ELSEVIER

Physica D 135 (2000) 305–330

PHYSICA D

www.elsevier.com/locate/physd

Time series analysis and prediction on complex dynamical behavior observed in a blast furnace

T. Miyano [a,*], S. Kimoto [d], H. Shibuta [b], K. Nakashima [b], Y. Ikenaga [c], K. Aihara [d]

[a] *Sumitomo Metal Industries, Ltd., Advanced Technology Research Laboratories, 1-8 Fusocho, Amagasaki, Hyogo 660, Japan*
[b] *Sumitomo Metal Industries, Ltd., Wakayama Steel Works, 1850 Minato, Wakayama 664, Japan*
[c] *Sumitomo Metal Industries, Ltd., Software Product Business Department, 3-11-36 Mita, Minato-ku, Tokyo 108, Japan*
[d] *Department of Mathematical Engineering and Information Physics, Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan*

Received 15 April 1996; received in revised form 11 November 1997; accepted 15 November 1997
Communicated by M. Mimura

Abstract

This paper describes a strategy for building a predictive model for actual complex time series. Time series data of temperature fluctuations observed in a blast furnace for iron-making are taken as an example. Chaotic features of the data are investigated with diagnostic algorithm for instability and parallelism of neighboring trajectories in phase space reconstructed from the time series data. Stationarity of the data is examined with diagnostic algorithm based on the KM_2O–Langevin equations developed by Okabe. A short time series for which no control actions were taken to the plant during measurement is diagnosed as possibly low-dimensional chaos, while for a long time series including many control actions during measurement, determinism is less visible and its predicted behavior exhibits a scaling property similar to self-affine random noise. Characteristic exponents are estimated from the scaling properties of the prediction error as a function of the prediction-time interval. Such information is exploited as prior knowledge for designing a generalized Gaussian radial basis function network as a predictor. The performance of the network is improved when linear algebraic polynomials are added to the network. The characteristic exponents estimated are used as reliability indices of forecasting future trends of the data.

PACS: 05.40.+j; 05.45.+b; 02.30.Mv

Keywords: Time series prediction; RBF networks; KM_2O–Langevin equation

1. Introduction

Predicting future behavior is a key ingredient in diverse fields from basic science to practical applications such as plant control and stock trading. In many actual plants, for example, past and current states of systems are monitored

* Corresponding author. Present address: Hirosaki University, Department of Intelligent Machines and Systems Engineering, 3 Bunkyocho, Hirosaki, Aomori 036-8561, Japan. Tel.: +81-172-39-3681; fax: +81-172-39-3513
E-mail address: miyano@cc.hirosaki-u.ac.jp (T. Miyano)

0167-2789/00/$ – see front matter
PII: S0167-2789(99)00135-9

as time series data of output signals from various sensors. Control procedures are usually chosen by reference to the history of observed data, to keep the system in prescribed conditions. If determinism is visible in the time series, one might expect that future behavior of the system would be so predictable to permit more effective control to avert system malfunctions occurring in any time interval. The understanding of chaos, however, has required us to alter such easy anticipation. In a chaotic system the influence of a small perturbation that has occurred at a certain instant will grow exponentially with time. As a result, the dynamical behavior of the system loses its long-term predictability. The situation, however, may not be hopeless. In some chaotic systems the decay of predictability may be so benign that future behavior may be predictable with a sufficient accuracy within a short time interval, when viewed on an appropriate timescale.

Nonlinear forecasting techniques such as neural networks [1] and radial basis function networks [2] have been recently shown to guarantee a good performance on short-term prediction about deterministic chaos. Suppose that a stationary time series is given and that determinism is visible with knowledge on irreducible degrees of freedom D and marginal prediction-time interval τ. Then one can design a network as input–output mapping from D-dimensional input vectors consisting of current and past values to output values τ time-interval into the future. Stationarity of the time series guarantees that the data used for learning reflect general dynamical behavior of the system. In most actual situations of interest, unfortunately, such information is not known in advance. In this context, characterizing both the deterministic feature and the stationarity of given data is of particular importance as a starting point for designing predictive models.

Discovering chaos from observational data is one of important issues with current interest. A prevalent method for this purpose is the Grassberger–Procaccia algorithm [3] that estimates the correlation dimension of a strange attractor from a time series. This approach, however, has a couple of flaws that the upper bound of the dimension estimates reliable depends on the size of data [4] and that it is prone to be fooled by colored noise stemming from stochastic processes even if a sufficiently large size of data is available [5]. On the other hand, nonlinear forecasting is becoming a popular approach as the alternative. Crucial progress in this discipline has been recently made by, e.g., Casdagli [2], Farmer and Sidorowich [6], and Sugihara and May [7]. The central idea of this approach is short-term predictability and long-term unpredictability of deterministic chaos. A prominent virtue of this method is that it may not require a large size of data like the Grassberger-Procaccia algorithm when the class of the approximating function of the predictor fits the dynamics underlying the data. In fact, Sugihara and May [7] have shown that chaos can be distinguished from uncorrelated random noise according to the decay of the correlation between predicted and actual values with increasing the prediction-time interval. However, their algorithm is also prone to be fooled by colored noise that also exhibits a decay of the prediction accuracy with the prediction-time interval. This should not be overlooked, since one should make use of a linear autoregressive predictor rather than a nonlinear predictor if given data represent colored noise (at least in our experience). To resolve this problem, Tsonis and Elsner [8] have recently proposed simple algorithm that permits distinguishing chaos from colored noise according to the scaling properties of the prediction error. This diagnostic algorithm works well when a time series includes no observational noise. A problem of misdiagnosis arises, however, when handling chaotic time series contaminated with additive random noise [9]. To fix the problem, another diagnostic algorithm based on the parallelism of neighboring trajectories in phase space has been shown to be useful [10,11]. The heart of the method is that if determinism is visible in a time series, neighboring trajectories reconstructed from the data should point to similar directions. This diagnostic algorithm seems tolerant to observational noise. Nonlinear forecasting associated with the diagnostic test for the parallelism may be a good tool for characterizing the dynamical nature of observational time series to acquire prior knowledge for designing a predictive model.

Moreover, interesting algorithm for estimating the degrees of stationarity of a time series has been recently developed on the basis of the KM_2O–Langevin equations by Okabe and his colleagues [12–18]. This approach views a time series as a realization of a Langevin equation. The degrees of stationarity are measured in terms of

the orthogonality of the fluctuation terms generated from the covariance of the data. A large size of data is not necessarily required in this method.

In this paper we describe a strategy for building predictive models for complex time series. Time series data of temperature fluctuations actually observed in a blast furnace for iron-making are taken as an example. The dynamical nature of the data as prior knowledge for designing the predictors is characterized by the diagnostic tests both for determinism and for stationarity. Generalized radial basis function networks as predictors are designed with the prior knowledge. The present paper is organized as follows. Section 2 provides theoretical preliminaries on the diagnostic tests and the predictive models. We resort to the Sugihara–May algorithm and the Kaplan–Glass algorithm to test for determinism and to the KM_2O–Langevin equations to test for stationarity. The theoretical foundation of generalized Gaussian radial basis function networks is given together with learning algorithm. A simple predictive device exploiting characteristic exponents as reliability indices is proposed for forecasting future trends of a time series. In Section 3 two kinds of actual time series are shown: a "short time series" for which no control procedures were taken to the plant during measurement, and a "long time series" including many control procedures during measurement. In Section 4 dynamical properties of the time series data are examined by the diagnostic algorithm. It is shown that the short time series is diagnosed as possibly stationary deterministic chaos, while for the long series determinism becomes less visible and stationarity is lost possibly due to control actions as nonstationary external noise. The characteristic exponents of both time series are estimated from the scaling property of the prediction error as a function of the prediction-time interval. In Section 5 the networks are designed in accordance with the diagnosis on the dynamical nature. The performance of the networks is shown to be improved when linear algebraic polynomials are added, although the null space of the stabilizer associated with Gaussian radial basis functions is zero. We also show the performance of the predictive device exploiting the characteristic exponents. Discussion is given in Section 6. Section 7 is conclusion.

2. Theoretical preliminaries

2.1. Diagnostic algorithm based on the instability of trajectories for determinism

Decay of predictability with time is a salient sign of deterministic chaos. This section describes diagnostic algorithm with nonlinear forecasting for the instability of neighboring trajectories reconstructed from a time series in phase space. The algorithm is based on the scaling property of the prediction error with the prediction-time interval. Given a time series $\{x(t)\}_{t=1}^{N}$ consisting of N data points of variable x equidistant in time, we construct phase space with delayed vectors generated from the time series as

$$\boldsymbol{x}(t) = (x(t), x(t - \Delta t), \ldots, x(t - (D-1)\Delta t)),$$

where Δt is an appropriately chosen sampling interval, and D denotes the embedding dimension related to irreducible degrees of freedom of the system. If determinism is visible in the data, future values at time $t + \tau\,\Delta t$ can be predicted from current and past values by an approximating function f that can be determined with example data representing general dynamical behavior of the series:

$$x(t + \tau\,\Delta t) = f[\boldsymbol{x}(t)] + \epsilon(\tau), \tag{1}$$

where τ is a certain integer corresponding to the prediction-time interval and $\epsilon(\tau)$ denotes random variables representing the approximation error. In the present work, the accuracy of deterministic prediction (as a measure of visible determinism) is measured in terms of the correlation coefficient between predicted and actual values and the root-mean-squared error normalized by the standard deviation of library data, denoted by $E(\tau)$. Suppose that an

appropriate function f is given. In a chaotic system, an infinitesimal distance between two nearby trajectories will grow exponentially with time to a finite magnitude. Such instability of trajectories is manifested in the exponential decay of the prediction precision with increasing the prediction-time interval, described by

$$\frac{E(\tau)}{E(1)} = \frac{\exp(\lambda\tau\Delta t)}{\exp(\lambda\Delta t)}, \tag{2}$$

where λ is the largest Lyapunov exponent [2,19]. Eq. (2) can be rewritten as

$$\log\frac{E(\tau)}{E(1)} = \lambda(\tau - 1)\Delta t. \tag{3}$$

The decay of the accuracy of prediction with τ, however, is not always the sign of deterministic chaos. A counterexample is self-affine random noise, i.e., colored noise stemming from a stochastic process. In such time series determinism would be seemingly visible to some extent due to sustaining autocorrelation. Fortunately, we can discriminate self-affine randomness from deterministic chaos according to the following scaling property of colored noise [9]:

$$\frac{E(\tau)}{E(1)} = \tau^H. \tag{4}$$

This can be rewritten as

$$\log\frac{E(\tau)}{E(1)} = H\log\tau \tag{5}$$

where H denotes the scaling exponent defined in relation to the statistical properties of self-affine random noise given by

$$\langle\kappa^{-H}|x(t+\kappa\Delta t) - x(t)|\rangle = \langle|x(t+\Delta t) - x(t)|\rangle, \tag{6}$$

where κ is the scaling factor, $\langle\cdot\rangle$ denotes the statistical moment in any order. The statistical properties are invariant under rescaling the timescale by κ and $x(t)$ by κ^{-H} [5]. This implies that there is no characteristic timescale for colored noise. Thus deterministic chaos can be distinguished from colored noise by comparing the degrees of linear correlation between the semilog plot (Eq. (3)) and the log–log plot (Eq. (5)). Note that the characteristic exponents can be estimated as the slope of the semilog or the log–log plot for smaller τ ($\tau \leq 6$ in the present work). The algorithm based on the scaling properties of colored noise was first introduced by Tsonis and Elsner [8]. The present expression of Eq. (5), however, may be more convenient for use as the counterpart of Eq. (3). In the present work, we apply the Sugihara–May predictor [7] as the approximating function, described in the next section, to this diagnostic algorithm.

