Measure of predictability

Weiguang Yao,* Christopher Essex,† Pei Yu,‡ and Matt Davison§
Applied Mathematics Department, University of Western Ontario, London, Ontario, Canada N6A 5B7
(Received 25 November 2003; published 14 June 2004)

Many techniques have been developed to measure the difficulty of forecasting data from an observed time series. This paper introduces a measure which we call the "forecast entropy" designed to measure the predictability of a time series. We use attractors reconstructed from the time series and the distributions in the regular and tangent spaces of the data which comprise the attractor. We then consider these distributions on different scales. We present a formula for calculating the forecast entropy. To provide a standard of predictability, we define an idealized random system whose forecast entropy will be maximal; we then use this measure to rescale the forecast entropy to lie in the range [0,1]. The time series obtained from several chaotic systems as well as from a pseudorandom system are studied using this measure. We present evidence that the forecast entropy can be used as a tool for determining optimal delays and embedding dimensions used for reconstructing better attractors. We also show that the forecast entropy of a random system has completely different characteristics from that of a deterministic one.

DOI: 10.1103/PhysRevE.69.066121 PACS number(s): 05.90.+m, 05.10.−a, 95.75.Wx, 05.45.−a

I. INTRODUCTION

In the study of a complicated physical system, data analysis of the time series of some physical quantities of the system is necessary in order to obtain the most important properties of the system. One of the fundamental problems in the study is how to measure the complexity of both local and global dynamical behaviors from the observed time series. There are two main approaches to quantifying the complexity of a distribution [1]. One approach has roots in dynamical systems theory and includes Lyapunov exponents and Kolmogorov-Sinai entropy [2]. The other stems from information theory and includes Shannon entropy [3] and algorithmic complexity [4]. For a review of the complexity measures, see the work of Shiner *et al.* [5]. For chaotic systems, various approaches to complexity measurements may be mutually complementary or may be contradictory. For example, the largest Lyapunov exponent (λ) of a chaotic attractor measures how fast two neighboring orbits diverge from one another. Thus, λ describes how fast information about an orbit is lost. It seems reasonable to conclude that the larger the λ, the more complex the system. On the other hand, the autocorrelation function of a time series gives the correlation between points. A shorter autocorrelation means that less information of the next point can be obtained from the current point. Therefore, it seems reasonable to say that the shorter the autocorrelation, the more complex the system. However, for the well-known chaotic Lorenz and Rössler systems, the largest Lyapunov exponent of the Lorenz system can be much larger than that of the Rössler system, but the autocorrelation of the Lorenz system is much longer than that of the Rössler system. So the Lyapunov exponent and the autocorrelation function approaches seem to contradict one another in this context.

A fractal-dimension-like correlation dimension [6] (d_c) may also indicate something about system complexity because it represents the density of the system's orbits in phase space. For a system in a stable equilibrium, $d_c=0$, in a stable periodic state, $d_c=1$, and in a chaotic state, $d_c>2$. It seems that the larger the d_c, the more complex the system. However, to determine which of the two systems with $d_c=2.1$ and 2.3, respectively, is more complex requires more information about the systems. However, the more information one has, the harder it may be to make a judgment. This suggests that *the complexity problem is itself complex*.

In this paper, we propose an alternative approach to measuring the predictability of the time series by considering the distribution of an observed time series in both the regular and tangent spaces. We introduce the idea of *forecast entropy* (F) to measure the predictability.

We introduce what we will call an "ideal random system" to act as the unpredictability standard. The maximum value of F is normalized to 1. At this maximum value, the system is totally unpredictable, having the maximum number of possible states not only in regular space but also in tangent spaces of any order. For any completely predictable system, such as periodic systems, $F=0$. For a real system or an attractor F lies in the interval $[0,1]$.

In Sec. II, our predictability problem will be framed. A commonly used procedure for measuring complexity will be analyzed to show its disadvantage in solving this predictability problem. To describe our procedure, an ideal random system will be introduced in Sec. III. A mathematical expression for F will be given in Sec. IV. Some simple cases of the predictability problem are enumerated in Sec. V to illustrate our approach. In Sec. VI we turn to some real chaotic systems. Our approach not only gives reasonable F's but also reveals important information about the system generating

*Present address: Department of Mathematics and Statistics, York University, 4700 Keele St., Toronto, Ontario, Canada M3J 1P3. Electronic address: wgyao@mathstat.yorku.ca
†Electronic address: essex@uwo.ca
‡Electronic address: pyu@pyu1.apmaths.uwo.ca
§Electronic address: mdavison@uwo.ca

the time series, such as the dynamical dimension of the system. A pseudorandom system and the noised Lorenz system will be investigated in Sec. VII. Finally, discussion and conclusions are given in Sec. VIII.

II. PREDICTABILITY PROBLEM

Our predictability problem is the following: Given a series of data, how difficult it is to predict the next point? Different techniques may be used for predictions. Frequently used techniques include neural networks [7], wavelets [8], return maps [9], and nonlinear dynamical forecasting [10]. The performance of these techniques may differ depending upon the given data. The task of studying the predictability problem is to show the general difficulty of predictions.

The predictability problem is often investigated by measuring the spatial complexity of the data. One begins by calculating the probability of finding a point in a specific neighborhood. Denoting by p_i the probability of finding the point in neighborhood i, one may define a *surprise* function

$$surprise = -\ln p_i, \tag{1}$$

which tells how much information is obtained by receiving p_i. The Boltzmann-Gibbs-Shannon entropy, here denoted by C, is just the expected value of this *surprise* over all states

$$C = -\sum_{i=1}^{n} p_i \ln p_i. \tag{2}$$

The problem with this technique is that it does not consider the spatial location of p_i at all. For instance, suppose that in four sequential positions 1, 2, 3, and 4, there are two probability distributions, given below:

space position: 1 2 3 4,

distribution I: {0.4, 0.4, 0.1, 0.1},

distribution II: {0.4, 0.1, 0.4, 0.1}.

Application of Eq. (2) to the two distributions yields the same answer, despite the very different spatial structure of the two probabilities, which could result in a large discrepancy in forecasting to which nearby position (1, 2, 3, or 4) we will step next. For distribution (I), one certainly will judge that the next data may be located *between* positions 1 and 2, while for distribution (II), one may say that the data may be located near position 1 or near position 3. Further, the resulting errors from the judgments may be different. Statistically, the error in the former distribution is less than that in the latter. Therefore, the complexity of these two distributions should be different.

In the above, we have proposed our predictability problem and analyzed an often-used complexity measure technique. This technique could not be safely used in the predictability problem because it does not consider the spatial structure of the probabilities. Our procedure is designed to overcome this. Before describing our procedure, we first introduce an ideal random system which plays an essential role in our procedure.

III. IDEAL RANDOM SYSTEM

Suppose there is a one-dimensional infinite time series $\{x(t_i)\}$ generated by a system, where $t_i = i\,\Delta t$, $i = 1, 2, \ldots$ and Δt is a constant. Denote the jth difference of $\{x(t_i)\}$ by

$$\{x^{(j)}(t_i)\} = [x^{(j-1)}(t_{i+1}) - x^{(j-1)}(t_i)]/\Delta t.$$

Here, $\{x^{(0)}(t_i)\} = \{x(t_i)\}$. The system is called an *ideal random system* if and only if for all $j = 0, 1, \ldots$, the series $\{x^{(j)}(t_i)\}$ is uniformly distributed in the regime $\{[a^{(j)}, b^{(j)}]\}$. In other words, $x^{(j)}(t_i)$ can be any value in $\{[a^{(j)}, b^{(j)}]\}$ in equal probability. Here, $a^{(j)}$ and $b^{(j)}$ are the smallest and largest values of $\{x^{(j)}(t_i)\}$.

For a finite series with equal time intervals, one may only consider the distributions up to some maximum order of finite differences. In this paper, we limit ourselves to considering the distribution of zeroth- and first-degree differences. The zeroth-degree difference corresponds to the usual space $\mathbf{R}^d$, where d is the embedding dimension used in reconstructing the attractor from the time series $\{x(t_i)\}$. The first-degree difference corresponds to the tangent space of $\mathbf{R}^d$.

