Discrimination power of measures for nonlinearity in a time series

Thomas Schreiber and Andreas Schmitz
Physics Department, University of Wuppertal, D-42097 Wuppertal, Germany

(Received 22 January 1997)

The performance of a number of different measures of nonlinearity in a time series is compared numerically. Their power to distinguish noisy chaotic data from linear stochastic surrogates is determined by Monte Carlo simulation for a number of typical data problems. The main result is that the ratings of the different measures vary from example to example. It therefore seems preferable to use an algorithm with good overall performance, that is, higher order autocorrelations or nonlinear prediction errors. [S1063-651X(97)16205-5]

PACS number(s): 05.45.+b

I. INTRODUCTION

The theory of nonlinear, deterministic dynamical systems provides powerful theoretical tools to characterize geometrical and dynamical properties of the attractors of such systems. Alongside the theoretical understanding of these systems, many of the typical phenomena have been realized in laboratory experiments. Many attempts have also been made to detect behavior characteristic of deterministic systems in field data, that is, time series recordings of real world phenomena. Not surprisingly, the coarse nature of these time series (finite number of points with finite resolution) makes it difficult to obtain unambiguous results. As a particular example, it has been pointed out [1] that linear stochastic processes with long range autocorrelations can lead to spuriously small estimates of the attractor dimension. (See also the discussion in Ref. [2].) The method of surrogate data [3] provides a rigorous statistical test for the null hypothesis that the data have been generated by a linear stochastic process. If this null hypothesis cannot be rejected, the results of a nonlinear analysis have to be regarded as spurious. In such a test, the value of some measure of nonlinearity is compared for the data and a number of randomized samples, the surrogates. The nonlinearity measure should be sensitive to the kind of nonlinearity suspected in the data, and it should be possible to estimate its value with low variance. In this paper we will numerically compare the performance of a selection of measures which have been proposed in the literature.

Apart from the mere detection of nonlinearity, nonlinear observables can be used to discriminate between distinct states of a system on the base of time series data. Most notably, claims have been made that measures derived from chaos theory are able to distinguish healthy patients from those with pathological biological rhythms, for example, cardiac arrhythmiae [4–6]. The results presented in this paper are also of relevance for the question of the preferable discriminating statistic in such a context. The most striking observation is that, although the simplest observables, notably simple prediction errors, show good overall performance, results differ immensely from application to application, which may explain the partially contradicting claims in the literature. If enough data are available to be split into a training set and a test set, and if a model for a reasonable alternative hypothesis can be constructed, then optimization of the test on typical data may be worthwhile.

II. TESTING FOR NONLINEARITY WITH SURROGATE DATA

Currently, the most general null hypothesis we know how to test against is that the data were generated by a stationary Gaussian linear stochastic process, maybe measured through an instantaneous measurement function [7]. Deviations from this null hypothesis can be detected by computing some nonlinear observable on the data. Since the probability distributions of such observables are generally not known analytically, they must be estimated by Monte Carlo resampling of the data. For this purpose one generates random data sets (surrogates) which conserve those properties of the data which are irrelevant for a given choice of the null hypothesis. For the hypothesis of a Gaussian linear stochastic process, the data and the surrogates must have the same autocorrelation function or, equivalently, the same power spectrum. For a nonlinearity test allowing for simple rescalings, the single time probability distribution also must be conserved. A (nonlinear) observable $t=t(\{x_n\})$ is estimated on the original data $\{x_n^0\}$ and all of the B surrogates $\{x_n^k\}$, $k=1,\ldots,B$. The distribution of t can be estimated from the values $t_k=t(\{x_n^k\})$. One can then test at a given level of significance for the assumption that $t_0=t(\{x_n^0\})$ was drawn from the same distribution. If this assumption is rejected, the original data $\{x_n^0\}$ are taken to be different from the linear surrogates, and are thus considered to be nonlinear at this level of significance.

The use of surrogate data has been promoted in the context of chaotic time series in Ref. [3]. Although this technique has made distinguishing chaos from noise much safer, some caveats remain. These will not be discussed in this paper; Refs. [8–10] provide noteworthy material. Throughout we will use examples where the known problems (nonstationarity, long coherence times) are of no concern.

