Markov models from data by simple nonlinear time series predictors in delay embedding spaces

Mario Ragwitz and Holger Kantz
Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Strasse 38, D-01187 Dresden, Germany
(Received 19 July 2001; published 15 April 2002)

We analyze prediction schemes for stochastic time series data. We propose that under certain conditions, a scalar time series, obtained from a vector-valued Markov process can be modeled as a finite memory Markov process in the observable. The transition rules of the process are easily computed using simple nonlinear time series predictors originally proposed for deterministic chaotic signals. The optimal time lag entering the embedding procedure is shown to be significantly smaller than the deterministic case. The concept is illustrated for simulated data and for surface wind velocity data, for which the deterministic part of the dynamics is shown to be nonlinear.

DOI: 10.1103/PhysRevE.65.056201 PACS number(s): 05.45.Tp, 02.50.−r, 05.40.−a

I. INTRODUCTION

Predictability of observed aperiodic data beyond their linear correlations is usually interpreted as a signature of deterministic structure. Based on the idea of reconstruction of phase spaces from scalar time series and on the hypothesis of deterministic chaos, the tools of nonlinear time series analysis allow one to evidence, characterize, and exploit determinism underlying the dynamics of the observable [1]. Unfortunately, deterministic chaos is only one possible origin of complex aperiodic time series, and intensive studies performed in the last years yielded ample evidence to show that the overwhelming majority of all real world data sets does not belong to this class. Typical phenomena of interest such as weather, climate, economy, biology or physiology either involve too many degrees of freedom to be resolved from scalar data, or the deterministic evolution of some macroscopic degrees of freedom is driven by the noise produced by other degrees of freedom. Therefore, often a nonlinear stochastic approach seems to be more appropriate. Recently, it has been shown that in certain situations nonlinear Langevin equations and Fokker-Planck equations can be derived from data [2]. It seems, however, that this procedure is restricted to Langevin equations with rather few degrees of freedom. Furthermore, one needs to record simultaneous measurements of all relevant degrees of freedom of a system in order to derive the equations of motion. Since this is an unrealistic starting point, we want to follow the spirit of embedding of scalar data. In this paper, we analyze the nature of the information stored in a scalar time series from a possibly multidimensional stochastic dynamical system, e.g., a multivariate Langevin equation. We propose a simple prediction scheme that can be interpreted as a Markov model for the observable.

In many data sets enhanced predictability was found by using nonlinear models living in reconstructed phase spaces. In fact, Casdagli [3] was even using the different predictive power of models ranging from local linear (i.e., globally nonlinear) to global linear ones in order to determine the degree of nonlinearity and determinism in data. If aperiodic data are best predictable by global linear models such as autoregressive processes (AR models), the best physical description is indeed the one given by such a process. If, in contrast, local linear models are superior, then there must be some structure in phase space to achieve this, despite the much reduced statistical robustness because of the locality. This structure is usually interpreted to be deterministic.

In this paper, we show that locally constant predictors in time delay embedding spaces are the natural way to extract a particular nonlinear stochastic process, namely, a Markov model of nontrivial order $m>1$ from observed scalar data. We introduce the prediction scheme, present its theoretical justification, discuss its essential parameters, and discuss its performance for numerically generated data. Finally, employing it to experimental time series data from surface wind velocities, we will show that locally constant predictors can be used to extract the nonlinear deterministic dynamics in boundary layer turbulence.

With the goal of modeling the stochastic systems, Paparella *et al.* [4] have successfully employed local predictors in reconstructed phase spaces for long-term simulations. As the present paper does, Ref. [4], relies on the extraction of the probability density function of the future value from data. As we will recall, in general modeling is different from predicting, and we will discuss the differences in Sec. III.

II. LOCALLY CONSTANT PREDICTORS AND MARKOV MODELS

The meanwhile classical approach of nonlinear time series analysis is the assumption that unpredictability and aperiodicity in data has its origin in a deterministic, chaotic dynamical system in some phase space. The scalar time series obtained by physical measurements is then a (nonlinear) projection of the phase space vectors $\vec{x}(t)$ onto the real numbers, $s_n=h[\vec{x}(t=n\Delta)]$, where Δ is the sampling interval. The concept of embedding [5,6] affirms that in the time delay embedding space of vectors $\vec{s}_n=(s_n,s_{n-\tau},\ldots,s_{n-(m-1)\tau})$ (m sufficiently large), equations of motion of the form $s_{n+1}=g(\vec{s}_n)$ exist. The function g can be reconstructed from the observed data under the assumption of its smoothness, where the pioneering work of Farmer & Sidorowich [7] introduced locally constant and locally linear approximations of g. In the remainder of this paper, we shall use the former and modifications thereof. First, a neighborhood diameter ϵ has to be fixed and neighborhoods $\mathcal{U}_n$ of $\vec{s}_n$ by $\mathcal{U}_n$

$=\{\vec{s}_k: \ ||\vec{s}_k-\vec{s}_n|| \leq \epsilon\}$ are formed. The locally constant predictor for the unobserved s_{n+1} is then

$$\hat{s}_{n+1}=\frac{1}{|\mathcal{U}_n|}\sum_{\vec{s}_k\in\mathcal{U}_n} s_{k+1}, \tag{1}$$

the mean of the ''futures'' of the neighbors. This is the maximum likelihood estimator of $\hat{s}_{n+1}$ under the assumption of Gaussian errors and a function $g(\vec{s})$ that is constant for $\mathcal{U}_n$, hence the name ''locally constant predictor.'' It can be straightforwardly generalized to a locally linear predictor by replacing $g(\vec{s})=\text{const}$ by $g(\vec{s})=\vec{a}\cdot\vec{s}+b$, an affine function.

The superiority of this locally constant or the locally linear fit over a global linear model [an autoregressive model of mth order AR(m)] of the form

$$x_{n+1}=\sum_{i=1}^{m} a_i x_{n-(i-1)\tau}+\xi_{n+1} \tag{2}$$

is usually interpreted as an indication for nonlinear determinism in the data, formalized, e.g., by the Casdagli test [3]. Here, the AR(m) model is a linear stochastic model, driven by random inputs ξ_n, which produces noise-driven damped harmonic oscillations [8].

A scalar Markov process of mth order in discrete time is defined by the fact that for any sequence of successive times $t_1, t_2, \ldots, t_n$ with $n>m$ all transition probabilities fulfill

$$\begin{aligned} &p(y_{n+1},t_{n+1}|y_n,t_n;y_{n-1},t_{n-1},\ldots,y_1,t_1) \\ &\quad=p(y_{n+1},t_{n+1}|y_n,t_n;y_{n-1},t_{n-1},\ldots,y_{n-m+1},t_{n-m+1}), \end{aligned} \tag{3}$$

i.e., the transition probability depends on the last m events only. Since the values of these transition probabilities can be arbitrary, such a Markov model is much more general than the AR(m) model mentioned above.

The purpose of this paper is to show that the locally constant predictor, originally based on the assumption of determinism, is in fact a particular predictor based on a Markov assumption. Its superiority with respect to a linear stochastic model can thus as well mean that the data are generated by a Markovian, nonlinear stochastic model. Apart from the conceptual difference, this has implications on the issue of modeling versus prediction: for a Markov model with non-δ-shaped transition probabilities, modeling and prediction are largely different tasks. But for nonstandard cost functions too, modified predictors can be useful.

First, we observe that the independent variables entering the transition probabilities of Eq. (3) are exactly the elements of a delay vector $\vec{s}_n$, if we identify the times with the corresponding integer multiples of the sampling interval, $t_k=k\Delta$. The times t_k will, therefore, be suppressed in the following, and the transition probabilities will be denoted by $p(y_{n+1}|\vec{y}_n)$. In order to extract these probabilities from data, we have again (as in the deterministic case above) to make the assumption that their dependence on $\vec{y}$ is smooth. Then it is reasonable to use the following approximation:

$$p(y_{k+1}|\vec{y}_k)\approx\hat{p}(y_{n+1}|\vec{y}_n) \quad \forall \ \vec{y}_k\in\mathcal{U}_n, \tag{4}$$

i.e., we use again a locally constant approximation. $\hat{p}(y_{n+1}|\vec{y}_n)$ can be estimated from the observed values of y_{k+1}, the ''futures'' of the elements $\vec{y}_k\in\mathcal{U}_n$, which form a sample according to $\hat{p}(y_{n+1}|\vec{y}_n)$.