2.2. The Sugihara–May predictor

The Sugihara–May predictor [7] is convenient for use because it does not require learning for adapting parameters included in the predictor, hence being computationally less expensive in comparison to neural networks and GRBF networks. Given N pairs of input vectors $\boldsymbol{x}(t)$ and the associated output values $x(t+\tau\Delta t)$, let the first n pairs be library patterns representing general dynamical behavior of the data. Then forecasts on the $N-n$ remaining input vectors, i.e. predictees, are made by

$$y(t+\tau\Delta t) = \frac{\sum_{k=1}^{D+1} x(t_k+\tau\Delta t)\exp(-d_k)}{\sum_{k=1}^{D+1}\exp(-d_k)}, \tag{7}$$

$$d_k = |\boldsymbol{x}(t) - \boldsymbol{x}(t_k)|, \tag{8}$$

where the summation is taken over the library patterns $\boldsymbol{x}(t_k)$ (k running from 1 to $D+1$) as $D+1$ closest neighbors forming the vertices of the smallest simplex including the predictee $\boldsymbol{x}(t)$ (t running from $n+1$ to N) in the D-dimensional Euclidean space. In this context, this predictive approach is local approximation. To find an appropriate value of D, various values of D are applied with $\tau = 1$ to search for the value of D that achieves the minimal error of prediction. Note that the time series should be stationary in order for the library data to represent general dynamical behavior.

2.3. Diagnostic algorithm based on the parallelism of trajectories for determinism

In nonlinear forecasting a true scaling property is often masked when a time series is contaminated with observational noise [9]. This suggests that it is dangerous to resort to a single diagnostic method to analyze observational time series. Another diagnostic test applied in this work is based on the algorithm developed by Wayland et al. [11] as a simpler variant of the Kaplan–Glass algorithm [10]. Let $\boldsymbol{y}(t)$ be the image of $\boldsymbol{x}(t)$, generated by time translation of the form: $\boldsymbol{y}(t) = \boldsymbol{x}(t + T\,\Delta t)$ with an appropriately chosen time interval T. We first find K nearest neighbors of $\boldsymbol{x}(t_0)$ in the sense of Euclidean distance, denoted as $\boldsymbol{x}(t_k)$. Let $\boldsymbol{y}(t_k)$ be the images of $\boldsymbol{x}(t_k)$. We next construct the translation vectors as

$$\boldsymbol{v}(t_k) = \boldsymbol{y}(t_k) - \boldsymbol{x}(t_k), \quad k = 0, 1, \ldots, K. \tag{9}$$

The central idea of this method is that the $K+1$ translation vectors should point in similar directions if determinism is visible in the data. The diversity in directions can be gauged by a useful measure referred to as the translation error E_{trans}:

$$E_{\text{trans}} = \frac{1}{K+1}\sum_{k=0}^{K}\frac{\|\boldsymbol{v}(t_k) - \langle\boldsymbol{v}\rangle\|^2}{\|\langle\boldsymbol{v}\rangle\|^2}, \tag{10}$$

where

$$\langle\boldsymbol{v}\rangle = \frac{1}{K+1}\sum_{k=0}^{K}\boldsymbol{v}(t_k).$$

The more visible determinism is, the smaller E_{trans} will be. To reduce the stochastic error associated with estimates, we seek the medians of E_{trans} for Q sets of N_0 randomly chosen $\boldsymbol{x}(t_0)$ and then take the average over the Q medians, according to Wayland et al. [11].

In this diagnostic algorithm the choice of D is important, since two neighboring trajectories with an inappropriate choice of D may not be close to each other in the true embedding space. It is thus necessary to examine the D-dependence of E_{trans}.

From numerical works on colored noises with various fractional power law spectral indices ($0 \leq \alpha \leq 2$), we tentatively determine the upper bound of visible determinism to be $E_{\text{trans}} = 0.5$ [9]. According to this criterion, time series stemming from stochastic processes are specified by E_{trans} exceeding 0.5. This algorithm seems tolerant to observational noise [9]. Besides estimating the degrees of visible determinism in an original time series, the algorithm is also applicable to testing for visible determinism in residuals in forecasting, for example, how well a linear predictor works to bleach noisy time series, as will be discussed in Section 6.

2.4. Diagnostic algorithm based on the KM_2O–Langevin equations for stationarity

Useful algorithm for estimating the degree of stationarity from a finite length of time series was introduced by Okabe and his collaborators [12–18]. The theory of KM_2O–Langevin equations is based on the fluctuation–dissipation

principle in non-equilibrium statistical physics [12–17]. This effective theory is composed of stationary analysis, causal analysis, and nonlinear prediction analysis. We review the stationary analysis and the method of application to observational time series in this subsection [12–18].

2.4.1. Stationary process

Let $\boldsymbol{Z} = (\boldsymbol{Z}(n); |n| \leq N)$ be any Q-dimensional stationary process on a probability space $(\Omega, \mathcal{B}, P)$. Then there exists a matrix function R: $\{-2N, -2N+1, \ldots, 2N\} \to M(Q; \boldsymbol{R})$ such that

$$\int_{\Omega} \boldsymbol{Z}(m)(\omega)^t(\boldsymbol{Z}(n)(\omega))\,\mathrm{d}P(\omega) = R(m-n), \tag{11}$$

where the matrix function $R(n)$ is the covariance matrix defined by

$$R(n) = \begin{pmatrix} R_{11}(n) & R_{12}(n) & \ldots & R_{1Q}(n) \\ R_{21}(n) & R_{22}(n) & \ldots & R_{2Q}(n) \\ \vdots & \vdots & \ddots & \vdots \\ R_{Q1}(n) & R_{Q2}(n) & \ldots & R_{QQ}(n) \end{pmatrix}. \tag{12}$$

The matrix function $R(\cdot)$ is called the covariance function associated with $\boldsymbol{Z}$, where the mean vector is assumed to be $\boldsymbol{0}$:

$$\int_{\Omega} \boldsymbol{Z}(m)(\omega)\,\mathrm{d}P(\omega) = \boldsymbol{0}. \tag{13}$$

2.4.2. KM_2O–Langevin fluctuation flow

We introduce a Q-dimensional flow $\boldsymbol{\nu}_+ = (\boldsymbol{\nu}_+(n) = {}^t(\nu_{+1}(n), \nu_{+2}(n), \ldots, \nu_{+Q}(n)); 0 \leq n \leq N)$ by

$$\boldsymbol{\nu}_+(0) \equiv \boldsymbol{Z}(0), \tag{14}$$

$$\boldsymbol{\nu}_+(n) \equiv \boldsymbol{Z}(n) - P_{\boldsymbol{M}_0^{n-1}(\boldsymbol{Z})}\boldsymbol{Z}(n) \quad (1 \leq n \leq N), \tag{15}$$

where $P_{\boldsymbol{M}_0^{n-1}(\boldsymbol{Z})}\boldsymbol{Z}(n)$ stands for the projection of $\boldsymbol{Z}(n)$ to the subspace $\boldsymbol{M}_0^{n-1}(\boldsymbol{Z})$ of $\boldsymbol{Z}$, defined by the closed linear hull of $\{\boldsymbol{Z}(l); 0 \leq l \leq n-1\}$. Similarly, we introduce a Q-dimensional flow $\boldsymbol{\nu}_- = (\boldsymbol{\nu}_-(l) = {}^t(\nu_{-1}(l), \nu_{-2}(l), \ldots, \nu_{-Q}(l)); -N \leq l \leq 0))$ by

$$\boldsymbol{\nu}_-(0) \equiv \boldsymbol{Z}(0), \tag{16}$$

$$\boldsymbol{\nu}_-(l) \equiv \boldsymbol{Z}(l) - P_{\boldsymbol{M}_{l+1}^{0}(\boldsymbol{Z})}\boldsymbol{Z}(l) \quad (-N \leq l \leq -1). \tag{17}$$

These two flows are called forward or backward KM_2O–Langevin fluctuation flows associated with $\boldsymbol{Z}$, respectively. The Kronecker product of $\boldsymbol{\nu}_\pm(n)$ is defined by

$$V_\pm(n) \equiv (\boldsymbol{\nu}_\pm(\pm n), {}^t\boldsymbol{\nu}_\pm(\pm n)) \qquad (0 \leq n \leq N). \tag{18}$$

The following relations (19)–(21) hold between $\boldsymbol{Z}$ and $\boldsymbol{\nu}_\pm$:

$$\begin{aligned} &\boldsymbol{M}_0^n(\boldsymbol{Z}) = \boldsymbol{M}_0^n(\boldsymbol{\nu}_+) \qquad (0 \leq n \leq N), \\ &\boldsymbol{M}_{-n}^0(\boldsymbol{Z}) = \boldsymbol{M}_{-n}^0(\boldsymbol{\nu}_-) \quad (0 \leq n \leq N), \end{aligned} \tag{19}$$

$$\begin{aligned} &(\boldsymbol{Z}(m), {}^t\boldsymbol{\nu}_+(n)) = 0 \qquad (0 \leq m < n \leq N), \\ &(\boldsymbol{Z}(-m), {}^t\boldsymbol{\nu}_-(-n)) = 0 \quad (0 \leq m < n \leq N), \end{aligned} \tag{20}$$

$$\begin{aligned}(\boldsymbol{\nu}_+(m), {}^t\boldsymbol{\nu}_+(n)) &= \delta_{m,n}\boldsymbol{V}_+(n) \quad (0 \le m, n \le N),\\ (\boldsymbol{\nu}_-(-m), {}^t\boldsymbol{\nu}_-(-n)) &= \delta_{m,n}\boldsymbol{V}_-(n) \quad (0 \le m, n \le N).\end{aligned} \tag{21}$$

2.4.3. KM_2O–Langevin equations and KM_2O–Langevin matrix

We define the following Toeplitz matrices as

$$\boldsymbol{T}_\pm(n) = \begin{pmatrix} R(0) & R(\pm 1) & \cdots & R(\pm(n-1)) \\ R(\mp 1) & R(0) & \cdots & R(\pm(n-2)) \\ \vdots & \vdots & \ddots & \vdots \\ R(\mp(n-1)) & R(\mp(n-2)) & \cdots & R(0) \end{pmatrix}. \tag{22}$$

We treat the case where the following condition is satisfied:

$$\boldsymbol{T}_\pm(n) \in GL(nQ; \boldsymbol{R}) \qquad (1 \le n \le N+1). \tag{23}$$

This condition is equivalent to the fact that two sets of vectors $\{\boldsymbol{Z}(n); 0 \le n \le N\}$ and $\{\boldsymbol{Z}(l); -N \le l \le 0\}$ are linearly independent, respectively. Under this condition there exists a unique system $\{\boldsymbol{\gamma}_+(n,t); 0 \le t < n \le N\}$ of $Q \times Q$-matrices such that

$$P_{\boldsymbol{M}_0^{n-1}(\boldsymbol{Z})}\boldsymbol{Z}(n) = -\sum_{t=0}^{n-1}\boldsymbol{\gamma}_+(n,t)\boldsymbol{Z}(t). \tag{24}$$

The forward and backward KM_2O–Langevin equations can hence be expressed as

$$\boldsymbol{Z}(\pm n) = -\sum_{t=1}^{n-1}\boldsymbol{\gamma}_\pm(n,t)\boldsymbol{Z}(\pm t) - \boldsymbol{\delta}_\pm(n)\boldsymbol{Z}(0) + \boldsymbol{\nu}_\pm(\pm n) \quad (1 \le n \le N). \tag{25}$$

The functions $\boldsymbol{\delta}_\pm(\cdot) \equiv \boldsymbol{\gamma}_\pm(\cdot, 0)$, $\boldsymbol{\gamma}_\pm(\cdot, *)$ and $\boldsymbol{V}_\pm(\cdot)$ are called the KM_2O–Langevin partial correlation function, the KM_2O–Langevin dissipation matrix, and the KM_2O–Langevin fluctuation matrix, respectively. Moreover, the system of these matrices is called the KM_2O–Langevin matrix associated with $\boldsymbol{Z}$ and defined by

$$\mathcal{L}M(\boldsymbol{Z}) \equiv \{\boldsymbol{\gamma}_\pm(n,t), \boldsymbol{\delta}_\pm(n), \boldsymbol{V}_\pm(l); 1 \le n \le N, 0 \le t \le n-1, 0 \le l \le N\}. \tag{26}$$

The following theorems hold on the KM_2O–Langevin matrix.

