Physica D 58 (1992) 50–76
North-Holland

PHYSICA D

# Identification and prediction of low dimensional dynamics

Leonard A. Smith [1,2]
*Department of Engineering, University of Warwick, Coventry CV4 7AL, UK*

Received 14 November 1991
Revised 15 March 1992
Accepted 15 March 1992

This contribution focuses upon extracting information from dynamic reconstructions of experimental time series data. In addition to the problem of distinguishing between deterministic dynamics and stochastic dynamics, applied questions, such as the detection of parametric drift, are addressed. Nonlinear prediction and dimension algorithms are applied to geophysical laboratory data, and the significance of these results is established by comparison with results from similar surrogate series, generated so as not to contain the property of interest. A global nonlinear predictor is introduced which attempts to correct systematic bias due to the inhomogeneous distribution of data common in strange attractors. Variations in the quality of predictions with location in phase space are examined in order to estimate the uncertainty in a forecast at the time it is made. Finally, the application of these methods to truly stochastic systems is discussed and the distinction between deterministic, stochastic, and low dimensional dynamics is considered.

## 1. Introduction

It is now generally recognized that complex dynamical behavior is not restricted to systems with many active degrees of freedom, and examples of low dimensional nonlinear systems with complex and apparently unpredictable behavior are commonly cited [1–5]. One imagines that Laplace would have no difficulty with chaos, for given the *exact* state of the universe, prediction of a modern chaotic future is no more difficult than Newton's laws. In Laplace's words, for an intellect "vast enough to submit this information to Analysis, ...nothing would remain uncertain, and the future, as well as the past, would lay before its eyes" [6]. Yet few would dispute that there exist many systems which are, in fact, not *deterministic* within any known physical framework. A recurring theme in time series analysis is the attempt to characterize these two types of systems in order to distinguish determinism and indeterminism, chaos and stochasticity. In many cases of interest, this cannot be resolved definitely and we shall focus on a simpler question: given a particular set of observations (and current technological constraints), can we detect low dimensional dynamics in a time series?

This question will be directed either at phenomena outside the lab or at experiments designed to investigate such phenomena. There is no question that numerical experiments have taught us much about the nature of chaos. A more interesting question now appears to be what chaos can teach us about Nature. The particular experiment discussed here investigates dynamical processes related to the motion of planetary atmospheres, and provides an instance of the reoccurring attempt to distinguish "red" noise from nonlinear determinism. This distinction is central in determining the direction of future research; the best model for an atmosphere which amplifies small scale stochas-

[1] Present address: Mathematics Institute, University of Oxford, Oxford, OX1 3LB, UK.
[2] E-mail address: lenny@uk.ac.warwick.eng.

tic disturbances differs from an optimal model of the complex deterministic interaction of a reasonable number of modes. Long term prediction of the first is possible only in a statistical sense, the present state does not define the future state after some (nonlinear) decorrelation time. Prediction in the second case is difficult, but possible in principle as the information needed is contained in the state of the system. In this paper, we discuss techniques for this test and report initial results indicating that there is low dimensional behavior in some systems (including the experiment) and not in others. The analysis of such systems is notoriously difficult; what we desire are tests for low dimensional behavior which are reliable, in the sense that they do not yield false positives.

There are now a large number of algorithms for detecting and quantifying low dimensional behavior and chaos. The known weaknesses of individual tests may be addressed through the analysis of *surrogate data*: non-deterministic time series constructed to be similar in appearance to the original data. Such an analysis aims to establish what aspect of the data set an algorithm is quantifying, by determining whether original data can be reliably distinguished from an ensemble of surrogates, when each data set is processed in precisely the same way. The construction of these surrogate data sets is discussed in section 2 (also see [7–9]). The usefulness of this test will depend on the quality of the generator of the surrogate sets, as a poor choice of surrogate generator will result in sets which are distinguished, not because of any underlying determinism, but by some other factor. Indeed we argue in section 6 that some stochastic series may be more predictable than surrogates generated with similar statistics.

Section 3 addresses the construction and evaluation of dynamic reconstructions from observational data, where the vector field is approximated in phase space. The desirability of this approach has been discussed previously (e.g. [10,11]). We note that by using additional information about the macroscopic state of the system, more useful dynamical reconstructions can be obtained from the same data set(s). Once a dynamic reconstruction is in hand, a variety of other questions may be asked. The existence of a good reconstruction provides an estimate of a minimal embedding dimension for the system; clearly if one has a six-dimensional flow by which the observed dynamics are well described, then a six-dimensional embedding is a practical one. In addition, many characteristics of the system can be estimated from the reconstructed flow much more easily than from the raw data directly, for example the spectrum of unstable periodic orbits, or Lyapunov exponents. As these quantities are well defined for a given reconstruction, one must address the question of how quickly the properties of the flow approach those of the underlying system. In the case of unstable periodic orbits, this can be very rapid [12].

Dynamic reconstructions can also clarify experimental uncertainties. In section 4, we analyze time series from a thermally stressed rotating fluid annulus [13]. Comparison with surrogate signals demonstrates that the organization in the reconstructed phase space dynamics is greater than that arising from either autocorrelation or simple advection alone. We also show how dynamic reconstructions offer a natural method for the detection of slow parametric drift. In addition, one may use the flow to make predictions providing a direct test of determinism. We stress the importance of what is predicted and, in general, of which aspect of the time series is reconstructed. The quality of predictions (the difference between the predicted and observed values) may vary with time due to differences in the volatility of different states of the system, variations in the quality of the predictor, or errors in observation. We demonstrate a method of estimating the expected uncertainty in a given prediction, and discuss how to distinguish between these various causes. Additional evidence that the time series are in fact

low dimensional is given in section 5, where we apply the Grassberger–Procaccia Algorithm (GPA) [14] to analyze the geometric structure of the experimental data sets and establish the significance of the results through comparison with those from surrogate data sets.

Finally, in section 6, the results of applying these prediction techniques to stochastic series is considered. Laplacian determinism requires that, in the limit of perfect initial data, the future of the system is uniquely defined, so the systems considered in this section are not deterministic in this sense. There is, however, either a low dimensional or a deterministic component in their evolution, due to which many stationary stochastic systems will appear deterministic relative to some surrogate series. We discuss some of the implications this holds for distinguishing between low dimensional determinism and stochasticity from time series.

## 2. The data sets: observed and manufactured

### 2.1. Experiments

The primary data sets considered in this paper come from the geophysically inspired experiments reported by Read et al. [13,15]. These experiments were performed within a fluid filled, rotating annulus with thermally conducting side walls and insulating boundaries top and bottom. A temperature difference was maintained between the inner (cooler) and outer side walls providing an infinite dimensional simulation of the mid-latitude circulation of the Earth's atmosphere. The temperature in the fluid was measured by an array of 32 thermocouples, uniformly distributed in azimuth at mid-height and mid-radius. By monitoring the flow rate (volume) and temperature of the coolant water, simultaneous measurements of the total heat transport through the inner boundary were obtained.

Of the many reported results, two realizations are considered here. They correspond to the temperature series $b$ and $d$ of table 1 of ref. [13] and are shown in figs. 1a and 1c. The heat flux differs from the temperature series as it is averaged around the entire annulus and thus does not display the roughly periodic structure seen in the local temperature probe; this structure is due to the advection of (an evolving) wave pattern around the annulus. Fourier spectra of these series are given by Read et al. [13]. Both Read et al. [13] and R. Smith [16] conclude that these time series are low dimensional; series $b$ coming from a strange attractor with a correlation dimension, $d_2 \approx 3$, while series $d$ with $d_2 \approx 2$, is considered to reflect a two-torus [13,15].

The isolation of an experiment from external forces is a major concern of experimentalists. To obtain the long time series for the rotating annulus, experimental runs of 20 hours were required. One may ask whether the environment has been sufficiently isolated, for example from diurnal temperature variations, so that no systematic parameter drift has occurred during the experiment. Might not an evolving three-torus present a geometric structure with properties similar to those observed? A method of detecting slow parametric drift with dynamic reconstructions is introduced in the next section. First we discuss the construction and use of surrogate data sets.

### 2.2. Surrogate data sets

A common objection to the dynamical systems analysis of data from poorly understood systems is that the significance of a given result is rarely established [17–21]. This objection can be addressed directly by considering a class of non-deterministic surrogate signals. The significance of a result is then established by comparing it with the outcome of the same test applied to these surrogate data sets.

The choice of surrogate signals will also depend on the known weaknesses of the algorithm

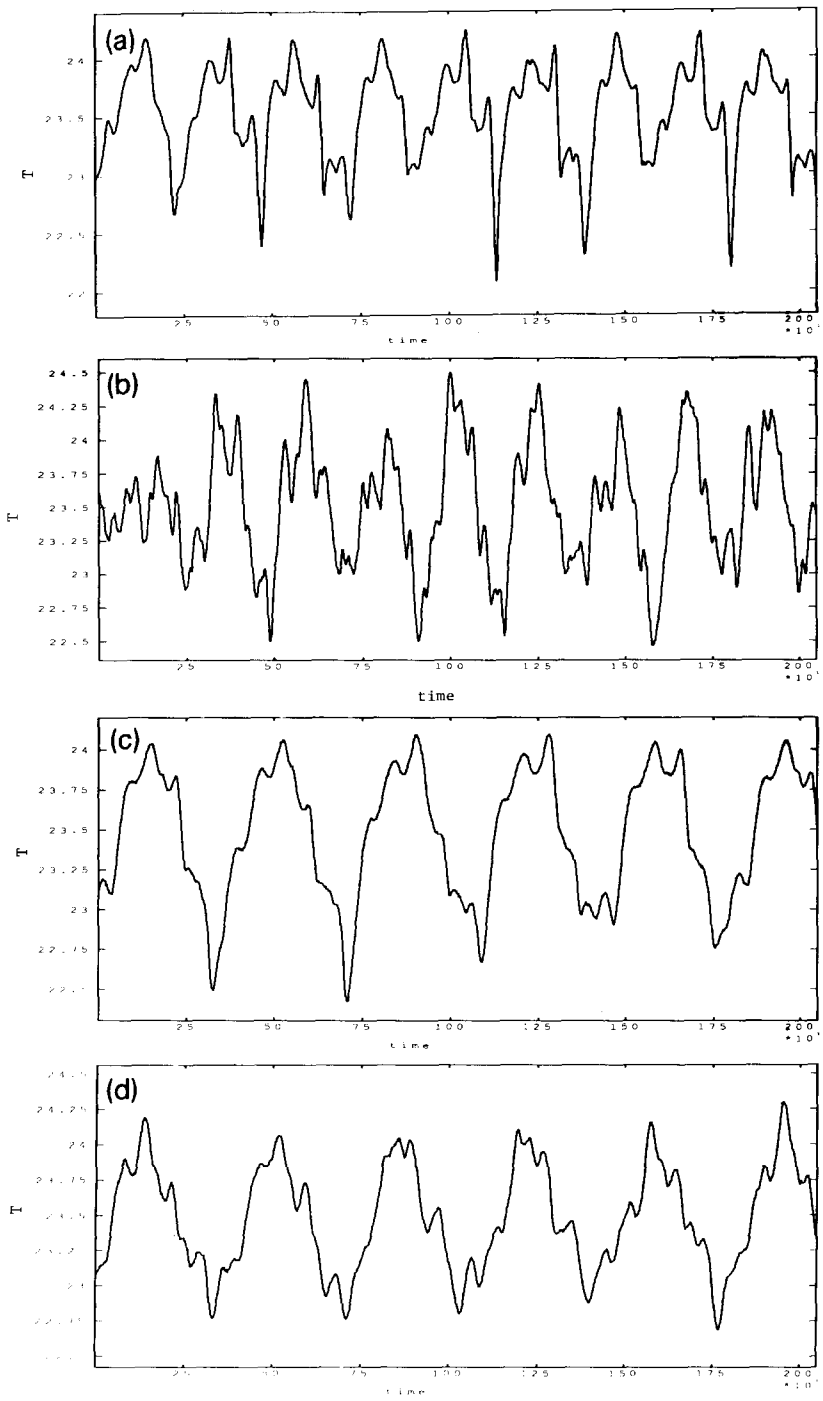


Fig. 1. A sample of the measurements of temperature, $T$, from the rotating annulus experiment and surrogate signals with the same autocorrelation function. (a) Time series $b$, (b) surrogate series for $b$, (c) time series $k$, (d) surrogate series for $k$.

employed to analyze the data. While any comparison with surrogate signals can quantify differences, only careful choice of the *surrogate generator* allows qualitatively new information. The goal is to generate signals with similar properties with respect to this weakness, but not containing "the physics" of the original signal. A result is significant with respect to the weakness tested if the algorithm can distinguish the true signal from the surrogates. In cases where a quantitative result (or distribution) is produced for both observed and surrogate signals, one may estimate the probability that observed value would occur by chance in an ensemble of surrogate realizations, and thereby evaluate the null hypothesis that the observed signal was a realization of the surrogate generator. Note that the second formulation is more difficult than the qualitative comparison of results, in that it requires the algorithm to converge for the surrogate signals. This poses a problem for dimension estimates via the GPA[#1].