In the deterministic case with sufficiently large m, the transition probabilities are δ shaped, $p(y_{n+1}|\vec{y}_n)=\delta(y_{n+1}-g(\vec{y}_n))$, and the estimate of Eq. (4) yields some narrow distribution, provided the neighborhood $\mathcal{U}_n$ is not too large in diameter. For a truly random setting this distribution might be broad and (if the sampling interval Δ is relatively large) even multimodal. In this latter case, the way how the knowledge of a sample of $\hat{p}(y_{n+1}|\vec{y}_n)$ is evaluated depends on the purpose.

When prediction is the goal, a typical cost function to be minimized is the mean squared prediction error $e^2=\Sigma(y_{n+1}-\hat{y}_{n+1})^2$. The best predictor is then the mean $\hat{y}_{n+1}=\int y'_{n+1}p(y'_{n+1}|\vec{y}_n)dy'_{n+1}$, i.e., exactly the locally constant predictor given by Eq. (1). Depending on the shape of $p(y_{n+1}|\vec{y}_n)$, the mean can be a value unlikely to be attained by y_{n+1}, and an iteration of this prediction scheme can yield a quite atypical sequence of y's, drastically different from the behavior of the true data. Modeling thus would require to choose a value at random from the observed distribution. This method was called local random analogue prediction in Ref. [4]. In cases where the mean value of $p(y_{n+1}|\vec{y}_n)$ is a particularly bad representative of the full distribution, modeling and prediction are hence two very distinct tasks. In such cases, it makes sense also to discuss other cost functions for predictions. If on the average the error should as often be positive as negative, the median is the optimal predictor. If we want to penalize large errors to their extreme, the cost function would be the maximum of all errors made. In this case, the optimal prediction is $\hat{y}_{n+1}=\frac{1}{2}(y_{max}-y_{min})$, where $y_{max/min}$ are the largest/smallest values of z for which $p(z|\vec{y})$ is nonzero.

Regardless of which cost function one uses and irrespective of whether the process is assumed to be stochastic or deterministic, the width of the distribution $p(y_{n+1}|\vec{y}_n)$ is a direct measure for the accuracy of the prediction: The more spread there is among the y_{k+1}, the larger is the local instability, and hence the larger might be the deviation of y_{n+1} from the mean of this distribution. This is illustrated in Fig. 1 for a deterministic chaotic model with measurement noise (where the spread is related to the position dependent exponential divergence of nearby trajectories, sometimes called local Lyapunov exponents), and for data from a nonlinear Langevin equation. We hence propose to use the variance of the transition probability distribution as a criterion for the reliability of the actual prediction.

III. WHEN IS A SCALAR OBSERVATION FROM A MULTIVARIATE LANGEVIN EQUATION MARKOVIAN?

If the hypothesis about a scalar time series is that it represents one observable from a vector-valued deterministic

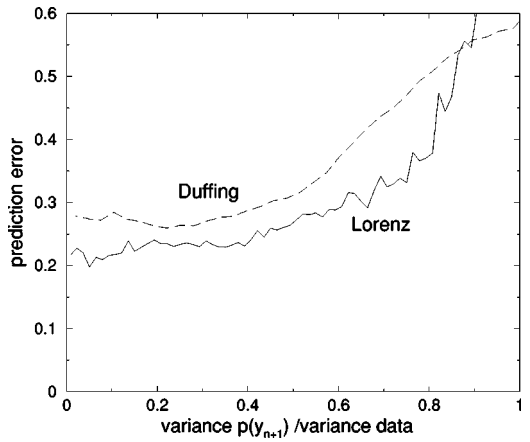

FIG. 1. The mean prediction error, $\sqrt{\langle(\hat{y}_{n+1}-y_{n+1})^2\rangle}$, as a function of the standard deviation of the distribution of $p(y_{n+1}|\vec{y}_n)$. Continuous line, chaotic time series of the Lorenz system [Eq. (9)] with additive measurement noise; dashed line, noise-driven Duffing oscillator [Eq. (6)].

system, the above-mentioned embedding theorems allow one to reconstruct a vector-valued space, the time delay embedding space, in which determinism is restored. The corresponding problem for Markov processes is the following: Given is a vector-valued, multidimensional Markov process and a single observable. Does the time series of values of this observable represent a scalar Markov process of some finite order m in time? Although this question is typically not posed, the general answer is well known to be negative [9]. Nonetheless, since in time series analysis one usually never has observations in the full phase space, this is a relevant issue that will be discussed in some detail in this section.

To make the relation to the deterministic setting as close as possible, we assume as a generator of the Markovian dynamics a Langevin equation of the type

$$\dot{\vec{x}}=\vec{f}(\vec{x}(t))+\mathbf{G}(\vec{x}(t))\vec{\Gamma}(t), \tag{5}$$

which is the generalization of a continuous time dynamical system, with an additional stochastic force $\mathbf{G}(\vec{x}(t))\vec{\Gamma}(t)$, where $\vec{\Gamma}(t)\in\mathbb{R}^l$ with $\langle\Gamma_k(t)\Gamma_{k'}(t')\rangle=\delta_{k,k'}\delta(t-t')$ as a l-dimensional Gaussian white noise and $\mathbf{G}(\vec{x})$ a $(n\times l)$-dimensional matrix function. For a n-dimensional state space with state vectors $\vec{x}(t)=(x_1(t),\ldots,x_n(t))$, Eq. (5) defines a Markov process of order n. For the deterministic limit $\mathbf{G}(\vec{x})\equiv 0$, the embedding theorem of Takens states that one can reconstruct the dynamics of the multidimensional process by using subsequent values of just a single scalar observable $h[\vec{x}(t)]$. Does a similar procedure exist for the multidimensional process generated by the Langevin equation? Can the information contained in the n-dimensional state vector $\vec{x}(t)$ somehow be reconstructed by measuring only s components with $s<n$, at subsequent times, employing a finite memory of those? In general the answer to this question is no. One cannot expect that the knowledge of only a few components is sufficient to fix the future probability distribution completely, not even for one of these s components.

However, we will demonstrate here that in certain situations the dynamics in a delay embedding space of a single observable is Markovian, and in many more situations it is approximately Markovian in the sense that the memory, albeit formally infinite, can be assumed to be finite in the sense that the errors thus introduced are smaller than the modeling errors stemming from the fact that all information about the process is extracted by statistical means from a finite amount of data.

Whether or not a scalar measurement from a multidimensional Langevin equation is Markovian depends on the system as well as on the measurement function. Let us first consider a simple example. The Duffing system

$$\frac{dx}{dt}=v(t),$$

$$\frac{dv}{dt}=av(t)-x^3(t)+x+b\Gamma(t) \tag{6}$$

describes the stochastic motion of a damped particle in a double well potential, where $v(t)$ is the velocity and $x(t)$ the position of the particle. This equation defines a Markov process of order 2 in (x,v). The particle keeps moving through the stochastic kicks $\Gamma(t)$. The change in the velocity of the particle is determined by the stochastic inputs as well as by the position of the particle. As argued by van Kampen [9] the position of the particle depends on the velocity at all previous times. Therefore, the velocity possesses an infinite memory. One has to know the velocity at all former times in order to determine the probability distribution for its future value. In the deterministic case (without stochastic forcing) the reconstruction relies on the fact that the second equation of Eq. (6) can be solved for x if a sufficient number of derivatives of y is given. This inversion property breaks down if the stochastic force is added and one, therefore, has to resolve x from the first equation in Eq. (6).