2.4.3.1. Dissipation–dissipation theorem. For any integer n, t $(1 \le t < n \le N)$,

$$\boldsymbol{\gamma}_\pm(n,t) = \boldsymbol{\gamma}_\pm(n-1, t-1) + \boldsymbol{\delta}_\pm(n)\boldsymbol{\gamma}_\mp(n-1, n-t-1). \tag{27}$$

2.4.3.2. Fluctuation–dissipation theorem For any integer n $(0 \le n \le N-1)$,

$$\boldsymbol{V}_\pm(n+1) = \boldsymbol{V}_\pm(n) - \boldsymbol{\delta}_\pm(n+1)\boldsymbol{V}_\mp(n)^t\boldsymbol{\delta}_\pm(n+1), \tag{28}$$

$$\boldsymbol{V}_+(n)^t\boldsymbol{\delta}_-(n+1) = \boldsymbol{\delta}_+(n+1)\boldsymbol{V}_-(n), \tag{29}$$

$$\boldsymbol{V}_+(n+1)^t\boldsymbol{\delta}_-(n+1) = \boldsymbol{\delta}_+(n+1)\boldsymbol{V}_-(n+1). \tag{30}$$

The following relations can be derived from Eqs. (28)–(30) for any integer n $(0 \le n \le N-1)$:

$$\boldsymbol{V}_+(0) = \boldsymbol{V}_-(0) = R(0), \tag{31}$$

$$\boldsymbol{V}_\pm(n+1) = (\boldsymbol{I} - \boldsymbol{\delta}_\pm(n+1)\boldsymbol{\delta}_\mp(n+1))\boldsymbol{V}_\pm(n). \tag{32}$$

2.4.4. A construction theorem

Conversely, let us suppose to be given a system $\{\boldsymbol{V}, \boldsymbol{\delta}_+(n);\ 1 \le n \le N\}$ of $Q \times Q$-matrices such that $\boldsymbol{V}$ is symmetric and positive definite. With the definition of $\boldsymbol{V}_+(0) = \boldsymbol{V}_-(0) = \boldsymbol{V}$, we can construct $(\boldsymbol{V}_+(n), \boldsymbol{\delta}_-(n), \boldsymbol{V}_-(n))$ from the fluctuation–dissipation theorem of Eqs. (28)–(30) by mathematical induction as follows:

$$\boldsymbol{V}_+(n) = \boldsymbol{V}_+(n-1) - \boldsymbol{\delta}_+(n)\boldsymbol{V}_-(n-1)^t\boldsymbol{\delta}_+(n), \tag{33}$$

$$\boldsymbol{\delta}_-(n)\boldsymbol{V}_+(n-1) = \boldsymbol{V}_-(n-1)^t\boldsymbol{\delta}_+(n), \tag{34}$$

$$\boldsymbol{V}_-(n) = \boldsymbol{V}_-(n-1) - \boldsymbol{\delta}_-(n)\boldsymbol{V}_+(n-1)^t\delta_-(n), \tag{35}$$

where $\boldsymbol{V}_+(n)$ are assumed to be non-negative definite $(1 \le n \le N)$. We next construct $\{\boldsymbol{\gamma}_\pm(n,t);\ 0 \le t < n \le N\}$ as follows:

$$\boldsymbol{\gamma}_\pm(n,t) = \boldsymbol{\gamma}_\pm(n-1,t-1) + \boldsymbol{\delta}_\pm(n)\boldsymbol{\gamma}_\mp(n-1,n-t-1). \tag{36}$$

Choose a set $\{w^\pm_{\pm i};\ 1 \le i \le Q(N+1)\}$ of $Q(N+1)$-dimensional random variables such that

$$(w^\pm_{\pm i}, w^\pm_{\pm j}) = \delta_{i,j} \tag{37}$$

and

$$(w^+_1, w^+_2, \dots, w^+_{Q(N+1)}) = (w^-_{-1}, w^-_{-2}, \dots, w^-_{-Q(N+1)}). \tag{38}$$

Then a Q-dimensional flow is defined as

$$\boldsymbol{\xi}_\pm(\pm n) \equiv^t (w^\pm_{\pm(nQ+1)}, w^\pm_{\pm(nQ+2)}, \dots, w^\pm_{\pm(nQ+Q)}), \tag{39}$$

which is a normalized white noise in such a sense that

$$(\boldsymbol{\xi}_\pm(\pm m),^t \boldsymbol{\xi}_\pm(\pm n)) = \delta_{m,n}\boldsymbol{I} \qquad (0 \le m, n \le N). \tag{40}$$

A Q-dimensional flow $\boldsymbol{\nu}_\pm = (\boldsymbol{\nu}_\pm(n);\ 0 \le n \le N)$ is also defined as

$$\boldsymbol{\nu}_\pm(\pm n) \equiv \boldsymbol{V}_\pm(n)^{1/2}\boldsymbol{\xi}_\pm(\pm n). \tag{41}$$

From Eqs. (40) and (41), we obtain

$$\boldsymbol{\nu}_+(0) = \boldsymbol{\nu}_-(0), \tag{42}$$

$$(\boldsymbol{\nu}_\pm(\pm m),^t \boldsymbol{\nu}_\pm(\pm n)) = \delta_{m,n}\boldsymbol{V}_\pm(n). \tag{43}$$

A Q-dimensional flow $\boldsymbol{Z} = (\boldsymbol{Z}(\pm n);\ 0 \le n \le N)$ is then constructed as

$$\boldsymbol{Z}(0) = \boldsymbol{\nu}_\pm(0), \tag{44}$$

$$\boldsymbol{Z}_\pm(\pm n) = -\sum_{t=0}^{n-1}\boldsymbol{\gamma}_\pm(n,t)\boldsymbol{Z}_\pm(\pm t) + \boldsymbol{\nu}_\pm(\pm n) \quad (1 \le n \le N). \tag{45}$$

Construction theorem. $\boldsymbol{Z}$ constructed above is stationary in the sense of weakly stationary process. From the fluctuation–dissipation theorem and the above construction theorem, therefore, the stationarity of $\boldsymbol{Z}_+$ is equivalent to the orthogonality of $\boldsymbol{\nu}_+$.

2.4.5. Stationary analysis of actual time series

This section describes the algorithm of stationary analysis by the KM_2O–Langevin equations for actual time series data. A Q-dimensional observational time series data $\{\boldsymbol{z}_+ = \boldsymbol{z}(n); 0 \le n \le N\}$ is regarded as a realization of $\boldsymbol{Z}$. The mean vector and the variance matrix are normalized to be $\boldsymbol{0}$ and $\boldsymbol{I}$, respectively.

Step 1. Calculate the sample covariance matrix function: $R^z = (R^z(n); -M \le n \le M)$

$$R^z_{ij}(n) \equiv \frac{1}{N+1}\sum_{m=0}^{M-n} z_i(n+m)\cdot z_j(m), \qquad R^z_{ji}(-n) \equiv R^z_{ij}(n), \tag{46}$$

where M is the reliable length of the covariance matrix function, given as $M = [3\sqrt{N+1}/Q] - 1$ [20]. We replace R in the previous equations by R^z hereafter.

Step 2. Set initial conditions:

$$\boldsymbol{V}^{\boldsymbol{z}}{}_{\pm}(0) = R^z(0), \tag{47}$$

$$\boldsymbol{\delta}^{\boldsymbol{z}}{}_{\pm}(1) = \boldsymbol{\gamma}^{\boldsymbol{z}}{}_{\pm}(1,0) = -R^z(\pm 1)R^z(0)^{-1}. \tag{48}$$

Calculate the sample KM_2O–Langevin matrices $\mathcal{L}M(\boldsymbol{z})$ recursively by

$$\boldsymbol{V}^{\boldsymbol{z}}{}_{\pm}(n) = (\boldsymbol{I} - \boldsymbol{\delta}^{\boldsymbol{z}}{}_{\pm}(n)\boldsymbol{\delta}^{\boldsymbol{z}}{}_{\mp}(n))\boldsymbol{V}^{\boldsymbol{z}}{}_{\pm}(n-1), \tag{49}$$

$$\boldsymbol{\delta}^{\boldsymbol{z}}{}_{\pm}(n+1) = -\left(R^z(\pm(n+1)) + \sum_{t=0}^{n-1}\boldsymbol{\gamma}^{\boldsymbol{z}}{}_{\pm}(n,t)R^z(\pm(t+1))\right)\boldsymbol{V}^{\boldsymbol{z}}{}_{\mp}(n)^{-1}, \tag{50}$$

$$\boldsymbol{\gamma}^{\boldsymbol{z}}{}_{\pm}(n+1,t) = \boldsymbol{\gamma}^{\boldsymbol{z}}{}_{\pm}(n,t-1) + \boldsymbol{\delta}^{\boldsymbol{z}}{}_{\pm}(n)\boldsymbol{\gamma}^{\boldsymbol{z}}{}_{\mp}(n,n-t) \qquad (1 \le t \le n < M) \tag{51}$$

where $\boldsymbol{\gamma}^{\boldsymbol{z}}{}_{\pm}(n,0) \equiv \boldsymbol{\delta}^{\boldsymbol{z}}{}_{\pm}(n)$.

Step 3. Construct the sample fluctuation flow: We can construct the sample fluctuation flow $\boldsymbol{\nu}^{\boldsymbol{z}}{}_{+i}$, where we label sample data by index i running from 0 to $N - M$.

$$\boldsymbol{\nu}^{\boldsymbol{z}}{}_{+}(0)_i = \boldsymbol{z}(i),$$

$$\boldsymbol{\nu}^{\boldsymbol{z}}{}_{+}(n)_i = \boldsymbol{z}(n+i) + \sum_{t=0}^{n-1}\boldsymbol{\gamma}^{\boldsymbol{z}}{}_{+}(n,t)\boldsymbol{z}(t+i) \qquad (1 \le n \le M). \tag{52}$$

Step 4. Check the orthogonality of sample fluctuation flow $\boldsymbol{\nu}^{\boldsymbol{z}}{}_{+i}$. Let us define one-dimensional data $\boldsymbol{\xi}_i = \{\xi(n)_i; 0 \le n \le Q(M+1)-1\}$ as $\boldsymbol{\nu}^{\boldsymbol{z}}{}_{+i}$ normalized by its variance $\boldsymbol{V}^{\boldsymbol{z}}{}_{+}$. Then the stationarity of $\boldsymbol{z}_+$ is equivalent to the orthogonality of $\boldsymbol{\xi}_i$, which can be expressed by the following set of three tests.