Hypothesis testing and model evaluation with surrogate data has a long history in statistics [22,23]. There are often two conflicting motivations in choosing surrogates: ease of analysis and similarity to the observed data. One of the most common surrogate series is independent, uniformly distributed white noise which has the advantage that the expected distributions can often be calculated analytically. Most time series, however, are not uniformly distributed, neither are consecutive observations independent; at reasonable sampling rates the series should display some structure. A simple IID surrogate generator provides the correct distribution by simply shuffling the data, while series longer than the observed signal may be obtained by randomly sampling the observed distribution.

Temporal correlations are reflected in the Fourier spectrum of the series. "White" series have a flat spectrum, while those with less power at high frequencies are called "red". The methods of the previous paragraph destroy temporal correlations in the original data. Perhaps the simplest method to amend this is to produce the time series where the next observation is chosen from a distribution determined by the current state. For digital data, the amount of data required to estimate this conditional distribution will depend on the resolution of the analog to digital conversion. In this paper we will be concerned primarily with yet another surrogate generator which preserves the autocorrelation function of the original data.

The point here is to show that there are a variety of methods available for constructing surrogate series and note that signals may be indistinguishable from one set, and not another. The insight gained from testing surrogate signals depends on the particular surrogate generator(s) adopted.

When searching for low dimensional deterministic dynamics, an alternate approach for generating surrogates is to use a simple stochastic model of the system. While care must be taken not to overfit the model to the data (e.g. to construct epicycles), this approach may be particularly useful in evaluating dimension calculations from short, highly structured series. Used by Grassberger when considering climate data [18,20], this approach is discussed below for the sunspot data.

Both predictability and correlation dimension estimates may be biased through autocorrelated signals. To test if a given result is significant with respect to signals with the same autocorrelation function the following surrogate generator (suggested by Osborne et al. [24]) may be used : first compute the Fourier transform of the original signal, then compute a set of random phases, and finally invert the transform using the orig-

[#1] In their discussion of the method of surrogate data, Theiler et al. [7] point out that, strictly speaking, an algorithm need not converge for the null hypothesis to be rejected. While this is true, it is important to note the difference between distinguishing the observed series from the surrogates and estimating a statistic (e.g. a dimension) from the observed series which describes the dynamics.

inal amplitudes and a particular set of random phases to generate a particular surrogate series. Since both the original and surrogate series will have the same Fourier amplitude spectrum, their autocorrelation functions will be identical. An ensemble of surrogate signals may then be subjected to exactly the same analysis as the original data; surrogates for the annulus data generated in this way are shown in figs. 1b and 1d. Note that we are concerned with relatively long series here, where many linear decorrelation times are available, so that calculation of the Fourier transform is not a problem. Signals from this FT surrogate generator will be used to demonstrate the significance of the dynamic reconstructions in section 4 and correlation integral calculations in section 5.

Note that it is not necessary for a surrogate generator to completely destroy the phase coherence in a signal, and in some cases, it is not desirable. Consider a signal containing some understood frequencies, such as a diurnal cycle in a temperature record. Complete phase randomization will distort the daily cycles obvious in the data and the true signal may be distinguished from the surrogates for this reason alone. A stronger result is obtained (in the sense that a more relevant null hypothesis is rejected) if the observed signal can be distinguished from an ensemble of surrogates which also contain a diurnal cycle. Such surrogate series may be generated by retaining a subset of the Fourier phases unaltered and randomizing the remainder. (Note that the fine structure of a fractal attractor will be destroyed simply by randomizing only the phases corresponding either to high frequencies or to those frequencies with relatively low power thus retaining some macroscopic structure.)

For some data sets, reconstructions *cannot* be distinguished from those of the surrogate generators discussed above; but for many interesting systems the construction of good surrogates will require a more detailed examination of the system. The underlying desire is often not to identify either nonlinearity or chaos, but low dimensional, deterministic dynamics. This may be pursued by employing the best stochastic model available for a given process as a source of surrogate signals. One must balance overfitting the model to the data (allowing unduly complex models) against setting up "straw man" surrogates. An excellent example is provided by the annual mean sunspot numbers. The basic asymmetries of the sunspot number (it is strictly positive, increases more rapidly than it decays, etc.) and the presence of events like the Maunder minimum [25] make the simple FT surrogates inappropriate for this series. It is nonlinear by inspection. Analysis calls for either a modification of the data set (e.g. Spiegel and Wolf [26]) or an improved surrogate generator. By modifying the dynamics of a linear ARMA model, Barnes et al. [27] have constructed a nonlinear stochastic model of sunspot number, which, fortuitously, produces Maunder minima. Treating this model as a surrogate generator, the significance of correlation integral results for the sunspot series is examined by Weiss [28], along with a discussion of solar aperiodicity in the context of nonlinear dynamical systems.

In summary, different surrogates will test different effects. The better the surrogate generator, the more relevant the class of signals that the data set (and by implication the system) can be distinguished from. Even then, the comparison with the best surrogate signals provides only a necessary condition for the detection of low dimensional dynamics; it is not sufficient. As we are showing what the signal is not, this approach cannot establish what the system is; in this sense proving moderate dimensional chaotic dynamics by this method is similar to proving true randomness, one only knows when one cannot do it.

## 3. Reconstructions

### 3.1. Static reconstructions

The methods of nonlinear dynamical systems theory discussed here require time series to be *reconstructed* in a geometrical framework [29]. Consider a single signal measured as a function of time, $s(t)$. Once the signal is recorded digitally in discrete time we have

$$s_i = s(i\tau_s), \quad i = 1,2,\ldots,n_s, \tag{3.1}$$

where $\tau_s$ is the sampling time (and $s_i$ is digitized to one of a finite number of values).

Consider a deterministic system with phase space dimension $M_s$. A trajectory, $\boldsymbol{x}(t)$, of this system is reconstructed in $M$ dimensions from a time series of a single observable, s(t), by the method of delays [30,31] to yield

$$\boldsymbol{x}(t) = (s(t), s(t-\tau_d), \ldots, s(t-(M-1)\tau_d)), \tag{3.2}$$

where $\tau_d$ is called the delay time. The delay time need not equal $\tau_s$ (although it must, of course, be an integer multiple of $\tau_s$). In fact the $M-1$ delays used in defining $\boldsymbol{x}(t)$ need not be equal, although they will be treated as such here. Methods for choosing $\tau_d$ vary (see e.g. [32–34] ); it is usually related to the decay of information in the signal with time, either from linear autocorrelation time ($\tau_{auto}$) or more general methods [35]. When constructing nonlinear predictors, the delay may be chosen to optimize the predictor as demonstrated in section 4. Breeden and Packard [36] discuss the case of time series sampled nonuniformly in time.

The arguments which follow do not depend on the use of this method of delays. We have achieved similar results with singular value decomposition (SVD) reconstructions (see [37,38,35]). Multi-variate series also work well, often with significantly shorter time series in terms of the total duration of the "experiment". This is easily understood as multivariate probes can distinguish states in phase space which appear similar to univariate probes due to projection effects. When working with finite data sets, the use of multi-probe data can add crucial information on the state of the system, either by directly characterizing macroscopic patterns or through direct (and much more efficient) evaluation of mode amplitudes. We return to this issue in section 7. Typically, each series is transformed to have zero mean and unit standard deviation, however the standard deviation can be varied to change the weighting between different variables in the interpolation scheme. When we are concerned with predicting a fixed period in the future, we consider a third time scale, $\tau_p$, the prediction time. Each point $\boldsymbol{x}(t)$ on the trajectory has a scalar image $s(t+\tau_p)$ and we wish to construct a predictor to determine this image for any $\boldsymbol{x}$. In other applications the time of the prediction is determined through some geometric constraint. For example, when working on a surface of section the time of the next crossing must be predicted as well as its location. Alternatively, when predicting recurrence times, the main goal of the analysis is to determine $\tau_p$.

### 3.2. Dynamic reconstructions

Recently there has been much interest in predicting nonlinear deterministic systems and a wide variety of approaches and algorithms have been proposed (see [39–43,33,44–48]). While these systems differ in detail they all attempt the same task, since in the context of deterministic analysis, prediction in time becomes a question of interpolation in phase space. To predict a deterministic system given a description of its current state, one is faced with the basic problem of interpolating the future behavior based on a sample of the "nearby" points. Like all interpolation problems, success depends on having a sufficient number of nearby points to satisfy the smoothness assumptions of the chosen algorithm. In the presence of noise, this require-

ment is increased so that the variations due to the noise may be, in some sense, averaged out. We shall use a global radial basis function predictor [43] and account for noise by fitting the predictor to the data in a least squares sense [41]. Further, we can account for the systematic bias introduced by extreme inhomogeneities in the distribution by adjusting the weighting scheme. This method provides a smooth flow over the entire region of the reconstruction (which may or may not be an "attractor"). As we are interested in finding global structures (e.g. periodic orbits), this smoothness, lost in most local methods, is important (see, however [48]).

When the underlying dynamics are not known in advance, we both construct and evaluate a dynamic reconstruction from the same data set. To do so, the data set is typically divided into two sections of unequal length: the *learning set* consisting of $n_L$ points from which a reconstruction is derived and the *test set* on which various reconstructions are evaluated. It is crucial that this distinction should be maintained for out-of-sample evaluation of the predictor. That is not to say that statistics from the ability to fit the learning set are not of interest, but that these two types of statistics measure essentially different things. Statistics generated within the learning set reflect how well the data can be forced into a given mold and may be useful, for example, for internal consistency checks and locating outliers. Those generated from the test set reflect how well the predictor generalizes from the learning set to new data. Only the latter are of use for cross-validation. Predictor "error" in the two sets is a very different quantity. For example, with the exact radial basis function predictor described below, the in-sample predictor error can be made zero for almost any data set.

The predictor is based upon a set of $n_c$ centers in an $M$-dimensional space:

$$\boldsymbol{x}_j^c,\ j = 1, 2, \ldots, n_c; \quad \boldsymbol{x}_j^c \in \mathbb{R}^M.$$

The choice of centers will be discussed below, but, in the simplest case, each center might correspond to a data point in the learning set. Associated with each of the $n_L$ points, $\boldsymbol{x}_i$, in the learning set is an observation, $s_i$; $s_i$ may be a future value of the system, a simultaneous value of another state variable, or even a past observation thought to contain noise [49]. In general, the problem is to construct a predictor (or map), $F(\boldsymbol{x})$: $\mathbb{R}^M \to \mathbb{R}^1$ which estimates $s$ for any $\boldsymbol{x}$. We will consider $F(\boldsymbol{x})$ of the form

$$F(\boldsymbol{x}) = \sum_{j=1}^{n_c} \lambda_j \phi(||\boldsymbol{x} - \boldsymbol{x}_j^c||), \tag{3.3}$$

where $\phi(r)$ are radial basis functions and the $\lambda_j$ are constants which are determined by observations in the learning set:

$$F(\boldsymbol{x}_i) \approx s_i. \tag{3.4}$$

Determining the $\lambda_j$ corresponds to the solution of the (linear) problem

$$\boldsymbol{b} = \mathbf{A}\boldsymbol{\lambda}, \tag{3.5}$$

where $\boldsymbol{\lambda}$ is a vector of length $n_c$ whose $j$th component is $\lambda_j$ and $\mathbf{A}$ and $\boldsymbol{b}$ are given by

$$A_{ij} = \omega_i \phi(||\boldsymbol{x}_i - \boldsymbol{x}_j^c||) \tag{3.6}$$

and

$$b_i = \omega_i s_i, \tag{3.7}$$

where $i = 1, \ldots, n_L$ and $j = 1, \ldots, n_c$. Traditionally, the weights $\omega_i$ reflect the varying confidence associated with the $i$th observation.