In the limit of $\delta t\to 0$ the x coordinate fulfills a second order Markov property, i.e., knowing the position at times $t-\delta t$ and t is sufficient to estimate the probability distribution of its next value. This is due to the fact that the knowledge of $x(t-\delta t)$ and $x(t)$ gives an estimate for the velocity $v(t)$, and knowing the position at $x(t-2\,\delta t)$ does not supply any additional information. In all cases where we speak about the Markov property for a discretely sampled observable of a time continuous system we implicitly refer to the limit $\delta\to 0$.

Applying the same arguments one can, for example, also understand the effect of dynamical noise coupled into the Rössler and the Lorenz systems. If the z variable of these systems is driven by noise one finds that the x coordinate is a Markov process of order 3 whereas the y coordinate possesses an infinite memory.

As a first example to investigate numerically, let us consider the noise-driven van der Pol oscillator that has a stable limit cycle as asymptotic solution

$$\dot{x} = y(t) + a\Gamma(t),$$

$$\dot{y} = [r - x(t)^2]y(t) - x(t) + b\Gamma(t). \tag{7}$$

The system was modeled by driving either the first or the second equation by the white noise inputs with unit variance (properly rescaled by the square root of the step width of the Euler integrator in the numerical simulation). So the parameters are either $(a,b)=(0.5,0)$ or $(a,b)=(0,0.5)$ and $r=3.0$. In the case where the second equation is driven by the noise ($a=0$), the x coordinate is a Markov process of second order in time. Deriving the first equation with respect to time, substituting $\dot{y}$ by the second equation, and replacing y by the use of the first one leaves us with a stochastic differential equation of second order in time for the variable x, which generates the Markov process.

Employing the same arguments as before, this gives rise to three different situations that we want to analyze numerically in the following: (a) The noise is added to an unobserved variable but the Markov property is valid for the observed coordinate. (b) The noise is added to the observed variable and hence destroys the Markov property. (c) The noise is added to an unobserved variable and the observed variable is not Markovian.

Let us first have a look at the reconstructed phase spaces for these three situations: For the case where the noise is coupled to the second equation, we show in Fig. 2(a) the phase portrait using the x variable for the embedding. In Fig. 2(b) we show the phase portrait in the case where the noise is coupled to the first equation, again using the x variable for the embedding. In Fig. 2(c) the noise is again added to the x variable but this time y is used for the reconstruction. In the first case we see only small deviations from the limit cycle. The second portrait appears like a random walk added to the limit cycle, whereas in the third plot the original limit cycle seems to have additional nontrivial structure.

These three cases will now be analyzed by means of the above introduced locally constant predictors. We will perform predictions as outlined above by varying the embedding dimension and the time delay. Usually, while increasing the embedding dimension one has to use larger neighborhood diameters ϵ in order to collect a certain number of neighbors in $\mathcal{U}_n$. This might penalize higher-dimensional embedding s. In order to rule out such an effect we require the same fixed number of neighbors within a fixed diameter of the neighborhood for each embedding dimension m and each time delay d, and vary the length of the time series within which the neighbors are sought for (we run through the time series backward in time until a certain number of neighbors have been found).

The result of the predictions is shown in Figs. 3(a)–3(c), where the prediction error (normalized by the standard deviation of the data) versus the time delay τ is shown for different embedding dimensions. We identify different memories of the time series depending on the way the noise is coupled into the system. For the first case (a), we find that the minimum of the prediction error does not depend on the embedding dimension for $m \geqslant 2$. In addition, the optimal embedding window, i.e., the time interval $(m-1)\tau$ spanned by a delay vector with optimal τ, is independent of m for $m \geqslant 2$, hence confirming that a second-order model is sufficient. In contrast we find for case (b) that the prediction error decreases for increasing embedding dimensions—each dimension adds information when predicting the future probability distribution. The improvement, however, amounts only to a small percentage of the total error. If we drive the x coordinate by the noise inputs and record the y coordinate (c) the y coordinate seems no longer to be Markovian. We find

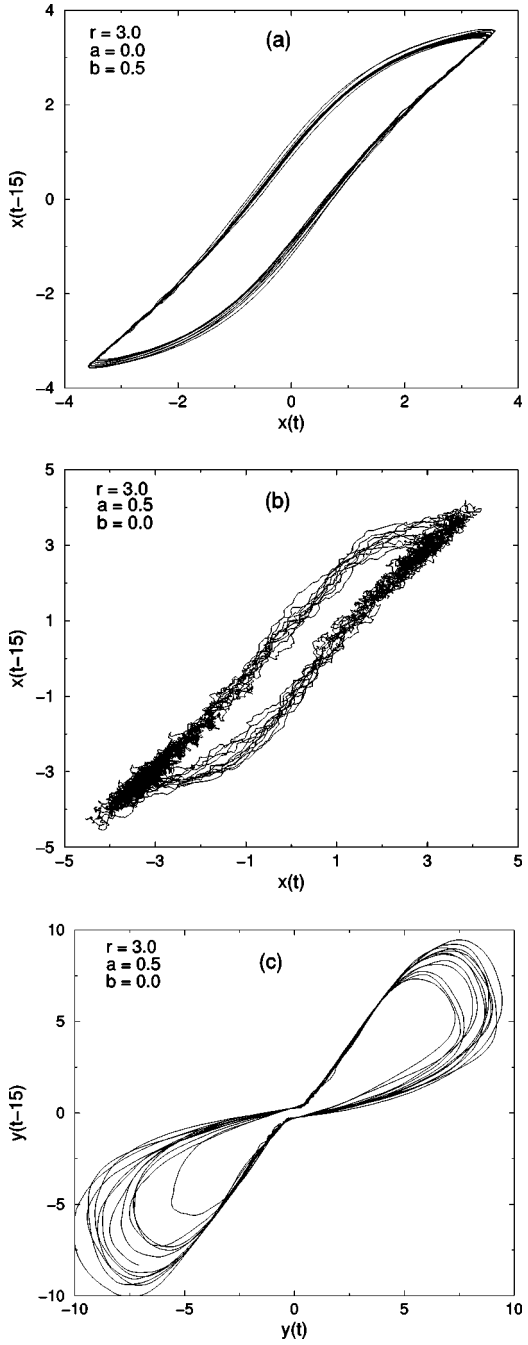

FIG. 2. Two-dimensional projection of the phase portrait of the van der Pol system for different configurations of the noise driving and different variables used for the embedding.

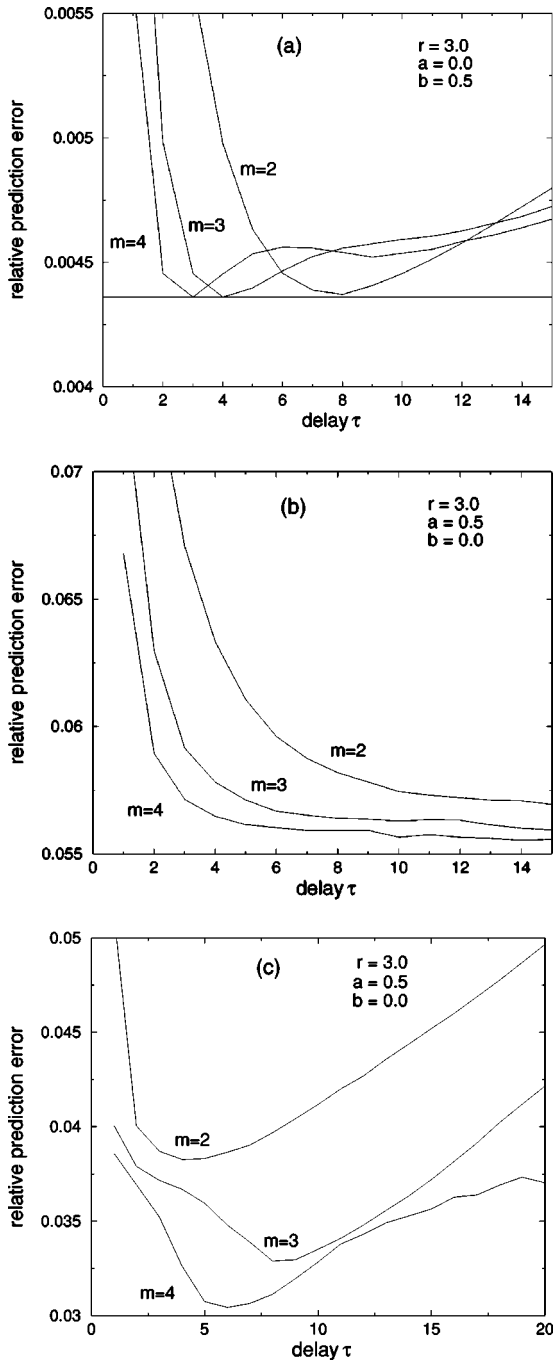

FIG. 3. Relative prediction errors (normalized by standard deviation of the data) versus time delay for different embedding dimensions. Noise configurations and coordinates used for embedding equal the corresponding values in Fig. 2.

higher-order memories in this variable. The improvement using higher-dimensional embedding is even significantly better than for the actually noise-driven variable (b).

