Neural Comput & Applic (2017) 28 (Suppl 1):S183–S195
DOI 10.1007/s00521-016-2306-8

ORIGINAL ARTICLE

Time series forecasting based on wavelet decomposition and feature extraction

Tianhong Liu[1] · Haikun Wei[1] · Chi Zhang[1] · Kanjian Zhang[1]

Received: 27 October 2015 / Accepted: 30 March 2016 / Published online: 30 April 2016

Abstract Time series forecasting is one of the most important issues in numerous applications in real life. The objective of this study was to propose a hybrid neural network model based on wavelet transform (WT) and feature extraction for time series forecasting. The motivation of the proposed model, which is called PCA-WCCNN, is to establish a single simplified model with shorter training time and satisfactory forecasting performance. This model combines the principal component analysis (PCA) and WT with artificial neural networks (ANNs). Given a forecasting sequence, order of the original forecasting model is determined firstly. Secondly, the original time series is decomposed into approximation and detail components by employing WT technique. Then, instead of using all the components as inputs, feature inputs are extracted from all the sub-series obtained from the above step. Finally, based on the extracted features and all the sub-series, a famous neural network construction method called cascade-correlation algorithm is applied to train neural network model to learn the dynamics. As an illustration, the proposed model is compared with two classical models and two hybrid models, respectively. They are the traditional cascade-correlation neural network, back-propagation neural network, wavelet-based cascade-correlation network using all the wavelet components as inputs to establish one model (WCCNN) and wavelet-based cascade-correlation network with combination of each sub-series model (WCCNN multi-models). Results obtained from this study indicate that the proposed method improves the accuracy of ANN and can yield better efficiency than other four neural network models.

Keywords Time series forecasting · Wavelet transform · Principal component analysis · Feature extraction · Neural network · Cascade-correlation algorithm

1 Introduction

Time series forecasting is an energetic research area which has drawn significant attention in a variety of domains such as economy, industry engineering and hydrology. By using time series forecasting, past observations of the same variable are collected and analyzed to develop a model describing the underlying relationships. Over the past several decades, many efforts have been devoted to the development and improvement of time series forecasting models. These models are developed from traditional linear prediction models and gradually improved and transited to nonlinear models.

Traditional statistical models have been focused on and applied because of their relative simplicity in understanding and implementation [1–5]. Linear prediction methods mainly include a series of classical statistical models based on autoregressive moving average (ARMA) model. Nonlinear prediction methods include complex models represented by all kinds of new types of machine learning methods. The limitation of linear models is that the underlying process studied is assumed to be linear, and consequently, the models may fail to capture nonlinear features commonly encountered in practice [6]. In recent years, machine learning methods have gradually become the main methods of solving the nonlinear and non-

✉ Haikun Wei
hkwei@seu.edu.cn

[1] Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, People's Republic of China

stationary complex time series prediction problems due to their strong nonlinear approximation ability. ANNs are one type of extensively used machine learning methods. Approaches based on ANNs for time series forecasting have produced convincing results [7–15], and the vast body of the literature is still growing.

Wavelet transform (WT) is a frequently used data preprocessing method. It can analyze a signal in both time and frequency domain, so that it surmounts the shortcomings of conventional Fourier transform. WT has been successfully embedded in various time series prediction models and has achieved satisfying results [16–19]. It provides an effective decomposition of time series; therefore, the sub-series can improve the performance of the forecasting models by seizing the available information on different resolution levels.

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observation of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (PCs). The number of PCs is less than or equal to the number of original variables. Y. Ouyang [20] used PCA to evaluate the water quality monitoring stations and showed that the number of monitoring stations can be reduced from 22 to 19. Wang [21] presented an improved method which integrates the PCA into a stochastic time strength neural network for forecasting financial time series. PCA extracted 2 PCs as the input data from six variables.

In many existing literature, sub-models are established separately according to the decomposition levels. Forecasting results are combined based on these sub-models finally. The combination of these results would improve forecasting accuracy compared with using original series without decomposition. However, the problems brought about are that both the calculation load and training time of the whole model would be magnified. Another problem is that the more sub-models are established, the more parameters need to be determined. This may lead to worse model generalization ability. The main objective of this paper is to provide an improved method for time series forecasting to solve these problems. A single simplified hybrid neural network model is proposed in this study. This model applies PCA to extract useful information from the wavelet components to construct model inputs. The model is trained by cascade-correlation algorithm. Combination of WT and PCA provides a method for solving feature extraction problems. The effectiveness of the proposed hybrid approach is demonstrated by the results obtained from both artificial data set and real-world data set. Forecasting of benchmark time series of Mackey-Glass as a hand-designed system is carried out. The mean daily flow of Oldman River near Brocket is used as real-world data set.

The rest of this paper is organized as follows. In Sect. 2, methodologies used in this paper are introduced. Section 3 illustrates the proposed hybrid model, and the modeling procedures are given. The experimental results are presented and discussed in Sect. 4. Conclusions based on the study are highlighted in Sect. 5.

2 Methodology

In this section, methodologies which will be used for constructing forecasting models are introduced. The WT and PCA methods are applied to decompose and feature extraction. An ANN model with cascade-correlation (CC) architecture as learning algorithm introduced in this section is employed as the prediction model. The principles of these methods are presented in detail in the following subsections.

2.1 Wavelet transform (WT)

WT is an essential time–frequency analysis tool for signal processing. It has been widely used because of its better performance compared with the Fourier transform. The basic aims of wavelet analysis are both to determine the frequency (or scale) contents of a signal and to assess and determine the temporal variation of this frequency content [22]. It has an advantage of having flexibility in choosing the mother wavelet in light of the characteristics of time series. Elaborate mathematical definitions of the wavelets are given in the literature [23].

WT can be divided into two categories: continuous wavelet transform (CWT) and discrete wavelet transform (DWT). The CWT for an original signal $f(t)$ with respect to a mother wavelet function $\psi(t)$ can be defined as [22]:

$$\mathrm{CWT}_f(a,b) = \langle f(t), \psi_{a,b}(t) \rangle = \frac{1}{\sqrt{|\alpha|}} \int_{-\infty}^{+\infty} f(t)\psi^* \left(\frac{t-b}{a} \right) \mathrm{d}t \tag{1}$$

where * is the complex conjugate of $\psi(t)$ and a and b denote the scale parameter and translation index, respectively. In CWT, wavelet coefficients are produced by continuously dilating and translating the mother wavelet, so that the wavelet coefficients are calculated for all possible scales and times. In DWT, it produces only the minimal number of coefficients necessary to reconstruct the original signal function $f(t)$. This reduction is achieved by the discretization of the parameters a and b, so that

$$\begin{cases} a = 2^{-s} \\ b = k2^{-s} \end{cases} \tag{2}$$

where s and k belong to the integer set Z and control the wavelet dilation and translation, respectively. Thus, a discrete version of CWT is obtained as

$$\begin{aligned} \mathrm{DWT}_f(a,b) &= \langle f(t), \psi_{a,b}(t) \rangle \\ &= \frac{1}{\sqrt{|2^{-s}|}} \int_{-\infty}^{+\infty} f(t)\psi^*\left(\frac{t-k/2^s}{1/2^s}\right)\mathrm{d}t \end{aligned} \tag{3}$$

By using wavelet discretization, timescale space can be sampled at discrete levels.

