International Journal of Bifurcation and Chaos, Vol. 10, No. 6 (2000) 1407–1422

MODEL SELECTION, CONFIDENCE AND SCALING IN PREDICTING CHAOTIC TIME-SERIES

ERIK M. BOLLT*
Department of Mathematics, 572 Holloway Rd., U.S. Naval Academy,
Annapolis, MD 21402-5002, USA

Received April 23, 1999; Revised October 5, 1999

Assuming a good embedding and additive noise, the traditional approach to time-series embedding prediction has been to predict pointwise by (usually linear) regression of the k-nearest neighbors; no good mathematics has been previously developed to appropriately select the model (where to truncate Taylor's series) to balance the conflict between noise fluctuations of a small k, and large k data needs of fitting many parameters of a high ordered model. We present a systematic approach to: (1) select the statistically significant neighborhood for a fixed (usually linear)model, (2) give an unbiased estimate of predicted mean response together with a statement of quality of the prediction in terms of confidence bands.

1. Introduction

Predicting the future evolution of dynamical systems has been a main goal of scientific modeling for centuries. The classic approach has been to build a global model, based on fundamental laws, yielding a differential equation which describes the motion of states. "This requires strong assumptions. A good fit of the data to the model validates the assumptions," [Weigenbend & Gershenfeld, 1993]. Weigend and Gershenfeld make a distinction between weak modeling (data-rich and theory-poor) and strong modeling (data-poor and theory-rich). This is related to, "... the distinction between memorization and generalization ...". It is always nice to have a general theory from which we may write down a global set of equations of motion. However, this is not always necessary.

If the time-series has been generated by a "chaotic" dynamical system, data-only based analysis, using the methods of embedding and attractor reconstruction, has become routine [Weigenbend & Gershenfeld, 1993; Abarbanel *et al.*, 1993; Kantz & Schrieber, 1997; Abarbanel, 1996; Farmer & Sidorowich, 1987, 1988]. Suppose that an autonomous dynamical system,

$$\dot{x} = F(x) \,, \quad x(t) \in \Re^n, \quad \text{and} \quad x(t_0) = x_0 \,, \tag{1}$$

has an invariant attractor A. In general, the experimentalist who does not know the underlying global model Eq. (1) does not even know which are the correct variables to measure. Generally, any single-channel data collected can be considered to be a scalar measurement function $h[x(t)] : \Re^n \to \Re$. Given a set of measurements $\{h[x(t_i)]\}_{i=0}^{N}$, taken at uniformly spaced times t_i, the method of time-delay embedding is to form the vector,

$$\begin{aligned}\mathbf{y}(t) = \langle h[x(t)],\, h[x(t-\tau)],\, h[x(t-2\tau)], \ldots,\\ h[x(t-d\tau)]\rangle \,,\end{aligned} \tag{2}$$

and one generally chooses τ to be some multiple of the sampling rate $\Delta t = t_{i+1} - t_i$. Takens proved [Takens, 1980] that, for topologically generic measurement function h, if the attractor A is a smooth

*E-mail: bollt@nadn.navy.mil; web: http://www.usna.edu/MathDept/faculty/emb.html

m-dimensional manifold, then if one chooses the delay dimension to be $d \geq 2m+1$, then Eq. (2) is an embedding, meaning there exists a one-to-one function $G : A \to \Re^d$, and G is a diffeomorphism. Sauer *et al.* [1991] proved an extension to allow for nonsmooth A, and even fractal A. To reconstruct the attractor, both of these results assume that the data is clean, and the data set is arbitrarily long. Neither assumption is physically realizable, but nonetheless, time-delay reconstruction has found many applications to nonlinear modeling and to prediction. See [Abarbanel, 1996; Kantz & Schrieber, 1997; Abarbanel *et al.*, 1993; Farmer & Sidorowich, 1987; Eckmann & Ruelle, 1985].

Local linear regression of the observed evolution of k-nearest neighbors $\{\mathbf{y}_j(t)\}_{j=1}^k$, to their images $\{\mathbf{y}_j(t+\tau)\}_{j=1}^k$, has emerged as the most popular method to predict "the next $y(t)$." The idea is that a Taylor's series of the (unknown) function f_τ, which evolves (flows) initial conditions $y(t)$, according to the differential equation, Eq. (1), is well approximated by the linear truncation, *if the near neighbors are "near enough."* Error analysis, such as that found in [Farmer & Sidorowich, 1988], is based on this local-truncation error, and therefore considers the Luyapunov exponents. However, little attention has been paid to the delicate balance of competing needs of accurate local regression predictions based on "nearby" observations:

1. Small local truncation error demands that neighborhoods be small, and therefore k must not be chosen too large, using a fixed (linear) model.
2. Statistical fluctuations demand that k be chosen large enough to infer a degree of smoothing.
3. Since a local polynomial model regresses the first several terms of a Taylor approximation, attempting to improve local truncation error by increasing the model degree (say to the quadratic term) comes at a cost of an explosion in the number k necessary to fit the many new parameters, and hence likely a decrease in the resulting smoothing.

A popular method [Walker, 1998; Lichtenberg & Lieberman, 1983] is to choose k equal to twice the number of parameters that is fitting, but we argue that this is no more than a "rule-of-the-thumb," as it does not adequately address issues 1–3. While some authors have noted only moderate success using local ordinary least squares to predict a noise-corrupted dynamical system (see [Kugiumtzis *et al.*, 1998] where the deterioration of local predictions for ill-chosen neighborhoods size k was recognized), we argue in this article that the failure of "OLS" is only a matter of choosing the correct scale, a matter which we aim to remedy. Sauer [1992] has recommended roughly choosing k to make the neighborhood, "... around the noise size of the data." A systematic method to choose k, due to Smith [1994], recognizes the delicate balance between local truncation errors, versus minimum data needs for statistical smoothing. Smith chose "optimal k" which minimizes *observed* prediction error, based on trials using a comparison of predictions of the data set divided into training and comparison sets. The technique we introduce in this article is philosophically different, in that we choose k to make a "statistically significant model" in the appropriate window.

A main purpose of this article is to present a systematic statistical analysis to choose, pointwise, the appropriate value of k near neighbors, to make *unbiased* predictions based on a statistically significant but unbiased polynomial regression. We will locally apply an hypothesis test to determine the critically last significant k. Then, using this optimal neighborhood size k to make predictions, in which the polynomial model is statistically significant, and hence unbiased, it is valid to complete the regression analysis with an analysis of variance (ANOVA) on the prediction. Prediction confidence is the most interesting application of the ANOVA when predicting nonlinear time-series. Giving the prediction, together with the, say 95%, confidence bands, is the second purpose of this article. We validate our analysis with real and simulated data sets.

2. Locality Analysis of Variance

For the sake of specificity to demonstrate the problem of scales when choosing k, we show a one-dimensional case of a noisy logistic map, $x_{n+1} = 4x_n(1-x_n)$, and $x_i \to x_i + \varepsilon_i$, where ε_i are independent normal random variable with standard deviation $\sigma = 0.1$. See Fig. 1, where in three successive window sizes (three values of k), the linear model is [Fig. 1(b)] biased, [Fig. 1(c)] significant and unbiased, [Fig. 1(d)] insignificant. Making the problem of scales particularly obvious, we can see that in the smallest window [Fig. 1(d)], the slope of the regressed line even has the wrong sign, and likewise the concavity of the regressed quadratic has the