To study the difference between these three cases by means of a well established approach for stochastic dynamical systems we will now use the concept of coarse grained dynamical entropies [10]. More precisely, we will exploit the properties of the conditional entropy $h_n(\epsilon)=H_{n+1}(\epsilon)-H_n(\epsilon)$, estimated by the correlation entropy

$$H_n(\epsilon)=-\ln C\left(n,\frac{\epsilon}{2}\right)=-\ln\left[\frac{2}{(N-n)(N+1-n)}\times\sum_{i<j}\Theta\left(\frac{\epsilon}{2}-|\vec{s}_i-\vec{s}_j|\right)\right]. \tag{8}$$

Here the $\Theta(\cdot)$ is the Heaviside step function, $\vec{s}_i$ are n-dimensional delay vectors and N is the length of the data sequence. It is well known that the conditional entropies $h_n(\epsilon)$ behave as $h_n(\epsilon)=c_n-\ln\epsilon$ for a stochastic process, where the constants c_n are monotonically decreasing with n. If the process is a Markov process of order m, $c_n=c_\infty$ for all $n>m$. Hence, in a logarithmic representation of the conditional entropies $h_1,\ldots,h_n$ for a Markov process of order m one finds m parallel but distinct lines. All curves for $n>m$ collapse onto the graph of h_m since there are no memories present of order $m+1$ and higher, which could reduce the entropy further.

The conditional entropies $h_1,\ldots,h_5$ are shown for the process in Eq. (7), for the cases where the noise is added to y [Fig. 4(a)] or to x [Figs. 4(b) and 4(c)]. In the case where the process is Markovian of order 2 (noise added to y, x recorded) we find h_2, h_3, h_4, and h_5 collapsing onto a single line for the range of ϵ values corresponding to the stochastic regime. In the case where the noise is added to the observed x variable directly we find that on small-length scales mainly the random motion around the limit cycle becomes visible and higher-order memories are difficult to detect using conditional entropies. Only in the last case we see that the noise introduces longer memories and creates nontrivial higher-dimensional structures in phase space.

In summary, only in exceptional cases, an observed scalar time series can be assumed to be Markovian. However, from the practical point of view, it seems that often (such as here) the memory, albeit formally infinite, decays fast and hence the process can be approximated by a finite-order Markovian process. If the error introduced by this approximation is smaller than other modeling errors caused by the finiteness of the data set, there is practically no difference between the Markov approximation and a hypothetical infinite memory model.

IV. OPTIMAL EMBEDDING PARAMETERS

In a deterministic system a strict lower bound for the dimension is $m\geqslant D$, where D is the number of active degrees of freedom [7]. The optimal value for the dimension can, however, be larger, since the sufficient embedding requirement to obtain an unfolded attractor is $m>2D$. An estimate of a proper value of m can be obtained using the method of false neighbors [11]. Using embedding dimensions higher than that necessary to unfold the attractor adds redundancy in the neighbor search and worsens the performance of the predictions due to the finite precision of the data and the limited length of the time series.

The time lag τ is not a subject of the embedding theo-

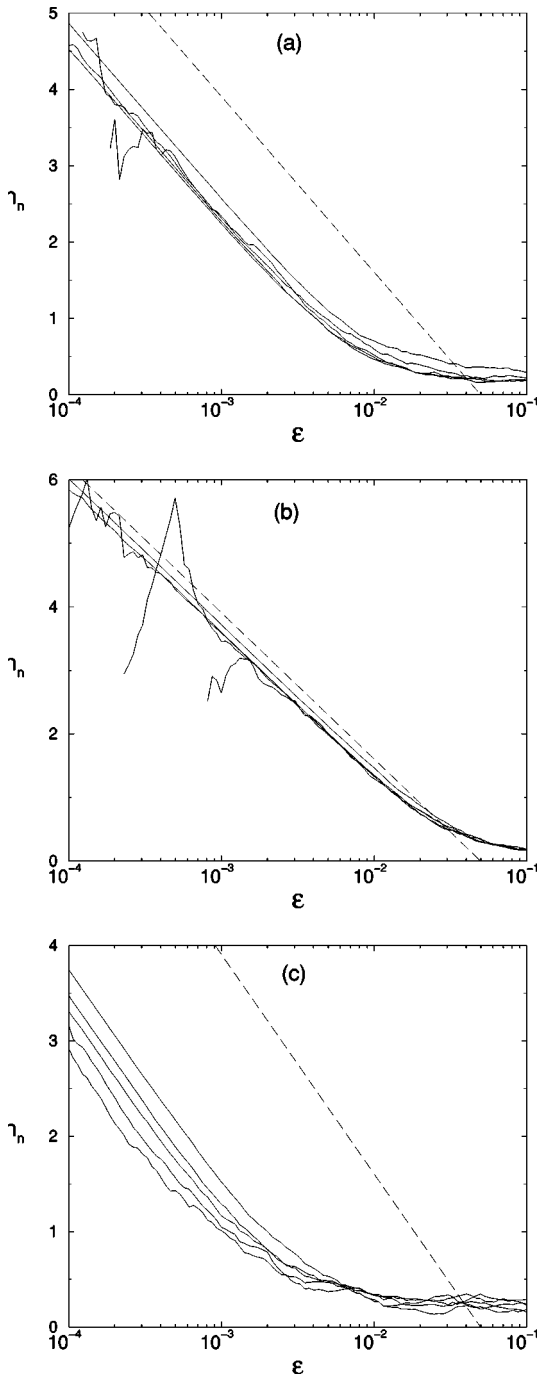

FIG. 4. Correlation entropies $h_1, \ldots, h_5$ for the attractors of the van der Pol system with delay $\tau = 35$. Noise configurations and coordinates used for embedding equal the corresponding values in Fig. 2. The dashed line shows the function $f(\epsilon) = -\ln(\epsilon) - 3$.

rems. In the limit of infinite precision of the data and infinitely long time series, all values of τ are equivalent. In a practical situation, however, a good choice of the delay is crucial. If τ is too large, successive elements of the embedding vector are almost independent and the vectors fill a large cloud in the reconstructed phase space. If τ is too small, successive elements of the embedding vector are strongly correlated and all vectors are clustered around the diagonal. Meaningful neighborhoods are difficult to obtain in

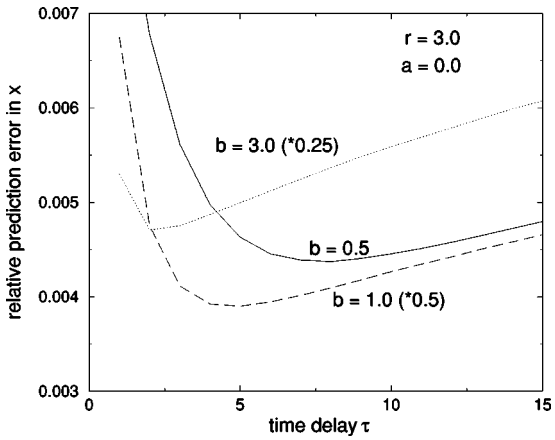

FIG. 5. Relative prediction error versus delay for different noise levels for the van der Pol system $b = 0.5,\ 1.0,\ 3.0$. The two curves for $b = 1.0$ and 3.0 are rescaled by a factor of 1/2 and 1/4, respectively.

both cases. These phase space considerations represent geometrical arguments to find a good delay, which have been formalized by statistics such as *fill factor* [12] or *displacement from diagonal* [13]. Since these are just recipes, it is often better to use the autocorrelation function (or the time delayed mutual information [14] to account for all nonlinear correlations). The simplest reasonable estimate of an optimal delay is the first zero of the autocorrelation function of the signal [1]. The striking point is that these estimates generally yield too large τ values for stochastic dynamical systems, as we will discuss here.