In this study, DWT is employed to process an original time series into a group of sub-series. An algorithm developed by Mallat [24] is employed in multi-solution computation, which represents an efficient way to implement the DWT using filters. The original series are passed through two kinds of filters. The low-pass filters which are associated with the scaling function allow the analysis of low-frequency components. The high-pass filters which are associated with the wavelet function allow the analysis of high-frequency components. An n levers wavelet decomposition showing the decomposition process is presented in Fig. 1. Signal A represents low-frequency approximation component, while signal D contains detail information of high-frequency component. The decomposition processes will be continued until reach the required number of n levels.

Many types of wavelets, including Daubechies, Symmlet, Gaussian, Mexican hat, Morlet and Shannon wavelets, can be used for wavelet-based time series analysis [31]. Daubechies wavelets are one of the most extensively used wavelets. They represent a collection of orthogonal mother wavelets with compact support, characterized by a maximal number of vanishing moments for some given length of the support. They often express as dbN, where N refers to the number of vanishing moment. The Daubechies wavelet transform is defined by computing running averages and differences via scalar products with scaling signals and wavelets [26]. The scaling signals and wavelets produce averages and differences using just a few more values from the signal. This provides a tremendous improvement in the capabilities of the Daubechies transforms. Daubechies wavelets are concerned in this paper.

2.2 Principal component analysis (PCA)

Feature extraction which contributes to removing redundant or irrelevant inputs features, not only reduces computing time for learning, but also improves forecasting accuracy. In order to make sure that the features of

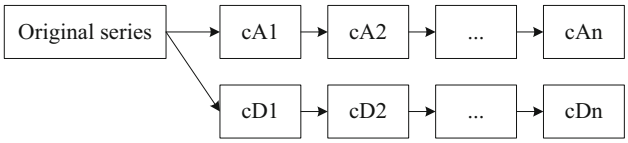

Fig. 1 Computational process of DWT

WCCNN model are useful and the scale is small, PCA is introduced to extract the feature inputs and reduce the data size. Results of a PCA are usually discussed in terms of components scores (the transformed variable values corresponding to a particular case in the data) and loading (the weight by which each standardized original variable should be multiplied to obtain the component score). Assume the data matrix with p variables, x_1, x_2,…, x_p, m times observations, the specific steps of PCA are as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mp} \end{bmatrix} \tag{4}$$

The decomposition components are normalized by using the following method:

$$Y_{ji} = (x_{ji} - \bar{x}_i)/S_i \tag{5}$$

where

$$\bar{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ji}, \quad S_i = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(x_{ji} - \bar{x}_i)^2}$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_\mathrm{p} \geq 0$ be the eigenvalues of covariance matrix of normalized data and $L_1, L_2, \dots, L_\mathrm{p}$ be the corresponding eigenvectors, the ith PC is such that

$$\mathrm{PC}_i = L_i^T X \tag{6}$$

where $i = 1,2, \dots, p$. The contribution rate of kth PC and the cumulative contribution rate of first k PCs are as follows:

$$\lambda_k / \sum_{i=1}^{p} \lambda_i, \quad \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$$

In PCA, the cumulative contribution rate represents the ability of synthesizing information. If the cumulative contribution rate exceeds some values, it can be considered that the first k PCs contain the most information of p original variables. These values need to be determined according to the concrete issues. The PCs reduce input variables and achieve the purpose of simplifying the problem. By selecting the most significant PCs, it is possible to identify certain relationships among the parameters of the data set under consideration.

2.3 Artificial neural networks (ANNs)

ANNs are data-driven and nonparametric models. A supervised learning algorithm for ANNs used in this paper is CC architecture, which was proposed by Fahlman and Lebiere (1900) [27]. This architecture is able to construct an appropriate small network automatically, which only needs to adjust the new weights. The processes of CC architecture are illustrated in Fig. 2. Black circles are frozen connection weights, and white circles are weights trained during output-training phase. The vertical lines sum all incoming activation. This architecture begins with a minimal network that has some inputs and one or more outputs units. The network which has no hidden units is a single layer network and is equivalent to a linear model (Fig. 2a). The hidden units are added to the network one by one as needed during learning, obtaining a multilayer structure (Fig. 2b, c). Each of the hidden units is placed into a new hidden layer. These units receive a connection from each of the network's original inputs and also from every preexisting hidden unit. This makes it possible to create high-order nonlinear feature detectors, customized for the problem at hand. There is also a bias input, permanently set to +1. Once a new unit has been added to the network, its input weights are fixed; only the output weights are trained repeatedly.

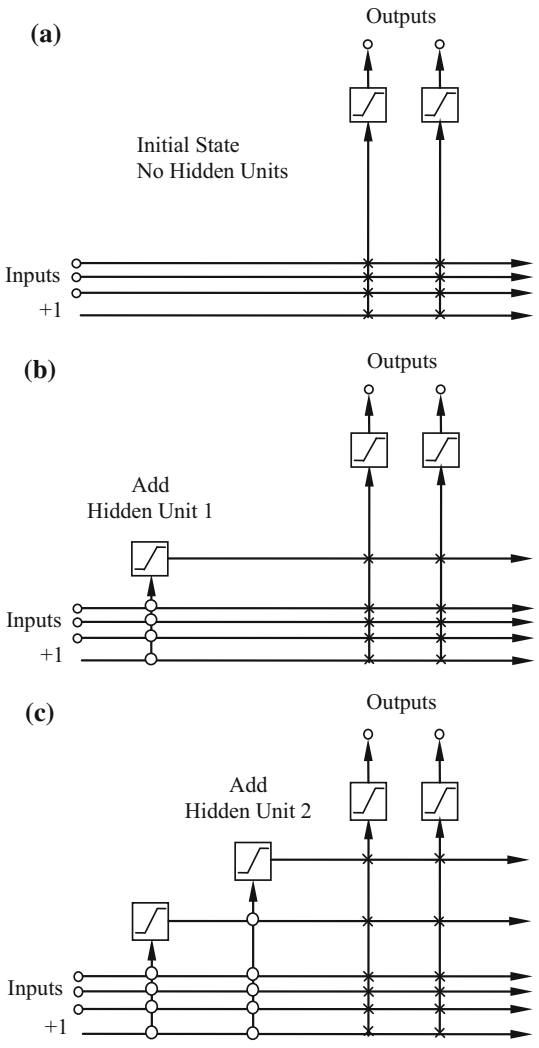

Fig. 2 Cascade-correlation (CC) architecture

CCNN was first developed in attempt to overcome certain problems and limitations of the popular back-propagation learning algorithm. The benchmark problem chosen for CCNN was "two-spirals," proposed by Alexis Wieland of MITRE Corp because it is an extremely hard problem for back-propagation algorithms; the network determines its own size and topology. A feed forward BP network model is established to compare the performance with the CC network in this study.

