
Monograph:
Billings, S.A. and Lee, K.L. (2002) The Effects of Noise Reduction on the Prediction Accuracy of Time Series. Research Report. ACSE Research Report 817 . Department of Automatic Control and Systems Engineering

---

The Effects of Noise Reduction on the Prediction Accuracy of Time Series

S A Billings and K L Lee
Department of Automatic Control and Systems Engineering
University of Sheffield
Mappin Street, Sheffield S1 3JD
United Kingdom

Research Report No. 817

February 2002

University of Sheffield

The Effects of Noise Reduction on the Prediction Accuracy of Time Series

S.A. Billings, K.L. Lee
Department of Automatic Control and Systems Engineering
University of Sheffield
Sheffield S1 3JD, UK

Abstract: A new iterative smoothing method based on the extended Kalman filter is introduced to smooth noisy chaotic time series. Two examples are given to illustrate the smoothing method. The smoothing method is then employed as a noise reduction pre-processing step to reduce measurement noise prior to identification and prediction. Three different prediction methods are introduced and the prediction performance is compared using three nonlinear examples. Superior predictive performance is obtained by the prediction method that employs the pre-processing step on the data.

1. Introduction

One of the most successful and widely used noise reduction methods is Kalman filtering (Davis 1977, Anderson and Moore 1979, Grewal and Andrews 1993). Recently Walker and Mees (1997) introduced a smoothing method based on the Kalman filter to reduce noise in chaotic data sets. The authors assumed that either the true model or a noise-free data set representative of the chaotic time series was available. Under these assumptions the smoothing method produced excellent noise reduction results. However a clear limitation in any real noise reduction problem is that the underlying dynamics are not usually known a priori and the model has to be learned from the noisy data as an integral part of the noise reduction process. Therefore in the present study an iterative Kalman smoothing method will be introduced which does not require the true model or a noise-free data set to be available. The iterative Kalman smoothing method will be tested on several chaotic examples.

Noise reduction methods are normally applied to reduce measurement noise on contaminated noisy outputs to enhance the future analysis of the data. For example

reducing the measurement noise on the output can facilitate the identification of dynamically valid models (Aguirre and Billings 1995, Aguirre et al. 1996) and the estimation of dynamical invariants of the time series (Kantz et al. 1993, Schreiber and Kantz 1996). Therefore it may be reasonable to assume that reducing measurement noise prior to computing predictions may have an advantageous effect on the prediction accuracy. Discussions based on a simple nonlinear autoregressive model will be used to show that the above belief may be true. Three different prediction methods are then introduced and compared. The first method does not employ any noise-reduction pre-processing step on the data. The second method employs a noise reduction pre-processing step on the training data set and the last method requires a noise reduction pre-processing step to be performed on the training and the testing data sets prior to identification and prediction. The multi-step ahead prediction performance for the three prediction methods is compared on three nonlinear examples.

The paper is organized as follows. The iterative Kalman smoothing method is introduced in Section 2, and simulation studies using this method to smooth two chaotic time series are presented in Section 3. The effects of reducing measurement noise on the prediction accuracy of the time series are discussed and three different prediction methods are introduced in Section 4. Multi-step ahead predictions for the three prediction methods on three nonlinear examples are performed and the results are presented in Section 5. Conclusions are given in Section 6.

2. The Iterative Kalman Smoothing Method

Consider the following model for the system equations

$$\mathbf{x}_{t+1} = g_t(\mathbf{x}_t) + \mathbf{w}_t \tag{1}$$

$$y_t = h_t(\mathbf{x}_t) + e_t \tag{2}$$

where $y_t \in \Re$ is the output, $\mathbf{x}_t \in \Re^m$ is the state of the system, m is the dimension of the state, $g_t(\bullet)$ and $h_t(\bullet)$ are nonlinear functions, $\mathbf{w}_t \in \Re^m$ and $e_t \in \Re$ are assumed to be independent and identically distributed noise processes with zero mean and covariances $Q \in \Re^{m\times m}$ and $\sigma_e^2 \in \Re$ respectively. Denote

$$G_t = \left.\frac{\partial g_t(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\hat{\mathbf{x}}_{t/t}} \quad \text{and} \quad H_t = \left.\frac{\partial h_t(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\hat{\mathbf{x}}_{t/t-1}} \tag{3}$$

so that the ij'th component of G_t represents the partial derivative of the i'th component of $g_t(\bullet)$ with respect to $\mathbf{x}_j$, and similarly for H_t, each derivative being evaluated at the point indicated.

If the nonlinear functions $g_t(\bullet)$ and $h_t(\bullet)$ are sufficiently smooth, these can be expanded in a Taylor series about the conditional expectations $\hat{\mathbf{x}}_{t/t}$ and $\hat{\mathbf{x}}_{t/t-1}$ as

$$g_t(\mathbf{x}_t) = g_t(\hat{\mathbf{x}}_{t/t}) + G_t(\mathbf{x}_t - \hat{\mathbf{x}}_{t/t}) + \ldots \tag{4}$$

$$h_t(\mathbf{x}_t) = h_t(\hat{\mathbf{x}}_{t/t-1}) + H_t(\mathbf{x}_t - \hat{\mathbf{x}}_{t/t-1}) + \ldots \tag{5}$$

Neglecting higher order terms and assuming knowledge of $\hat{\mathbf{x}}_{t/t}$ and $\hat{\mathbf{x}}_{t/t-1}$, the signal model can be approximated as

$$\mathbf{x}_{t+1} = G_t\mathbf{x}_t + \mathbf{w}_t + g_t(\hat{\mathbf{x}}_{t/t}) - G_t\hat{\mathbf{x}}_{t/t} \tag{6}$$

$$y_t = H_t\mathbf{x}_t + e_t + h_t(\hat{\mathbf{x}}_{t/t-1}) - H_t\hat{\mathbf{x}}_{t/t-1} \tag{7}$$

The following time update and measurement equations for the extended Kalman filter can be derived, see for example Davis (1977), Anderson and Moore (1979), Grewal and Andrews (1993), where Σ_t denotes the covariance of $\mathbf{x}_t$.

