Neurocomputing 82 (2012) 186–195

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Prediction for noisy nonlinear time series by echo state network based on dual estimation

Chunyang Sheng, Jun Zhao*, Ying Liu, Wei Wang

School of Control Science and Engineering, Dalian University of Technology, China

ARTICLE INFO

Article history:
Received 29 July 2011
Received in revised form
3 November 2011
Accepted 5 November 2011
Communicated by Z. Wang
Available online 3 January 2012

Keywords:
Echo state network
Dual estimation
Kalman filter
Time series
Prediction

ABSTRACT

When using echo state networks (ESNs) to establish a regression model for noisy nonlinear time series, only the output uncertainty was usually concerned in some literature. However, the unconsidered internal states uncertainty is actually important as well. In this study, an improved ESN model with noise addition is proposed, in which the additive noises describe the internal state uncertainty and the output uncertainty. In terms of the parameters determination of this prediction model, a nonlinear/linear dual estimation consisting of a nonlinear Kalman filter and a linear one is proposed to perform the supervised learning. For verifying the effectiveness of the proposed method, the noisy Mackey Glass time series and the generation flow of blast furnace gas (BFG) in steel industry practice are both employed. The experimental results demonstrate that the proposed method is effective and robust for noisy nonlinear time series prediction.

1. Introduction

Theoretically, recurrent neural networks are universal approximators, and as such, have an excellent ability of approximating any nonlinear mapping to any degree of accuracy [1]. Echo state network, a kind of recurrent neural network, exhibits good performance for the prediction of nonlinear or non-Gaussian dynamic system [2]. The dynamic reservoir of ESN, instead of the hidden layer of generic neural network, involves a large number of sparsely connected neurons that shows a sound memory characteristic; furthermore, only the output connections need to be determined during the learning process, which simplifies the establishment of the network. Recently, such network had been successfully applied to time series prediction [3,4], short-term load forecasting [5], signal processing [6,7] and automatic control [8].

On the other hand, ESNs fundamentally suffer from two basic limitations despite its well popularity. First, the ill-conditioned solution associated with linear regression or recursive least squares method could hardly be avoided in the learning process, and such solution might deviate from the real system. As mentioned in [9], the large output weights impaired the generalization capability of the model that means the model input slightly deviated from the training data such that the relatively poor results would occur. In addition, the model with large weights might lead to the lack of stability when ESN features the output feedback. The second issue is the unsatisfactory low performance when the sample uncertainties involved. Since the core of ESN lies in the dynamic reservoir driven by the external input, the update of internal states often accompany with the uncertainty from input. Thus, an inaccurate target value might be obtained. In practice, the uncertainty in real-world dynamic system is fairly prevalent; for example, consider the situation where the various noises are often introduced into an industrial system by sensors. The low accuracy of ESNs in industrial time series prediction was addressed in [10], and the similar demonstrations had been reported in other fields such as sea clutter prediction [11], nonlinear system modeling [12], and nonlinear filtering [13].

To avoid the ill-condition, the eigenvalue spread of the correlation matrix of reservoir activation signals was proposed in [9], where the ill-condition phenomenon was prevented by adding noise to the reservoir during training so as to promote the stability and robustness of the network. However, this method failed to avoid the ill-condition thoroughly and the additive noise might impair the prediction accuracy. Subsequently, the singular value decomposition [10], the biologically motivated learning method [12] and the swarm intelligence optimization [13] were adopted to train the network. All of these methods avoided the ill-condition and improved the accuracy for noiseless time series more or less, but as for the noisy nonlinear series, the quality of these methods had not been proved or demonstrated. In [10] presented, the empirical mode decomposition was used for noise reduction; yet, the noise-reduced sample still comprised of the uncertainties. In [14] reported, the Bayesian method that focused on the distribution of output weights was

* Corresponding author.
E-mail address: zhaoj@dlut.edu.cn (J. Zhao).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.11.021

employed to perform the network training. Although the prediction accuracy with the Bayesian based ESN was improved thanks to the considered output uncertainty, the uncertainty of the internal states resulted from the intrinsic sample was ignored. In addition, a kind of decoupled ESN based on the thought of multiple reservoirs was proposed in [11] and [15], in which those methods were somewhat suitable for complex problems, but the internal states uncertainty was also not mentioned.

In this study, an improved ESN model with noise addition is proposed to predict the noisy nonlinear time series, in which the uncertainties from internal states and outputs are meanwhile considered in accordance with the industrial practice. For the optimal model parameters, a nonlinear/linear dual estimation method is designed, in which a nonlinear Kalman filter serves to estimate the internal states, and a linear one estimates the output weights. The contribution of this paper lies in the following three aspects. First, the ill-conditioned solution could be avoided by using the dual estimation for the parameters determination. Second, the implementation process of the proposed method is much easier than that of the classical dual estimation based RNN. Finally, due to the consideration of internal states uncertainty, the accuracy and the robustness of the model are greatly enhanced. To demonstrate the accuracy and the robustness of the proposed method, the standard and noisy Mackey-Glass time series are first employed as the testing examples. And, the proposed method is further adopted to predict the generation flow of blast furnace gas in steel industry. The experimental results indicate the proposed method presents a good performance for the prediction of noisy industrial time series.

The rest of the paper is organized as follows. Section 2 presents an improved ESN model with additive noise for prediction and elaborates how to determine the parameters of this model. In Section 3, several types of nonlinear Kalman filters are introduced to estimate the internal states of the established model. And, the corresponding output weights estimation is described in Section 4. Section 5 carries out the two classes of simulation experiments to verify the effectiveness of the proposed method. Finally, we summarize the paper and give the future work in Section 6.

2. Improved ESN based on dual estimation

We review the standard form of ESN first in this section and present an improved version considering the noises involvement.

2.1. Improved ESN with noise addition

Echo state network contains the input layer, the reservoir and the output layer. Many sparsely connected neurons in the reservoir guarantee the echo property of the network [2]. The classical recursive formula of ESN reads as

$$\mathbf{x}_k = f(\mathbf{W}^{in}\mathbf{u}_k + \mathbf{W}\mathbf{x}_{k-1} + \mathbf{W}^{back}\mathbf{y}_{k-1}) \tag{1}$$

$$\mathbf{y}_k = f^{out}(\mathbf{W}^{out}(\mathbf{u}_k, \mathbf{x}_k, \mathbf{y}_{k-1})) \tag{2}$$

where $\mathbf{u}_k$ is the exogenous input; the dimension of input is equal to m. $\mathbf{x}_k$ is the internal states, the dimensionality of $\mathbf{x}_k$ is N. $\mathbf{y}_k$ is the output, the dimensionality of $\mathbf{y}_k$ is L. $\mathbf{W}^{in} = (W^{in}_{i,j}) \in \mathbb{R}^{N\times m}$ denotes the input weights; $\mathbf{W} = (W_{i,j}) \in \mathbb{R}^{N\times N}$ denotes the internal weights of the neurons in reservoir. To provide proper memorization capabilities, $\mathbf{W}$ should be sparse whose connectivity level ranges from 1% to 5% and its spectral radius should be less than 1; and $\mathbf{W}^{out} = (W^{out}_{i,j}) \in \mathbb{R}^{L\times(m+N+L)}$ denotes the output weights. f and f^{out} are the activation functions of internal neurons and output neurons, respectively.

When using echo state network to establish a regression model for noisy time series, the output uncertainty used to be considered in the existing literature. In general, the independent Gaussian noise sequences reflecting the difference between the observation and the expected output are introduced into the output formula (2). However, as for as noisy time series, the internal states are still uncertain. In this study, an improved ESN considering the noise addition becomes

$$\mathbf{x}_k = f(\mathbf{W}^{in}\mathbf{u}_k + \mathbf{W}\mathbf{x}_{k-1}) + v_{k-1} \tag{3}$$

$$y_k = \mathbf{W}^{out} \cdot [\mathbf{u}_k, \mathbf{x}_k] + n_k \tag{4}$$

where y_k is a scalar quantity that shows the network is a single-output model, $v_{k-1} \in \mathbb{R}^{N\times 1}$ and n_k are independent white Gaussian noise sequences. And, $\forall k$, $E[v_{k-1}] = E[n_k] = \mathbf{0}$; $\forall i,j$, $E[v_i v_j^T] = \mathbf{R}^v$, $E[n_i n_j^T] = \sigma_n^2$. In (3), $v_{(k-1)}$ reflects the uncertainty of internal states, and n_k reflects the output uncertainty in (4).

