MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1447 August, 1994
C.B.C.L. Memo No. 101

Forecasting Global Temperature Variations by Neural Networks

Takaya Miyano and Federico Girosi

Abstract

Global temperature variations between 1861 and 1984 are forecast using regularization network, multilayer perceptrons, linear autoregression, and a local model known as the simplex projection method. The simplex projection method is applied to characterize complexities in the time series in terms of the dependence of prediction accuracy on embedding dimension and on prediction-time interval. Nonlinear forecasts from the library patterns between 1861 and 1909 reveal that prediction accuracies are optimal at the embedding dimension of 4 and deteriorate with prediction-time interval. Regularization network, backpropagation, and linear autoregression are applied to make short term predictions of the meteorological time series from 1910 to 1984. The regularization network, optimized by stochastic gradient descent associated with colored noise, gives the best forecasts. For all the models, prediction errors noticeably increase after 1965. These results are consistent with the hypothesis that the climate dynamics is characterized by low-dimensional chaos and that the it may have changed at some point after 1965, which is also consistent with the recent idea of climate change. However, care must be taken of such an interpretation in that a time series of colored noise with few data points that has zero mean and many degrees of freedom can also show a similar behavior.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences, at the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology and at the Advanced Technology Research Laboratories, Sumitomo Metal Industries, Japan. This research is sponsored by grants from the Office of Naval Research under contracts N00014-92-J-1879 and N00014-93-1-0385; and by a grant from the National Science Foundation under contract ASC-9217041 (this award includes funds from ARPA provided under the HPCC program). Additional support is provided by the North Atlantic Treaty Organization, ATR Audio and Visual Perception Research Laboratories, Mitsubishi Electric Corporation, Sumitomo Metal Industries, and Siemens AG. Support for the A.I. Laboratory's artificial intelligence research is provided by ARPA contract N00014-91-J-4038.Takaya Miyano is with Advanced Technology Research Laboratories, Sumitomo Metal Industries, LTD., Amagasaki, Japan.

1 Introduction

In this paper we will apply some linear and nonlinear regression techniques to the analysis of the time series of global temperature variations between 1861 and 1984. We use a data set published by Jones et al. (1986), who synthesized global mean surface air temperature differences between successive years by correcting non-climatic factors in near-surface temperature data over the land and the oceans of both hemispheres. As pointed out by Jones et al., this time series has the interesting feature of having a long timescale warming trend that is remarkable in the 1980s, and that is in the right direction and of the correct magnitude in terms of recent ideas of global warming (see figure 1). It has been recently conjectured that it exists a global climate dynamical system, that is chaotic and whose attractor has a low dimensionality (4-7). If this is the case it should be possible to model the time series of global temperature variations with a model of the type

$$x(t+1) = f(x(t), x(t-1), \ldots, x(t-n)) + \xi_t \qquad (1)$$

where f is some unknown function, the time t is expressed in years, n is a small number (of the order of the dimension of the chaotic attractor) called *embedding dimension*, and ξ_t are random independent variables representing uncertainty in the measurements. In this paper we want to investigate how well a model of type (1) can fit and predict the data, and will use different techniques to reconstruct the unknown function f, representing the dynamics underlying the data.

The paper is organized as follows. In section 2 we introduce the general problem of time series prediction, and discuss the particular case of the global temperature time series, pointing out where the major difficulties are. In section 3 we use a technique by Sugihara and May (1990) to estimate the minimal embedding dimension for the global temperature time series (the number n in eq. 1) and to test the hypothesis of chaotic dynamics. In section 4 we present some experimental results obtained applying regularization networks, linear autoregression and multilayer perceptrons to the prediction of this time series. In section 5 we discuss the results and describe some future work. In the appendices we describe the different forecasting techniques that we used for our analysis.

2 Basic time series analysis

Making predictions is one of the basic subjects in science. Building a model for forecasting a time series is one of the tools that we can use to analyize the mechanism that generated the data. Given the time series $\{x(t)\}_{t=1}^N$ of observations of the variable x at differents points in time there are two extreme situations that we can happen to encounter in its analysis:

1. the value of the variable x at time $t+\tau$ is uniquely determined by the values of the variable at certain times in the past. For example a relation of the type

$$x(t+\tau) = f(x(t), x(t-\tau), \ldots, x(t-n\tau))$$

could hold for some integer n and some function f. In this case the system is fully deterministic, and, in principle, its behaviour is predictable. One could derive the mapping f from first principles, provided that all the interactions between elements in the system are clearly known, or reconstruct its shape from the data $\{x(t)\}_{t=1}^N$.

2. the values of $x(t)$ are independent random variables, so that the past values of x do not influence at all its future values. There is no deterministic mechanism underlying the data, and prediction is not possible at all.

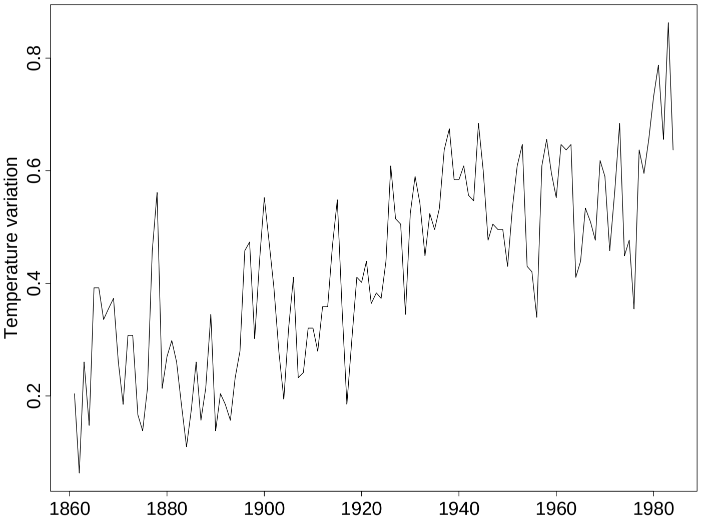

Figure 1: The global temperature variation as reported by Jones et al. (1986).

