CHAPTER 4

LOCALLY OPTIMIZED PREDICTION OF NONLINEAR SYSTEMS: STOCHASTIC AND DETERMINISTIC

Leonard A. Smith
Mathematical Institute
University of Oxford
Oxford OX1 3LB, U.K.
email: lenny@maths.ox.ac.uk

Abstract

Forecasting the behavior of chaotic dynamical systems tends to be more difficult at some times than at others. This arises for two basic reasons: first, that the intrinsic predictability of a chaotic system usually varies with the state of the system, and second, the information we have about different regions of state space varies as well. The intrinsic variability implies, for instance, that Lyapunov exponents need not reflect any practical "limit of prediction." A method of locally optimizing the parameters of prediction schemes, which takes advantage of variability from both sources, is introduced and illustrated on both stochastic and deterministic nonlinear systems; difficulties in distinguishing the two are noted. It is also demonstrated that even very good (but not perfect) predictors will leave information in the time series of the residual prediction errors from chaotic signals: it is not possible to reduce the residuals to a sequence of independent, identically distributed random variables. A straightforward method to search for this residual predictability is demonstrated.

1. Introduction

Variability and unpredictability are two of the hallmarks of deterministic chaos; it is our aim here to show how the first can be exploited to reduce the second. We begin by examining this variability in the Lorenz attractor (Lorenz 1963), and considering the types of uncertainty in initial condition that give rise to prediction error, even when a perfect model is at hand. In section 3, we cast the prediction problem as one of interpolation in state space; recalling previous work on local linear prediction, we illustrate a weakness in this approach. We also show that irreducible approximation error will, in general, leave information in the residual prediction errors from chaotic systems. The predictability of a chaotic system varies with the location of the initial condition; the main result of this paper is to introduce a prediction scheme which embraces the variability both of dynamics and geometry. This approach is used to optimize local prediction in section 4, where it is illustrated on time series data from a low dimensional, presumably deterministic laser system, and a nonlinear stochastic model for sunspots. In these examples, local linear prediction is optimized by allowing the number of points used in constructing the predictor to vary with location in state space. Next, in section 6, we demonstrate why a time series from a chaotic system (defined as a deterministic system with at least one positive Lyapunov exponent) need not be difficult to predict all the time; the issue here is one of the uniformity of stretching around state space. The penultimate section contains a discussion of residual predictability, proposing a simple test to detect systematic prediction error, which indicates that further improvement in prediction accuracy is possible. The test is straightforward to apply, and has been found very effective in practice. Our conclusions, along with a pointer toward ensemble predictions and the limit of predictability, are given in section 8. In short, accurate prediction of a nonlinear system from limited data requires sensitivity to the variation of the system's properties in state space; examining this variability can cast new light on the "limits of predictability" as well as individual predictions.

2. Variations in Predictability with Initial Condition

The cover illustration shows the error-doubling time for a variety of initial conditions on the Lorenz attractor (Smith 1994; Ziehmann-Schlumbohm 1994). The colors reflect the minimum time required for an infinitesimal uncertainty to double; the red points double within one time unit, the orange within two, the yellow three and so on. This illustrates the main points of this paper: that the predictability of the flow is both variable and highly organized. Points with short error doubling times will tend to be more difficult to predict than those with long ones: the predictability of a chaotic system varies with initial condition (Benzi and Carnevale 1989; Doerner *et al.* 1991; Nese 1989; Palmer 1993; Smith 1992; Ziehmann-Schlumbohm 1994). By exploiting this variability, we can significantly improve the reliability of our predictions. Moreover, arguments based on uniform error growth are misleading. The argument linking the largest Lyapunov exponent to the "prediction horizon" is a good example; Lyapunov exponents need not reflect practical limits of prediction.

Before diving into the details of predicting chaotic dynamical systems, we consider two more general forecasting issues: the types of initial uncertainty we are likely to encounter and the implications of the statistic adopted to evaluate a given set of predictors. The short term error growth we observe will depend on the nature of the initial uncertainty in our observation. In the case of a perfect model, error arises solely from uncertainty in the initial conditions. When dealing with data on a strange attractor, the growth of uncertainty is often quantified through the divergence of nearest neighbors, whose likely orientation varies systematically around the attractor. A different structure is observed when this uncertainty is oriented by the largest (global) Lyapunov exponent (as on the cover), which will differ yet again from cases where it is determined either by the locally fastest growing direction, or by a random displacement. In systems like the Earth's atmosphere, we must consider systematic errors both in the observations themselves and from any preprocessing required by the model (as discussed by Palmer *et al.* 1994 and in this volume). In addition, arguments based on perfect predictors overlook the systematic effects of optimal, yet imperfect predictors, an issue examined in section 3 below. Finally, we note that the vast majority of uncertainty growth studies concentrate exclusively on infinitesimal initial uncertainties. The quantities of interest when determining the limitations of prediction must, by definition, consider the regime where the magnitude of the uncertainty becomes comparable with the range of the observable.

To evaluate a set of predictors, we must choose a criteria for comparison. For nonlinear systems, the result will be much more sensitive to the particular statistic chosen than an intuition based on independent, identically distributed (IID) gaussian residuals would suggest. Both the root mean square (RMS) error and the average absolute error may be dominated by a few large errors (Benzi and Carnevale 1989). Using the median absolute error or throwing out a small fraction of the worse predictions (Casdagli *et al.* 1992; Farmer and Sidorowich 1987), reduces this sensitivity, but clouds interpretation of the results. In practice, the predictor which is most frequently the most accurate may be wildly wrong when it is inaccurate. Indeed, it is the application for which the predictions are made that determines, for example, whether the prediction which is most frequently optimal is preferred over the one which is least frequently horrid. While it is obvious that the cost function will play a central role in determining the "best" predictor, the sensitivity to this choice in nonlinear systems is observed to be much greater than in linear systems. Prediction errors from chaotic systems are neither gaussian nor independent; they are correlated both in time and state space, and, as we shall see, they are usually chaotic themselves.

3. Prediction as Nonlinear Interpolation

The recent success of prediction methods for many chaotic dynamical systems (Abarbanel *et al.* 1993; Eubank and Farmer 1990; He and Lapedes 1993; Sauer 1993; Smith 1992; Sugihara and May 1990; Tong 1990; Ziehmann-Schlumbohm 1994; Ziehmann-Schlumbohm *et al.*1995) is due, in large part, to the successful translation of an extrapolation problem to an interpolation problem (Eckmann and Ruelle 1985). Ideally, these methods consider a point in the state space of a deterministic physical system, where the future of each initial condition is uniquely determined and, if the equations of motion are known, may be estimated. Alternatively, *if* enough data are available to provide a sufficient number of recurrent trajectories passing near the initial condition in question, then we can interpolate the future trajectory given only the observations. Yet the functions involved in chaotic dynamical systems are, at

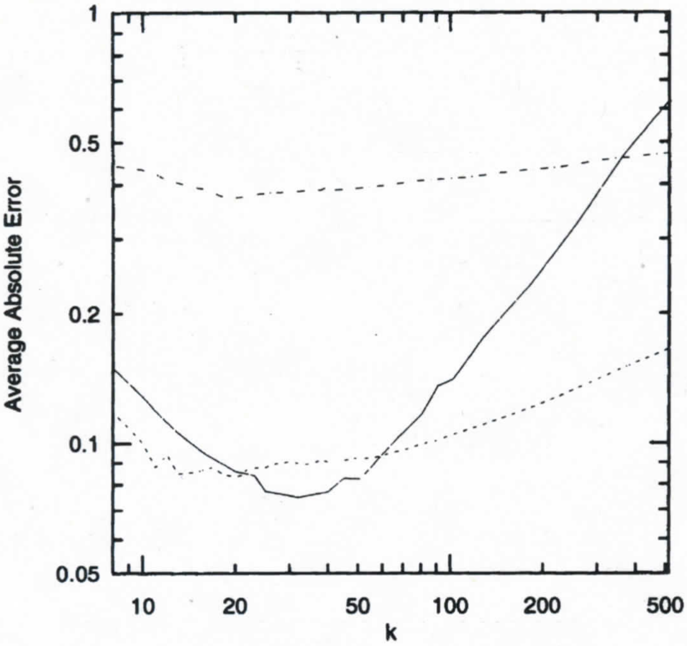

Fig. 1. Average absolute forecast error, E, as a function of the number of near neighbors, k, used for out-of-sample prediction of (solid) x_i of the Ikeda map, (dashed) the laser system, and (dot-dashed) the stochastic sunspot model. $E(k)$ is normalized by the average deviation.

their simplest, nonlinear, and the data are typically distributed on a strange attractor. We are faced with a high dimensional (> 2) interpolation of a complicated function sampled on an inhomogeneous distribution. With noise.