There are two important parameters which characterize the performance of a statistical test. One is its *size* α, which is the probability that the null hypothesis is rejected, although it is in fact true. Specifying a *level of significance* $1-p$ of the test amounts to the statement that its size does not exceed p. It is customary to specify p *a priori*, and design the test accordingly. The important question of whether the surrogate data test indeed has the specified size has been previously addressed; see Refs. [7,8,10]. If the actual probability of a false rejection is larger than p, the test yields incorrect results. The above references give examples where this situation can occur with surrogate data tests. While excessive size renders the test useless, an actual size which is smaller than p is formally admissible. However, it can result in a dramatic decrease in discrimination power. In such cases (for example, if a fitted linear model is run to generate surrogates), it is therefore advisable to calibrate the test by using ''surrogate surrogate data'' [10]. Since the size of the test may depend on the particular realization of the null hypothesis, this calibration is usually quite cumbersome. We verified the correct test size for all the numerical examples in this paper by performing a series of tests on surrogate data fulfilling the null hypothesis.

While the size predominantly assesses the quality of the surrogate data sets, in this paper we want to evaluate the abilities of different *observables t* to detect nonlinearity. This property is quantified by the *power β* of the test. It is defined as the probability to reject the null hypothesis correctly when it is indeed false. The power of a statistical test can be determined empirically by repeating the test many times on different realizations of the data. Since we cannot make strong assumptions about the distributions of the observables, there is no alternative to this computationally expensive approach. However, in order to limit the computational effort, we performed tests at a rather low level of significance, for which only a few surrogate data sets are necessary.

III. MEASURES OF NONLINEARITY

We evaluated a number of different nonlinear observables. Most of them are at least inspired by the theory of nonlinear dynamical systems, and rely on a time delay embedding of the scalar time series. Embedding vectors in m dimensions are formed as usual: $\vec{x}_n=(x_{n-(m-1)\tau},\dots,x_n)$, where τ is the delay time. Since the Grassberger-Procaccia correlation dimension D_2 [11] seems to be among the most popular measures, we considered several variants of this algorithm. The correlation sum $C(\epsilon)$ at a scale ϵ is given by

$$C(\epsilon)=\text{const}\times\sum_{|i-j|>t_{\min}}\Theta(\|\vec{x}_i-\vec{x}_j\|-\epsilon). \tag{1}$$

Dynamically correlated pairs are discarded as usual, and const refers to the normalization. Since none of the examples in this study would allow for the identification of a true scaling region, we will choose the length scales for good discrimination power. Of course, this will make an interpretation as a fractal dimension or complexity measure impossible. In particular, we implemented two ways of turning $C(\epsilon)$ into a single number:

(1) A maximum likelihood (ML) estimator of the Grassberger-Procaccia correlation dimension D_2 is given by

$$t^{\text{ML}}(m,\tau,\epsilon)=\frac{C_m(\epsilon)}{\int_0^\epsilon \frac{C_m(\epsilon')}{\epsilon'}d\epsilon'}. \tag{2}$$

This expression is taken from Ref. [12]. The maximum likelihood estimation of the correlation dimension goes back to Ref. [13]. Therefore such quantities are generally referred to as the *Takens' estimator*.

(2) Brock *et al.* (BDS) [14] showed that for a sequence of independent random numbers, $C_m(\epsilon)=C_1(\epsilon)^m$ holds, where m is the embedding dimension. In the same paper, a formal test for this property was also introduced. Instead of the original BDS statistic, which was introduced in order to be able to give the asymptotic form of the probability distribution, we use the simpler expression

$$t^{\text{BDS}}(m,\tau,\epsilon)=C_m(\epsilon)/C_1(\epsilon)^m. \tag{3}$$

Other choices we tried are values of $C(\epsilon)$ at fixed length scales, which gave consistently less power, and dimension estimators based on pointwise dimensions. In the latter case, the scaling exponent of neighbor distances is determined for each point separately. The actual observable is then the mean or the median of these values [6,15]. Since we did not find any interesting deviations from the power of the maximum likelihood estimator t^{ML}, we did not include detailed results in this paper.