Casdagli [43] was the first to solve this problem in the context of predicting chaotic systems, considering the special case of exact interpolation where centers are chosen from the learning set and only their images are considered in eq. (3.4). In this case the interpolation on the centers is exact; the matrix $\boldsymbol{A}$ is square and the solution for $\boldsymbol{\lambda}$ depends on the $\boldsymbol{A}$ being nonsingular. This is guaranteed when the $\boldsymbol{x}_j^c$ are distinct and $\phi(r)$ is a radial basis function [50,51].

Typical radial basis functions are $\phi(r) = r$, $r^3$, $\sqrt{r^2 + c}$, $1/\sqrt{r^2 + c}$, and $e^{-r^2/c}$ where $c$ denotes a constant often based on the average distance between neighboring centers.

Casdagli demonstrated the effectiveness of this approach and showed that interpolation in both time and parameter space was possible. There are, however, several drawbacks when applying it to noisy data; in this form, the interpolation fits the centers exactly, and no information from the points in the learning set not chosen as centers, is used. It is desirable to use this information, and important to avoid overfitting or fitting the noise in data exactly (especially since when making the choice of centers, one may tend to select outliers). Even with numerical systems, computational constraints limit the number of centers used.

In order to include the information available from the learning set, Broomhead and Lowe [41] solved this problem in a least squares sense and studied the behavior of the logistic map. For the least squares case, the entire learning set is included in eq. (3.5) ($\boldsymbol{b}$ is of length $n_L > n_c$), but a smaller number of centers are employed, and thus $\mathbf{A}$ is not square. Further there is no need to know the images of the centers, so they need not correspond to observations from the series. We seek a $\lambda$ which minimizes $\chi^2 = ||\boldsymbol{b} - \mathbf{A}\lambda||^2$

Choosing the solution which also minimizes $||\lambda||^2$ corresponds to

$$\lambda = \mathbf{A}^+ \boldsymbol{b}, \tag{3.8}$$

where $\mathbf{A}^+$ is the Moore–Penrose pseudo-inverse of $\mathbf{A}$. Efficient methods of calculating $\mathbf{A}^+$ via singular value decomposition are discussed in [23]. Noting that the guaranteed solubility of the original system is lost in this generalization, Broomhead and Lowe [41] quantified the effect of increasing the number of centers and considered this modeling approach as a special case of a neural network with a guaranteed learning rule.

We continue this approach, introducing the weights $w_i$, and investigate the sensitivity of the solution to the choice of centers and the effects of observational noise. Given the inhomogeneous (often singular) distribution of data on a chaotic attractor (see e.g. [52]), these weights can also be used to provide a more uniform prediction error across the attractor by reducing the importance of the dense regions of the reconstruction.

One common objection to the use of radial basis function interpolation arises from the large number of free parameters employed, one per center in the original formulation. The least squares formulation addresses this question of parsimony. When determining the coefficients of eq. (3.5), we perform a SVD of the matrix $\mathbf{A}$. In doing so, a tolerance is set as to the smallest meaningful value an eigenvalue can take [23]. Values below this threshold are considered superfluous and suppressed (set equal to zero). This prevents "extra" degrees of freedom in the model from overfitting "noise fluctuations". As the threshold is raised, the estimated uncertainty in the modeling parameters (the $\lambda_i$) decreases dramatically, with little effect on the $\chi^2$ or the in-sample predictor error. This might be taken to mean that higher tolerances were preferred to avoid fitting noise in the learning set. Out-of-sample prediction error statistics often conflict with this interpretation: there are examples for which, although the estimated uncertainty of the $\lambda_i$ is greater, low threshold models consistently yield better out-of-sample prediction statistics. This implies that the model is *not* fitting noise in the learning set. It is here that a difference between radial basis functions is observed: for a fixed tolerance and an identical choice of centers, models with $\phi(r) = e^{-r^2/c}$ consistently use fewer degrees of freedom than those with $\phi(r) = r^3$ or $\phi(r) = r$.

Predictions more than one sampling time into the future can be made either by *direct forecasts*, constructing a predictor for this time scale, or through *iterative forecasts* repeatedly using a predictor which forecasts a smaller time step. Farmer and Sidorowich [49] conclude that iterative forecasts are generally better than direct

forecasts, but also present an experimental example where the reverse is observed. Stokbro [48] also compares the two as a function of the forecast time and comments on the choice for the basic time step. Direct forecasts are used in this paper.

As noted above, minimization of $||\boldsymbol{b} - \mathbf{A}\boldsymbol{\lambda}||^2$ with all $\omega_i$ equal results in a bias in favor of the frequently visited regions of phase space. In order to achieve a good reproduction throughout phase space, such weighting is not desirable, as argued in the next section. One method to account for this is to partition the phase space and allow only a specified number of points in each partition. In the presence of noise it is preferable to retain all the observations and adjust the $\omega_i$ so that the partitions are more equally weighted.

*3.2.1. The choice of reconstruction centers*

We now consider the question of how to determine the centers. Four approaches to this question are to choose the centers either

(i) randomly (or uniformly) in the region of phase space explored by the data,

(ii) with respect to the probability density (measure) on the reconstruction,

(iii) spatially uniform on the reconstruction,

(iv) with respect to the local divergence on the reconstruction.

Methods (iii) and (iv) appear the most robust in terms of providing good reconstructions with a limited number of centers, but unfortunately, these results appear to be system dependent. There are several shortcomings in methods (i) and (ii) . Placing the centers uniformly in space works well when the system explores the entire region, as with the logistic map in one dimension. In higher dimensional spaces, there are often large lacunae into which the system does not venture; placing many centers in such gaps is counter-productive, at least when localized basis functions are employed [#2] . Numerical experiments indicate this is particularly relevant in cases where the underlying dynamics is not determined by a simple analytic formulation (contrast a true surface of section of a flow with that of an analytically defined map) perhaps due to the smoothness of the dynamics.

Centers may be placed uniformly with respect to the probability density on the reconstruction either by choosing them equally spaced in time or randomly sampling the series. This initially attractive idea often yields poor results. One reason for this can be understood in the case of flows where the speed in phase space varies from point to point. An ideal illustration of this effect is provided by the Duffing oscillator near homoclinic bifurcation (see [53,54]). In this case, a trajectory spends most of its time near the fixed point, while the centers "should" be distributed over the relatively rare excursions. With maps, inhomogeneities in the measure also result in a poor distribution of centers.

One method to distribute the centers uniformly on the reconstruction is simply to disallow centers closer than some nearest center distance $d_{\mathrm{nc}}$. This succeeds in the Duffing case and will, in general, avoid the accumulation of centers in the slow moving regions of the reconstruction which are relatively easy to predict. But that is the real point. Once the basic skeleton of the reconstruction is covered, it is reasonable to place additional centers in regions where the fine structure of the flow is greatest and where prediction is most difficult. Note that this need not correspond to the fine structure of the probability density or geometry, the fine structure here is in the vector field of the phase space flow, not that of the attractor. Also note that, while the location of additional centers allows the predictor to develop fine structure, this will not occur unless the data are weighted toward the recovery of that fine structure. Returning to the Duffing oscillator, we would like to combine (iii) and (iv) , covering the excursions and also the region about the unstable manifold near the origin so that the beginning of an ex-

[#2] I would like to thank James Theiler for pointing out this qualification.

cursion is predicted. Locations where the flow is contracting need not be sampled so densely.

The importance of these effects in a given reconstruction may be determined by dividing the reconstruction space into partitions and examining the errors made in each region. When the centers are distributed uniformly on the reconstruction, a straightforward way to partition the reconstruction is to classify each point according to the center to which it is nearest. This is now demonstrated for rotating annulus data.

## 4. Applications to laboratory data

We now apply the ideas of the last section to the rotating annulus data. Consider first a dynamic reconstruction of data set series $b$ built from a learning set consisting of 2K points ($1K = 2^{10}$) from the first 16K data points of this $50K$ point data set. The reduction from 16K to 2K was achieved by increasing the sampling time by a factor of 8, thus all time steps considered in reconstruction will be multiples of $8\ \tau_s$. A total of $n_c = 128$ centers were chosen such that no two were closer than a nearest center distance, $d_{nc}$; this was implemented by choosing an initial value of $d_{nc}$ large enough so that less than half the desired number of centers were found on the first pass through the learning set. The value of $d_{nc}$ was then decreased by a factor of 0.7 and the process repeated iteratively in order to avoid over-sampling any one segment of the learning set. In this case, the delay time $\tau_d = 4\ (8\tau_s)$ and an embedding dimension, $M = 5$, were also chosen taking into account the results of the correlation integral calculations of section 5. We shall refer to this model with $\phi(r) = e^{-r^2/c}$ as reconstruction A.

The initial results are presented in fig. 2. The three panels show (a) the observed (solid) and predicted (symbol) time series as a function of time, (b) the absolute value of the prediction error, and (c) the distance between the point being predicted and the nearest center to it, $d_{nc}$. The prediction time ($\tau_p = 18\ (8\tau_s)$) was chosen as twice $\tau_{auto}$ (the first zero of the linear autocorrelation function). Each of the predictions was made at this fixed distance into the future, the series shown is taken from the beginning of the test set and represents completely out-of-sample testing. In panel 2a, the prediction time is just over $\frac{1}{3}$ of the separation of the tick marks. Comparing the first 2 panels, it is observed that large errors often correlate with extreme values of the measured signal. Occasionally, there are episodes of poor predictions (not shown) which do not correspond to extreme values of the signal but do correspond to large values of $d_{nc}$; this implies that the trajectory is located in a region of phase space not explored during the learning set. Predictions in such regions are extrapolations and generally not reliable.

The effect of varying reconstruction parameters and choice of basis function is illustrated in fig. 3, which shows the cumulated predictor error profile, $P(\epsilon)$, for three different reconstructions. These graphs display the fraction of the learning set predicted to within a given error. For example, the solid line denoting the reconstruction of fig. 2 shows that half the learning set was predicted with $\log_2(\text{error}) < -3$ corresponding to an accuracy of approximately 6 bits. The right most (short-dashed line) corresponds to a similar reconstruction with $\tau_d = 8\tau_s$ (a factor of 4 shorter than reconstruction A). The corresponding predictions are about 0.5 bits worse; more so for small errors. $\tau_d$ was chosen to optimize this distribution, although for $\tau_d$ slightly greater than the chosen value the variation was small.