Now let us examine how the model parameters should be chosen in the case that the time series has been generated by a Langevin process. As pointed out in the preceding section one has to consider in general two different cases. One possibility is that the time series represents a Markov process of the order of the original multidimensional process given by the Langevin equation. In this case the optimal embedding dimension is the order of the process and every further information, i.e., higher-dimensional embedding, only adds redundancy. The second possibility is that the time series possesses an infinite memory due to the stochastic driving force in the Langevin equation. In this case every additional dimension adds information and increases the predictability.

The way in which the unrecorded variables are recovered by the delay embedding tells us about the optimal delay. Formally, we reproduce the hidden variables by introducing higher derivatives of the measured variable. Since these derivatives are practically replaced by the difference between two values of the time series in the limit of vanishing time difference between these values, the delay should be in principle as small as possible and hence the sampling interval of the time series. We will see from our examples that this is the case for high noise levels. For very small noise levels we evidently find that the optimal delay assumes the value found for deterministic case, e.g., the first zero of the autocorrelation function. For intermediate noise levels the optimal delay interpolates monotonically in between these two limits. This is illustrated in Fig. 5 for the van der Pol system [Eq. (7)],

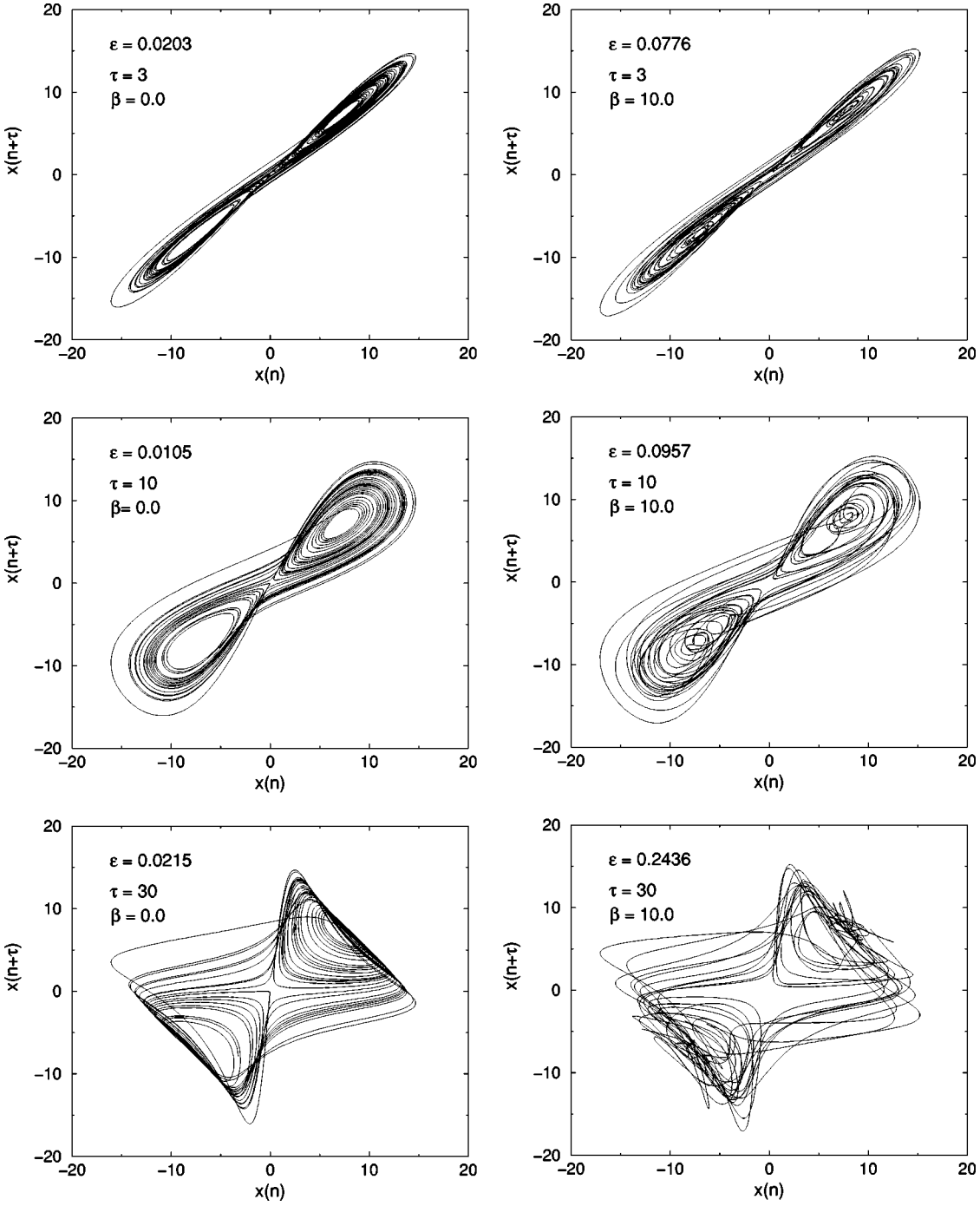

FIG. 6. Two-dimensional projection of the phase portrait of the Lorenz system for different values of the delay $\tau=3$, 10, and 30, vanishing noise level [(a)–(c)] and noise level $\beta=10.0$ [(d)–(f)]. The relative prediction error ϵ is also given in the graphs.

when the second equation is driven by the noise and x is measured.

As a second example we want to analyze the noise-driven Lorenz system. The equations of motion are

$$\dot{x}=6(y-x),$$

$$\dot{y}=28x-y-xz, \tag{9}$$

$$\dot{z}=xy-(13/6)z+\beta\Gamma.$$

A time series of length $N=100\,000$ was generated by integrating these equations and sampling the x component every 0.1 units of time. In Figs. 6(a)–6(c) we show three different values of the delay and for vanishing noise level a two-dimensional projection of the phase portrait of this system. For a small delay of $\tau=3$ the phase portrait is centered around the diagonal since consecutive values of the time series are very similar. For an intermediate delay of $\tau=10$ the attractor seems well unfolded. Larger delay times such as $\tau=30$ lead to complicated intersecting graphs. For vanishing

noise level the apparently simplest phase portrait [Fig. 6(b)] leads to the minimal relative prediction error ϵ, where the absolute error is normalized by the variance of the data σ. However, the optimal value for the delay decreases as the noise level is increased. In Figs. 6(d)–6(f) we show phase portraits of the Lorenz system with the same values of the delay as in Figs. 6(a)–6(c) but for a noise level of $\beta=10.0$. Still the phase plot with $\tau=10$ shows the best unfolded attractor by visual inspection. But this time the minimal prediction error is achieved with a delay of $\tau=3$. Therefore, for stochastic dynamical systems the visual inspection of the phase portrait as well as the first zero of the autocorrelation function might not be a good condition for choosing an optimal value of the delay. We hence recommend an explicit optimization of the prediction errors with respect to the time lag τ.