2.2. The dual estimation

For the established model above, both the internal states and the output weights are unknown. Hence, it is a tough task for traditional learning methods, such as linear regression or recursive least squares method, to estimate the internal states and output weights

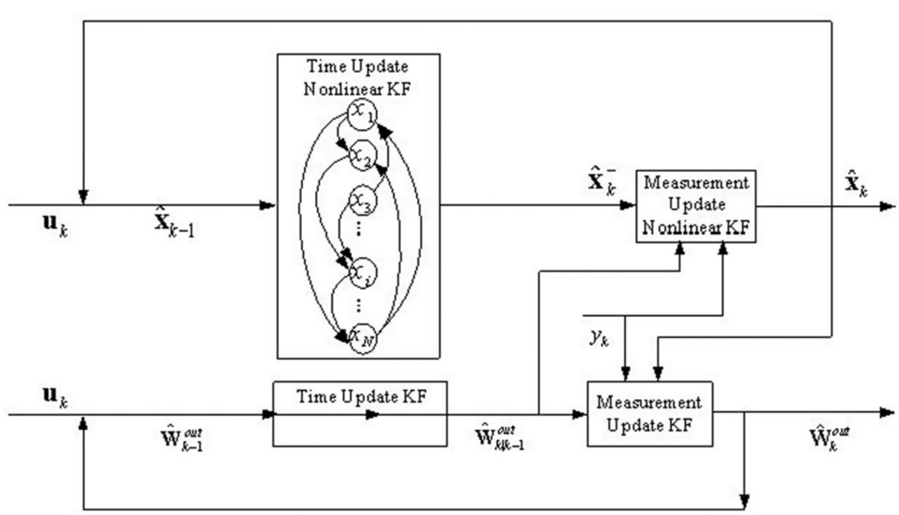

Fig. 1. The proposed dual estimation architecture.

simultaneously. The dual estimation works by alternating between the model estimation and the process state estimation, whose main advantage lies on the model can be estimated with the unknown process state [16]. Making full use of this advantage, a nonlinear/linear dual estimation is proposed in this study to estimate the internal states and the output weights of the established ESN. Fig. 1 describes the brief structure of the proposed dual estimation method. The nonlinear/linear dual estimation is regarded in nature as an optimization algorithm that recursively determines the internal states $\hat{\mathbf{x}}_k$ and output weights $\hat{\mathbf{W}}_k^{out}$ via minimizing the cost function, a joint binary function of $\hat{\mathbf{x}}_k$ and $\hat{\mathbf{W}}_k^{out}$ given by

$$J(\mathbf{x}_k,\mathbf{W}_k^{out}) = \sum_{k=1}^{n}\{(\sigma_n^2)^{-1}(y_k-\mathbf{W}_k^{out}\cdot[\mathbf{u}_k,\mathbf{x}_k])^T(y_k-\mathbf{W}_k^{out}\cdot[\mathbf{u}_k,\mathbf{x}_k]) + (\mathbf{x}_k-\hat{\mathbf{x}}_{k|k-1})^T(\mathbf{R}^v)^{-1}(\mathbf{x}_k-\hat{\mathbf{x}}_{k|k-1})\} \tag{5}$$

where $\hat{\mathbf{x}}_{k|k-1} \stackrel{\Delta}{=} f(\hat{\mathbf{x}}_{k-1},\mathbf{u}_k,\mathbf{W}^{in},\mathbf{W})$. It is obvious that the priori estimation of internal states is the function of input weights $\mathbf{W}^{in}$ and internal weights $\mathbf{W}$ that are determined in advance and fixed in the process of estimation. To minimize the cost function with respect to $\hat{\mathbf{x}}_k$ and $\hat{\mathbf{W}}_k^{out}$, the partial derivatives of (5) are calculated as

$$\frac{\partial J(\mathbf{x}_k,\mathbf{W}_k^{out})}{\partial \mathbf{W}_k^{out}} = \sum_{k=1}^{n}\{2(\sigma_n^2)^{-1}(y_k-\mathbf{W}_k^{out}[\mathbf{u}_k,\mathbf{x}_k])\}\cdot(-[\mathbf{u}_k,\mathbf{x}_k]) \tag{6}$$

$$\frac{\partial J(\mathbf{x}_k,\mathbf{W}_k^{out})}{\partial \mathbf{x}_k} = \sum_{k=1}^{n}\{2(\sigma_n^2)^{-1}(y_k-\mathbf{W}_k^{out}[\mathbf{u}_k,\mathbf{x}_k])\cdot\mathbf{W}_k^{out,\mathbf{x}}+\mathbf{Q}(\mathbf{x}_k)\} \tag{7}$$

$$\mathbf{Q}(\mathbf{x}_k) = (\mathbf{R}^v)^{-1}(\mathbf{x}_k-\hat{\mathbf{x}}_{k|k-1})+((\mathbf{R}^v)^{-1})^T(\mathbf{x}_k-\hat{\mathbf{x}}_{k|k-1}) \tag{8}$$

where $\mathbf{W}_k^{out,\mathbf{x}}$ is a partitioned matrix of $\mathbf{W}_k^{out}$. Since the matrix $\mathbf{R}^v$ is diagonal, the inverse matrix of $\mathbf{R}^v$ equals to the transpose of the inverse of $\mathbf{R}^v$, i.e. $(\mathbf{R}^v)^{-1} = ((\mathbf{R}^v)^{-1})^T$. Then the formula (8) can be described as

$$\mathbf{Q}(\mathbf{x}_k) = 2(\mathbf{R}^v)^{-1}(\mathbf{x}_k-\hat{\mathbf{x}}_{k|k-1}) \tag{9}$$

If the minimum of $J(\mathbf{x}_k,\mathbf{W}_k^{out})$ exists, Eqs. (6) and (8) should be equal to zero. Considering $[\mathbf{u}_k,\mathbf{x}_k] \neq \mathbf{0}$, we have

$$y_k-\mathbf{W}_k^{out}[\mathbf{u}_k,\mathbf{x}_k] = \mathbf{0} \tag{10}$$

$$\mathbf{x}_k = \hat{\mathbf{x}}_{k|k-1} \tag{11}$$

(10) and (11) show that the estimation for the internal states is unique as long as the optimum of (5) exists, and the uniqueness of output weights $\hat{\mathbf{W}}_k^{out}$ depends on the positive-definiteness of matrix $[\mathbf{u}_k,\mathbf{x}_k]$. In addition, (10) also explains that the linear regression is a limited method for the parameters determination of ESN due to its ill-condition, so the Kalman filter is employed in this study to estimate the output weights. Based on the above analysis, the nonlinear/linear dual estimation proposed in this study consists of two parts: the internal states estimation with a nonlinear Kalman filters and the output weights estimation.

3. Internal states estimation with nonlinear Kalman filters

Many of nonlinear Kalman filters had been studied in the related fields, in which three of the most typical ones are the extended Kalman filter (EKF) [31], the unscented Kalman filter (UKF) [20] and the cubature Kalman filter (CKF) [22]. In this section, the nonlinear Kalman filters are respectively analyzed for the internal states estimation structure.

3.1. Extended Kalman filter

EKF is prevalent due to its simplicity and practicality, which results from the linearization of the nonlinear functions based on the first order Taylor series expansion. When estimating the internal states with EKF, Eq. (3) has to be linearized as

$$\mathbf{x}_k \approx \tilde{\mathbf{x}}_k + \mathbf{F}_{\mathbf{x}}(\hat{\mathbf{x}}_{k-1},\mathbf{u}_k)(\mathbf{x}_{k-1}-\hat{\mathbf{x}}_{k-1}) \tag{12}$$

where $\hat{\mathbf{x}}_{k-1}$ is the posterior estimation in time step $k-1$. $\tilde{\mathbf{x}}_k$ can be obtained from (13). $\mathbf{F}_{\mathbf{x}}(\hat{\mathbf{x}}_{k-1},\mathbf{u}_k)$ is the Jacob matrix of the partial derivative of f with respect to $\mathbf{x}$, i.e.,

$$\tilde{\mathbf{x}}_k = f(\mathbf{W}^{in}\mathbf{u}_k+\mathbf{W}\hat{x}_{k-1}) \tag{13}$$

$$[\mathbf{F}_{\mathbf{x}}(\hat{\mathbf{x}}_{k-1},\mathbf{u}_k)]_{i,j} = \left.\frac{\partial f_i(\mathbf{x},\mathbf{u}_k)}{\partial \mathbf{x}_j}\right|_{\mathbf{x}=\hat{\mathbf{x}}_{k-1}} \tag{14}$$

To facilitate the calculation, (4) can be rewritten as

$$y_k = \mathbf{W}_{k|k-1}^{out,\mathbf{u}}\mathbf{u}_k+\mathbf{W}_{k|k-1}^{out,\mathbf{x}}\mathbf{x}_k+n_k \tag{15}$$

As such, the nature of EKF can be viewed as a form of feedback control. And, the equations for the EKF are divided into two parts, time update equations and measurement update equations. The time update equations are responsible for obtaining the priori estimates of the next time step, in which the priori $\hat{\mathbf{x}}_{k|k-1} = \tilde{\mathbf{x}}_k$ with the covariance $\mathbf{P}_{\mathbf{x}_{k|k-1}}$. The measurement update equations are responsible for the posterior estimates through correcting the priori based on the current noisy measurement y_k.

