Environmental Pollution xxx (2017) 1–13

Contents lists available at ScienceDirect

Environmental Pollution

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

Daily air quality index forecasting with hybrid models: A case in China☆

Suling Zhu [a], Xiuyuan Lian [b, *], Haixia Liu [b], Jianming Hu [b], Yuanyuan Wang [b], Jinxing Che [c]

[a] School of Public Health, Lanzhou University, Lanzhou 730000, Gansu, China
[b] School of Mathematics & Statistics, Lanzhou University, Tianshuinanlu 222, Lanzhou, China
[c] School of Science, Nanchang Institute of Technology, Nanchang 330099, JiangXi, China

ARTICLE INFO

Article history:
Received 18 April 2017
Received in revised form
15 August 2017
Accepted 18 August 2017
Available online xxx

Keywords:
Hybrid model
Air pollution indexes
Forecasting

ABSTRACT

Air quality is closely related to quality of life. Air pollution forecasting plays a vital role in air pollution warnings and controlling. However, it is difficult to attain accurate forecasts for air pollution indexes because the original data are non-stationary and chaotic. The existing forecasting methods, such as multiple linear models, autoregressive integrated moving average (ARIMA) and support vector regression (SVR), cannot fully capture the information from series of pollution indexes. Therefore, new effective techniques need to be proposed to forecast air pollution indexes. The main purpose of this research is to develop effective forecasting models for regional air quality indexes (AQI) to address the problems above and enhance forecasting accuracy. Therefore, two hybrid models (EMD-SVR-Hybrid and EMD-IMFs-Hybrid) are proposed to forecast AQI data. The main steps of the EMD-SVR-Hybrid model are as follows: the data preprocessing technique EMD (empirical mode decomposition) is utilized to sift the original AQI data to obtain one group of smoother IMFs (intrinsic mode functions) and a noise series, where the IMFs contain the important information (level, fluctuations and others) from the original AQI series. LS-SVR is applied to forecast the sum of the IMFs, and then, S-ARIMA (seasonal ARIMA) is employed to forecast the residual sequence of LS-SVR. In addition, EMD-IMFs-Hybrid first separately forecasts the IMFs via statistical models and sums the forecasting results of the IMFs as EMD-IMFs. Then, S-ARIMA is employed to forecast the residuals of EMD-IMFs. To certify the proposed hybrid model, AQI data from June 2014 to August 2015 collected from Xingtai in China are utilized as a test case to investigate the empirical research. In terms of some of the forecasting assessment measures, the AQI forecasting results of Xingtai show that the two proposed hybrid models are superior to ARIMA, SVR, GRNN, EMD-GRNN, Wavelet-GRNN and Wavelet-SVR. Therefore, the proposed hybrid models can be used as effective and simple tools for air pollution forecasting and warning as well as for management.

1. Introduction

Air pollution is an important environmental problem in many parts of the world (Kurt and Oktay, 2010). On the one hand, along with rapid development, the emission of substandard industrial pollutants is a major cause of severe air pollution in industrialized areas. On the other hand, there has been a dramatic increase in population coupled with rapid economic development. Atmospheric pollution in China is very serious and is mainly reflected as a high concentration of suspended particles in the urban atmospheric environment. For example, there is a high level of sulfur dioxide (SO_2) and particulate matter (PM), along with a rapid increase in automotive exhaust emissions, and an aggravating tendency in nitrogen oxides (NO_x) pollution. In recent years, air pollution has led to increasingly hazy weather in China.

The public has become increasingly concerned about the atmospheric environment because air quality affects everyone (Zhang et al., 2012). Furthermore, air pollution may have serious impacts on human health, including asthma, impaired lung function, cardiopulmonary illnesses (Yahya et al., 2014), obstacles to physiological functions, and increased mortality rates (Mindell and Joffe, 2004). Therefore, there are many suggestions to protect the population from heavy pollution, for example: i) the general public should reduce outdoor activities; ii) some vulnerable populations,

☆ This paper has been recommended for acceptance by Dr. Hageman Kimberly Jill.
* Corresponding author.
E-mail address: 819612640@qq.com (X. Lian).

http://dx.doi.org/10.1016/j.envpol.2017.08.069
0269-7491/

including children, pregnant women, the elderly, and patients with respiratory disease and cardiovascular diseases, should stay indoors as much as possible and stop outdoor activities; and iii) primary schools and kindergartens should reduce outdoor physical education and outdoor activities (Sheng and Tang, 2015). Therefore, air pollution forecasting plays a vital role in people's daily life, as well as in warning and controlling air pollution.

The air quality index (AQI) and $PM_{2.5}$ are two important indicators among pollution indexes, where $PM_{2.5}$ is particulate matter with a diameter of less than or equal to *2.5* μm. AQI is an indicator of air quality which reflects and evaluates the air quality status, which simplifies the concentrations of several pollutants into one single numerical form. The AQI is calculated with reference to the new ambient air quality standards (GB3095-2012), which covers six pollutants, including sulfur dioxide (SO_2), nitrogen dioxide (NO_2), $PM_{2.5}$, PM_{10} (particulate matter with a diameter less than or equal to 10 μm), ozone (O_3), and carbon monoxide (CO) (Sheng and Tang, 2015). There existed different definitions of AQI, such as fuzzy-based air quality index (FAQI, Sowlat et al., 2011) which defined the AQI by weighting the CO, PM_{10}, SO_2, NO_2, O_3 with fuzzy criterion, air pollution index (API, Yuan and Liu, 2014; Chen et al., 2016) which only covered CO, PM_{10}, SO_2, NO_2 and O_3. The popular pollution index AQI covers six pollutants ($PM_{2.5}$, CO, PM_{10}, SO_2, NO_2, O_3), so it is analyzed in this research. The definition of the used AQI is presented in the appendix (Yuan and Liu, 2014). Intuitively, it can be seen that the daily AQI time series shows the average pollution trend changes along with days. For the general public, the AQI is an important index that can be used to easily understand whether the air quality is bad or good. It is also helpful in data interpretation for decision making processes related to pollution mitigation measures and air quality management (Kumar and Goyal, 2011). According to air quality standards (GB3095-2012) and various impacts on human health, the AQI is divided into six classes (Fig. 1). Corresponding to the six classes of air quality, as the AQI increases, the level of pollution increases. In terms of different physical qualities and AQI, there are different suggestions for different people (Fig. 1).