The time update equations are

$$\hat{\mathbf{x}}_{t+1/t} = g_t(\hat{\mathbf{x}}_{t/t}) \tag{8}$$

$$\Sigma_{t+1/t} = G_t\Sigma_{t/t}G_t^T + Q \tag{9}$$

where the superscript T denotes transpose. The measurement equations are

$$\hat{\mathbf{x}}_{t+1/t+1} = \hat{\mathbf{x}}_{t+1/t} + \Sigma_{t+1/t}H_{t+1}^T\left(H_{t+1}\Sigma_{t+1/t}H_{t+1}^T + R\right)^{-1}\left[y_{t+1} - h_{t+1}(\hat{\mathbf{x}}_{t+1/t})\right] \tag{10}$$

$$\Sigma_{t+1/t+1} = \Sigma_{t+1/t} - \Sigma_{t+1/t}H_{t+1}^T\left(H_{t+1}\Sigma_{t+1/t}H_{t+1}^T + R\right)^{-1}H_{t+1}\Sigma_{t+1/t} \tag{11}$$

There are many types of Kalman smoother (Anderson and Moore 1979) and the fixed lag smoother of lag one, whose form is particularly suitable for the later prediction part of this study, was chosen. The smoothing equations of the fixed lag smoother can be derived as

$$M_{t/t+1} = \Sigma_{t/t}G_t^T H_{t+1}^T\left(H_{t+1}\Sigma_{t+1/t}H_{t+1}^T + R\right)^{-1} \tag{12}$$

$$\hat{\mathbf{x}}_{t/t+1} = \hat{\mathbf{x}}_{t/t} + M_{t/t+1}\left[y_{t+1} - h_{t+1}(\hat{\mathbf{x}}_{t+1/t})\right] \tag{13}$$

$$\Sigma_{t/t+1} = \Sigma_{t/t} + \Sigma_{t/t}G_t^T\Sigma_{t+1/t}^{-1}\left(\Sigma_{t+1/t+1} - \Sigma_{t+1/t}\right)\left(\Sigma_{t/t}G_t^T\Sigma_{t+1/t}^{-1}\right)^T \tag{14}$$

The smoothing procedure using the extended Kalman filter is performed as follows. Initially the filter is run forwards in time by applying the time update and measurement equations to filter the data from $t=1$ to $t=N$, where N is the total length of the data. Then the filter is run backwards in time by using the smoothing equations to smooth the data from $t=N$ to $t=1$. Therefore each noise reduction iteration consists of one forward and one backward pass of the extended Kalman filter.

The true model of the system $(g_t(\bullet)$ and $h_t(\bullet))$, are usually unknown and have to be identified. In this study, polynomial functions are chosen to approximate $g_t(\bullet)$ and $h_t(\bullet)$ and the Nonlinear AutoRegressive (NAR) model, which is a special case of the Nonlinear Autoregressive Moving Average (NARMA) model (Leontaritis and Billings 1985a, 1985b, Chen and Billings 1989) of the form

$$y_t = f\left(y_{t-1},\ldots, y_{t-n_y}\right)+\xi_t \tag{15}$$

where n_y is the maximum output lag, ξ_t are the residuals and $f(\bullet)$ is some nonlinear function, will be used. The NAR model allows the use of faster and easier identification algorithms compared to the NARMA model which requires the use of an iterative estimation approach. The forward regression orthogonal least squares algorithm (Chen et al. 1989) will be employed to identify the NAR model. Assume that a piece of data $z_t, t=1\ldots N$ is available and that the following model has been identified using the forward regression orthogonal least squares algorithm

$$z_t = 0.5z_{t-1}z_{t-2}+0.3z_{t-1}+\xi_t \tag{16}$$

Then the functions $g_t(\bullet)$ and $h_t(\bullet)$ can be approximated using

$$g_t(\mathbf{x}_{t-1})=\begin{bmatrix}0.5z_{t-1}z_{t-2}+0.3z_{t-1}\\ z_{t-1}\end{bmatrix},\ h_t(\mathbf{x}_t)=[z_t] \text{ and } \mathbf{x}_t=\begin{bmatrix}z_t\\ z_{t-1}\end{bmatrix}$$

$$\text{with } G_{t-1}=\begin{bmatrix}0.5z_{t-2}+0.3 & 0.5z_{t-1}\\ 1 & 0\end{bmatrix} \text{ and } H_t=[1\quad 0]^T$$

Having estimated the unknown functions $g_t(\bullet)$ and $h_t(\bullet)$, two more factors need to be considered before the Kalman time update, measurement and smoothing equations can be applied. These are the values of the state and measurement noise covariances Q and σ_e^2 respectively. The noise covariances Q and σ_e^2 are assumed to be known or can be approximated using the method suggested in Heald and Stark (2000). In the present

study the noisy data are assumed to be corrupted by only measurement noise and hence the value of Q should ideally be set to zero. However, the identified model is only an approximation to the true model and therefore it is advisable not to put Q equal to zero so as to compensate for the inaccuracy in the identified model. The value of Q should be set higher at the initial noise reduction iterations and should decrease at subsequent iterations to match the reduction in the noise level. In this study, the values of Q for each noise reduction iteration were simply set as

$$Q^{(i)} = \beta^{i+1}\sigma_e^2 I \tag{17}$$

where i is the noise iteration, $0 < \beta < 1$ and $I \in \Re^{m \times m}$ is the identity matrix.

If the output data is very noisy, a high value of β should be selected. A rough guide to an appropriate choice of β can be obtained by comparing the values of the estimated noise covariance σ_e^2 and the minimum mean squared prediction errors estimated from the testing data set. If the minimum mean squared prediction errors are much larger than σ_e^2 (a few times higher than σ_e^2), this implies that the data are very noisy and the identified model is very likely to be a poor representation of the true model, therefore β should be set to a high value such as 0.5. Alternatively, if both the values of the minimum mean squared prediction errors and σ_e^2 are close (the minimum mean squared prediction errors are less than twice the value of σ_e^2) then a small value of β can be used, such as 0.15.