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k|k-1}+\mathcal{K}_k(y_k-\hat{\mathbf{W}}_{k|k-1}^{out}\cdot[\mathbf{u}_k,\hat{\mathbf{x}}_{k|k-1}]) \tag{16}$$

where $\mathcal{K}_k$ is the Kalman gain, $\mathcal{K}_k = \mathbf{P}_{\mathbf{x}_{k|k-1}}\mathbf{H}_k^T(\mathbf{H}_k\mathbf{P}_{\mathbf{x}_{k|k-1}}\mathbf{H}_k^T+\sigma_n^2)^{-1}$, and $\mathbf{H}_k = \mathbf{W}_{k|k-1}^{out,\mathbf{x}}$.

The first order EKF may be biased in ‘highly’ nonlinear system, which is the obvious disadvantage [17]. As a solution to mitigate the biased error, the second order EKF and the central difference Kalman filter were proposed in [18,19], respectively. However, these filters were limited to the differentiable functions.

3.2. Unscented Kalman filter

Based on the unscented transformation, UKF is not necessarily limited to the usage of differentiable functions. Using UKF, the state random variable has to be redefined as the concatenation of internal states and noise variables, $\mathbf{x}^a = [\mathbf{x}^T v^T n^T]^T$, namely an augmented variable, whose covariance $\mathbf{P}^a$ can be represented as $\mathrm{diag}(\mathbf{P}_{\mathbf{x}},\mathbf{R}^v,\sigma_n^2)$. Referencing the report of [20], the intuition of UKF was that it is easier to approximate a probability distribution compared to the estimation of arbitrary nonlinear function or transformation. Following this intuition, a set of sigma points is generated.

$$\chi_{k-1}^a = [\hat{\mathbf{x}}_{k-1}^a\ \hat{\mathbf{x}}_{k-1}^a+\sigma_{\mathbf{x}}\ \hat{\mathbf{x}}_{k-1}^a-\sigma_{\mathbf{x}}] \tag{17}$$

where $\sigma_{\mathbf{x}} = \sqrt{(n_{\mathbf{x}}^a+\lambda)\mathbf{P}_{k-1}^a}$, $n_{\mathbf{x}}^a$ is the dimensionality of $\mathbf{x}^a$ that is the sum of the dimensionality of $\mathbf{x}$, v and n. In this way, the mean and the covariance of these points equal to those of the augmented variable, denoted as $\hat{\mathbf{x}}_{k-1}^a$ and $\mathbf{P}_{k-1}^a$ respectively. Each sigma point is instantiated through the nonlinear function f to yield a set of alternatives.

$$\chi_{k|k-1}^{\mathbf{x}} = f(\mathbf{W}^{in}\mathbf{u}_k+\mathbf{W}\chi_{k-1}^{\mathbf{x}})+\chi_{k-1}^{v} \tag{18}$$

Then, the prior mean and covariance of internal states exhibit as

$$\hat{\mathbf{x}}_{k|k-1} = \frac{1}{2(n_{\mathbf{x}}^a+\lambda)}\sum_{i=1}^{2n_{\mathbf{x}}^a}\chi_{i,k|k-1}^{\mathbf{x}} \tag{19}$$

$$\mathbf{P}_{\mathbf{x}_{k|k-1}} = \frac{1}{2(n_{\mathbf{x}}^a+\lambda)}\sum_{i=1}^{2n_{\mathbf{x}}^a}[\chi_{i,k|k-1}^{\mathbf{x}} - \hat{\mathbf{x}}_{k|k-1}][\chi_{i,k|k-1}^{\mathbf{x}} - \hat{\mathbf{x}}_{k|k-1}]^T \tag{20}$$

Due to the linearity of the output equation, the expected output $\hat{\mathbf{y}}_{k|k-1} = \mathbf{W}_{k|k-1}^{out} \cdot [\mathbf{u}_k, \hat{\mathbf{x}}_{k|k-1}]$. Then the measurement update of UKF can be implemented similar as that of EKF. In such a way, the computational complexity of UKF can be much more reduced without unscented transformation in the measurement update. With respect to the accuracy, UKF is usually superior to that of EKF, but the computational efficiency is far from satisfactory [21]. In addition, to generate a set of sigma points, the Cholesky decomposition needs to compute the square-root covariance matrix. However, the UKF-computed covariance matrix cannot always be positive definite. Thus, the filter process might be instable.

3.3. Cubature Kalman filter

CKF is a novel approximate solution to cope with the drawbacks that the existing filters suffered from [22]. The experimental results showed that CKF was a more effective method for training RNN [23], in which the internal states could be estimated in two update steps. First, the priori statistic features of internal states can be computed by using the priori distribution and the historical internal states, i.e.

$$\hat{\mathbf{x}}_{k|k-1} = \int_{\mathbb{R}^{n_{\mathbf{x}}}} f(\mathbf{W}^{in}\mathbf{u}_k + \mathbf{W}\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}; \hat{\mathbf{x}}_{k-1}, \mathbf{P}_{\mathbf{x}_{k-1}}) d\mathbf{x}_{k-1} \tag{21}$$

$$\mathbf{P}_{\mathbf{x}_{k|k-1}} = \int_{\mathbb{R}^{n_{\mathbf{x}}}} f(\mathbf{W}^{in}\mathbf{u}_k + \mathbf{W}\mathbf{x}_{k-1}) f^T(\mathbf{W}^{in}\mathbf{u}_k + \mathbf{W}\mathbf{x}_{k-1}) \times p(\mathbf{x}_{k-1}; \hat{\mathbf{x}}_{k-1}, \mathbf{P}_{\mathbf{x}_{k-1}}) d\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k|k-1}(\hat{\mathbf{x}}_{k|k-1})^T + \mathbf{R}^v \tag{22}$$

where $p(\mathbf{x}_{k-1}; \hat{\mathbf{x}}_{k-1}, \mathbf{P}_{\mathbf{x}_{k-1}})$ is the probability density of $\mathbf{x}_{k-1}$. Second, the corresponding posterior estimation is computed by the current noisy measurement y_k.

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k|k-1} + \mathcal{K}_k(y_k - \hat{y}_{k|k-1}) \tag{23}$$

where $\hat{y}_{k|k-1}$ is the expected output of ESN calculated by $\hat{y}_{k|k-1} = \hat{\mathbf{W}}_{k|k-1}^{out} \cdot [\mathbf{u}_k, \hat{\mathbf{x}}_{k|k-1}]$.

The key of CKF is to compute the integrals in (21) and (22). Here, we assume the probability density in the formulas satisfy the Gaussian density $\mathcal{N}(\mathbf{x}_{k-1}; \hat{\mathbf{x}}_{k-1}, \mathbf{P}_{\mathbf{x}_{k-1}})$, i.e., to compute its mean and covariance. Then, the Gaussian weighted integrals can be computed by a third-degree cubature rule [23]. For instance, approximated by a Gauss–Laguerre formula, we have

$$\int_{\mathbb{R}^{n_{\mathbf{x}}}} f(\mathbf{x}) \mathcal{N}(\mathbf{x}; \mu, \Sigma) d\mathbf{x} \approx \frac{1}{2n_{\mathbf{x}}}\sum_{i=1}^{2n_{\mathbf{x}}} f(\mu + \sqrt{\Sigma}\xi_i) \tag{24}$$

where a square-root factor of the covariance Σ satisfies $\Sigma = \sqrt{\Sigma}\sqrt{\Sigma}^T$, and the set of $2n_{\mathbf{x}}$ cubature points are given by

$$\xi_i = \begin{cases} \sqrt{n_{\mathbf{x}}}\mathbf{e}_i, & i = 1, 2, \ldots n_{\mathbf{x}} \\ -\sqrt{n_{\mathbf{x}}}\mathbf{e}_{i-n_{\mathbf{x}}}, & i = n_{\mathbf{x}}+1, n_{\mathbf{x}}+2, \ldots 2n_{\mathbf{x}} \end{cases} \tag{25}$$

where $\mathbf{e}_i \in \mathbb{R}^n$ denotes the ith elementary column vector.

Since the probability density of non-Gaussian process can be estimated by a Gaussian mixture density with a finite number of weighted sums of Gaussian densities $p(\mathbf{x}) \sim \sum_{i=1}^{n} a_i \mathcal{N}(\overline{\mathbf{x}}_i, \mathbf{P}_{\mathbf{x}_i})$, the CKF can be extended to non-Gaussian problems.