In most practical and interesting cases one has to deal with time series with properties that lie in between these two extremes. For example, one could have a series of observations of a variable whose time evolution is governed by a deterministic set of equations, but the measurement are affected by noise. In this case a more appropriate model for the time series would be:

$$x(t+\tau) = f(x(t), x(t-\tau), \ldots, x(t-n\tau)) + \xi_t \qquad (2)$$

where $\{\xi_t\}$ is a set of random variables. If one could exactly model the function f the state of the system at time $t+\tau$ could be predicted within an accuracy that depends only the variance of the random variables $\{\xi_t\}$. A system of this type is therefore intrinsically deterministic but some stochasticity enters at the measurement level.

In other cases, as for the time series generated by the noise in a semiconductor device, the system could be intrinsically stochastic. However, since the observation are correlated in time, it might be still predictable to some extent.

It is usually very difficult to discover, from a time series, what kind of mechanism generated the data, but building models of the type (1) is often useful to understand the relevant variables of the system and its degree of randomness. Until not many years ago most of the models used to analyze time series were linear, and therefore limited to succeed on the class of linear problems. However, recent progress in theoretical and computational aspects in nonlinear problems has enabled us to characterize the degrees of randomness in nonlinear systems and to forecast nonlinear dynamical behavior. A number of authors, among which Farmer and Sidorowich (1987), Casdagli (1989), Sugihara and May (1990), have applied and developed novel nonlinear regression techniques to predict chaotic time series, and shown that short-term predictability can often be achieved with a good degree of accuracy. They have also demonstrated that low-dimensional chaos can be distinguished from white noise according to nonlinear forecasts of the dynamical behavior.

One could find the minimal embedding dimension and valid prediction-time interval through applying a variety of network structures by trial and error of cross-validation techniques. Such procedure, however, is likely to be computationally expensive, and one of the ways to circumvent such problem is to resort to local approximation such as the algorithm of Sugihara and May (1990). In local approximation, predictions are made from only nearby states in the past, so that it consumes less computational time. The local approximation, however, has shortcomings that the mapping is discontinuous, which results in less sufficient prediction accuracy than global approximation such as neural networks. Thus the association of the local approximation with neural networks is an effective way for building a model to make predictions of complex time series. One could find the minimal embedding dimension and the valid prediction interval using the local approximation without consuming a lot of computational time and forecast the time series with good prediction accuracy using neural networks the structures of which are determined according to the forecasts by the local approximation.

2.1 Global temperature time series and the climate attractor

It has been controversial whether the climate is low-dimensional chaos or not (Nicolis and Nicolis, 1984; Grassberger, 1984; Essex, Lookman and Nerenberg, 1987; Lorenz, 1991). Nicolis and Nicolis (1984) first claimed the existence of a low-dimensional climate attractor in terms of the correlation dimension of the time series synthesized by interpolations from an isotope record of deep-sea cores. Grassberger (1984), however, argued that their estimate may reflect not the actual climatic dynamics but the artifact due to the interpolation. Meanwhile, Essex et al. (1987) published a calculation on the correlation dimension for non-filtered time series of daily geopotential observations. They agreed with the existence of such climate attractor. Recently Lorenz (1991) expressed doubts on the interpretations of the previous calculations. He claimed hat a low-dimensional attractor is unlikely to exist for the global climate, although the climatic subsystems could be low-dimensional chaos.

3 Is there a global climatic attractor?

In this section we first discuss the plausibility of the hypothesis that there is low-dimensional chaotical system underlying the global temperature time series, and then present some results about its short-term predictions with a number of methods. We use a data set published by Jones et al. (1986), who synthesized global mean surface air temperature differences between successive years by correcting non-climatic factors in near-surface temperature data over the land and the oceans of both hemispheres. Although some questions such as the influence of urbanization effect are raised about the significance of the data (Wu, Newell and Hsuing, 1990) we still think it is of interest to see what kind of information can be extracted from this data set.

3.1 Testing for chaos

We have used the technique proposed by Sugihara and May (1990) to understand the nature of the time series we decided to analyze. The technique consists in studying the behaviour of the correlation coefficient between the prediction and the target as a function of how many steps in the future we are trying to predict, and of the embedding dimension. Sugihara and May argued that looking at the rate of decrease to zero of the correlation coefficient it is possible to distinguish chaotic systems from non-chaotic ones. In our experiments, following Sugihara and May, predictions have been done according to the simplex projection technique described in appendix (A.4).

Figures (2) and (3) show the Sugihara-May technique applied to a number of synthetic time series. Figure (2) shows a plot of the correlation coefficient as a function of embedding dimension for predictions of: a) white noise (open triangles and dashed line); b) $f^{-1.6}$-noise (solid triangles and dotted line); c) Hènon mapping superimposed on white noise (solid squares and short-dashed line), d) sine wave superimposed on white noise (open circles and long-dashed line). The white random noise has been synthesized using equation (8) in appendix (B) with $\alpha = 0$. In Figure (2), predictions have been made on a test set of 150 vectors from a data set of 150 examples.

Figure (3) shows a plot of the correlation coefficient as a function of prediction-time step for the time series shown in figure (2). In figure (3), predictions have been done assuming an embedding dimension $n = 3$, in similar way to Figure (2).