Many quantities which have been proposed in the literature for nonlinearity testing in some way or the other quantify the nonlinear predictability of the signal. Examples include genuine forecasting methods (e.g., Ref. [16]) but also the statistic proposed by Kaplan and Glass [17] and to some extent the false nearest neighbor techniques [18]. We use a particularly stable representative of the class of predictability measures:

(3) A nonlinear prediction error with respect to a locally constant predictor F can be defined by

$$t^{\text{PE}}(m,\tau,\epsilon)=\left(\sum[x_{n+1}-F(\vec{x}_n)]^2\right)^{1/2}. \tag{4}$$

The prediction over one time step is performed by averaging over the future values of all neighboring delay vectors closer than ϵ in m dimensions.

In Ref. [19] a nonlinear Volterra-Wiener model is claimed to be superior to other techniques when applied to short noisy signals. We compared the maximal feasible noise level for a detection of nonlinearity quoted in Ref. [19] to the performance of the locally constant predictor above for the Hénon, Ikeda, and Lorenz series. We found that t^{PE} gave either better (Hénon, Ikeda) or comparable (Lorenz) performance, and therefore did not include the Volterra-Wiener model in this study.

Further, we used the following nonlinear observables:

(4) Linear (two point) autocovariances can be generalized by introducing more than one lag. In the spectral domain, this generalization leads to the bispectrum and polyspectra [20]. Our (somewhat arbitrary) choice of a higher-order autocovariance (or cumulant) is

$$t^{\text{C3}}(\tau)=\langle x_n x_{n-\tau} x_{n-2\tau}\rangle. \tag{5}$$

(5) A simple quantity which is frequently used to detect deviations from time-reversibility is

$$t^{\text{REV}}(\tau)=\langle (x_n-x_{n-\tau})^3\rangle. \tag{6}$$

We explicitly indicated the adjustable parameters which can be chosen using several different strategies. One possibility is to optimize the adjustable parameters. This has to be done either for data which is not subsequently used for the test, or it has to be done individually for each data set and surrogate. The former requires a knowledge of the correct answer for the ''training data'' which is rather uncommon. The latter is computationally extremely expensive, and care has to be taken in order to avoid overfitting of the data. Note, for example, that minimizing prediction errors does not necessarily optimize the discrimination power.

In the present work, we fix as many parameters as possible to reasonable *ad hoc* values prior to the tests. Before each test, a brief survey was performed as to which embedding dimensions and delay times lead to satisfactory results for each quantity. We feel that this procedure comes closest to what one can do in practice, where also a formal optimization of the discrimination power is impossible. The length scale ϵ was either determined as a fixed fraction ($\frac{1}{4}$) of the root-mean-squared (1), (2) or the peak-to-peak amplitude (3) of the data.

IV. IMPLEMENTATION AND RESULTS

The surrogate data sets will be generated as described in Ref. [7], which is the appropriate method when the null hypothesis is that the data have been generated by a Gaussian linear stochastic process, possibly measured through a monotonic, instantaneous, time-independent measurement function. In brief, the method is based on an ordinary phase randomized surrogate series $S=\{s_n, n=1,\dots,N\}$ which has the same sample power spectrum as the time series $X=\{x_n, n=1,\dots,N\}$. Such a surrogate is obtained by taking the Fourier transform of X, randomizing the phases, and inverting the transform. Now the following two steps are iterated alternatingly:

(1) The surrogate series is brought to the sample distribution of X by rank ordering,

$$s'_n = x_{\mathrm{index}(\mathrm{rank}(s_n))}. \tag{7}$$

Here, $\mathrm{rank}(s_n)=k$ and $\mathrm{index}(k)=n$ if s_n is the kth smallest value in S. After this step, S' and X have the same distribution of values, but the power spectrum may have changed.

(2) The Fourier amplitudes of $S'=\{s'_n, n=1,\dots,N\}$ are replaced by those of X. The resulting series S'' has the same sample power spectrum as X. This step may, however, alter the distribution of values.

In Ref. [7], numerical evidence and heuristic arguments are given that this scheme indeed converges to a sequence with the same distribution *and* the same power spectrum as the data. While formal convergence can only be expected for infinitely long sequences, the approximation is satisfactory for finite data length. If the deviation from a Gaussian distribution or linear correlations in the time series are not too strong, the usual amplitude-adjusted phase-randomized surrogates [3] yield an accurate test as well. Our results do not explicitly depend on the particular method of generating constrained Monte Carlo realizations.

