</sup> Vapnik, V. N., *The Nature of Statistical Learning Theory, Second Edition*, Springer Verlag 1999

**Suggested improvement to the manuscript:** Figure 1 might be included in the SOM and explained.

## Objection 1

In the remainder I treat the less serious objections. I first rephrase the reviewer's first objection: *the idea of using a randomly connected network to yield a reservoir of basis functions has no significant originality – it is known since the beginnings of neural network research and has been tested often and with limited success.*

### Refutation:

- **Randomly connected networks as reservoirs of basis functions were mostly of the feedforward type.** It is true that the forefather of modern neural networks, the Perceptron, had neurons with fixed, random connections (although not among each other but to the input field). Similarly, some versions of the well-known radial basis function networks can be considered randomly connected. But all of these are feedforward networks, very different objects from the recurrent networks considered in the manuscript. Mathematically, feedforward networks are *functions*, whereas recurrent networks are *dynamical systems*.
- **If similar approaches exist, they are not well-known.** I cannot preclude the possibility that randomly connected recurrent neural networks were used previously in ways that bear similarity to my method; specifically, by training the output connection weights only. However, if this is the case, these approaches are not well-known in the field of machine learning / artificial neural networks. The only exception known to me (now) is the 1981 paper by Geoffrey Hinton, pointed out by the first reviewer. Since the first conception of ESNs in 2000, I discussed my ideas with a number of prime researchers in the field (Helge Ritter, Geoffrey Hinton, Danil Prokhorov, Wolfgang Maass, Peter Dayan, Zoubin Ghahramani, Wolfgang Singer, Jun Tani, Volker Tresp, Jürgen Schmidhuber) and none of them expressed concern about lack of originality. If the second reviewer could be asked to give references indicating to what approaches s/he is alluding, presenting and discussing them would certainly be instructive and benefit the manuscript.
- **My approach rests on original mathematical insights.** I did not expand on the mathematical aspects in the body of the manuscript because I felt that this would be inappropriate for the wide-spanning readership of Science. However, I give references to online techreports where interested readers may learn more about the following points:
  - several abstract (some not entirely trivial) characterizations of the "echo state property" that randomly connected networks must have to make the learning method work,
  - a formal analysis of algebraic properties of the network weight matrix that ensure the echo state property,
  - a formal analysis of the short-term memory capacity of echo state networks. This short-term memory capacity is the main reason for the good performance of echo state networks for learning systems which have memory.
- **Echo state networks are of interest as models of biological neural networks.** In the field of computational neuroscience, we currently see a surge of interest in models of neural information processing that are very similar to the basic idea of echo state networks. Specifically, this concerns the "liquid state" model proposed by Wolfgang Maass and Henry Markram. Both in echo state and liquid state networks, the basic idea is

that a random reservoir is "tapped" by trainable readout neurons. Wolfgang Maass (with whom I now co-operate) developed these ideas simultaneously and independently. I briefly point out these correspondences to biological neural networks at the end of my submission paper.

**Suggested improvement to the manuscript:** (i) I will comment and reference the paper indicated by the first reviewer, (ii) I could elaborate more on the mathematical results underpinning my approach, and (iii) I could expand more on the emerging correspondences to biological neural networks.

### Objection 3

If I understand it correctly, the third objection can be paraphrased as follows: *ESNs are random structures, whose construction is determined by a few global parameters (number of hidden units, sparseness of connectivity, spectral radius of weight matrix). For a given learning task, optimal values of these parameters must be found. I should supply a method for determining these values; specifically, this method must not make use of the testing data. Furthermore, I should demonstrate that the results do not critically depend on the specific random connection structure. Finally, I should use real-world data and not synthetic textbook systems.*