4. Weights estimation

Although filters are common to use for learning the parameters of neural network, most of the existing learning methods are based on nonlinear filters, such as EKF [24–26] and UKF [27]. This is mainly because the hidden states and the parameters are coupled for the feed-forward NNs or the RNNs. In [28], the gene regulatory network viewed as a stochastic dynamic model was established, whose parameters were identified by the Kalman filtering. Based on the issues, when no coupled relationship exists in the estimations of the process states and the parameters, the filter can also identify the parameters accurately. For the proposed ESN, the unknown parameters are only the output weights, so the Kalman filtering can be used to estimate the output weights. Thus, a separate state-space formulation for the underlying output weights is rewritten as

$$\mathbf{W}_k^{out} = \mathbf{W}_{k-1}^{out} + \mathbf{q}_{k-1} \tag{26}$$

$$y_k = \mathbf{W}_k^{out} \cdot [\mathbf{u}_k, \hat{\mathbf{x}}_k] + n_k \tag{27}$$

where the state transition matrix in (26) is simply an identity. Since the estimation of the output weights can be regarded as the feedback form of measurements, the equations of the KF (28)–(32) can be divided into two parts: time update equations (28) and (29) and measurement update equations (30)–(32). The time update equations use the current output weights and its covariance estimation to obtain a priori of output weights for the next time step.

$$\hat{\mathbf{W}}_{k|k-1}^{out} = \hat{\mathbf{W}}_{k-1}^{out} \tag{28}$$

$$\mathbf{P}_{\mathbf{W}_{k|k-1}^{out}} = \mathbf{P}_{\mathbf{W}_{k-1}^{out}} + \mathbf{R}^{\mathbf{q}} \tag{29}$$

The measurement update equations aim at the feedback which is incorporated into the priori for the improved posterior estimation. Another interpretation of the feedback is the difference between the current noisy measurement y_k and the expected output $\hat{y}_{k|k-1}$.

$$\hat{\mathbf{W}}_k^{out} = \hat{\mathbf{W}}_{k|k-1}^{out} + \mathcal{K}_k \cdot (y_k - \hat{y}_{k|k-1}) \tag{30}$$

According to (30), the Kalman gain $\mathcal{K}_k$ and the expected output $\hat{y}_{k|k-1}$ need to be determined as

$$\mathcal{K}_k = \mathbf{P}_{\mathbf{W}_{k|k-1}^{out}}[\mathbf{u}_k, \hat{\mathbf{x}}_k]([\mathbf{u}_k, \hat{\mathbf{x}}_k]^T \mathbf{P}_{\mathbf{W}_{k|k-1}^{out}}[\mathbf{u}_k, \hat{\mathbf{x}}_k] + \sigma_n^2)^{-1} \tag{31}$$

$$\hat{y}_{k|k-1} = \hat{\mathbf{W}}_{k|k-1}^{out} \cdot (\mathbf{u}_k, \hat{\mathbf{x}}_k) \tag{32}$$

The linear regression or recursive least squares method for solving the parameters of ESN could accompany with the ill-condition, which is mainly because the parameters identification is related to the inverse of the singular matrix. In contrast, Kalman filter has no need to solve the inverse of the singular matrix since the value of $([\mathbf{u}_k, \hat{\mathbf{x}}_k]^T \mathbf{P}_{\mathbf{W}_{k|k-1}^{out}}[\mathbf{u}_k, \hat{\mathbf{x}}_k] + \sigma_n^2)^{-1}$ is a scalar that can be directly obtained. And then the ill-condition can be effectively avoided. All in all, the Kalman filter can estimate the output weights of network accurately without ill-condition, and the estimation process becomes very simple and has a lower computational complexity.

5. Simulations

To verify the performance of the proposed ESN model, two categories of instances are presented in this section including the noise additive Mackey-Glass time series and the practical generation flow of blast furnace gas (BFG) in steel industry. Both of the two problems belong to the prediction of noisy nonlinear time series.

5.1. Mackey-Glass chaotic time series

Mackey-Glass system [29] is a time-delay differential system formulated as

$$\frac{dx(t)}{dt} = \frac{ax(t-\tau)}{1+x(t-\tau)^{10}} - bx(t) \tag{33}$$

where the parameters are typically set as $a=0.2$, $b=-0.1$ and $\tau=17$. To sample the time series, we numerically integrate by using the forth-order Runge–Kutta method with sampling period of 2-s and the initial condition $x(0)=1.2$. To verify the accuracy and robustness of the proposed model, a standard Mackey-Glass time series and a noisy one are employed. The noisy Mackey-Glass is generated by the standard one with additive white Gaussian noise with the variance 0.01. Thus, two sets of time series are obtained, each of which is with the identical form of $\{u(1), u(2), \ldots u(n), \ldots\}$, where $n \geq 2000$. The sample $u(k+\Delta)$ can be estimated from a properly chosen time series $\{u(k), u(k-\Delta), \ldots, u(k-(d_E-2)\Delta), u(k-(d_E-1)\Delta)\}$, where d_E and Δ denote the embedding dimensionality and the delay, respectively.

For the prediction of noisy time series, the input dimension of the improved ESN is empirically set to 50. The activation function of the internal neurons uses $\phi(v) = \tanh(v)$, while a linear activation is set on output. The internal states and output weights are initialized with zero-mean Gaussian with diagonal covariance of $0.005\,\mathbf{I_x}$ and $0.005\,\mathbf{I}_{\mathbf{W}^{out}}$, respectively. And, $\mathbf{W}^{in}$ and $\mathbf{W}$ are randomly generated. According to issues of [30], the spectral radius of $\mathbf{W}$ is set as 0.8 and its sparse connectivity equals to 2%.

To train the ESN model, we firstly construct a set of samples denoted as $\{(\mathbf{u}_i, y_i) | i=1,2,\ldots,1000\}$, where $[y_i] = [u(i+\Delta)]$, $d_E=50$, $\Delta=1$, $[\mathbf{u}_i] = [u(i-(d_E-1)\Delta)u(i-(d_E-2)\Delta)\cdots u(i)]$. Besides training, the rest of data are used to test. Here, we make $\mathbf{R}^v$ and $\mathbf{R^q}$ decay such that $\mathbf{R}_0^{\mathbf{v}} = (1/\lambda - 1)\mathbf{P}_{\mathbf{x}_0}$ and $\mathbf{R}_0^{\mathbf{q}} = (1/\lambda - 1)\mathbf{P}_{\mathbf{W}_0^{out}}$ with λ equals to 0.9995, and the value of σ_n^2 is determined by the intrinsic noise of the sample data.

To clarify the impact of the number of internal units on the prediction accuracy, a comparative experiment based on cross-validation is conducted, where the number of samples for cross-validation is 1000. Fig. 2 shows the cross-validation results of the nonlinear/linear dual estimation based ESN with different number of internal units. From this figure, when the number of

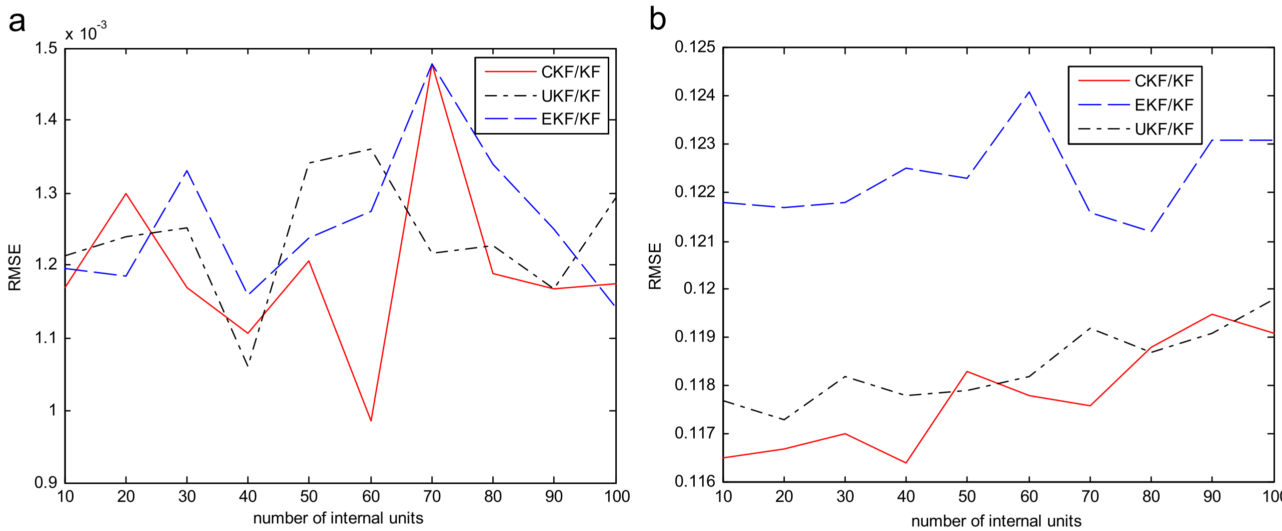

Fig. 2. Prediction accuracy of the dual estimation based ESN with different number of internal units. (a) Standard Mackey-Glass time series, (b) noisy Mackey-Glass time series.