The two basic approaches to this problem consider either global interpolation functions for the entire state space, or restrict attention to local regions. The global approach requires a complicated interpolation scheme (*e.g.* radial basis functions or neural nets). Here we will consider the "local" approach (Farmer and Sidorowich 1987; Sugihara and May 1990; Tong 1990), which allows much simpler interpolation schemes. Following Casdagli *et al.* (Casdagli 1992; Casdagli *et al.* 1992), we can test for nonlinearity by evaluating a series of local predictors based on the k nearest neighbors and determining the value, k_c, at which the observed prediction error is smallest. For linear stochastic systems, k_c should correspond to the largest k, while for noise free deterministic systems, k_c should be of the order of the dimension of the system, given enough data. As illustrated below, for both deterministic systems with noise and nonlinear stochastic systems, a minimum at "moderate" values of k is to be expected.

As a concrete example, consider the chaotic, 2 dimensional Ikeda map

$$\mathbf{D}_{\text{ikeda}}(x,y) = (1.0 + \mu\ [x\ \cos(t) - y\ \sin(t)], \mu\ [x\ \sin(t) + y\ \cos(t)]) \tag{1}$$

where $t = 0.4 - \frac{6.0}{x^2+y^2+1}$ and $\mu\ = 0.90$. We will make one step ahead predictions of the value of x, given 1024 observations and base points (x, y) with gaussian noise*

*For simplicity, we consider additive white noise, which is represented through an independent and

(σ = 0.125) added to the observations. The solid line in figure 1 shows the out-of-sample, normalized average absolute error, $E(k)$ indicating $k_c \approx 32$. Yet if we consider only the 5 predictors with k = 8, 16, 32, 64 and 128, then we find the $k = 32$ predictor is the most accurate just 27% of the time (the distribution being 19, 23, 27, 23, and 8%, respectively). This illustrates a shortcoming of selecting a predictor with this approach: a global choice of k fails to account for the variation in the length-scales of either the dynamics or the data distribution. These two effects are shown schematically in figures 2 and 3. Figure 2 shows a one-dimensional local linear approximation to a polynomial curve. The circle shows the radius at which the expected value of the noise is equal to the error introduced by the linear approximation of the true curve. The optimal local radius within which data should be used, r_{opt}, depends upon the statistics of the noise, the data density and the local curvature; and this is our point: that even with uniformly distributed data, r_{opt} will change with the local curvature of the function estimated. Figure 3 illustrates additional complications due to the nonuniform distribution of data; these effects change not only with location in state space, but also at the same location with different data sets. In a given realization, there are likely to be regions where the local data density is too low to accurately estimate the local component of the dynamics and, simultaneously, other regions where the data density is sufficient to obtain a very good estimate. The (locally) optimal neighborhood to fit will vary between these regions, regardless of whether it is parameterized by k or r. Our goal is to develop a scheme which adjusts k to minimize the expected prediction error when the local curvature is not known. We will return to local linear predictors shortly, but first we consider the general question of irreducible approximation error in optimal predictors.

3.1. Imperfect Predictions: Colorfast chaos

Let $\mathbf{D}(\mathbf{x})$ define a deterministic dynamical system at a point $\mathbf{x}$ in state space. For simplicity, we consider a scalar measurement function $D(\mathbf{x})$ of $\mathbf{D}(\mathbf{x})$ representing the future property of $\mathbf{D}(\mathbf{x})$ which we wish to predict. The problem is then one of function approximation; we wish to approximate $D(\mathbf{x})$ given a particular family of predictors, F, with parameters λ. We may divide D heuristically into two parts:

$$D(\mathbf{x}) = F(\lambda, \mathbf{x}) + E_F(\mathbf{x}) \tag{2}$$

where $F(\lambda, \mathbf{x})$ represents an optimal fit to D and $E_F(\mathbf{x})$ represents the deterministic structure in $D(\mathbf{x})$ orthogonal to $F(\lambda, \mathbf{x})$. $F(\lambda, \mathbf{x})$ is optimal in the sense that the remaining approximation error is due to the structure of F itself. For example, if F is a linear model, it reproduces the linear behavior of $D(\mathbf{x})$ exactly; in this case $E_F(\mathbf{x})$ would consist of the quadratic and higher order terms in $D(\mathbf{x})$.

The main point of this section is that, in general, we expect $E_F(\mathbf{x})$ to be a good measurement function: when Taken's theorem (Sauer *et al.* 1991; Takens 1981)

identically distributed random variable added to each observation. The case of red (autocorrelated) noise can be handled in a similar manner, when the time scales of the noise are shorter than the recurrence time of the underlying dynamics.

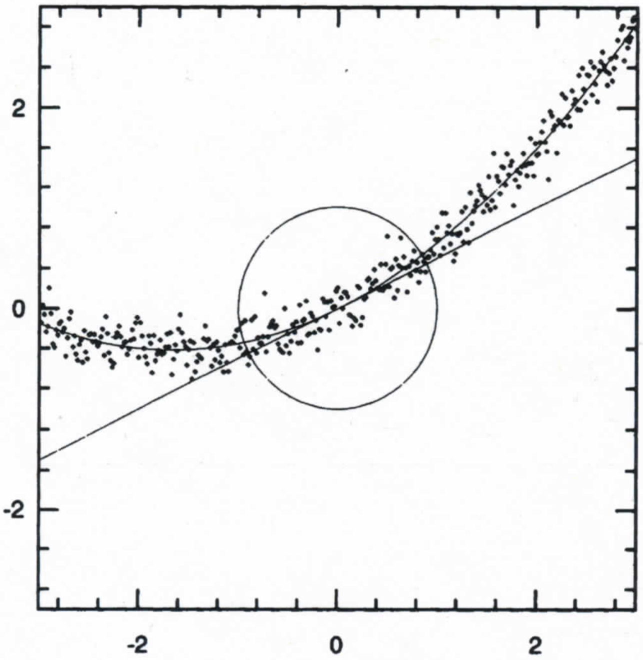

Fig. 2. Schematic diagram showing the need to balance the noise level against the local curvature in a linear fit to a polynomial function. The circle shows the radius at which the deviation due to the linear approximation equals the expected value of the noise.