The results in figure (2) and figure (3) reveal characteristics of complexities in each time series. The Hènon mapping superimposed on white noise has a peak embedding dimension and its prediction accuracy deteriorates quickly with prediction-time step. This is typical of low-dimensional chaos. The sine wave superimposed on white noise has long-term predictability independently of embedding dimension. White noise is unpredictable at all. It should be noticed that the $f^{-1.6}$-noise has short-term predictability independently of embedding dimension. According to some recent work (Miyano, 1994; Miyano et al. 1992), such characteristics are also found for an actual colored noise observed for a semiconductor device. The flat relation between the correlation coefficient and embedding dimension is an important indicator in distinguishing low-dimensional chaos from colored noise.

We applied the same techniques described above to the time series observed by Jones et al. (1986). The time series has been obtained by hand-scanning the original figures on the paper. Therefore, the present time series includes some *read error*. We use the first 45 input vectors, i.e., the data from 1861 to 1915, as training data, being motivated by the idea that a climate change may have occurred around the mid 20th-century by factors such as increasing energy consumption and environmental pollution. The correlation coefficient as a function of the embedding dimension for predictions of the meteorological time series is presented in Figure (4). Forecasts up to 1944 and up to 1984 are shown by solid circles and solid line, and by open circles and dashed line, respectively. Notice that the plot has a peak at $n = 4$, suggesting that this time series has been generated by a dynamical system with a low dimensional attractor of dimension 4.

A plot the correlation coefficient versus the prediction-time step is shown in Figure (5), where the embedding dimension has been set to 4, according to the results of figure (4). Forecasts from 1910 to 1944 and from 1910 to 1984 are indicated by solid circles and solid line, and by open circles and dashed line, respectively. The prediction accuracy deteriorates rapidly with increasing time step in both cases. This indicates that the time series has only short-term predictability. The plots shown in Figure (4) and Figure (5) appear to be typical of a low-dimensional chaotic time series. Such diagnosis, however, is dangerous, since colored noise with many degrees of freedom can also provide a similar trend in prediction, when handling relatively small number of data points.

Figure (6) and Figure (7) show a plot of the correlation coefficient versus the embedding dimension and prediction-time step respectively for $f^{-1.8}$ (Miyano, 1994; Miyano et al. 1992). The fractal dimension was estimated using the algorithm developed by Higuchi (1988). The power spectrum of the random noise can be described as $f^{-1.8}$ with respect to frequency f, according to the relation between the fractal dimension D and the power law index α of $f^{-\alpha}$: $D = (5 - \alpha)/2$. In both Figure (6) and Figure (7), solid circles and solid line correspond to training and testing set size of 50, while open circles and dashed line correspond to training and test size of 500. Notice the sensitivity to the number of data.

4 Forecasting the global temperature time series

We tested three different approximation techniques to make predictions one step ahead, with an embedding dimension equal to 4:

1. Gaussian Hyper Basis Functions with a linear term. A network of 5 Gaussian units has been used, to which a linear term has been added. The linear term has been computed first, and then the residuals have been approximated with the gaussian network. Minimization of the mean square error was run over the coefficients, the centers and the variances of the gaussians, for a total of 30 free parameters.
2. Multilayer Perceptron with one hidden layer. A standard Multilayer Perceptron network was used for the prediction, with 4 hidden sigmoidal units, for a total of 24 parameters.
3. linear regression. A linear regression model was fitted to the data, using the statistical package Splus (Becker, Chambers and Wilks, 1988).

We use the first 45 data points, i.e., the time series from 1861 to 1909, as training set, and tested the resulting approximation on three different test sets: 1910–1944, 1910–1964, 1910–1984. For each experiment we measured the root mean square error ε:

$$\varepsilon = \sqrt{\sum_{\alpha=1}^{k}(x_\alpha - \hat{x}_\alpha)^2}$$

where the sum runs over the elements of the set being tested, x_α are test values and $\hat{x}_\alpha$ are the values predicted by the model. We also measured for each test set the variance

$$\sigma = \sqrt{\sum_{\alpha=1}^{k}(x_\alpha - <x>)^2}$$

where $<x>$ is the average value of the test set. Notice that the quantity $\frac{\varepsilon}{\sigma}$ is particularly significant, because if it has value zero then predictions are perfect, while if it is equal to 1 then predictions are no better that the average of the targets.
The experimental results for ε and σ have been reported in table (1), while the forecasts are shown in figures (8), (9) and (10) respectively. In the upper part of the figures we displayed the observed time series, represented by a dashed line, and the trained model, represented by the solid line. Solid circles have been used for the test set and white circles for the training set. In the lower part of the figures the residuals of the approximation have been shown.

It is clear that the Hyper Basis Function model makes best forecasts and that the linear regression makes the worst. This suggest that the dynamical behavior of the time series is nonlinear. It should be noticed, however, that prediction error increases remarkably after 1965 for all the models, although it is more pronounced in the

	Linear	HBF	MLP
e (training)	0.10	0.09	0.90
e (1910–1944)	0.14	0.10	0.12
e (1910–1964)	0.14	0.10	0.13
e (1910–1984)	0.16	0.12	0.15
e/σ (training)	0.84	0.83	0.81
e/σ (1910–1944)	1.13	0.82	1.01
e/σ (1910–1964)	1.25	0.88	1.11
e/σ (1910–1984)	1.31	0.98	1.18

Table 1: Root-mean-squared prediction error (e) and normalized Root-mean-squared prediction error (e/σ) for the 3 technique we tested. See text for explanation

Multilayer Perceptron and linear regression. This is not inconsistent with the idea of global warming suggested by Jones et al. In fact, assume that a model is successful in learning the dynamics underlying the meteorological time series that correspond to the period 1861–1909. Then, one possible interpretation for the increase in the prediction error after 1965 is that a change in the climate dynamics took place at some point after 1965.