1. Test $\mathcal{M}$: whether or not the mean of $\boldsymbol{\xi}_i$ is 0.
2. Test $\mathcal{V}$: whether or not the variance of $\boldsymbol{\xi}_i$ is 1.
3. Test $\mathcal{O}$: whether or not the covariance of $\boldsymbol{\xi}_i$ is 0.

Step 5. Calculate the degree of stationarity R_s:

Here we describe in more detail Steps 4 and 5. The mean, the variance, and the covariance of $\boldsymbol{\xi}_i$ are given by

$$\mu^{\xi}_i = \frac{1}{Q(M+1)}\sum_{t=0}^{Q(M+1)-1}\xi(t)_i, \tag{53}$$

$$(v^{\xi}-1)_i = \frac{\sum_{t=0}^{Q(M+1)-1}(\xi(t)_i^2 - 1)}{\sqrt{\sum_{t=0}^{Q(M+1)-1}(\xi(t)_i^2-1)^2}}, \tag{54}$$

Table 1
Criteria for stationarity

	Criterion 1 (95%)	Criterion 2 (50%)
$\mathcal{M}$	$-1.96 < e_i^{\mathcal{M}} < 1.96$	$-0.68 < e_i^{\mathcal{M}} < 0.68$
$\mathcal{V}$	$-1.96 < e_i^{\mathcal{V}} < 1.96$	$-0.68 < e_i^{\mathcal{V}} < 0.68$
$\mathcal{O}$	$-0.29 < e_i^{\mathcal{O}} < 0.46$	$-0.14 < e_i^{\mathcal{O}} < 0.11$

$$R^{\xi}(n,m)_i = \frac{1}{Q(M+1)} \sum_{t=m}^{Q(M+1)-1-n} \xi(t)_i \xi(t+n)_i \qquad (1 \le n \le L, 0 \le m \le L-n), \tag{55}$$

where $L = \left[2\sqrt{Q(M+1)}\right]$ is the reliable length for the covariance of $\boldsymbol{\xi}_i$ [20]. Then $\overline{R^{\xi}}_i$ is defined as

$$\overline{R^{\xi}}_i = \frac{1}{H} \sum_{n=1}^{L} \sum_{m=0}^{L-n} R^{\xi}(n,m)_i, \tag{56}$$

where $H = L(L+1)/2$.

We check up the three criteria of Step 4 to judge the degree of stationarity in terms of the following estimative values:

$$e_i^{\mathcal{M}} = \sqrt{Q(M+1)} \cdot \mu_i^{\xi}, \tag{57}$$

$$e_i^{\mathcal{V}} = (v^{\xi} - 1)_i, \tag{58}$$

$$e_i^{\mathcal{O}} = \sqrt{Q(M+1)} \cdot \overline{R^{\xi}}_i. \tag{59}$$

To gauge the degree of stationary, we count the number of samples whose estimative values lie in the ranges theoretically expected for stationary time series, as summarized in Table 1, then represent the fractional number of the samples as $E_{\mathcal{M}1}$ and $E_{\mathcal{M}2}$, $E_{\mathcal{V}1}$ and $E_{\mathcal{V}2}$, $E_{\mathcal{O}1}$ and $E_{\mathcal{O}2}$, corresponding to test $\mathcal{M}$, $\mathcal{V}$, and $\mathcal{O}$, respectively. We further assume a logistic function to represent a measure of stationarity as follows:

$$\mathcal{X} = f(E_{\mathcal{X}}) = \frac{1}{1 + \exp\{-(E_{\mathcal{X}} - (E_{\mathcal{X}}(\text{theory}) - E_{\mathcal{X}}^0))/T_0\}}, \tag{60}$$

where $E_{\mathcal{X}}$ ($\mathcal{X} = \mathcal{M}1, \mathcal{M}2, \mathcal{V}1, \mathcal{V}2, \mathcal{O}1$, or $\mathcal{O}2$) denotes the fractional number of samples fitting each criterion, $E_{\mathcal{X}}(\text{theory})$ is the corresponding value theoretically expected for the normal distribution, i.e., 0.95 or 0.5 for criterion 1 or 2, respectively, and $E_{\mathcal{X}}^0$ is a constant given by

$$E_{\mathcal{X}}^0 = a_{\mathcal{X}} \cdot 2^{-\log_{10}(N+1)},$$

where $a_{\mathcal{X}}$ is a parameter whose values are shown in Table 2 [18]. T_0 is a constant given by

$$T_0 = \frac{E_{\mathcal{X}}^0}{\ln\{(1-\epsilon)/\epsilon\}},$$

Table 2
Values of $a_{\mathcal{X}}$

	Criterion 1	Criterion 2
$\mathcal{M}$	$a_{\mathcal{M}1} = 1.0$	$a_{\mathcal{M}2} = 1.2$
$\mathcal{V}$	$a_{\mathcal{V}1} = 1.2$	$a_{\mathcal{V}2} = 1.5$
$\mathcal{O}$	$a_{\mathcal{O}1} = 1.5$	$a_{\mathcal{O}2} = 1.7$

where ϵ is set to 10^{-10}. Finally, we take the product of the estimative values $\mathcal{M}1$, $\mathcal{M}2$, $\mathcal{V}1$, $\mathcal{V}2$, $\mathcal{O}1$, and $\mathcal{O}2$ obtained by Eq. (60), to evaluate the degree of stationarity, R_s, as the indicator lying between 0 and 1 [18] as follows:

$$R_s = \mathcal{M}1 \cdot \mathcal{M}2 \cdot \mathcal{V}1 \cdot \mathcal{V}2 \cdot \mathcal{O}1 \cdot \mathcal{O}2. \tag{61}$$

A stationary time series in the exact sense of weakly stationary process corresponds to $R_s = 1$.

2.5. Generalized radial basis function networks and learning rule

Radial basis function (RBF) networks are a class of approximation technique based on regularization theory [21–23]. The networks generate a regression surface as an input–output mapping from sparse examples under a prior assumption on the smoothness of the mapping. Let $\{(\boldsymbol{x}(t), x(t+\tau\Delta t)) \in \boldsymbol{R}^D \times \boldsymbol{R}\}_{t=1}^N$ be given examples from which the approximating function f is estimated. To evaluate the performance of the network as a predictor, we define the risk functional as

$$H[f] = \frac{1}{N}\sum_{t=1}^{N}[x(t+\tau\Delta t) - f(\boldsymbol{x}(t))]^2 + \lambda_{\mathrm{reg}}\psi[f], \tag{62}$$

$$\psi[f] = \|\hat{P}f\|^2, \tag{63}$$

where $\lambda_{\mathrm{reg}} \geq 0$ is an appropriately chosen regularization parameter, $\|\cdot\|$ denotes an L^2 norm, and $\hat{P}$ is a certain differential operator associated with prior knowledge about the smoothness of f, referred to as a stabilizer. The first term of the right hand side of Eq. (62) measures how well the network reproduces library examples, while the second term generates penalty associated with oscillating behavior in f. The optimal solution of f for the support of $\psi[f]$, Supp $\psi = \{f|\hat{P}f \neq 0\}$, can be derived from

$$\frac{\delta H[f]}{\delta f} = 0.$$

The kernel Ker $\psi = \{f|\hat{P}f = 0\}$, i.e., the null space of the stabilizer consisting of the k derivative is usually given by algebraic polynomials of degree $k-1$ or less. Thus the approximating function space is the direct sum of the support and the kernel of the stabilizer, Ker $\psi \oplus$ Supp ψ. In many applications of interest, f belonging to Supp ψ can be given by a linear combination of RBFs, $G(\|\boldsymbol{x}(t) - \boldsymbol{x}(t_k)\|)$, where the centers $\boldsymbol{x}(t_k)$ with k running from 1 to N correspond to given examples. Poggio and Girosi [21,22] extended the RBF networks technique to reduce the number of RBFs to fewer than the number of examples, instead permitting the centers to move during learning. Such an approximation scheme is cited as generalized radial basis function (GRBF) networks. In the present work we make use of Gaussians as the basis functions:

$$f[\boldsymbol{x}(t)] = \sum_{h=1}^{N_h} c_h \exp\left\{-\beta_h \sum_{i=0}^{D-1}[x(t-i\Delta t) - \theta_{hi}]^2\right\}, \tag{64}$$

where c_h, β_h, and θ_{hi} are parameters to be determined through learning from examples, and N_h denotes the number of Gaussians, i.e., hidden nodes of the network. Note that the null space associated with Gaussians is zero, so that no polynomial terms may be requested to complement Ker ψ [21,22]. Nevertheless, we may add linear terms corresponding to linear autoregression (AR) to the right hand side of Eq. (64):

$$f[\boldsymbol{x}(t)] = \sum_{h=1}^{N_h} c_h \exp\left\{-\beta_h \sum_{i=0}^{D-1}[x(t-i\Delta t) - \theta_{hi}]^2\right\} + \sum_{i=0}^{D-1} a_i x(t-i\Delta t) + d, \tag{65}$$

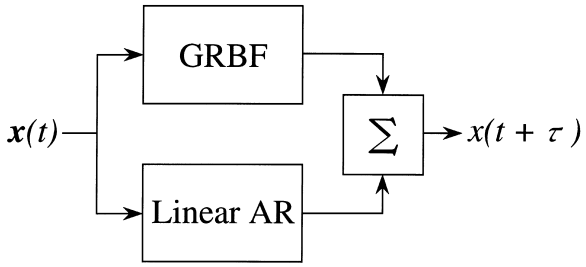

Fig. 1. Schematic diagram of GRBF networks with linear algebraic polynomials.

where a_i and d are parameters to be optimized (Fig. 1). As will be shown in Section 5, the performance of the networks appears to be improved by adding the linear terms when handling chaotic time series contaminated with observational random noise.

Optimizing the parameter values included in GRBF networks is rather computationally expensive. Once the optimization is performed, however, it consumes little time to make predictions. Another virtue of GRBF networks is the capability of machine-learning, i.e., on-line learning in much the same way as multilayer perceptrons. In the present work stochastic gradient descent is applied to adjust c_h, β_h, and θ_{hi} so as to minimize the empirical risk functional, i.e. $H[f]$ with $\lambda_{\text{reg}} = 0$ [24,25]:

$$\frac{\mathrm{d}c_h}{\mathrm{d}s} = -\omega_1 \frac{\partial H[f]}{\partial c_h}, \tag{66}$$

$$\frac{\mathrm{d}\beta_h}{\mathrm{d}s} = -\omega_2 \frac{\partial H[f]}{\partial \beta_h} + \eta(s), \tag{67}$$

$$\frac{\mathrm{d}\theta_{hi}}{\mathrm{d}s} = -\omega_3 \frac{\partial H[f]}{\partial \theta_{hi}} + \eta(s), \tag{68}$$

where s is the time parameter in the iteration loop of the optimization process, $\omega_1, \omega_2, \omega_3 > 0$ are appropriately chosen learning rates, and $\eta(s)$ can be Gaussian white noise or self-affine random noise with power law spectral index $\alpha = 1$. The empirical risk functional is not convex with respect to β_h and θ_{hi}. This implies that the network would get stuck in local minima in $H[f]$. The random noise $\eta(s)$ is expected to make the network jump out of such local minima. For GRBF networks with linear terms, a_i and d are determined using least-mean-squared-error fitting to minimize the empirical risk functional. The optimization of the linear parameters precedes training of the GRBF network that views the linear parameters as constants. This means that the GRBF network is optimized so as to minimize the risk functional consisting of residuals of the linear AR predictor. In this context, the present network architecture may be reminiscent of projection pursuit regression [26].