In order to prove the advantages of proposed model, the flowcharts of WCCNN and WCCNN multi-models which are the conjunction models of wavelet decomposition and CCNN are illustrated in Fig. 3.

In WCCNN model, the original time series are decomposed by DWT after identifying model order. The approximation and detail components obtained are then used as the CCNN model inputs. In WCCNN multi-models, the approximation and detail components are applied to set up sub-models, respectively. Forecasting values of these models are added to obtain the final results.

3 Proposed PCA-WCCNN model

A hybrid PCA-WCCNN model is proposed in this paper. The purpose of decomposition and feature extraction is, on one hand, to improve the predictive accuracy and, on the other hand, to present a single simplified model to reduce the training time.

Before constructing input–output pairs, the potential orders are needed to be determined. The Box–Jenkins methodology is adopted to achieve this objective [28]. Box–Jenkins methodology is a five-step process for identifying, selection and assessing conditional mean models for discrete, univariate time series data. The sample autocorrelation function (ACF) and partial autocorrelation function (PACF) are used in this paper. If time series are not stationary, successively difference series to attain stationarity is necessary. The ACF and PACF of a stationary series cut off completely after a few lags. In time series analysis, PACF gives the partial correlation of a time series with its own lagged values, controlling for the values of the time series at all shorter lags. The use of these functions is introduced as part of the Box–Jenkins approach to time series modeling. By plotting the partial autocorrelative

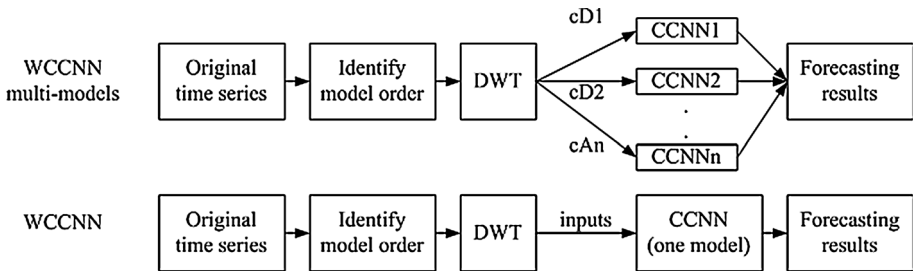

Fig. 3 Flowcharts of WCCNN multi-models and WCCNN

functions, one could determine the appropriate lags, which give the potential model orders.

Multilevel wavelet decomposition is considered to decompose the original time series into an approximation series and a set of detail series using DWT technique. In the feature selection process, models require that the input variables should have poor correlation. Because strong correlation between input variables implies that they carry more repeated information, and it may increase the computational complexity and reduce the prediction accuracy of the model. The essence of PCA is the rotation of space coordinates that does not change the data structure. The obtained PCs are the linear combination of variables, reflecting the original information to the greatest degree. These PCs are uncorrelated with each other. The framework of forecasting steps of PCA-WCCNN model is illustrated in Fig. 4.

4 Experiments

For evaluating the performance of the proposed forecasting model, the Mackey-Glass time series [29] and a mean daily flow of Oldman River near Brocket from Time Series Data Library (TSDL) [30] are used in this study. The time series forecasting based on the chaotic Mackey-Glass differential equation is a standard benchmark in the areas of neural networks for comparing the learning and generalization abilities of different algorithms. River flow modeling and prediction is one of the earliest forecasting problems to have attracted the interest of a good number of scientists and is one of the most frequently analyzed problems in hydrology.

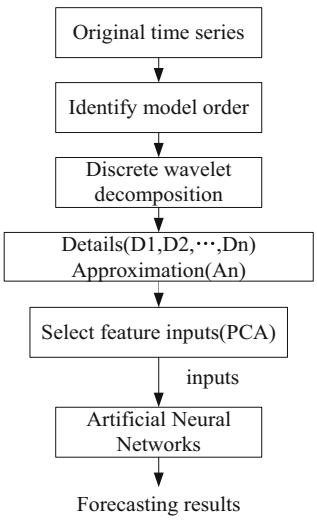

Fig. 4 Framework of forecasting steps of PCA-WCCNN model

4.1 Performance evaluation

To evaluate the accuracy of the proposed model, two different criterions including the root-mean-squared error (RMSE) and the mean absolute error (MAE) are used in the experiments. These performance indexes can be written as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i - y_i]^2} \tag{7}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{8}$$

where N is the number of observed values, $\hat{y}_\mathrm{i}$ is the predicted value and y_i is the actual value. The elapsed time is calculated in the experiments in order to compare the training speed of the models.

4.2 Application to the Mackey-Glass time series forecasting

4.2.1 Analysis

The Mackey-Glass equation was first introduced as a model of white blood cell production equation. One interesting feature of the Mackey-Glass problem is that real-value outputs are required instead of the discrete output values found in most neural network benchmarks. Time series which will be predicted with network algorithm is generated from the following equation

$$\frac{\mathrm{d}(x)}{\mathrm{d}t} = \beta x(t)\frac{\alpha x(t-\tau)}{1+x(t-\tau)^{10}} \tag{9}$$

with parameters $\alpha = 0.2$, $\beta = -0.1$, and $\tau = 17$ [31]. Three thousand and five hundred data points are generated with an initial condition of $x(0) = 1.2$ and $x(t) = 0$ when $t < 0$ based on the second-order Runge–Kutta method to discrete the differential equation. Figure 5 shows the portion of the series used for this study. The standard method for this prediction is to create a mapping f from D points of the time series spaced Δ apart, i.e., $[x(t - (D - 1)\Delta), \ldots, x(t - \Delta), x(t)]$, to predict future point $x(t + P)$. The embedding dimension $D = 2$ as the model orders are two in this case. Previous studies have used the value $\Delta = 6$ and prediction interval values of $P = 6$ and $P = 85$ [32, 33]. The characteristic time constant of $x(t)$ is $t_{\text{char}} = 50$, which makes it particular difficult to forecast $x(t + P)$ with $P > t_{\text{char}}$ [31]. By choosing $P = \Delta$, it is possible to predict the value of time series at any multiple of Δ time steps in the future by feeding the output back into the input and iterating the solution. Based on the analysis above, the model inputs consist of two past values of $x(t)$, i.e., $(x(t - 6), x(t))$ and the model is to predict the value of $x(t + 6)$ in this study.