The long-dashed line in fig. 3 shows the distribution for a reconstruction similar to reconstruction A, but using $\phi(r) = r^3$; as was often observed, the exponential provided a slightly better fit. Although we shall not discuss the effect of different basis functions further, it is interesting to note that the two predictors tend to yield similar predictions across phase space. Indeed, they are in closer agreement with each other than with the observations. This is shown in fig. 4b

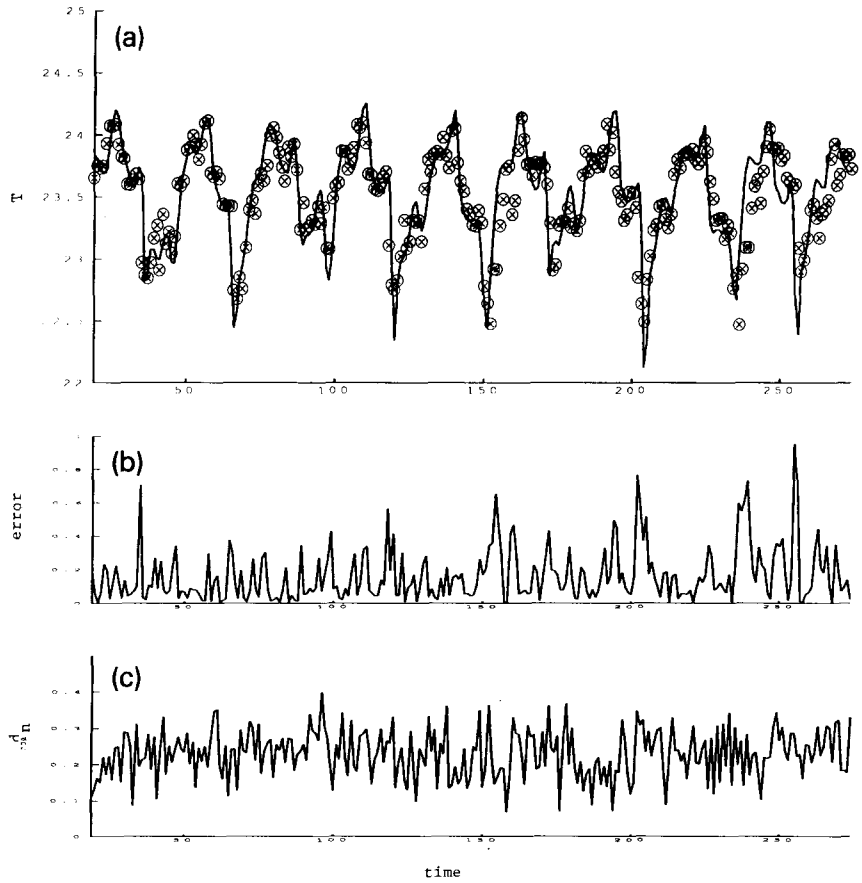


Fig. 2. An extract of the results of applying reconstruction A to time series *b*. (a) the observed (solid) and predicted (symbol) temperature values, (b) the error, and (c) nearest center distance for the point from which the prediction was made.

which is a scatter plot of the prediction of reconstruction A, where $\phi(r) = e^{-r^2/c}$, against that with $\phi(r) = r^3$. Panel 4a is a similar plot with reconstruction A against the observations. For small observed values, the two predictors remain in rough agreement although both are inaccurate; the inaccuracy results, in part, from the low weighting the least squares fit assigns to the less commonly observed values. By adjusting the the weights, $w_i$, we can force the distribution of errors to be more uniform over the reconstruction.

The cumulated predictor error profiles show a slow decay in predictability as $\tau_p$ increases comparable to Read's Lyapunov estimate of $1.79 \times 10^{-3}$ bits per second (or one bit per advection period). As these values are small, it may be argued that they are numerically zero and the system is not, in fact, chaotic. One alternative is the parametric drift mentioned above. We give evidence below that this is not the case. Of course, the significance of the Lyapunov exponent can be addressed directly via comparison with the distribution of values obtained from surrogate sets; where the Lyapunov exponent estimates from the surrogate signals are used to define the expected range of values to be considered as computationally equivalent to zero. It should also be noted that a different predictor was constructed for each of these prediction times (i.e. a direct predictor for each value of $\tau_p$). The decay of predictability with time in this instance may differ from the case of

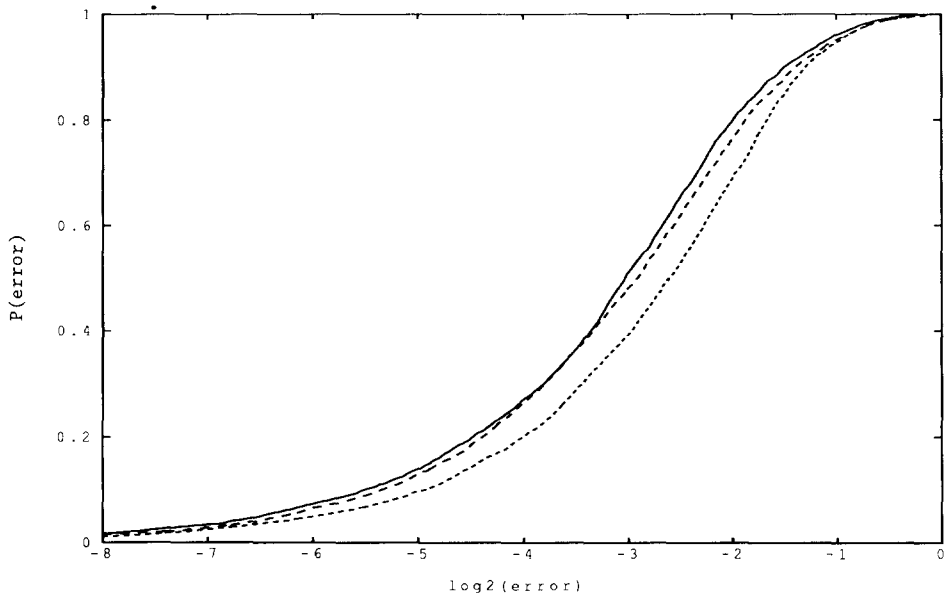


Fig. 3. Cumulated predictor error profile for a reconstruction with $n_c = 128$, $M = 5$, $\tau_p = 18$ $(8\tau_s)$ and $\phi(r) = e^{-r^2/c}$, $\tau_d = 4$ $(8\tau_s)$ (solid), $\phi(r) = r^3, \tau_d = 4$ $(8\tau_s)$ (long dashed), $\phi(r) = e^{-r^2/c}, \tau_d = 1$ $(8\tau_s)$ (short dashed). The horizontal axis is the base-2 logarithm of the error.

an iterated fixed-step predictor.

Rather than estimate the Lyapunov exponents of surrogate series, we investigate the significance of observing this level of predictability. In particular, whether predictions of similar accuracy would be found in other signals with the same autocorrelation function. To do so we construct surrogate series with a FT surrogate generator and consider the prediction of a reconstruction with parameters identical to those of reconstruction A above. The resulting cumulative error profiles are shown in fig. 5. Eight surrogate series, each with a different set of random phases was analyzed, the results shown have the lowest (best) average absolute predictor error. Considering the error to which 50% of the series is predicted, reconstruction A is almost one bit lower, implying that the prediction error is almost a factor of 2 less, easily distinguishing the observed data from the surrogate series.

Noting that the distributions do not appear to be Gaussian, we can reject the hypothesis that these two realizations either have the same mean (via the t-test) or were generated from the same distribution (via the Kolmogorov–Smirnov test) at well over the 0.99 confidence level. We wish to stress both the significance and limitations of this statement. The surrogate generator here preserved the autocorrelation function of the data set, and the radial basis function predictor easily distinguished five-dimensional reconstructions of these two signals. What we have really shown is that this 5D reconstruction is more coherent than that of these surrogate data sets. This is somewhat different from establishing that the data arise from a deterministic five-dimensional system. Further evidence that the data do in fact reflect low dimensional dynamics in provided by the correlation integral results.

As an additional test, we construct a "surrogate predictor" to determine whether the predictability of this signal is only due to the advection of a slowly evolving signal. Look again at the observational and the surrogate version of series *b* in fig. 1; the coherence between one "wave" and the next is stronger in the real signal than in the surrogate data. The surrogate predictor sim-

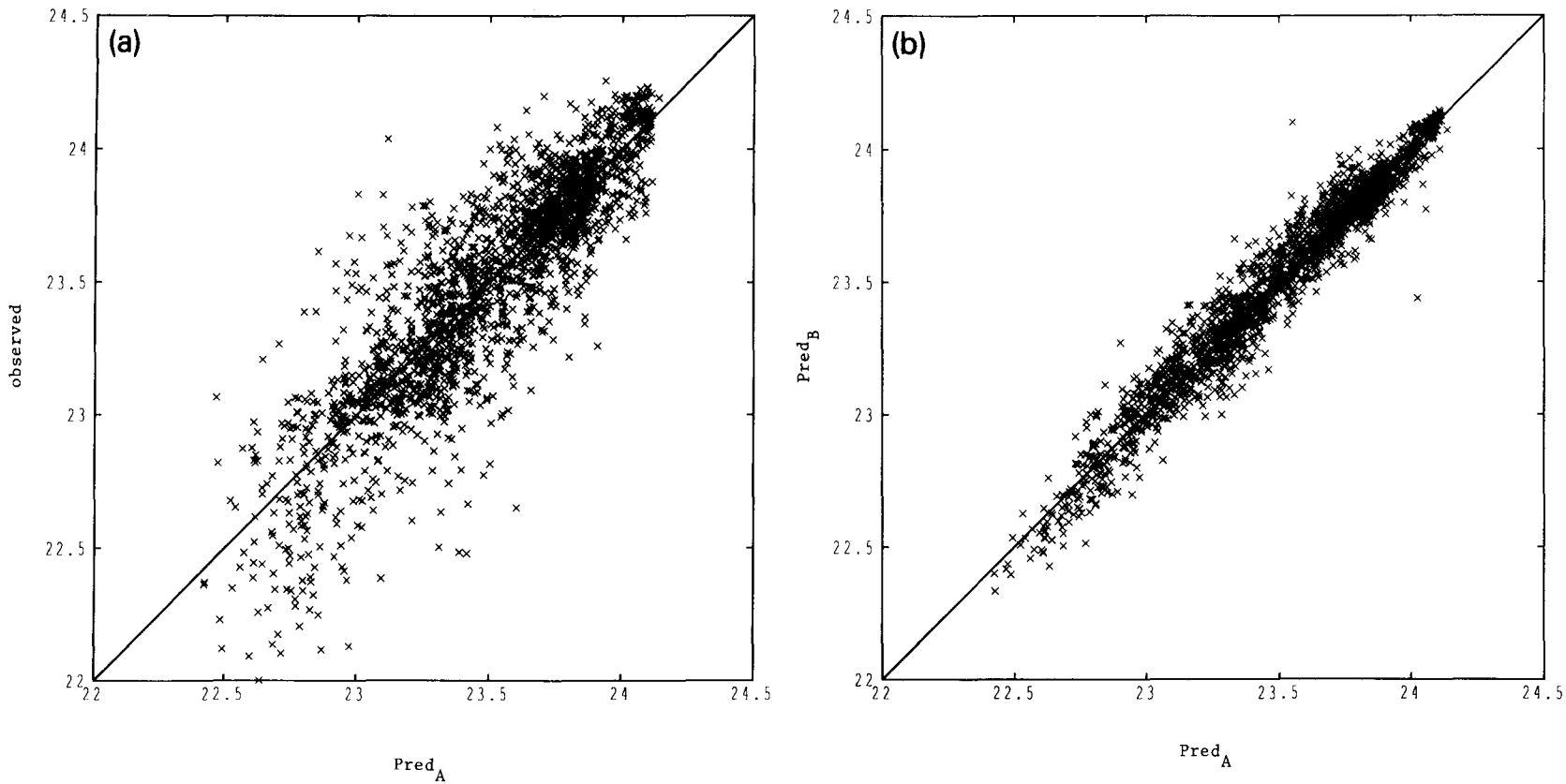


Fig. 4. Scatter plots of the predictions of reconstruction A (horizontal) against (a) the observed value and (b) the prediction of a similar reconstruction, $\text{Pred}_B$, using $\phi(r) = r^3$.