2.6. A predictive device for forecasting future trends

In many applications it may be sufficient to forecast future trends of a time series instead of forecasting the exact values that the series will take. When handling complex time series, one must keep in mind that the accuracy of prediction will inevitably suffer from deterioration with increasing the prediction-time interval. We propose a simple predictive device taking such deterioration into account. Let $\Delta y(\tau)$ be the differences between predicted values $y(t + \tau\Delta t)$ and $y(t + (\tau - 1)\Delta t)$ with τ running from 1 through $\tau_{\max}$, where $y(t)$ corresponds to the actual value at a current instant: $y(t) = x(t)$. We consider a trend index $\tau_{\max}$ time-steps into the future, $P(\tau_{\max})$ defined by

$$P(\tau_{\max}) = \frac{\sum_{\tau=1}^{\tau_{\max}} w(\tau)\Delta y(\tau)}{\tau_{\max}\sigma}, \tag{69}$$

where $w(\tau)$ is a weight coefficient representing deterioration in the prediction precision with τ, and σ is the standard deviation of the library data used for optimizing the predictor. We may define $w(\tau)$ with the largest Lyapunov exponent λ as

$$w(\tau) = \exp(-\lambda\tau) \tag{70}$$

for chaotic time series and

$$w(\tau) = \tau^{-H} \tag{71}$$

with the scaling exponent H for colored noise. The characteristic exponents can be estimated as the initial slope of the semilog or log–log plot, as described in Section 2.1. It should be noted that the predictive device would work for colored noise with higher α, since for such time series determinism is seemingly visible to some extent due to high autocorrelation sustaining over a certain period. We may interpret the meaning of $P(\tau_{\max})$ as follows. If $P(\tau_{\max}) = 1$, then there will be rising trends in time elapsing of $\tau_{\max}$ by an amount of σ relative to a current value $x(t)$. If $P(\tau_{\max}) = -1$, there will be sinking trends in time elapsing of $\tau_{\max}$ by an amount of σ relative to $x(t)$. The performance of the device will be tested in Section 5.

3. Observed time series

A blast furnace is a huge reactor for iron-making, e.g. 100 m in height and 10 m in diameter. The plant is equipped with cooling apparatus for water-cooling its side wall to prevent from damage by heat. The temperature difference between input and output cooling water is an indicator of importance for controlling the plant. Various actions for plant control are usually taken by reference to the trends of the temperature sequences. Fig. 2 shows a time series of the temperature difference for which no control actions were taken during measurement. The series consists of 528 data points observed at a sampling time of 10 min, labeled as “short time series”. The autocorrelation function and the power spectrum of the series are illustrated in Figs. 3(a) and (b), respectively. Fig. 4 depicts the first 500 data

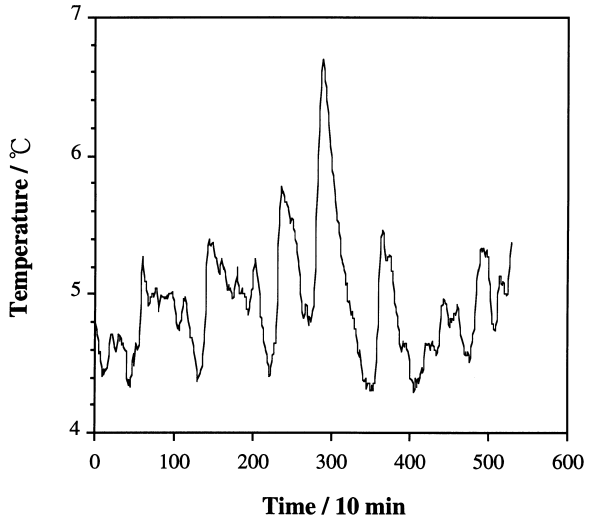

Fig. 2. Time series of the temperature difference (labeled as “short time series”) for which no control actions were taken during measurement, consisting of 528 data points. The sampling time Δt is 10 min.

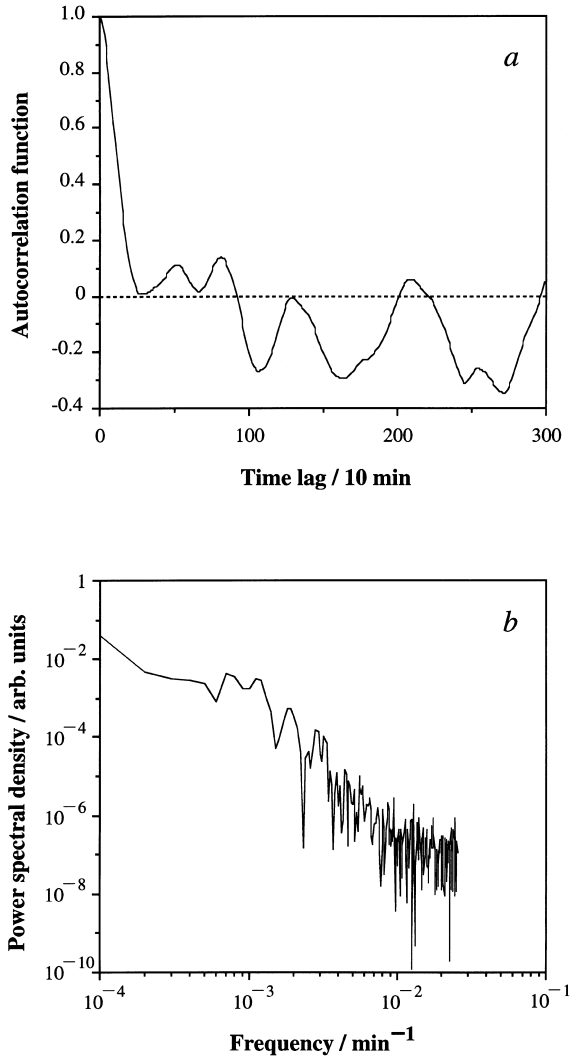

Fig. 3. Autocorrelation function (a) and power spectrum (b) of the short time series. A Hanning window is used in (b).

points of another time series for which many control actions were taken during measurement. The series consists of 8189 data points, labeled as “long time series”. The autocorrelation function and the power spectrum of the series are illustrated in Figs. 5(a) and (b), respectively. Both time series were observed for one of the cooling apparatus, so that the data represent local information at the same part of the system. No periodic signals are included in both data.

4. Characterization of the dynamical nature of the time series

4.1. Test for determinism

The diagnostic algorithm based on nonlinear forecasting is first applied to characterize the dynamical nature of the short time series. The Sugihara–May predictor is used with library patterns generated from the first half of the

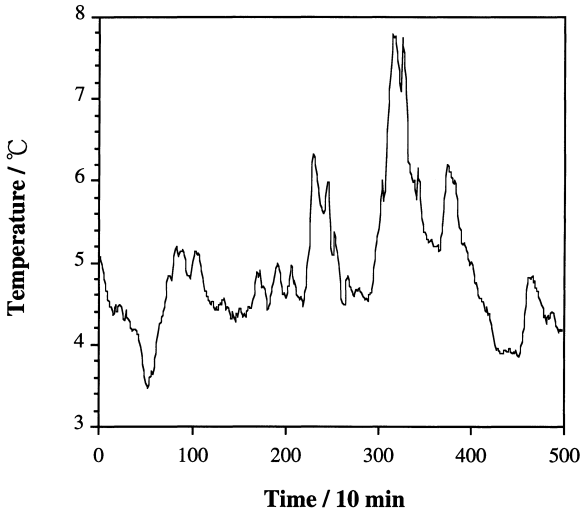

Fig. 4. The first 500 data points of a time series of the temperature difference (labeled as "long time series") for which many control actions were taken during measurement, consisting of 8189 data points. The sampling time Δt is 10 min.

Table 3
Degrees of linear correlation for the semilog plot and estimates of the Lyapunov exponent for the short time series, see Fig. 6

Embedding dimension D	Correlation coefficient γ^2	Lyapunov exponent λ ($\times 0.1\text{min}^{-1}$)
3	0.999	0.108
6	1.000	0.098
10	0.999	0.088

Table 4
Degrees of linear correlation for the log–log plot and estimates of the scaling exponent for the long time series, see Fig. 7

Embedding dimension D	Correlation coefficient γ^2	Scaling exponent H
3	0.999	0.577
6	0.992	0.576
10	0.979	0.458

series. Fig. 6(a) illustrates the correlation coefficient between predicted and actual values as a function of D with $\tau = 1, 3, 10$. The optimal value of D corresponding to the peak correlation coefficient is around 2–3 independently of τ. Figs. 6(b)–(d) show the correlation coefficient as a function of τ, the semilog plot of τ versus $\log[E(\tau)/E(1)]$, and the log–log plot of $\log \tau$ versus $\log[E(\tau)/E(1)]$, respectively. The short time series exhibits linear correlation in the semilog plot irrespectively of the choice of D. The correlation coefficient γ^2 indicating the degrees of linearity of the semilog plot and the largest Lyapunov exponent λ estimated as the slope of the straight line are summarized in Table 3. The corresponding results on the long time series are shown in Figs. 7(a)–(d). The optimal value of D is around 4. The long time series exhibits linear correlation in the log–log plot, in contrast to the short time series. The degrees of linearity of the log–log plot and the scaling exponent H estimated as the slope of the straight line are summarized in Table 4. Nonlinear forecasting suggests that the statistical properties of the prediction error are characteristic of low-dimensional deterministic chaos for the short time series and of self-affinity like colored noise for the long time series.

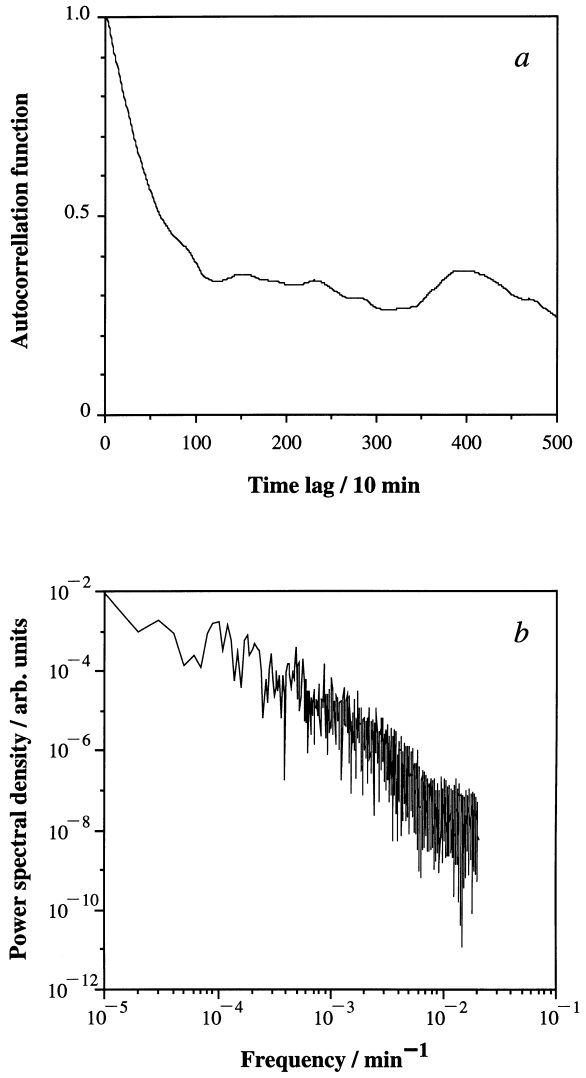

Fig. 5. Autocorrelation function (a) and power spectrum (b) of the long time series. A Hanning window is used in (b).