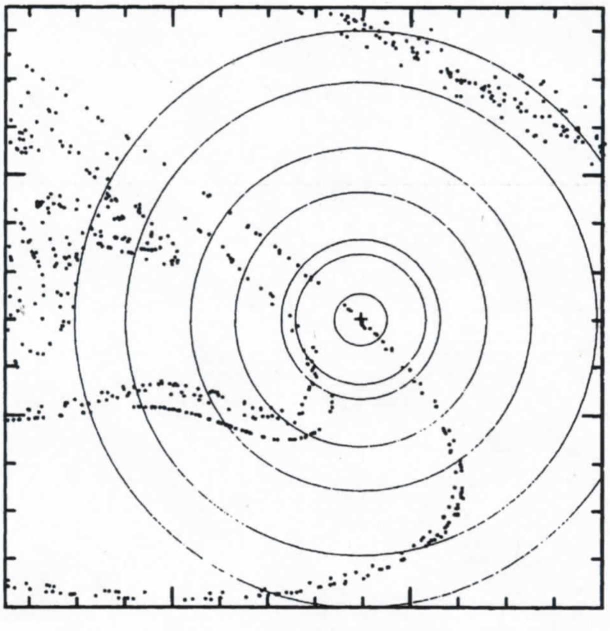

Fig. 3. Circles containing 9, 17, 33, 65, 129 and 257 points near the point to be predicted (+). The data is not evenly distributed within the circles, tending to be almost linear for $k = 8$ and remaining completely skewed to the one side of the prediction point until $k = 256$.

applies to a chaotic data stream, it will also apply to the time series of residuals from our prediction scheme. This implies that the residuals will be chaotic, albeit with more complicated macroscopic structure and a smaller signal to noise ratio, and indicates the impossibility of truly bleaching chaotic data (Theiler and Eubank 1994); chaos is colorfast in the sense that the residuals of non-perfect predictors will, in general, be chaotic. Removing the global linear structure of a chaotic signal (*i.e.* "pre-whitening" or "bleaching") will not result in IID residuals (Brock *et al.* 1991), although it will obscure the results of a nonlinear analysis, as stressed by Theiler and Eubank (Theiler and Eubank 1994). Similar results hold for more complex filters (Broomhead *et al.* 1992; Sauer *et al.* 1991), where the filter or predictor influences the particular value that we observe, but the dynamics are determined by the underlying system![†] Ultimately, the goal of nonlinear prediction is to reduce residuals to an IID sequence; we now examine obstacles to such a reduction even in the absence of noise.

As an explicit example of the separation in (2), consider local polynomial prediction near a point $\mathbf{x_0}$. We expand $D(\mathbf{x})$ about $\mathbf{x_0}$; setting $r = ||\mathbf{x} - \mathbf{x_0}||$ and denoting angular orientation by the vector $\mathbf{\Theta}$ yields

$$D(\mathbf{x}) = \underbrace{D(\mathbf{x_0}, 0) + a_1(\mathbf{x_0}, \mathbf{\Theta})\ r}_{F(\lambda,\mathbf{x})} + \underbrace{a_2(\mathbf{x_0}, \mathbf{\Theta})\ r^2 + a_3(\mathbf{x_0}, \mathbf{\Theta})\ r^3 + \ldots}_{E_F(\mathbf{x})} \tag{3}$$

where we have shown $F(\lambda, \mathbf{x})$ and $E_F(\mathbf{x})$ for local linear prediction. For higher order polynomial predictors, additional terms would be shifted to $F(\lambda, \mathbf{x})$. For other nonlinear predictors (*e.g.* a particular radial basis function fit or a specific neural net) this decomposition may be difficult to write down, but holds in principle as long as $D(\mathbf{x})$ is not a member of the family of functions which may be approximated exactly by F. If $D(\mathbf{x})$ is such a function, then $F(\lambda, \mathbf{x}) = D(\mathbf{x})$, and F is called a perfect predictor.

For local linear prediction of systems in which $a_2(\mathbf{x}, \mathbf{\Theta})$ varies significantly with $\mathbf{x}$, quantities like the optimal radius will be a function of $\mathbf{x}$ even when the learning data is uniformly distributed: r_{opt} will tend to be smaller where $a_2(\mathbf{x}, \mathbf{\Theta})$ is large. Typically, $a_2(\mathbf{x}, \mathbf{\Theta})$ does vary; figure 4 reflects this variation in the case of the Ikeda map. Here the value of $a_2(\mathbf{x})$, the absolute value of $a_2(\mathbf{x}, \mathbf{\Theta})$ averaged over $\mathbf{\Theta}$, is shown. Variations in data density will enhance this effect, particularly on strange attractors where the distribution is typically fractal and extremely inhomogeneous.

In the noise-free case, k_c is the minimum number of neighbors required to solve for $F(\lambda, \mathbf{x})$ (assuming the data is not distributed in a degenerate way). In the presence of noise, the number of neighbors (or equivalently, the optimal radius, r_{opt}) will depend on the noise level and the local curvature of D at $\mathbf{x}$, as well as the local data distribution. It is this balance between curvature and noise which is illustrated in figure 2. This approach is applicable to all local prediction schemes; it is the aim of local optimal prediction to vary the parameters of F (*e.g.* the value of k) in order to both balance the local structure of E_F against that of the noise, and adapt to the details of the data distribution, in cases where the analytic structure of D is unknown.

[†] In contrast to the series obtained by repeatedly iterating a non-perfect predictor.

A similar approach might be applied to global schemes as well, but given that global schemes are more costly to construct and, usually, quite easy to evaluate, it might be preferable to stack (Wolpert 1992) global predictors or even to choose the best reconstruction as a coarse function of location in state space (Smith 1992), and then employ ensemble prediction (Palmer 1992; Palmer 1993; Smith 1994)

Returning to local predictors, we note that the effects of variation in curvature can be shown by analyzing, for a given $\mathbf{x}$, an ensemble of data sets, each consisting of uniformly distributed learning data with gaussian noise. By determining the value of the radius which provides the best fit for each realization and observing this distribution for different points on the attractor, the $a_2(\mathbf{x})$ dependence of r can be seen, for example, by plotting the distribution of r against the local value of $a_2(\mathbf{x})$ (not shown).

Thus we have seen that the optimal number of neighbors will vary with the local curvature of $D(\mathbf{x})$ in state space, the local data density, (which will itself vary "strangely" for dissipative chaotic attractors) and, of course, the details of the noise. This ı .st observation implies not only that the details will be system dependent and vary with location in state space, but they may also change at a given point $\mathbf{x}$ from realization to realization; if, by chance, a number of low noise data points are the near neighbors of the point of interest, the optimal radius for this realization may be much smaller than that expected from the general arguments above; local optimal prediction adapts to these special cases.

4. Locally Optimizing Linear Prediction

We now present an algorithm to determine the optimal k from the data alone. In each of the cases discussed below, the data set is divided into a learning set from which the predictors are constructed, and a test set for out-of-sample evaluation. The learning set is analyzed to estimate the global parameters k_c and $k_{\mathbf{max}}(>> k_c)$, the largest neighborhood to be considered. The basic idea is to consider a small number of points in the learning set close to the point to be predicted, $\mathbf{x_0}$, and then employ local "drop-one-out" predictors with various k to predict the (known) image of each of these points in turn (Smith 1993). The best value of k for this $\mathbf{x_0}$ is then determined from the observed errors in predicting these points from the learning set.