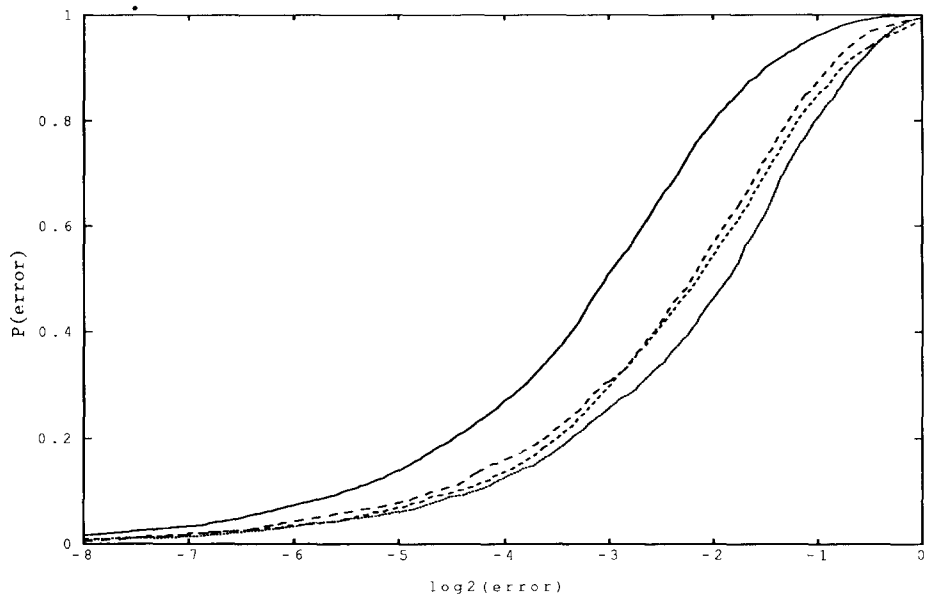


Fig. 5. Cumulated predictor error profile for reconstruction A on (a) observed (solid) and (b) surrogate (long dashed) series. And the period 1 surrogate predictor in (a) observed (short dashed) and (b) surrogate (dotted) series.

ply projects forward the last observed point an integer number of advection periods in the past. The cumulative error profile for this predictor is shown in fig. 5 for both the observed data set (short-dashed line) and the surrogate (the right most, fine dotted line). Although it is not as accurate as reconstruction A, this one dimensional predictor clearly differentiates between the true signal and these surrogates, demonstrating one limitation of the FT surrogate generator in this case. (This originates, in part, from the loss of phase coherence of the periodic advection signal in the FT surrogates, as noted above.) To establish that a system is chaotic through surrogate signals, we would have to reject all nonchaotic surrogates; this is clearly not feasible, and highlights the importance of the selection of surrogate signals.

The inhomogeneity of the spatial distribution of points in the series $b$ reconstruction is reflected in fig. 6 which shows a histogram of the number of times each center is nearest to the point from which a prediction is made in the test set. It is convenient to use this partition of the phase space by nearest center to examine the variation of predictor error with location as well. This is shown in fig. 7 which will be used to predict the error associated with each prediction of the time series below. Examining the distribution of errors in the learning set about individual centers provides examples where the data density is high and the average error is below the global average. Simultaneously, the error distribution about some other center with fewer associated points may be very broad. On the assumption that this is due to additional structure in the flow near the latter center, additional resolution should be placed in this region instead of the denser region where the flow is already well described.

### 4.1. *Predicting predictability*

We have seen that the predictability of a reconstruction varies with location, due to both the underlying dynamics of the system and the weighting scheme used in the construction of the predictor itself. We now quantify this variability and, in so doing, estimate the uncertainty associated with each individual prediction. This will

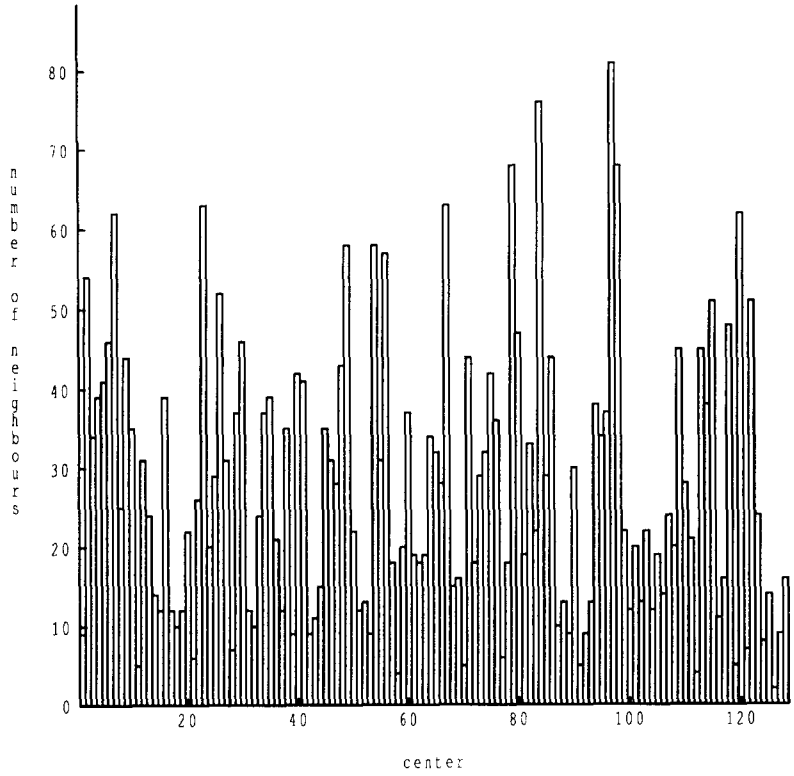


Fig. 6. Histogram of number of predictions made as classified by the nearest center to the point from which the prediction was made. The data is from the test set of reconstruction A.

allow us to estimate the error associated with a prediction at the time the prediction is made and thus forecast error bars as well as expected values.

When estimating the probable errors associated with each location in the reconstruction we again use the partition provided by the centers. Let $n_N(\boldsymbol{x})$ equal the index of the center nearest to the point $\boldsymbol{x}$, that is

$$n_N(\boldsymbol{x}) = j, \tag{4.1}$$

where $(||\boldsymbol{x}-\boldsymbol{x}_j^c||)$ is the minimum value of $(||\boldsymbol{x}-\boldsymbol{x}_k^c||)$ over all centers $k$.

At the beginning of the test set, initial estimates can be drawn from the histograms of the learning set. When very large quantities of data are available, one may estimate the mean and standard deviation error associated with each center (or even examine each distribution). We note that, in some examples, the distributions are far from Gaussian (e.g. bimodal) and the distribution of positive errors is very different from that of negative errors. In shorter sets where many centers may have only a few ($< 3$) tests, we have found it useful to define the average positive predictor error associated with the $j$th center as

$$E_j^+ = \frac{\sum_k e_k \delta_{j,n_N(\boldsymbol{x}_k)}}{\sum_k \delta_{j,n_N(\boldsymbol{x}_k)}}, \tag{4.2}$$

where $\delta$ is the Kronecker delta, $e_k$ is the error associated with the $k$th prediction (i.e. the predicted value minus the observed) and the sum is over all $k$ such that $e_k > 0$. $E_j^-$ is similarly defined for $e_k < 0$. These average positive and negative errors are then used as predicted error bars for future $\boldsymbol{x}$ such that $n_N(\boldsymbol{x}) = j$. Positive and negative errors are considered separately, so

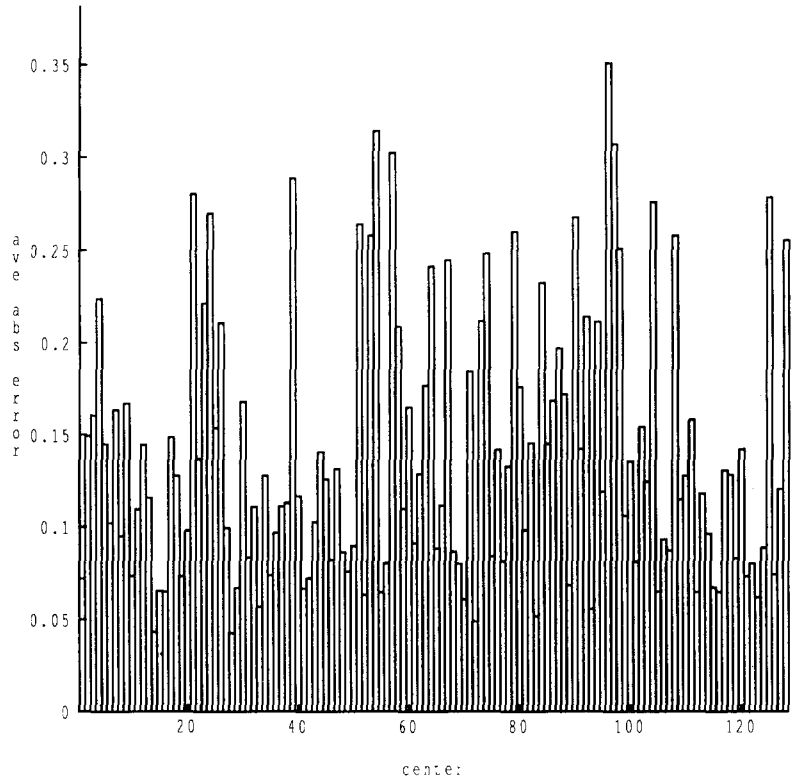


Fig. 7. Histogram of local error from the test set of reconstruction A classified according to the center nearest to the point predicted as in fig. 6.

asymmetries in the predictability are preserved.

Two short sequences taken from the annulus temperature series $b$ are shown in fig. 8 where the time scale is increased from that of fig. 2 to make individual predictions more clear. In this case, the predictions were made 18 ($8\tau_s$) steps ahead corresponding to almost two horizontal tick marks. Panel A shows a typical result; note that the expected error is often asymmetrically distributed about zero. This implies that the absolute value of the average prediction error could be reduced by adding a constant to each prediction, the value of which was dependent upon the nearest center, $n_N(\boldsymbol{x})$. This would improve the predictions but at the cost of a smooth predictor (local nonlinear predictors should provide even lower predictor error, and can function at lower data densities than local linear methods). Modifying the weighting scheme used in constructing the predictor provides an alternative global approach which preserves smoothness.

The more striking result is the reliability of the estimated error; predictions which lie in portions of the time series with sharp vertical displacements have large estimated errors, the slowly changing portions expected to be more predictable tend to have smaller estimated errors which are reflected in the observed error.

In addition to their practical value, these estimates can be used to identify regions of the reconstruction with greater instability and to distinguish outliers from variation due to this instability. The only instances of persistently misleading results noted thus far occur when the trajectory explores a portion of the phase space not visited in the learning set; this condition can often be identified by an increase in the nearest neighbor distance as noted above. Persistently

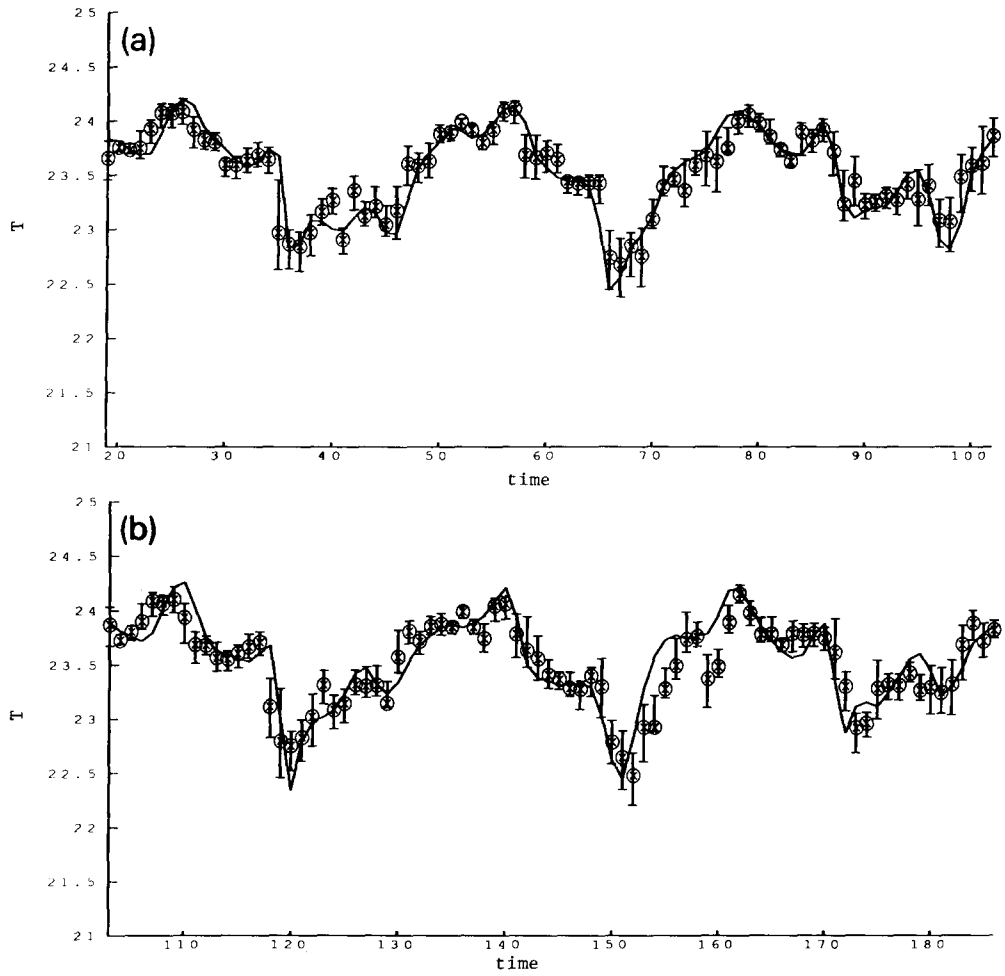


Fig. 8. A section from the learning set of reconstruction A where the temperature, its forecast value and the predicted uncertainty are plotted. Here $\tau_p = 18$ $(8\tau_s)$.

poor predictions and error estimates may also indicate sensor failure or a change in the dynamics of the physical system, an application developed in refs. [55,56].