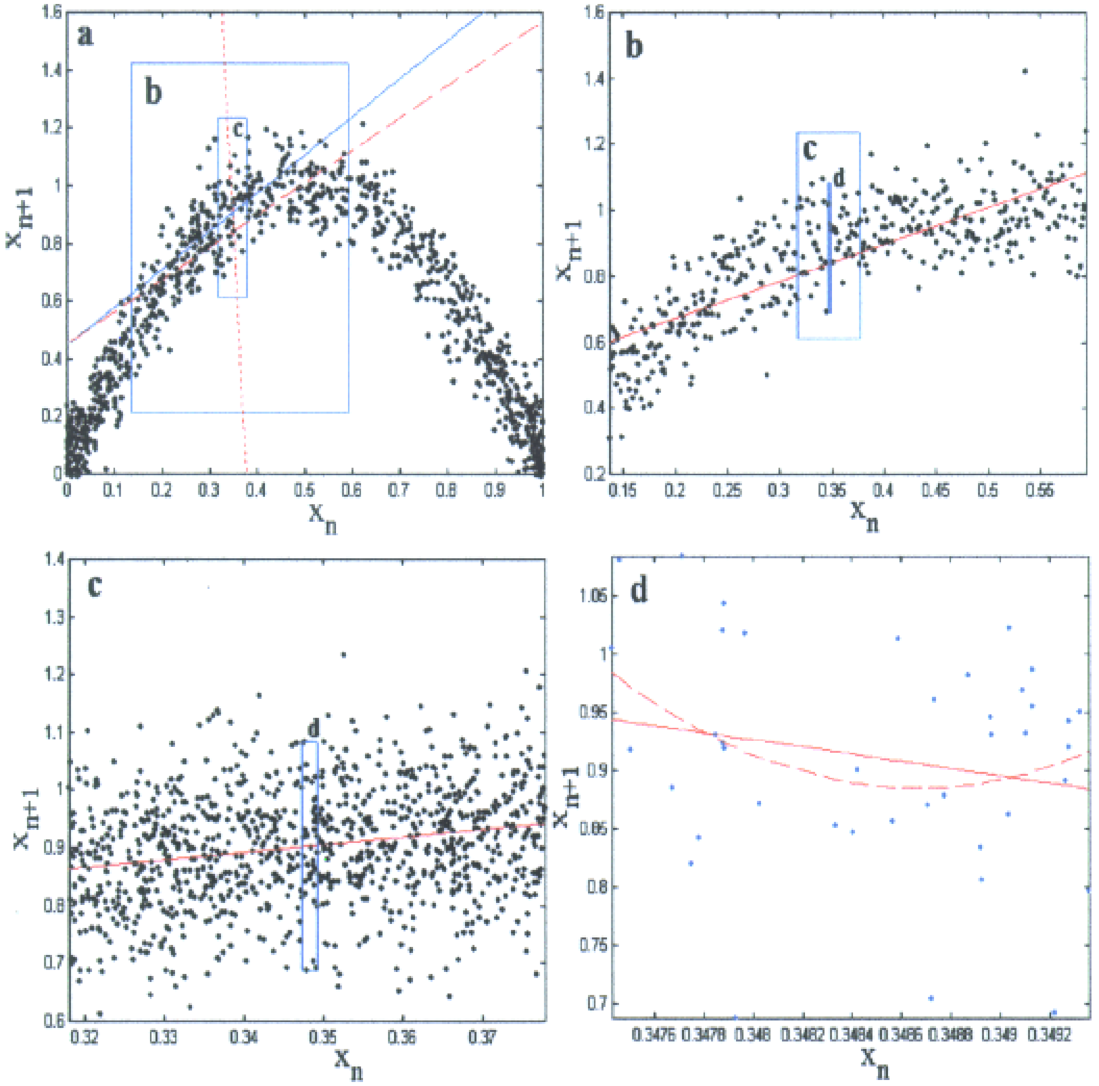


Fig. 1. Four progressively smaller windows of logistic map. (a) Full unit interval, (b) linear model is biased by strong quadratic curvature, (c) best window for a linear model since a quadratic adds no "significant" modeling contribution to a linear prediction, (d) window is too small for a linear or a quadratic model, as we see that these regressed models even have the wrong slope and concavity. The significant model in window (d) would be the constant model $f(x) = c$.

wrong sign. We stress that this is only an accident of the given sample of random variables, and in the small window, regression is overly sensitive to the noise. It can be said that the box is so thin, that the vertical dimension of the box is dominated by noise volume, σ, which does not diminish as we zoom. In this smallest window, only a constant term model, the mean value, is statistically significant. On the other hand, in the big window, [Fig. 1(b)], the regressed line becomes biased by curvature.

We rename the embedded data vector, Eq. (2), $\mathbf{y}(t)$ at time t to be $\mathbf{z}_i$, and the flow which advances $\mathbf{y}(t)$ to $\mathbf{y}(t+\tau)$ can be identified as a map T on the embedded manifold,

$$\mathbf{z}_{i+1} = T(\mathbf{z}_i)\,. \tag{3}$$

Hence, we have discrete orbit data $\{\mathbf{z}_j\}_{j=1}^{N-d}$. There are two kinds of noise which we consider. Additive measurement noise is added to the data generated by the deterministic rule Eq. (3), $\mathbf{z}_j \to \mathbf{z}_j + \varepsilon_j$. On the other hand, modeling error is described by a stochastic model in which noise is added before each next iterations, $\mathbf{z}_{i+1} = T(\mathbf{z}_i) + \varepsilon_i$.

Suppose we wish to predict the next state of an initial condition $\mathbf{w}$. A window is defined by the k-nearest neighbors to $\mathbf{w}$, this region $U(\mathbf{w}, k)$ can be taken to be the convex-hull of the closest points from the data-set $\{\mathbf{z}_j\}_{j=1}^{N-d}$. We will assume that in a fixed-sized window, $U(\mathbf{w}, k)$, that there is a "statistically significant" and unbiased polynomial model of a truncated Taylor's series of $T|_{\mathbf{w}}$. Note that to respect the topological embedding described by Eqs. (2) and (3), we maintain the multivariate description of the map $\mathbf{z}_{i+1} = T(\mathbf{z}_i) \in \Re^d$, even though statistically, this is not necessary. Inspection of Eq. (2) reveals that in terms of regression, only the first position of the vector, $[\mathbf{z}_{i+1}]_1$, requires prediction. The rest of the "new" delay vector is simply a shifted image of the previous state, $\{[\mathbf{z}_{i+1}]_j = [\mathbf{z}_i]_{j-1}\}_{j=2}^{d}$. Therefore, the regression Eq. (4) amounts to a linear solve in those corresponding parameters, and hence their solution has zero variance.

Since a general multivariate polynomial of degree s, in d-dimensional space has, $d(s+d)!/s!d!$, coefficients, which grows quickly with s and d, we describe the following only for comparison of linear (affine) models versus quadratic models, which already requires that $k \geq 60$ for our $d =$ four-dimensional examples, but the algorithm generalizes with essentially no modification. An affine model of $T|_{\mathbf{w}}$ is $\mathbf{z} = \mathbf{T_0} + \mathbf{DT} \cdot \mathbf{h}$, where $\mathbf{T_0}$ is the average of $\{\mathbf{z}_{\mathbf{k}_j}\}_{j=1}^{k}$ over $U(\mathbf{w}, k)$, $\mathbf{DT}$ is related to the Jacobian derivative averaged over $U(\mathbf{w}, k)$, and $\mathbf{h} = \mathbf{w} - \mathbf{z}$. (We index the k points in $U(\mathbf{w}, k)$ by k_j.) A quadratic model of $T|_{\mathbf{w}}$ is $\mathbf{z} = \mathbf{T_0} + \mathbf{DT} \cdot \mathbf{h} + (1/2)\mathbf{h}^t \cdot \mathbf{H} \cdot \mathbf{h}$, where $\mathbf{H}$ is related to the Hessian matrix of second derivatives. The d parameters of $\mathbf{T_0}$ and the d^2 parameters of $\mathbf{DT}$ may be found by least squares according to the normal equations [Neter *et al.*, 1996],

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{4}$$

which is the convenient matrix form of linear regression, and maintains the same vector form regardless of the degree of the fitted polynomial. For an unbiased model, expectation of the RV is $E(\boldsymbol{\varepsilon}) = \mathbf{0}$. The word "linear" refers to the linearity of coefficients which combine multiple linearly independent terms in combinations. For the affine model, one chooses,

$$\mathbf{Y} = \begin{bmatrix} \mathbf{z}_{k_1+1}^t \\ \mathbf{z}_{k_2+1}^t \\ \cdots \\ \mathbf{z}_{k_k+1}^t \end{bmatrix}, \quad \text{and} \quad \mathbf{X} = \mathbf{X}_1 = \begin{bmatrix} 1 & : & \mathbf{z}_{k_1}^t \\ 1 & : & \mathbf{z}_{k_2}^t \\ \cdots & & \\ 1 & : & \mathbf{z}_{k_k}^t \end{bmatrix}, \tag{5}$$

while for the quadratic model,

$$\mathbf{X} = [\mathbf{X}_1 : \mathbf{X}_2]\,, \tag{6}$$

and,

$$\mathbf{X}_2 = \begin{bmatrix} \mathbf{z}_{k_1,1}\mathbf{z}_{k_1}^t & \mathbf{z}_{k_1,2}\mathbf{z}_{k_1}^t & \cdots & \mathbf{z}_{k_1,d}\mathbf{z}_{k_1}^t & \\ \mathbf{z}_{k_2,1}\mathbf{z}_{k_2}^t & \mathbf{z}_{k_2,2}\mathbf{z}_{k_2}^t & \cdots & \mathbf{z}_{k_2,d}\mathbf{z}_{k_2}^t & \cdots \\ \cdots & & & & \\ \mathbf{z}_{k_k,1}\mathbf{z}_{k_k}^t & \mathbf{z}_{k_k,2}\mathbf{z}_{k_k}^t & \cdots & \mathbf{z}_{k_k,d}\mathbf{z}_{k_k}^t & \end{bmatrix}, \tag{7}$$

is the convenient way to write quadratic terms, and not including the self-terms $\mathbf{z}_{k_i,j}\mathbf{z}_{k_i,j}$ in order not to overestimate. Now formally, in this matrix notation, the fitted parameters are,

$$\mathbf{b} = (\mathbf{X}^t \cdot \mathbf{X})^{-1} \cdot \mathbf{X} \cdot \mathbf{Y}\,, \tag{8}$$

but since normal equations from linear regression are often highly ill-conditioned, the more numerically stable way to solve Eq. (4) is by Pensrose–Pseudo Inverse, or SVD [Press *et al.*, 1992; Golub & Van Loan, 1989].

Nothing mechanical will stop a scientist from performing the above regression for any data set size, as long as k is chosen large enough so that Eq. (4) is over-determined, $k \geq d(d+1)$ for the linear case, or $k \geq (d+2)!/2!(d-1)!$ for quadratic. But as we already discussed, an ill-chosen k gives bad results. Consider the case that k is too big. We wish that Eq. (4) is not ill-conditioned. In terms of expectation,

$$E(\mathbf{b}) = \boldsymbol{\beta}\,. \tag{9}$$

However, if the true model is,

$$\text{Full Model: } \mathbf{Y} = \mathbf{X}_1 \cdot \boldsymbol{\beta}_1 + \mathbf{X}_2 \cdot \boldsymbol{\beta}_2\,, \tag{10}$$

but we omit some vector of terms $\mathbf{X}_2 \cdot \boldsymbol{\beta}_2$ by only assuming the submodel,

$$\text{Submodel: } \mathbf{Y} = \mathbf{X}_1 \cdot \boldsymbol{\beta}_1\,, \tag{11}$$

then it can be shown [Draper & Smith, 1981] that there is an introduced "bias",

$$E(\mathbf{b}_1) = \boldsymbol{\beta}_1 + \mathbf{A} \cdot \boldsymbol{\beta}_2\,,$$
$$\text{where} \qquad \mathbf{A} = (\mathbf{X}_1^t \cdot \mathbf{X}_1)^{-1} \cdot \mathbf{X}_1^t \cdot \mathbf{X}_2\,, \tag{12}$$

is the so-called alias or bias matrix. For the linear submodel, take $\mathbf{X}_1$ to be given in Eq. (5). For the quadratic full model, take $\mathbf{X}_2$ to be all of the quadratic terms in Eq. (6). In our setting, bias corresponds to choosing k too large, in which case the nontrivial local curvature is too large and the varied $\mathbf{X}_2$ in $\mathbf{A}$, Eq. (12), is a nontrivial bias. See Fig. 1(c) for the 1-D illustration.