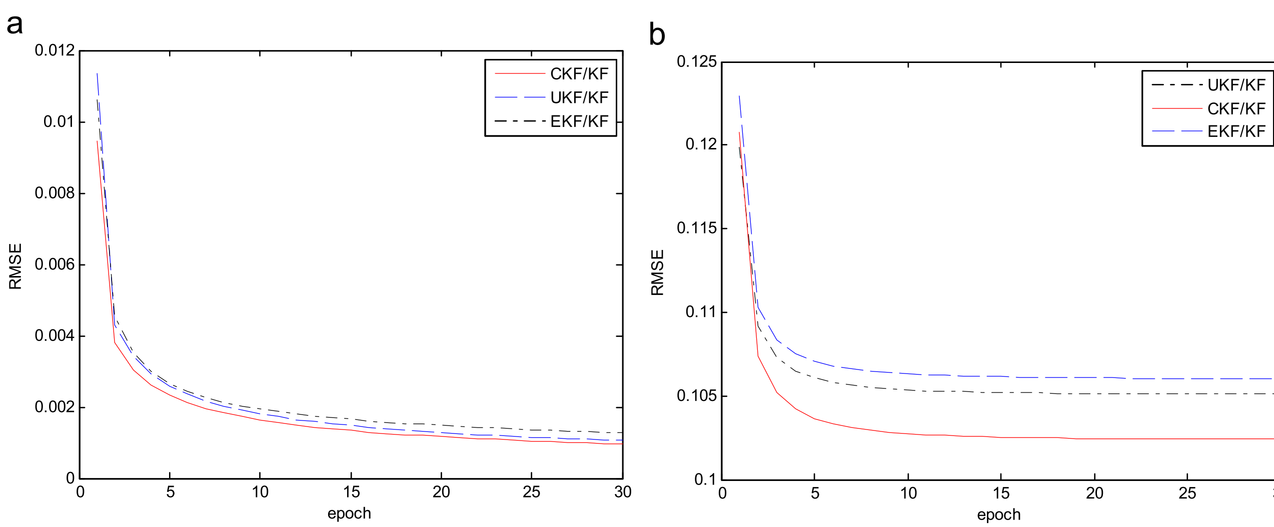

Fig. 3. The corresponding training curve of the proposed ESN. (a) Standard Mackey-Glass time series, (b) noisy Mackey-Glass time series.

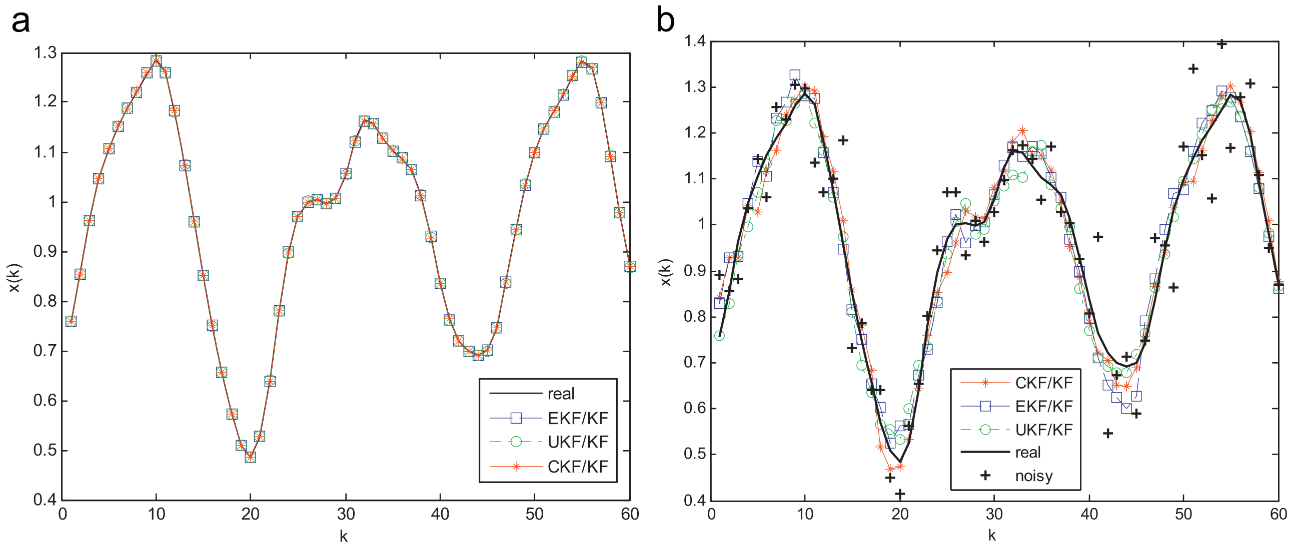

Fig. 4. Prediction results of the three nonlinear/linear dual estimation based ESN. (a) Standard Mackey-Glass time series, (b) noisy Mackey-Glass time series.

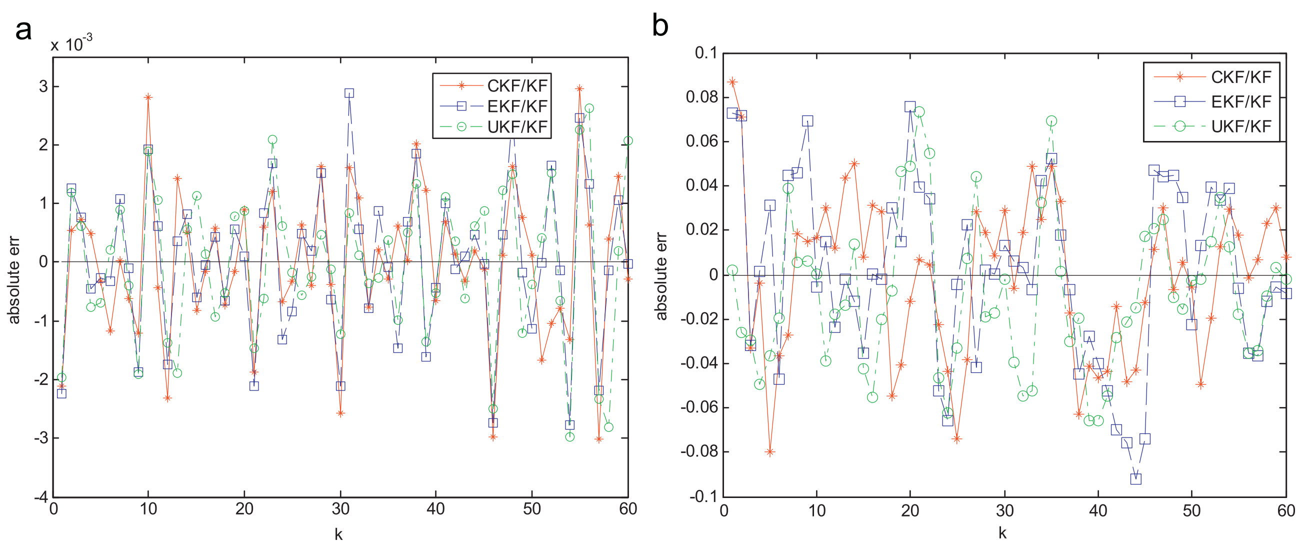

Fig. 5. Prediction error of the three nonlinear/linear dual estimation based ESN. (a) Standard Mackey-Glass time series, (b) noisy Mackey-Glass time series.

internal units is equal to 60 and 40, the improved ESN based on CKF/KF dual estimation presents the best performance for the standard and the noisy Mackey-Glass time series, respectively.

The training curves for the standard and the noisy Mackey-Glass time series are shown in Fig. 3, where the relationship between the training epoch and the training error evaluated by RMSE is illustrated. It can be seen that the nonlinear/linear dual estimations are convergent to train the improved ESN. A 60-min prediction curve is comparatively presented by the three dual estimation based networks in Fig. 4, and their absolute prediction errors are shown as Fig. 5. From Fig. 4, the proposed model is very effective for the standard Mackey-Glass with high accuracy and it can further play good performance for noisy time series. On the perspective of the average absolute error, the proposed CKF/KF dual estimation is the lowest for the noisy time series but not apparent for the standard ones. The corresponding quantified evaluation for the prediction performance can be listed in Table 1.

To further represent the importance of the state estimation to ESN, the prediction results produced by three methods, the proposed ESN based on CKF/KF dual estimation, the ESN based on KF without state estimation, and the generic ESN, are comparatively presented in Fig. 6. From this figure, the proposed method with state estimation presents the best performance on the respect of prediction accuracy, evaluated by the absolute error in Fig. 7. From this figure, the generic ESN is the most effective

Table 1
The comparative results of prediction performance.