Over the past few decades, air pollution warnings have caught the world's attention. Currently, many people make decisions on outdoor actions according to the air pollution forecasts to avoid the effects of atmospheric pollution on human health. Therefore, developing an accurate and effective AQI forecasting model is an important topic. The air and pollutants move in different ways and directions, mainly through natural causes and atmospheric phenomena (Kurt and Oktay, 2010). In fact, the atmosphere is an intricate dynamic system that is quite difficult to model (Kurt and Oktay, 2010). Therefore, forecasting the AQI or any other pollution index is not easy.

Faced with such a problem, some researchers have proposed techniques to forecast PM_{10}, $PM_{2.5}$ and other pollution indicators. According to the pattern of data processing, the related forecasting models can be classified into two classes: empirical models and chemical transport models (CTMs) (Cobourn, 2010). Konovalov et al. (2009) noted that chemical transport models are worse than empirical models for forecasting the air index PM_{10} because CTMs can simulate some components reasonably well, but cumulative errors from poorly modeled or missing components lead to relatively large errors in simulated $PM_{2.5}$ concentrations (Cobourn, 2010). Therefore, the forecasting models for AQI and other air pollution indexes should take full advantages of individual models or revise the results of CTMs according to the residual sequence of forecasting (Cobourn, 2010).

In terms of empirical models, statistical models are widely applied to forecast various air pollution indexes, for example, the ARIMA (autoregressive integrated moving average) model, multiple linear model (MLR), artificial neural networks (ANNs), support vector regressions (SVRs) and hybrid models. The ARIMA model is a classical statistical modeling technique for analyzing nonlinear

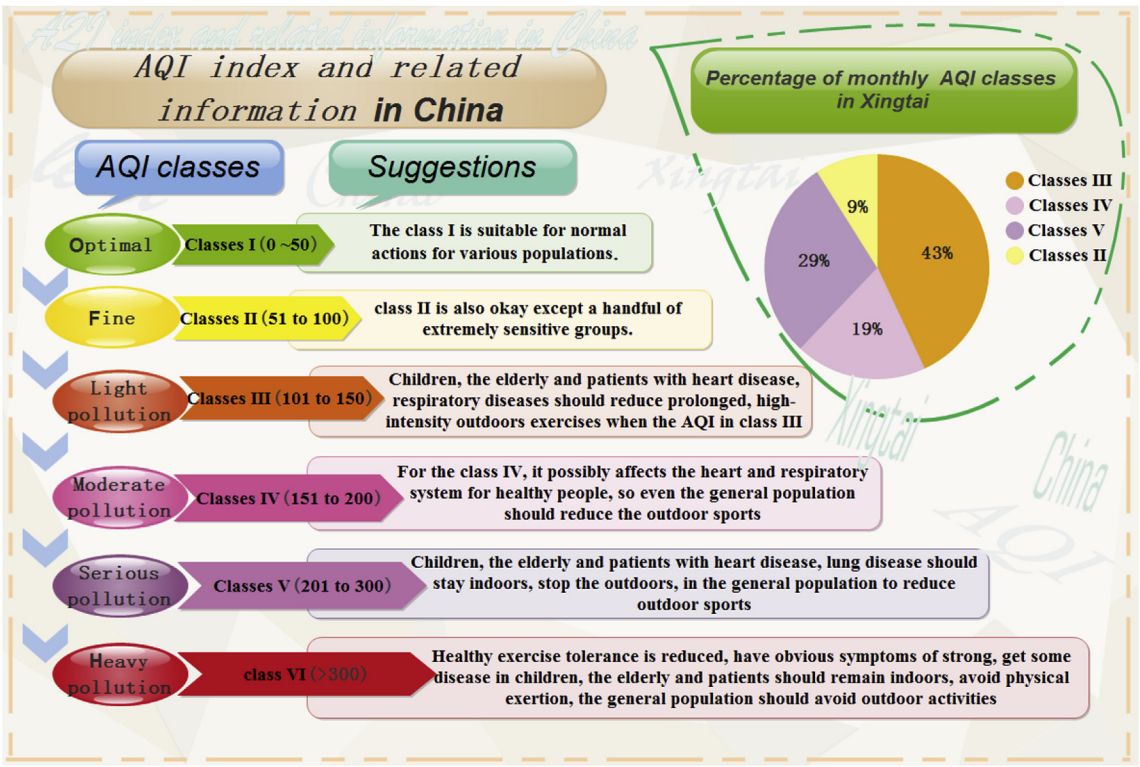

Fig. 1. The AQI index and related information in China.

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

time series data (Jian et al., 2012; Reikard, 2012; Slini et al., 2002; Song et al., 2015). Although multiple linear regression model may be somewhat dull compared to other approaches, the method remains useful and widely used in the prediction field due to its accuracy of interpretation (Zhou et al., 2014). For example, a MLR was used for air quality forecasting in Delhi (Goyal et al., 2006). Zhou et al. (2014) applied a MLR to forecast the one-day-ahead $PM_{2.5}$ levels based on meteorological data and air pollution indexes.