The new iterative Kalman based smoothing procedure can be summarised as follows. The data set $y_t^{(i)}$ is defined as the smoothed data set obtained at iteration i, similarly $Q^{(i)}$ and $\left(\sigma_e^2\right)^{(i)}$ are the estimated covariances of the state and measurement noise at iteration i respectively. Initially set $i = 0$ and hence $y_t^{(0)} = y_t$, $\left(\sigma_e^2\right)^{(0)} = \sigma_e^2$ and $Q^{(0)} = Q$.

i) Divide the noisy output equally into training and testing data sets.

ii) Set the initial values of n_y, l (degree of polynomial function) and n_m (number of terms in the model) to be say $n_y = 1$, $l = 3$ and $n_m = 4$ (see (iii) below).

iii) Use the forward regression orthogonal least squares algorithm (Chen et al. 1989) to identify the model from the training data set by selecting the best n_m important

terms from the candidate term set in the polynomial expansion and estimate the associated parameters.

iv) Calculate the mean squared prediction errors over the testing data set using the identified model in step (iii).

v) Increase n_m by one and repeat steps (iii) to (v). Go to step (vi) when $n_m = 50$.

vi) Increase n_y by one, set $n_m = 4$ and repeat steps (iii) to (vi). Go to step (vii) when $n_y = 15$.

vii) Record the values of l, n_y and n_m of the model with the minimum mean squared prediction errors calculated over the testing data set.

viii) Re-identify a global model from all the data $y_t^{(i)}$ (training +testing data sets) using the values of l, n_y and n_m obtained in step (vii).

ix) Apply the extended Kalman time update, measurement and smoothing equations (eqns (8)-(14)) using $g_t(\bullet)$, $h_t(\bullet)$, $\left(\sigma_e^2\right)^{(i)}$ and $Q^{(i)}$, to smooth the output to get the first smoothed data set $y_t^{(i+1)}$.

x) Increase i by one and calculate $\left(\sigma_e^2\right)^{(i)} = \sigma_e^2 - \frac{1}{N}\sum_{t=1}^{N}\left(\xi_t^{(T)}\right)^2$ where $\xi_t^{(T)} = y_t^{(0)} - y_t^{(i)}$ and $Q^{(i)} = \beta^{i+1}\sigma_e^2 I$.

xi) Repeat steps (viii) to (xi) until $\left(\sigma_e^2\right)^{(i)}$ is small.

In step (v) the maximum allowable terms in the model has been set to be 50 because simulations suggest that the smallest mean squared prediction errors calculated over the test data usually occur at values of n_m much smaller than 50. Similarly the maximum allowable n_y which has been set to 15 and $l = 3$ have been chosen based on practical application of the algorithm but other choices can be used if these prove unsatisfactory. In step (xi), the iterative Kalman smoothing procedure is terminated when $\left(\sigma_e^2\right)^{(i)}$ is very small compared with σ_e^2. Sometimes $\left(\sigma_e^2\right)^{(i)}$ may incorrectly be negative if $\frac{1}{N}\sum_{t=1}^{N}\left(\xi_t^{(T)}\right)^2 > \sigma_e^2$ in step (x), this means that the amount of noise removed from the noisy output is greater than the estimated variance σ_e^2 and hence the procedure has to be terminated.

3. Simulation Studies Using the Iterative Kalman Smoothing Method

The iterative Kalman smoothing method described above will be illustrated using two well-known examples, the Lorenz time series and the Henon map.

3.1 The Lorenz Time Series

The equations of the Lorenz chaotic time series are

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= rx - y - xz \\ \dot{z} &= -bz + xy \end{aligned} \tag{18}$$

with $\sigma = 10, b = 8/3$ and $r = 28$. A fourth order Runge-Kutta method was employed to integrate the equations with a fixed step size of 0.001s. One thousand data points of the x-coordinate were collected with a sampling rate of 0.05s and this served as the noise-free data x_t. A zero mean uncorrelated noise sequence e_t with $\sigma_e^2 = 0.65$ was added to the noise-free data to obtain the noisy output, $y_t = x_t + e_t$. The resulting output data set had a signal-to-noise ratio of 40dB. The first 500 points of the output were then used as the training data set and the next 500 points were used as the testing data set. The iterative Kalman smoothing procedure described in Section 2 was then implemented to smooth the noisy output.

The following model structure gave the minimum mean squared prediction errors over the testing data set, $l = 3, n_y = 11$ and $n_m = 13$. The minimum mean squared prediction errors calculated over the testing data set was 1.20, which was less than twice the value of σ_e^2, and therefore the estimated model was assumed to be a good approximation of the true model. Hence a small value of $\beta = 0.15$ was employed. The extended Kalman time update, measurement and smoothing equations were then applied to smooth the output and the procedure was terminated at the seventh iteration because the value of $\left(\sigma_e^2\right)^{(7)}$ was negative. The smoothed data set $y_t^{(6)}$ was taken as the final smoothed data set. The phase portrait plots of the noise-free data set, noisy output and the final smoothed data set are illustrated in Figure 1. A measure of the effectiveness of the smoothing algorithm can be obtained by computing

$$D = \sqrt{\frac{\sum_{t=1}^{N}\left(y_t^{(0)} - x_t\right)^2}{\sum_{t=1}^{N}\left(y_t^{(n_p)} - x_t\right)^2}} \tag{19}$$

where $y_t^{(n_p)}$ is the final smoothed data. A higher value of the index D means that more noise has been removed from the noisy output. Clearly this index D is of theoretical interest only since, in practice, the noise-free data x_t will not be available. The value of D obtained for the Lorenz example was 3.34. The results above show that the iterative Kalman smoothing method has successfully reduced the noise contamination in the Lorenz time series.

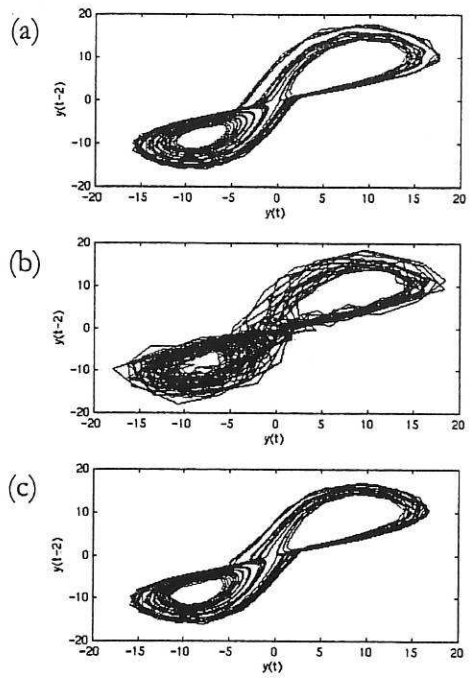

Figure 1 Phase portrait plots of the Lorenz time series with a sampling rate of 0.05s for (a) the noise-free data set, (b) the noisy output and (c) the final smoothed data set.