On the other hand, k to small is just as dangerous, as statistical flucations in observed responses $\mathbf{y}$ will cause significant fluctuations in observed values of the random variables $\mathbf{b}$; see Fig. 1(d) for example. Each random sample of $\mathbf{y}$ will in principle lead to a different sampled $\mathbf{b}_1$, which upon repeated sampling, fills-out a typically elliptical cloud in $\boldsymbol{\beta}$-parameter space, with ellipse center at the (Full model, or unbiased) mean $\boldsymbol{\beta}$. More data sampled gives a better point-estimate of $\mathbf{b}$ to $\boldsymbol{\beta}$.

Balancing these concerns of large versus small scales, we are thus motivated to make the following statement of goal when choosing k:

Statement of Goal: *Choose k as large as possible so that the submodel, Eq. (11), is "significant," but the full model, Eq. (10), is insignificant.*

To well-define "significant," we resort to a statistical hypothesis test [Neter *et al.*, 1996],

$$H_0 : \mathbf{b}_2 = \mathbf{0}\,, \tag{13}$$

or all $k(k+1)/2$ extra coefficients of the full quadratic model, Eq. (10), are "essentially" zero. If this is not found to be true, one concludes the alternative hypothesis,

$$\begin{aligned} H_a : \ & \text{some } [\mathbf{b}_2]_j \neq 0\,, \\ & \text{for some } j = 1,\, 2, \ldots,\ \frac{k(k+1)}{2}\,, \end{aligned} \tag{14}$$

and hence the quadratic part of the model is required.

Given $\mathbf{w}$, a point to predict, our algorithm to find the critical k-neighborhood $U(\mathbf{w},\, k)$, satisfying the stated goal, is as follows. Choose k so large that T is obviously not well approximated by an affine model, say 10% of the data set, and so that Eq. (5) is highly overdetermined. We sort these k-nearest neighbors by distance from $\mathbf{w}$. Then one (or several if the set is large) at a time, we prune this list until we first conclude H_0, at the critical k_{cr}, defining the window $U(\mathbf{w},\, k_{\text{cr}})$.

All that is left is the discussion of statistically concluding H_0 to a given significance level α. In practice, one does not find that modeling noisy data gives exactly $\mathbf{b}_2 = \mathbf{0}$. Rather, one asks that an $1-\alpha$ confidence region (ellipsoid) around the sampled $\mathbf{b}_2$, given by Eq. (8) in the $\mathbf{b}_2$ projection in coefficient space, includes the origin. The statistical F-test for multiple regression [Neter *et al.*, 1996] decides whether the proposed full model is statistically significant relative to a proposed submodel, to significance α. If,

$$\begin{aligned} F* &= \frac{\Delta\text{SSE}}{\Delta(\text{DF})} \div \frac{\text{SSE}_{\text{full}}}{\text{DF}_{\text{full}}} \\ &\leq F(1-\alpha;\, \Delta(\text{DF}),\, \text{DF}_{\text{full}})\,, \end{aligned} \tag{15}$$

then one concludes H_0. Here, $\text{SSE}_{\text{full}} = (\mathbf{Y} - [\mathbf{X}_1 : \mathbf{X}_2] \cdot [\mathbf{b}_1^t : \mathbf{b}_2^t]^t) \cdot (\mathbf{Y} - [\mathbf{X}_1 : \mathbf{X}_2] \cdot [\mathbf{b}_1^t : \mathbf{b}_2^t]^t)^t$, $\text{SSE}_{\text{sub}} = (\mathbf{Y} - \mathbf{X}_1 \cdot \mathbf{b}_1) \cdot (\mathbf{Y} - \mathbf{X}_1 \cdot \mathbf{b}_1)^t$, $\Delta\text{SSE} = \text{SSE}_{\text{full}} - \text{SSE}_{\text{sub}}$, $\text{DF}_{\text{full}} = [\#$ degrees of freedom of Eq. (6)$] = \#\textit{rows} - \#\textit{columns}$ of $\mathbf{X}_1 : \mathbf{X}_2]$, $\text{DF}_{\text{sub}} = [\#$ degrees of freedom of Eq. (5)$] = \#\textit{rows} - \#\textit{columns}$ of $\mathbf{X}_1$ $\text{DF} = \text{DF}_{\text{full}} - \text{DF}_{\text{sub}}$, and $F(1-\alpha;\, \Delta(\text{DF}),\, \text{DF}_{\text{full}})$ is percentiles of the F-distribution; see [Neter *et al.*, 1996; Draper & Smith, 1981].

Said in terms of a local Principal Component Analysis (PCA) [Hediger *et al.*, 1990], using the Singular Value Decomposition (SVD) of the full model Eq. (10), there are $d + d^2 + d^2(d+1)/2$ singular values ω_i, but if the $d^2(d+1)/2$ singular values [Press *et al.*, 1992; Golub & Van Loan, 1989] corresponding to the augmentation term $\mathbf{X}_2 \cdot \boldsymbol{\beta}_2$ are "small" then one accepts the null hypothesis H_0, and small is defined to be "statistically insignificant to level-α," according to Eq. (15).

Finally, once one has determined that the submodel Eq. (5) is statistically significant relative to the full model Eq. (6) then the ANOVA on the submodel is unbiased by the full model. In particular, we are interested in the predicted mean response, $\hat{\mathbf{Y}} = \mathbf{X}_1 \cdot \mathbf{b}_1$, together with the corresponding $1-\alpha$ confidence bands.

We are now in the position to write [Neter *et al.*, 1996],

$$\text{Var}(\hat{\mathbf{Y}}) = \mathbf{w}^t \cdot \boldsymbol{\sigma}^2\{\mathbf{b}\} \cdot \mathbf{w}\,. \tag{16}$$

In practice, one uses the unbiased point estimator of the variance–covariance matrix $\boldsymbol{\sigma}^2\{\mathbf{b}\}$; $\mathbf{s}^2\{\mathbf{b}\} = \text{MSE}_{\text{sub}}(\mathbf{X}_1^t \cdot \mathbf{X}_1)^{-1}$, where mean square error is in terms of sum square of error, $\text{MSE}_{\text{sub}} = \text{SSE}_{\text{sub}}/\text{DF}_{\text{sub}}$. So the estimator of Eq. (16) is $s^2\{\hat{Y}\} = \mathbf{w}^t \cdot \mathbf{s}^2\{\mathbf{b}\} \cdot \mathbf{w}$. Finally, for a single response, one can write the predicted mean new observed response in terms of confidence bands,

$$\hat{\mathbf{Y}} \pm t(1-\alpha/2;\, \text{DF}_{\text{sub}})\sqrt{1 + \mathbf{s}^2\{\mathbf{b}\}}\,, \tag{17}$$

which is in terms of the "student-t" distribution. In this paper, we choose $\alpha = 0.05$, a standard confidence level of 95%.

3. Forecasting

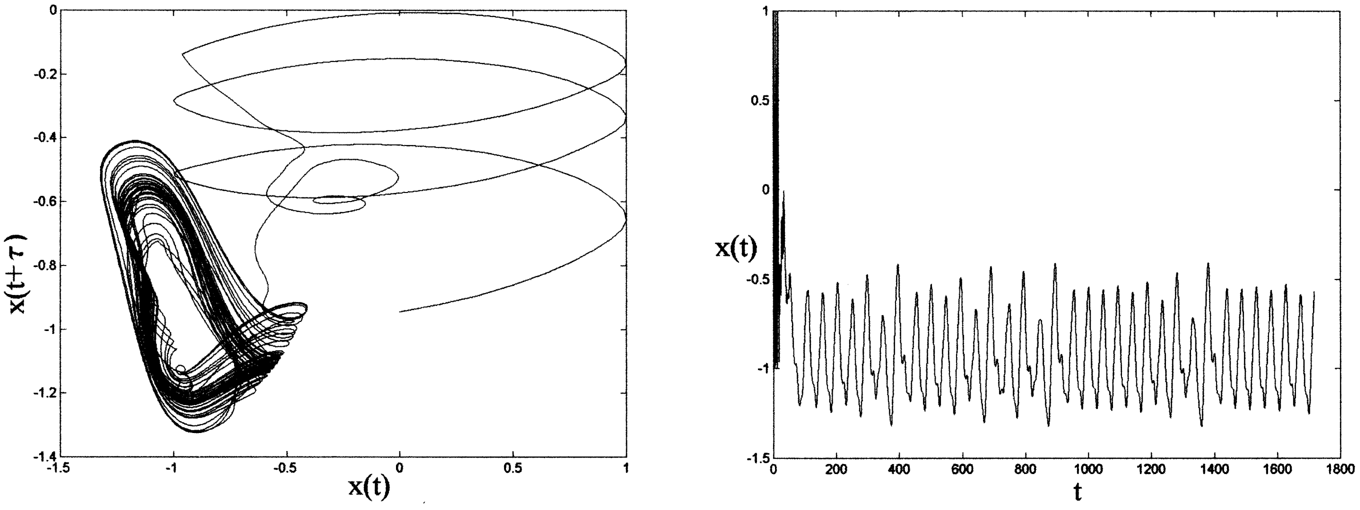
We validate our results with real and simulated data. Also, we check that the confidence bands reported by Eq. (17) do in fact contain the true responses close to the predicted (95%) percentage of the predictions.