Such straightforward interpretation could be, however, a misdiagnosis, since the trend in prediction error could be due to failure in generalizing the underlying dynamics. In fact, according to our recent work (Miyano, 1994), a similar trend in prediction error can also be present in forecasting colored noise with relatively few data points. Figure (11) shows the $f^{-1.8}$-noise observed for a semiconductor device (Miyano et al. 1992). In Figure (12), we present results of learning and predictions of the random noise by regularization network with 3 input nodes and 5 hidden nodes without linear terms. Figure (13) shows residuals obtained by subtracting the corresponding target from each prediction. We use first 250 points in the training set. In this case, the optimal embedding dimension is assumed to be 3 according to the results shown in Figure (7). The predictions agree well with the targets except for the portion from *time* = 300 to 350 during which the network forecasts lower values than observed. From Figures (12) and (13), one would make a misdiagnose that the time series was low-dimensional chaos and that the discrepancy between *time* = 300 and 350 indicated some change in the dynamics. The discrepancy is, however, clearly due to failure in generalization.

5 Discussion and open problems

The present considerations on the meteorological time series leads to two possible interpretations of the global climate: one is that a low-dimensional climate attractor may exist and that the climate dynamics may have altered at some point after 1965; the other is that the temperature variations may be colored noise with many degrees of freedom. The latter interpretation would lead to the following forecast of the future trend in the climate: the global temperature would begin to decrease at some point in the future, since colored noise has a zero mean in a long time period. Within the framework of the present study we can dismiss neither of the interpretations. The present work is still in a preliminary stage. In a future paper, we plan to forecast non-hand-scanned meteorological time series with more data points and to clarify whether the climate is low-dimensional chaos or not.

Acknowledgment

We are grateful to Tomaso Poggio, Reginald Newell, Jim Hutchinson, John Harris, and Yoshikazu Tobinaga for valuable discussions.

A Regression techniques

In this section we describe the different regression techniques that we have used to forecast and analyze the global temperature time series. The first three are based on the assumption that the data can be well approximated by a parametric function of the form

$$f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha H(\mathbf{x}; \mathbf{w}_\alpha)$$

where the c_α and the $\mathbf{w}_\alpha$ are considered free parameters and are found by a least squares technique. Models of this type are usually called “neural networks”, because they can be represented by a network with one layer of hidden units. The last technique is a local technique, similar in spirit to nearest neighbor models and to kernel regression.

A.1 Linear model

The simplest model of the form (1) that we can build is one in which the function f is linear, and therefore the relation between the past values of x and the future ones is of the type:

$$x(t) = a_0 + a_1 x(t-1) + a_2 x(t-2) + \cdots + a_n x(t-n) + \xi_t \quad (3)$$

This is the so called AR (autoregressive) model, and is one of the many linear models that have been developed in time series analysis. There is clearly a huge literature about linear models (see for example Myers, 1986; Draper and Smith, 1981; Searle, 1971) and we will not spend more time on this topic in this paper. In our experiments the coefficients of the model have been computed according to standard least square routines, that correspond to assuming that the random variables ξ_t in eq. (3) are independent and have Gaussian distribution with zero mean. The statistical package Splus (Becker, Chambers and Wilks, 1988) has been used to perform this calculation.

A.2 Gaussian Hyper Basis Functions with a linear term

One of the parametrizations we used for the prediction task had the following form:

$$f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha e^{-w_\alpha \|\mathbf{x}-\mathbf{t}_\alpha\|^2} + \mathbf{a}\cdot\mathbf{x} + d \ . \qquad (4)$$

The coefficients $c_\alpha, \mathbf{a}, d$, the centers $\mathbf{t}_\alpha$, the widths w_α of the gaussians were considered as free parameters. Techniques of this type are quite common in the neural network literature (Moody and Darken, 1989; Poggio and Girosi, 1990), although the linear term has not beed used very often. When the widths of gaussians have all the same value the technique is a particular case of is a particular case of a general class of approximation techniques, called regularization networks, that share the property that they can be justified in the framework of regularization theory (Girosi, Jones and Poggio, 1993) and can be represented by a network with one layer of hidden units.

Usually the parameters of the network are found by minimizing the mean square error by some numerical minimization technique. In our case, since the number of data points available is very small, we do not expect to be able to model surfaces much more complicated than an hyperplane. Therefore, we first fit an hyperplane to the data (the term $\mathbf{a}\cdot\mathbf{x} + d$), and then estimate the parameters of the rest of the expansion by fitting the residuals of the hyperplane approximation, choosing a small number n of basis functions. The gaussian basis functions are therefore to be considered as a "corrective term" to a linear approximation technique. The minimization technique we used for estimating the parameters of the gaussian basis functions is a stochastic gradient descent algorithm, described in appendix B.

A.3 Multilayer perceptrons

Multilayer perceptrons (MLP) are an approximation technique that is based, in its most common form, on the following parametric representation:

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i \sigma(\mathbf{x}\cdot\mathbf{w}_i + \theta_i) \ ,$$

where σ is a sigmoidal function, that in our case we set to $\frac{1}{1+e^{-x}}$, that is one the most common choices. In our computations the parameters of the network have been estimated using backpropagation and the generalized delta rule (Rumelhart, Hinton and Williams, 1986; Hecht-Nielsen, 1989), that is a form of stochastic gradient descent.