V. NONLINEAR FLUCTUATIONS IN STOCHASTIC SYSTEMS

The issue of whether or not a stochastic process is linear has strong implications on the magnitude of fluctuations as a function of time. Nonlinear fluctuations in Markovian processes can be identified using the suggested prediction algorithm. The idea is that if we use a linear process to predict an intrinsically nonlinear process, the predicted fluctuations are too small on an average. We hence analyze the probability density function (PDF) of the increments $\Delta\hat{s}_T=\hat{s}^{model}_{n+T}-s_n$ predicted by the AR model and by the locally constant model, where $\hat{s}^{model}_{n+T}$ is the value of the signal a time T ahead as predicted by either of the two models. Although it is possible that the difference of the average prediction error of an AR model and a nonlinear model is small, the PDF's of the predicted increments can differ clearly. This is due to the fact that long tails of the PDF of a nonlinear stochastic process have their origin in stochastic fluctuations as well as in nonlinear correlations. The latter can be modeled by a nonlinear scheme but not by a linear algorithm.

For the noise-driven Duffing oscillator [Eq. (6)] with $a=-0.5$ and $b=0.5$ the motion is essentially created by the stochastic term—without noise the particle would come to rest—and the deterministic part of the dynamics is nonlinear. A time series of this process is shown in Fig. 7 using a sampling rate of $\delta t=0.1$. In Fig. 8 we show the PDF's of the linearly and nonlinearly predicted increments $\Delta\hat{s}_T=\hat{s}^{model}_{n+T}-s_n$. The locally constant predictor was run in a two-dimensional embedding space with optimal delay of $\tau=10\delta t$, and the linear predictions were performed by using an AR(2) model with equal time lag. The prediction horizon was $T=20\delta t$.

We also show the PDF of the increments of the actual time series and of the increments modeled by knowing the exact deterministic part of the equations of motion (6). Whereas the AR model is unable to capture the large increments of the signal, the locally constant scheme gives a good approximation to the PDF produced by the deterministic part of the equations of motion.

If the locally constant model was able to capture the deterministic part of the Markovian dynamics, all predictions should be correct on average. In the same sense, the predicted increment, i.e., the difference between the predicted value $\hat{s}_{n+T}$ and the actual value s_n, should be correct on average. This means that if one predicts an increment $\Delta\hat{s}_T=\hat{s}_{n+T}-s_n$ in a number of k situations the average of the actually measured increments Δs_T in these situations should converge towards $\Delta\hat{s}_T$ for large k.

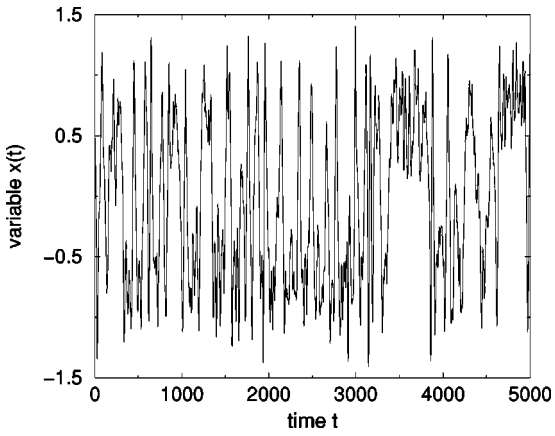

FIG. 7. Variable x versus time t for the Duffing oscillator.

Using this statistics we observe a significant difference between the two prediction schemes as is shown in Fig. 9. Whereas the locally constant predictor gives the accurate increments on average, the linear model systematically underestimates the fluctuations. The AR(2) model is unable to fit the nonlinear deterministic part of the process. Thus it is shown that a locally constant model captures the nonlinear deterministic dynamics of this second-order Markovian process and the suggested statistics can be used to detect nonlinearity in a time series, whereas in this case the average prediction errors of the two models are very similar.

The scheme developed above will now be applied to surface wind velocities of the atmospheric boundary layer. As studied in Ref. [15] for the prediction of surface wind veloci-

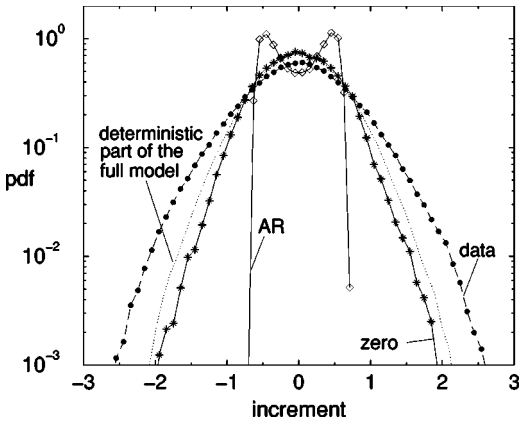

FIG. 8. PDF of the increments of the x variable of system 6. Also shown are the increments predicted by the linear (AR) and by the nonlinear models (zero) as well as the increments predicted by knowing the exact deterministic part of the equations of motion.

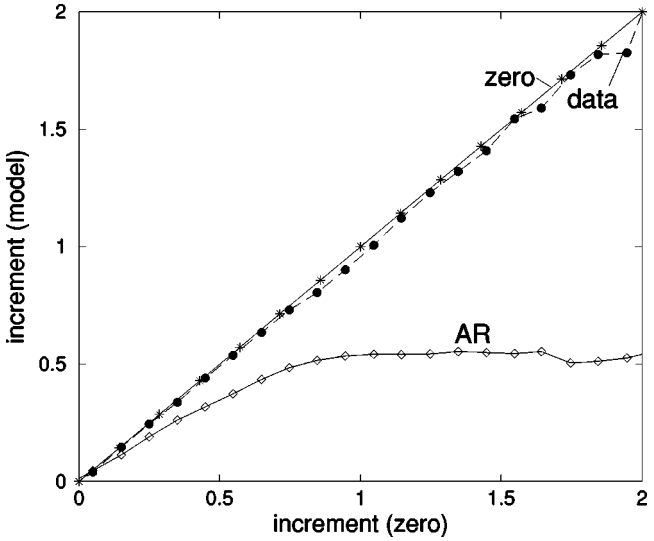

FIG. 9. Actually measured increment Δs_T and AR-predicted increment $\Delta \hat{s}_T^{AR}$ versus the increment $\Delta \hat{s}_T^{zero}$ predicted by the locally constant scheme.

ties, no reduction of the average prediction error can be achieved using a nonlinear scheme instead of a linear scheme. Improvement of the nonlinear model is possible, however, in situations where a large increase of the wind speed is predicted by the nonlinear algorithm. This behavior can be understood in a more general fashion using the framework of Markov models.

We use data recorded on the Lammefjord on the island Seeland in Denmark. The terrain around the measurement station is very flat and no major obstacles interfere with the fluid flow. One component of the wind velocity was recorded with a sampling rate of 8 Hz using an ultrasonic anemometer located at an altitude of 10 m during a period of 24 h. A typical time series of the wind velocity is shown in Fig. 10.

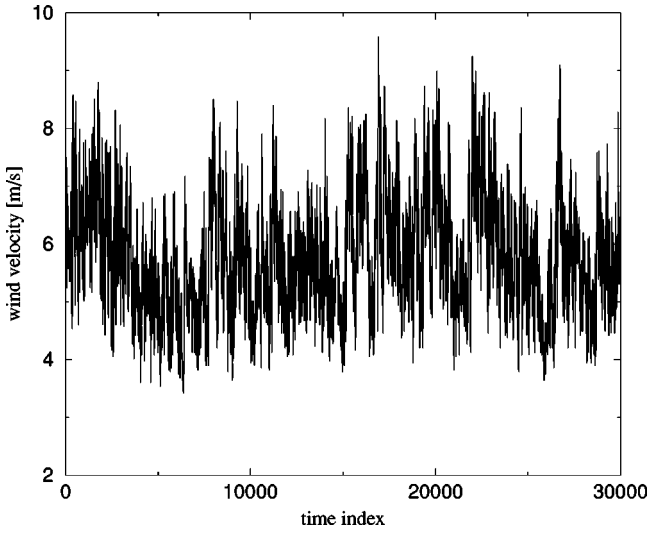

FIG. 10. Time series of the total wind velocity during a period of 1 h recorded on the Lammefjord in Denmark with a sampling rate of 8 Hz.