More specifically, to determine the best value of k at the point $\mathbf{x_0}$, determine the $k_{\mathbf{max}}$ nearest neighbors of $\mathbf{x_0}$ in the learning set. From this subset, select the $N_{\mathbf{drop}}(\approx 8)$ points nearest to $\mathbf{x_0}$ with the requirement that these test points are well separated in time; this requirement is crucial to avoid picking consecutive points from the same segment of the trajectory, which leads to highly correlated (and misleading) estimated errors[‡] In the example shown in figure 3, these $N_{\mathbf{drop}}$ points lie within the smallest circle. For several values of k (corresponding to the circles in figure 3), construct $N_{\mathbf{drop}}$ distinct predictors by dropping out, in turn, each of the $N_{\mathbf{drop}}$ test points and predicting the point omitted. Finally, combine these results for each k to

[‡] A similar phenomena is known to bias dimension estimates (Theiler 1986)

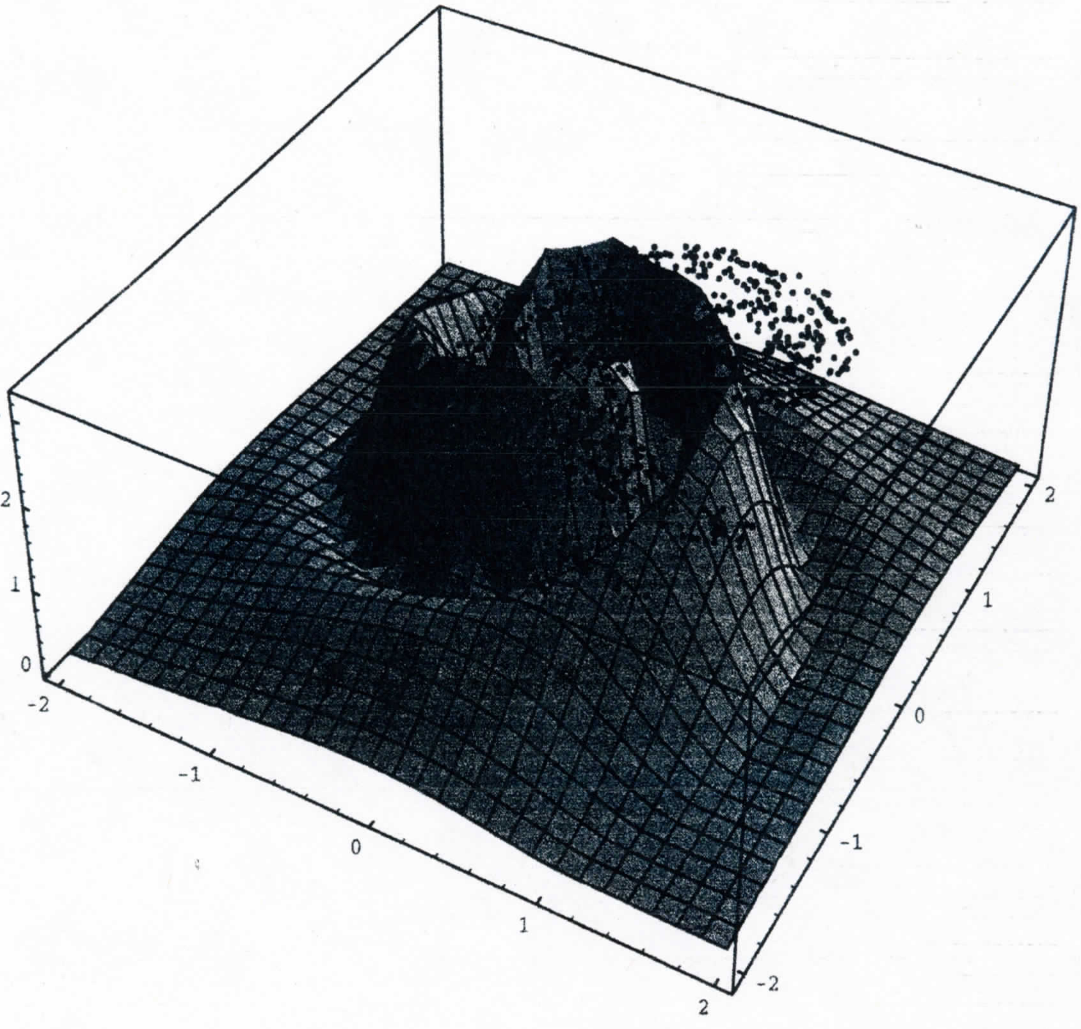

Fig. 4. This surface reflects the variation in the quadratic component for perfect, one step ahead prediction of the x value of the Ikeda map. The height corresponds to the absolute value of the quadratic term, averaged over angle. The dots show the location of the attractor.

obtain an estimated prediction error at $\mathbf{x_0}$ for a k neighbor predictor, and determine the optimal radius at $\mathbf{x_0}$ by comparing these estimates.

The choice of optimal radius is non-trivial, and detailed results will be presented elsewhere (Smith and Drysdale 1994). Frequently, choosing the value of k with the minimum estimated error improves the predictor a large fraction of the time, but *increases* the overall RMS error by making a few bad choices. A second approach is to consider the estimated error as a function of k and find the local minimum nearest k_c. Alternatively, we can partition the entire space, and use the partition element to define the initial guess for k (see Smith 1992; Smith 1993). When considering local minima, it is useful to check that there is not a significantly (in terms of the standard error of the estimated prediction error) deeper minima elsewhere. In the results presented below, the second method is used.

We consider data from a deterministic laser system (Weigend and Gershenfeld 1993) and a nonlinear, stochastic model (Barnes *et al.* 1980). To apply these ideas when only a single time series, s_i, is available, we first form a reconstruction space via the method of delays (Sauer *et al.* 1991). This yields the series of M-dimensional vectors, $\mathbf{x}_i$

$$\mathbf{x}_i = (s_i, s_{i-j}, \ldots, s_{i-j(M-1)}) \tag{4}$$

where j is the delay time. Under ideal circumstances, Taken's theorem (Sauer *et al.* 1991) assures us that many properties from the true state space dynamics are preserved by this reconstruction. Of course, this theorem suggests that a single observed data series can be used, but the use of multivariate data can remove projection effects commonly found in finite series, although it introduces the question of defining a metric (literally, an apples and oranges question) . In each case, we use previously published reconstruction parameters M and j to make direct, fixed step, local linear predictions, and evaluate the models out-of-sample. For delay reconstructions, this model is

$$F(\lambda, \mathbf{x}) = \lambda_0 + \sum_{\ell=0}^{M-1} \lambda_{\ell+1}\ s_{i-j\ell} \tag{5}$$

where the λ_i are determined by least squares fit to the k nearest neighbors of $\mathbf{x}$. The variation of the predictor with $\mathbf{x}$ can be made smooth by weighting the contributions of points as a function of their distance from $\mathbf{x}$.

The laser system has been widely discussed (Weigend and Gershenfeld 1993). We take a test set of 1000 points (figure 5a) and a learning set of 8192 points with $M = 4$ and $j = 1$, and predict 4 steps ahead. The dashed line in figure 1 shows the expected behavior of the out-of-sample prediction error with k with a minimum at $k = 20$. Using the drop-one-out scheme of the previous section to allow k to vary with $\mathbf{x}$ improves the average absolute error, E, relative to the k_c predictor by 6% (median absolute error by 10%). If we suppress the worst 10% predictions of each predictor, following Casdagli *et al.*(Casdagli *et al.* 1991), this becomes a 6.5% reduction of E; omitting these points also reduces the RMS error for the k_c predictor by an order of magnitude. When we examine, after the fact, how often each predictor from the set k = 8, 11, 16, 23, 32, 45 and 64 was the most accurate out-of-sample, we find they were