### 4.2. Parametric drift

To conclude this section we address the question of slow parametric drift over the course of the experiment. If this were to occur, the dynamics at different points in time would be best determined from learning sets located nearby in time. To see if this is indeed the case, the learning set above (from the initial 4096 points) was tested on the remaining time series divided into thirds. Computing the Kolmogorov–Smirnov statistic, $d$, between the out-of-sample error distributions from a given predictor on different sections of the series [23], we find that the null hypothesis that the observed distributions arise from the same distribution function cannot be rejected at the 90% level of confidence. If this test is applied directly to the two halves of series $b$, we have $d_{obs} \approx 0.04$ and probability$(d > d_{obs}) \approx 0.86$. When applied to the data series, the Kolmogorov–Smirnov test indicates that the range and distribution of the data did not change; when applied to the error distributions,

it indicates that the quality of the predictor did not change and hence is evidence against slow parametric drift. It does not, of course, rule out recurrent parametric drift on time scales short relative to the length of the series. There exists no definitive method to do so as, in practice, the distinction between "parameters" and "system variables" becomes a philosophical question when both vary on small time scales.

## 5. Correlation dimension

The Grassberger–Procaccia Algorithm or GPA [14], which estimates the correlation dimension, $d_2$, provides a direct measure of the geometry of a distribution and has become perhaps the most widely used tool in the search for low dimensional dynamics. It is described in detail in [57,58]. Briefly stated, one wishes to estimate the correlation integral $C_2(\ell)$ of a distribution of points $\boldsymbol{x}$:

$$C_2(\ell) = \frac{\text{number of pairs of points separated by less than } \ell}{\text{total number of pairs of points}} = \text{probability}(\|\boldsymbol{x}_i - \boldsymbol{x}_j\| < \ell), \tag{5.1}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are two randomly chosen points in the set. It is implicitly assumed here that one is selecting from the set of all possible pairs of points on the attractor. This is not the case with reconstructions from time series when the spatial separation between a pair of points reflects that they are close in time. Theiler [59] demonstrated that for smooth dynamical systems, consideration of points close in time can lead to one-dimensional "knees" in correlation integral estimates. More recently, Osborne and Provenzale [60] have found finite correlation dimensions for power law noises, but these are another case of this same effect and need not foil dimension estimates in practice [61]. A simple test for detecting such effects is given in ref. [9]. Taking care that these effects are minimal, the correlation integral is approximated as

$$C_2(\ell) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \Theta(\ell - |\boldsymbol{x}_i - x_j|), \tag{5.2}$$

where $\ell$ is the length scale and $\Theta(x)$ is the Heaviside function which is equal to zero for negative argument and one otherwise. When the limit is not taken, the sums over $i$ and $j$ should be restricted so that $|i - j| > W$ [59]. Numerically efficient methods for evaluating the correlation integral are available [62,63].

In the limit of small $\ell$, we expect $C_2(\ell)$ to be scaling, that is

$$C_2(\ell) \sim \chi(\ell)\, \ell^{d_2}, \tag{5.3}$$

which defines $d_2$, the correlation exponent and $\chi(\ell)$ accounts for lacunarity effects [64,65]. At finite length scales, one can inspect the local slope of $\log_2 C_2(\ell)$ as a function of $\log_2(\ell)$ for a scaling range over which to estimate $d_2$. When estimating $d_2$, the $i = j$ terms in the sum should be neglected, although it is useful to compute $C_2(\ell)$ with both normalizations (this involves negligible computational cost) and compare their slopes as functions of $\log \ell$. Both curves provide useful information in judging the quality and evolution of reconstructions with changes in the embedding dimension.

There has been a great deal of discussion in the literature regarding the amount of data required to obtain a meaningful estimate of the characteristics of chaotic dynamical systems. For the correlation exponent, several authors [66–69] have provided estimates of the minimum "number of data points" required. Unfortunately, it is not easy to determine the number of data points in a time series in this sense. The difficulty lies in assumptions which require the data to be spread uniformly with respect to some underlying probability density (measure). Appeals to ergodicity are of no use when the sampling rate is such that consecutive measurements are dynamically correlated, for this biases the correlation integral by

increasing the probability of a pair of points at separation $\ell + \delta$, given that there is a pair with separation $\ell$. The dynamical correlation time is also very difficult to estimate *a priori*, it is certainly not the linear correlation time (i.e. the first zero of the linear autocorrelation function, $\tau_{\text{auto}}$) or the first minimum of the one dimensional mutual information [68]. A second drawback of these scaling arguments is that they provide necessary but not sufficient conditions, and the former are of much less use than a measure of success. It is easily shown that through smooth deformation of a reconstruction, one can always increase the number of data points required to obtain a good estimate of a dimension.

Like all analysis techniques, the GPA must be applied with some insight. There has been much discussion about the *a priori* knowledge of the system required to apply this algorithm. We would liken application of the GPA to that of the Fast Fourier Transform (FFT): one must understand the algorithm and its limitations when interpreting the results. To push the analogy, one rarely hears reports of strong power in a frequency beyond Nyquist limit, or public arguments over whether it is really necessary to have a stationary time series to apply the FFT. Such results would reveal a false application of the algorithm, not a flaw in it. Nor is it claimed that one must understand the physics of the system to gain useful information from a power spectrum. The analogy holds.

The analogy fails when the difficulty of coding the algorithm is considered, and the general knowledge of its limitations. Particular errors of application are well documented [59,66,58,57,70] although they are still common. Even so, these are again necessary, not sufficient conditions. Even when precautions are taken, one would like to estimate the probability that a given result, with specified reconstruction parameters (delay times or SVD window lengths, etc) is not due to such factors as the length of the data set. This is the strength of employing surrogate signals.

### 5.1. *Surrogate signals and the correlation integral*

The FT surrogate generator is now used to evaluate the results obtained for the rotating annulus data. Rather than attempt to automate the choice of a scaling region, fig. 9 presents the slope of the correlation integral from series *b* data along with that of a representative surrogate set. The solid (short-dashed) lines represent the slope of the $\log_2 C_2(\ell)$ against $\log_2 \ell$ including (excluding) the $i = j$ points in the sum. The regular long-dashed line is the expected slope for white noise with the same diameter (see [66]). The difference between observed and surrogate graphs is striking both in the value of the plateau (if one can be said to exist for the surrogate set) and the relative location of the nearest neighbor distances as reflected by the value of $\log_2(\ell)$ at which the curve including the $i = j$ points returns to zero. It is tempting to define a scaling range and determine the probability of observing the value $d_2 \approx 3$ for the surrogate sets. Such a calculation is questionable as the algorithm has not converged in the case of this surrogate series and there is no saturation as the embedding dimension is increased. This detracts little from the argument that the observed series has a significantly different correlation integral than expected due to its autocorrelation function.

We now examine the series *d* data set, considered to be quasi-periodic (two incommensurate frequencies) by Read et al. [13]. The slope of the correlation integral with length scale for this set is shown in fig. 10. Note that it does appear to be about two-dimensional. The feature at length scales $-2.5 < \log_2(\ell) < -1$, reflects the macroscopic structure of the reconstruction and not an artifact of the analysis. Assuming that the reconstruction is a two-torus, this would indicate that it has greater extent in "one direction", the true dimension of the distribution is not observed until length scales smaller than $\frac{1}{8}$ are reached. Even if a surrogate generator preserves the geometry of the two-torus, this particular macroscopic struc-

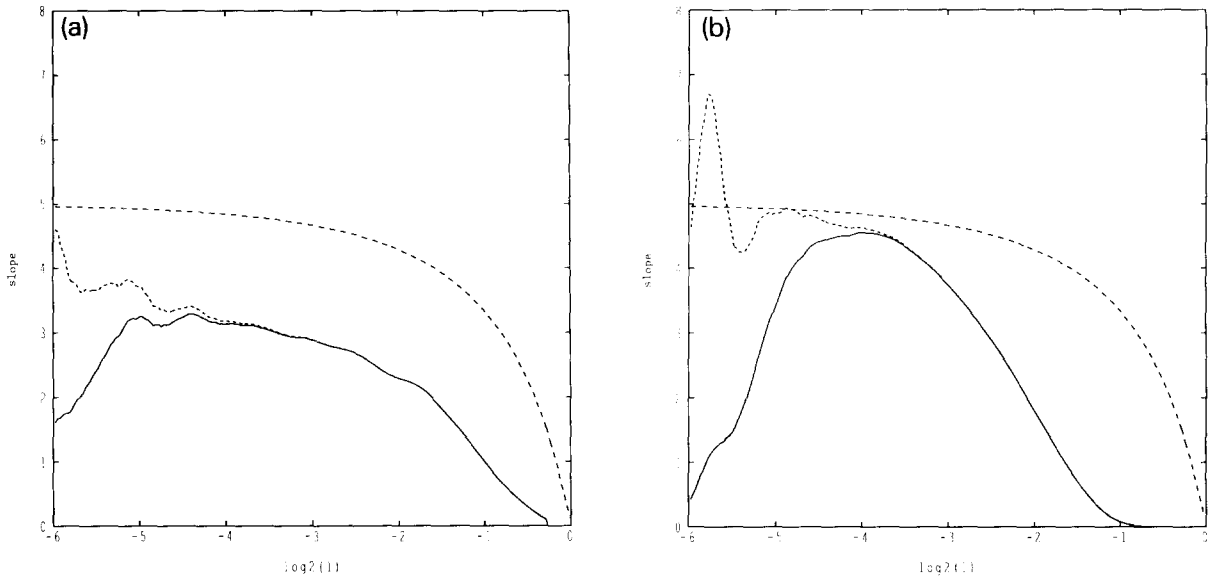


Fig. 9. Local slope of the correlation integral from series $b$ This analysis utilized $2^{15}$ data points. (a) observed and (b) and surrogate.

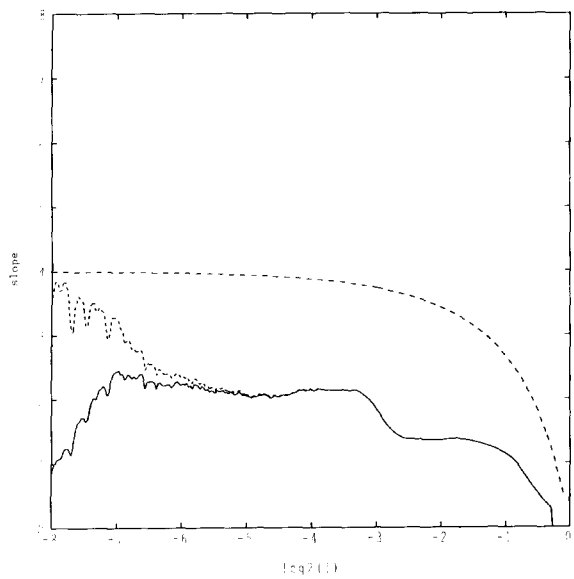


Fig. 10. Local slope of the correlation integral from series $k$.

ture need not be maintained; macroscopic distortions will shift the scaling range, confounding attempts to compare correlation dimension estimates between observed and surrogate signals even in a case where they both converge in the limit of small $\ell$. The point here is that two topologically equivalent distributions with different macroscopic structure will have different correlation integrals at large scales. This is a fundamental limitation inherent in the geometric analysis of reconstructions, and provides an example where the lower bounds on data requirements for dimension calculations are vast *under*estimates of the true amount of data required for this type of analysis. (In this particular case, the Fourier spectrum indicated a quasi-periodic attractor.) It also provides an example of where surrogate series can provide misleading results if a fixed scaling range is used.