A.4 Simplex projection method

The simplex projection method, that has been used by Sugihara and May (1990) to tell chaotic time series from non chaotic ones, is a local approximation method, very close to the k-nearest neighbour technique and kernel regression. Suppose that a data set of N data points in d dimensions has been given, consisting of input-output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ that have been obtained by randomly sampling an unknown function f in presence of noise. When the value of f at a point $\mathbf{x}$ that does not belong to the data set has to be computed, first its closest $d+1$ points $\mathbf{x}_{i_1}, \ldots \mathbf{x}_{i_{d+1}}$ in the data set are found. Then the value of the function at $\mathbf{x}$ is estimated by a weighted average of the values of the function at the data points $\mathbf{x}_{i_1}, \ldots \mathbf{x}_{i_{d+1}}$, that is the values $y_{i_1}, \ldots y_{i_{d+1}}$. Points that are more far from $\mathbf{x}$ receive a smaller weight, according to an exponential decay. In formulas, the estimated value of f at $\mathbf{x}$ is

$$f(\mathbf{x}) = \frac{\sum_{\alpha=1}^{d+1} y_{i_\alpha} e^{-\sigma d_\alpha}}{\sum_{\alpha=1}^{d+1} e^{-\sigma d_\alpha}}$$

where we have defined

$$d_\alpha = \|\mathbf{x} - \mathbf{x}_{i_\alpha}\|$$

and σ is a parameter that define the locality of the technique, and can be set with cross-validation techniques. This technique can be considered as an approximation of kernel regression with e^{-x} as a kernel. In fact in this case kernel regression would have the form:

$$f(\mathbf{x}) = \frac{\sum_{1=1}^{N} y_i e^{-\sigma\|\mathbf{x}-\mathbf{x}_i\|}}{\sum_{i=1}^{N} e^{-\sigma\|\mathbf{x}-\mathbf{x}_i\|}} \qquad (5)$$

The method we use in this paper is equivalent to take only the closest $d+1$ points in the expansion (5).

B Stochastic gradient descent

Let $H(\xi)$ be a function that has to be minimized with respect to the set of parameters ξ. The standard gradient descent algorithm is an iterative algorithm in which the set of parameters ξ evolves toward the minimum according to the following law:

$$\frac{d\xi}{dt} = -\omega \frac{\partial H(\xi)}{\partial \xi} \qquad (6)$$

where t is a time parameter in the iteration loop of the optimization process and $\omega > 0$ is called *learning rate.* One of the many inconveniences of this algorithm is that, if converges, it converges to a local minimum. Since in the optimization problems arising from the training of a neural network it is known that multiple local minima are present, it is important to have a technique that is able to escape at least some of the local minima and get close to the global minimum. A simple way to achieve this consists in adding a stochastic term in eq. (6), that becomes:

$$\frac{d\xi}{dt} = -\omega \frac{\partial H(\xi)}{\partial \xi} + \eta(t) \qquad (7)$$

where $\eta(t)$ is random noise. As an effect of the addition of the noise term η the set of parameters ξ will not always follow the direction of the gradient, and will sometime go uphill instead of downhill, having a chance to escape some local minima. In our case the random noise $\eta(t)$ was synthesized by the following equation:

$$\eta(t) = \sum_{k=1}^{X} (2\pi ck)^{-\frac{\alpha}{2}} [Acos(2\pi ckt) + Bsin(2\pi ckt)] \quad (8)$$

where c and X are set to 0.01 and 20000, respectively, and A and B random numbers lying between 0 and 1. The power spectrum of $\eta(t)$ is given as $f^{-\alpha}$ with respect to the frequency f. In order to verify the validity of the algorithm, we adapt it to optimizing the regularization network to predict a chaotic time series synthesized by Hènon mapping (embedding dimension is set to 2. White noise (i.e., $\alpha = 0$), f^{-1}-noise (pink noise), and f^{-2}-noise are used as perturbation. It is found that f^{-1}-noise works as well as white noise, while f^{-2}-noise does not. The value of the normalized root mean square error ε/σ for a test set of 1000 points and a training set of 1000 examples was 0.097 for a network with 5 basis functions and no linear term, in the case in which f^{-1}-noise on 1000 examples. This good result confirmed that the stochastic gradient descent algorithm worked correctly and achieved some good local minimum.

References

[1] R.A. Becker, J.M. Chambers, and A.R. Wilks. *The new S language.* Computer Science Series. Wadsworth & Brooks/Cole Advanced Books and Software, Pacific Grove, CA, 1988.

[2] M. Casdagli. Nonlinear prediction of chaotic time-series. *Physica D*, 35:335–356, 1989.

[3] N.R. Draper and H. Smith. *Applied Regression Analysis.* Wiley, New York, 1981. (second edition).

[4] C. Essex, T. Lookman, and M. A. H. Nerenberg. The climate attractor over short timescales. *Nature*, 326:64–66, 1987.

[5] J.D. Farmer and J.J. Sidorowich. Predicting chaotic time series. *Physical Review Letters*, 59:845, 1987.

[6] F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. A.I. Memo No. 1430, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

[7] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, February 1995. (to appear).

[8] P. Grassberger. Do climatic attractors exist? *Nature*, 323:609–612, 1984.

[9] R. Hecht-Nielsen. Theory of backpropagation neural network. In *Proceedings of the International Joint Conference on Neural Networks*, pages I–593–I–605, Washington D.C., June 1989. IEEE TAB Neural Network Committee.

[10] P. D. Jones, T. M. W. Wigley, and P. B. Wright. Global temperature variations between 1861 and 1984. *Nature*, 322:430–434, 1986.