Example 1. First we numerically simulate a "noisy" dynamical system by adding white noise to the Mackey–Glass differential delay equations [Mackey & Glass, 1977],

$$x'(t) = \frac{ax(t-t_d)}{1+[x(t-t_d)]^c} - bx(t) + \varepsilon\,, \qquad (18)$$

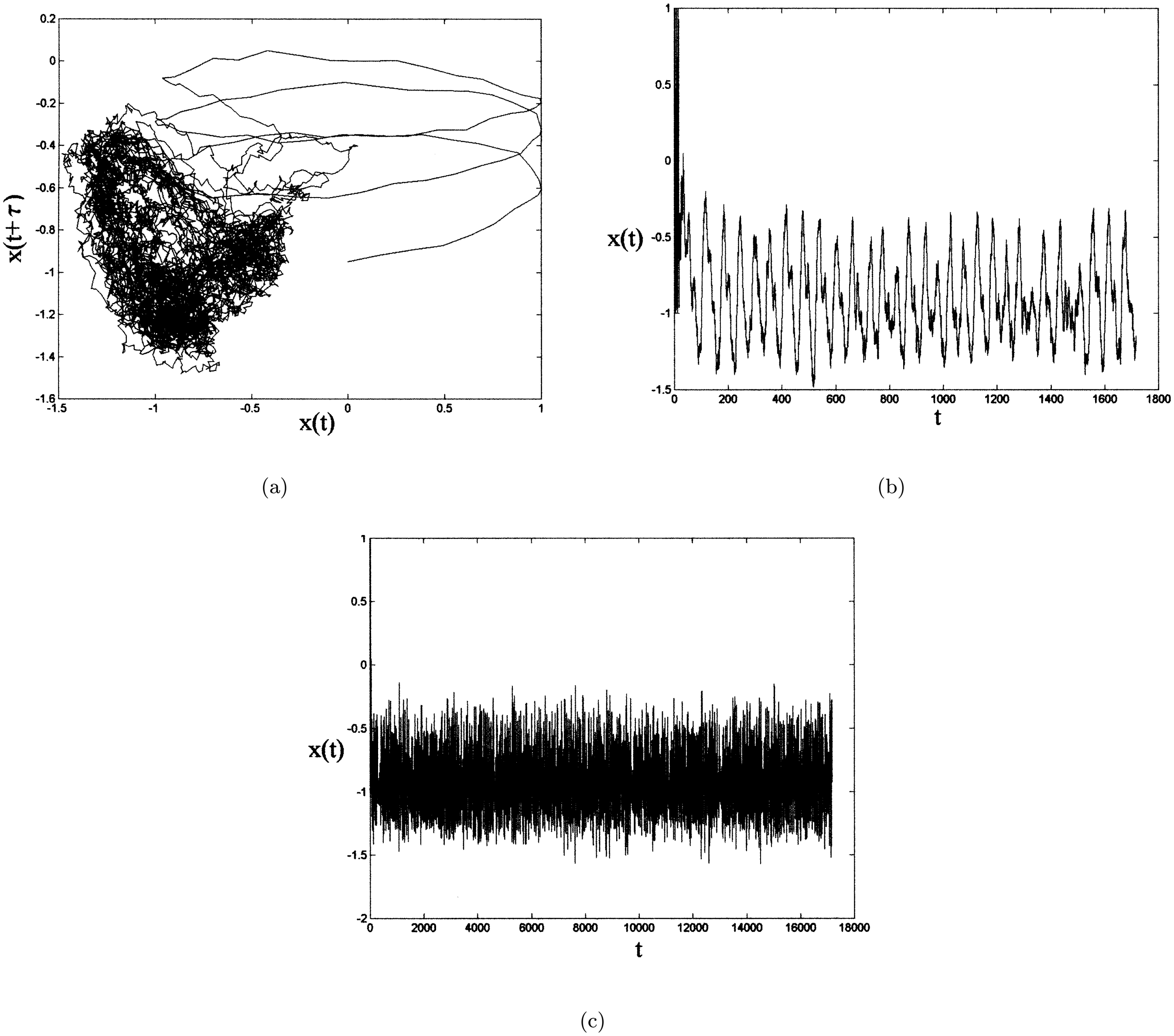
which has become a standard example in time-series analysis [Farmer, 1982; Lichtenberg & Lieberman, 1983] of a high (infinite) dimensional dynamical system with a low-dimensional attractor. This type of stochastic perturbation models "dynamic" noise. We have chosen parameters $t_d = 17$, $a = 0.2$, $b = 0.1$, $c = 10.0$ which give an embedding dimension of $d = 4$. We use integration steps of $\Delta t = t_d/100$ throughout. In Fig. 2(a) we show a short segment of a clean, $\varepsilon = 0$, two-delay projection of the delay attractor, and Fig. 2(b) shows a short segment of the corresponding clean time-series. We have added the stochastic forcing of a normal RV ε with standard deviation $\sigma = 0.1$, for the rest of this study. Figure 3(a) shows this noisy stochastic attractor. Figure 3(b) shows the corresponding stochastic time-series, using the same initial condition as was used in Fig. 2(b). The noise error, however, in Eq. (18) is dynamic. In Fig. 3(c), we show the full $N = 10^5$ data points time-series used in our forecasting experiment. While we will vary delay's τ as allowed by the embedding theorem, we will choose $\tau = 6$ to correspond to one time unit, the first minimum of average mutual information [Farmer & Sidorowich, 1987].

As is usual benchmarking practice, we split the full time-series in Fig. 3(c) into halves, a training set, and a validation set. We then made numerous validations of prediction for multiples of embedding τ = six − steps ahead. Figure 4 shows a small segment of one such experiment, predicting ahead 4τ, using the methods of model and scale selection of k near neighbors as described above (as in Fig. 3(c), displaying too long a segment conveys little information.) Furthermore, using the appropriate k-scale, the local-linear model is unbiased, to significance level $\alpha = 0.05$, and Eq. (17) is expected to give a good confidence estimate. The blue bands are the (locally in embedding space) predicted 95% confidence bands, and the center of the bands is also marked blue, signifying the estimated signal response. The red line is the true response, which



(a) (b)

Fig. 2. Clean Mackay–Glass, Eq. (18), (a) attractor and (b) short time-series, $t_d = 17$, $a = 0.2$, $b = 0.1$, $c = 10.0$, and $\varepsilon = 0$. See Example 1.



(a)

(b)

(c)

Fig. 3. Noisy Mackay–Glass, Eq. (18), (a) attractor, (b) short time-series, and (c) full $N = 10^5$, $\Delta t = t_d/100$ time-series for prediction experiment. $t_d = 17$, $a = 0.2$, $b = 0.1$, $c = 10.0$, and $\varepsilon = 0$. See Example 1.

we can see is in fact usually between the 95% bands. We observed in this experiment that the so-called 95% confidence bands, derived by Eq. (17) and the above procedure, in fact bounded the (red) true response 96.12% of the time. Furthermore, as predicted, we see that the blue bands capture most of the red's noise variance, and when it does wander outside the confidence bands, it usually does not stray far. Please note that the red line wanders widely around the predicted mean response center blue line, thus strengthening the argument that a prediction is only useful together with a statement of the prediction's quality.

Example 2. We now study real infrared NH3 laser data [Huebner *et al.*, 1989], contributed by U. Huebner to the Sante-Fe Institute prediction contest [Weigenbend & Gershenfeld, 1993]. This data set contains $N \approx 10\,000$ points, shown in Fig. 5, for which we find a good embedding in dimension $d = 4$, and we choose the delay $\tau = 2$ to be equal to one time-unit. Again, for benchmarking, we divide the set into halves, training and validation sets. In Fig. 6 we show a short segment of the predicted mean response (center dashed blue line), 95% confidence bands (outer dashed blue lines) and true response (solid-red line). Again, the estimated 95%

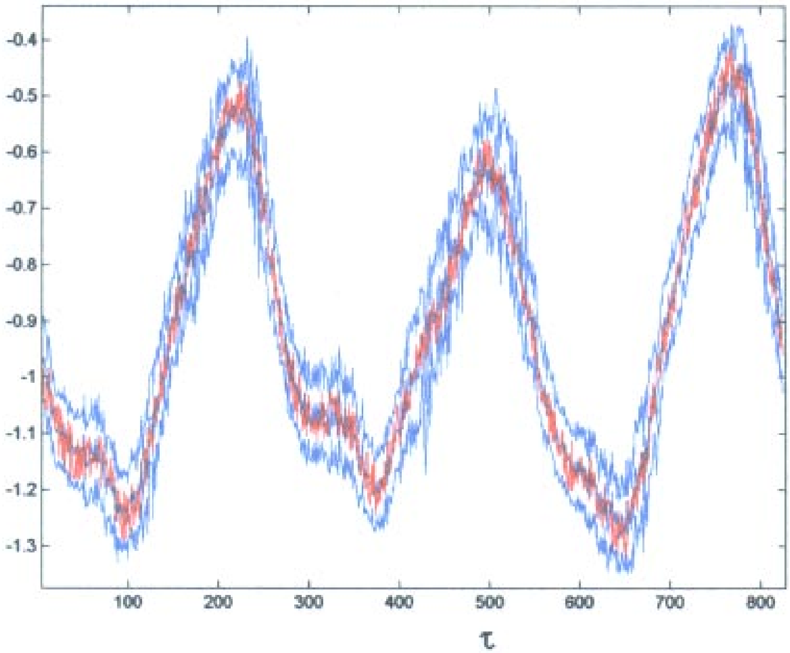


Fig. 4. A short segment of predicted mean responses (center blue line) using the first half, $N/2 = 50,000$, points in Fig. 3(c), together with estimated 95% confidence bands (outer blue lines), and the true observed response (red line) from validation set, data in Fig. 3. See Example 1.

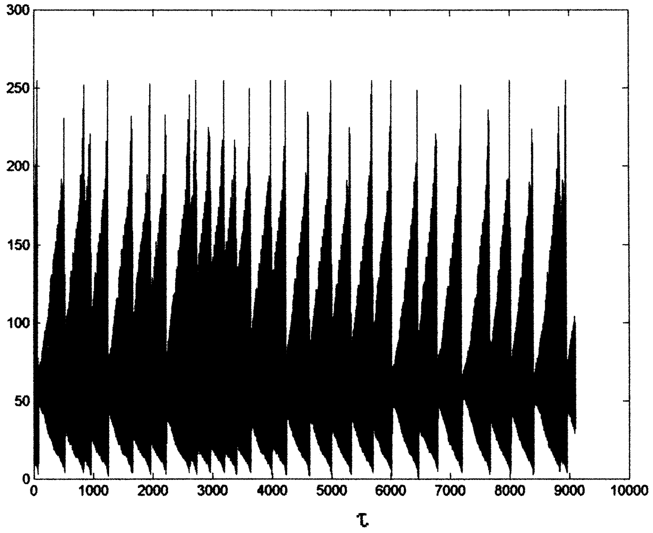


Fig. 5. The $N = 9,093$ points NH3 laser data set from the Sante-Fe Institute prediction contest [Huebner *et al.*, 1989; Weigenbend & Gershenfeld, 1993]. See Example 2.