As mentioned above, we do not know the probability distributions of the nonlinear observables used in this paper. In particular, Gaussianity cannot be assumed. Therefore we have to employ a nonparametric, rank-based test, as has been suggested in Ref. [9]. A test is called *one sided* if the null hypothesis is rejected only if the data deviate from the surrogates in a specified direction. In this case and at a given size α, we create $B=1/\alpha-1$ surrogate data sets, and compute the test statistic t_0 on the original data set and its value $t_k, k=1,\dots,B$ on each of the surrogates. Since we have a total of $1/\alpha$ sets, the probability for each of them to have the smallest value of t by chance is just α, as desired. For two-sided tests, we generate $B=2/\alpha-1$ surrogates. The probability for any of the $2/\alpha$ sets to have either the smallest or largest value of t is then again α.

TABLE I. Maximal feasible noise level for the detection of nonlinearity with $\beta=0.95$ (0.7). Results for the Hénon map.

Statistic	Parameters	Feasible noise level $a_{\max}$ $\beta=0.95$	$\beta=0.7$
t^{ML}	$m=2$	0.7	0.9
t^{BDS}	$m=3$	1.1	1.3
t^{PE}	$m=3$	1.2	1.5
t^{C3}	$\tau=1$	1.1	1.5
t^{REV}	$\tau=1$	1.4	1.8

For the nonlinearity measures inspired by the theory of deterministic dynamical systems [(1)–(3) above], we expect nonlinearity in the data to result in lower values. Thus it is natural to perform one-sided tests. For the remaining two measures we perform two-sided tests. In order to limit the computational burden, all tests are carried out at the 90% level of significance; that is, with nine (19 for two-sided tests) surrogates. For practical applications, at least a 95% confidence is usually required. The power can be increased by performing tests based on more than the minimal number of surrogate data sets.

For purely deterministic signals, we would almost invariably obtain a discrimination power of $\beta=1$. Therefore we contaminate deterministic sequences $\{x_n\}$ with noise $\{\eta_n\}$, which consists of a phase-randomized copy of the sequence. Thus the noise is random, but with the same power spectrum as the data (*in-band noise*). The noisy data are given by

$$s_n = \left(\frac{1}{1+a^2}\right)^{1/2} (x_n + a\,\eta_n). \tag{8}$$

The way the noise is generated and added guarantees that the power is not dominated by changes in the autocorrelations or the variance of the data.

One sequence of tests is performed at different noise levels in order to determine the maximal feasible noise level

TABLE II. Fraction of successful rejections out of 1000 tests, noisy Lorenz data. The errors are based on the assumption of a binomial distribution for independent trials. No significant rejection is possible with t^{C3} and t^{REV}.

Statistic	Parameters	Power β
t^{ML}	$m=3$	0.25 ± 0.02
t^{BDS}	$m=2$	0.24 ± 0.02
t^{PE}	$m=4$	0.66 ± 0.02
t^{C3}	$\tau=3$	0.09 ± 0.01
t^{REV}	$\tau=3$	0.10 ± 0.01

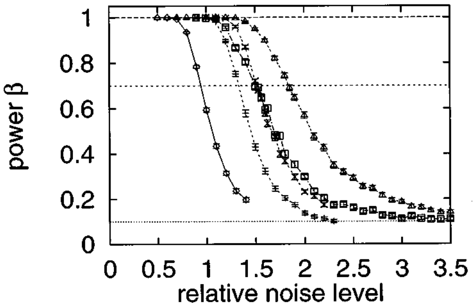

FIG. 1. Comparison of discrimination power for different nonlinearity measures and noise levels, for Hénon data with in-band noise. Curves from the left: correlation statistics t^{ML} and t^{BDS}, prediction error t^{PE} (crosses), third-order cumulant t^{C3}, and time asymmetry t^{REV}. The size of the test was taken to be 0.1.

which allows for the detection of nonlinearity with a power of $\beta = 0.95$ (0.7). The practical usefulness of tests with power less than $\beta = 0.7$ seems questionable. In this sequence, 2000 individual tests with Hénon time series of length 2048 were carried out for each point. The results are summarized in Table I and Fig. 1. For this discrete time system, a unit time delay seems most appropriate.