[11] E. N. Lorenz. Dimension of weather and climate attractor. *Nature*, 353:241–244, 1991.

[12] T. Miyano. 1994. (in preparation).

[13] T. Miyano, M. Fujito, K. Fujimoto, , and A. Sanjoh. Fractal analysis of leakage-current fluctuations of Si metal-oxide-semiconductor capacitors for the characterization of dry-etching damage. *Appl. Phys. Lett.*, 61:2521, 1992.

[14] J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

[15] R.H. Myers. *Classical and Modern Regression with Applications.* Duxbury, Boston, 1986.

[16] C. Nicolis and G. Nicolis. Is there a climatic attractor? *Nature*, 311:529–532, 1984.

[17] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), September 1990.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(9):533–536, October 1986.

[19] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Parallel Distributed Processing.* MIT Press, Cambridge, MA, 1986.

[20] S. R. Searle. *Linear Models.* J. Wiley, New York, 1971.

[21] G. Sugihara and R.M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344:734–741, 1990.

[22] Z. Wu, R. E. Newell, and J. Hsuing. Possible factors controlling marine temperature variations over the past century. *J. Geophysical Research*, 95:11799–11810, 1990.

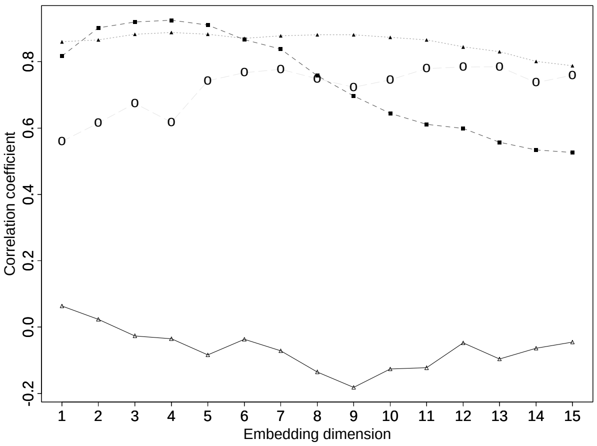

Figure 2: Plot of the correlation coefficient as a function of embedding dimension for various synthetic time series: white noise (open triangles and dashed line); $f^{-1.6}$-noise (solid triangles and dotted line); Hènon mapping superimposed on white noise (solid squares and short-dashed line); and sine wave superimposed on white noise (open circles and long-dashed line). Predictions on a test set of 150 points were made from a training set of 150 examples.

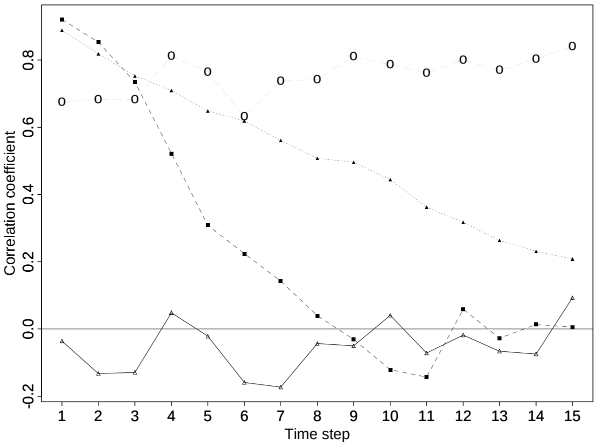

Figure 3: Plot of the correlation coefficient as a function of prediction-time steps for various synthetic time series; white noise (open triangles and dashed line); $f^{-1.6}$-noise (solid triangles and dotted line); Hènon mapping superimposed on white noise (solid squares and short-dashed line); and sine wave superimposed on white noise (open circles and long-dashed line). Predictions on a test set of 150 points were made from a training set of 150 examples. The embedding dimension is set to 3.

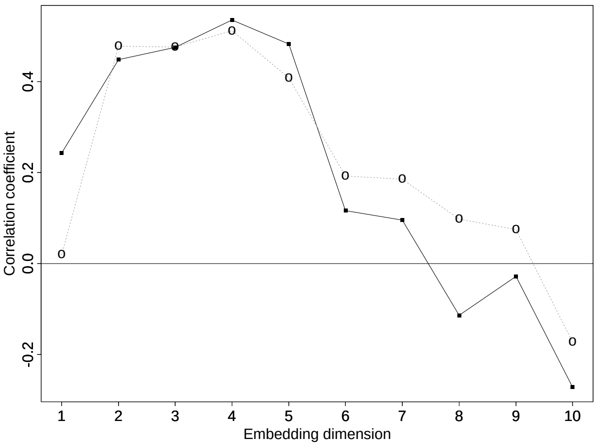

Figure 4: Plot of the correlation coefficient as a function of embedding dimension for the meteorological time series. Predictions are made from the first 45 data points, i.e., the data from 1861 to 1915. Forecasts up to 1944 and up to 1984 are shown by solid circles and solid line, and by open circles and dashed line, respectively.

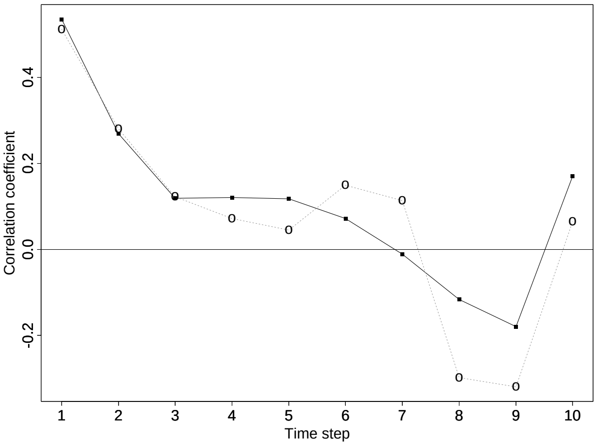

Figure 5: Plot of the correlation coefficient as a function of prediction-time step for the meteorological time series. Forecasts are made from the first 45 data points, i.e., the data from 1861 to 1909. Forecasts from 1910 to 1944 and from 1910 to 1984 are shown by solid circles and solid line, and by open circles and dashed line, respectively. The embedding dimension has been set to 4, according to the results of fig. (4).