3.2 The Henon Chaotic Time Series

In this example the well-known Henon map will be studied. The Henon map is described by the equation

$$x_t = 1 - 1.4x_{t-1}^2 + 0.3x_{t-2} \tag{20}$$

One thousand data points were generated and these points were assumed to be the noise-free data set. A zero mean uncorrelated noise sequence with $\sigma_e^2 = 0.0055$ was added to the noise-free data set to obtain the noisy output. The resulting output had a

signal-to-noise ratio of 40dB. The first 500 points were used as the training data set and the next 500 points were used as the testing data set. The minimum mean squared prediction errors calculated over the testing data set was 0.030. This was much higher than the value of σ_e^2 and hence β was set as 0.5. The iterative Kalman smoothing procedure described in Section 2 was then applied. The iterative smoothing procedure was terminated at the 13th iteration because the value of $\left(\sigma_e^2\right)^{(13)}$ was negative. Therefore the smoothed data at the 12th iteration was taken as the final smoothed data. The phase portrait plots of the noise-free data, the output and the final smoothed data are shown in Figure 2. The value of D obtained for the Henon case was 3.71.

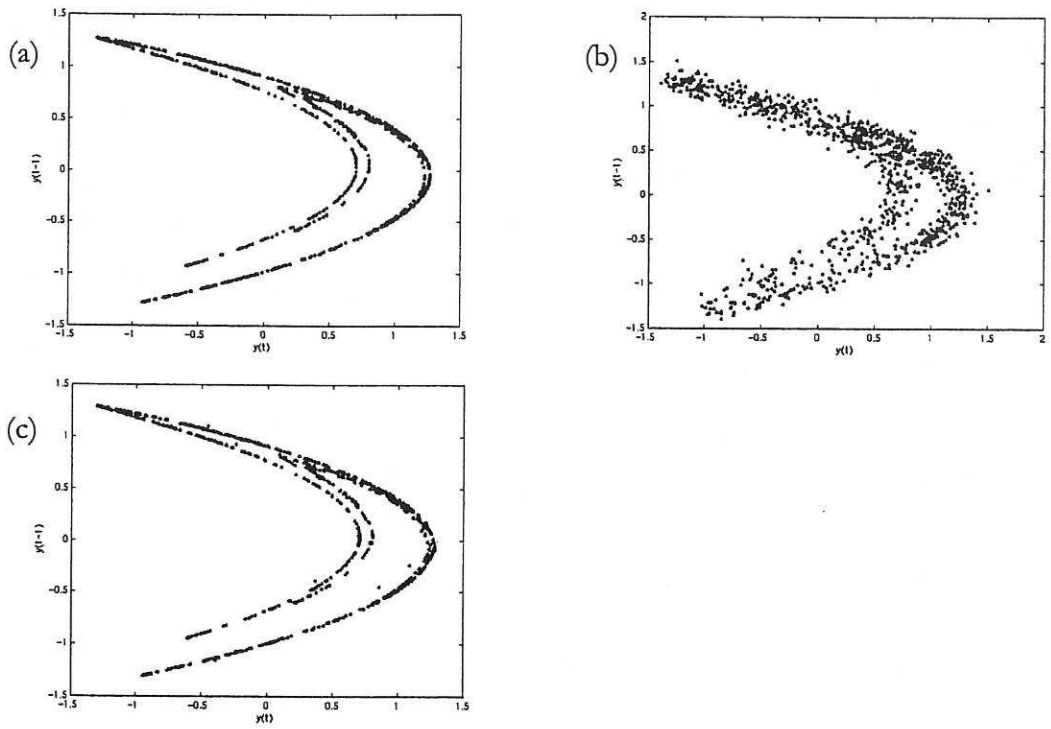

Figure 2 Phase portrait plots of the Henon map for (a) the noise-free data set, (b) the noisy output and (c) the final smoothed data set.

3.3 Discussion on the Iterative Kalman Smoothing Method

Walker and Mees (1997) noted that if the model has to be identified from the noisy output, the repeated application of the extended Kalman filter tends to produce a smoothed data set which displays the dynamics of the identified model rather than the underlying dynamics of the time series. However the present paper shows that the extended Kalman filter can still be used successfully to smooth noisy outputs even though the true model or a noise-free data set of the time series is not available. An

important point is not to set $Q^{(i)}$ equal to zero, this then compensates for the inaccuracies in the identified model even though the output is corrupted by measurement noise only. The iterative Kalman smoothing method has also been tested on other well-known chaotic time series including Chua's circuit (Chua et al. 1986), the Mackey-Glass (Mackey and Glass 1977) and the Duffing-Holmes (Holmes 1979) chaotic time series and excellent noise reduction results were obtained. From the simulation studies a crucial factor that can affect the performance of the iterative Kalman smoothing method is to obtain an accurate estimate of the measurement noise variance σ_e^2 which is usually unknown for real time series. Therefore future studies should investigate how such a parameter will affect the results and should develop methods to estimate σ_e^2. The rest of the paper will concentrate on the second and main objective of this paper that is to investigate the effects of reducing measurement noise using the iterative Kalman smoothing method on the prediction accuracy of the time series.

4. Effects of Reducing Measurement Noise on Multi-step Ahead Predictions for Nonlinear Time Series

One of the most important objectives of time series analysis is to accurately compute forecasts or predictions of future values. In many application areas, such as economics, stock prediction, sales data forecasting, electric load forecasting, hydrology, astronomy etc, obtaining good predictions are usually not trivial because many factors can affect the predictions.

One factor is the measurement noise contamination of the data set. The measurement noise contaminated in the output has been shown to hinder the identification of dynamically valid input-output models (Aguirre and Billings 1995, Aguirre et al. 1996) and the estimation of dynamical invariants of the time series (Kantz et al. 1993, Schreiber and Kantz 1996). The measurement noise may also have an adverse effect on the prediction accuracy of the time series. If the underlying dynamics of the time series are deterministic, reducing the measurement noise before computing the predictions may improve the prediction accuracy as shown in Cao et al. (1998). However the objective in Cao et al. (1998) was to investigate the determinism in human posture control data using nonlinear prediction. The human posture control data set was smoothed using the simple noise reduction method of Schreiber (1992) and the smoothed data points were fed into

the identified model to compute predictions. This would not be possible in real time prediction because only past data points can be smoothed and present data points to be fed into the model to make predictions can only be filtered. In this study the real time prediction scenario will be studied. Multi-step ahead predictions will be considered. After a predictive model has been identified, the one-step ahead predictions can be obtained using

$$\hat{y}_{t+1/t} = \hat{g}\left(y_t, y_{t-1}, \ldots, y_{t-n_y+1}\right) \tag{21}$$

where $\hat{g}(\bullet)$ is the identified predictive model and $\left(y_t, y_{t-1}, \ldots, y_{t-n_y+1}\right)$ are the arguments of the model. The multi-step ahead predictions are normally computed in an iterative manner by feeding back previous predicted values. For example the two-step ahead predictions can be calculated as

$$\hat{y}_{t+2/t} = \hat{g}\left(\hat{y}_{t+1/t}, y_t, \ldots, y_{t-n_y+2}\right) \tag{22}$$

The effects of reducing measurement noise on multi-step ahead predictions will be studied next in Section 4.1 & 4.2 based on a simple nonlinear example.