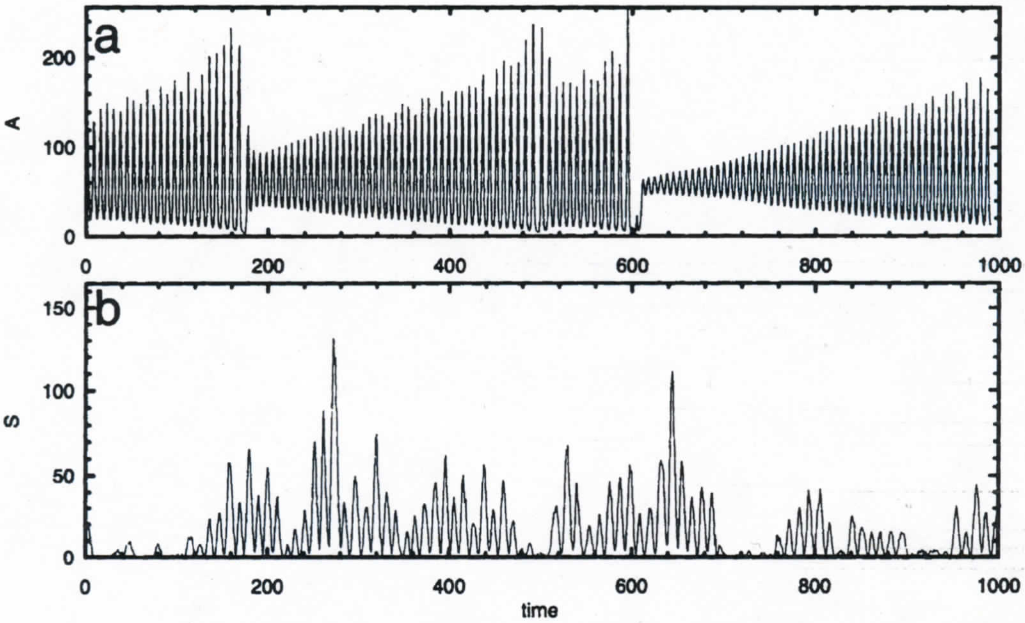

Fig. 5. Time series data from (a) a laser experiment and (b) the stochastic sunspot model.

best 20, 14, 14, 12, 11, 11 and 18% of the time, respectively; the optimal k clearly varies with location.

We observe that large errors are likely to occur when exploring new regions of state space (*i.e.* the macroscopic structure of the attractor not represented in the learning set); this is to be expected, in as much as, in new regions of state space, our procedure for prediction by interpolation reverts to an extrapolation problem. In regions of very low data density, reverting to the "zeroth order" method of simply taking the image of the nearest neighbor (often called analog prediction (Lorenz 1969)) can significantly improve the overall prediction error. An advantage of drop-one-out schemes is that, even when extrapolating, they provide an estimate of the expected error which differs from that available from previous predictions in the same neighborhood (Smith 1992). Local optimal prediction can also outperform fixed k prediction in data dense regions; small fixed k local predictors, like iterated predictors, get lost in data-dense regions where the dynamics are simple, but the series is slowly varying and nearest neighbors are determined by chance (the noise). We expect significant improvement using local optimal prediction, for example, in the chemical experiments presented by Prof. Olsen (Olsen *et al.* 1994).

While we do not know how to construct the analog of figure 4 for this four dimensional reconstruction of experimental data, we exploit the quantization of the data to quickly search for "exact" returns in the reconstruction space, and then examine the images of these points. This reveals that the majority of exact returns have very similar images, but also shows several cases of exact recurrence (*i.e.* to the digital resolution of the data), which have vastly different images (comparable to the range of the data). Given the available resolution, the function we are interpolating is multi-valued at these locations, and no interpolation scheme can assure success. We could lift these degeneracies by increasing M (given enough data), altering the delay parameter, j , in these regions (Smith 1993), or combining the results of several different prediction schemes (Sauer 1993; Smith 1993; Wolpert 1992). Note that in many regions of state space, the surface we are interpolating becomes more contorted (indeed, multi-valued) as the prediction time is increased beyond the sampling rate; in such regions we expect iterated, one step ahead prediction to out perform the direct four step ahead predictor used above.

5. Aleatoric Dynamics

While the prediction techniques discussed above were developed in the context of nonlinear deterministic systems and then generalized to handle noise (Broomhead and Lowe 1988; Casdagli 1992; Farmer and Sidorowich 1987), they are applicable in some nonlinear stochastic systems. In the past, successful nonlinear prediction has been interpreted as evidence of deterministic dynamics. This interpretation is incorrect for a class of nonlinear, but fundamentally stochastic systems which do **not** meet the criteria of Laplacian determinism; the current state of such a system does not define a unique future (or past). We call such systems "aleatoric" since, while the underlying driving mechanism appears not to be deterministic, the dynamics are governed in large part by deterministic laws, as with a roll of the dice. The boundaries of what should be included under this heading are not clear, save that the dynamics should be nonlinear (and not trivially made linear), and the random contribution to the dynamics should be essential. A chaotic system with dynamic noise (*i.e.* noise which effects the state of the system, not just the observation of that state) falls into this category, but the concept of aleatoric systems was motivated by stochastic systems with no chaotic analog. One such system was proposed by Barnes *et al.*(Barnes *et al.*1980) as a model for annual mean sunspot numbers, Y (figure 5b). Based on an ARMA(2,2) model with nonlinear modifications to ensure that Y remains positive and tends to increase more rapidly than it decreases, the model is

$$Z_n = \phi_1 Z_{n-1} + \phi_2 Z_{n-2} + a_n - \theta_1 a_{n-1} - \theta_2 a_{n-2} \tag{6}$$

$$Y_n = Z_n^2 + \alpha (Z_n^2 - Z_{n-1}^2)^2 \tag{7}$$

where $\phi_1 = 1.90693, \phi_2 = -0.98751, \theta_1 = 0.78512, \theta_2 = -0.40662$, $\alpha = 0.03$ and the a_n are IID gaussian random variables with zero mean and $\sigma = 0.4$. This system was used by the author and Weiss (Weiss 1990) as a surrogate process (Smith 1992; Theiler *et al.*1992) to illustrate the difficulty of estimating the correlation dimension of the observed sunspot record. It also illustrates the advantage of high data densities with an "incorrect" model, over low data densities from the true model: forecast errors of local linear predictors fit to 6000 years of model data can outperform the same predictor fit to the first half of the observed sunspot series, when the evaluation is based on the RMS error of the residuals of 1 year ahead predictions of the second half of the observed series.

To avoid any bias from our knowledge of the underlying system, the reconstruction parameters of Casdagli *et al.*(Casdagli *et al.*1992) ($M = 3$, $j = 1$) for the observed sunspot series were used to make one year ahead predictions. The dot-dashed curve in figure 1 traces the out-of-sample average absolute error as a function of k showing a minimum at $k = 20$, a behavior similar to that expected from a deterministic system.[§] Further, the data density and sensitivity of the dynamics vary tremendously in different regions of reconstruction space, so local optimal prediction improves prediction; in addition, the time series of prediction errors is far from IID. While Casdagli *et al.* (Casdagli *et al.* 1992) explicitly note that nonlinear stochastic systems may display this behavior, they imply that one can distinguish a minimum at "medium" k from one at "small" k. In practice, such a distinction may be invisible (*e.g.* in figure 1), and attempts to distinguish stochastic aleatoric dynamics from deterministic chaotic dynamics may require huge data sets, in order to get good statistics on very near returns in reconstruction space and to quantify the decay of predictability more directly (Casdagli *et al.*1992; Sugihara and May 1990). While of philosophical interest, the distinction may have few, if any, practical implications, and the ultimate resolution of this question requires the evaluation of limits that cannot be determined from any quantity of finite precision measurements. In contrast, the question of whether a given system is better modeled by a stochastic approach or a deterministic approach can be profitably addressed.

6. Lyapunov Exponents and Limits on Predictability

We now return to deterministic chaos and support our earlier claim that Lyapunov exponents can provide a misleading indication of predictability. Lyapunov exponents represent a growth rate averaged over the attractor. In a simple example, we contrast this average with the average time taken for an initial uncertainty to double; when the rate of uncertainty growth is not uniform, the two may be very different. We then consider the implications this holds for iterated and direct prediction schemes.