The optimal decomposition levels and mother wavelet must be selected in advance to determine the performance in the wavelet domain. Several researchers have used an empirical equation to determine the decomposition level [34, 35]. The empirical equation is

$$L = \text{int}(\log(N)) \tag{10}$$

where L is the decomposition level, N is the sample numbers and int is the integer-part function. The trial-and-error procedure is used to determined the decomposition levels in this study and three decomposition levels are obtained which is coincide with the result of the empirical equation. The original time series are decomposed using Daubechies wavelets, and sub-series components are *D1, D2, D3* and *A3*. There are many Daubechies wavelets, but they are all very similar. Definition and advantages of *db4* wavelet can be found in the literature [25]. Based on the analysis of this time series, the db4 WT will be concerned. Original and decomposed time series are shown in Fig. 6. The values of detail *D1* are too small. In order to see it more clearly, only a part of the series are given. Since the other four series have significant regularity, more data than *D1* are used to draw this picture.

To compare with the CC algorithm, BP algorithm is used to ANNs. Chester [36] and Zhang [37] studies suggest that the ideal number of hidden layers in BP architecture is often one or two, and it is accepted that a network with three layers connected toward ahead can approximate any continuous function in a reasonable way [38]. BPNN used this study has one input layer, one output layer and one hidden layer between the input and output layers. Although there are some experimental suggestions for determining the number of hidden neurons, the trial-and-error procedure used in the current research which is more reliable in spite of its time consume inherence.

4.2.2 *Forecasting results*

The original time series is decomposed into sub-series. All the wavelet components are as follows:

$$[x_{D1}(t-6), x_{D1}(t), x_{D2}(t-6), x_{D2}(t), x_{D3}(t-6), x_{D3}(t), x_{A3}(t-6), x_{A3}(t)]$$

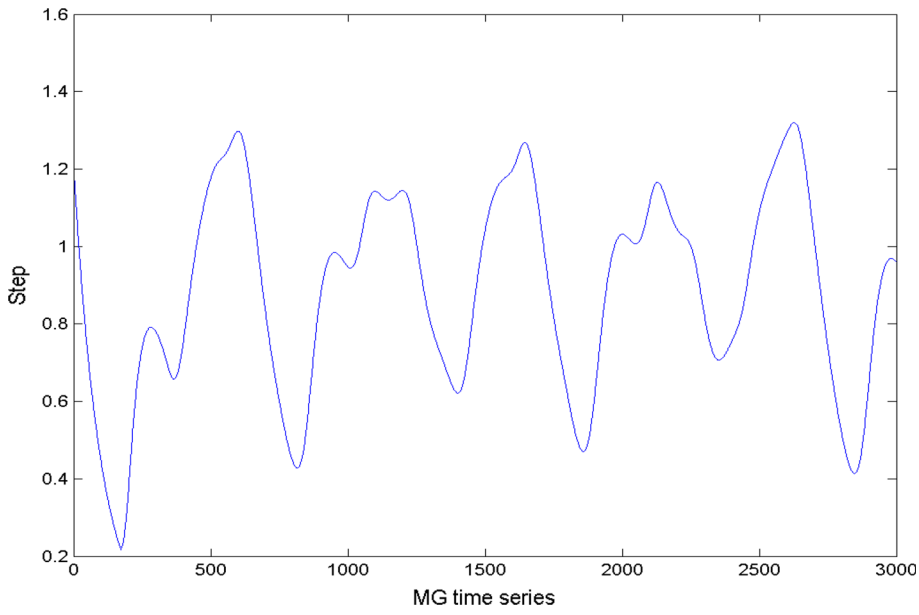

Fig. 5 Mackey-Glass time series

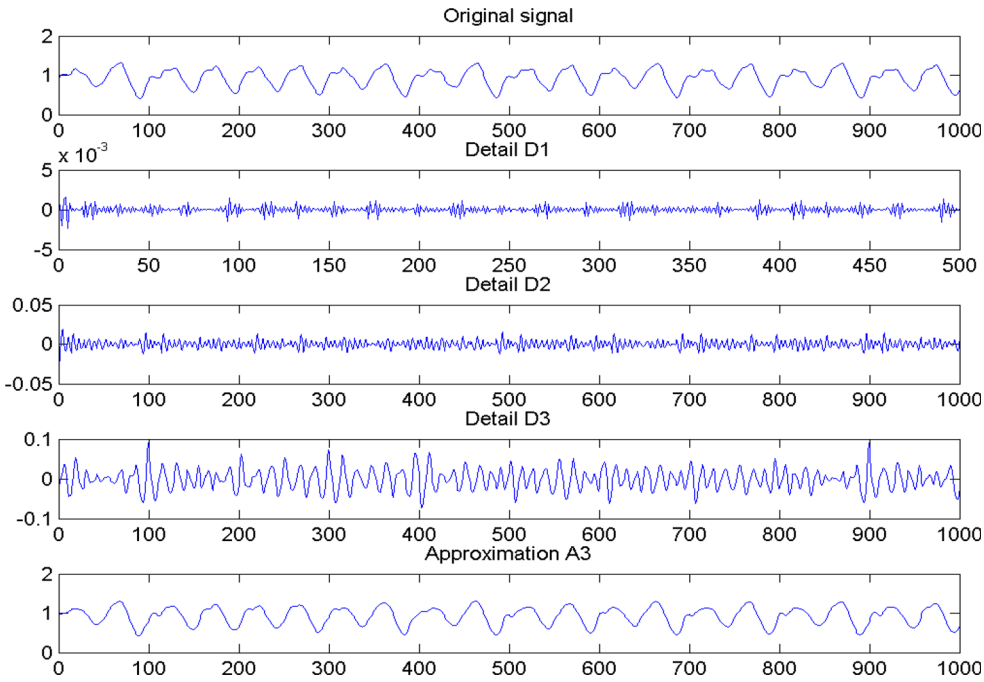

Fig. 6 Original and decomposed MG time series using db4 wavelet

Table 1 PCA results of MG time series components

Component	Eigenvalues	Contribution rate (%)	Cumulative contribution rate (%)
1	1.7056	0.2132	0.2132
2	1.6830	0.2104	0.4236
3	1.5103	0.1888	0.6124
4	1.0726	0.1341	0.7464
5	0.8924	0.1115	0.8580
6	0.4896	0.0612	0.9192
7	0.3313	0.0414	0.9606
8	0.3152	0.0394	1.0000

PCA is applied to extract the PCs from the above wavelet components. The contribution rate and cumulative contribution rate of the principal components are presented in Table 1. In this paper, if the cumulative contribution rate exceeds 85 %, it can be considered that the first k PCs contain the most information of p original variables.