We will use this as the standard of unpredictability and to normalize F of real observed time series. Of course, such an ideal random system cannot exist—it is impossible for $\{x(t_i)\}$ to be iid (independent and identically distributed) and uniform and for the differences to also be iid and uniform. But no other system will have as big a forecast entropy as this hypothetical system, so it obtains a normalization constant.

In the following, we only consider the zeroth-degree situation, since a jth-degree case can follow the same procedure based on the distribution of $\{x^{(j)}(t_i)\}$.

IV. CALCULATION OF FORECAST ENTROPY

As we discussed in Sec. II, a more meticulous version of the predictability problem should consider the spatial organization of the probabilities. The method based on Eq. (2) does not do this. One way of incorporating spatial information is to separate the group into subgroups or observe the distribution in different scales. For example, for distribution (I), if we decompose the distribution into two subgroups $\{p_1, p_2\}$ and $\{p_3, p_4\}$, the difference of the probabilities between these two subgroups appears, indicating spatial information. A more rigorous analysis of this basic idea leads to our approach described below.

A. F in the one-dimensional case ($d=1$)

Suppose there is a segment of the one-dimensional time series $\{x(t_i)\}$, $i = 1, 2, \ldots, n$. A probability distribution can be obtained based on the values of these data points. For simplicity and without loss of generality, suppose $n = 2^m$ (see more discussion later), where m is a positive integer. If the time series is generated by the ideal random system, the distribution is uniform in $x \in [a, b]$, where a and b are the smallest and largest values of $\{x(t_i)\}$. Equally partition the interval $[a, b]$ into n subintervals so that for the time series from the ideal random system we expect a single point in each of the subintervals as shown in Fig. 1. If the time series

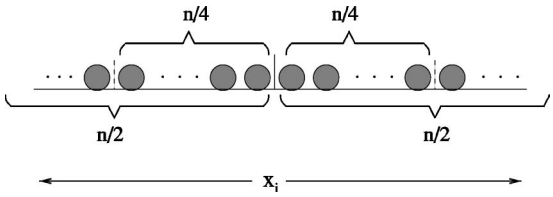

FIG. 1. Distribution of a one-dimensional ideal random system.

is not from the ideal random system, the distribution will not be uniform. Thus, we may expect many points in one subinterval and none in other subintervals. We describe the procedure as follows.

Step 1. Cut the line at the center of $[a,b]$, and count the number of points on the left and right half lines, respectively; then, denote the numbers by n_L and n_R, respectively. The location (left or right half of the line) of the next point may be predicted based on the fractions $p_L=n_L/(n_L+n_R)$ and $p_R=n_R/(n_L+n_R)$, respectively, which we interpret as probabilities stemming from some process. At this scale, the difficulty of prediction may be measured by the Boltzmann-Gibbs-Shannon entropy

$$S_1 = -p_L \ln p_L - p_R \ln p_R. \tag{3}$$

Here, the subscript 1 of S indicates the first level. For the ideal random system, $n_L=n_R=n/2$. Therefore, $p_L=p_R=1/2$ and then $S_1=\ln 2$.

Step 2. Again cut the left and right half lines at their centers, and count the number of points on the new intervals, as shown in Fig. 1. Denote the resulting numbers by n_{LL}, n_{LR}, n_{RL}, and n_{RR}.

To predict whether the next point will be located in the regime *LL* or *LR*, one may use the fractions, interpreting them as conditional probabilities, $p_{LL}=n_{LL}/(n_{LL}+n_{LR})$ and $p_{LR}=n_{LR}/(n_{LL}+n_{LR})$. The difficulty of the prediction can again be measured by

$$S_L = -p_{LL} \ln p_{LL} - p_{LR} \ln p_{LR}. \tag{4}$$

Similarly, in the regime including *RL* and *RR*, the probabilities are $p_{RL}=n_{RL}/(n_{RL}+n_{RR})$ and $p_{RR}=n_{RR}/(n_{RL}+n_{RR})$, respectively. The Boltzmann-Gibbs-Shannon entropy in this regime is

$$S_R = -p_{RL} \ln p_{RL} - p_{RR} \ln p_{RR}. \tag{5}$$

For the ideal random system, $p_{LL}=p_{LR}=p_{RL}=p_{RR}=1/2$ and $S_L=S_R=\ln 2$. We now add to get S_2, the total Boltzmann-Gibbs-Shannon entropy of level 2. $S_2=S_L+S_R=2\ln 2$.

Step 3. Repeat the process on the shorter and shorter intervals to the *m*th level. For the ideal random system, we expect only one datum in each of the subgroups. Obviously, at the *k*th level, the total $S_k=2^{k-1}\ln 2 (k\leqslant m)$.

Define F as

$$F = \sum_{k=1}^{m} \alpha_k S_k, \tag{6}$$

where $\alpha_k, k=1,\ldots,m$, are parameters dependent only on the number of points, n. Therefore, F is the summation of weighted S's up to the *m*th level. For the ideal random system, $F=F_{ideal}$, given by

$$F_{ideal} = \ln 2 \sum_{k=1}^{m} \alpha_k 2^{k-1}. \tag{7}$$

To determine the parameters α_k's, we consider two distributions with the same number of points, $n=2^m (m\geqslant 2)$. One is generated by the ideal random system, with the F of that distribution calculated by Eq. (7). Suppose another system with the same number of points, n, has the pattern $\{2,0,2,0,\ldots,2,0\}$where the digit (0 or 2) means the number of points in the position. This distribution is special because the difference between the ideal distribution and this case appears only at the last step in the decision tree, and it doubles the probabilities of the ideal case. We use a factor of 2 because we are working with powers of 2. The expression for the F of this distribution, F', is the same as Eq. (7), except that $S_m=0$:

$$F' = \ln 2 \sum_{k=1}^{m-1} \alpha_k 2^{k-1}. \tag{8}$$

F' shows a special property, in which one term is lost compared with Eq. (7). As indicated, to arrive at a specific point in the first distribution, the fraction is $p_{ideal}=1/n$, while in the second distribution, $p'=2/n$. The difficulty of the prediction for the first distribution is 2 times that of the second one. One should expect this to be reflected in that F_{ideal} of the first distribution is also double the second one:

$$F_{ideal} = 2F'. \tag{9}$$

From Eqs. (7)–(9),

$$\alpha_m = \frac{1}{2^{m-1}} \sum_{k=1}^{m-1} \alpha_k 2^{k-1}. \tag{10}$$

Thus we have obtained the relation of α_m to $\alpha_1,\alpha_2,\ldots,\alpha_{m-1}$. Similarly, we can find the relation of α_i to $\alpha_1,\alpha_2,\ldots,\alpha_{i-1}$, where $i\leqslant m$, by considering two distributions with the same number of points, $n=2^m$, $m\geqslant 2$. One is the same as that of the ideal random system at the *i*th level—namely, $\{2^{m-i},2^{m-i},\ldots,2^{m-i}\}$. The other is $\{2^{m-i+1},0,2^{m-i+1},0,\ldots,2^{m-i+1},0\}$. For the former distribution, $F=F_1$:

$$F_1 = \ln 2 \sum_{k=1}^{i} \alpha_k 2^{k-1}. \tag{11}$$

For the latter distribution, $F=F_2$:

$$F_2 = \ln 2 \sum_{k=1}^{i-1} \alpha_k 2^{k-1}. \tag{12}$$

Because $F_1 = 2F_2$, from Eqs. (11) and (12), one has

$$\alpha_i = \frac{1}{2^{i-1}} \sum_{k=1}^{i-1} \alpha_k 2^{k-1}. \tag{13}$$

When $i=2$, it follows from Eq. (13) that

$$\alpha_2 = \alpha_1/2. \tag{14}$$

When $i=3$, using Eqs. (13) and (14), one obtains $\alpha_3 = \alpha_1/2$. Repeating the process to deduce all the parameters, finally one has

$$\alpha_k = \alpha_1/2, \quad k = 2, \ldots, m. \tag{15}$$

One may use other distributions with the same number of points to obtain the relation of the parameters. For example, one distribution is $\{2^{m-i}, 2^{m-i}, \ldots, 2^{m-i}\}$ and the other is $\{2^{m-i+2}, 0, 0, 0, 2^{m-i+1}, 0, 0, 0, \ldots, 2^{m-i+1}, 0, 0, 0\}$. In this case, for the latter distribution, $F = F_3$:

$$F_3 = \ln 2 \sum_{k=1}^{i-2} \alpha_k 2^{k-1} \tag{16}$$

and

$$F_1 = 4F_3. \tag{17}$$

We then lose two terms in the expansion (16) with respect to Eq. (11). Finally, we will have two free parameters. (In practice, only one is free because the other will be determined by normalization.) The uniqueness of the solution requires the α's to be independent of the distribution chosen. The values of the parameters are chosen to satisfy Eq. (15).