4.1 One-step Ahead Prediction Case

In the following discussion, the dynamics of the system are assumed to be deterministic and the noise-free data set is corrupted by measurement noise to give a noisy output. Consider a simple one-dimensional example

$$x_t = g\left(x_{t-1}\right) \tag{23}$$

$$y_t = x_t + e_t \tag{24}$$

Assume that the nonlinear function $g(\bullet)$ is known or can be identified. The optimal one-step ahead prediction is

$$y_{t+1/t}^{opt} = g\left(x_t\right) \tag{25}$$

The one-step ahead prediction for the noisy output case $\hat{y}_{t+1/t}$ is computed by feeding the noisy data y_t into the nonlinear function $g(\bullet)$ to give

$$\hat{y}_{t+1/t} = E[g(y_t)] \tag{26}$$

Substituting y_t from eqn (24) into eqn (26) and applying a Taylor series expansion by retaining up to second-order terms gives

$$\hat{y}_{t+1/t} = E[g(x_t + e_t)] = E\left[g(x_t) + e_t g'(x_t) + \frac{1}{2} e_t^2 g''(x_t) \right]$$

$$\approx g(x_t)+\frac{1}{2}\sigma_e^2 g''(x_t) \tag{27}$$

where $g'(x_t)$ and $g''(x_t)$ are the first and second derivatives of g, and σ_e^2 is the variance of the noise e_t. Now assume that the noise reduction method has been applied and this has successfully reduced the noise so that the noise-reduced data s_t is

$$s_t = x_t + \eta_t \tag{28}$$

where η_t is assumed to be zero mean uncorrelated noise and $\sigma_\eta^2 << \sigma_e^2$. Following the same analysis as above yields the one-step ahead prediction for the noise-reduced case $\hat{s}_{t+1/t}$ as

$$\hat{s}_{t+1/t} = E[g(s_t)] \approx g(x_t)+\frac{1}{2}\sigma_\eta^2 g''(x_t) \tag{29}$$

Since $\sigma_e^2 >> \sigma_\eta^2$, therefore $\left|\frac{1}{2}\sigma_e^2 g''(x_t)\right| >> \left|\frac{1}{2}\sigma_\eta^2 g''(x_t)\right|$. It can be deduced from eqns (27) and (29) that $\hat{s}_{t+1/t}$ should be closer to the optimal prediction than $\hat{y}_{t+1/t}$.

4.2 Multi-step Ahead Predictions Case

Next consider the two-step ahead prediction task. The optimal two-step ahead prediction can be obtained as

$$y_{t+2/t}^{opt} = g(g(x_t)) \tag{30}$$

The two-step ahead prediction for the noisy case $\hat{y}_{t+2/t}$ can be computed using the iterative approach by feeding the one-step ahead prediction $\hat{y}_{t+1/t}$ into the nonlinear function $g(\bullet)$ to give

$$\begin{aligned}
\hat{y}_{t+2/t} &= E[g(\hat{y}_{t+1/t})] = E[g[g(y_t)]] = E[g[g(x_t+e_t)]] \\
&\approx E\left[g\left(g(x_t)+e_t g'(x_t)+\frac{1}{2}e_t^2 g''(x_t)\right)\right] \\
&\approx E\left[g(g(x_t))+\left(e_t g'(x_t)+\frac{1}{2}e_t^2 g''(x_t)\right) g'(g(x_t)) \right. \\
&\qquad \left. +\frac{1}{2}\left(e_t g'(x_t)+\frac{1}{2}e_t^2 g''(x_t)\right)^2 g''(g(x_t))\right] \\
&\approx g(g(x_t))+\frac{1}{2}\sigma_e^2\left[g''(x_t)g'(g(x_t))+(g'(x_t))^2 g''(g(x_t))\right]
\end{aligned} \tag{31}$$

where only second order terms in the Taylor series expansion are retained. For the noise-reduced case the two-step ahead prediction $\hat{s}_{t+2/t}$ is

$$\hat{s}_{t+2/t} \approx g(g(x_t)) + \frac{1}{2}\sigma_\eta^2 \left[g''(x_t) g'(g(x_t)) + (g'(x_t))^2 g''(g(x_t)) \right] \tag{32}$$

Knowing that $\sigma_e^2 >> \sigma_\eta^2$, $\hat{s}_{t+2/t}$ should again be a better prediction of $y_{t+2/t}^{opt}$ than $\hat{y}_{t+2/t}$. Additional multi-step ahead predictions may be tedious to perform but the results above suggest that the noise-reduced case should perform better than the noisy case in terms of prediction accuracy with respect to noise-free data.

4.3 Three Different Prediction Methods

Three different prediction methods will be discussed below and the prediction results will be compared in the simulation studies.

4.3.1 Method 1

This is the simplest and the most conventional method where no pre-processing of the data is performed. The procedure of Method 1 can be summarised as follows.

i) Divide the output training data set into estimation and validation data sets.

ii) Set the initial values of l and n_y to be 3 and 1 respectively.

iii) Set the initial important terms to be included in the model $n_m = 4$.

iv) Use the forward regression orthogonal least squares algorithm to identify the model from the estimation data set by selecting the best n_m important terms from the candidate terms of the polynomial expansion and estimate the associated parameters.

v) Calculate the mean squared prediction errors over the validation data set.

vi) Increase n_m by one and repeat steps (iv) to (vi). Go to step (vii) when $n_m = 50$

vii) Increase n_y by one and repeat steps (iii) to (vii). Go to step (viii) when $n_y = 15$

viii) Record the values of l, n_y and n_m with the minimum mean squared prediction errors calculated over the validation data set.

ix) Re-identify a model $\hat{g}^{(1)}(\bullet)$ from the training data set (estimation + validation data sets) using the values of l, n_y and n_m recorded in step (viii).

x) Compute one-step ahead prediction of Method 1 $\hat{y}_{t+1/t}^{(1)}$ using the identified model in step (ix) with the arguments $\left(y_t, y_{t-1}, ..., y_{t-n_y+1}\right)_{test}$ where $(y_t)_{test}$ are the testing data points.

xi) Update the arguments as $\left(\hat{y}_{t+1/t}^{(1)}, y_t, ..., y_{t-n_y+2}\right)_{test}$ and compute the two-step ahead predictions using the iterative approach.

xii) Repeat until k-step ahead predictions have been obtained. Repeat steps (x) to (xii) for all the points in the testing data set.