Further, we evaluated the different quantities for a number of particular data problems, time series from the Lorenz equations, an NMR laser experiment, and an assembly of uncoupled tent maps. In Table II we show the results for time series of the Lorenz system at standard parameter values. 2048 samples of the x coordinate were recorded every 0.08 time units. We added noise of amplitude $a = 1.3$. It was checked for each of the different observables (but with fewer tests) that other choices of the lag time and the embedding dimension did not lead to significantly better results.

A long experimental time series from a NMR laser experiment [21] was split into 600 segments with 1000 points each. We added in-band noise of amplitude $a = 0.8$. Results are shown in Table III.

Finally, we consider an assembly of uncoupled tent maps. Each individual map is given by $x_{n+1} = 2x_n$ if $x < 0.5$ and $x_{n+1} = 2 - 2x_n$ if $x \geqslant 0.5$. The recorded variable is the sum of the variables of N individual tent maps. No noise is added. The discrimination power is measured as a function of N. In Fig. 2 we show the results for the time asymmetry, prediction error, and ML statistics. Table IV shows the results for all the nonlinearity measures at $N = 16$. In this example, the time asymmetry statistic is doing extremely well. The prediction error also gives a reasonable power, while all other quantities basically fail, although different settings for m were considered.

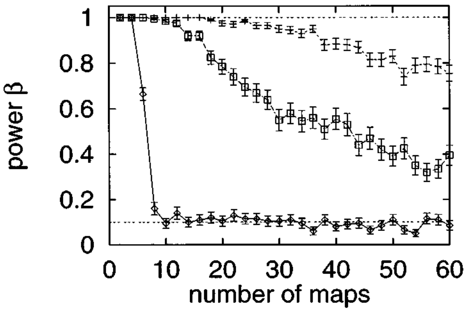

FIG. 2. Discrimination power for uncoupled tent maps. In this figure, results are shown for three selected nonlinearity measures. From above: time asymmetry t^{REV}, prediction error t^{PE}, $m = 4$, and ML dimension estimator t^{ML}, $m = 2$. The number of maps was varied in steps of two, and each point was obtained with 200 tests. The size of the test was taken to be 0.1.

V. CONCLUSIONS

The results presented in Tables I–IV and the figures suggest that the root-mean-squared error of a simple nonlinear predictor gives consistently good discrimination power. Other nonlinearity measures give even better performance in some cases, but fail in others. In particular, the time, reversal asymmetry does very well most of the time. but can also fail completely. Asymmetry under time reversal is a sufficient and powerful indicator of nonlinearity, but not a necessary condition. Which algorithm is to be preferred in a particular situation depends on the availability of an independent check for the discrimination power. In the typical situation that only few precious data sets, or even just one recording (as in long-term geophysical observations) is available, it seems advisable to use a robust, general purpose statistic with few adjustable parameters, for example a simple prediction error. If asymmetry under time reversal appears under visual inspection of the data, a simple statistic like t^{REV} will probably give the best results.

The null hypothesis we adopted in this work was chosen since it is the most general one that excludes nonlinear determinism and that can be tested for properly. If we are in fact looking for deterministic structure in a signal, then simple statistics like t^{REV} and t^{C3} which are based on higher-order cumulants are not very attractive because they are also quite sensitive to those deviations from the null hypothesis we are *not* looking for. The formal test discussed in this paper answers the question if *any* deviation from a (rescaled) Gaussian linear stochastic process can be detected. Surrogate data tests have, however, been mostly used with the question in mind if it is legitimate and useful to use methods from

TABLE III. Fraction of successful rejections out of 600 tests, noisy NMR laser data.

Statistic	Parameters	Power β
t^{ML}	$m=3$	0.61±0.03
t^{BDS}	$m=3$	0.86±0.02
t^{PE}	$m=3$	0.79±0.02
t^{C3}	$\tau=3$	0.45±0.03
t^{REV}	$\tau=1$	0.35±0.02

TABLE IV. Fraction of successful rejections out of 100 tests, sum of 16 uncoupled tent maps.

Statistic	Parameters	Power β
t^{ML}	$m=2$	0.11±0.03
t^{BDS}	$m=2$	0.10±0.03
t^{PE}	$m=4$	0.92±0.03
t^{C3}	$\tau=1$	0.10±0.03
t^{REV}	$\tau=1$	1.00±0.00

dynamical systems theory. This amounts to specifying a particular class as an alternative hypothesis. In such a case we should choose the discriminating statistic accordingly, that is, from the arsenal of dynamical systems methods.