6.1. Baker's Apprentice Maps

In the Baker's Map (see *e.g.* Tong 1990 and references therein), the unit square is stretched by a factor of 2 in the x direction, compressed by a factor of 2 in y, and the resulting rectangle is then cut and stacked to form an area preserving map. In this uniform case, the largest Lyapunov exponent, λ_1, is 1 bit per iteration. The error doubling time for a given initial condition is simply the minimum time required for an initial uncertainty to increase by a factor of two. In this case, all initial uncertainties in the expanding direction double on each iteration, hence the average error doubling time, τ_{double}, for the Baker's Map is equal to one iteration. Contrast this situation with a class of generalized Baker's Maps, the family of Baker's Apprentice Maps (Smith

Indeed, this system is in a sense "low-dimensional," as discussed in reference (Smith 1992).

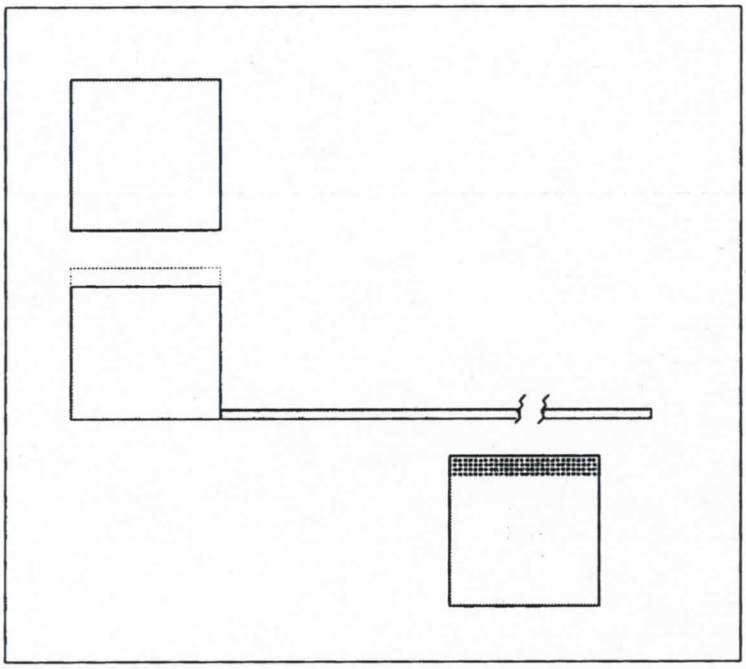

Fig. 6. Schematic representation of a Baker's Apprentice Map showing how a small portion of the unit square is stretched a great deal, then cut and stacked, while the majority of initial conditions are displaced only slightly.

1993), given by

$$x_{i+1} = \begin{cases} \frac{1}{\alpha} x_i & \text{if} \quad 0 \leq x_i < \alpha \\ \beta \, (x_i - \alpha) \quad \text{mod } 1 & \quad \alpha \leq x_i < 1 \end{cases}$$

$$(9)$$

$$y_{i+1} = \begin{cases} \alpha y_i & \text{if} \quad 0 \leq x_i < \alpha \\ \alpha + \frac{1}{\beta}(\lfloor \beta(x_i - \alpha) \rfloor + y_i) & \quad \alpha \leq x_i < 1 \end{cases}$$

where $\alpha = \frac{2^n-1}{2^n}$, $\beta = 2^{2^n}$ and $\lfloor z \rfloor$ denotes the greatest integer less than or equal to z. A schematic diagram illustrating the action of these maps is shown in figure 6 of the "dough" is stretched a great deal (2^{2^n}) before being cut and stacked, while the majority of the initial conditions are displaced only slightly (a factor of $(\frac{1}{\alpha}) \approx 1$). For each positive integer n, equations (9) define an area preserving map whose positive Lyapunov exponent, λ_1 $(= 1 - \alpha \log_2 \alpha)$ is greater than 1 bit per iteration. Thus the largest Lyapunov exponent of each of these maps is *greater* than that of the Baker's Map, yet for the majority of initial conditions, the Apprentice Maps are much more predictable.

To quantify this predictability, we compute the doubling time for infinitesimal uncertainties in the expanding direction of an Apprentice map; points of equal doubling time fall in vertical bands, reminiscent of those seen in the Lorenz system on the cover. In particular, initial conditions with x in the range $\alpha \leq x < 1$ will double after one iteration, those in the preimage of this band (*i.e.* $\alpha^2 \leq x < \alpha$) after two iterations, and so on until we reach the number of iterations, call it k, required for the uncertainty to double without having reached the region of extreme stretching. But k is then just the smallest integer such that $\left(\frac{1}{\alpha}\right)^k \geq 2$; initial conditions in the band $0 \leq x < \alpha^k$ will have doubling time k. Hence the doubling time for the n^{th} Apprentice Map is

$$\tau_{\text{double}} = k\alpha^k + \sum_{i=1}^{k} i\,[(1-\alpha)\alpha^{i-1}] = \sum_{i=0}^{k-1} \alpha^i = \frac{1-\alpha^k}{1-\alpha}$$

where $k = \lceil -\frac{1}{\log_2(\alpha)} \rceil$, and $\lceil z \rceil$ is the smallest integer greater than or equal to z. For large n, $\tau_{\text{double}} \approx 2^{n-1}$. And this happens fairly quickly: at $n = 4$, $\tau_{\text{double}} = 8.133$ while by $n = 8$, $\tau_{\text{double}} = 128.45$. The chaotic maps corresponding to large n are easily predicted (most of the time), although their largest Lyapunov exponent is always greater than that of the Baker's Map.

Predictability is reflected in the error doubling time, but not by the Lyapunov exponents. The extent to which this occurs in other systems will depend on how localized (and densely populated) the regions of maximum stretching are; here we note that there are extended regions within which all uncertainties *decrease* with time for chaotic attractors in the Lorenz, Moore-Spiegel and Rössler systems (Smith *et al.*1994; Ziehmann-Schlumbohm 1994; Ziehmann-Schlumbohm *et al.*1995). In such regions, any initial uncertainty will be damped, an unexpected occurrence in these paradigm chaotic attractors. This not only eases the prediction of chaotic systems, but may also be of practical importance, for example, in determining to best times at which to perturb a system one is attempting to manipulate. In addition, characterizing the decay of predictability by the growth of errors will be complicated by the lack of independence between consecutive errors. In fact, the correlation of errors in state space provides the basis for a test for residual predictability, which we will consider in the next section.

6.2. Modes of prediction : Direct and iterative forecasts

Variations in predictability will also influence the choice between constructing iterated predictors (which forecast a short time ΔT ahead and then are iterated $T/\Delta T$ times to get a period T forecast) and direct predictors (for which $\Delta T = T$). Farmer and Sidorowich (Farmer and Sidorowich 1988) argue that, in the limit of good predictors, iterative forecasts are preferred. As they note, the argument assumes a simple relation between the error growth and Lyapunov exponents; it relies on uniformity in the stretching in state space and fails, for example, in data dense regions where the magnitude of the predicted change in one time step is small compared to the observational noise. The choice between iterative and direct prediction will vary in a manner quite like that of the optimal neighborhood size; given sufficient data the choice of ΔT should also be made locally.

7. Detecting Residual Predictability

We began by examining the structure of error doubling times on a strange attractor and observed organization in this structure. To conclude, we suggest adopting the same approach to look for residual predictability, but using a "color code" based on the (signed) prediction error, rather than the doubling time. Organization in these errors means improved prediction is possible. Figure 8 provides an example based on 1024 test points on the Ikeda attractor and a global radial basis function prediction of the next value of x; those points at which the prediction error is positive are denoted by a plus, while those with negative errors are marked with dot. The large scale structure in the distribution of prediction error is clear (although the magnitude may, of course, be quite small). In higher dimensions, where visualization is more difficult, we test for structure in the residuals by considering nearest neighbor pairs in the reconstruction space, and simply count the number of pairs where both of the associated prediction errors have the same sign. If the residuals are independent and identically distributed (IID), then the expected number of pairs with like signs is easily computed. As an IID sequence should remain IID under any general regrouping (Dawid 1984; von Mises 1957), contrasting the expected and observed number of pairs with like signs provides an immediate, quantitative evaluation of the prediction scheme. There are, of course, more powerful tests for structure in the residuals; one example being the BDS test, which considers delay reconstructions of the series of the residuals themselves (Brock *et al.*1991). The major advantages of the test presented below are its simplicity and the use of the delay space coordinates of the original data, allowing immediate access to the small length scales in state space where traces of predictability are most likely to be found.