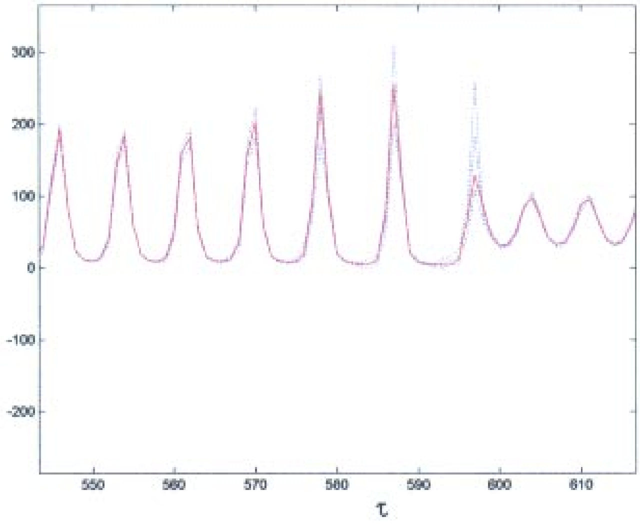


Fig. 6. Predicting laser data ahead $\tau = 2$ time-steps. Predicted mean response (center blue-dashed line), 95% confidence bands (out dashed blue lines) and true response, the solid red line. See Example 2.

confidence bands are close to the observed 97% of the time bounding of the (red) true responses. Observe that the confidence bands are usually quite tight, denoting that this was a very low-noise experiment. Further note that sometimes, particularly at the end of "building segments," the confidence bands are not as tight, denoting either a noisy part of the experiment, and/or low data density in this region of phase space. This is consistent either with wide local variations of Luyapunov exponents as observed in [Farmer & Sidorowich, 1987, 1988], or a region of phase space with a relatively low invariant measure, which in turn results in few near neighbors. In such a case, our technique forces settling on a small value of k, with corresponding high statistical variation on estimated parameters, but nonetheless with perhaps large truncation error. Such would not be a desirable situation, but we have argued that such a situation is nonetheless the best estimate of parameters that the noisy and finite-length data set will support. It is useful to know the scales of prediction accuracy.

4. Long Term Forecasting

There are two techniques whereby one can make forecasts over a time-interval τ. These are m short iterative forecasts over times τ/m, or one direct forecast over the full time interval τ [Farmer & Sidorowich, 1987, 1988; Abarbanel, 1993]. Based on a local-truncation analysis, Farmer and Sidorovich found the counter-intuitive result that iterative forecasts are superior, with rms error scaling as $N^{\frac{-q}{D_1}} e^{\sigma_{\max} T}$. This contrasts to rms error of direct forecasts, $N^{\frac{-q}{D_1}} e^{q\sigma_{\max} T}$, due to the extra q in the exponent, where $q-1$ is the degree of the polynomial, N is the *fixed* neighborhood size, $\sigma_{\max}$ is the largest Luyapunov exponent, and D_1 is the information dimension. Their analysis does not apply to our approach, since they assume a fixed neighborhood size N, which we consider to be a major factor in making good long term predictions. In fact, they typically choose N to be twice the number of parameters to be fit, which according to our results, is often low. Likewise, others [Kugiumtzis *et al.*, 1998] have also found that iterative forecasting is not always superior to direct forecasting, even for fixed neighborhood size.

We now discuss direct long term forecasting. Given data from a chaotic dynamical system, it should be expected that errors grow in time, and hence our ability to predict should degrade in time. Equivalently, our confidence in predictions should degrade with time. We find this reflected by the fact that our 95% confidence bands tend to get wider with time. We choose a simple 1-D model to illustrate issues of scaling, data density and time.

Example 3. We generate $N = 10^3$ data points with a noisy logistic map, $x_{n+1} = 3.8x_n(1-x_n)+\varepsilon_n$, where ε_n are i.i.d. Gaussian RV's with $\sigma = 0.01$. Notice that this is "dynamic noise." Again, the data set is divided into halves, for training and validation. For data generated by a map, rather than a flow, the "natural" delay is $\tau = 1$, one iterate. Choosing $\tau > 1$ shows higher iterates of the map. In Fig. 7, we show direct forecasts over progressively longer times. The three dotted-blue lines are the 95% confidence bands (outer) and mean predicted response (center). The solid-red line is the true response. Confidence deteriorates in time, as reflected by exponentially widening 95% confidence bands, necessary to contain the (red) true response; note

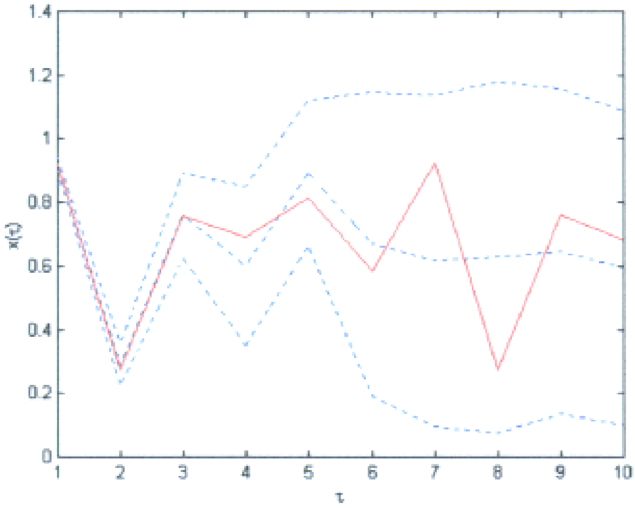


Fig. 7. Direct forecasts of noisy logistic map, $x_{n+1} = 3.8x_n(1-x_n)+\varepsilon$, over progressively longer times, τ. 95% confidence bands, outer blue-dashed lines, predicted mean response, center blue-dashed line, and true response, solid-red line. See Example 3.

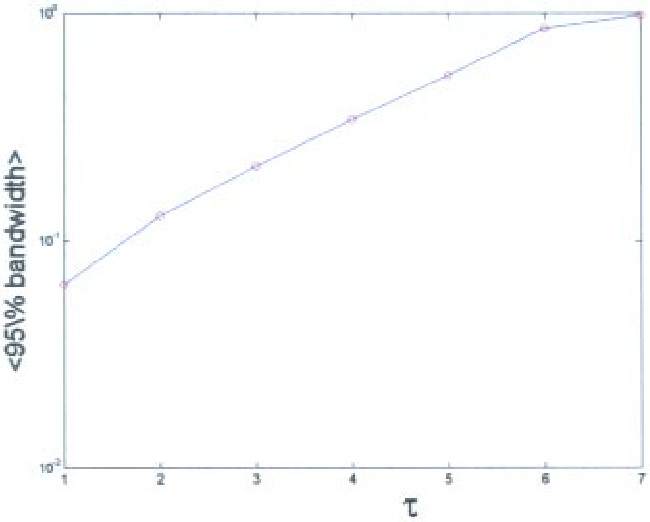


Fig. 8. The mean bandwidth of the noisy logistic map, $\langle 95\%\ \mathit{bandwidth}(\tau)\rangle$, averaged over predicting each of the 500 points of validation set, versus forecast time τ. See Example 3.