4.3.2 Method 2

The noisy training data set is smoothed using the iterative Kalman smoothing method discussed in Section 2. Then a predictive model is identified from the noisy training data using the smoothed training data points $(s_t)_{train}$ as targets. The predictive model is identified as

$$(s_t)_{train} = \hat{g}^{(2)}\left(y_{t-1}, ..., y_{t-n_y}\right)_{train} + \xi_t^{(2)} \tag{33}$$

where $\hat{g}^{(2)}(\bullet)$ is the identified model in Method 2 and $\xi_t^{(2)}$ are the residuals. The predictions over the testing data set are obtained by feeding the noisy testing data points into the identified model as

$$\hat{y}_{t+1/t}^{(2)} = \hat{g}^{(2)}\left(y_t, ..., y_{t-n_y+1}\right)_{test} \tag{34}$$

The procedure of Method 2 can be summarised as follows.

Steps (i) to (viii) are the same as Method 1.

ix) Perform the iterative Kalman smoothing procedure described in Section 2 on the training data set to obtain the smoothed training data set.

x) Re-identify a model $\hat{g}^{(2)}(\bullet)$ using the values of l, n_y and n_m recorded in step (viii) with the smoothed training data points as targets.

xi) Compute the one-step ahead prediction of Method 2 $\hat{y}_{t+1/t}^{(2)}$ using the identified model in step (x) with the arguments $\left(y_t, y_{t-1}, ..., y_{t-n_y+1}\right)_{test}$.

xii) Update the arguments as $\left(\hat{y}_{t+1/t}^{(2)}, y_t, ..., y_{t-n_y+2}\right)_{test}$ and compute the two-step ahead predictions using the iterative approach.

xiii) Repeat until k-step ahead predictions have been obtained. Repeat steps (xi) to (xiii) for all points in the testing data set.

4.3.3 Method 3

In Method 3 the noisy training data set is smoothed using the iterative Kalman smoothing method. This is followed by the identification of the predictive model from the smoothed training data set as

$$(s_t)_{train} = \hat{g}^{(3)}\left(s_{t-1},\ldots,s_{t-n_y}\right)_{train} + \xi_t^{(3)} \tag{35}$$

where $\hat{g}^{(3)}(\bullet)$ is the identified model in Method 3 and $\xi_t^{(3)}$ are the residuals. The predictions over the testing data set are obtained using

$$\hat{s}_{t+1/t}^{(3)} = \hat{g}^{(3)}\left(s_t,\ldots,s_{t-n_y+1}\right)_{test} \tag{36}$$

To illustrate how this can be implemented consider the following example. Assume that the training data of size 500 has been smoothed to obtain $s_1, s_2,\ldots, s_{500}$. The one-step ahead prediction is

$$\hat{s}_{501/500}^{(3)} = \hat{g}^{(3)}\left(s_{500}, s_{499},\ldots, s_{501-n_y}\right) \tag{37}$$

To compute the next one-step ahead prediction $\hat{s}_{502/501}^{(3)}$, the data points $\left(y_{501}, y_{500},\ldots, y_{502-n_y}\right)$ are filtered using the time update and measurement equations of (8), (9), (10) and (11) to obtain $\left(s_{501}, s_{500},\ldots, s_{502-n_y}\right)$. Then the backward pass smoothing equations (12), (13) and (14) have to be applied to further smooth the data points $\left(s_{500},\ldots, s_{502-n_y}\right)$. The prediction is then obtained by computing

$$\hat{s}_{502/501}^{(3)} = \hat{g}^{(3)}\left(s_{501},\ldots, s_{502-n_y}\right) \tag{38}$$

Therefore to make a prediction, the latest data point (y_{501}) can only be filtered but the earlier data points $\left(y_{500},\ldots, y_{502-n_y}\right)$ can be smoothed. For the current prediction task, the data points to be fed into the model to make prediction are filtered and smoothed once only. The procedure of Method 3 can be summarised as follows.

Steps (i) to (ix) are the same as Method 2.

x) Re-identify a predictive model $\hat{g}^{(3)}(\bullet)$ from the smoothed training data set using the values of l, n_y and n_m recorded in step (viii).

xi) Use the time update and measurement eqns (8)-(11) to filter the testing data points to obtain the filtered testing data points $\left(s_t, s_{t-1},\ldots, s_{t-n_y+1}\right)_{test}$. Then apply the backward pass smoothing eqns (12)-(14) to further smooth the data points $\left(s_{t-1}, s_{t-2},\ldots, s_{t-n_y+1}\right)_{test}$

xii) Compute the one-step ahead prediction of Method 3 $\hat{s}_{t+1/t}^{(3)}$ using the identified model in step (x) with the argument $\left(s_t, s_{t-1},\ldots, s_{t-n_y+1}\right)_{test}$.

xiii) Update the argument as $\left(\hat{s}^{(3)}_{t+1/t}, s_t, \ldots, s_{t-n_y+2}\right)_{test}$ and compute the two-step ahead predictions using the iterative approach.

xiv) Repeat until k-step ahead predictions have been obtained. Repeat steps (xi) to (xiv) for all points in the testing data set.

5. Simulation Studies on Prediction Accuracy

The procedures outlined in the preceding section for the three prediction methods will be illustrated by means of three nonlinear examples. The first two examples consider the Lorenz and the Henon chaotic time series. The 1000 noise-free data points of both time series generated in Section 3 were used. The third example considers a non-chaotic time series obtained by integrating the Duffing-Holmes equation

$$\ddot{x} + \gamma\dot{x} - \alpha\left(x - x^3\right) = F\cos(wT)) \tag{39}$$

with $\gamma = 0.168, \alpha = 0.5, F = 0.05$ and $w = 1$. The equation was integrated using a fourth order Runge-Kutta method and the noise-free data set of 1000 data points was collected with a sampling rate of 0.1s.