Generally, ANNs are capable of modeling non-linear relationships between input and output variables and are often used in forecasting variables in complex systems (Gardner and Dorling, 1999). ANNs are also widely applied to forecast air quality indexes (Díaz-Robles et al., 2008; Elangasinghe et al., 2014; Feng et al., 2015; Hooyberghs et al., 2005; Jiang et al., 2004; Kolehmainen et al., 2001; Ordieres et al., 2005; Pérez et al., 2000; Voukantsis et al., 2011; Fernando et al., 2012; Yi and Prybutok, 1996). However, an ANN is unable to present a clear formula for the forecasting model. SVRs are statistical models that are presented after ANNs and also employed to analyze pollution indexes (Ortiz-García et al., 2010; Yeganeh et al., 2012; Lu and Wang, 2008).

The popular ANNs and SVRs can perform better than traditional statistical models; nevertheless, in most realistic cases, the pollution indexes data are non-stationary or chaotic, and thus, application of only the ANN or SVR models rarely occurs. It is worth emphasizing that the hybrid model was successfully and widely employed in pollution index forecasting. Díaz-Robles et al. (2008) applied the hybrid model with ARIMA and artificial neural networks to forecast particulate matter in Temuco (Díaz-Robles et al., 2008), which showed that the hybrid model had the minimal MAE (mean absolute error) among the multiple linear model, ARIMAX (ARIMA with inputs), ANN and hybrid model. More specifically, the MAE values for the multiple linear model, ARIMAX, ANN and hybrid model are *20.83, 19.87, 20.65* and *6.74*, respectively. Many research papers use the hybrid model to forecast pollution (Wang et al., 2017a, b; Wang et al., 2015; Yeganeh et al., 2012; de Mattos Neto et al., 2014; Niu et al., 2016; Zhou et al., 2014).

From a review of the literature, the advantages of hybrid models are: first, hybrid models can make full use of various forecasting models; second, data preprocessing techniques are effective at information extraction and improving forecasting precision; and third, hybrid models can generally obtain more accurate prediction results.

Data preprocessing transforms or projects raw data into smaller spaces, including through wavelet transform and EMD (empirical mode decomposition) methods, which are helpful for modeling and forecasting. Wavelet transform decomposes the original sequence into components with different frequencies and amplitudes, but the different components generally depend on fixed frequency bands. It is not an easy task to select the appropriate frequency bands. Therefore, the essence of the wavelet transform is the band-pass filter, and the result of decomposition is worse than EMD. EMD, proposed by Huang et al. (1998), is an effective preprocessing technique for nonlinear and non-stationary time series data, which is essentially a tranquilization of a data sequence or signal processing. It is widely applied in various areas, such as forecasting TAIEX (Cheng and Wei, 2014), exploring Parkinson's disease (Rojas et al., 2013), conducting similarity analyses of DNA sequences (Zhang et al., 2011), performing geophysics research (Huang and Wu, 2008), and analyzing financial issues (Huang et al., 2003). Essentially, EMD is used to decompose signals or time series into a few intrinsic mode functions (IMFs) through a sifting process (Huang et al., 1998; Zhou et al., 2014). The IMFs contain important information in the original series, and some IMFs present tendency or seasonality, so they can be easily forecasted by statistical models.

For AQI forecasting, in terms of our knowledge, there are some research papers employed hybrid models (Yang and Wang, 2017; Wang et al., 2017a, b). However, EEMD-ANN hybrid models (Yang and Wang, 2017) and two-phase decomposition technique and modified extreme learning machine (Wang et al., 2017a) are without error revisions. Thus, this study proposes two hybrid models (EMD-SVR-Hybrid and EMD-IMFs-Hybrid) that are integrated with EMD and S-ARIMA to forecast the AQI, where S-ARIMA is employed to forecast the residual series from the two hybrid models. The main steps of the EMD-SVR-Hybrid are as follows: the data preprocessing technique EMD (empirical mode decomposition) is utilized to sift the original AQI data to obtain one group of smoother IMFs (intrinsic mode functions) and noise series, in which the IMFs contain the important information from the original AQI series. The SVR is applied to forecast the sum of the IMFs, and then, S-ARIMA is employed to revise the residual sequence. In addition, EMD-IMFs-Hybrid first separately forecasts the IMFs with statistical models and then models the residual data with S-ARIMA.

The characteristics of the proposed hybrid model and the aims of this study are as follows:

- The EMD is applied to decompose the AQI time series into one group of IMFs with better performance, i.e., $AQI = \sum_i IMF_i + noise$, where the size of the IMFs is selected according to original AQI data and the IMFs can be more easier modeled and forecasted using statistical theory.
- The $\sum_i IMF_i$ is modeled and forecasted by SVR, and the residual series between AQI and the forecasting data of $\sum_i IMF_i$ are revised by the hybrid model, that is, EMD-SVR-Hybrid. The forecasting results of EMD-SVR- Hybrid are the final forecasting results for AQI. Obviously, the EMD-SVR-Hybrid takes full advantages of EMD, SVR and S-ARIMA. In particular, the characteristics of EMD-SVR-Hybrid are as follows: i) EMD effectively sifts the effective information in the original AQI time series; ii) SVR can address forecasting with nonlinear trends well; and iii) the hybrid technique can capture the missing information of SVR compared with the original data.
- The selected IMFs from EMD are separately modeled and forecasted in terms of their unique properties, and the residual data between the sum of the IMF forecasting and original AQI data is forecasted by S-ARIMA, that is, EMD-IMFs-Hybrid model. The forecasting results of EMD-IMFs-Hybrid are the final forecasting results for AQI. Obviously, the EMD-IMFs-Hybrid model takes full advantage of EMD as well as the properties of IMFs and S-ARIMA. Characteristics i) and iii) of EMD-IMFs-Hybrid are same as those of the EMD-SVR-Hybrid model. The unique characteristic of EMD-IMFs-Hybrid is that the forecasting of IMFs involved optimal statistical models for each IMF, so it fully processes the information of IMFs.
- Based on the popular forecasting assessment measures, i.e., MAE and MAPE (mean absolute percentage error), the proposed EMD-IMFs-Hybrid and EMD-SVR-Hybrid models more accurately forecast AQI than ARIMA, SVR, Wavelet-SVR, Wavelet-GRNN and EMD-GRNN. Therefore, it can be used to make air pollution forecasting and warnings.