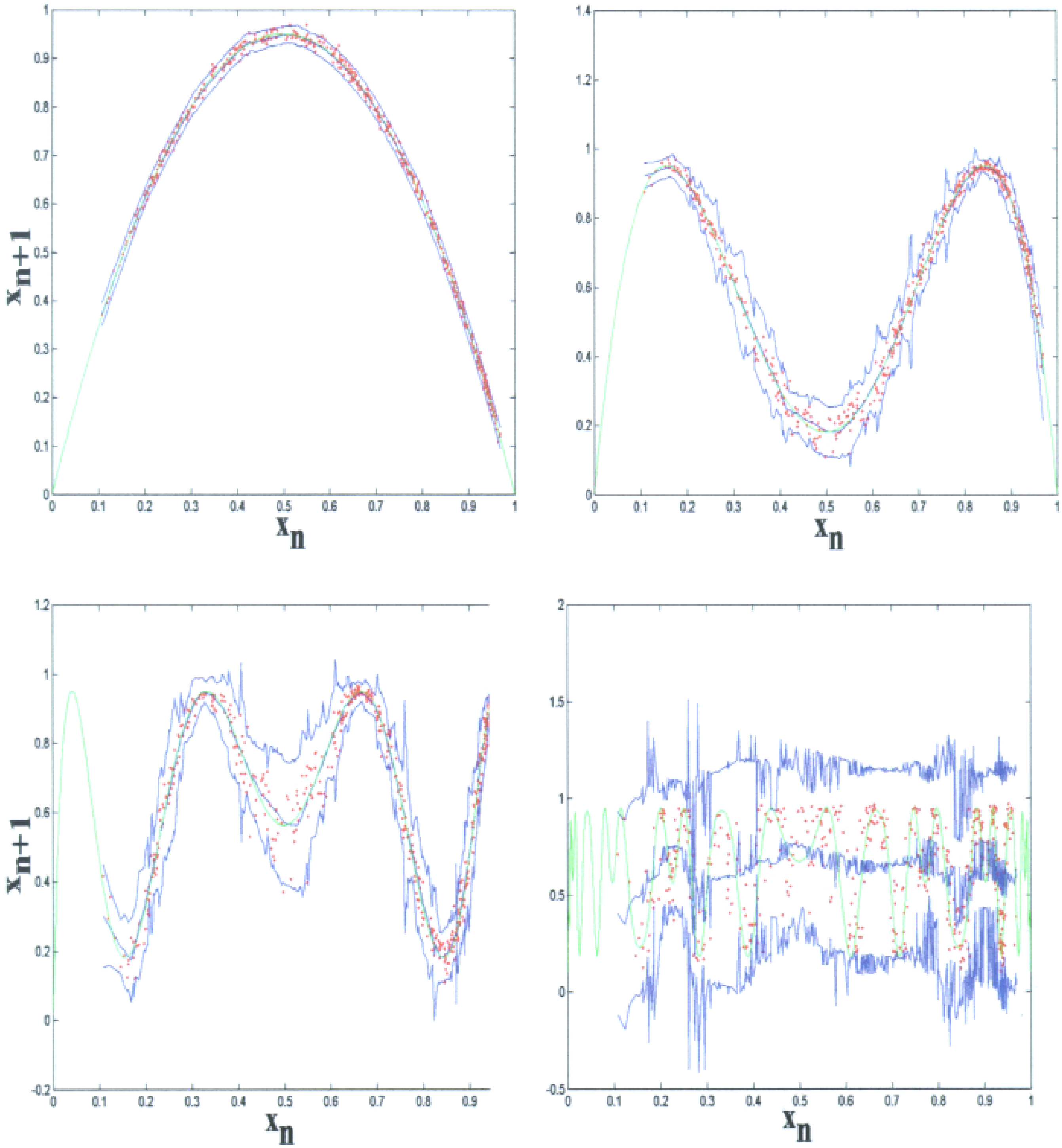


Fig. 9. Noisy logistic map delay embeddings of data (red points) as pairs $(x_i,\, x_{i+\tau})$ for progressively increasing τ. The green curves show clean versions ($\varepsilon = 0$, $x_{n+1} = 3.8x_n(1-x_n)$), of iterates of the map. (a)–(e) show progressively increasing iteration delays, $\tau = 1$, 2, 3, 6, and hence increasingly fast oscillations of the function, and data. Also shown are the 95% confidence bands (outer blue curves), and predicted mean response (center blue). By $\tau = 6$, the data appears stochastic, and hence prediction is the mean. See Example 3.

also that error (center blue-dotted line versus solid-red line) increases with time. By the sixth iterate, the confidence bands have saturated to the entire attractor. In Fig. 8, we show $\langle 95\%\ \mathit{bandwidth} \rangle$, the mean bandwidth averaged over predicting each of the 500 points of validation set, versus forecast time τ. The almost perfect line on the log scale, before saturation, is a reflection of the positive Luyapunov exponent.

In Figs. 9(a)–9(d), we show the delay embedding of data (red points) as pairs $(x_i, x_{i+\tau})$ for progressively increasing τ, $\tau = 1, 2, 3, 6$. The experimentalist has only data points, but to help guide our eyes, we have drawn in green curves, a clean version ($\varepsilon = 0$, $x_{n+1} = 3.8x_n(1 - x_n)$, even though the data was generated with noise) of iterates of the logistic map which generated the data. Progressively higher iterates of the quadratic give progressively higher degree polynomials; the nth-iterate is a 2^n degree polynomial, which requires progressively faster oscillations to fit the extra turning points in the unit interval. This was essentially accounted for by Farmer and Sidorowich's local truncation analysis, though the interpretation was different, as reflected by their use of a fixed number N of near neighbors. Also in Figs. 9(a)–9(d), we show our calculated 95% confidence bands; correctly calculating these bands required that at each iteration, a progressively smaller neighborhood was chosen so that a line significantly fits the increasingly faster oscillating function. This is true until saturation. More data would be necessary to resolve the map in a smaller interval. Our algorithm suffers this forecasting horizon, as seen by the fact that when $\tau = 6$, the predicted mean response is approximately the mean value line of the data, $x \approx 0.5$, which is the center blue line, and the 95% confidence bands are at the extreme of the unit interval. Consider Fig. 10, which shows $\langle k(\tau) \rangle$, the mean number of near neighbors selected to predict the validation set. Neighborhood sizes decrease with increasing iteration, due to faster oscillations and more turning points of a higher degree polynomial, until the minimum at $\tau = 5$. This is due to the data density limit of a fixed training set size of $N = 500$ points. For $\tau > 5$, the noise saturated models have more significance as horizontal lines (the constant model is the lowest degree possible submodel), which automatically tends to select larger neighborhoods.

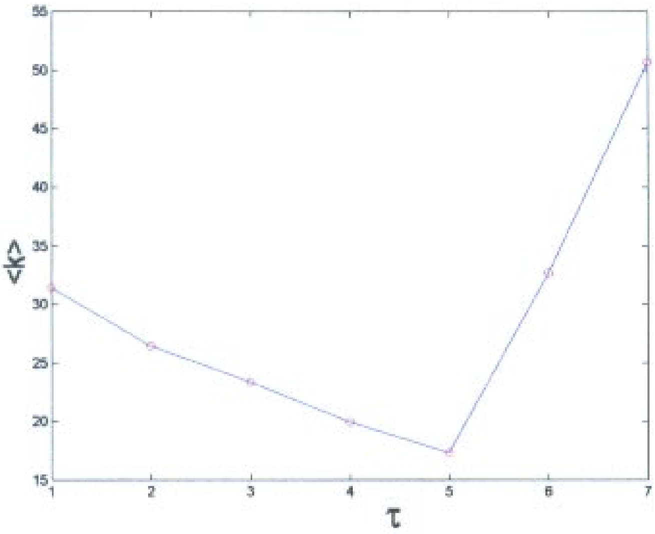


Fig. 10. Mean number of near neighbors selected to predict the validation set, $\langle k(\tau) \rangle$. Neighborhood sizes decrease with increasing iteration, until the minimum at $\tau = 5$. For $\tau > 5$, the noise saturates. See Example 3.

Finally, we discuss a symmetry issue, involving the significance level, which we found only in one-dimensional modeling. Consider Figs. 11(a)–11(c). In Fig. 11(a), we show as red dots, the $k = 36$-near neighbors to predict the pink-star at $x = 0.18$; these are the data points which regress a linear model better than a quadratic model, to significance level $\alpha = 0.05$. In Fig 11(b), we show the $k = 120$ data points in red for the fourth iterate $\tau = 4$, which our (starting from k too large and then decreasing) algorithm selected as more significant than the quadratic model, to significance level $\alpha = 0.05$. A fairly horizontal line was found to be more significant than any improvement due to adding a quadratic term. The reason for this "obviously" too big a neighborhood is due to the fact that we are comparing a linear submodel $y = b_0 + b_1 x$, to a quadratic full model $y = b_0 + b_1 x + b_2 x^2$, even though the full model is truly a 16th degree polynomial $y = \Sigma_{i=0}^{16} b_i x^i$. Rejecting the null-hypothesis H_0 that $b_2 = 0$ does not imply that we are in a small enough neighborhood that a linear approximation is sufficient; it is possible that we are still in a relatively large cubic (or higher) degree region, but $b_2 = 0$. In other words, rejecting H_0, and hence concluding H_a, that the quadratic term is not needed does not necessarily imply that a linear submodel is better than a full model which includes the rest of the truncated Taylor's series: $y = b_0 + b_1 + \Sigma_{i=3}^{\infty} b_i x^i$.

Rather than running an F-test of a linear submodel versus the extra coefficients of a very high degree polynomial, which is generally the safer assumption when truncating a Taylor's series, we have