In all cases, the noise-free data sets were corrupted additively by zero mean uncorrelated noise to give noisy output data sets with a 40dB signal-to-noise ratio. The first 500 output data points were used as the training data set with the first 400 points used as the estimation data set and the next 100 points used as the validation data set. The last 500 points of the output were used as the testing data set. The three prediction methods described in Section 4.3.1, 4.3.2 & 4.3.3 were used to perform the multi-step ahead predictions on the three nonlinear examples. A measure of the prediction accuracy achieved by the different methods can be obtained by computing the mean squared prediction errors with respect to the noise-free data $(mse)_{noise-free}$ as

$$(mse)_{noise-free} = \frac{1}{N}\sum_{t=1}^{N}\gamma_t^2 \tag{40}$$

where $\gamma_t = \hat{y}_t - x_t$, $\hat{y}_t$ is the obtained prediction by the different prediction methods. The results for the three examples are presented in Tables 1,2 & 3.

Of all the three different prediction methods, Method 3 that performs the noise reduction pre-processing step on the training and the testing data was the best. The improvement in prediction accuracy was very significant and in the range of 30%-60%

compared to Method 1 where no pre-processing of the data was performed. These results agree well with the discussions in Section 4.1 & 4.2. For the Henon map, the k-step ahead predictions for Method 1 and 2 were unstable and explosive when $k \geq 3$. This however did not occur in Method 3. For the Henon example, therefore reducing the measurement noise prior to identification and prediction may help to reduce the possibility of the predictions exploding. Method 2 performs better than Method 1 in one-step ahead predictions but mixed results were obtained for multi-step ahead predictions. In the results presented in Tables 1,2 & 3, the value of the measurement noise variance σ_e^2, which was required in the noise reduction pre-processing step, was assumed to be known or to have been estimated accurately. This may not be possible for real time series. Therefore the predictions using Method 3 were repeated on all three examples but assuming that σ_e^2 was either overestimated or underestimated by 30%. The result for the Lorenz example is illustrated in Figure 3, which shows that Method 3 can still perform better than Method 1 even if the variance σ_e^2 is over/under estimated by 30%. Similar findings were obtained for the Henon and the Duffing-Holmes examples. Therefore the results presented suggest that the use of noise reduction methods to reduce the measurement noise on the output prior to identification and prediction can improve the prediction accuracy.

Table 1. The multi-step ahead mean squared prediction errors with respect to noise-free data $(mse)_{noise-free}$ for the three methods on the Lorenz time series. The values in brackets are the percentage improvement in prediction accuracy compared to Method 1.

Prediction Step	$(mse)_{noise-free}$ *for Method 1*	$(mse)_{noise-free}$ *for Method 2*	$(mse)_{noise-free}$ *for Method 3*
1	0.5384	0.5120 (4.9%)	0.2995 (44.4%)
2	0.8140	0.7899 (3.0%)	0.4509 (44.6%)
3	1.0513	1.1496 (-9.4%)	0.6466 (38.5%)
4	1.4453	1.4509 (-0.4%)	0.8899 (38.4%)
5	2.0441	1.9134 (6.4%)	1.2852 (37.1%)
6	2.9238	2.5696 (12.1%)	1.9092 (34.7%)
7	4.1863	3.3826 (19.2%)	2.7891 (33.4%)

8	6.0673	4.8746 (19.7%)	4.0834 (32.7%)
9	8.7947	7.4665 (15.1%)	6.0496 (31.2%)
10	13.2784	11.4968 (13.4%)	9.2237 (30.5%)
11	18.8704	16.0494 (14.9%)	12.7287 (32.5%)
12	24.3072	19.4540 (20.0%)	15.4579 (36.4%)
13	28.7963	21.8238 (24.2%)	17.4699 (39.3%)
14	33.1627	24.7884 (25.3%)	19.2220 (42.0%)
15	37.6871	27.8241 (26.2%)	19.2247 (49.0%)
16	41.0155	29.3264 (28.5%)	17.3575 (57.7%)
17	42.3975	30.4848 (28.1%)	15.9795 (62.3%)
18	42.4926	33.7762 (20.5%)	15.7786 (62.9%)
19	42.0857	37.9083 (9.9%)	16.5657 (60.6%)
20	42.5461	40.5634 (4.7%)	17.9891 (57.7%)

Table 2. The multi-step ahead mean squared prediction errors with respect to noise-free data $(mse)_{noise-free}$ for the three methods on the Henon map. The values in brackets are the percentage improvement in prediction accuracy compared to Method 1. The values of $(mse)_{noise-free}$ were explosive for Methods 1 & 2 on prediction steps greater than 2, hence the percentage improvements were not computed and were indicated by (-----).

Prediction Step	$(mse)_{noise-free}$ *for Method 1*	$(mse)_{noise-free}$ *for Method 2*	$(mse)_{noise-free}$ *for Method 3*
1	0.0313	0.0311 (0.6%)	0.0241 (23.0%)
2	0.1223	0.1251 (-2.3%)	0.0669 (45.3%)
3	1.4125	1.6056 (-----)	0.1165 (-----)
4	715.24	1032.0 (-----)	0.2077 (-----)
5	2×10^{13}	2×10^{15} (-----)	0.3879 (-----)

Table 3. The multi-step ahead mean squared prediction errors with respect to noise-free data $(mse)_{noise-free}$ for the three methods on the Duffing non-chaotic time series. The values in brackets are the percentage improvement in prediction accuracy compared to Method 1. All entries were to the power of 10^{-3}.

Prediction Step	$(mse)_{noise-free}$ *for Method 1*	$(mse)_{noise-free}$ *for Method 2*	$(mse)_{noise-free}$ *for Method 3*
1	0.0743	0.0609 (18.0%)	0.0417 (43.9%)
2	0.0814	0.0685 (15.8%)	0.0432 (46.9%)
3	0.0886	0.0767 (13.4%)	0.0443 (50.0%)
4	0.0964	0.0861 (10.7%)	0.0443 (54.0%)
5	0.0992	0.0940 (5.2%)	0.0458 (53.9%)
6	0.1086	0.1006 (7.4%)	0.0469 (56.8%)
7	0.1089	0.1059 (2.8%)	0.0470 (56.9%)
8	0.1142	0.1094 (4.2%)	0.0473 (58.6%)
9	0.1158	0.1129 (2.5%)	0.0481 (58.4%)
10	0.1202	0.1122 (6.7%)	0.0480 (60.1%)
11	0.1094	0.1130 (-3.3%)	0.0413 (62.2%)
12	0.1086	0.1108 (-2.0%)	0.0400 (63.2%)
13	0.1058	0.1069 (-1.0%)	0.0375 (64.5%)
14	0.1000	0.1008 (-0.8%)	0.0358 (64.2%)
15	0.0969	0.0940 (3.0%)	0.0338 (65.1%)
16	0.0804	0.0848 (-5.4%)	0.0316 (60.8%)
17	0.0669	0.0765 (-14.3%)	0.0278 (58.4%)
18	0.0637	0.0690 (-8.3%)	0.0268 (57.9%)
19	0.0559	0.0635 (-13.6%)	0.0254 (54.6%)
20	0.0514	0.0601 (-16.9%)	0.0240 (53.3%)