The paper is arranged as follows. Section 2 gives a simple introduction to the study city and data set. General descriptions of several models are given in Section 3 and 4. The forecasting performance and end-point analysis are presented in Section 5 and 6, respectively.

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

2. Study area and data set

Xingtai is located in the south of Hebei province in China, located at north latitude 36°50′~37°47′, east longitude 113°52′~115°49′ (Fig. 2), which is close to Beijing. There are several major factors that lead to the poor air quality in Xingtai. First, its terrain is low-lying and the wind speed is low. Second, fog is common and the area is large. Third, hundreds of coal enterprises distributed in Xingtai lead to waste and serious pollution emissions. Therefore, the geographic location severely affects the air quality. Additionally, China's cities with the worst air pollution in 2014, the air quality in Xingtai ranked second. In particular, the monthly distribution of the AQI from August 2014 to January 2015 showed that class II was *9%*, class III was *43%*, class IV was *19%* and class V was *29%* (Fig. 1). The data are provided by web sites at http://113.108.142.147:20035/emcpublish/ and http://www.aqistudy.cn/historydata/index.php. These data show that Xingtai is a heavily polluted city in China, so we select Xingtai to model and analyze the AQI. In this study, the data set of daily AQI is partitioned into two segments: the training set (*06/01/2014-07/23/2015*) and validation set (*07/24/2015-08/22/2015*). There are *415* data points in the training set: the mean is *140.1807*, minimum is *35*, maximum is *500* and standard deviation is *75.46973*. There are *30* data points in the testing set: the mean is *118.6207*, minimum is *63*, maximum is *192* and standard deviation is *31.55088*.

3. Air quality forecast techniques

3.1. Seasonal ARIMA model

The seasonal ARIMA model (S-ARIMA) was proposed by Box and Jenkins (1976) and has been widely used in various forecasting domains. For example, Chaudhuri and Dutta (2014) adopted the ARIMA model to predict the concentrations of NO_2, CO, SO_2, O_3 and PM_{10}. In this research, the ARIMA model is employed to model the AQI data and S-ARIMA models are employed to forecast IMF_5 as well as analyze the residual sequences of EMD-SVR-Hybrid and EMD-IMFs-Hybrid.

The S-ARIMA model is designed for a time series with seasonal changes. The ARIMA model is the special case of the *S-ARIMA(p,d,q)(P,D,Q)* model, where p is the order of autoregressive, q is the order of moving average polynomials, d is the order of regular differences, D is the number of seasonal differences, and P and Q are the seasonal autoregressive and moving average orders, respectively. Obviously, the *ARIMA(p,d,q)* model is the case of *S-ARIMA(p,d,q)(0,0,0)*. For one time series with seasonal changes, we should obtain the seasonal difference to obtain one stationary series. The other steps are similar to the ARIMA model.

3.2. Holt-Winters additive model

The Holt-Winters model is one technique for smoothing data, i.e., an exponential smoothing data analysis tool. The basic principle of exponential smoothing is the nearer to historical data, the greater impact on the future. The Holt-Winters additive model is used for forecasting future data based on estimating the trend and season factors. The formula of the Holt-Winters additive model is $\hat{y}_{t+m} = L_t + b_t{*}m + I_{t-l+m}$, where $\hat{y}_{t+m}$ is the forecasted data for time $t + m$ based on the data before time $t + 1$, L_t is the level, b_t is trend slope, and I_{t-l+m} is the seasonal component with season length l. Furthermore, formulas for the three components are $L_t = \alpha(y_t - I_{t-l}) + (1-\alpha)(L_{t-1} + b_{t-1})$, $b_t = \beta(L_t - L_{t-1}) + (1-\beta){*}b_{t-1}$ and $I_t = \delta(y_t - L_t) + (1-\delta){*}I_{t-l}$, where α, β and δ are smoothing parameters for the level, trend and season, respectively.