For the ideal random system, substituting the parameters in Eq. (15) into Eq. (7) yields

$$F_{ideal} = \alpha_1 \ln 2 \left(1 + \frac{1}{2} \sum_{k=2}^{m} 2^k \right) = \alpha_1 2^m \ln 2. \tag{18}$$

If we define $F_{ideal} = 1$ for the ideal random system, we then obtain

$$\alpha_1 = \frac{1}{2^m \ln 2}. \tag{19}$$

For any distribution with the number of points, $n = 2^m \geqslant 4$, from Eq. (6), we have

$$F = \frac{1}{2^{m+1} \ln 2} \left(2S_1 + \sum_{k=2}^{m} S_k \right). \tag{20}$$

$F \in [0,1]$ because a distribution under consideration is always in comparison with the distribution of the ideal random system which has *the same number of points*. The magnitude of F is thus normalized in terms of something that is unbounded. Hence, this does not mean that absolute entropy is bounded.

If $n = 3^m$, we may cut the line (interval) into three equal intervals each time. Suppose the probabilities of finding a point in the three intervals are p_1, p_2, and p_3. Then $S = -\sum_{k=1}^{3} p_k \ln p_k$. For the ideal random system, $p_k = 1/3$, and $S = \ln 3$. Similarly, up to the mth level, using a similar process to the case $n = 2^m$, we obtain

$$F = \sum_{k=1}^{m} \alpha_k S_k = \alpha_1 \left(S_1 + \frac{2}{3} \sum_{k=2}^{m} S_k \right). \tag{21}$$

For the ideal random system,

$$F_{ideal} = \alpha_1 3^{m-1} \ln 3. \tag{22}$$

Taking $\alpha_1 = (3^{m-1} \ln 3)^{-1}$ results in $F_{ideal} = 1$ for the ideal random system. Substituting α_1 into Eq. (21) gives the formula for calculating F for any distribution.

The above procedure can be generalized to the case $n = l^m$, where l is the smallest positive integer that satisfies $l = n^{m^{-1}}$ for any possible positive integer m. For example, if $n = 16 = 2^4 = 4^2$, l should be taken as 2, not 4. In this general case, for the ideal random system,

$$F_{ideal} = \alpha_1 \, l^{m-1} \ln l. \tag{23}$$

Taking $\alpha_1 = 1/(l^{m-1} \ln l)$ yields $F_{ideal} = 1$ for the ideal random system. For any distribution with $n = l^m$, we have

$$F = \sum_{k=1}^{m} \alpha_k S_k = \frac{1}{l^{m-1} \ln l} \left(S_1 + \frac{l-1}{l} \sum_{k=2}^{m} S_k \right). \tag{24}$$

Obviously, when $m=1$, F reduces to C in Eq. (2) except a coefficient $(\ln n)^{-1}$, not a good measure as we have stated in Sec. II. Therefore, when n is such that m must be 1 if one follows the formula $n = l^m$, one should use $n' = l'^{m'}$ instead, where $n' > n$ and $m' > 1$. Then these n points are distributed on the line by a scale n'/n. Finally, F can be calculated according to the distribution.

B. F in the multiple-dimensional case ($d \geqslant 2$)

For simplicity, we may first calculate F of the ideal random system along each coordinate (see more discussion later), then define the system's F to be the average of these forecast entropies. For example, for the ideal random system, the distribution is uniform along each coordinate, and then the forecast entropy is 1 on the distribution, which results in the average of the forecast entropies being 1.

C. F of a system up to the jth difference

Denote F_i as the forecast entropy of the system at the ith difference, where $i \in [0, j]$; then, the F of the system up to the jth difference is defined as

$$F = \frac{1}{j+1} \sum_{i=0}^{j} F_i. \tag{25}$$

For the ideal random system, $F = 1$ because $F_i = 1$ for all i.

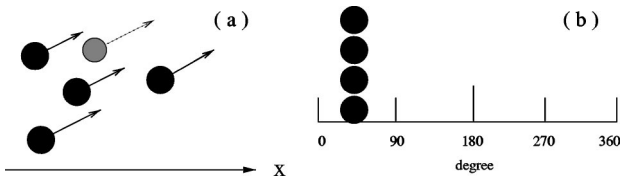

FIG. 2. Distribution of four balls on a plane in case 1.

V. SOME BASIC CASES

In some forecasting techniques such as Short's "nonlinear dynamical forecasting" [10], the prediction is based on the local manifold centered at the predicting point in regular space. The manifold can be determined by the first-degree differences of the points. The difficulty of predicting the value of the next point depends on the distribution of the points in the first-degree difference space—i.e., the tangent space. In the cracking technique proposed by Pérez and Cerdeira [9], however, the difficulty relies on the distribution of points in regular space only. The first-degree differences of the points are zero.

Next, we investigate F for some basic cases. For simplicity we consider a local region where there are only four points on a plane. As discussed above, to obtain F of the distribution on a plane, one might first project the distribution onto two rectangular coordinates to get two distributions on the two lines, respectively, then calculate F's of the distributions separately, and finally average F's as the expected F of the planar distribution.

If one considers instead the distribution in tangent space, one may use the distribution of the *angles* spanned between the tangent directions of the points in regular space in terms of some reference direction (such as the x axis), instead of the tangents themselves. The angles are then mapped onto an interval scaled between 0° and 360°. The scale is determined by the number of points. If there are n points, the line is scaled to n equal intervals (so that for the ideal random case, there is one point in each interval on average). After that, one has to adjust the direction of the coordinates on the plane so that most angles are distributed between 0° and 180°. After the adjustment, the coordinate on a plane is unique. One need not consider the angular distribution of other coordinates.

Case 1. Suppose the four points move with the same velocity and in the same direction as depicted in Fig. 2(a). Obviously, if one predicts x_{i+1} from x_i based on this manifold, x_{i+1} can be obtained without error, as the dashed line in Fig. 2(a) shows. F is calculated based on the distribution of the angles spanned between these directions (actually, only one direction in this case) and the x coordinate, as shown in Fig. 2(b). All the angles have the same value. We then have $F=\alpha_1(-1\ln 1-0\ln 0)=0$, where α_1 is a constant determined by calculating F of an ideal random system with the same number of points. We have taken $0\ln 0=0\times(\lim_{x\to 0}x\ln x=0)$ in calculating F in this case.

The average distribution of four points from an ideal random system is shown in Fig. 3(a). The arrows indicate the directions in which the points move. The distribution of the angles mapped onto a line is shown in Fig. 3(b). From Eq. (7), the forecast entropy of the ideal random system of the distribution is

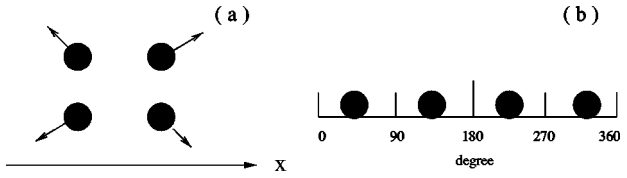

FIG. 3. Distribution of four uniformly distributed balls on a plane.

$$F_{ideal}=4\alpha_1\ln 2. \tag{26}$$

Letting $F_{ideal}=1$, one has

$$\alpha_1=\frac{1}{4\ln 2}. \tag{27}$$

From this case, we may answer the question: Why do we need to consider the distribution in tangent space? Not only can a random distribution in range of regular space be quite uncomplicated in the tangent space, it is the tangent directions that hamper forecasting events most when one has a cloud of data, but no idea which one to go to next. One can use such prediction techniques as nonlinear dynamical forecasting to forecast the next data without difficulty. Therefore, a careful measure of the predictability needs to consider this dynamical aspect of signal data.