Methods	Mackey-Glass time series	RMSE	MAPE	Computational cost(s)
EKF/KF based ESN	Standard	0.0013	0.1046	34.46
	Noisy	0.04304	3.8283	39.606
UKF/KF based ESN	Standard	0.0013	0.1109	246.34
	Noisy	0.04231	3.8386	249.237
CKF/KF based ESN	Standard	0.0012	0.1006	84.943
	Noisy	0.04025	3.6515	89.159
KF based ESN	Standard	0.0017	0.1440	23.45
	Noisy	0.06572	5.9925	22.628
Generic ESN standard	Standard	0.000044	0.004	1.091
	Noisy	0.05111	4.7834	1.176

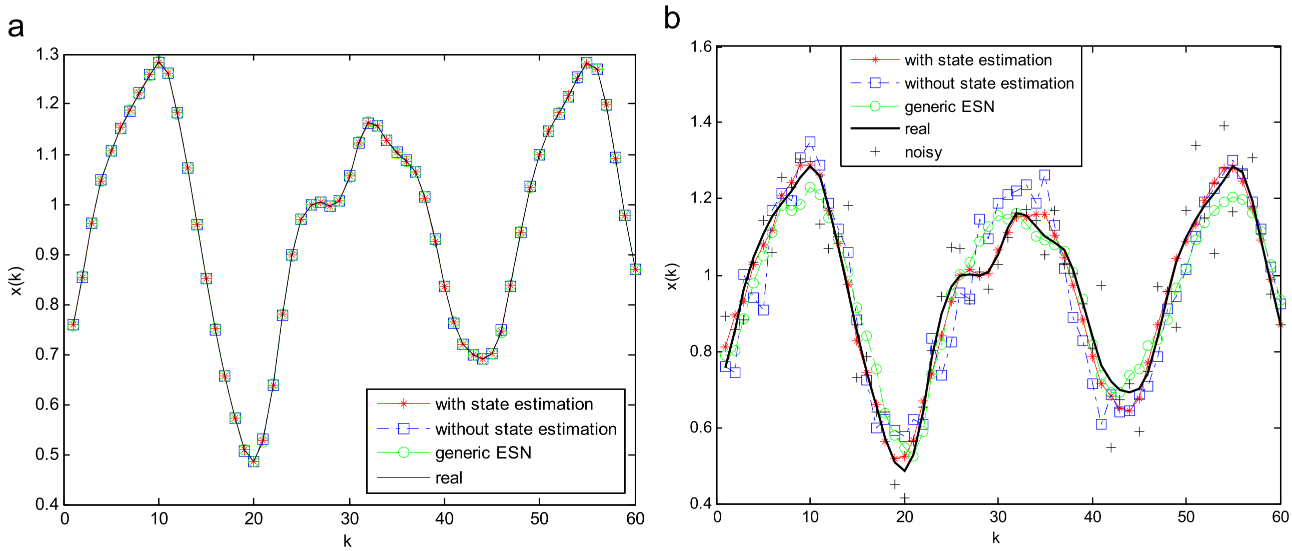

Fig. 6. Comparison of prediction results produced by proposed ESN based on CKF/KF dual estimation, ESN based on KF without internal states estimation and generic ESN. (a) Standard Mackey-Glass time series, (b) noisy Mackey-Glass time series.

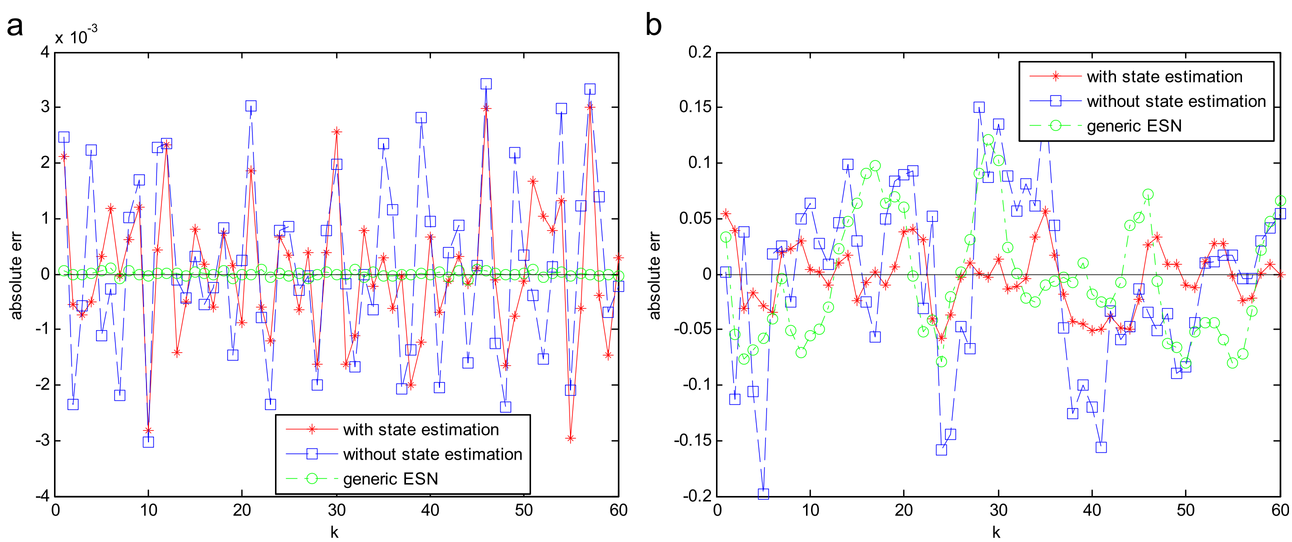

Fig. 7. Comparison of prediction error produced by proposed ESN based on CKF/KF dual estimation, ESN based on KF without internal states estimation and generic ESN. (a) Standard Mackey-Glass time series, (b) noisy Mackey-Glass time series.

one for the standard Mackey-Glass but perform worst for the noisy ones, and the KF based ESN shows a bad performance for both the standard and the noisy time series. For a quantified statistics, the detailed results are also listed in Table 1, where the root of mean square error (RMSE), the mean absolute percentage error (MAPE) and the computational time of these methods are exhibited.

$$RMSE = \sqrt{\sum_{i=1}^{n}(Y_i - F_i)^2 \Big/ n} \tag{34}$$

$$MAPE = \sum_{i=1}^{n}\left(\frac{|Y_i - F_i|}{Y_i} \times 100\right) \Big/ n \tag{35}$$

From Table 1, it is clear that the prediction accuracy of the nonlinear/linear dual estimation based ESN is definitely higher than that of the other methods for the noisy Mackey-Glass time series. In particular, the CKF/KF dual estimation based network gets the best result. For the standard Mackey-Glass time series, all these methods are effective and the generic ESN has the highest accuracy. But for the noisy ones, the proposed method is more effective and performs better than the others. That is to say, the proposed method is more robust than the other methods. With regard to the computational efficiency, the solving speed of the generic ESN is higher than the others; while, the UKF/KF based one performs the highest computational cost. Under the comprehensive consideration, the proposed CKF/KF based ESN is a suitable method with respect to the accuracy and the computing efficiency, which gives a sound outcome in the field of noisy nonlinear time series prediction.

5.2. Application on prediction for BFG flow

Steel industry is usually accompanied with high energy consumption and environmental pollution, in which the byproduct gas is viewed as one of the most useful energy resources. Hence, it

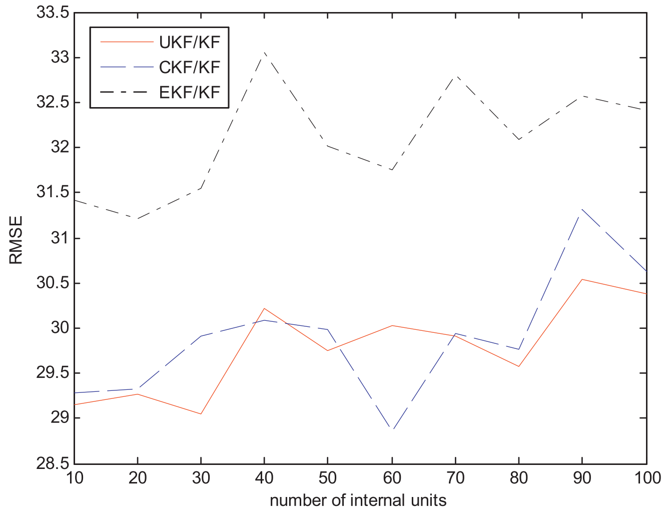

Fig. 8. Cross-validation results of the improved ESN with different number of internal units.