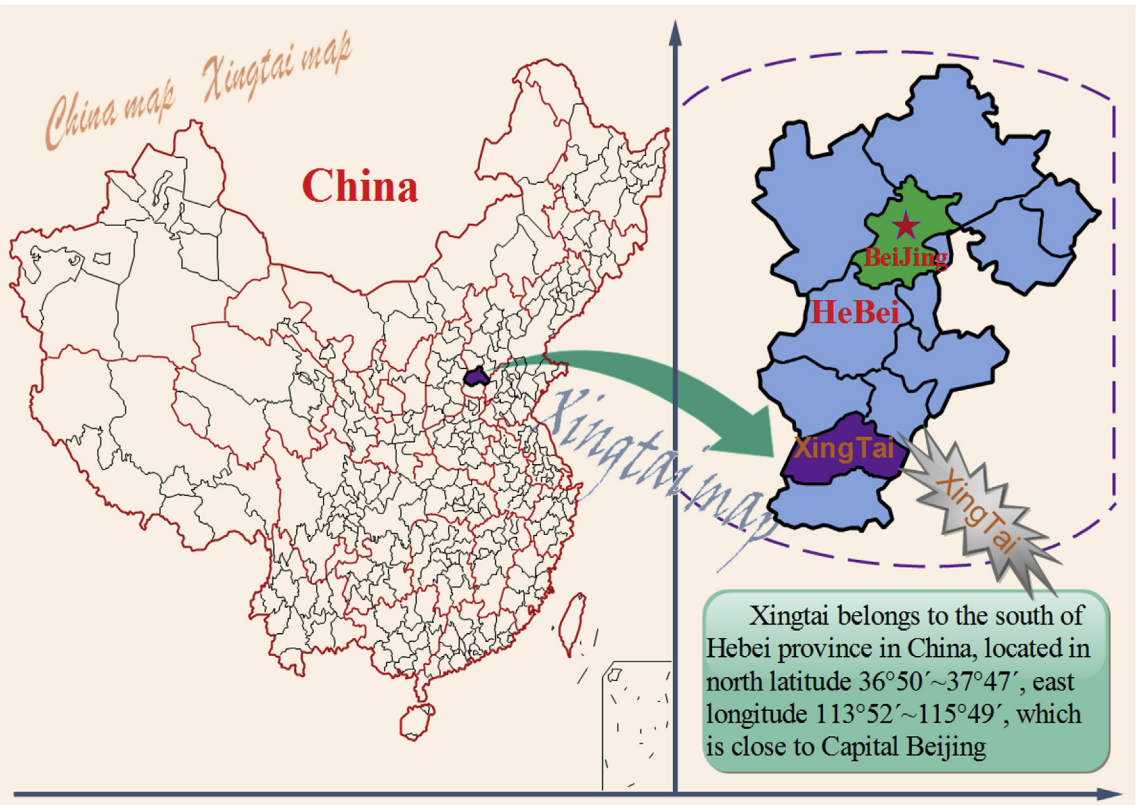

Fig. 2. Geographic location of Xingtai in China.

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

The initial values for the level, trend and season are $L_l = \frac{y_1+y_2+\ldots+y_l}{l}$, $b_l = \frac{y_{l+1}-y_1+y_{l+2}-y_2+\ldots+y_{l+l}-y_l}{l^2}$ and $I_i = y_i - L_l (i = 1, 2, \ldots, l)$, respectively.

The Holt-Winters model has been used in different forecasting domains. For example, Tratar and Strmčnik (2016) applied the Holt-Winters method to monthly short-term heat load forecasting. In this paper, it is employed to forecast IMF_4 and IMF_6.

3.3. GM model

The GM (Grey model) forecasting model refers to a gray model or grey model (Tien, 2009). The core of the GM is to establish differential equations. *GM(1,1)* is the most commonly used gray model, which is composed of a first-order differential equation with a single variable. Thereby, the establishment of *GM(1,1)* only needs one series y_s $(s = 1, \ldots, n)$. In this section, let $x^0 = \{x_s^0 = y_s | s = 1, 2, \ldots, n\}$. The constructional steps of the GM(1,1) model are as follows: Suppose the original data sequence to be $x^0 = (x_1^0, x_2^0, \ldots, x_n^0)(x_i^0 > 0, i = 1, \ldots, n)$, and accumulating the original sequence once yields the *1-AGO* (accumulated generating operation) sequence $x^1 = (x_1^1, x_2^1, \ldots, x_n^1)$. *IAGO* (inverse accumulated generating operation) is used to calculate the predicted value of GM, which is defined as (Wang et al., 2005): $\begin{cases} \widehat{x}_1^0 = \widehat{x}_1^1 = \widehat{x}_1^0 \\ \widehat{x}_k^0 = \widehat{x}_k^1 - \widehat{x}_{k-1}^1 \ (k = 2, 3, \ldots, n) \end{cases}$, where $\widehat{x}_k^0$ $(k = 1, 2, \ldots, n)$ is the forecasted data of GM. Obviously, if *IAGO* is applied to the sequence x^1, we will obtain the original sequence x^0.

GM(1,1) is defined as (Wang et al., 2005):

$$\frac{dx_t^1}{dt} + ax_t^1 = b,$$

where the gray development coefficient a and gray control coefficient b are the parameters to be estimated. The parameters of *GM(1,1)* according to the least squares method are $(a, b)^T = (B^T B)^{-1} B^T I$, where $I = (x_2^0, \ldots, x_n^0)^T$ and $B = -\frac{1}{2}\begin{pmatrix} x_1^1 + x_2^1 & -2 \\ x_2^1 + x_3^1 & -2 \\ \cdots\cdots & \\ x_{n-1}^1 + x_n^1 & -2 \end{pmatrix}$. Therefore, the solution to equation *GM(1,1)* is $\widehat{x}_t^1 = \left(x_1^0 - \frac{\widehat{b}}{\widehat{a}}\right) e^{-\widehat{a}\,(t-1)} + \frac{\widehat{b}}{\widehat{a}}$, which can be written in the discrete form $\widehat{x}_{k+1}^1 = \left(x_1^0 - \frac{\widehat{b}}{\widehat{a}}\right) e^{-\widehat{a}k} + \frac{\widehat{b}}{\widehat{a}}$ $(k = 1, 2, \ldots)$. This is a specific formulation of the *GM(1,1)* model to predict future data. Let $\widehat{x}_0^1 = 0$, and the forecasting formula with *IAGO* is:

$$\widehat{x}_k^0 = \widehat{x}_k^1 - \widehat{x}_{k-1}^1 = \left(x_1^0 - \frac{\widehat{b}}{\widehat{a}}\right)\left(1 - e^{\widehat{a}}\right) e^{-\widehat{a}(k-1)}, \quad k = 1, 2, 3, \ldots.$$

The GM model has been used in different forecasting domains. In this paper, it is employed to forecast IMF_7.