This table indicates that the cumulative contribution rates of the first five PCs exceed 85 %, namely the first five PCs contain 85% information of the wavelet components. These five PCs are conducted as the inputs of the PCA-WCCNN model instead of using all the wavelet components. The training sample is 3000, while 500 data are used to test the model.

In BPNN, the original time series are used without decomposition. A number of experiments are carried out to set parameters, viz. initial input–hidden–output nodes, learning rate, epochs, activation function and learning error

Table 2 Parameters and their values during learning processes of the BPNN and CCNN

Parameter	BPNN value	CCNN value
Initial nodes	2-20-1	–
Learning rate	0.0001	0.01
Epochs	10000	1000
Activation function	Sigmoid	Sigmoid
Learning error	0.05	0.025

to obtain the optimal results. The values that exhibit the best behavior in terms of accuracy have been chosen. The determined optimal values of all these parameters including BP and CC algorithms are listed in Table 2. The blank in the table means that different models using CC algorithm have different initial node values. That should be determined in the specific model.

The testing results of PCA-WCCNN, WCCNN, WCCNN multi-models, CCNN and BPNN models are illustrated in Fig. 7. The elapsed time of each model is computed under the same condition. The performance of the five models and the elapsed time are shown in Table 3.

4.3 Application to the mean daily flow forecasting

4.3.1 Analysis

The second data set used in this paper is the mean daily flow of Oldman River near Brocket from Time Series Data

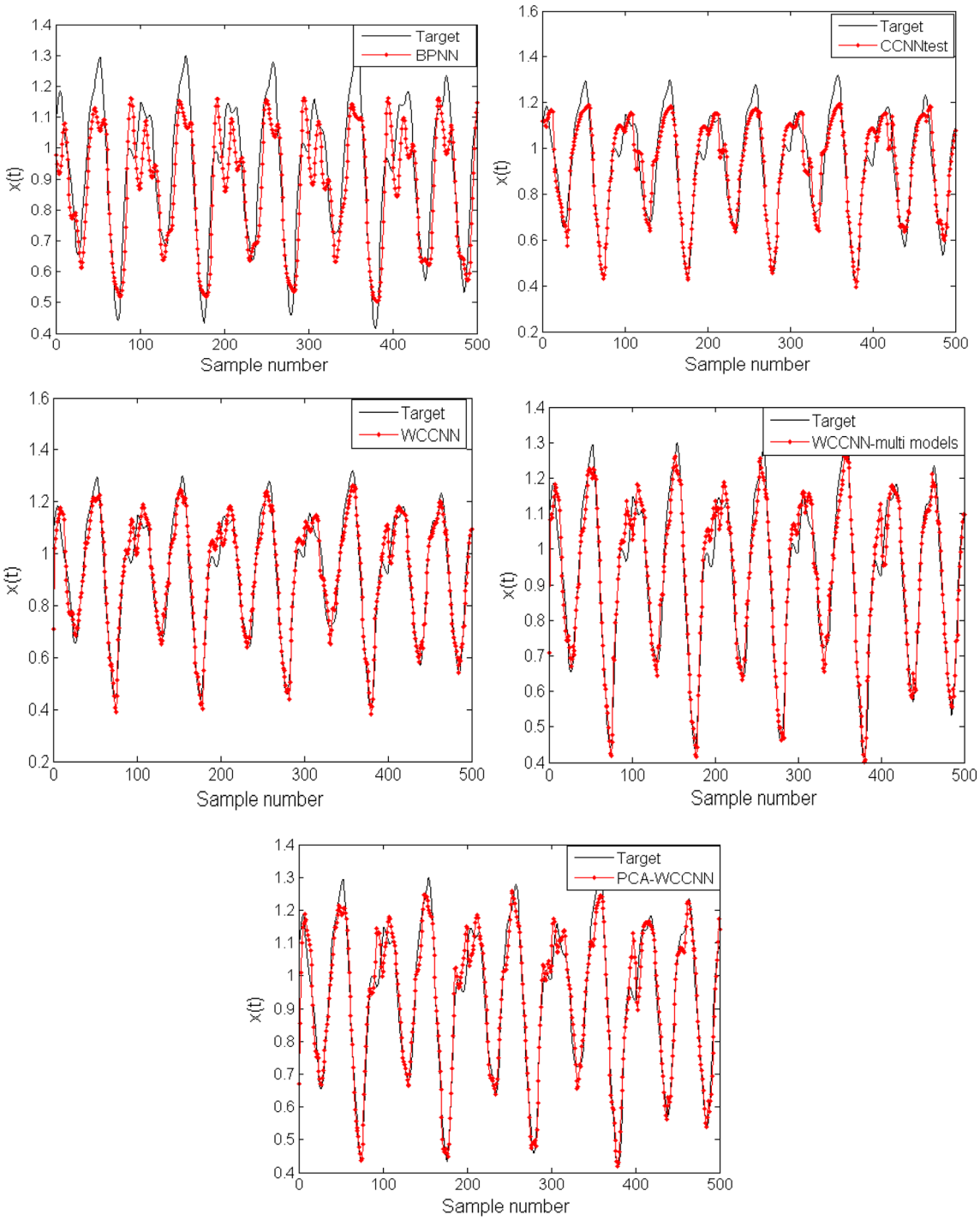

Fig. 7 Forecasting results of five models

Library (TSDL), which was created by Rob Hyndman, professor of the statistics at Monash University, Australia. This data set has 1460 fact values in one time series, from January 01, 1988, to December 31, 1991. The original time series are shown in Fig. 8.