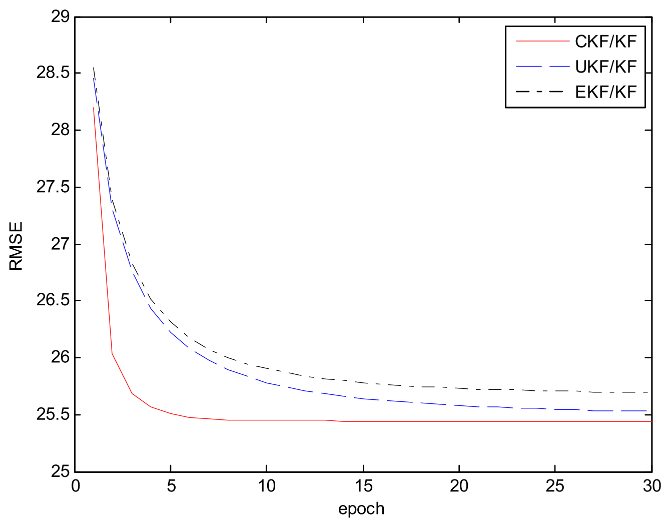

Fig. 9. Training curve of the three nonlinear/linear dual estimation based improved ESN.

is very meaningful to real-time predict the gas flow for energy conservation and the pollution reduction. Due to the complexity of gas generation process, it is very hard to establish physics based model for the flow prediction. In production practice, the current forecasting process mainly relies on the personal experience coming from the scheduling workers. In this study, the flow prediction for blast furnace gas (BFG) is studied. We chose the #1 blast furnace of Shanghai Baosteel Co. Ltd., China as the practical example to verify the effectiveness of the proposed method, and the real sample data covers the gas flow in August, 2010. The sampling interval of the gas flow series is 1 min. Considering the fact that the flow variation within 1 h can fundamentally reflect the dynamics feature of gas flow, the input dimensionality of the proposed ESN is set to 60. We randomly select the flow data of continuous 1060 min for the model training and testing, and use 30 epochs per run to train the constructed network. Namely, each epoch consists of 1000 samples and the length of each sample is 61. Thus, we obtain a set of samples denoted as $\{(\mathbf{u}_i, y_i) | i = 1, 2, \ldots, 1000\}$, where $[\mathbf{u}_i] = [u(i-(d_E-1)\Delta)u(i-(d_E-2)\Delta) \cdots u(i)]$, $[y_i] = [u(i+\Delta)]$, $d_E = 60$, $\Delta = 1$.

To obtain the optimal number of neurons in reservoir, the cross-validation method is employed and Fig. 8 shows the validation error presented by RMSE corresponding to the different number of internal units when the proposed ESN are trained. From Fig. 8, when the number of internal units is equal to 60, the network based on CKF/KF dual estimation exhibits the best performance. The training curves of gas generation time series using the proposed method are shown in Fig. 9, where the error is strictly convergent after 30 epochs training and the convergent speed of CKF/KF dual estimation is obviously much higher than the other two methods.

Table 2
The comparative analysis of prediction performance.

Methods	RMSE (km^3/h)	MAPE	Computational time (s)
EKF/KF based ESN	30.7087	5.8729	38.91
UKF/KF based ESN	29.2645	5.4194	260.91
CKF/KF based ESN	28.3257	5.2379	80.85
KF based ESN	35.1557	6.3798	21.56
Generic ESN	34.2332	6.3621	2.36

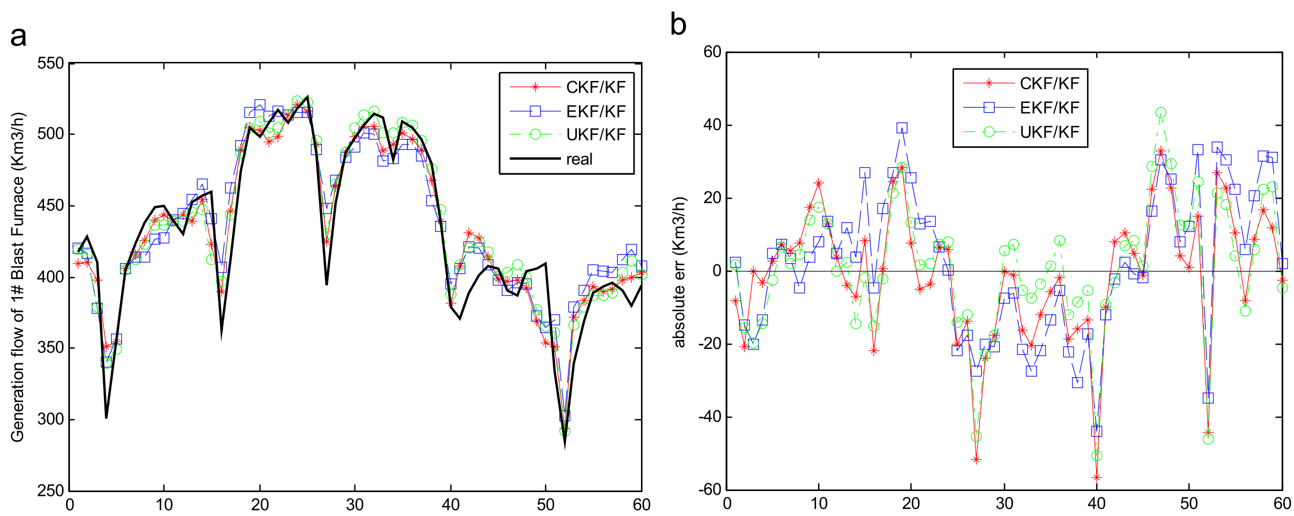

Fig. 10. Prediction results for the gas flow prediction by using the three dual estimation based ESN. (a) Prediction curve, (b) prediction error.

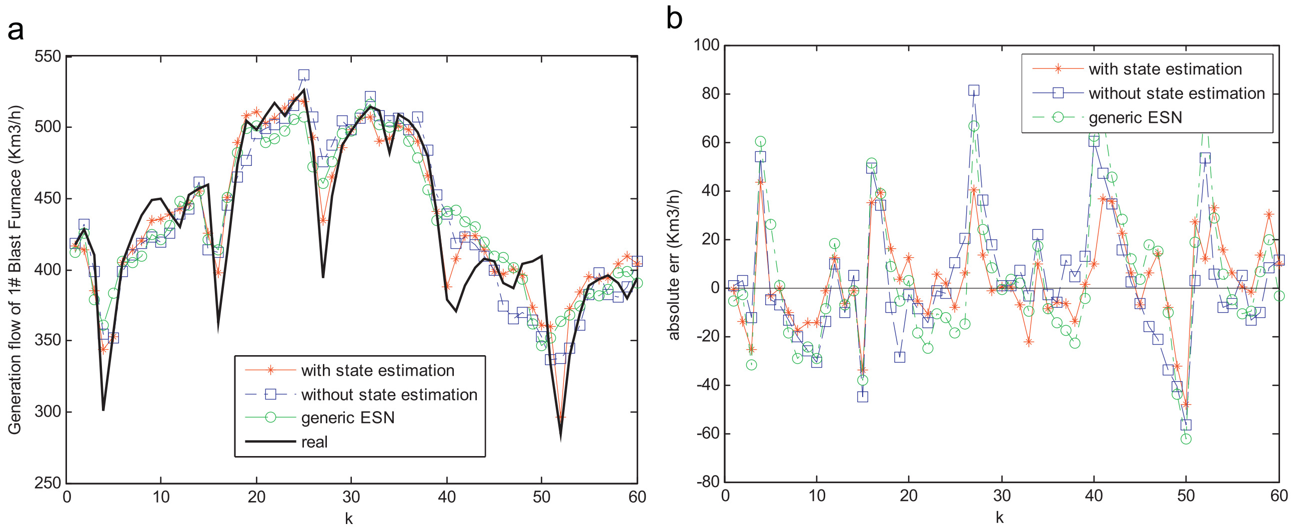

Fig. 11. Comparison of prediction results produced by three methods. (a) Prediction curve, (b) prediction error.

Furthermore, a group of comparative prediction results for the gas flow randomly selected also in August, 2010, are shown in Fig. 10. From Fig. 10(a), a 60-min prediction is presented by the three dual estimations based ESNs and their absolute prediction errors are shown as Fig. 10(b). Likewise, the average absolute error by the proposed CKF/KF dual estimation one is the lowest. The corresponding quantified indexes for the prediction performance can be listed in Table 2.

Here we similarly represent the effort by the state estimation, the prediction experiments also involve the three methods, the proposed ESN, the ESN based on KF without state estimation and the generic ESN, whose comparative results are presented in Fig. 11. From this figure, similar to the simulations using the above noisy Mackey-Glass, the proposed method exhibits the best performance on prediction accuracy, and the absolute error is meanwhile depicted in Fig. 11(b). Table 2 gives the quantified statistics for the prediction results, where RMSE, MAPE and computational time serve to evaluate the running quality. Likewise, the accuracy of the first three methods is definitely higher than that of other methods. On the comprehensive perspective of practical application, the proposed CKF/KF based improved ESN is a more effective method for the prediction problem of the industrial BFG generation flow.