3.4. LS-SVR based on state space construction

Support vector machines are effective tools for small sample classification and regression problems. LS-SVR (least squares support vector regression machines) is a tool that is simply applied to the SVR for prediction, and it is described in Qin et al. (2010). In this study, we apply the SVR to analyze AQI and IMFs time series data with the state space construction.

State space construction: Assume that $Y_t = \{y_t | t = 1, 2, \ldots N, N+1, \ldots, N^*\}$ is the time series; then, the state space of Y_t is:

$$X_{M*p} = \begin{bmatrix} y_1 & y_2 & \cdots & y_p \\ y_2 & y_3 & \cdots & y_{p+1} \\ \vdots & \vdots & \cdots & \vdots \\ y_M & y_{M+1} & \cdots & y_{p*M} \end{bmatrix},$$

where X_{M*p} is composed of the historical time series data and $p*M = N$. Obviously, there are p columns in the matrix X_{M*p}, i.e., p predictors. If we renew one data for the series, the matrix can be rewritten as:

$$X'_{M*p} = \begin{bmatrix} y_2 & y_3 & \cdots & y_{p+1} \\ y_3 & y_4 & \cdots & y_{p+2} \\ \vdots & \vdots & \cdots & \vdots \\ y_{M+1} & y_{M+2} & \cdots & y_{M*p+1} \end{bmatrix}.$$

Therefore, the time series model can be modeled by multiple regression with LS-SVR. For example, to obtain the one-ahead forecasted data at time $N^* + 1$ from the SVR, we only need to establish the relationship between X_{N^*-1} and the response vector $Y_{N^*-1} = (y_{N^*-M+1}, y_{N^*-M+2}, \ldots, y_{N^*})$. For the *h-ahead* forecasted data, the formula is similar to the idea of the one-ahead forecasting process. For simplicity, the pairs of X and Y are abbreviated as $\{(X_i, Y_i) | i = 1, 2, \ldots, \}$.

4. The hybrid models based on EMD

4.1. EMD

EMD is based on the following assumptions: first, there are at least two extreme values, a maximum and a minimum; second, the local time domain characteristics for the data are uniquely determined by the time scale between the extreme points; and third, if there are no extreme point in the series, but there is an inflection point, we can find the extreme value for the differential series and then obtain the decomposed results by integration. The purpose of the EMD algorithm is to decompose the signal into IMFs, and the IMFs must meet the following two conditions: i) in the total dataset, the number of zero-crossings and extreme points (maximum or minimum) must be equal or differ at most by one; ii) the mean value, which is composed of local maxima and local minima envelopes, is zero at any point. The calculation steps of the EMD algorithm for a signal *s(t)* are as follows (Huang et al., 1998):

Step 0: Let $s_0(t) = s(t)$, where $s(t)$ is the original AQI time series.
Step 1: Calculate all of the local extreme points of signal $s_0(t)$.
Step 2: Link all of the local maximal points with a cubic spline to form the upper envelope $u_0(t)$. Similarly, all of the minimal points develop the lower envelope $v_0(t)$.
Step 3: The mean value of the upper and lower envelope is $m_0(t) = \frac{u_0(t) + v_0(t)}{2}$. Compute the difference between the signal $s_0(t)$ and $m_0(t)$: $h_0(t) = s_0(t) - m_0(t)$.
Step 4: Judge whether $h_0(t)$ meets the two conditions of IMFs. If satisfied, $c_1(t) = h_0(t)$ is the first IMF; otherwise, let $s_0(t) = h_0(t)$ and then repeat step 1 - step 3.
Step 5: Remember $r_1(t) = s_0(t) - c_1(t)$ as a new standby signal.
Step 6: If $r_1(t)$ has at least two extreme values (Zhou et al., 2014), let $s_0(t) = r_1(t)$, and then repeat steps 1 to 5 and obtain the second IMF $c_2(t)$.
Step 7: Similarly, we repeat steps 1–6 n times, and then obtain the IMFs, $c_1(t), \ldots, c_n(t)$. Therefore, the original signal can be

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

expressed as $AQI = \sum_i IMF_i + noise$, where $\sum_i IMF_i = \sum_{k=1}^{n} c_k(t)$ and $noise = s_0(t) - \sum_i IMF_i$.

4.2. EMD-SVR-hybrid model

According to the theory of EMD, we remove the IMFs with the high frequency noted as noise and then sum the remaining IMFs noted as $D_t = \sum_i IMF_i$. Then, LS-SVR is employed to forecast D_t, which is called EMD-SVR. The residual sequence between EMD-SVR and AQI is modeled by S-ARIMA in terms of its characteristics. The sum of EMD-SVR and S-ARIMA is the AQI forecasting, which is noted as EMD-SVR-Hybrid, and the procedures of EMD-SVR-Hybrid are shown in Fig. 3. It is worth noting that the final step of the EMD-SVR-Hybrid is a model test. If the proposed model passes the test, the forecasting results are output.

4.3. EMD-IMFs-hybrid model

The IMFs from EMD are separately modeled and forecasted by different statistical models in terms of their unique properties, and the sum of IMFs forecasting results are noted as EMD-IMFs. There also may be a large difference between EMD-IMFs and the actual AQI data. Therefore, S-ARIMA is applied to forecast the residuals from EMD-IMFs, which is named the EMD-IMFs-Hybrid, and the procedures are shown in Fig. 3. The forecasting result of AQI is the sum of EMD-IMFs and S-ARIMA. In our research, the IMF_4, IMF_5, IMF_6 and IMF_7 are selected from IMFs, and they are, respectively, forecasted by *Holt-Winters(0.9,0.2,0.3)*, *S-ARIMA(1,1,1)(0,1,1)*, *Holt-Winters(0.1,0.2,0.5)* and *GM(1,1)*, and the residual data are modeled by S-ARIMA, as referenced in Fig. 3.