Case 2. Suppose the four points are not uniformly distributed [Fig. 4(a)]. The angular distribution is shown in Fig. 4(b). Based on the distribution, we have

$$F=-\alpha_1\left[2\left(\frac{1}{4}\ln\frac{1}{4}+\frac{3}{4}\ln\frac{3}{4}\right)+\frac{1}{3}\ln\frac{1}{3}+\frac{2}{3}\ln\frac{2}{3}\right]=0.6352, \tag{28}$$

where Eq. (27) has been used.

Case 3. Consider the distribution of one point in case 2. If it were changed slightly [see Fig. 4(a)] according to the dashed line, the F of the distribution would still be 0.6352. But it makes sense that F in this case ought to be different from that in case 2. Clearly, to treat this we need to determine F using higher-degree difference spaces, in order to obtain sufficient differences in the distributions.

We have studied three kinds of distributions. Any other distributions can be studied similarly. In a real chaotic attractor, the case often encountered is that the number of points changes in different local regions. One may have to partition the space in different ways. In some prediction techniques the points are used to construct a map. The number of points

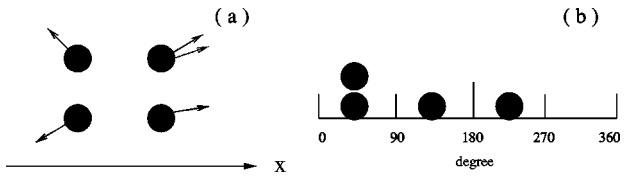

FIG. 4. Distribution of four balls on a plane in case 2.

is determined by the number of unknown parameters in the map. If the number of parameters is 25, at least 25 neighboring points are needed. In this case, we may search the neighboring points centered on the reference point in regular space to obtain 2^5 such points. (Of course, one may also choose $l=5$ and $m=2$ for the 25-point case.)

VI. *F* OF SOME CHAOTIC SYSTEMS

Consider a time series $s(t)$ coming from a chaotic system. When the forecast entropy F is based on the attractor reconstructed from $s(t)$, three factors may affect the measure.

The first factor is the value of the delay τ. An improper value of τ will cause the reconstructed attractor to be more tangled than it needs to be.

Second, the dimension will affect the local structure and therefore affect the set of the neighbors, while F is calculated by averaging the local F's in local regions. For example, when the attractor is reconstructed to be two dimensional, the coordinate of the ith point on the plane is $(s(t_i), s(t_i+\tau))$. Whether the jth point is a neighbor of the ith point in regular space is determined by the distance—e.g., $\sqrt{[s(t_i)-s(t_j)]^2+[s(t_i+\tau)-s(t_j+\tau)]^2}$. If one chooses three dimensions, the coordinate of the ith point is $(s(t_i), s(t_i+\tau)s(t_i+2\tau))$, and the distance between points i and j is changed.

Third, the size of the neighborhood or the number of neighbors has an effect. Obviously, if the size $\epsilon \to 0$ or there is no neighbor of the reference point, then $F=0$. On the other hand, if $\epsilon \to \infty$ or all the points in the attractor are neighbors of the predicting point, then F may be quite close to 1 because there may be a large number of tangent directions.

Clearly, to compare F's of different time series—i.e., of different chaotic attractors—the number of points in each series must be the same. Further, one needs to consider the number of orbits consisting of N points. Every system has its own natural time scale. When integrating with a single step size some systems oscillate very quickly, while others oscillate slowly. Thus, use of the same sampling rate to obtain the time series across different systems may cause problems in calculating F. For example, for two $N=1024$ time series, the Lorenz system oscillates quickly and represents 20 orbits, while the Rössler system oscillates slowly and represents just 3 orbits. The information provided by the first series may be enough to describe the dynamical behavior of the system, but that provided by the second one may be insufficient. Therefore, one has to use different sampling rates for different systems so that the number of orbits consisting of N points is roughly the same. Of the systems considered in this paper, the Lorenz system oscillates the fastest. We use a sampling rate of 0.01 time unit for the Lorenz system and larger rates of the other systems. For all time series in this paper, $N=2^{14}$.

To calculate F of a multiple-dimensional attractor, we first project the attractor to each plane, then calculate F of the distributions on these planes and use the average as the attractor's F. For example, for a d-dimensional attractor with coordinates $(x_1, x_2, \ldots, x_d)$, where $x_i = x[t+(i-1)\tau]$, $i=1,2,\ldots,d$, we may consider as many as C_d^2 planes. In practice, it is enough to consider a smaller number of planes to obtain F. When $d \ll N\Delta t/\tau$, where Δt is the sampling rate, the phase diagram projected from the reconstructed attractor on the (x_i, x_{i+1}) plane is the same as that on the (x_1, x_2) plane, where $i=2,3,\ldots, d-1$. Including the plane itself, there are in total $d-1$ similar diagrams as on the (x_1, x_2) plane. Similarly, there are $d-i$ diagrams similar to (x_1, x_{i+1}) plane including the plane itself, where $i=1,2,\ldots, d-1$. Denoting by F^i the forecast entropy of the distribution on the (x_1, x_{i+1}) plane, we obtain the forecast entropy of the d-dimensional attractor by averaging the entropies of the distributions on each plane:

$$F_d = \frac{2}{d(d-1)} \sum_{i=1}^{d-1} (d-i) F^i. \tag{29}$$

To calculate F^i, we first find n nearest neighbors of point j in regular space, then calculate forecast entropy of the local distribution of these neighbors in tangent space—that is, according to their angles as discussed in Sec. V. Let point j experience each point of the time series. The average of these local F's is taken as F^i of the distribution in the plane.

To choose n, the number of neighbors, we consider two cases. First, we choose $n=2^{d+2}$ where d is the dimension. We want to investigate for which value of d F reaches its lower bound so that one can use a d-dimensional reconstructed attractor for optimal prediction. Second, $n=16$. We investigate the complexity of local structures of the attractor.

A. Chaotic Lorenz system

The chaotic Lorenz system considered here was first introduced in [11] and is

$$\dot{x} = 10(y-x),$$

$$\dot{y} = 25\,x - y - x\,z,$$

$$\dot{z} = x\,y - 2.667\,z. \tag{30}$$

To obtain an observed time series $s(t)=x(t)$ from the system, we first use the fourth-order Runge-Kutta method to integrate it. The step size is 0.01. To reconstruct an attractor from the time series, one needs a delay τ, which may be determined by calculating the autocorrelation function [12,13], mutual information [14,15], and mutual redundancy [16].

However, for the Lorenz system, these techniques do not give a satisfactory τ. Instead, one often finds that $\tau=0.1$ is best to reconstruct the attractor, because for this value the attractor looks spread out like the original, projected on the (x,y) plane. The time series $s(t)$ is shown in Fig. 5(a). The reconstructed attractor when $\tau=0.1$ is depicted in Fig. 5(b).

Is this τ also best from the viewpoint of predictability? This question may be answered by calculating F, which is calculated based on the distribution in the tangent space of the series.

Case 1. The number of neighbors $n=2^{d+2}$. In this case, we focus on the predictability based on the reconstructed attractors with different embedding dimensions. Figure 6(a) dis-

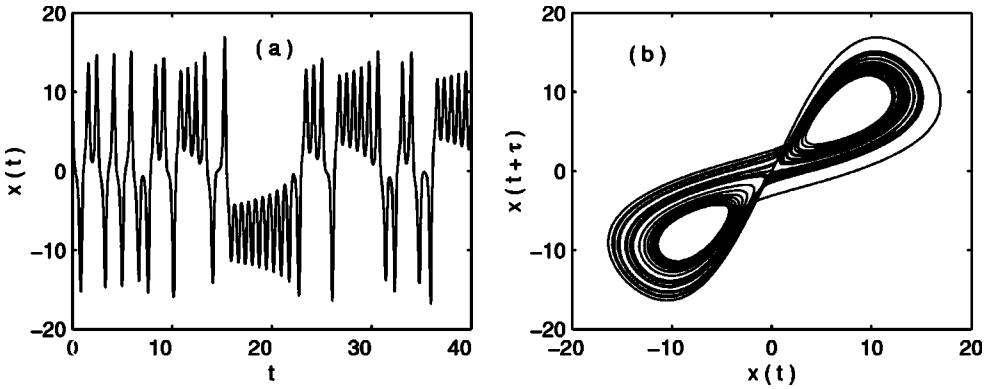

FIG. 5. (a) A piece of time series $x(t)$ from the Lorenz system. (b) The reconstructed attractor from the series when delay $\tau=0.1$.

plays F as a function of τ when $d=2,3,4$, and 5, respectively. Denote F_d as F when the dimension is d. The following information may be obtained from the figure [17].