The test is based on the observation that errors in the predictions of nonlinear systems are likely to be correlated in state space due to systematic biases (*e.g.* E_F). To determine whether nearest neighbor pairs in reconstruction space have independent residuals:

i) Consider original delay reconstruction coordinates as a collection of points in M dimensional space.

ii) Classify each point into one of two groups depending on whether the (signed) prediction error at that location was positive or negative (alternatively, if it is above or below its median value).

iii) Determine unique nearest neighbor pairs, and count the number pairs in which both neighbors belong to the same group.

iv) Test the resulting distribution (number of pairs the same, number different) against the null hypothesis that the points are randomly classified.

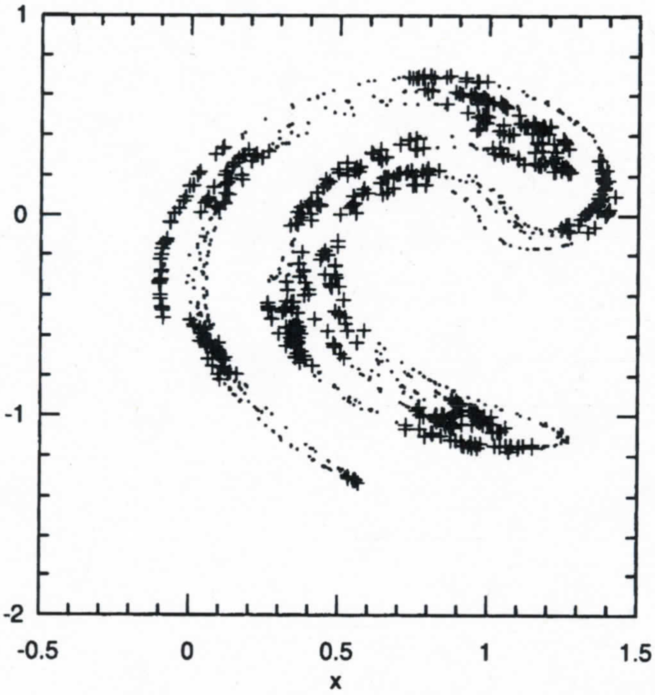

Fig. 7. Spatial organization in the residuals of the Ikeda map for a global radial basis function predictor. Points marked "+" have positive residuals, while those marked "." are negative. Note the large scale structure within which points tend to cluster into a fairly small number of regions of like type.

When coherence is detected, the test may be repeated for larger collections of points (triples,...) to determine the scale of the correlations. Figures 8 and 9 show the results for a global radial basis function fit and a local quadratic predictor of the Ikeda map, respectively. The correlation in the signs of the errors is apparent by eye. This large scale variation in the residuals of global fits is commonly observed when the embedding dimension is small enough to view the fit. The correlation lengths in the local quadratic predictor are smaller, but the null hypothesis of independent errors can still be easily rejected. Rejection of the null hypothesis is, of course, sufficient to show that significant residual predictability remains in a reconstruction, but it is **not** a necessary condition. While less powerful than the BDS test, it is simpler to implement and interpret. Our test is most likely to err on the side of underestimating the remaining predictability, indicating that a given prediction scheme is doing better than it truly is; we have not yet found such "false positive" results to be a problem. Indeed, we detect residual predictability in almost all of the chaotic systems we have examined.

8. Conclusions

The difficulty of predicting the evolution of a nonlinear system varies with the state of the system. Fortunately, this variation is highly organized, and we may exploit it

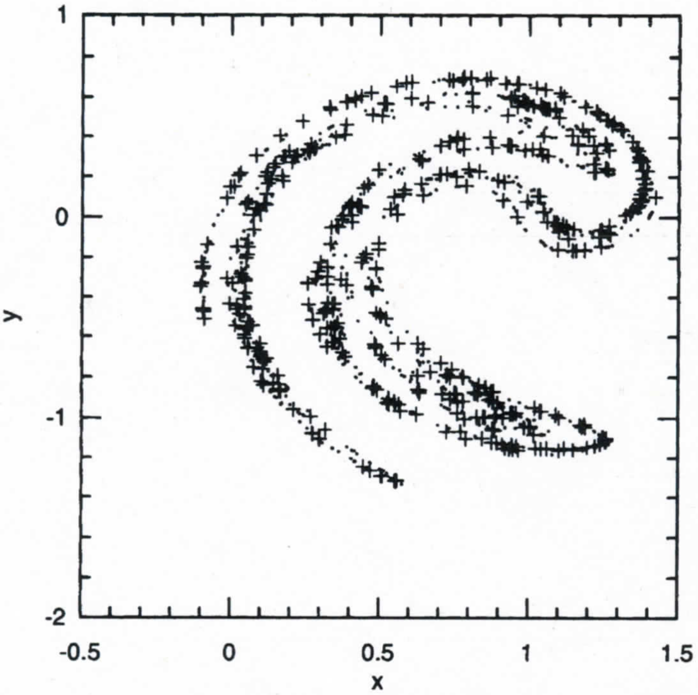

Fig. 8. Spatial organization in the residuals of the Ikeda map for 8 neighbor, local quadratic prediction. Once again, points marked "+" have positive residuals, while those marked "." correspond to negative residuals. In contrast to figure 7, the points with positive and negative error are much better mixed, but still tend to cluster into groups of like type.

both to improve our predictions and our estimate of the degree to which we believe them. This organized variability also requires careful consideration when attempting to distinguish aleatoric and chaotic signals, and when interpreting global "bounds" on predictability and the statistics of error growth; the average error doubling time may paint a very different picture of predictability than the time at which the average error has doubled.

In this paper, we have discussed the growth of small, usually infinitesimal, uncertainty in an initial condition. In practice, any functional definition of a "limit of predictability" should allow an uncertainty to attain a magnitude comparable with the range of observed values. Consider the evolution of an ensemble of initial conditions on the Lorenz attractor, all initially within a region of state space which makes them *indistinguishable*, given a series of three, 8 bit measurements. The behavior of such ensembles reveals a wealth of phenomena (Palmer 1993; Smith 1994; Smith *et al.*1994), we draw attention to only three here. First, the distribution may initially sharpen, indicating that in the short term, initial uncertainties decrease with time (Smith *et al.*1994; Tong and Moeanaddin 1988; Ziehmann-Schlumbohm 1994; Ziehmann-Schlumbohm *et al.*1995). Second, the distribution will become multimodal fairly quickly; this will cause difficulty if we tune our predictors by minimizing the squares of the prediction error. The "best" least squares predictor will predict the mean of the distribution, which may not reflect the value of any element in the ensemble. Third, even after large times when the distribution is spread over the entire range of observable values, as long as the distribution can be distinguished from the projection of the invariant measure onto our measurement function, then meaningful conclusions can be drawn from the distribution. The limit of predictability is reached when the ensemble can no longer be distinguished from this projection of the invariant measure (Smith 1994). Needless to say, this limit will vary with the location of the initial conditions.

The goal of nonlinear prediction is to remove all recognizable structure from the time series. For chaotic systems, it is not possible eliminate all structure in the series of out of sample prediction errors without a perfect predictor; to detect this structure in the residuals, it is useful to work within the framework of the original reconstruction space of the observations. Ultimately, our goal is to exploit the structure in the dynamics to minimize structure in the residuals.