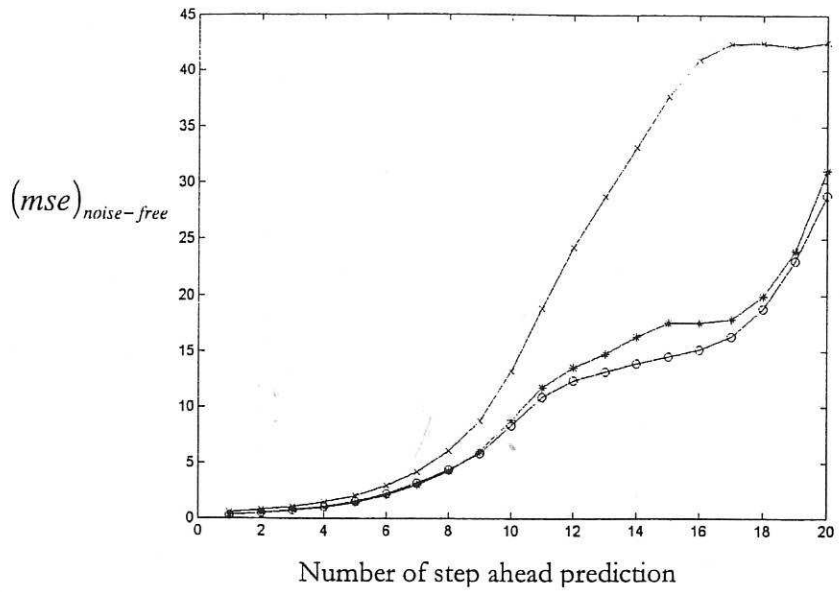

Figure 3. The multi-step ahead mean squared prediction errors with respect to the noise-free data $(mse)_{noise-free}$ on the Lorenz example for prediction Method 1 (×), Method 3 with σ_e^2 underestimated by 30% (o) and overestimated by 30% (*).

6. Conclusions

A new iterative Kalman smoothing based method based on the extended Kalman filter has been introduced. Excellent noise reduction results were obtained using the smoothing method on two well-known chaotic examples where the underlying dynamics or noise-free data sets were assumed to be unknown or unavailable. The effects of reducing the measurement noise on the prediction accuracy of time series were also investigated. Discussions based on a simple nonlinear example showed that higher prediction accuracy should be obtained by reducing the measurement noise prior to identification and prediction. The iterative Kalman smoothing method was employed as the noise reduction method in the pre-processing step. Three different prediction methods were then presented. The first prediction method was the simplest with no pre-processing of the data. The second prediction method employed the noise reduction pre-processing step on the training data set only and the last method used the noise reduction pre-processing step on both the training and the testing data sets. Multi-step ahead predictions for the three methods on three nonlinear examples were performed. Significant improvements were achieved by method 3 in predicting the noise-free data

points compared with method 1. The results suggest that employing a noise reduction pre-processing step on the noisy data prior to identification and prediction is beneficial.

Acknowledgements

SAB gratefully acknowledges that part of this work was supported by EPSRC.

References

Aguirre, L.A. and Billings, S.A. (1995). Identification of models for chaotic systems from noisy data: implication for performance and nonlinear filtering. Physica D, 85, 239-258.

Aguirre, L.A., Mendes, E.M. and Billings, S.A. (1996). Smoothing data with local instabilities for the identification of chaotic systems. International Journal of Control, 63, 3, 483-505.

Anderson, B.D.O. and Moore, J.B. (1979). Optimal filtering, Prentice-Hall, Englewood Cliffs, N.J.

Cao, L.Y., Kim, B.G., Kurths, J. and Kim, S. (1998). Detecting determinism in human posture control data. International Journal of Bifurcation and Chaos, 8, 1, 179-188.

Chen, S. and Billings, S.A. (1989). Modelling and analysis of nonlinear time series. International Journal of Control, 50, 6, 2151-2171.

Chen, S, Billings, S.A. and Luo, W. (1989). Orthogonal least squares methods and their application to nonlinear system identification. International Journal of Control, 50, 1873-1896.

Chua, L.O., Komuro, M. and Matsumoto, T. (1986). The double scroll family. IEEE Transactions on Circuits and Systems, 33, 1072-1118.

Davies, M.H.A. (1977). Linear estimation and stochastic control. Chapman and Hall.

Grewal, M.S. and Andrews, A.P. (1993). Kalman filtering; Theory and practice. Prentice Hall Information and Systems Sciences series. Prentice-Hall, Englewood Cliffs, N.J.

Heald, J.P.M. and Stark, J. (2000). Estimation of noise levels for models of chaotic dynamical systems, Physical Review Letters, 84, 11, 2366-2369.

Holmes, P.J., (1979). A nonlinear oscillator with a strange attractor. Philosophical Transaction of the Royal Society London A, 292, 419-448.

Kantz, H., Schreiber, T., Hoffmann, I., Buzug, T., Pfister, G., Flepp, L.G., Simonet, J., Badii, R. and Brun, E. (1993). Nonlinear noise reduction: A case study on experimental data. Physical Review E, 48, 2, 1529-1538.

Leontaritis, I.J. and Billings, S.A. (1985a). Input-output parametric models for nonlinear systems part I: deterministic nonlinear systems. International Journal of Control, 41, 303-328; (1985b) Input-output parametric models for nonlinear systems part II: stochastic nonlinear systems. International Journal of Control, 41, 329-344.

Mackey, M.C. and Glass, L., (1977). Oscillation and chaos in physiological control systems. Science, 197, 287-289.

Schreiber, T. and Kantz, H. (1996). Observing and predicting chaotic signal: Is 2% noise too much? In Kravtsw, Y. and Kadtke, J. editors, Predictability of Complex Dynamical Systems, Springer Series in Synergetics No.69, 1-24, Springer New York.

Walker, D.M. and Mees, A. (1997). Noise reduction of chaotic systems by Kalman filtering and by shadowing. International Journal of Bifurcation and Chaos, 7, 3, 769-779.