(i) F_2 oscillates around F_3, and $F_3<F_4<F_5$. This indicates that a three-dimensional embedding space is enough to reconstruct the attractor for prediction purposes. Surprisingly, when $\tau<0.3$, $F_2<F_3$. Therefore, it may be better to do a prediction in a two-dimensional attractor if one uses the information of the neighbors in the tangent space. In practice, however, one may add some conditions on the choice of neighbors to predict better [10]. In that case, it may be better to use a three-dimensional attractor. As we will see in case 2, the local structure of the three-dimensional Lorenz attractor is simpler than that of the two-dimensional one.

(ii) The minimum of F_2 is 0.0082 at $\tau=0.18$, and the minimum of F_3 is 0.022 at $\tau=0.1$. Therefore, in order to obtain the best forecasting result, one should adjust the value of the delay as d changes.

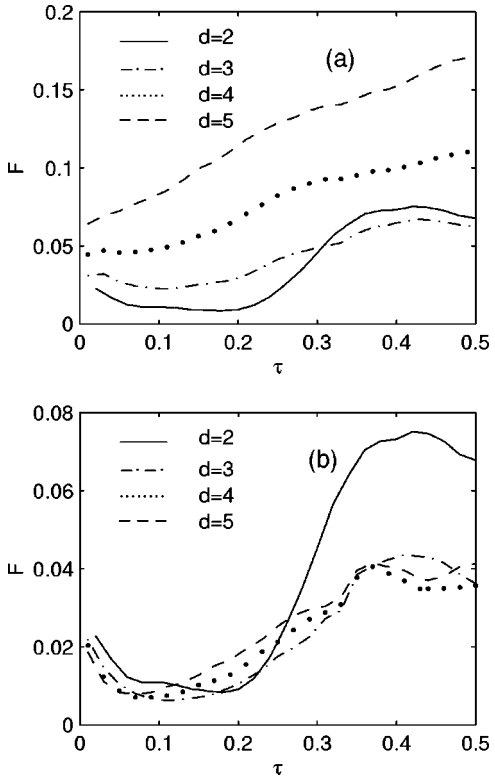

FIG. 6. F vs τ of the Lorenz system (31) when $d=2,3,4$, and 5, respectively. (a) $n=2^{d+2}$, (b) $n=16$.

Case 2. The number of neighbors is fixed at 16. We want to investigate the characteristics of the local structures of reconstructed attractors with the same number of neighbors. In this case, as shown in Fig. 6(b), we may obtain the following conclusions.

(i) Except around $\tau=0.2$, $F_3<F_2$, which indicates that the local structure of the three-dimensional attractor is simpler than that of the two-dimensional one.

(ii) The minimum of $F=0.0062$ appears at $d=3$ and $\tau=0.1$. This indicates that the delay $\tau=0.1$ is the best candidate to reconstruct the attractor when one uses delayed coordinates. Thus, for the time series from the chaotic Lorenz system, F in tangent space has solved the problem of the value of delay, while the autocorrelation function in regular space cannot.

(iii) F_d does not decrease further as d increases after $d>2$.

It may indicate that one cannot simplify the local structures by increasing the dimension of an attractor after $d>2$. In other words, it is enough to use three dimensions to describe the system. The number of dimensions is exactly equal to the dynamical dimension of the Lorenz system. This result is reasonable: A nonlinear coupled system cannot be decoupled. The observed time series of any variable contains the information of the others and the whole system.

Therefore, by calculating F of the distribution of an observed time series in tangent space, we have obtained (a) the dynamical dimension of the system that produces the time series, (b) the value of τ to best reconstruct an attractor, and (c) the embedding dimension for optimal predictions.

The main purpose of statistics is to capture useful information from an observed time series [18]. When delayed coordinates are employed to reconstruct an attractor, the value of the delay and the embedding dimension must be determined so that the reconstructed attractor shares some properties with the original one, such as no correlation between its coordinates, no crossing of its orbits, and simplicity of its local structure. Kennel, Brown, and Abarbanel [19] determined d by studying the "noise" behavior of the neighbors about a reference point. They studied an observed finite time series from the Lorenz system and found that, in a prop-

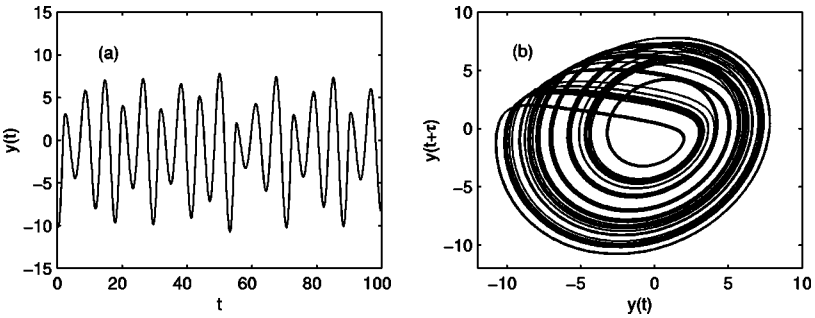

FIG. 7. (a) A piece of time series $y(t)$ from the Rössler system. (b) The reconstructed attractor from the series when delay $\tau = 1.45$ determined by the autocorrelation function.

erly reconstructed attractor, the "noise" is minimum. Their work suggested that $d=4$ for the series. Cenys and Pyragas [20], however, suggested that $d=1+[\text{integer part of } (d_A)]$ where d_A is the fractal dimension of the attractor. They used a similar technique to that of Kennel *et al.* [19], but considered the behavior of the neighborhood, not neighbors. For an observed time series from the Lorenz system or some other three-dimensional chaotic system, the result of Cenys and Pyragas [20] is perfect because $d_A \in (2,3)$ for these systems, and we have $d=3$. But for some higher-dimensional systems, the fractal dimension of their chaotic attractors may be still low—for example, $d_A \in (2,3)$. In this case, the result of Cenys and Pyragas [20] is unsuitable. There are many other arguments in the literature for determining these values [21,22]. Our work suggests that when more information can be used from the time series—i.e., the distribution of the series in tangent space and properly measure this distribution—these values may be well determined.

In the remaining part of this section, we use three other chaotic systems whose dynamical dimensions are 3 or 4 to show that the advantages of F hold not only in the Lorenz system, but in other chaotic systems as well.

B. Chaotic Rössler system

The chaotic Rössler system, introduced in [23], is

$$\dot{x} = -y - z,$$

$$\dot{y} = x + 0.2\,y,$$

$$\dot{z} = 0.2 + z\,(x - 5.7). \tag{31}$$

The time series $s(t)=y(t)$ and the reconstructed attractor are depicted in Fig. 7.

To calculate F, the series is sampled using $\Delta t=0.06$ so that there are approximately the same number of orbits in the series ($N=2^{14}$) as that for the Lorenz system case. The result is shown in Fig. 8.

C. Four-dimensional chaotic system

A very entangled chaotic system described in [24] is

$$\dot{x} = y + z,$$

$$\dot{y} = \beta_1\,y - x\,w + \gamma\,w,$$

$$\dot{z} = (1-\alpha)x - (1+\beta_2)z + xy^2,$$

$$\dot{w} = y + z - \beta_2\,w. \tag{32}$$

When $\alpha=38.2$, $\beta_1=\beta_2=0.2$, and $\gamma=-0.54$, the system is in the chaotic state. Figure 9 shows the chaotic attractor in two-dimensional phase space. Figures 10(a) and 10(b) display a time series $s(t)=y(t)$ and the reconstructed attractor when $\tau=1.5$ calculated by using the autocorrelation function. The sampling rate is 0.04.

By comparing Fig. 10(b) with Fig. 9, it is seen that the reconstructed attractor is not like the original and seems quite entangled. In fact, it is very difficult to find a τ which results in plausible reconstruction of this attractor. We will give a possible explanation for this behavior using the concept of forecast entropy.