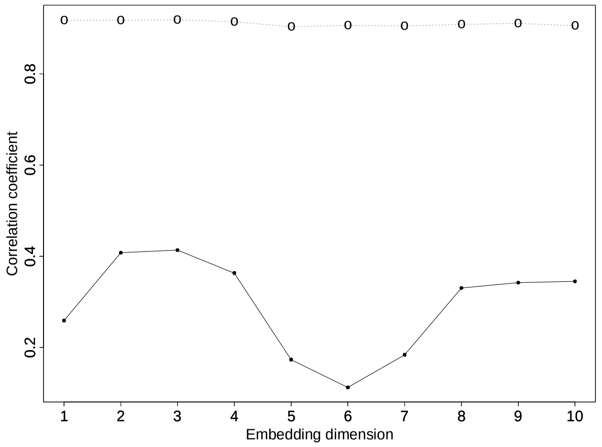

Figure 6: Plot of the correlation coefficient as a function of embedding dimension for colored noise with fractal dimension of 1.603. The random noise is not synthetic but a leakage-current fluctuations observed for a semiconductor device (Miyano et al., 1992). Solid circles and solid line: Predictions on a test set of 50 points are made from a training set of 50 examples. Open cicles dashed line: Predictions on a test set of 500 points and a training set of 500 examples.

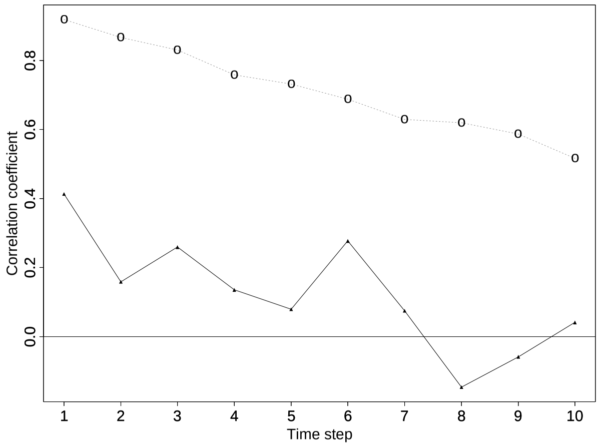

Figure 7: Plot of the correlation coefficient as a function of prediction-time step for the random noise observed in a semiconductor device. The embedding dimension is set to 3. Solid circles and solid line: Predictions on a test set of 50 points are made from a training set of 50 examples. Open cicles dashed line: Predictions on a test set of 500 points and a training set of 500 examples.

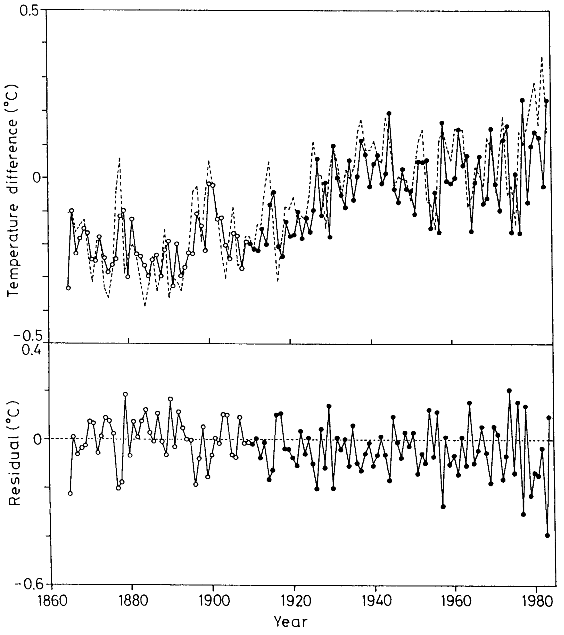

Figure 8: Forecasts of the meteorological time series by regularization network: forecasts(upper part); and residuals obtained by subtracting the corresponding target from each prediction (lower part). The network has been trained on the first 45 data points, i.e., the time series from 1861 to 1909. Solid circles and solid line indicate predictions on the input vectors that the network has not seen. Open circles and solid line indicate results of training. The observed time series is shown by dashed lines.

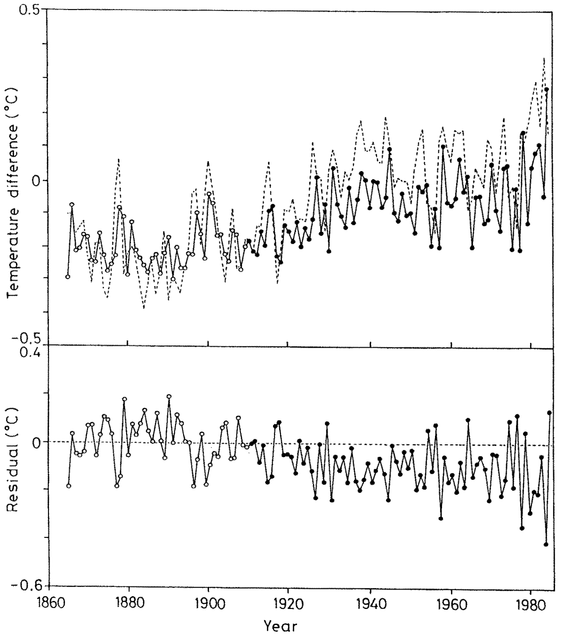

Figure 9: Forecasts of the meteorological time series by backpropagation network: forecasts(upper part); and residuals obtained by subtracting the corresponding target from each prediction (lower part). The network has been trained on the first 45 data points, i.e., the time series from 1861 to 1909. Solid circles and solid line indicate predictions on the input vectors that the network has not seen. Open circles and solid line indicate results of training. The observed time series is shown by dashed lines.