The mean daily flow data set is used for PCA-WCCNN, WCCNN, WCCNN multi-models, CCNN and BPNN models, respectively. The training samples are accounted for 80 %, and the rest 20 % are used as testing samples. The original and decomposed mean daily flow time series using db4 wavelet are presented in Fig. 9. Before training, the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) are used to determine the model orders. Figure 10 plots the sample ACF and PACF

Table 3 The performances of the four models forecasting MG time series

Model	RMSE	MAE	Elapsed time (s)
PAC-WCCNN	0.0511	0.0381	35.81
WCCNN	0.0611	0.0437	33.25
WCCNN multi-model	0.0561	0.0419	130.64
CCNN	0.0740	0.0548	27.36
BPNN	0.1066	0.0898	35.02

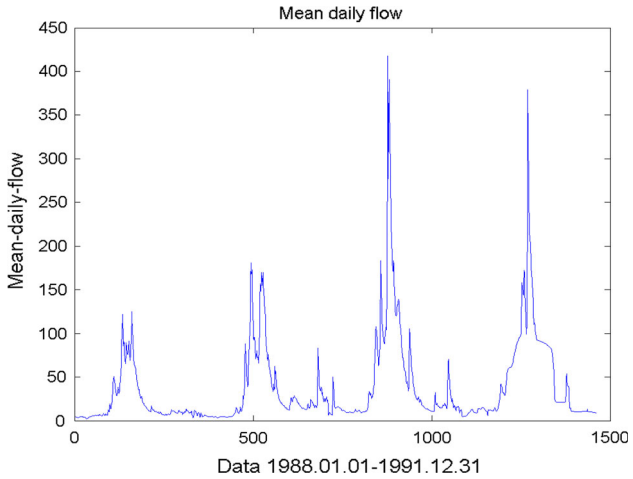

Fig. 8 Original time series of the mean daily flow (m^3/s)

for the series and the significant, linearly decaying ACF indicate a nonstationary process (Fig. 10a). In order to remove the linear trend, we take a first difference of the data and plot the sample ACF and PACF of the differenced series as shown in Fig. 10b. The blue lines are the upper and lower confidence bounds. The red dots are ACF and PACF values.

The differenced series appear more stationary. The sample ACF of the differenced series decays more quickly. The sample PACF cuts off after lag 3. Therefore, the model orders are identified as 3. Assume that all the wavelet components are expressed as follows:

$$[y_{D1}(t-2), y_{D1}(t-1), y_{D1}(t), y_{D2}(t-2), y_{D2}(t-1), y_{D2}(t), y_{D3}(t-2), y_{D3}(t-1), y_{D3}(t), y_{A3}(t-2), y_{A3}(t-1), y_{A3}(t)]$$

New feature inputs for WCCNN model from the wavelet components must be constructed. The same method applied in the MG time series is also used in this case.

4.3.2 *Forecasting results*

Table 4 gives the computed results of the contribution rate and cumulative contribution rate of the principal components. It can be seen that the cumulative contribution rate of the seventh component exceeds 85 %. The first to seventh components contain more than 85 % information of the wavelet components. These seven PCs are conducted as the input data of the PCA-WCCNN model instead of the wavelet components.

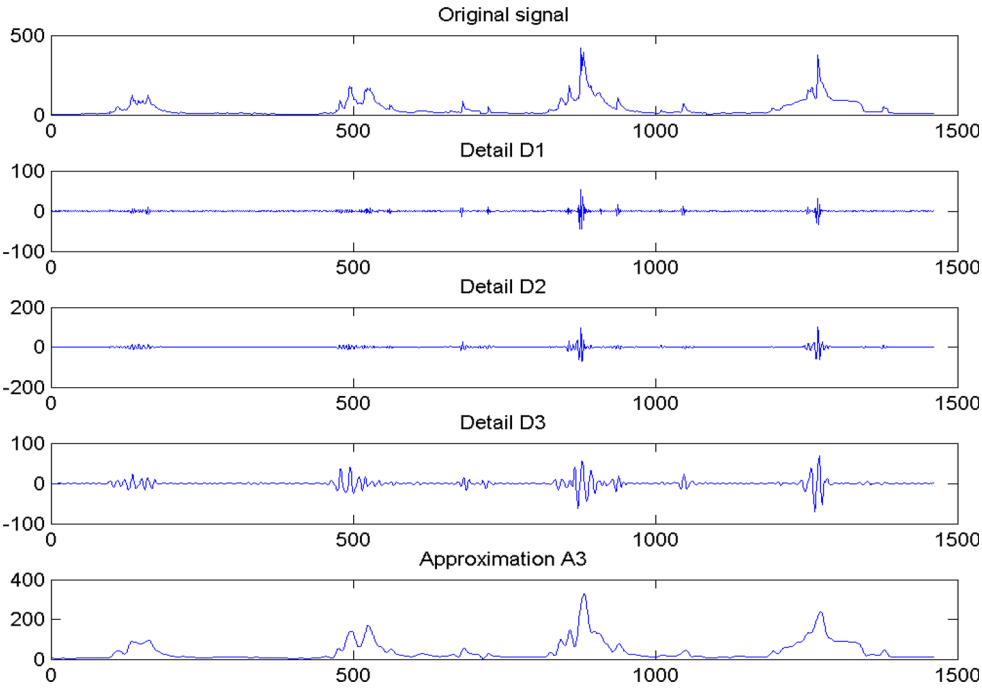

Fig. 9 Original and decomposed time series using db4 wavelet

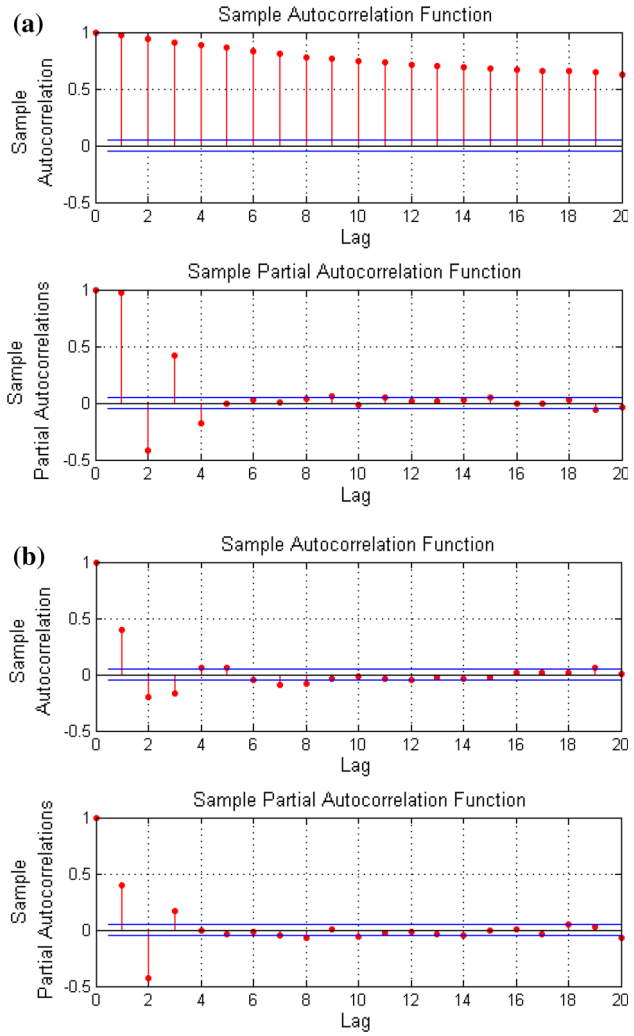

Fig. 10 Order identification for mean daily flow time series

Table 4 The PCA results of mean daily flow time series components

Component	Eigenvalues	Contribution rate (%)	Cumulative contribution rate (%)
1	2.9812	0.2484	0.2484
2	2.0443	0.1704	0.4188
3	1.6292	0.1358	0.5546
4	1.3104	0.1092	0.6638
5	1.0596	0.0883	0.7521
6	0.9740	0.0812	0.8332
7	0.7367	0.0614	0.8946
8	0.6791	0.0566	0.9512
9	0.3431	0.0286	0.9798
10	0.2238	0.0186	0.9984
11	0.0183	0.0015	0.9999
12	0.0003	0.0001	1.0000