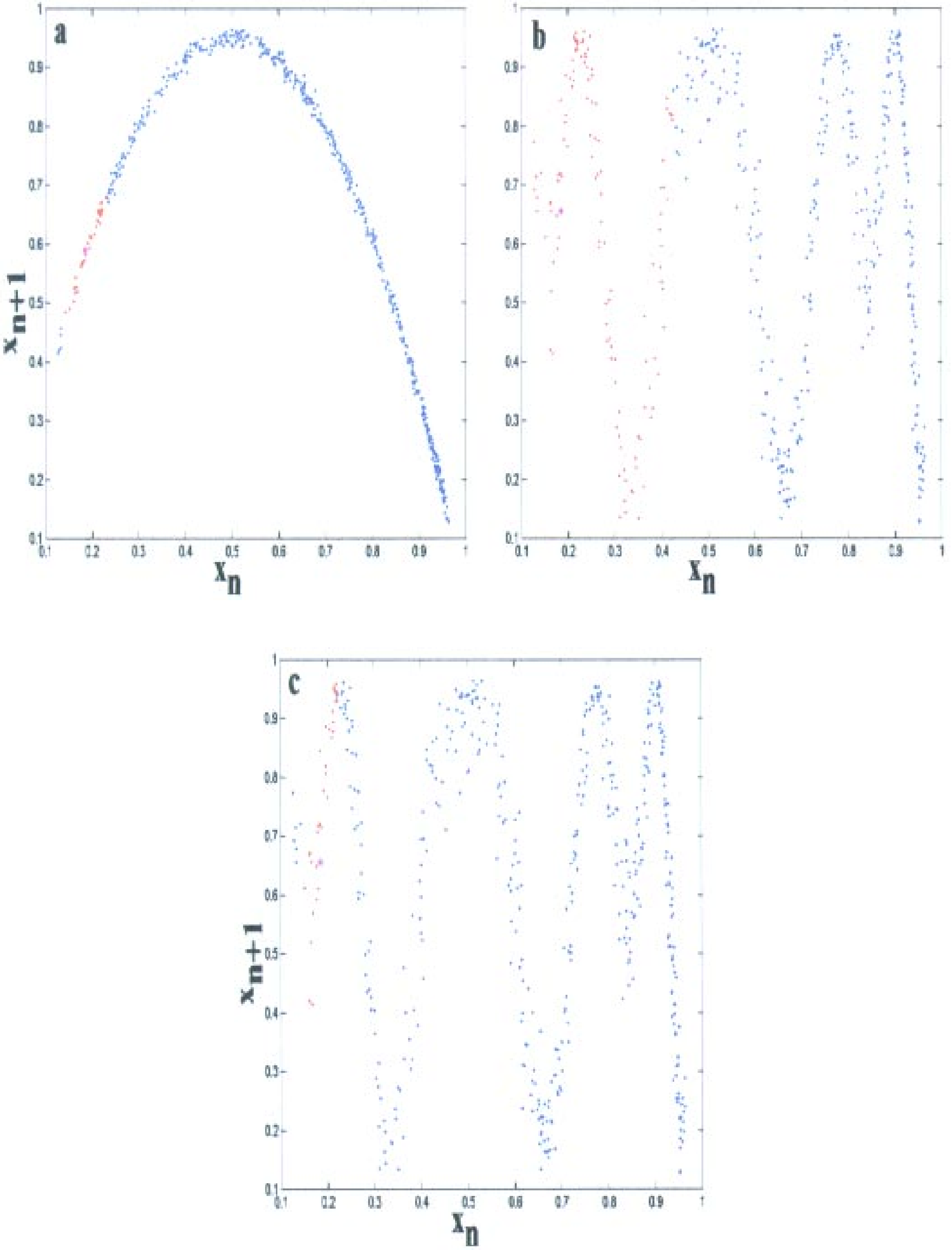


Fig. 11. (a) Given the noisy logistic map, the $k = k_{\mathrm{cr}_1} = 36$-near neighbors (red) to predict the pink-star at $x = 0.18$, which regress a linear model better than a quadratic model, to significance level $\alpha = 0.05$. (b) The $k = k_{\mathrm{cr}_1} = 120$ data points in red, for the 4th iterate selected to significance $\alpha = 0.05$, due to nonquadratic nonlinearity, $y = b_0 + b_1 + \Sigma_{i=3}^{\infty} b_i x^i$. (c) The $k = (k_{\mathrm{cr}_0} + k_{\mathrm{cr}_1})/2 = 32$ points selected with $\alpha = 0.1$. See Example 3.

found the following "hack" gives good results. First, we increase our willingness to make a Type I error, $\alpha = P(\text{Type I error})$, where a Type I error means that H_0 is true, but we incorrectly reject H_0. By adjusting α up slightly from $\alpha = 0.05$ when $\tau = 1$ linearly to $\alpha = 0.1$ when $\tau \geq 4$, we make allowance that the function is likely to be curvier, and we become more pessimistic about evidence that $b_2 = 0$, that the function is not curvy (even though the nonlinearity may in fact be cubic or higher). This tends to push the k_{cr_1} down. Second, we find another k_{cr_0}, which we define to be the point when the constant submodel $y = b_0$ (i.e. the average) is just as significant as a linear full-model, $y = b_0 + b_1 x$. Then we average these two k_{cr}, which gives the effect of placing k in the center of neighborhood sizes which we would call significantly linear. In Fig. 11(c) we show the $k = 32$ points which were selected when predicting the same point, the pink star, with $\alpha = 0.1$ and $k = (k_{\mathrm{cr}_0} + k_{\mathrm{cr}_1}/2)$. We have found this technique to be unnecessary for higher dimensional predictions, whence the symmetry issue is extremely unlikely since it is an uncommon point for all coefficients of the Hessian matrix in the quadratic form of the local Taylor's expansion to be zero; when $d > 1$, we always choose fixed $\alpha = 0.05$ and $k = k_{\mathrm{cr}_1}$.

Example 4. We return to the numerically generated noisy Mackey–Glass data set, Example 1, and the laser data set, Example 2, with all of the same parameters as previously discussed. We now make long-term direct forecasts in higher dimensions. In Fig. 12, we see a long-term forecast of the Mackey–Glass data, with 95% confidence bands, and likewise, we see forecasts of the laser data in Fig. 13, both for increasing delay time τ. In both cases, the confidence bands increase on average with time but not monotonically, as seen in Fig. 14 for MG and laser data respectively. There are two factors which lend to nonmonotonicity: (1) Luyapunov exponents measure asymptotic growth rate averages; a positive exponent does not contradict short-time negative exponents, (2) prediction confidence also depends on data density which varies significantly

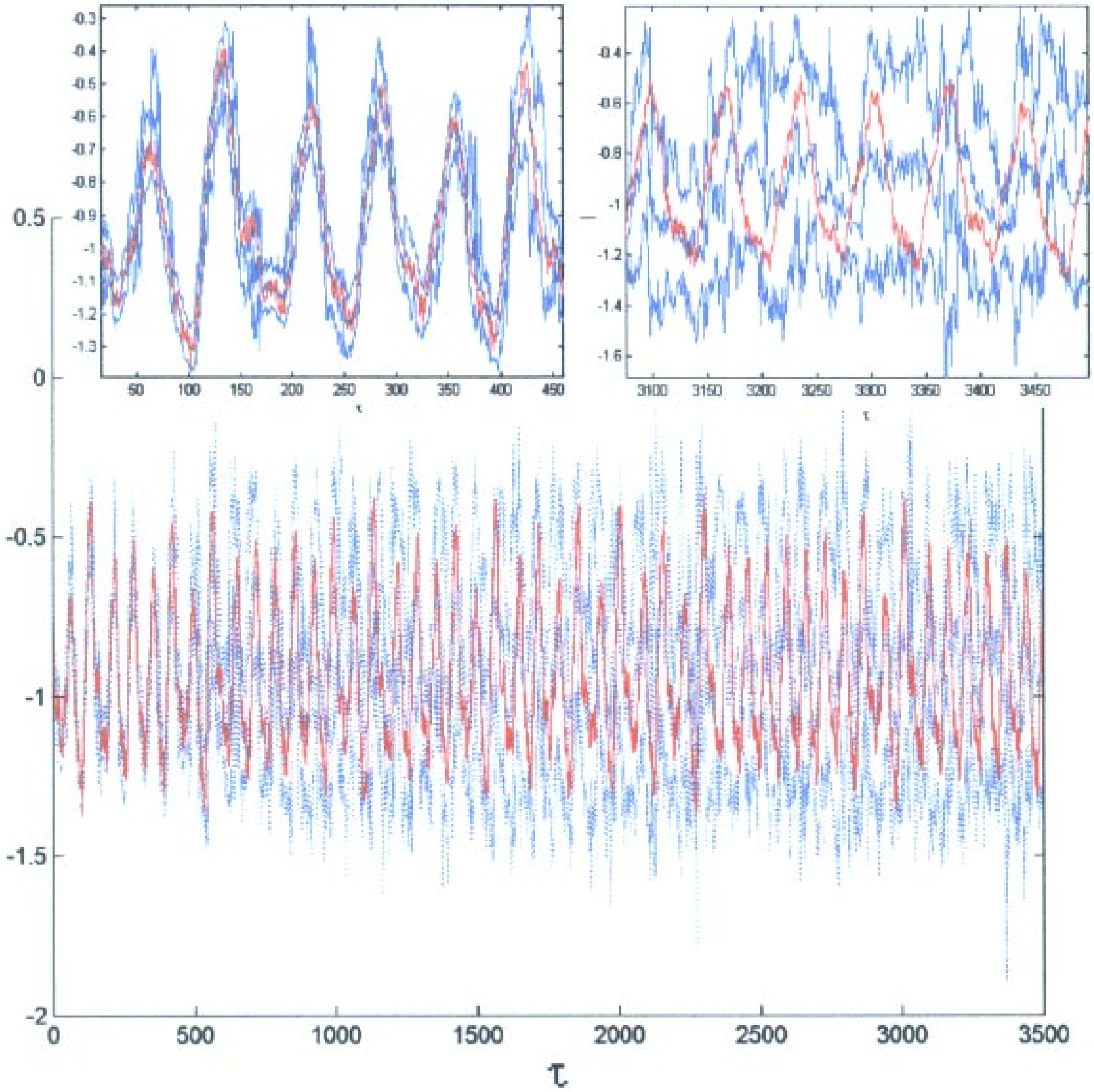


Fig. 12. Long-term prediction of noisy Mackey–Glass data. Prediction (center blue), 95% confidence bands (outer blue), and true response (red) versus time τ. Note that confidence bands get wider with time, as the effect of noise grows.