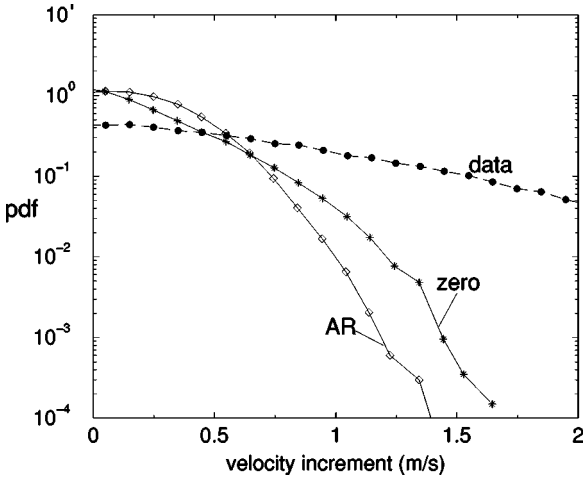

FIG. 11. Positive branch of the PDF of the increments of the time series of the surface wind. Also shown are the increments predicted by the linear [AR(10)] and by the nonlinear model (zero) with embedding dimension $m=10$.

The signal appears highly disorganized and presents structures on all time scales.

Our analysis again relies on the prediction of the wind velocities using a locally constant predictor in ten-dimensional embedding space and an AR(10) model (employing time lag of unity), with a prediction horizon T of 20 sampling intervals. We show in Fig. 11 the right-hand branch of the PDF of the predicted increments as well as the PDF of the increments of the time series. As in the Duffing system, one can see a difference between the PDF's of the increments predicted by the linear and by the nonlinear models. The latter predicts larger fluctuations of the signal. Next we will show that these large fluctuations predicted by the nonlinear scheme give on an average a better representation of the real increments than the linear model in these situations.

Figure 12 shows the measured increments and the increments predicted by the AR model and the nonlinear model

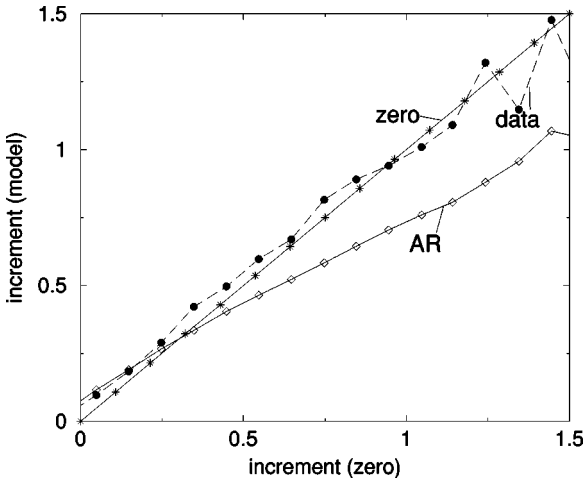

FIG. 12. Actually measured increment Δs_T and AR-predicted increment $\Delta \hat{s}_T^{AR}$ versus the increment $\Delta \hat{s}_T^{zero}$ predicted by the locally constant scheme with embedding dimension $m=10$.

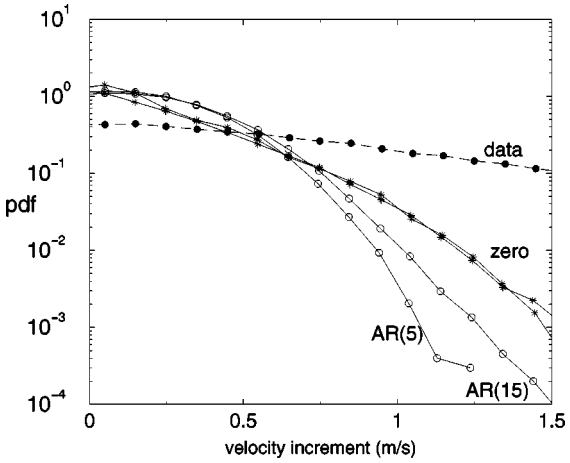

FIG. 13. Positive branch of the PDF of the increments of the signal and that given by the predictors as in Fig. 11, but for embedding dimensions of $m=5$ and $m=15$.

versus the increment predicted by the nonlinear scheme. We observe a similar behavior as for the Duffing oscillator, whereas the locally constant model gives a satisfying representation of the increments that the AR model significantly underestimates on an average. This shows that the data of surface wind velocities are essentially nonlinear and a nonlinear model is able to fit nonlinear fluctuations.

We have chosen the embedding dimension $m=10$ because this value is large enough to obtain nontrivial predictions and sufficiently low in order to keep the computational effort manageable. Since there is no *a priori* optimal value for m in higher-dimensional systems as for the surface wind, we consider the last m measurements to contain the dominant information on the transition probabilities and the earlier events to be corrections thereof. We want to demonstrate now that the result presented above is valid for a range of different values of m. This is shown in Figs. 13 and 14 for the values $m=5$ and $m=15$. The qualitative behavior is the same as in Figs. 11 and 12.

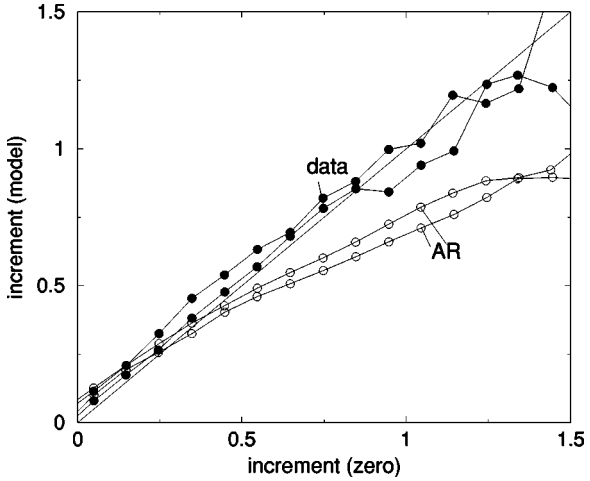

FIG. 14. Actually measured increment Δs_T and AR-predicted increment $\Delta \hat{s}_T^{AR}$ versus the increment $\Delta \hat{s}_T^{zero}$ predicted by the locally constant scheme for embedding dimensions of $m=5$ and $m=15$.

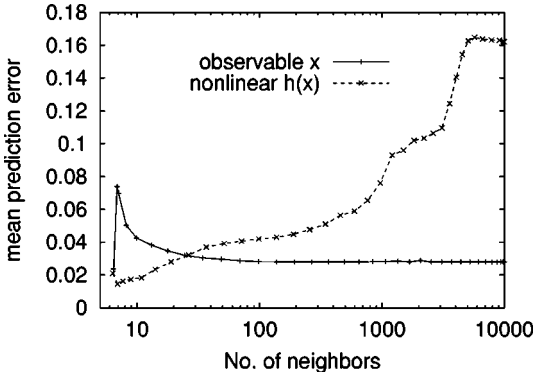

FIG. 15. The average forecast errors of local linear models as a function of the number of neighbors used for the fits, for $\{x_n\}$ (continuous) and $\{s_n\}$ (dashed) of the AR(2) process (see text).

VI. NONLINEAR MEASUREMENT FUNCTION

An important and typical situation is that the output of a linear system is measured by a nonlinear function. This induces nonlinear correlations in the data. Especially if the measurement function is unknown or not invertible, one has to resort to a nonlinear algorithm to model the time series. To compare the predicting power of a nonlinear model and a linear model on data transformed by a nonlinear function we use a method suggested by Casdagli [3]. Using this statistics one can tune between a globally linear model and a local model by computing the one-step prediction error for the linear approximation as a function of the neighborhood size diam($\mathcal{U}$). For small neighborhood size diam, one has a local model but for neighborhoods in the limit of the attractor size the predictions are given by the AR model. Let us investigate this statistics for an AR(2) process, namely, $x_{n+1}=a_1x_n+a_2x_{n-1}+\Gamma_n$ with the measurement function $s_n=\mathrm{sgn}(x_n)\sqrt{|x_n|}$ and $a_1=1.985$ and $a_2=-0.995$. In Fig. 15 the average forecast errors of the local linear models as a function of the number of neighbors used for fitting the models are shown for both the sequences $\{s_n\}$ and $\{x_n\}$.