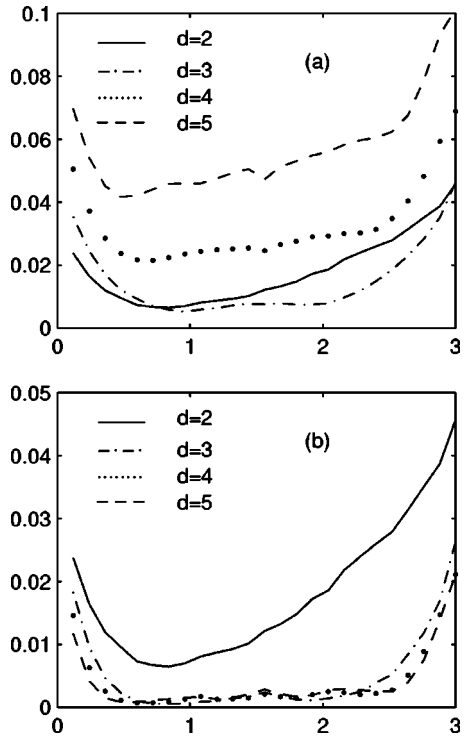

FIG. 8. F vs τ of the Rössler system (32) when $d=2,3,4$, and 5, respectively. (a) $n=2^{d+2}$. Following results are observed. (i) The minimum $F=0.0052$ appears in $d=3$ at $\tau=1.0$ and F_3 changes very slowly when τ is between 1.0 and 2.0. (ii) $d=3$ is the most suitable dimension for predictions. (iii) The minimum F is smaller than that of the Lorenz system (31), which indicates that the Lorenz system is more complicated than the Rössler system from the predictability viewpoint. (b) $n=16$. $F_2 > F_3$, and F_3, F_4, and F_5 are almost the same. Thus $d=3$ is the dynamical dimension of the Rössler system.

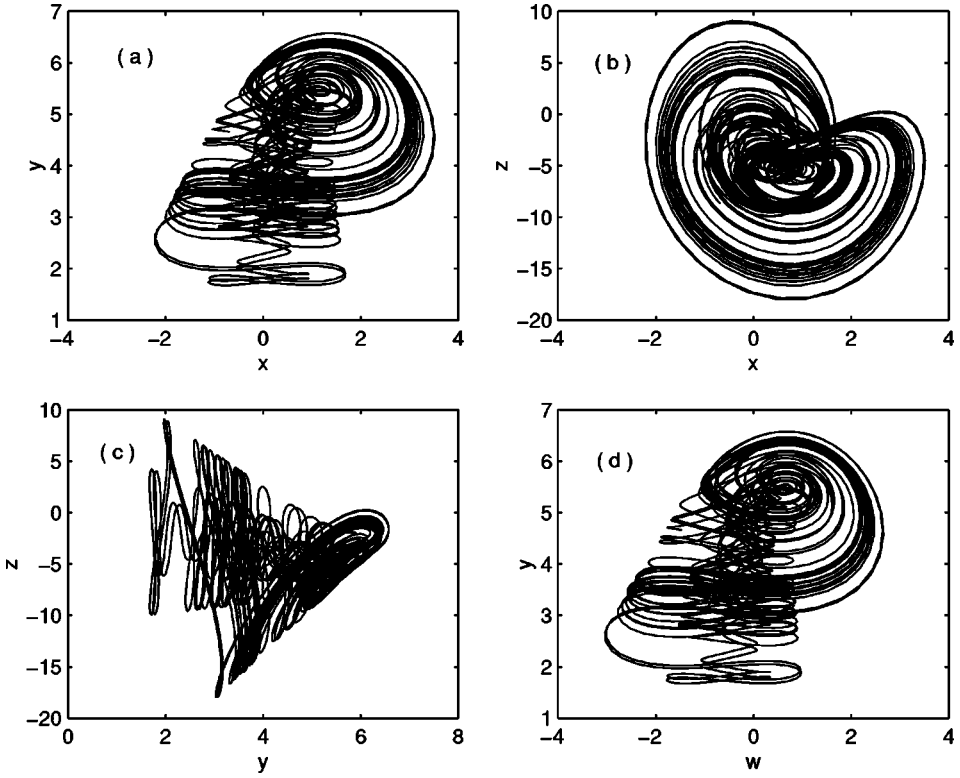

FIG. 9. Some phase portraits of the chaotic attractor of system (32).

Case 1. $n=2^{d+2}$. F in this case is shown in Fig. 11(a). From the figure, the following is observed: (i) The minimum $F<0.07$ appears for $d=2$ as $\tau\to 0$. (ii) $d=3$ is the most suitable dimension to do predictions if $\tau>0.15$. (iii) The minimum of F_3 is 0.11 when $\tau=0.32$. The minimum is much larger than those of the Lorenz and Rössler systems. Therefore, it is much more difficult to do predictions. The result agrees with what we obtained in Ref. [24]. (iv) Unlike the Lorenz and Rössler systems, here F_2 does not arrive at its minimum monotonically as τ increases from 0.

Case 2. $n=16$. F in this case is shown in Fig. 11(b). Obviously, $F_2>F_3>F_4$, while F_4,F_5, and F_6 are almost the same. $d=4$ is the dynamical dimension of the system.

D. Another four-dimensional chaotic system

The system is described by [25]

$$\ddot{x} = -(\alpha + y^2)x + y,$$

$$\ddot{y} = -(\beta + x^2)y + x. \tag{33}$$

When $\alpha=0.1,\beta=0.101$ and initial condition $(x(0),\dot{x}_0(0),y(0),\dot{y}_0(0))=(0.1,0.1,-0.1,-0.1)$, the four-dimensional system is highly chaotic. The attractor is shown in Fig. 12. It is observed that the local structure of the attractor is more tangled than those of the Lorenz and Rössler attractors.

Figures 13(a) and 13(b) display a time series $s(t)=x(t)$ and the reconstructed attractor when $\tau=6.0$. The result from the autocorrelation function is $\tau=150.0$, which is bad from the viewpoint of reconstructing an attractor. In fact, there is no reasonable τ to reconstruct an attractor like the original. We have used $\tau=6.0$ just because the reconstructed attractor is *a little* like the original. The sampling rate is 0.1.

Case 1. $n=2^{d+2}$. F in this case is shown in Fig. 14(a). From the figure, the following is seen: (i) The minimum $F<0.08$ appears at $d=2$ as $\tau\to 0$. (ii) A three-dimensional (3D) attractor is enough for optimal predictions. (iii) The

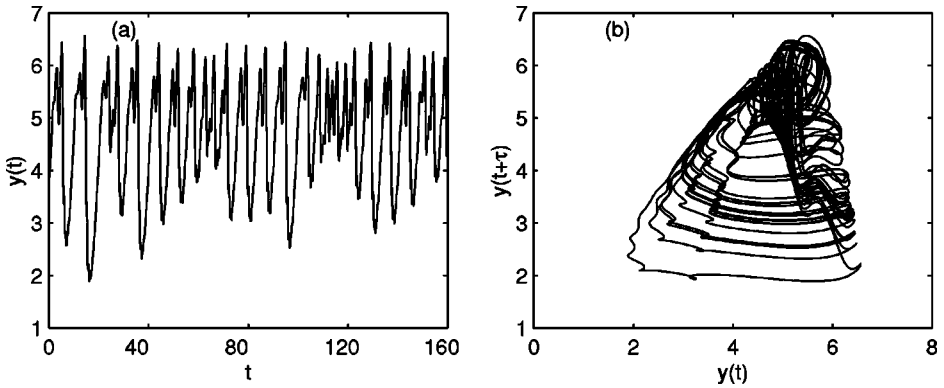

FIG. 10. (a) A piece of time series $y(t)$ from system (32). (b) The reconstructed attractor from the series when delay $\tau=1.5$.