The Wayland algorithm is next applied to the data. Fig. 8 illustrates the average over the medians of E_{trans} as a function of D. The dash–dotted line indicates the upper bound of E_{trans} for visible determinism. According to the criterion, determinism is fairly visible for the short time series. On the other hand, determinism becomes less visible for the long time series. However, E_{trans} is below the upper bound for visible determinism, which indicates that the long time series includes deterministic ingredients more than expected for random noise. Although the statistical properties in the prediction error exhibit self-affinity like colored noise, the long time series may not be characterized entirely as colored noise. In numerical experiments on noisy chaotic time series such as a Lorenz attractor [9], $E_{\text{trans}} \ll 0.1$ in the case of no observational noise contamination, increasing up to 0.5 with increase in the noise level. We thus infer that both time series may be basically related to deterministic dynamics contaminated possibly with additive random noise that is enhanced by control actions. Hence a GRBF network with a linear autoregressive predictor may be a good predictive model to approximate the dynamic behavior of the time series.

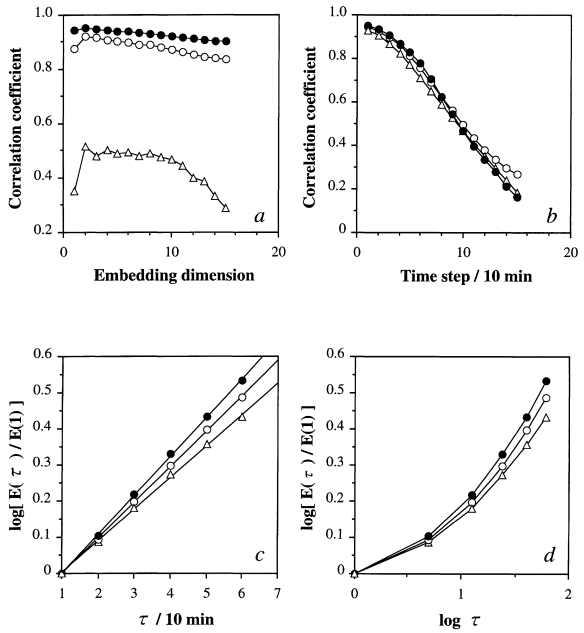

Fig. 6. Nonlinear forecasts by the Sugihara–May predictor about the short time series. Library patterns ($\Delta t = 10$ min) are generated from the first half (264 data points) of the series to forecast the last half. (a) The correlation coefficient between predicted and actual values as a function of D with $\tau = 1$ (●), 3 (○), 10 (△). (b) The correlation coefficient as a function of τ. (c) The semilog plot. Straight lines indicate linear fitting in the least-squared-error sense. (d) The log–log plot. In (b)–(d) $D = 3$ (●), 6 (○), 10 (△). See Table 3.

Table 5
Degrees of stationarity for the short time series

	$x(t)$	$x(t)^2$	$x(t)^3$	arct($x(t)$)	log($x(t)+1$)
R_s	0.994	0.000	0.000	1.000	0.999

Table 6
Degrees of stationarity for the long time series

	$x(t)$	$x(t)^2$	$x(t)^3$	arct($x(t)$)	log $x(t)$
R_s	0.000	0.000	0.000	0.000	0.000

4.2. Test for stationarity

The diagnostic algorithm based on the KM_2O–Langevin equations is applied to test the stationarity. Estimates of the degrees of stationarity, R_s, for the short time series are summarized in Table 5 where stationarity is also examined for nonlinear transformations with arctan $x(t)$, log $x(t)$, $x^2(t)$, and $x^3(t)$. The results show that the short time series is stationary in the sense of weakly stationary process. In Table 6 the corresponding estimates are summarized for the long time series. The long time series as a whole is non-stationary, in contrast to the short time series. However, some parts of the long time series exhibit weak stationarity, as shown in Table 7.

Table 7
Degrees of stationarity for each part of the long time series

Period	$x(t)$	$x(t)^2$	$x(t)^3$	arct($x(t)$)	log $x(t)$
0–527	0.806	0.000	0.000	1.000	0.996
528–1055	0.030	0.000	0.000	0.325	0.559
1056–1583	0.999	0.000	0.000	1.000	1.000
1584–2111	0.998	0.000	0.000	1.000	0.998
2112–2639	1.000	0.000	0.000	1.000	1.000
2640–3167	1.000	0.000	0.000	1.000	0.831
3168–3695	0.415	0.000	0.000	0.929	0.000
3696–4223	0.002	0.000	0.000	0.652	0.000
4224–4751	0.000	0.000	0.000	0.999	0.000
4752–5279	1.000	0.000	0.000	1.000	1.000
5280–5807	0.000	0.000	0.000	0.998	0.000
5808–6335	0.000	0.000	0.000	0.000	0.000
6336–6863	0.000	0.000	0.000	0.694	0.000
6864–7391	0.999	0.000	0.000	1.000	1.000
7392–7919	1.000	0.487	0.000	1.000	1.000
7920–8188	0.953	0.001	0.000	1.000	0.970

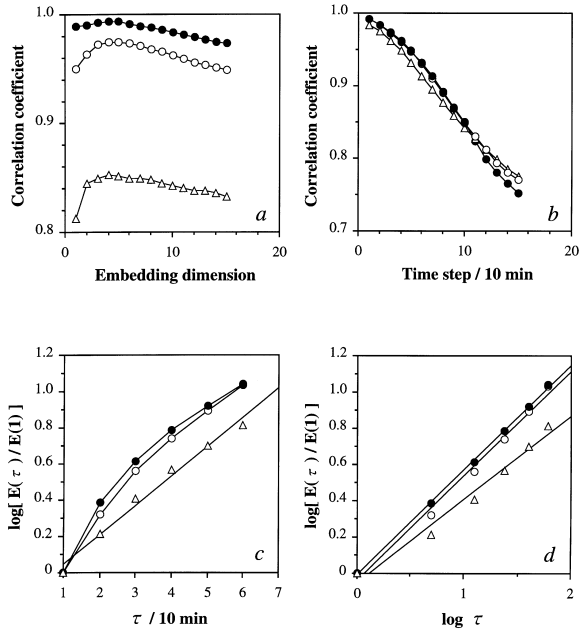

Fig. 7. Nonlinear forecasts by the Sugihara–May predictor about the long time series. Library patterns ($\Delta t = 10$ min) are generated from the first half (4000 data points) of the data to forecast the last half. (a) The correlation coefficient between predicted and actual values as a function of D with $\tau = 1$ (●), 3 (○), 10 (△). (b) The correlation coefficient as a function of τ. (c) The semilog plot. (d) The log–log plot. In (b)–(d) $D = 3$ (●), 6 (○), 10 (△). Straight lines indicate linear fitting in the least-squared-error sense. See Table 4.

5. Prediction by generalized radial basis function networks

The preceding diagnosis indicates that the short time series exhibits stationarity and visible determinism providing the instability of trajectories characteristic of deterministic chaos. A GRBF network as a nonlinear predictor with

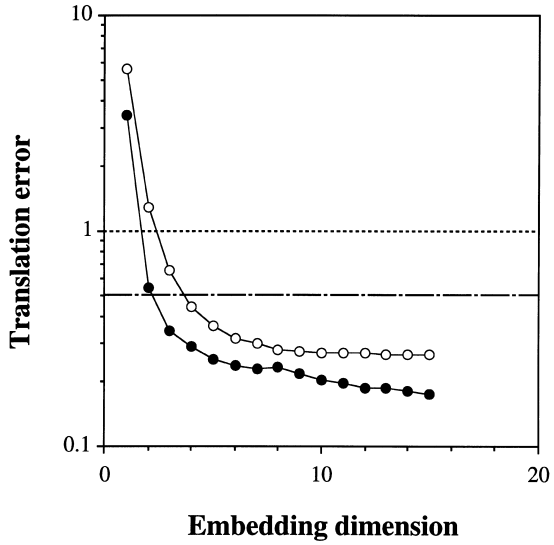

Fig. 8. The average over 20 medians of E_{trans} as a function of D for the short time series (●) and the long time series (○). The phase space is generated with $\Delta t = 30$ min. The medians are estimated for 20 sets of 500 randomly chosen vectors with 4 nearest neighbors and the associated images generated by time translation with $T = 5$. The dash–dotted line indicates the upper bound of E_{trans} for visible determinism.

the choice of $D = 3$ may be appropriate for short-term prediction of the short time series. Although for the long time series determinism is less visible and there is no stationarity as a whole possibly due to external perturbations of control actions, the time series includes some stationary parts and exhibits more determinism than expected for random noise. Hence a similar architecture of GRBF networks with $D = 3$ may be also applicable to the long time series. Note that for the long time series a linear predictor such as an AR model could be more appropriate, which may depend on how much nonlinearity was masked by the control actions. Later we will compare the difference in performance between the GRBF network and the AR predictor to estimate the degrees of nonlinearity in the data. The AR predictor is of the form

$$f[\boldsymbol{x}(t)] = \sum_{i=0}^{D-1} a_i x(t - i\Delta t) + d. \tag{72}$$

We set $\tau = 1$ for the predictors, which corresponds to forecasting the nearest future behavior of the series. When making forecasts $\tau \geq 2$ time-steps into the future, iterative forecasting is applied in which predicted values at a current time step are fed back to the input node at the next time step and this procedure is iterated $\tau - 1$ times. The linear parameters a_i and d are determined by least-mean-squared-error fitting. The AR predictor is used as the linear terms of the GRBF network.

Figs. 9(a) and (b) show the performance of a GRBF network (2 hidden nodes) with the linear terms for the short time series. The network has been trained for the library patterns generated from the first 300 data points. For comparison, the performances of the GRBF network with the linear terms, a GRBF network (7 hidden nodes) without the linear terms, and the linear AR predictor are summarized in Table 8. The time series sustains high autocorrelation over a certain period, so that a trivial predictor $y(t + \tau\Delta t) = x(t)$ can provide seemingly good accuracy of prediction, which is also shown in Table 8. In order for the predictors to make sense, the prediction error should be less than that of the trivial predictor. Note that the GRBF network achieves better performance when the linear terms are added, the implication of which will be discussed in Section 6. For $\tau = 1$ the AR predictor works as well as the GRBF network with the linear terms. As τ increases, however, the GRBF with the linear terms surpasses

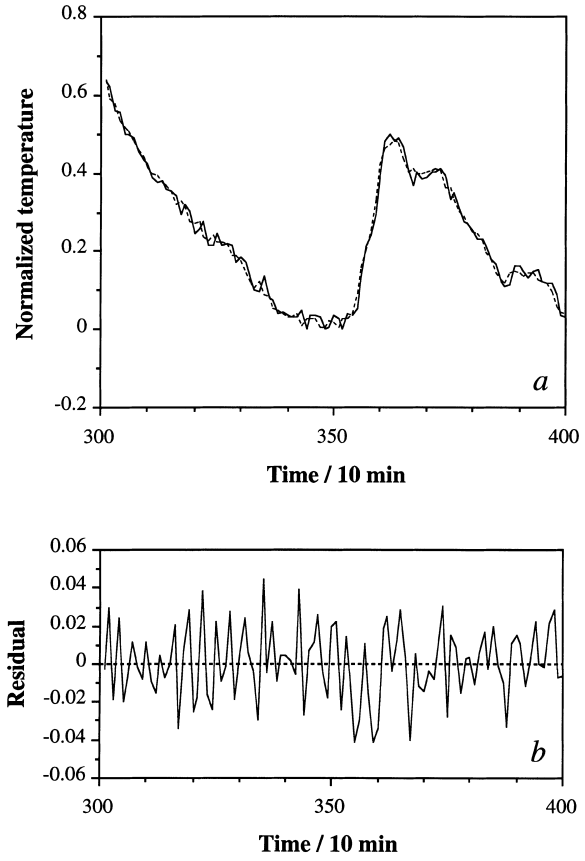

Fig. 9. (a) The first 100 points of 228 forecasts ($\tau = 1$) for the short time series by a GRBF network with linear algebraic polynomials. Predicted and the actual values are indicated by solid and dashed curves, respectively. The temperature is normalized with respect to the difference between the maximum and the minimum values. The network has been trained for the library patterns generated from the first 300 data points. (b) Residuals obtained by subtracting actual values from predicted values.