## 6. Nonlinear prediction of stochastic systems

The arguments above demonstrate that nonlinear predictors can distinguish dynamical systems with a structured phase space flow from those whose motion in phase space is incoherent. If we identify the former systems as deterministic and the latter as stochastic, we have a good test for determinism. Unfortunately such a classification will consider many classic "stochas-

tic" systems as deterministic as it fails to distinguish "determinism" from stochastic systems which are "low dimensional" in the sense that they are associated with a probabilistic flow in a low dimensional phase space.

In this section we shall consider two systems, the disturbed pendulum of Yule [71] and the Ornstein–Uhlenbeck process[72,73]. The disturbed pendulum provides an example of a deterministic system where noise feeds back into the dynamics (dynamic noise) rather than being superimposed on the measurements (observational noise). The Ornstein–Uhlenbeck process has become a paradigm of stationary stochastic systems. We demonstrate that nonlinear deterministic predictors provide a good approximation to optimal prediction of this system and indicate the difficulties this implies for tests of determinism using surrogate signals. We are not interested here in establishing whether one type of nonlinear deterministic predictor is better than another, but in their common properties.

Yule considered two simple models to account for the lack of simple periodicity in the 15 sunspot cycles then available. Both models are based on observations of a pendulum. In the first, the observations of perfect periodicity are subject to *superposed fluctuations* or observational noise. In this case, for sufficiently long series, Fourier analysis will detect the underlying periodicity. Any deterministic predictors which allow for observational noise should do so as well. In the second case, the observational noise is considered negligible, but *disturbances* to the pendulum's motion (caused by boys with pea shooters) change the energy of the pendulum and feed back into the systems dynamics. When the shocks are well separated in time, nonlinear deterministic predictors will give excellent predictions (between the shocks) due to the structure of the underlying two-dimensional phase space of the pendulum. Good predictions are possible as long as the expected time interval between impacts, $\Delta t$, is not small relative to the sum of the reconstruction window and prediction time or

$$\Delta t \gtrsim (m-1)\tau_{\mathrm{d}} + \tau_{\mathrm{p}}. \tag{6.1}$$

The Ornstein–Uhlenbeck process models the velocity of a Brownian particle. From a dynamical systems perspective, it is preferable to consider the velocity, $u(t)$, rather than the displacement, as the velocity time series is stationary. The change in the velocity, $\mathrm{d}u(t)$, is given by

$$\mathrm{d}u(t) = -\beta u(t)\,\mathrm{d}t + \sigma\gamma(t)\sqrt{\mathrm{d}t}, \tag{6.2}$$

where $\gamma(t)$ is a random Gaussian process with zero mean and unit variance, $\mathrm{d}t$ is the time step, and the parameters $\beta$ and $\sigma$ are related to frictional drag and the driving impacts respectively. The optimal (statistical) predictor for this process is known; given the initial condition $u_0$, the expected value of $u(t)$ is

$$P_{\mathrm{theory}}(u(t)) = E(u(t) \mid u(0) = u_0) = u_0\, \mathrm{e}^{-\beta t}. \tag{6.3}$$

Estimates for the variance are also available [73]. To test whether the reconstructed dynamics finds this structure, a 2048 point learning set ($\beta = 0.5$, $\sigma = 1.0$, $\mathrm{d}t = 0.05$) with $m = 1$, $n_{\mathrm{c}} = 64$ and $\phi(r) = r$ was constructed and tested out-of-sample on an additional 2048 points. The point here is not whether this radial basis function predictor is optimal, but is merely to demonstrate that *any* good dynamic reconstruction should identify this structure in an Ornstein–Uhlenbeck series (or a series from a stochastic model like that of Barnes et al. [27]).

We compare the predictor with $P_{\mathrm{theory}}$ by plotting the predicted future value against the current observed value in fig. 11. The solid line corresponds to $P_{\mathrm{theory}}$. A scatter plot of the observed future value against the current observed value shows a wide distribution. Both the agreement and disagreement between the deterministic predictor and the expected value displayed in fig. 11 is understood. The largest values of $u$ in the

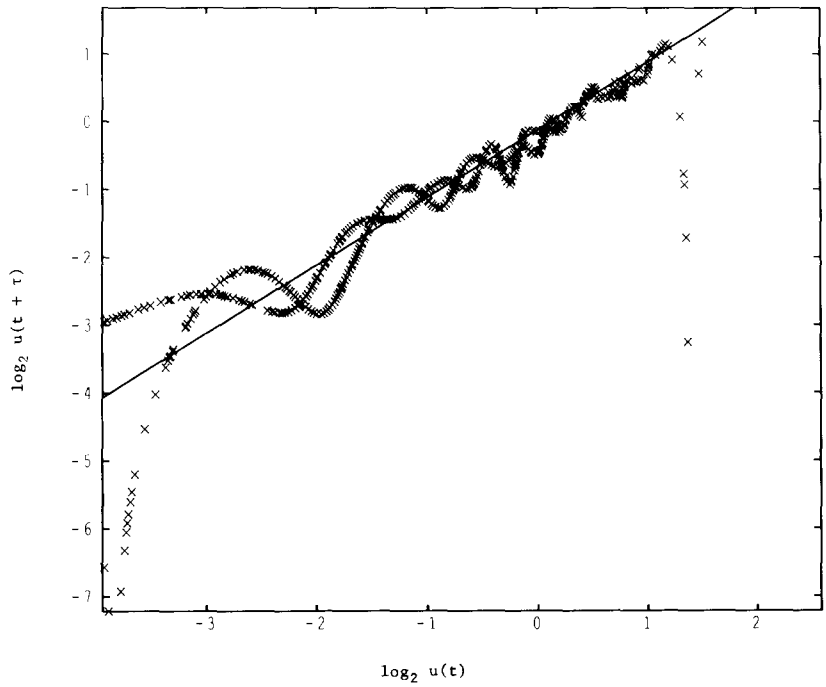


Fig. 11. Prediction of an Ornstein–Uhlenbeck process. The solid line denotes the optimal prediction. The predictor appears to be double valued because predictions for both positive and negative $x$ are superimposed. The very inaccurate predictions at large $x$ occur when the system explores a region of phase space not visited in the learning set. At small $x$, the variance in the expected value is large and the ability of the predictor to find the expected value is diminished.

test set correspond to very poor predictions (the markers on the right side of the plot well below the ideal line); these points have large nearest center distances and correspond to values of $u$ not visited in the learning set. For slightly smaller $u_0$ there is good agreement between the two predictors; the two images of the deterministic predictor are the superimposed values for positive and negative $u_0$. For small $u_0$ there is poor agreement between the two predictors; due to both the magnification of small distortions by the logarithmic scales and the increase in uncertainty of future values of initial conditions near $u = 0$.

To the extent that these systems are deterministic, the dynamic reconstructions quantify their behavior. Yet they are stochastic in the sense that the current state of the system does not completely define its future. Once the deterministic structure is quantified, the quality of the predictions should not improve regardless of increases in the amount of data available. The lack of improved precision with increasing data, embedding dimension or changing delay time (for infinite data sets) provides an indication of the stochastic component of the process but is difficult to establish with finite data sets. In these particular cases, examining the predictor error series of the pendulum, and its spatial variation in the O-U process, could help identify the dynamics of the processes. The situation is more complicated in stochastic systems with more complex (higher-dimensional) phase space structure. Here nonlinear dynamics comes from nonlinear structure in the governing equations regardless of whether they are stochastically or deterministically driven. As the structure of the governing equations increases, the nature of the stochastic forcing may become less apparent. This holds implications for the use of surrogate data, in that surrogate generators which destroy this structure will be distinguished regard-

less of whether the underlying driving force is stochastic or deterministic.

One approach to recover this distinction is to look to longer time scales. Stone [54] has considered a Duffing oscillator driven either sinusoidally (chaos) or by random perturbations (stochasticity) and shown that the signals from these two systems are similar in terms of power spectra and symbolic dynamics. The short time predictability of these signals is similar as well; however if one considers long time phenomena the two cases can be distinguished. In particular, the series of time intervals between departures from the origin is distinguished by the predictor presented here. These return times are effectively independent and identically distributed in the stochastically forced system while the chaotic system is, initially, predictable and the distribution of predictor error for return times appears to relax to the same distribution as the stochastic case as predictions are made farther into the future. As the nonlinear structure of these two systems is identical, they provide a useful example of the similarities and differences of stochastic and deterministic behavior.

## 7. Discussion

It is customary to consider low dimensional determinism and stochasticity as two clearly distinct types of behavior. As we have seen, distinguishing between these alternatives is sometimes difficult. We have presented a general approach to evaluating algorithms, which attempts this distinction through contrasting the results a given algorithm produces on the observed data with those produced from surrogate data. The importance of choosing a good surrogate generator has been stressed and the general effectiveness of this approach has been demonstrated for correlation exponent calculations and prediction algorithms on laboratory data.

We have focused our attention primarily on the rotating annulus experiments and established that these data sets differ significantly from the surrogate series considered. This gives us confidence that dynamical systems techniques can provide a better understanding of this system, in particular in determining the nature of the underlying driving mechanisms. This goal is difficult to obtain with the data in the form presented here for, while it may display deterministic, "low dimensional" behavior, the physics in delay space is not at all simple. Dynamical systems texts often give the impression that a system which evolves on a low dimensional (say $d_2 < 5$) attractor has simple physics. This is somewhat misleading. For a set of five ordinary differential equations (ODE's) it is true, perhaps even for a set of 10 ODE's which collapse onto such an attractor.

For a large physical system with many degrees of freedom, the dynamics in 5D is certainly more simple than not under such restriction, but the physics is a mess in 5D. The equations of motion need not correspond to the macroscopic physical properties of the system and will almost certainly not correspond to a set of simple ODE's. While a great deal can be learnt from such systems, it is misleading to imply that the physics, in a traditional sense, will become clear. Indeed, we may need to develop a new way of interpreting physics and it is tempting to draw an analogy with the way statistical mechanics answers different questions than classical dynamics. An alternate approach which we are currently pursuing with the annulus data is to recast the data into a form in which the physics is more assessable. The spatial distribution of probes allows a spatial Fourier transform into wavenumber space. With the data in this form, a multivariable reconstruction can address the general question of predictability directly, as well as particular questions concerning which mode interactions drive the dynamics of the system. For example, suppose the data is recast into a multivariate series of the amplitudes of modes A, B, C, .... Using the predictor discussed above, we plan to examine the extent to which the energy in modes A

and B determine the future behavior of mode C, thereby directly testing cascade and mode–mode interaction hypotheses. Through this type of study we hope to clarify what physical processes dominate the dynamics of this system.

We have also shown how dynamic reconstructions can be used to address open questions concerning the experiment itself. By demonstrating that predictors formed from one segment of the time series yield statistically indistinguishable errors when applied to different segments of the time series, we have shown that the complexity observed is not due to slow parametric drift over the duration of the experiment. The statistically significant difference, between both the predictability and the correlation integrals of the observed signals and surrogates generated with identical autocorrelation functions, provides a strong case for low dimensional dynamics in this system. This case is further supported by the demonstration that a simple prediction by memory scheme, while capable of distinguishing between the FT surrogates and the observations, is much less accurate than the deterministic radial basis function predictor.