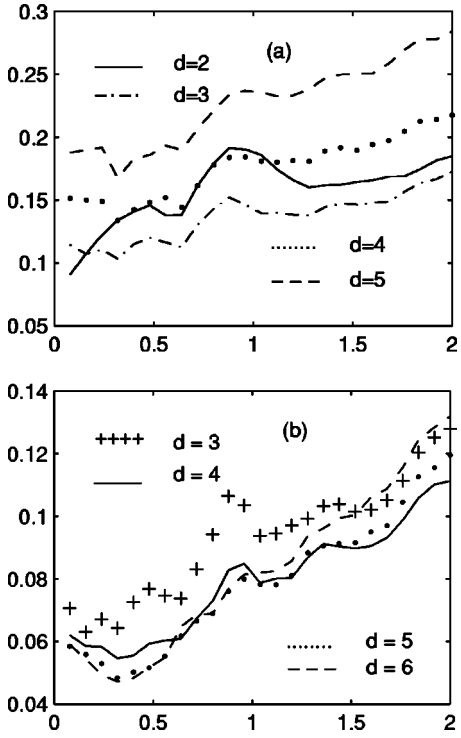

FIG. 11. F vs τ of system (32). (a) $n=2^{d+2}$, (b) $n=16$.

minimum value of $F_3<0.1$ when $\tau\to 0$. (iv) F_3 of this system is much larger than the corresponding forecast entropy (F_3) of the other systems studied above. Therefore, it is very difficult to do predictions based on the reconstructed attractor. (v) Another distinctive aspect of F of this system in terms of the Lorenz and Rössler systems studied above is that, like F_2 of the former 4D system, the minimal value of F_d exists as $\tau\to 0$. This characteristic, as we will see in the next section, is similar to that of a pseudorandom system. It may indicate that it is difficult to reconstruct an attractor similar to the original based on the time series from system (33).

Case 2. $n=16$. F in this case is shown in Fig. 13(b). Obviously, $F_2>F_3>F_4$, while F_4, F_5, and F_6 are almost the same. $d=4$ is the dynamical dimension of the system.

From the above examples, we may conclude that our forecast entropy technique is a convincing measure of the difficulty of prediction based on an observed time series. The technique may also capture some important information such as the dynamical dimension of the system producing the time series when the system is deterministically chaotic. In the next section, we shall investigate a pseudorandom number generator and a noised time series from the Lorenz system.

VII. *F* OF NONDETERMINISTIC SEQUENCES

A system-supplied functions *rand*$(\cdot)$ is almost always a *linear congruential generator*, which generates a series of integers I_j each between 0 and $m-1$ by the recurrence relation [26]

$$I_{j+1} = aI_j + c(\mathrm{mod}\ m). \tag{34}$$

Here m is called the modulus, which determines the maximal length of the pseudorandom number sequence. a and c are positive integers called the multiplier and increment, respectively. A "minimal standard" generator proposed by Park and Miller [27] is based on the following choices:

$$a = 7^5 = 16807, \quad m = 2^{31} - 1 = 2147483647, \quad c = 0. \tag{35}$$

In our calculation, we use the series of the pseudorandom numbers distributed in $[0,1]$, which is then given by $s_j = I_j/m$.

We investigate the second case, where the characteristics of the local structure of the reconstructed attractors are considered (if we still call them *attractors*) with the same number of neighbors. Let us take $n=16$, $N=1024$, and the sampling rate of one unit—i.e., sampling the output from the generator continuously. F in this case is shown in Fig. 15(a). The figure shows that (i) F monotonically increases with d when $d=2, 3, 4, 5$ and (ii) F increases with τ until $\tau=3$ units and remains unchanged when τ increases further. The shape of F in the pseudorandom case is clearly different from that of the deterministic chaotic cases.

From the forecast entropy point of view, the pseudorandom system is not "ideal." The distribution of s_j in regular space is quite uniform, but not in tangent space, as displayed in Fig. 16 where $s'_j = s_{j+1} - s_j$ for any integer j.

To investigate a higher-dimensional reconstructed attractor, we calculate the F at $\tau=30$ units and $d=2,3,\ldots,30$. The result is shown in Fig. 15(b). It is found that F monotonically increases with d maybe to its limit 1 as $d\to\infty$. Of course, F cannot keep increasing for any linear congruential generator with a modulus m when $d\geqslant m$. In fact, F would decrease rapidly because the periodic sequence with maximal length m cannot fill up an m-dimensional space [26].

A practical observed time series may be contaminated by noise. To show the ability of forecast entropy to capture in-

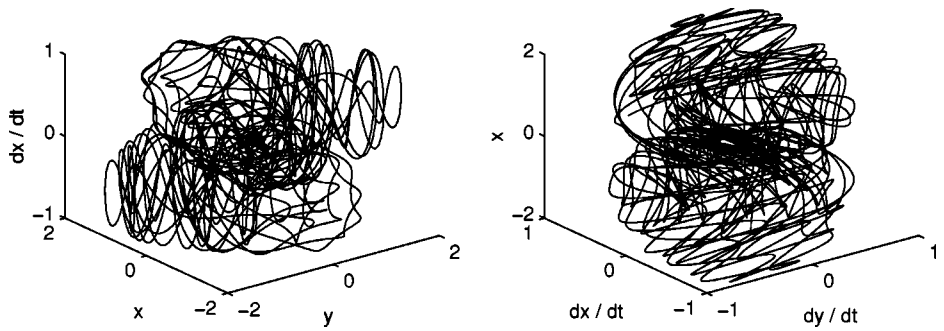

FIG. 12. Phase portraits of the chaotic attractor of system (33).

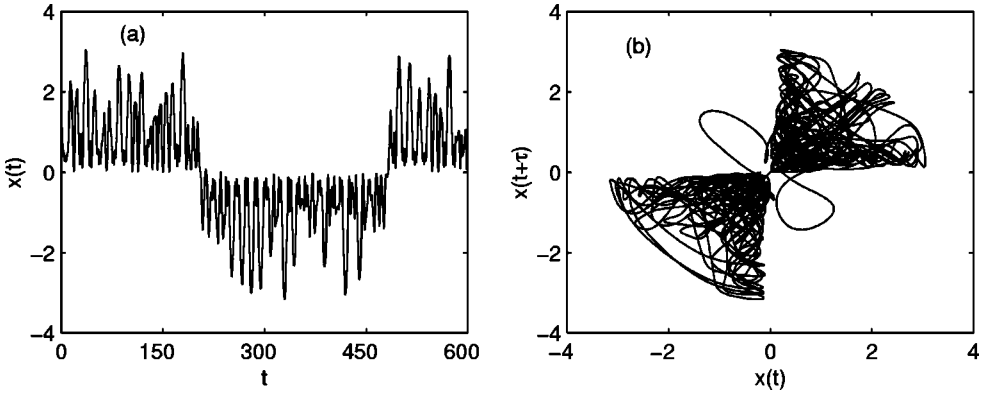

FIG. 13. (a) A piece of time series of $x(t)$ of system (33). (b) The reconstructed attractor from the time series when delay $\tau=6.0$.

formation in a contaminated series, we use the series $s(t_j)=x(t_j)+s_j$, where $x(t_j)$ is the time series from the Lorenz system and s_j the random series studied above. The signal noise rate (SNR) is about 30.

When $n=16$, F is displayed in Fig. 17. It is seen that when $\tau<0.03$ (or 3 units in the pseudorandom generator), similar to the pseudorandom case, F increases as τ does; however, when $\tau\geqslant 0.03$, the shape of F is very similar to that of the noncontaminated time series from the Lorenz system except F is much larger now, and still one can determine the dynamical dimension ($d=3$) of the deterministically chaotic system from the noisy series.

VIII. CONCLUSION AND DISCUSSION

We have proposed a *forecast entropy* which measures the difficulty of predicting an observed time series. Unlike existing entropy-based measures, which depend only on the distribution in regular space, forecast entropy is based on the distributions of the time series in difference spaces up to some maximum order. In this paper, we have focused on the distribution in both regular and tangent spaces.

To measure the distribution in tangent space, our procedure considers the distribution from the coarsest to the finest resolutions. We have shown several examples involving systems of various dimensionality and ranging from deterministic to pseudorandom. We may conclude that our procedure can determine the difficulty of prediction. This is equivalent to determining the complexity of the local structure of the reconstructed attractor.