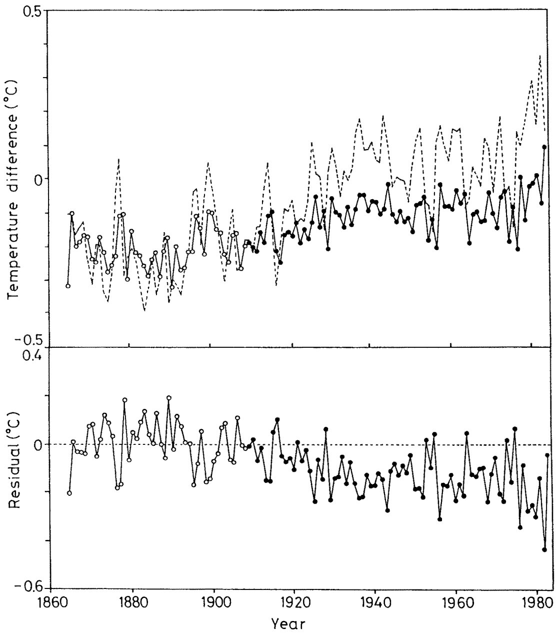

Figure 10: Forecasts of the meteorological time series by linear autoregression: forecasts(upper part); and residuals (lower part). The model has been trained on the first 45 data points, i.e., the time series from 1861 to 1909. Solid circles and solid line indicate predictions on the input vectors that the network has not seen. Open circles and solid line indicate results of fitting. The observed time series is shown by dashed lines.

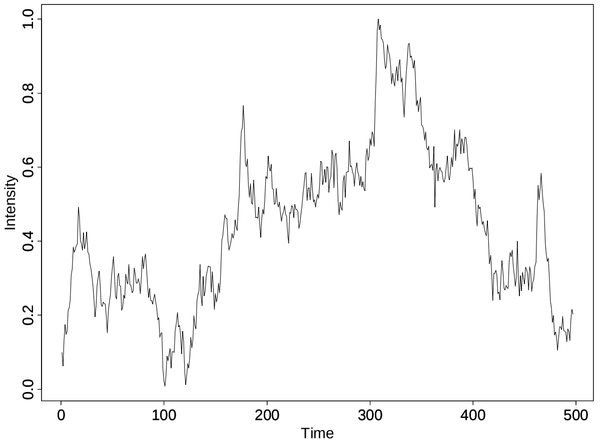

Figure 11: The $f^{-1.8}$-noise observed for a semiconductor device (Miyano, 1994; Miyano et al., 1992).

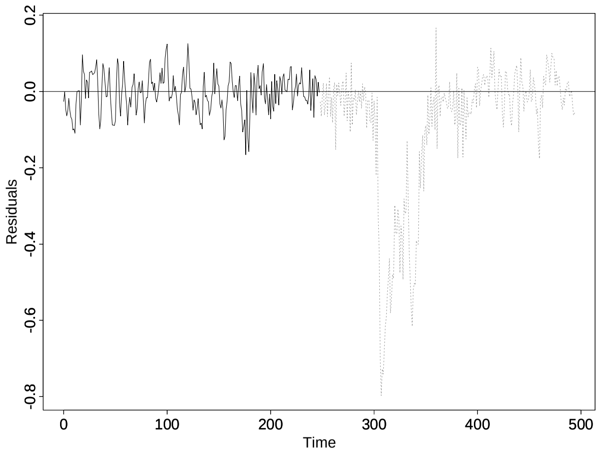

Figure 13: Residuals of the approximation of $f^{-1.8}$-noise by a Gaussian Hyper Basis Functions.

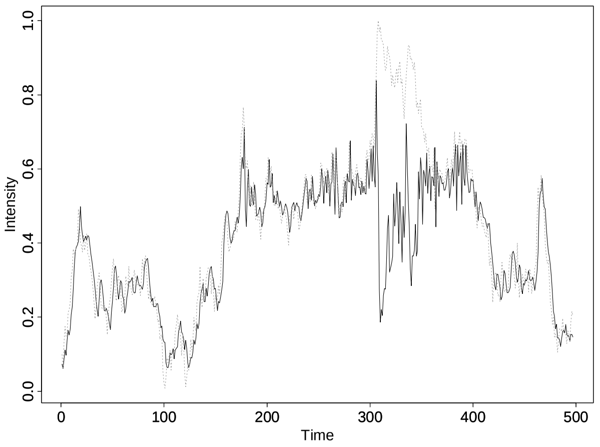

Figure 12: Results of training (from 0 to 249)and predictions (from 250 to 500) of the $f^{-1.8}$-noise using Gaussian Hyper Basis Functions without linear terms. The embedding dimension is set to 3. The network is trained for first 250 library patterns, i.e., the time series from $time$ = 0 to 250. The predictions agree well with the targets except for the portion from $time$ = 300 to 350 during which the network forecasts lower values than observed.