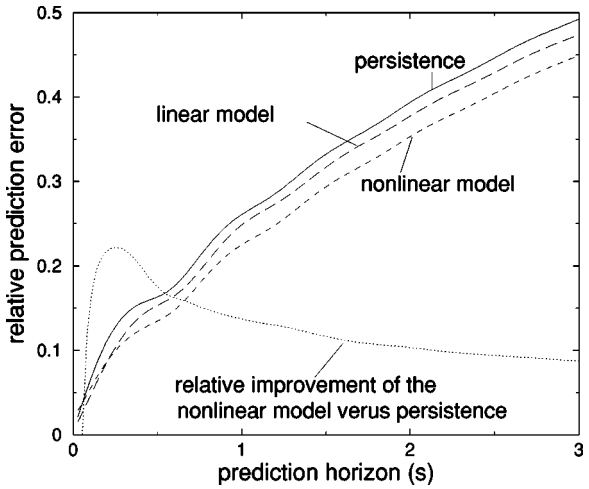

FIG. 16. Relative prediction error of the linear model (AR) and of the nonlinear model (zero) for the power output of a wind turbine versus prediction horizon.

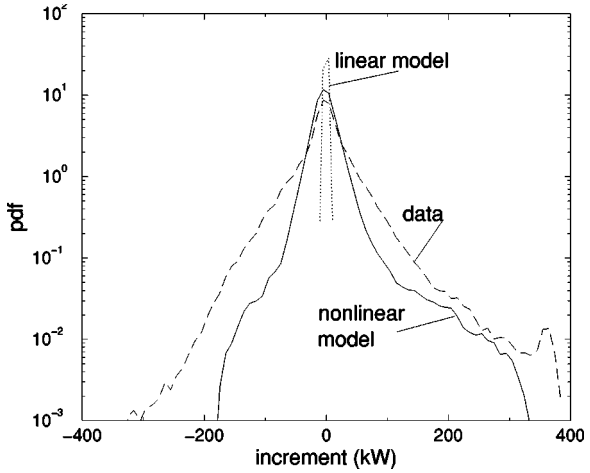

FIG. 17. PDF's of the increments of the data, of the linear model (AR), and of the nonlinear model (zero) for the power output for a prediction horizon of $1s$.

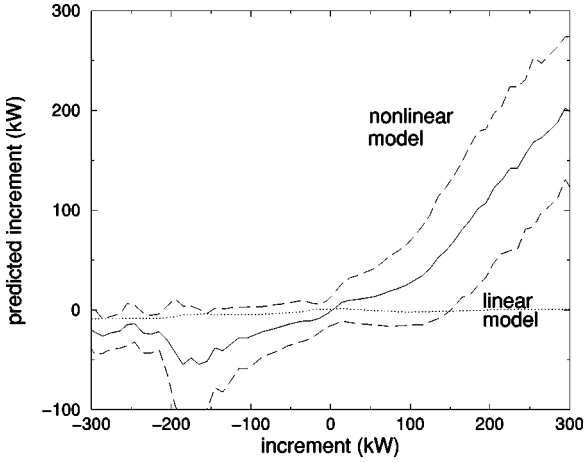

FIG. 18. Increments predicted by the nonlinear model (plus/minus standard deviation) and by the linear model versus the increments of the data.

For the direct output $\{x_n\}$ of the linear process the AR model gives slightly better predictions than the local model due to its higher statistical robustness. However, if the output of the linear system is nonlinearly transformed a nonlinear model can be significantly better.

As an example of practical importance we consider now the power output of a wind turbine, which behaves as the third power of the wind speed. Consequently, if one considers the longitudinal component of the velocity of the atmosphere v_l as the independent variable of a dynamical system the power output P would be a transformation of that variable through measurement function $P=v_l^3$. The actual measurement function is more complicated, however, due to cutoffs at a minimal and a maximal velocity and because of additional technical details of a wind turbine. Linear correlations in the velocity signal are transformed into nonlinear ones by the action of the nonlinear measurement function. This on an average leads to improved predictability of the power signal if one uses a nonlinear model despite the fact that the mean prediction error of the velocity signal itself does not decrease by the use of a nonlinear scheme. For a time series of the power output of a wind turbine, the relative forecast errors $\varepsilon_{model}/\sigma$ versus the prediction horizon are shown in Fig. 16 for the linear as well as for the nonlinear models. Also shown is the prediction error using the last value as prediction for the next one (persistence). The improvement using the nonlinear scheme amounts on an average upto 10% of the prediction error of the linear model. More important than this averaged improvement is this behavior of the nonlinear model when large fluctuations occur in the time series. For this we show the PDF of the increments of the data, of the nonlinearly predicted increments, and the PDF of the increments predicted by the AR model in Fig. 17 for a prediction horizon of $1s$. Whereas the AR model is unable to capture the large fluctuations, the nonlinear scheme almost resembles the data. Finally, we want to show that the large fluctuations predicted by the nonlinear scheme really occur correlated with the increments of the actual data. In Fig. 18 the average increment predicted by either of the models (plus/minus the standard deviation for the nonlinear model) is shown versus the actual increments of the data. An optimal predictor would correspond to the diagonal in this figure. The graphs show an asymmetry that is due to the fact that the power output has an upper cutoff. Therefore, a decrease of the power data is often impossible to predict, since it is not preceded by a typical pattern of the time series.

VII. CONCLUDING REMARKS

We have discussed the application of a locally constant predictor in a reconstructed phase space to stochastic data. In contrast to previous work where enhanced predictability of time series data by such a scheme was interpreted as a signature of determinism, we show here that this modeling scheme represents an empirical Markov model for the data. Despite the fact that scalar time series typically does not represent a Markov process, the approximation is rather good in many applications. We have compared the practical issues of modeling stochastic data to the modeling of deterministic data, where, e.g., one surprising result is the need of much shorter values of the time lag τ in the embedding procedure. The width of the transition probability allows one to estimate the precision of the prediction, and the statistics of the fluctuations gives an estimate of the degree of nonlinearity in the data. We have applied this scheme with considerable success to field measurements with low predictability, namely, surface wind velocities.

[1] H.I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer, New York, 1996); H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 1997).
[2] S. Siegert, R. Friedrich, and J. Peinke, Phys. Lett. A **243**, 275 (1998).
[3] M. Casdagli, J. R. Stat. Soc. Ser. B. Methodol. **54**, 303 (1991).
[4] F. Paparella, A. Provenzale, L. A. Smith, C. Taricco, and R. Vio, Phys. Lett. A **235**, 233 (1997).
[5] F. Takens, *Detecting Strange Attractors in Turbulence*, Lecture Notes in Mathematics Vol. 898 (Springer, New York, 1981).
[6] T. Sauer, J. Yorke, and M. Casdagli, J. Stat. Phys. **65**, 579 (1991).
[7] J. D. Farmer and J. J. Sidorowich, Phys. Rev. Lett. **59**, 845 (1987).
[8] For example, G. E. Box and G. M. Jenkins, *Time Series Analysis* (Holden-Day, San Francisco, 1976).
[9] N. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier Science, Amsterdam, 1992).
[10] P. Gaspard and X.-J. Wang, Phys. Rep. **235**, 291 (1993); H. Kantz and E. Olbrich, Physica A **280**, 34 (2000).
[11] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, Phys. Rev. A **45**, 3403 (1992); R. Hegger and H. Kantz, Phys. Rev. E **60**, 4970 (1999).
[12] Th. Buzug and G. Pfister, Physica D **58**, 127 (1992).
[13] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, Physica D **73**, 82 (1994).
[14] A. M. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986).
[15] M. Ragwitz and H. Kantz, Europhys. Lett. **51**, 595 (2000).