Table 5 Parameters and their values during learning processes of the BPNN and CCNN

Parameter	BPNN value	CCNN value
Initial nodes	3-20-1	–
Learning rate	0.0001	0.01
Epochs	10,000	1000
Activation function	Sigmoid	Sigmoid
Learning error	0.05	0.025

The values that exhibit the best behavior in terms of accuracy of BP and CC algorithms have been chosen are illustrated in Table 5. The blank in the table means that different models using CC algorithm have different initial node values. That should be determined in the specific model.

Since the inputs are determined, the input–output pairs are extracted in order to train and test model. The testing results of five different models are illustrated in Fig. 11. The mean daily flow forecasting performances and elapsed time computed under the same condition are shown in Table 6.

5 Discussion

For each example, five models are constructed: CCNN, BPNN, WCCNN, WCCNN multi-models and PCA-WCCNN. The first two are single models, which are built using the original time series. The other three are wavelet decomposition-based models. Compare the performance of the five models from Tables 1, 2, 3, 4, 5, 6, conclusions can be summarized as follows:

1. CCNN and BPNN: The elapsed time illustrated that CCNN learns quickly than BPNN in these examples. The RMSE and MAE of the two models indicated that CCNN has better performance. This is because CC algorithm requires no back-propagation of error signals through the connections of the network. Moreover, it dose not need to decide the hidden layer structure beforehand as the network determines its own size and topology.
2. WCCNN and CCNN: Wavelet decomposition-based CCNN model outperform CCNN in terms of RMSE and MAE. This reveals the effectiveness of the wavelet decomposition. By selecting suitable filters, WT can greatly reduce the correlation between different characteristics of the original time series. As for training speed, WCCNN model is a little bit slower than

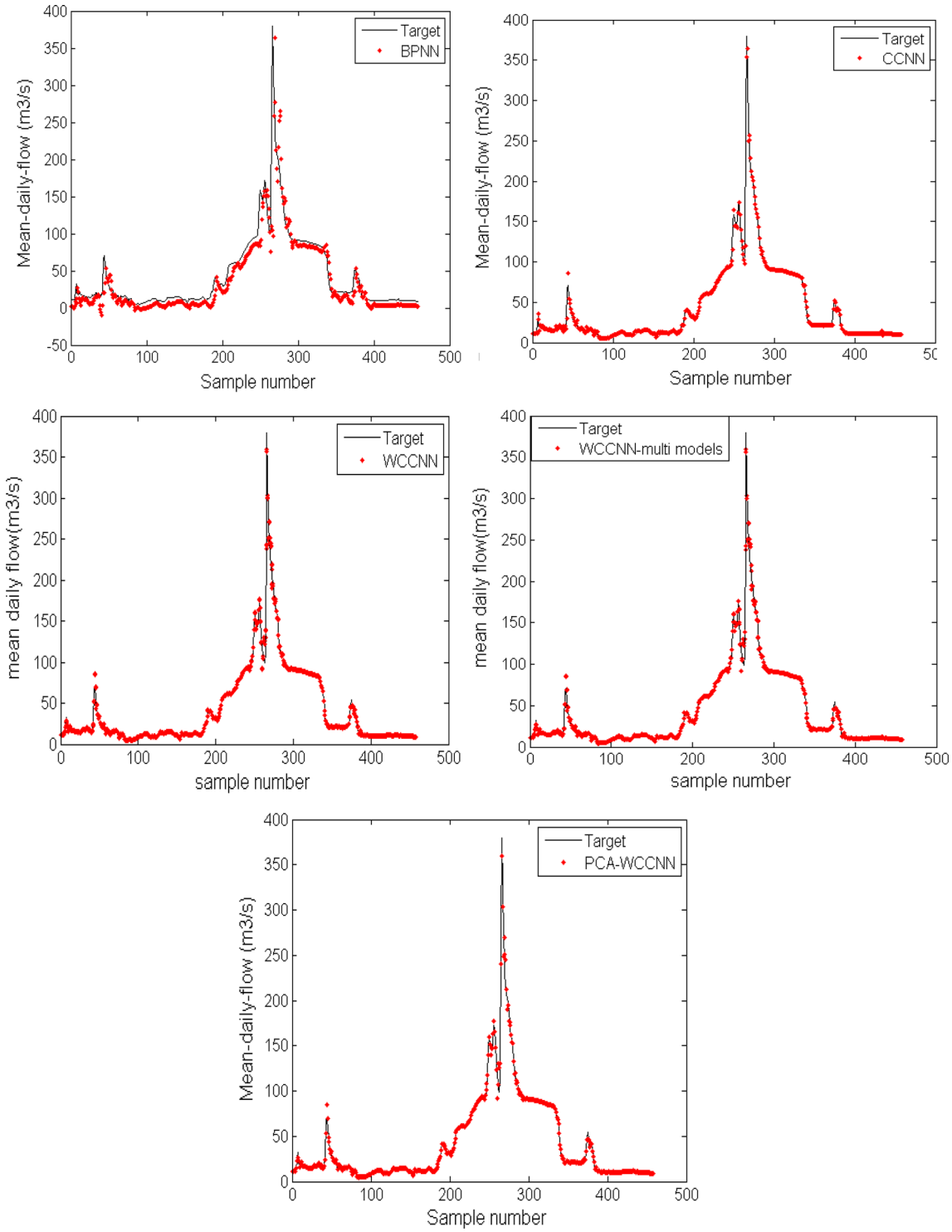

Fig. 11 Forecasting results of five models

CCNN. This can be interpreted as time taken by wavelet decomposition.