Table 8
Predictive performances of a GRBF network (2 hidden nodes) with linear polynomials, a GRBF network (7 hidden nodes) without linear polynomials, an AR predictor, and a trivial predictor $y(t + \tau\Delta t) = x(t)$ for the short time series

Prediction error	Predictive model			
$E(\tau)$	GRBF with linear terms	GRBF	Linear AR	$y(t + \tau) = x(t)$
$E(1)$	0.134	0.162	0.133	0.166
$E(3)$	0.324	0.349	0.329	0.422

the AR predictor in performance, as shown in Fig. 10. This means that the same approximation error one time-step into the future does not necessarily guarantee that the distinct predictors have learned the same dynamical behavior. We interpret that these observations are the signature of nonlinearity of the dynamics underlying the time series.

The Lyapunov exponent of the short time series has been inferred to be $\lambda \approx 0.10$ ($\times 0.1\,\mathrm{min}^{-1}$) from the scaling properties of the prediction error. We exploit the estimate in the predictive device described in Section 2.5. Fig. 11 shows the performance of the predictive device, where predictions $\tau = 1 - 3$ time-steps into the future are made iteratively by the GRBF with the linear terms. The predictive device seems to successfully reproduce the corresponding actual trends.

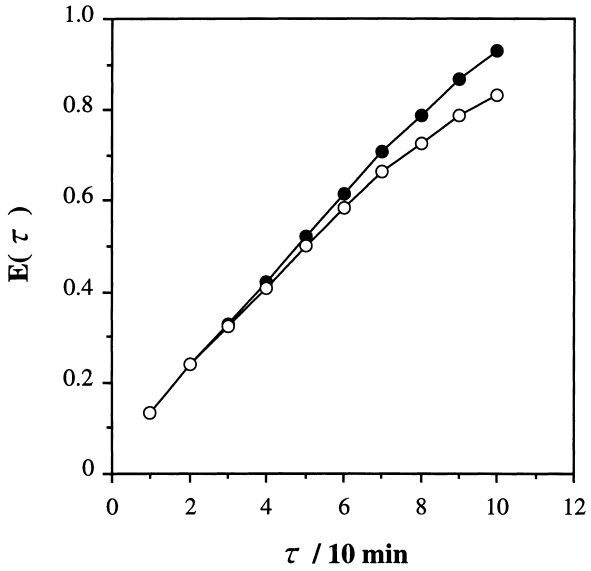

Fig. 10. Comparison of the predictive performances between the GRBF network with the linear terms (○) and the linear AR predictor (●) for the short time series.

Table 9
Predictive performances of a GRBF network (3 hidden nodes) with linear polynomials, an AR predictor, and a trivial predictor $y(t+\tau\Delta t) = x(t)$ for the long time series

Prediction error	Predictive model		
$E(\tau)$	GRBF with linear terms	Linear AR	$y(t+\tau) = x(t)$
$E(1)$	0.079	0.079	0.085
$E(3)$	0.171	0.172	0.187

Figs. 12(a) and (b) illustrate the performance of a GRBF network (3 hidden nodes) with the linear terms for the long time series. The network has been trained with the library patterns generated from the first 4000 data points. In this case the size of the library examples is larger by one order than in the case of the short time series. Thus we may add one more hidden node to the network to enhance its representation capacity within a similar generalization error [27]. The performances of the GRBF network with the linear terms, the linear AR predictor, and the trivial predictor are summarized in Table 9. Fig. 13 shows the performances of the GRBF network and the AR predictor as a function of τ. In this case the GRBF network works slightly better than the AR predictor for larger τ. This suggests that nonlinearity is not entirely masked in the long time series in spite of control actions as external random noise. According to the scaling properties of the prediction error, a scaling exponent of $H = 0.58$ has been estimated for the long time series, although the dynamical nature of the data may not be entirely ascribed to colored noise. The estimate of H is exploited to forecast the future trends of the long series, shown in Fig. 14 where the predictive device is based on the GRBF network with the linear terms. Predicted trends do not match the corresponding actual trends, although the predictor appears to provide statistically good accuracy of prediction for the nearest-future predictions.

6. Discussion

The dynamical behavior of the temperature sequences reflecting local states of the blast furnace seems to be chaotic and stationary when no control actions are taken to the system. However, determinism becomes less visible

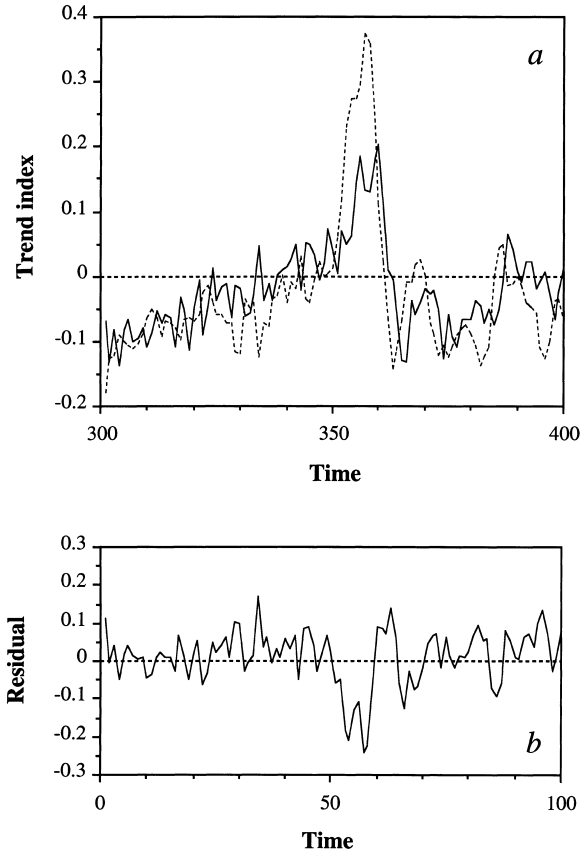

Fig. 11. Future trend forecasts by the predictive device of Eq. (69) about the short time series ($\tau_{\max} = 3$). The Lyapunov exponent is $\lambda = 0.10$. The GRBF network with the linear terms is used iteratively to make forecasts $\tau = 1$–3 time-steps into the future. Solid curve and dashed curve indicate predicted trends and the corresponding actual trends, respectively. The actual trends are calculated with the reliability factor $w(\tau) = 1$.

when many control actions are added to the system. Interestingly, the predicted behavior exhibits a scaling property like self-affine random motion and its stationarity is violated, although the origin of the self-affinity is unclear. In spite of the scaling property of the prediction error, the dynamical nature of the long time series may not be ascribed entirely to colored noise because of the degrees of visible determinism estimated by the Wayland algorithm. These observations indicate that it is hazardous to resort to a single diagnostic method to characterize dynamical properties of complex time series.

A remaining issue to be discussed is the effect of the linear algebraic polynomials on the Gaussian GRBF networks. Regularization theory shows that the networks do not require algebraic polynomials when Gaussians are used as the

Fig. 12. (a) The first 100 points of 4186 forecasts ($\tau = 1$) about the long time series by a GRBF network with linear algebraic polynomials. Predicted and the actual values are indicated by solid and dashed curves, respectively. The temperature is normalized with respect to the difference between the maximum and the minimum values. The network has been trained for the library patterns generated from the first 4000 data points. (b) Residuals obtained by subtracting actual values from predicted values.

Fig. 13. Comparison of the predictive performances between the GRBF network with the linear terms (○) and the linear AR predictor (●) for the long time series.

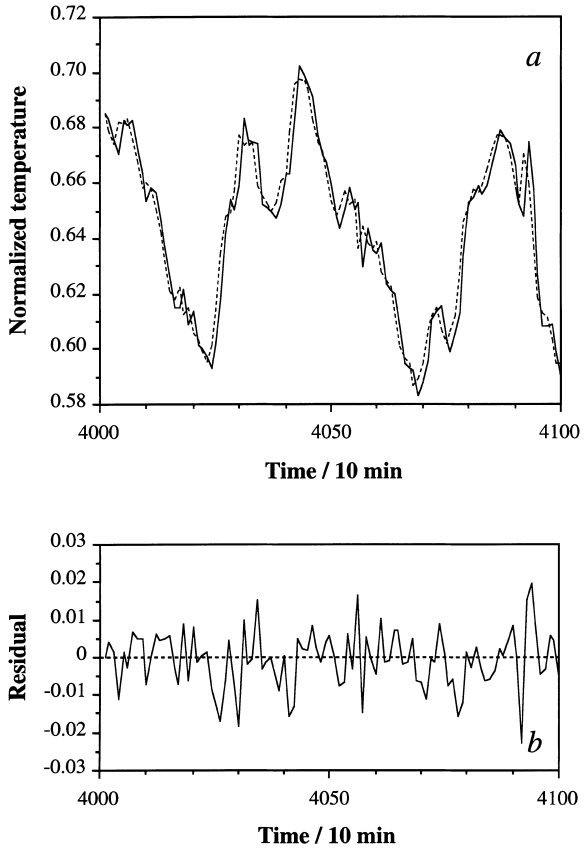
a
Normalized temperature
0.72
0.70
0.68
0.66
0.64
0.62
0.60
0.58
4000
4050
4100
Time / 10 min
b
Residual
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
4000
4050
4100
Time / 10 min

Fig. 12.

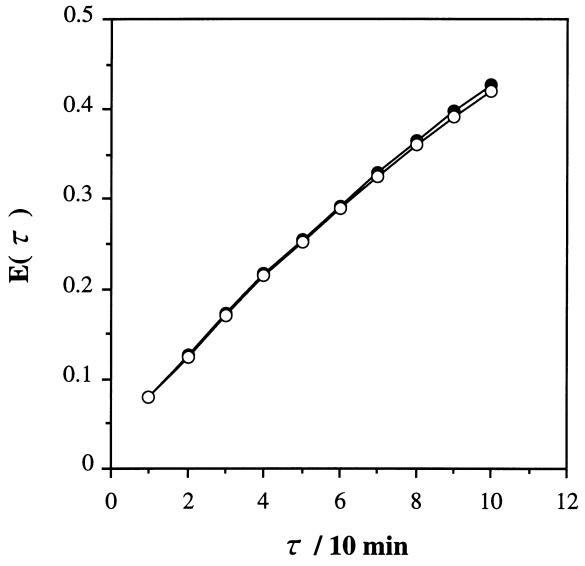
E(τ)
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
τ / 10 min

Fig. 13.