Acknowledgements

The research presented in this paper has greatly benefited from many exchanges with C. Ziehmann-Schlumbohm. It is also a pleasure to acknowledge both technical assistance and mathematical discussions with D. Drysdale, A. Fowler, K. Fraedrich, and M. Muldoon. I am grateful to S. Ellner, B. LeBaron, B. Seifert, R. Smith and H. Tong for both suggestions and directions through the statistical literature. Issues related to residual predictability have been enlightened by conversations with D. Broomhead and J. Huke and the observations of R. Jones. This work was supported by a Senior Research Fellowship at Pembroke College.

References

Abarbanel,H. D. I., Brown, R., Sidorowich, J. J. and Tsimring, L.S.1993 The analysis of observed chaotic data in physical systems. *Rev Mod Phys*, **65(4)**, 1331–1392.

Barnes, J. A., Sargent, H. H. and Tryon, P. V. 1980 Sunspot cycle simulation using random noise. In *The Ancient Sun* (ed. R.O. Pepin, J.A. Eddy, and R.B. Merrill), pp. 159 - 163, New York, Pergamon.

Benzi, R. and Carnevale, G. F. 1989 A possible measure of local predictability. *Journal of the Atmospheric Sciences*, **46**, 3595-3598.

Brock, W. A., Hsieh, D. and LeBaron, B. 1991 *Nonlinear Dynamics, Chaos, and Instability: Statistical Theory and Economic Evidence.* MIT Press, Cambridge, MA.

Broomhead, D. S. and Lowe, D. 1988 Multivariable functional interpolation and adaptive networks. *J. Complex Systems*, **2**, 321-355.

Broomhead, D. S., Huke, J. P. and Muldoon, M. R. 1992 Linear filters and nonlinear systems. *J. R. Stat. Soc.*, **B(54)**, 373-382.

Casdagli, M. 1992 Chaos and deterministic *versus* stochastic non-linear modeling. *J. R. Statist. Soc. B*, **54(2)**, 303-328.

Casdagli, M., *et al.* 1992 Nonlinear modeling of chaotic time series. In *Applied Chaos* (ed. J. H. Kim and J. Stringer), pp. 335-380, New York, John Wiley.

Dawid A. P. 1984 Statistical theory: The prequential approach. *J. R. Statist. Soc. A*, **147(2)**, 278-292.

Doerner, R., *et al.* 1991 Predictability portraits for chaotic motinons. *Chaos, Solitons and Fractals*, **1**, 553-571.

Eckmann, J. P. and Ruelle, D. 1985 Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, **57**, 617-656.

Eubank, S. and Farmer, D. 1990 An introduction to chaos and randomness. In *Proc. SFI Summer School* (ed. E. Jen), Addison-Wesley.

Farmer, J. D. and Sidorowich, J. 1987 Predicting chaotic time series. *Phys. Rev. Lett.*, **59**, 8.

Farmer, J. D. and Sidorowich, J. 1988 Exploiting chaos to predict the future and reduce noise. In *Evolution, Learning, and Cognition* (ed. Y. C. Lee), pp. 277. World Scientific.

He, X. and Lapedes, A. 1993 Nonlinear modeling and prediction by successive approximation using radial basis functions. *Physica D*, **70**, 289-301.

Lorenz, E. N. 1963 Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130.

Lorenz, E. N. 1969 Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636-646.

Nese, J. M. 1989 Quantifying local predictability in phase space. *Physica D*, **35**, 237-250.

Olsen, L. F., Valeur, K. R., Geest, T., Tidd, C. W. and Schaffer, W. M. 1994 Nonlinear forecasting of non-uniform chaotic attractors in an enzyme reaction. *Phil. Trans. R. Soc. Lond.* **A, 348(1688)**, 421-430.

Palmer, T. N, 1992 Ensemble prediction. *Meteorological*, **(58)**, 5-15 (June).

Palmer, T. N. 1993 Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteo. Soc.*, **74(1)**, 49 (June).

Palmer, T. N., Buizza, R., Molteni, F., Chen, Y. C. and Corti, S. 1994 Singular vectors and the predictability of weather and climate. *Phil. Trans. R. Soc. Lond.* **A**, **348(1688)**, 459-475.

Sauer, T. 1993 Time series prediction using delay coordinate embedding. In *Predicting the Future and Understanding the Past: A Comparison of Approaches* (ed. A. Weigend and N. Gershenfeld), **volume XV** of *SFI Studies in Complexity*, pp. 175-194, New York, Addison-Wesley.

Sauer, T., Yorke, J. A. and Casdagli, M. 1991 Embedology. *J. Stat. Phys.*, **65**, 579-616.

Smith, L. A. 1992 Identification and prediction of low-dimensional dynamics. *Physica D*, **58**, 50-76.

Smith, L. A. 1993 Do Lyapunov exponents limit predictabilty? Preprint.

Smith, L. A. 1993 Does a meeting in Santa Fe imply chaos? In *Predicting the Future and Understanding the Past: A Comparison of Approaches* (ed. A. Weigend and N. Gershenfeld), **volume XV** of *SFI Studies in Complexity*, pp. 323-344, New York, Addison-Wesley.

Smith, L. A. 1994 Visualising predictability with chaotic ensembles. In *Advanced Signal Processing: Algorithms, Architectures and Implementations* (ed. F.T. Luk), **volume 2296**, Bellingham, WA, SPIE. to appear.

Smith, L. A., Ziehmann-Schlumbohm, C. and Fraedrich, K. 1994 The limits of predictability: The decrease of uncertainty and return of skill in the lorenz system. Pre-print.

Smith, L. A. and Drysdale, D. 1994 Local optimal prediction of low dimensional dynamics. In preparation.

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

Takens, F. 1981 Detecting strange attractors in fluid turbulence. In *Dynamical Systems and Turbulence* (ed. D. Rand and L. S. Young), **volume 898**, pp. 366, New York, Springer-Verlag.

Theiler, J. 1986 Spurious dimension from correlation algorithms applied to limited time-series data. *Phys. Rev. A*, **34(3)**, 2427-2432.

Theiler, J. and Eubank, S. 1994 Don't bleach chaotic data. *Chaos*, To appear.

Theiler, J., Eubank, S., Longtin, A., Galdrikan, B. and Farmer, J. D. 1992 Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, **58**, 77.

Tong, H. 1990 *Non-Linear Time Series Analysis.* Oxford Univ. Press, Oxford.

Tong, H. and Moeanaddin, R. 1988 On multi-step non-linear least squares prediction. *The Statistician*, **37**, 101-110.

von Mises, R. 1957 *Probability Statistics and Truth.* George Allen and Unwin, London.

Weigend, A. and Gershenfeld, N. (editors 1993) *Predicting the Future and Understanding the Past: A Comparison of Approaches*, **volume XV** of *SFI Studies in Complexity.* Addison-Wesley, New York.

Weiss, N. O. 1990 Periodicity and aperiodicity in solar magnetic activity. *Phil. Trans. R. Soc. Lond.*, **A 330**, 617-625.

Wolpert, D. 1992 Stacked generalization. *Neural Networks*, **5**, 241-259.

Ziehmann-Schlumbohm, C. 1994 *Vorhersagestudien in chaotischen Systemen und in der Praxis - Anwendung von Methoden der nichtlinearen Systemanalyse.* PhD. Thesis, *Freie Universität Berlin. Meteorologische Abhandlungen N.F. Serie A Monographien.*

Ziehmann-Schlumbohm, C., Fraedrich, K., and Smith, L. A. 1995 Ein internes vorhersagbarkeitsexperiment im lorenz-modell. *Meteorologische Zeitschrift N.F.*, **14**.