In this paper, we have applied a global, nonlinear predictor based on radial basis function interpolation which explicitly considers noise in the data set and the inhomogeneity of the reconstruction in phase space. This type of predictor may be improved in several ways. For example, the reconstruction may be altered to include known physics in the problem at hand: in a problem where diurnal cycles are known to be important, the time of day could be included by placing the reconstruction on a circle. In systems like the annulus, the predictability may be improved by recasting the data set into a form in which the physics is more assessable as discussed above. It is often the case that additional information regarding the macroscopic state of the system is available in addition to time series data. Examples under investigation include laboratory data where the phase of a forcing function is known [74], and meteorological series, where the general structure of the regional weather pattern is included to improve the prediction of local temperature series [75]. For finite, noisy data sets, considerations such as these may be crucial to obtaining a significant result.

In addition to better embeddings, improvements in the prediction scheme are also possible but are likely to involve system specific answers. For example, the choice between iterative forecasting and direct forecasting may vary with the particular dynamical system, the data density, the noise level and even the details of the predictor itself (local or global, linear or nonlinear, ...). The system specific nature of this problem is likely to reoccur in other details of reconstructions, such as the importance of the method employed for choosing centers. For the predictor presented here, one may improve the method used to account for noise; we have applied a straightforward least squares approach. Implicit in this approach is the assumption that the "noise" is located in the quantity being predicted ($s$), not the base point ($\boldsymbol{x}$). For delay reconstructions this is certainly not the case, the same noise level is present in the base point as in the prediction. One approach to this problem would be to consider total least squares. This is analogous to performing an SVD fit in two dimensions rather than a least squares fit when it is known that there is error in both coordinates.

In the attempt to distinguish between deterministic and stochastic dynamics through prediction, one complication has been noted: the ability of deterministic predictors to identify the expected values for some stochastic systems, and thereby differentiate them from (some) surrogates. This is particularly true in effectively low dimensional stochastic systems, systems which exhibit stochastic motion within a structured low dimensional phase space. (Although as stochastic systems they remain, of course, infinite dimensional.) While such systems clearly fail to follow strict Laplacian determinism, it is not clear how they are best classified. Their detection by nonlinear prediction will depend

on the particulars of what quantities and length scales are analyzed, and useful classification in the presence of these effects may require consideration of second order properties of the predictor. In the event, these distinctions may require a more precise definition of what constitutes determinism.

## Acknowledgements

I thank M. Muldoon for many stimulating discussions and mutual information calculations. I am happy to acknowledge helpful discussions and disagreements with J. Theiler, D. Broomhead, K. Fraedrich, G. King and B. Mestel, and would like to thank P. Read, M. Gaster, T. Mullin for supplying data the analysis of which provided the insights reported here. M. Chappell, C. Lanone, P. Read and J. Theiler have provided very useful criticisms of a draft of this paper. Finally, I thank N. Weiss for discussions about chaos in the sun and the sunspot record. This research has been supported by the Science and Engineering Research Council and the US Office of Naval Research.

## References

[1] E.N. Lorenz, J. Atmos. Sci. 20 (1963) 130.
[2] D.W. Moore and E. A Spiegel, Astrophys. J. 143 (1966) 871–887.
[3] J. Crutchfield, J.D. Farmer, N.H. Packard and R.S. Shaw, Sci. Am. 254 (1986) 46–57.
[4] N. Gershenfeld, in: Directions in Chaos, ed. B.-l. Hao (World Scientific, 1988) pp. 310–383.
[5] S. Eubank and D. Farmer, An introduction to chaos and randomness, in: Proc. SFI Summer School, ed. E. Jen (Addison-Wesley, 1990).
[6] Marquis de Laplace and Pierre-Simon, Théorie Analytique des Probabilitiés (Paris, 1820), reproduced in the Oeuvres complètes de Laplace, Volume 11 (Paris, 1886).
[7] J. Theiler, B. Galdrikan, A. Longtin, S. Eubank and J. D. Farmer, in: Nonlinear Prediction and Modelling, eds. M. Casdagli and S. Eubank (Addison-Wesley, 1992) pp. 163–188.
[8] J. Theiler, S. Eubank, A. Longtin, B. Galdrikan and J. D. Farmer, Physica D 58 (1992) 77, these Proceedings.
[9] A. Provenzale, L.A. Smith, R. Vio and G. Murante, Physica D 58 (1992) 31, these Proceedings.
[10] E. Kostelich and J. Yorke, Physica D 41 (1990) 183–196.
[11] L.A. Smith, Quantifying chaos with predictive flows and maps: locating unstable periodic orbits, in: Measures of Complexity and Chaos, eds. N.B. Abraham et al. (Plenum, 1990).
[12] L.A. Smith, Applied chaos: computing unstable periodic orbits through predictive flows and maps, in: Information Dynamics, eds. H. Atmanspacher et al. (Plenum, 1991).
[13] P. Read, M.J. Bell, D.W. Johnson and R.M. Small, Chaotic regimes in rotating baroclinic flow, J. Fluid Mech. (1991), to appear.
[14] P. Grassberger and I. Procaccia, Phys. Rev. Lett. 50 (1983) 346.
[15] P. Read, Applications of singular value decomposition to the analysis of "baroclinic chaos", Physica D 58 (1992) 455, these Proceedings.
[16] R.L. Smith, in: Nonlinear Prediction and Modelling, eds. M Casdagli and S. Eubank (Addison-Wesley, 1992) pp. 115–136.
[17] C. Nicolis and G. Nicolis, Nature 311 (1984) 529.
[18] P. Grassberger, Nature 323 (1986) 609.
[19] C. Nicolis and G. Nicolis, Nature 326 (1987) 523.
[20] P. Grassberger, Nature 326 (1987) 524.
[21] I. Procaccia, Nature 333 (1988) 498–499.
[22] R. von Mises, Probability Statistics and Truth, (George Allen and Unwin, London, 1957).
[23] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, Numerical Recipes, (Cambridge Univ. Press, Cambridge, 1987).
[24] A.R. Osborne, A.D. Kirwan, A. Provenzale and L. Bergamasco, Physica D 23 (1986) 75–83.
[25] J.A. Eddy, Science 192 (1976) 1182–1202.
[26] E.A. Spiegel and A. Wolf, Chaos and the solar cycle, in: Chaos in Astrophysics Ann. NY Acad. Sci. Vol. 497 (1987) pp. 55–60.
[27] J.A. Barnes, H.H. Sargent and P.V. Tryon, Sunspot cycle simulation using random noise, in: The ancient sun eds. R.O. Pepin, J.A. Eddy and R.B. Merrill (Pergamon, New York, 1980).
[28] N.O. Weiss, Phil. Trans. R. Soc. London A 330 (1990) 617–625.
[29] D.S. Broomhead and R. Jones, Time-series analysis, Proc. R. Soc. London 423 (1989) 103–121.
[30] N.H. Packard, J.P.Cruchfield, J.D. Farmer and R.S. Shaw, Phys. Rev. Lett. 45 (1980) 712.
[31] F. Takens, in: Dynamical Systems and Turbulence, eds. D. Rand and L.-S. Young (Springer, 1981) p. 366.
[32] Th. Buzug, T. Reimers and G. Pfister, Europhys. Lett. 13 (1990) 605–610.
[33] M. Casdagli, Chaos and deterministic versus stochastic non-linear modeling, J. R. Statist. Soc. B submitted (1991).
[34] T. Sauer, J.A. Yorke and M. Casdagli, J. Stat. Phys. 65 (1991) 579–616.

[35] A.M. Fraser, Physica D 34 (1989) 391–404.
[36] J.L. Breeden and N.H. Packard, Nonlinear analysis of data sampled nonuniformly in time, Technical Report CCSR-91-15, Center for Complex Systems Research, Urbana, Il 61801. (1991).
[37] D.S. Broomhead and G.P. King, Physica D 20 (1986) 217.
[38] G. King, R. Jones and D.S. Broomhead, Nucl. Phys. B (Proc. Suppl.) 2 (1987) 379.
[39] J.D. Farmer and J. Sidorowich, Phys. Rev. Lett. 59 (1987) 8.
[40] J. Crutchfield and B.S. McNamara, J. Compl. Syst. 1 (1987) 417–452.
[41] D.S. Broomhead and D. Lowe, J. Compl. Syst. 2 (1988) 321–355.
[42] A.I. Mees, Modelling complex systems, in: Proc. Conf. on Modelling Complex Systems, eds. L.S. Jennings, A.I. Mees and T.L. Vincent (Birkhaüser, 1989).
[43] M. Casdagli, Physica D 35 (1989) 335–356.
[44] N.H. Packard, J. Compl. Syst. 4 (1990) 543.
[45] A.S. Weigend, B.A. Huberman and D. E. Rumelhart, Predicting the future: A connectionist approach, Intern. J. Neural Syst. (1990), submitted.
[46] G. Sugihara and R.M. May, Nature 344 (1990) 734.
[47] H. Tong, Non-Linear Time Series Analysis, (Oxford Univ. Press, 1990).
[48] K. Stokbro, Predicting chaos with weighted maps, Nordita preprint (1991).
[49] J.D. Farmer and J. Sidorowich, in: Evolution, Learning, and Cognition, ed. Y.C. Lee (World Scientific, 1988) p. 277.
[50] M.J.D. Powell, Radial basis functions for multivariate interpolation: a review, in: Proc. IMA Conf. on Algorithms for the Approximation of Functions and Data (RMCS Shrivenham, 1985).
[51] C.A. Michelli, Constr. Approx. 2 (1986) 11–22.
[52] A. Arneodo, G. Grasseau and E.J. Kostelich, Phys. Lett. A 131 (1987) 426.
[53] J. Guckenheimer and P. Holmes, Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields, Vol. 42 of Applied mathematical sciences (Springer, 1983).
[54] E. Stone, Phys. Lett. A 148 (1990) 434–442.
[55] L.A. Smith, K. Godfrey, P. Fox and K. Warwick, A new technique for fault detection in multi-sensor probes, in Control 91 (1991) p. 1062.
[56] L.A. Smith and K. Godfrey, Nonlinear methods of fault detection in multi-sensor probes, in preparation (1992).
[57] P. Grassberger, T. Schreiber and C. Schaffrath, Nonlinear time sequence analysis, Preprint WUB 91-14 (1991).
[58] J. Theiler, J. Opt. Soc. Am. A7 (1990) 1055–1073.
[59] J. Theiler, Phys. Rev. A 34 (1986) 2427–2432.
[60] A.R. Osborne and A. Provenzale, Physica D 35 (1989) 357–381.
[61] J. Theiler, Phys. Lett. A 155 (1991) 480-493.
[62] J. Theiler, Phys. Rev. A 36 (1987) 4456–4462.
[63] P. Grassberger, Phys. Lett. A 148 (1990) 63.
[64] R. Badii and A. Politi, Phys. Lett. A 104 (1984) 303–305.
[65] L.A. Smith, J.-D. Fournier and E.A. Spiegel, Phys. Lett A 114 (1986) 465.
[66] L.A. Smith, Phys. Lett. A 133 (1988) 283.
[67] D. Ruelle, Proc. R. Soc. London A 427 (1990) 241–248
[68] J. Theiler, Phys. Rev. A 41 (1990) 3038–3051.
[69] C. Essex and M. and Nerenberg, Proc. R. Soc. London A 435 (1991) 287–292.
[70] K. Judd and A.I. Mees, Estimating dimensions with confidence, Aust. J. Bif. Chaos (June 1991).
[71] G.U. Yule, Phil. Trans. R. Soc. London A 226 (1927) 267–298.
[72] G.E. Uhlenbeck and L.S. Ornstein, Phys. Rev. 36 (1930) 823–841.
[73] D.R. Cox and H.D. Miller, The theory of stochastic procesesses, (Chapman and Hall, New York, 1965).
[74] M. Gaster, Proc. R. Soc. London A 430 (1990) 3–24.
[75] K. Fraedrich, C. Ziehmann-Schlumbohm and L.A Smith, Estimating state dependent predictability: Some meteorological applications, Ann, Geophys. (1992) 1992 General Assembly Supplement Volume.