Let us finally remark that a couple of tests for nonlinear properties of time series have been proposed which use surrogate data in a different way or not at all. Rather than estimating the distribution of the observable t from a randomized sample, it is sometimes calculated on the base of some assumptions. If the null hypothesis is that of a purely Gaussian linear random process (without distortion), significance levels for higher-order correlation functions can be derived. Some authors, e.g., Refs. [6,19], observe that most observables t are temporal averages over individual quantities t_n determined for each point in a time sequence. In order to derive the distribution of t from the knowledge of $\{t_n\}$, however, one has to make certain assumptions. Reference [6] assumes Gaussianity and independence of the sequence $\{t_n\}$, while Ref. [19] needs only independence. We do not see what should justify the assumption of point-to-point independence of $\{t_n\}$ for autocorrelated time series data; indeed, we empirically find the assumption to be wrong at least for prediction errors and pointwise dimensions. The common positive correlation among the t_n leads to an underestimation of the variance of the average t, and thus to a dangerous overestimation of the significance of the test.

ACKNOWLEDGMENTS

We thank James Theiler, Daniel Kaplan, Peter Grassberger, and Holger Kantz for stimulating discussions. This work was supported by the SFB 237 of the Deutsche Forschungsgemeinschaft.

[1] A. R. Osborne and A. Provenzale, Physica D **35**, 357 (1989); J. Theiler, Phys. Lett. A **155**, 480 (1991).

[2] C. Nicolis and G. Nicolis, Nature (London) **311**, 529 (1984); P. Grassberger, *ibid.* **323**, 609 (1986); C. Nicolis and G. Nicolis, *ibid.* **326**, 523 (1984); P. Grassberger, *ibid.* **326**, 524 (1987).

[3] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).

[4] G. E. Morfill and G. Schmidt, Phys. Bl. **50**, 156 (1994).

[5] J. Kurths, A. Voss, P. Saparin, A. Witt, H. J. Kleiner, and N. Wessel, CHAOS **5**, 88 (1995).

[6] J. E. Skinner, M. Molnar, and C. Tomberg, Integr. Physiol. Behav. Sci. **29**, 217 (1994).

[7] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).

[8] J. Theiler, P. S. Linsay, and D. M. Rubin, in *Time Series Prediction: Forecasting the Future and Understanding the Past*, edited by A. S. Weigend and N. A. Gershenfeld, SFI Studies in the Sciences of Complexity Proceedings Vol. XV (Addison-Wesley, Reading, MA, 1993), p. 429.

[9] J. Theiler and D. Prichard, Physica D **94**, 221 (1996).

[10] J. Theiler and D. Prichard, Fields Inst. Commun. **11**, 99 (1997).

[11] P. Grassberger and I. Procaccia, Physica D **9**, 189 (1983).

[12] J. Theiler, Phys. Lett. A **135**, 195 (1988).

[13] F. Takens, in *Dynamical Systems and Bifurcations*, edited by B. L. J. Braaksma, H. W. Broer, and F. Takens, Lecture Notes in Mathematics Vol. 1125 (Springer, Heidelberg, 1985).

[14] W. A. Brock, W. D. Dechert, J. A. Scheinkman, and B. LeBaron, *A Test for Independence Based on the Correlation Dimension* (University of Wisconsin Press, Madison, 1988); W. A. Brock, D. A. Hseih, and B. LeBaron, *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence* (MIT, Cambridge, MA, 1991).

[15] N. Birbaumer, W. Lutzenberger, H. Rau, G. Mayer-Kress, and C. Braun, Int. J. Bifurc. CHAOS **6**, 267 (1996).

[16] G. Sugihara and R. May, Nature (London) **344**, 734 (1990).

[17] D. T. Kaplan and L. Glass, Phys. Rev. Lett. **68**, 427 (1992).

[18] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, Phys. Rev. A **45**, 3403 (1992).

[19] M. Barahona and C.-S. Poon, Nature (London) **381**, 215 (1996).

[20] T. Subba Rao and M. M. Gabr, *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Lecture Notes in Statistics Vol. 24 (Springer, New York, 1984).

[21] M. Finardi, L. Flepp, J. Parisi, R. Holzner, R. Badii, and E. Brun, Phys. Rev. Lett. **68**, 2989 (1992).