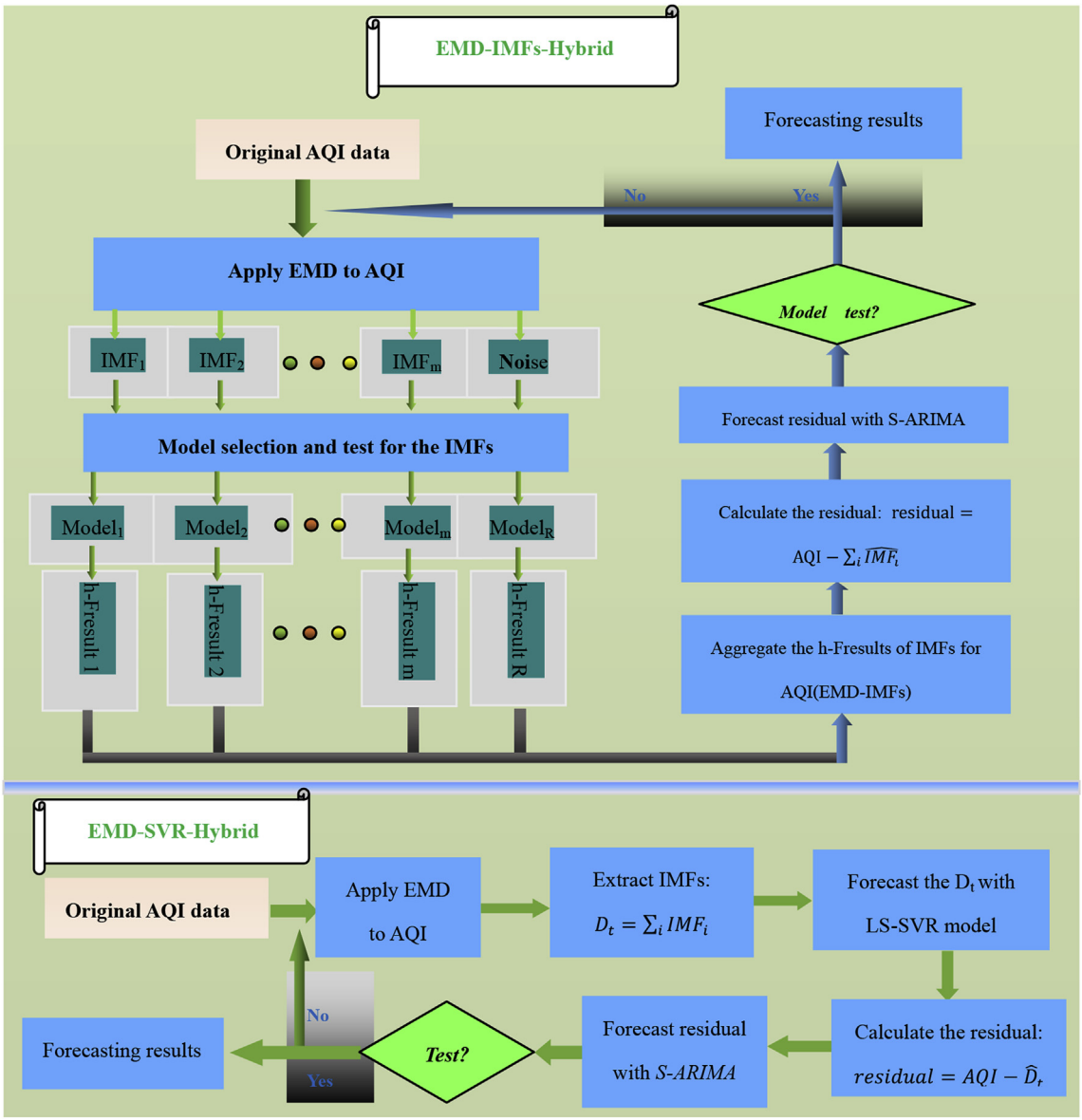

Fig. 3. The procedures of EMD-SVR-Hybrid and EMD-IMFs-Hybrid (*h-Fresult* is *h-ahead* forecasting result).

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

5. Data analysis

5.1. Statistical measures for forecasting performance

There are many error evaluation metrics to assess the accuracy of air pollution forecasting models, for example, MSE (Niu et al., 2016), MAPE (Bai et al., 2016), MAE (Feng et al., 2015), and RMSE (Feng et al., 2015). However, no uniform standard method has been identified as the best evaluation criterion. In this study, we applied the popular error measures MAE, RMSE, MAPE, ARE (absolute relative error) and index of agreement (IA) to compare the forecasting ability of different models. The principles of MAE, RMSE MAPE, ARE and IA are, respectively, defined as: $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|$, $RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2/n}$, $MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\widehat{y}_i|}{|y_i|} \times 100\%$, $ARE = \frac{|y_i-\widehat{y}_i|}{|y_i|} \times 100\%$, and $IA = 1 - \frac{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{\sum_{i=1}^{n}(|\widehat{y}_i-\overline{y}|+|y_i-\overline{y}|)^2} \times 100\%$, where y_i is the true data, $\widehat{y}_i$ is the predicted data, $\overline{y}$ is the mean of true data and n is forecasting data size.

To present the forecasting comparisons clearly, we introduce the accuracy improvement (AI) index in this study. AI (Zhu et al., 2011) is defined as: $AI_p = \frac{S-S_p}{S} \times 100\%$, where S and s_p are error measures for the fixed model and specified model, respectively. It can be concluded that if $AI_p > 0$, the specified forecasting model does better, and if $AI_p < 0$, the specified model does not beat the fixed model.