### Refutation:

- **I did (of course) not use test data for parameter optimization.** All experiments reported in the submitted manuscript proceeded by first optimizing the few global model parameters by hand, using only the training data, then training an ESN on the training data, and finally testing the obtained model on independent test data.
- **Some hand-tuning of parameters cannot be avoided in machine learning algorithms.** It is standard practice, if not unavoidable in working with artificial neural networks to experiment by hand in order to determine appropriate ranges of structural network parameters. Typically, the number of neurons, the global network topology, learning rates and/or learning rate decay schemes (sometimes elaborate), control parameters for gradient descent speedup mechanisms, termination criteria for iterative optimization schemes have to be found by experimentation. Requiring a fully automated learning scheme wherein *all* parameters are automatically determined just is demanding too much. The ESN approach compares favourably with other neural network learning techniques in that the actual learning is a constructive algorithm that needs no parametrization itself – other techniques all rely on iterative schemes where the learning mechanism itself needs heavy tuning. In fact, there are only four parameters that I have found to be important: network size, spectral radius, input and output scaling. In an online techreport referenced in the submitted paper I give detailed practical hints to optimize these. Because the number of hand-tuned parameters is small and they have broad tuning curves, a few iterations of hand-tuning typically suffice.
- **The specific random connectivity of an ESN plays not a very important role.** It is true that given a fixed set of structural parameters, independently created ESNs perform differently. The reviewer rightly demands that I consider this point and provide statistics to show how strong is the effect of random variations between individual ESNs, and I will do so if the occasion arises. However, the effect is not large and in practice can be neglected. This is probably due to the fact that I use relatively large networks, where the effects of random internal structure smooth out. For instance, when I train 100 independently created ESNs on the MG prediction task considered in the submitted

manuscript, all on the same training data, I obtain an average  $\log_{10}$  prediction NRMSE error for the 84 step prediction of about  $-4.2$ , with a standard deviation of  $0.38$ . Given that the best previous techniques featured a  $\log_{10}$  prediction error of  $-1.7$ , this means that the entire performance "cloud" of random ESNs is separated by more than 2 orders of magnitude from previous techniques.

- **Using synthetic data is appropriate for first introducing a novel modelling technique.** Only with synthetic data that can be systematically varied it is possible to evaluate the principle properties of a learning technique. For instance, in the equalizer example in the manuscript I could systematically vary the amount of noise and see how the model copes. Likewise, only for synthetic chaotic systems it is possible to obtain sharp estimates of the Lyapunov exponent. Therefore I believe it is justified (and standard practice, too) to use synthetic data in an early publication on a new learning method. However, the reviewer is certainly right that the final criterium for usefulness is the performance of a modelling technique on real-world data. One such example is given in the supplementary manuscript (the Laser data set).

**Suggested improvement to the manuscript:** I will provide the statistics of ESN performance scattering for both the Mackey-Glass and the channel equalization examples.

## Appendix: review #2.

This manuscript reports two examples of using artificial neural networks, which the authors call Echo State Networks, to signal prediction and filtering. Although the reported numbers may sound appearing, this work leaves many points to be clarified and would not be of particular interest to the wide readership of Science.

1) The idea of diversifying input signals by a randomly connected network to build basis functions for synthesizing the output has been around for many years, probably since the days of original Perceptron. It's use for temporal signal processing has also been tested, often with limited results. What differentiates ESNs from previous works are the use of sparse connections and the scaling of connection weights by computing the spectral radius. These two points are not adequately stated, especially the latter is only mentioned in the SOM.

2) How far ahead one can predict a chaotic time series is fundamentally limited by the size of Liapunov exponent (Kuo et al., 1992). It is very surprising that the MG system was predicted for a long time as reported here, if the system is in a strongly chaotic regime. At the reported parameter setting of MG system, what was the value of the largest Liapunov exponent? Unfortunately the SOM does not report the numerical integration method used, but with a quite large time step of  $0.1$ , the simulated time course would have been rather different from that of the genuine delay-differential dynamics of MG system. A knowledgeable reader would suspect that the reason why long-term prediction was possible was because the simulated system was not strongly chaotic (e.g., quasi-periodic) due to a poor simulation method. In order to reject such a suspicion, the Liapunov exponent of the simulated system should be reported, and the success of prediction should be measured with respect to the theoretical limit.

3) The critical parameters of ESN implementation are the number of hidden units, the sparseness of connectivity, and the spectral radius of the connection weights. In the benchmark tasks for which the TRUE

outputs are known, tuning of these parameters are easy since we can just tune them to minimize the the prediction error. However, a real challenge is to set these parameters for real problems for which we don't know the true output. This is why competitions of signal predictions have been in practice, in which the true answers are withheld. In light of the state of the art of signal processing, the value of any new method is recognized only through its application with real problems with real uncertainties, rather than its application to textbook examples to which true answers are known. Although parameter tuning is always a headache in signal prediction, an additional problem with ESN is to decide which instance of random connections to use, since not all the random networks perform equally, as demonstrated in Figure 4d. If these are not actually the problems, the authors should either report systematic procedures for setting those parameters for novel problems, or demonstrate in systematic simulations that performance of ESNs do not critically depend on parameter settings.

#### Reference

J. Kuo, J. Principe, and B. deVries. Prediction of chaotic time series using recurrent neural networks. In Proc. 1992 IEEE Workshop of Neural Networks in Signal Processing , pages 436-443, 1992.