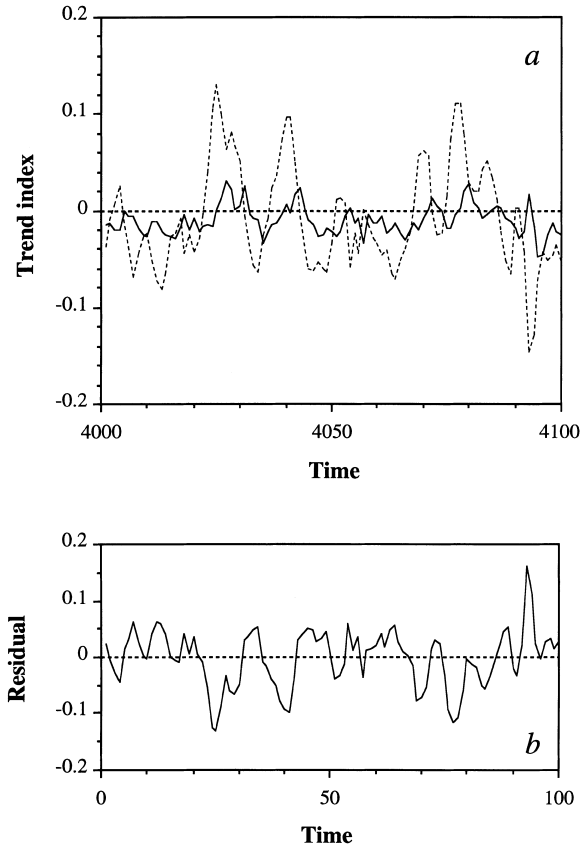

Fig. 14. Future trend forecasts by the predictive device of Eq. (69) about the long time series ($\tau_{\max} = 3$). The scaling exponent is $H = 0.58$. The GRBF network with the linear terms is used iteratively to make forecasts $\tau = 1$–3 time-steps into the future. Solid curve and dashed curve indicates predicted trends and the corresponding actual trends, respectively. The actual trends are calculated with the reliability factor $w(\tau) = 1$.

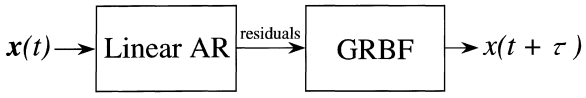

Fig. 15. Schematic diagram of a GRBF network connected to a linear filter in series.

basis functions [21,22]. Nevertheless, the predictive performance was improved by adding the linear terms. Note also that the number of hidden nodes was saved by adding the linear terms. The linear terms seem to help reconstruct a hyperplane that represents the behavior of observational noise included in the data. In fact, actual time series data are contaminated with external noise such as the measurement error. Such observational noise is considered to form no characteristic structure in the phase space. In contrast, a chaotic attractor generally has a global characteristic structure that a GRBF network can approximate but a globally linear predictor cannot. This may be the case with the short time series. This conjecture seems to be consistent with the predictions about the long time series. The long series is considered to include a considerable amount of stochastic ingredients that may have masked the chaotic and nonlinear feature. This may cause little difference in the predictive performances with increasing τ between the GRBF and the linear predictors for the long time series. One might raise a question as to why the linear predictor had not been used to filter out the random noise and why the residuals consisting (mainly) of chaotic (nonlinear)

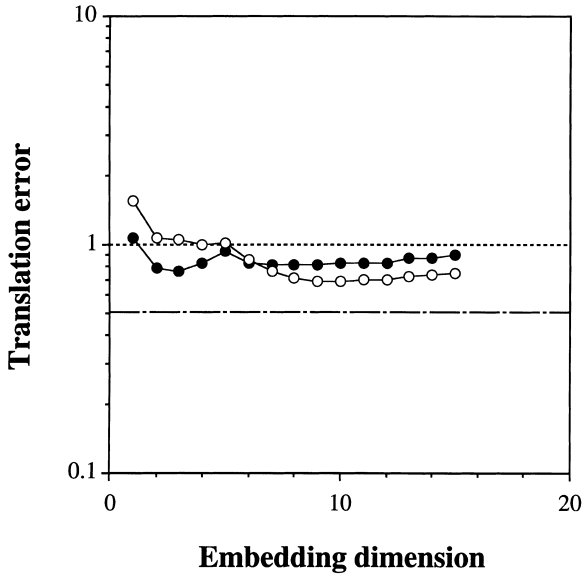

Fig. 16. The average over the medians of E_{trans} as a function of D about the residuals of the linear AR predictor for the short time series (●) and the long time series (○). The phase space is generated with $\Delta t = 10$ min. The medians are estimated for 10 (20) sets of 100 (500) randomly chosen vectors with 4 nearest neighbors and the associated images generated by time translation with $T = 5$ for the short (long) time series. The dash–dotted line indicates the upper bound of E_{trans} for visible determinism.

ingredients had not been input to the GRBF part as shown in Fig. 15. Such filtering is harmful, however, in that it causes a significant loss of determinism from the data. Fig. 16 illustrates results of the Wayland test on the residuals of the linear AR predictor for the short and long time series. In each case, E_{trans} exceeds the upper bound for visible determinism. The residuals have no deterministic feature that could be captured by function approximation.

7. Conclusion

It seems to be a good strategy to make combinational use of distinct diagnostic methods to characterize complexities in observational time series. The diagnostic test based on nonlinear forecasting is capable of capturing visible determinism as a function of the embedding dimension and the prediction-time interval. Scaling exponents can be estimated from the scaling properties of the prediction error. Misdiagnosis due to noise contamination of data is avoidable by estimating the degrees of parallelness of trajectories in phase space. The estimates can be a measure indicating how predictable the time series is. Analysis of deterministic features helps design an appropriate architecture of a predictive model. The diagnostic test for stationary guarantees that library examples to be used for optimizing the predictor reflect general dynamical behavior of the system. The strategy was successfully applied to the temperature fluctuations observed in a blast furnace to suggest possible existence of chaotic dynamics. Gaussian GRBF networks with linear algebraic polynomials are appropriate for forecasting chaotic time series contaminated with observational noise. The linear terms seem to successfully capture the dynamical behavior of the additive noise. A virtue of this networks architecture is that the contribution of the GRBF parameters is reduced automatically during learning when nonlinearity is invisible in the library examples. Such an effect has been also observed for GRBF networks consisting of another type of bell-shaped radial basis functions such as $(1+\cosh x)^{-1}$ [28]. Bleaching a noisy time series with a linear AR predictor is harmful in that it induces a substantial loss of determinism from the data.

Acknowledgements

The authors would appreciate Prof. Yasunori Okabe of the University of Tokyo for his guidance on the theory of the KM_2O–Langevin equations and many helpful discussions, and Masahiro Kashiwada for technical support.

References

[1] A.S. Lapedes, R. Farber, Nonlinear signal processing using neural networks: prediction and system modeling, Technical Report LA-UR-87-2662 (1987), Los Alamos National Laboratory.
[2] M. Casdagli, Nonlinear prediction of chaotic time series, Physica D 35 (1989) 335.
[3] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractor, Phys. Rev. Lett. 31 (1983) 346.
[4] D. Ruelle, Deterministic chaos; the science and the fiction, Proc. Roy. Soc. London A 427 (1990) 241.
[5] A.R. Osborne, A. Provenzale, A search for chaotic behavior in large and mesoscale motions in the pacific ocean, Physica D 35 (1989) 357.
[6] J.D. Farmer, J.J. Sidorowich, Predicting chaotic time series, Phys. Rev. Lett. 59 (1987) 845.
[7] G. Sugihara, R.M. May, Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series, Nature 344 (1990) 734.
[8] A.A. Tsonis, J.B. Elsner, Nonlinear prediction as a way of distinguishing chaos from random fractal sequences, Nature 358 (1992) 217.
[9] T. Miyano, Time series analysis of complex dynamical behavior contaminated with observational noise, Int. J. Bifur. Chaos 6 (1996) 2031.
[10] D.T. Kaplan, L. Glass, Coarse-grained embeddings of time series: random walks, Gaussian random processes, and deterministic chaos, Physica D 64 (1993) 431.
[11] R. Wayland, D. Bromley, D. Pickett, A. Passamante, Recognizing determinism in a time series, Phys. Rev. Lett. 70 (1993) 580.
[12] Y. Okabe, On Stochastic Difference Equations for the Multi-Dimensional Weakly Stationary Time Series, Prospect of Algebraic Analysis, Academic Press, New York, 1988, p. 601.
[13] Y. Okabe, Application of the theory of KM_2O–Langevin equations to the linear prediction problem for the multi-dimensional weakly stationary time series, J. Math. Soc. Jpn. 45 (1993) 277.
[14] Y. Okabe, A new algorithm derived from the view point of the fluctuation–dissipation principle in the theory of KM_2O–Langevin equations, Hokkaido Math. J. 22 (1993) 199.
[15] Y. Okabe, Y. Nakano, The theory of KM_2O–Langevin equations and its applications to data analysis (I): Stationary analysis, Hokkaido Math. J. 20 (1991) 45.
[16] Y. Okabe, A. Inoue, The theory of KM_2O–Langevin equations and its applications to data analysis (II): causal analysis, Nagoya Math. J. 134 (1994) 1.
[17] Y. Okabe, T. Ootsuka, The theory of KM_2O–Langevin equations and its applications to the nonlinear prediction problem for the one-dimensional strictly stationary time series, J. Math. Soc. Jpn. 47 (1995) 349.
[18] S. Kimoto, T. Ikeguchi, T. Matozaki, K. Aihara, In: M. Yamaguti (Ed.), Deterministic Chaos and its Stationary Analysis, Towards the Harnessing of Chaos, Elsevier, Amsterdam, 1994, p. 373.
[19] M. Casdagli, D. des Jardins, S. Eubank, J.D. Farmer, J. Gibson, N. Hunter, J. Theiler, Nonlinear modeling of chaotic time series; theory and applications, Technical Report LA-UR-91-1637, Los Alamos National Laboratory, 1991.
[20] H. Akaike, T. Nakagawa, Statistical Analysis and Control for a Dynamical System, Science Company, 1973, in Japanese.
[21] T. Poggio, F. Girosi, Networks for approximation and learning, Proc. IEEE 78 (1990) 1481.
[22] F. Girosi, M. Jones, T. Poggio, Regularization theory and neural networks architectures, Neural Computation 7 (1995) 219.
[23] G. Wahba, Spline Models for Observational Data, Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, 1990.
[24] T. Miyano, F. Girosi, Forecasting global temperature variations by neural networks, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, A. I. Memo No.1447 (1994).
[25] T. Miyano, H. Morita, A. Shintani, T. Kanda, M. Hourai, Characterization of complexities in Czochralski crystal growth by nonlinear forecasting, J. Appl. Phys. 76 (1994) 2681.
[26] J.H. Friedman, W. Stuetzle, Projection pursuit regression, J. Am. Statist. Assoc. 76 (1981) 817.
[27] P. Niyogi, F. Girosi, On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, A. I. Memo, No.1467 (1994).
[28] T. Miyano, K. Aihara, Forecasting complex time series by bell-shaped regularization networks, Proceedings of International Symposium on Artificial Life and Robotics (AROB'97), 1997, p. 150.