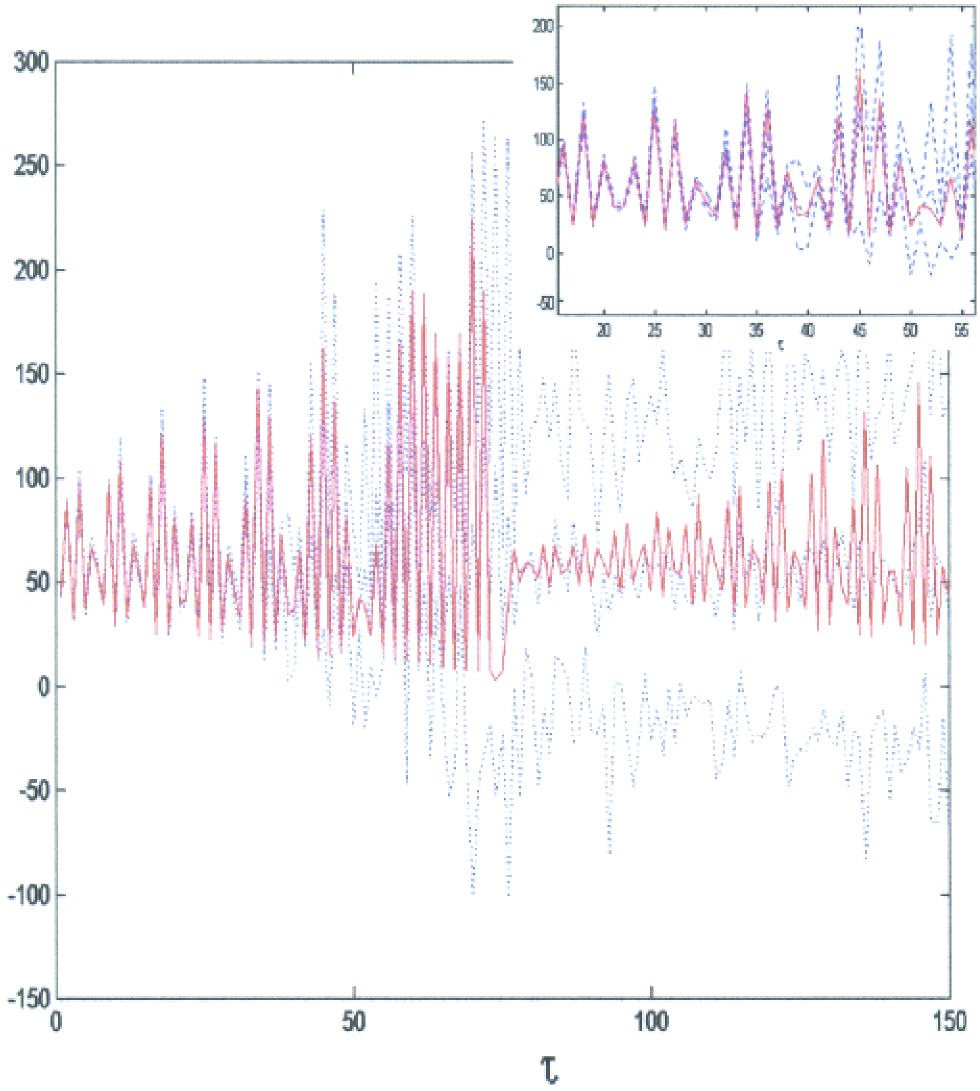
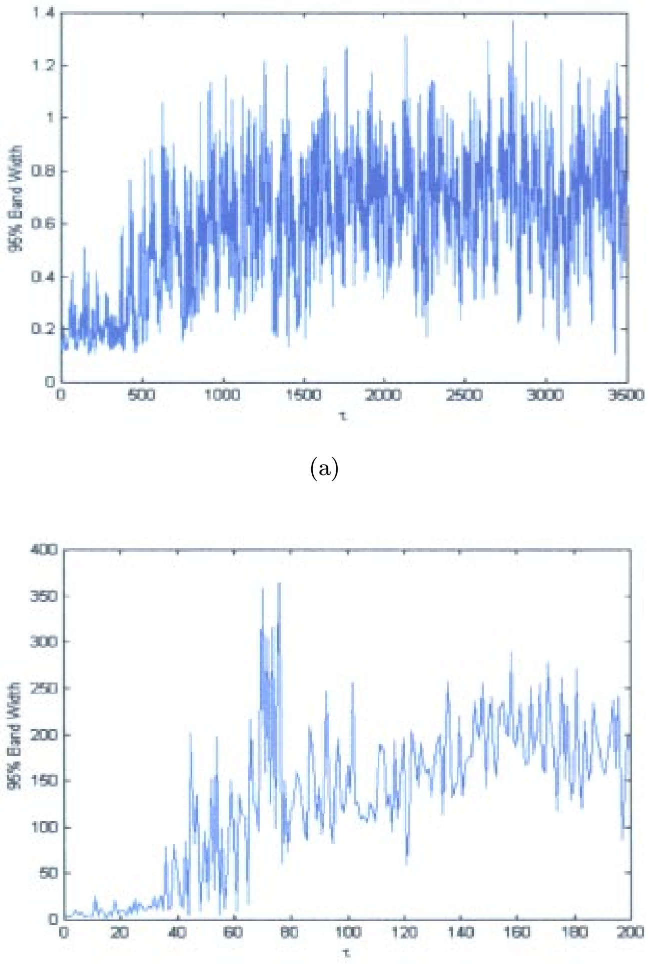


Fig. 13. Long-term prediction of infrared NH3 laser data. Prediction (center blue), 95% confidence bands (outer blue), and true response (red) versus time τ. Note that confidence bands get wider with time, as the effect of noise grows. Laser intensity is never negative, which further restricts these confidence bands.

over the attractor, according to a typically nonuniform invariant measure. Note that confidence bands can often be further restricted on physical grounds. For example, the laser intensity data can never be negative.

As with any statistical analysis, we have made implicit assumptions when making use of the F-statistc in Eq. (15) and t-distribution in Eq. (17). There is the usual assumption that ε is identically independently distributed (i.i.d.) [Neter *et al.*, 1996; Draper & Smith, 1981]. More likely, there is a complicated relationship between external noise and modeling "noise" of an unknown invariant measure, with resulting unknown multiple-step conditional probabilities on the noise-term. While, the assumptions are at best approximations even in the sense of averaging over the attractor, we observe excellent results with our methods. Even in more benign examples, such as linear regression of height versus weight data in human population samples, i.i.d. is the common assumption in ANOVA, even though short people tend to have smaller weight standard deviations than do tall people; validation of the assumptions of the *model* is in terms of good results.

In conclusion, we wish to emphasize that any nonlinear forecasting of noisy data is incomplete without an ANOVA. As far as we know, this is



(a)

(b)

Fig. 14. Long-term prediction error of (a) noisy Mackey–Glass data, (b) infrared NH3 laser data. Average widths of the 95% confidence bands, shown in Figs. 12 and 13 respectively, grow exponentially until saturation.

the only presentation that gives ANOVA, with validation of the predicted confidences, for predicting chaotic time-series. A prediction without analysis describing its quality is, at best, no more than a shot in the dark at a (hopefully) unbiased estimator. In contrast, predictions together with a statement like "the 95% confidence bands are very narrow," show that the prediction is very likely to be of high quality. On the other hand, a prediction together with a qualifying statement that the 95% confidence band is very wide, shows that the prediction is of (known) questionable quality and variability. If the data does not support an accurate prediction, that is important to know. Often, regression results are reported without discussion as to how well the predicted response will describe the true response.

We have given a systematic analysis to choose appropriate models, together with the appropriate scale for the model, and then to report the result with expected confidence. Finally, we have shown that a widely used value to select k-near neighbors as twice the number of parameters being fitted leads to typically either biased, or stochastically variable results, and the best k must scale with both τ, data density, and data set size. Hence, implicit with these techniques is an assumption that there is enough data available, that pruning neighborhood sizes is an option. If the inherent noise of the data set, and curvature of the map, requires a very small neighborhood, but the data set is so short that typically there are no k-enough close neighbors, then our algorithm "bottoms-out"; the suggested k becomes close to the minimum number of parameters being fitted, and therefore, their estimation becomes questionable. This becomes particularly prevalent for high-dimensional attractors; we have had no problem predicting real one-dimensional chemical reaction data with only 78 data points, but 1000 data points of laser data is difficult in $d = 4$ dimensions. We consider this to be an unavoidable, but now known, local feature of a short data set.

Acknowledgments

We thank Gary Fowler, Peter Turner, Celso Grebogi and Tim Sauer for helpful discussions. We would especially like to thank the author's student Midshipman J. Wood who did an honor's project [Wood, 1999] closely related to this article. The author is supported by the NSF under grant No. DMS-9704639.

References

Abarbanel, H. [1996] *Analysis of Observed Chaotic Data* (Springer, NY).

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

Draper, N. R. & Smith, H. [1981] *Applied Regression Analysis*, 2nd edition (John Wiley, NY).

Eckmann, J.-P. & Ruelle, D. [1985] "Ergodic theory of chaos and strange attractors," *Rev. Mod. Phys.* **57**(3), Part 1, 617–656.

Farmer, J. [1982] "Chaotic attractors of an infinite-dimensional system," *Phys.* **D4**, 366–393.

Farmer, J. D. & Sidorowich, J. J. [1987] "Predicting chaotic time series," *Phys. Rev. Lett.* **59**(8), 845–848.

Farmer, J. & Sidorowich, J. [1988] "Exploiting chaos to predict the future and reduce noise," in *Evolution, Learning, and Cognition*, ed. Lee, Y. C. (World Scientific, Singapore), p. 277.

Golub, G. & Van Loan, C. [1989] *Matrix Computations*, 2nd edition (Johns Hopkins University Press, Baltimore).

Hediger, T., Passamante, A. & Farrell, M. E. [1990] "Characterizing attractors using local intrinsic dimensions calculated by singular-value decomposition and information-theoretic criteria," *Phys. Rev.* **A41**, 5325–5332.

Huebner, U., Abraham, N. B. & Weiss, C. O. [1989] "Dimensions and entropies of chaotic intensity pulsations in a single-mode far-infrared NH3 laser," *Phys. Rev.* **A40**, 6354–6365.

Kantz, H. & Schrieber, T. [1997] *Nonlinear Time Series Analysis* (Cambridge University Press, NY).

Kugiumtzis, D., Lingjaerde, O. C. & Christopherson, O. C. [1998] "Regularized local linear prediction of chaotic time-series," *Phys.* **D112**, 344–360.

Lichtenberg, A. J. & Lieberman, M. A. [1983] *Regular and Stochastic Motion* (Springer-Verlag, NY).

Mackey, M. C. & Glass, L. [1977] *Science* **197**, 287.

Neter, J., Kutner, M. H., Nachtsheim, C. J. & Wasserman, W. [1996] *Applied Linear Regression Models*, 3rd edition (IRWIN, Chicago).

Press, W., Teukolsky, S., Vetterling, W. & Flannery, B. [1992] *Numerical Recipes in Fortran 77*, *The Art of Scientific Computing*, 2nd edition (Cambride University Press, Melbourne, Australia).

Sauer, T. [1992] "A noise reduction method for signals from nonlinear systems," *Phys.* **D58**, 193–201.

Sauer, T., Yorke, J. A. & Casdagli, M. [1991] "Embedology," *J. Stat. Phys.* **65**(3/4), 579–616.

Smith, L. [1994] "Local optimal prediction: Exploiting strangeness and the variation of sensitivity to initial conditions," *Phil. Trans. R. Soc. Lond.* **A348**, 371–381.

Takens, F. [1980] in *Dynamical Systems and Turbulence*, eds. Rand, D. & Young, L.-S., Lecture Notes in Mathematics **898** (Springer, Berlin), p. 366.

Walker, D. M. [1998] "Local filtering of noisy nonlinear time series," *Phys. Lett.* **A249**, 209–217.

Weigenbend, A. S. & Gershenfeld, N. A. [1993] *Times Series Prediction*: *Forecasting the Future and Understanding the Past*, (Addison-Wesley Reading, MA).

Wood, J. [1999] "Error statistics of time-delay embedding prediction on chaotic time-series," US Naval Academy Trident Scholar Honors Thesis.