3. WCCNN and WCCNN multi-models: WCCNN multi-models use all the wavelet components as inputs to construct model for each sub-series. Although the RMSE and MAE in this time series are slightly better than WCCNN, the training time is much longer then WCCNN model. This greatly reduces the model efficiency.
4. PCA-WCCNN: The proposed PCA-WCCNN performs better than other model as illustrated by the calculated

Table 6 The performances of the five models forecasting mean daily flow

Model	RMSE (m^3/s)	MAE (m^3/s)	Elapsed time (s)
PCA-WCCNN	4.314	1.623	27.31
WCCNN	4.625	1.738	27.08
WCCNN multi-models	4.406	1.641	92.85
CCNN	4.854	2.382	25.64
BPNN	5.519	3.235	28.46

values of RMSE and MAE from tables. Appropriate inputs are selected through the PCA to extract significant information from wavelet components and insure the choice is better than others. PCA-WCCNN is concise compared to the WCCNN multi-models and more accurate than WCCNN. It also can help to shorten training time although the elapsed time is a little bit longer than WCCNN and CCNN models. This can be interpreted as time taken by wavelet decomposition and PCA processes.

6 Conclusions

From the above studies, it can be found that the model performance has much to do with feature inputs and learning algorithms. On the one hand, the combination of wavelet decomposition and feature selection improve the performance of neural networks significantly. On the other hand, CCNN takes shorter elapsed time then BPNN. It can be concluded that CC algorithm learns faster than traditional BP algorithm. Thus, CC algorithm can speed up the learning. Moreover, the feature selection also simplified model structural. Given consideration to forecasting accuracy and training efficiency, the proposed neural network model has the advantage of applying to time series forecasting. Further studies can also investigate the models using different kinds of mother wavelets.

Acknowledgments The authors gratefully acknowledge the financial support of this research by the National Natural Science Foundation of China (Grant No. 61374006), the Major Program of National Natural Science Foundation of China (Grant No. 11190015) and the Natural Science Foundation of Jiangsu (Grant No. BK20131300).

References

1. Kumar Jain VK (1999) Autoregressive integrated moving averages (ARIMA) modeling of a traffic noise time series. Appl Acoust 58(3):283–294
2. Ediger VS, Akar S (2007) ARIMA forecasting of primary energy demand by fuel in Turkey. Energy Policy 35(3):1701–1708
3. Zhang GP (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50:159–175
4. Khashei M, Bijari M (2011) A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Appl Soft Comput 11:2664–2675
5. Gan M, Cheng Y, Liu K, Zhang G (2014) Seasonal and trend time series forecasting based on a quasi-linear autoregressive model. Appl Soft Comput 24:13–18
6. Chen R, Tsay RS (1993) Functional-coefficient autoregressive models. J Am Stat Assoc 88(421):298–308
7. Gan M, Peng H, Dong X (2012) A hybrid algorithm to optimize RBF network architecture and parameters for nonlinear time series modeling. Appl Math Model 36(7):2911–2919
8. Silva CGD (2008) Time series forecasting with a non-linear model and the scatter search meta-heuristic. Inf Sci 178:3288–3299
9. Zhang GP, Kline DM (2007) Quarterly time-series forecasting with neural networks. IEEE Trans Neural Netw 18(6):1800–1814
10. Gerald C, Dimitri S (2007) Knowledge-based modularization and global optimization of artificial neural network models in hydrological forecasting. Neural Netw 20(4):528–536
11. Hippert HS, Taylor JW (2010) An evaluation of Bayesian techniques for controlling model complexity and selecting inputs in a neural network for short-term load forecasting. Neural Netw 23(3):386–395
12. Talaee PH (2014) Multilayer perceptron with different training algorithms for streamflow forecasting. Neural Comput Appl 24:695–703
13. Adhikari R (2015) A neural network based linear ensemble framework for time series forecasting. Neurocomputing 157: 231–242
14. Donate JP, Cortez P, Sánchez GG, Miguel AS (2013) Time series forecasting using a weighted cross-validation evolutionary artificial neural network ensemble. Neurocomputing 109:27–32
15. Firmino PRA, Neto PSGDM, Ferreira TAE (2014) Correcting and combining time series forecasters. Neural Netw 50:1–11
16. Joo TW, Kim SB (2015) Time series forecasting based on wavelet filtering. Expert Syst Appl 42:3868–3874
17. Seo Y, Kim S, Kisi O, Singh VP (2015) Daily water level forecasting using wavelet decomposition and artificial intelligence techniques. J Hydrol 520:224–243
18. Liu H, Tian H, Pan D, Li Y (2013) Forecasting models for wind speed using wavelet, wavelet packet, time series and artificial neural networks. Appl Energy 107:191–208
19. Karthikeyan L, Nagesh Kumar D (2013) Predictability of non-stationary time series using wavelet and EMD based ARMA models. J Hydrol 502:103–119
20. Ouyang Y (2005) Evaluation of river water quality monitoring stations by principal component analysis. Water Res 39:2621–2635
21. Wang J, Wang J (2015) Forecasting stock market indexes using principle component analysis an stochastic time effective neural networks. Neurocomputing 156:68–78

22. Heil CE, Walnut DF (1989) Continuous and discrete wavelet transforms. SIAM Rev 31(4):628–666
23. Gencay R, Selcuk F, Whitcher B (2001) an introduction to wavelets and other filtering methods in finance and economics. Academic Press, Elsevier
24. Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal 11(7):674–693
25. Walker JS (2008) A primer on wavelets and their scientific applications. CRC Press
26. Amiady N, Keynia F (2009) Short-term load forecasting of power systems by combination of wavelet transform and neuro-evolutionary algorithm. Energy 34(1):46–57
27. Fahlman SE, Lebiere C (1990) The cascade-correlation learning architecture, in advances in neural information processing systems 2. Morgan Kaufmann, San Mateo, pp 524–532
28. Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. Wiley
29. Mackey MC, Glass L (1977) Oscillation and Chaos in physiological control systems. Science 197:287
30. https://datamarket.com/data/set/235b/mean-daily-flow-oldman-rivernear-brocket-jan-01-1988-to-dec-31-1991
31. Crowder RS (1991) Predicting the mackey-glass time series with cascade-correlation learning, connectionist models. In: Proceedings of the 1990 summer school, pp 117–123
32. Zhao JS, Yu XJ (2015) Adaptive natural gradient learning algorithms for Mackey-Glass chaotic time prediciton. Neurocomputing 157:41–45
33. Mohammadi R, Fatemi Ghomi SMT, Zeinali F (2014) A new hybrid evolutionary based RBF networks method for forecasting time series: a case study of forecasting emergency supply demand time series. Eng Appl Artif Intell 36:204–214
34. Nourani V, Alami MT, Aminfar MH (2009) A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. Eng Appl Artif Intell 22(3):466–472
35. Adamowski J, Chan HF (2011) A wavelet neural network conjunction model for groundwater level forecasting. J Hydrol 407(1):28–40
36. Chester DL (1990) Why two hidden layers are better than one? In: Proceedings of the international joint conference on neural networks, pp 1265–1268
37. Zhang X (1994) Time series analysis and prediction by neural networks. Optim Method Softw 4:151–170
38. Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2(5):359–366