6. Conclusion

In this study, an improved ESN based on nonlinear/linear dual estimation is proposed to predict a class of noisy nonlinear time series, in which the dual estimation is used to estimate the internal states and the output weights of the network for the purpose of overcoming the drawbacks of the generic ESN when the uncertainties involved in the network. The experimental results show that the proposed method is accurate and robust for noisy nonlinear time series prediction and performs well for the flow prediction in steel industry.

The future work should be focused on three aspects. First, the model parameters such as the sparsity of reservoir and the spectral radius of internal state matrix selected by the empirical experiments should be further studied theoretically. Second, the state estimation is only for neural network in this study, the state estimation for stochastic networks [32] and complex networks [33,34] with randomly varying nonlinearities should be considered in future. Third, some practical problems that probably accompanies with not only the Gaussian noise but also that of other forms should be further studied, in which the noise level might be much higher than what presented in this article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (nos. 61034003 and 61104157). The cooperation of energy center of Shanghai Baosteel Co. Ltd, China, in this work is greatly appreciated.

References

[1] K. Hornik, M. Stinchcombe, H. White, Multilayer feed-forward networks are universal approximators, Neural Networks 2 (1989) 359–366.
[2] H. Jaeger, H. Haass, Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless telecommunication, Science 304 (2004) 78–80.
[3] F. Zhai, X. Lin, Z. Yang, Financial time series prediction based on echo state network, in: 2010 Sixth International Conference on Natural Computation, 2010, pp. 3983–3987.
[4] F. Wyffels, B. Schrauwen, A comparative study of reservoir computing strategies for monthly time series prediction, Neurocomputing 73 (2010) 1958–1964.
[5] H. Showkati, A.H. Hejazi, S. Elyasi, Short term load forecasting using echo state networks, in: The 2010 International Joint Conference on Neural Networks, 2010, pp. 1–5.
[6] Y. Xia, B. Jelfs, M.M. Van Hulle, An augmented echo state network for nonlinear adaptive filtering of complex noncircular signals, IEEE Trans. Neural Networks 22 (2010) 1–10.
[7] S. Seth, M.C. Ozturk, J.C. Principe, Signal processing with echo state networks in the complex domain, in: 2007 IEEE Workshop on Machine Learning for Signal Processing, 2007, pp. 408–412.
[8] P.G. Ploger, A. Arghir, T. Gunther, Echo state networks for mobile robot modeling and control, in: Robot Soccer World Cup 3020, 2003, pp. 157–168.
[9] H. Jaeger, Reservoir riddles: suggesting for echo state network research, in: 2005 IEEE International Joint Conference on Neural Networks, 2005, pp. 1460–1462.
[10] Y. Liu, J. Zhao, W. Wang, Improved echo state network based on data-driven and its application in prediction of blast furnace gas output, Acta Automatica Sin. 35 (2009) 731–738.
[11] Y. Xue, L. Yang, S. Haykin, Decoupled echo state networks with lateral inhibition, Neural Networks 20 (2007) 365–376.
[12] J.J. Steil, Online reservoir adaptation by intrinsic plasticity for backpropagation-decorrelation and echo state learning, Neural Networks 20 (2007) 353–364.
[13] Q. Song, Z. Feng, M. Lei, Stable training method for echo state networks with output feedback, in: 2010 International Conference on Networking, Sensing and Control, 2010, pp. 159–164.
[14] M. Han, D. Mu, Multi-reservoir echo state network with sparse Bayesian learning, Adv. Neural Networks (2010) 450–456.

[15] B. Zhang, Y. Wang, Echo state networks with decoupled reservoir states, in: 2008 International Workshop on Machine Learning for Signal processing, 2008, pp. 444–449.
[16] E.A. Wan, A.T. Nelson, Dual extended Kalman filter methods, in: S. Haykin (Ed.), Kalman Filtering and Neural Networks, Wiley, 2001, pp. 123–174.
[17] R. Mehra, A comparison of several non-linear filters for re-entry vehicle tracking, IEEE Trans. Autom. Control 16 (1971) 307–319.
[18] M. Athans, R.P. Wsihner, A. Bertolini, Suboptimal state estimation for continuous-time nonlinear systems from discrete noise measurements, IEEE Trans. Autom. Control 13 (1968) 504–514.
[19] T.S. Schei, A finite difference method for linearizing in nonlinear estimation algorithms, Automatica 33 (1997) 2051–2058.
[20] S. Julier, J. Uhlmann, A new method for the nonlinear transformation of means and covariances in filters and estimators, IEEE Trans. Autom. Control 45 (2000) 477–482.
[21] S. Julier, J. Uhlmann, Unscented filtering and nonlinear estimation, in: IEEE Invited Paper, Proceedings of the IEEE, vol. 92, 2004, pp. 401–422.
[22] I. Arasaratnam, S. Haykin, Cubature Kalman filters, IEEE Trans. Autom. Control 54 (2009) 1254–1269.
[23] I. Arasaratnam, S. Haykin, Nonlinear Bayesian filters for training recurrent neural networks, in: A. Gelbukh, E. Morales (Eds.), Advances in Artificial Intelligence, Springer, 2008, pp. 12–33.
[24] Z. Wang, X. Liu, Y. Liu, J. Liang, V. Vinciotti, An extended Kalman filtering approach to modeling nonlinear dynamic gene regulatory networks via short gene expression time series, IEEE/ACM Trans. Comput. Biol. Bioinformatics 6 (2009) 410–419.
[25] H. Yang, J. Li, F. Ding, A neural network learning algorithm of chemical process modeling based on the extended Kalman filter, Neurocomputing 70 (2007) 625–632.
[26] D. Simon, Training radial basis neural networks with the extended Kalman filter, Neurocomputing 48 (2002) 455–475.
[27] E.A. Wan, R. Merwe, The unscented Kalman filter, in: S. Haykin (Ed.), Kalman Filtering and Neural Networks, Wiley, 2001, pp. 221–280.
[28] Z. Wang, F. Yang, D.W.C. Ho, S. Swift, A. Tucker, X. Liu, Stochastic dynamic modeling of short gene expression time series data, IEEE Trans. NanoBiosci. 7 (2008) 44–55.
[29] M.C. Mackey, L. Glass, Oscillation and chaos in physiological control systems, Science 197 (1977) 287–289.
[30] G. Venayagamoorthy, B. Shishir, Effects of spectral radius and settling time in the performance of echo state networks, Neural Networks 22 (2009) 861–863.
[31] N. Zeng, Z. Wang, Y. Li, M. Du, X. Liu, Inference of nonlinear state-space models for sandwich-type lateral flow immunoassay using extended Kalman filtering, IEEE Trans. Biomed. Eng. 58 (2011) 1959–1966.
[32] B. Shen, Z. Wang, Y.S. Hung, G. Chesi, Distributed H-infinity filtering for polynomial nonlinear stochastic systems in sensor networks, IEEE Trans. Ind. Electron. 58 (2011) 1971–1979.
[33] J. Liang, Z. Wang, X. Liu, Distributed state estimation for discrete-time sensor networks with randomly varying nonlinearities and missing measurements, IEEE Trans. Neural Networks 22 (2011) 486–496.
[34] B. Shen, Z. Wang, X. Liu, Bounded H-infinity synchronization and state estimation for discrete time-varying stochastic complex networks over a finite-horizon, IEEE Trans. Neural Networks 22 (2011) 145–157.

Chunyang Sheng received the B.E. degree in 2008 from Shenyang institute of Technology, China. He is currently a Ph.D. candidate student in the School of Control Science and Engineering of Dalian University of Technology.

His research interests include intelligent optimization, data based prediction and scheduling and machine learning.

Jun Zhao received the B.S. degree in control theory from Dalian Jiaotong University, Dalian, China, and the Ph.D. degree in engineering from Dalian University of Technology, Dalian, China, in 2003 and 2008, respectively. He is currently an Assistant Professor with the Research Center of Information and Control, Dalian University of Technology.

His research interest covers industrial production scheduling, computer integrated manufacturing, intelligent optimization, and machine learning.

Ying Liu received the B.S. and Ph.D. degrees from Dalian University of Technology, Dalian, China, in 2004 and 2010, respectively. She currently holds a Post-Doctoral position with the Department of Control Science and Engineering, Dalian University of Technology.

Her research interest covers integrated production planning and scheduling, system simulation and modeling, intelligent optimization and application, and artificial neural network.

Wei Wang received the B.S., M.S., and Ph.D. degrees from Northeastern University, Evanston, IL, in 1982, 1986, and 1988, respectively, all in industrial automation. He is currently a Professor with the Research Center of Information and Control, Dalian University of Technology, Dalian, China. Previously, he was a Post-Doctor with the Division of Engineering Cybernetics, Norwegian Science and Technology University during 1990–1992, Professor and Vice Director of the National Engineering Research Center of Metallurgical Automation of China during 1995–1999, and a Research Fellow with the Department of Engineering Science, University of Oxford during 1998–1999.

His research interest covers adaptive control, computer integrated manufacturing, and computer control of industrial process.