Further, forecast entropy can also capture some important information such as good values of the delay, the embedding

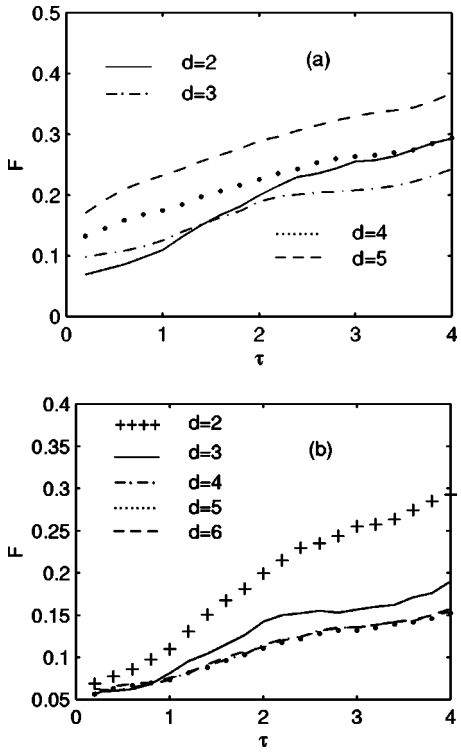

FIG. 14. F vs τ of system (33), when $d=2,3,4,5$, and 6, respectively. (a) $n=2^{d+2}$, (b) $n=16$.

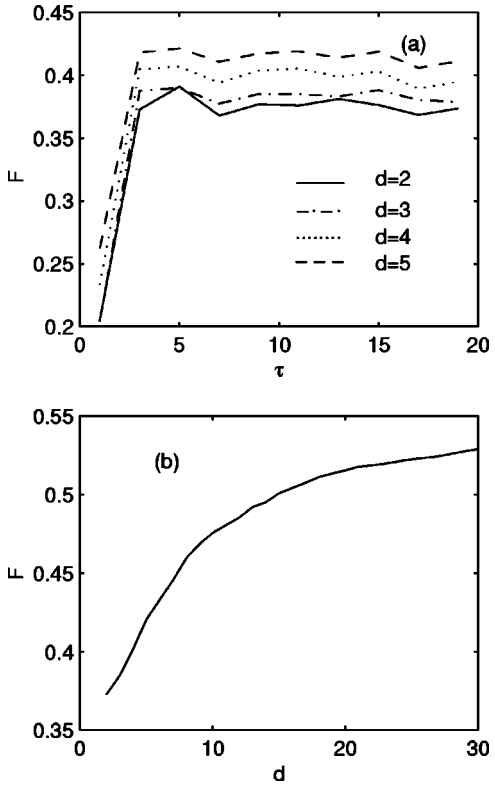

FIG. 15. (a)F vs τ of system (34) when $n=16$. (b) F vs d of system (34) when $\tau=30$ units and $n=16$.

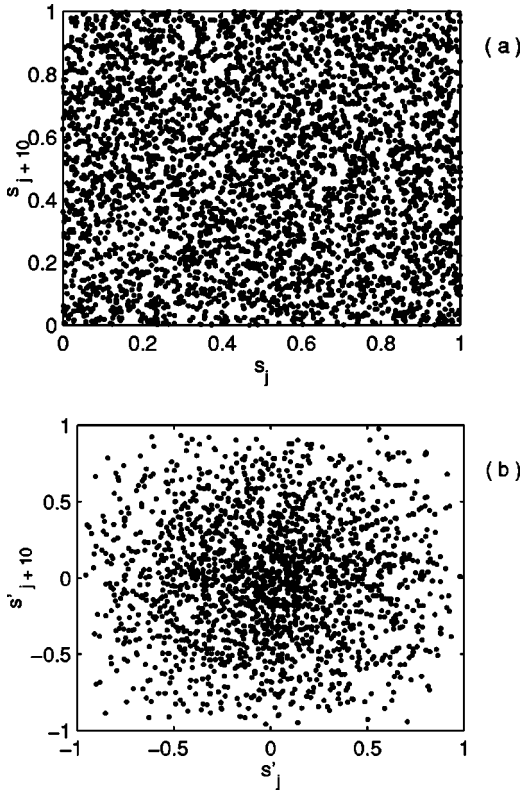

FIG. 16. The distributions of the pseudorandom system in (a) regular space and (b) tangent space.

dimension for optimal prediction, and especially the dynamical dimension of the system generating the observed time series. Finally, the forecast entropy procedure may be useful to distinguish whether an observed time series is random or chaotic. In our examples, the forecast entropy of a deterministically chaotic time series was clearly different from that of a random sequence. However, it should be noted that the distinction between deterministic chaotic systems and random systems might not be so obvious. Recent work in ergodic theory suggests that one may put these systems in one frame [28]. Philosophically, with proper measurement, a completely random series could yield a deterministic process.

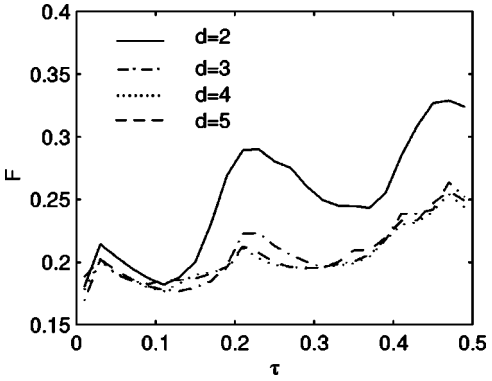

FIG. 17. F vs τ of the noisy time series from the Lorenz system when $n=16$.

When the series is from a high-dimensional system, our procedure is unable to determine the dynamical dimension. This is reasonable: one cannot infer everything that happens in the whole world based on observing the motion of an ant. But if we have more than one series of data generated from the system, we may obtain more information by calculating the forecast entropy.

ACKNOWLEDGMENT

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

[1] G. Boffetta, M. Cencini, M. Falcioni, and A. Vulpiani, Phys. Rep. **356**, 367 (2001).
[2] J.-P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).
[3] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).
[4] A. N. Kolmogorov, Probl. Inf. Transm. **1**, 1 (1965).
[5] J. S. Shiner, M. Davison, and P. T. Landsberg, Phys. Rev. E **59**, 1459 (1999).
[6] M. A.H. Nerenberg and C. Essex, Phys. Rev. A **42**, 7065 (1990).
[7] K. Hornik, M. Stinchcombe, and H. White, Neural Networks **2**, 359 (1989).
[8] S. S. Rao and V. Ramamurti, in *Neural Networks: Theory, Technology, and Applications*, edited by P. K. Simpson (IEEE Press, New York, 1986), pp. 98–103.
[9] G. Pérez and H. A. Cerdetra, Phys. Rev. Lett. **74**, 1970 (1995).
[10] K. M. Short, Int. J. Bifurcation Chaos Appl. Sci. Eng. **4**, 959 (1994).
[11] E. Lorenz, J. Atmos. Sci. **20**, 130 (1963).
[12] A. M. Albano, J. Muench, C. Schwartz, A. I. Mees, and P. E. Rapp, Phys. Rev. A **38**, 3017 (1988).
[13] A. I. Mees, P. E. Rapp and L. S. Jennings, Phys. Rev. A **36**, 340 (1987).
[14] A. M. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986).
[15] A. M. Fraser, IEEE Trans. Inf. Theory **35**, 245 (1989).
[16] A. M. Fraser, Physica D **34**, 391 (1989).
[17] W. Yao, C. Essex, P. Yu, and M. Davison (unpublished).
[18] J. Rissanen, *Stochastic Complexity in Statistical Inquiry* (World Scientific, Singapore, 1989).
[19] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, Phys. Rev. A **45**, 3403 (1992).
[20] A. Cenys and K. Pyragas, Phys. Lett. A **129**, 227 (1988).
[21] C. Essex and M. A. H Nerenberg, Am. J. Phys. **58**, 986 (1990).
[22] E. Ott, W. D. Withers, and J .A. Yorke, J. Stat. Phys. **36**, 687 (1984).
[23] O. E. Rössler, Phys. Lett. **57A**, 397 (1976).
[24] W. Yao, C. Essex, and P. Yu, Dynamics of Continuous, Discrete and Impulsive Systems, Series B: Applications and Algo-

rithms, Vol. 10, pp. 221–234 (2002).
[25] W. Yao, Ph.D. thesis, the University of Western Ontario, 2002.
[26] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, in *Numerical Recipes in C—The Art of Scientific Computing*, 2nd ed. (Cambridge University Press, New Rochelle, 1992), pp. 275–286.
[27] S. K. Park and K. W. Miller, Commun. ACM **31**, 1192 (1988).
[28] D. S. Ornstein and B. Weiss, Bull., New Ser., Am. Math. Soc. **24**, 11 (1991).