5.2. The model descriptions

According to the characteristics of IMFs, we separately select the optimal models to forecast the IMFs (IMF_4, IMF_5, IMF_6, IMF_7; Fig. 4). For example, IMF_5 shows seasonal fluctuation, while *S-ARIMA(1,1,1)(0,1,1)* is employed to forecast it. Holter-Winters is applied to forecast IMF_4, which shows the level and seasonality. Detailed information of optimal models is provided in Table 1.

5.3. Forecast comparisons

The actual and forecasted values are listed in Table 2. Roughly, the experimental results show that the error of the ARIMA model is very high, followed by the SVR model. The model *ARIMA(5,2,3)* is applied to forecast the original AQI data. The forecasting results vary greatly from the actual data. The MAPE of ARIMA is *186.4%*, so it is worse than any expert's guess, which illustrates the failure of the forecasting of the traditional ARIMA model. Data processing, which has two times the difference and three times the moving average, leads to a large discrepancy between the processed data and original data. It also leads to a partial loss of information from the original data. This also shows the complex properties of the original AQI data. The prediction ability of the SVR model is better compared with ARIMA, whose MAPE is only *30.3%*. In Table 2, the extreme pollutions are effectively forecasted by the two proposed hybrid models, which are marked in bold.

The different forecasting performances of different models are listed in Table 3. Obviously, we can clearly observe that the model EMD-SVR-Hybrid is better than ARIMA and SVR without EMD and error revision. Obviously, the three models (Wavelet-SVR, Wavelet-GRNN and EMD-GRNN) with data preprocessing make better progress compared with ARIMA. The rankings of the seven models are as follows: EMD-IMFs-Hybrid, EMD-SVR-Hybrid, EMD-GRNN, Wavelet-GRNN, GRNN, SVR, Wavelet-SVR and ARIMA (in terms of MAE); EMD-SVR-Hybrid, EMD-IMFs-Hybrid, EMD-GRNN, Wavelet-GRNN, GRNN, SVR, Wavelet-SVR and ARIMA (in terms of RMSE); EMD-IMFs-Hybrid, EMD-SVR-Hybrid, EMD-GRNN, Wavelet-GRNN, GRNN, SVR, Wavelet-SVR and ARIMA (in terms of MAPE); and EMD-SVR-Hybrid, EMD-IMFs-Hybrid, EMD-GRNN, Wavelet-GRNN, GRNN, Wavelet-SVR, SVR and ARIMA (in terms of IA). That is, different forecasting precision indexes lead to different ranks for forecasting models. Generally, EMD-SVR-Hybrid and EMD-IMFs-Hybrid models are superior the other models. MAPE is the most important precision index for forecasting assessment. It is worth mentioning that the proposed EMD-IMFs-Hybrid has the best MAPE precision among the eight models. Obviously, the two hybrid models are able to forecast the extreme AQIs which are marked in bold.

Table 4 shows the contingency table for the forecasted and observed AQI levels. The AQI class is categorized into four groups: class *I* and class *II* (as the first group), class *III* (as the second group), class *IV* (as the third group), and class *V* and class *IV* (as the fourth group). *F%* represents the percentage of the observed days in each class that were forecasted to be in that class (Zhou et al., 2014), which is listed in the last column in Table 4. *T%* provides the percentage of the forecasted days in each class that actually occurred (Zhou et al., 2014). In Table 4, the diagonal numbers in bold correspond to successful classifications (Zhou et al., 2014), which means that the level of the forecasting value is equal to the level of the observed value. Table 4 reveals that the overall correct rate of the EMD-SVR-Hybrid model is *80%*, followed by the EMD-IMFs-Hybrid (*70%*), EMD-GRNN (*63.33%*), GRNN (*63.33%*), Wavelet-GRNN (*60%*), SVR (*30%*), Wavelet-SVR (*16.67%*) and ARIMA (*0%*). From above information, the EMD-IMFs-Hybrid and EMD-SVR-Hybrid models have higher performances than the other six models. Moreover, the ARIMA model has the worst performance, with an overall correct classification rate of *0%*, and the SVR with wavelet decomposition did not perform better than the single SVR. EMD-GRNN performs better than Wavelet-GRNN, which shows the EMD is more effective than wavelet in AQI information extraction and forecasting. In this study, the EMD-IMFs-Hybrid and EMD-SVR-Hybrid models can recognize class IV (the third group) with a probability *100%*, which proved that the two proposed hybrid models effectively forecast extreme pollution. The correct forecasting of AQI levels is very significant for air pollution warnings, people's health and outdoor plans.

5.4. Model test

The desired result, in theory, is for the forecasting data to be equivalent to the actual data, i.e., the forecasting data and the real data should be in line, with a slope of *1* in the two-dimensional coordinate system. However, there are large or small differences between forecasting results and actual data. Therefore, we tested the eight models with the above idea. For the eight models, we built linear models without intercepts in the software R (Fig. 5). For the eight models: $ARIMA = 2.5011 \times AQI$ with $SD = 0.1478$ and $t = 16.92 (p < 2 \times 10^{-16})$ (Fig. 5a), $SVR = 0.78217 \times AQI$ with $SD = 0.05045$ and $t = 15.5 (p < 1.43 \times 10^{-15})$ (Fig. 5b), $Wavelet - SVR = 0.77175 \times AQI$ with $SD = 0.04868$ and $t = 15.86 (p < 7.97 \times 10^{-16})$ (Fig. 5c) $GRNN = 0.97118 \times AQI$ with $SD = 0.05437$ and $t = 17.86 (p < 2 \times 10^{-16})$ (Fig. 5d), $Wavelet - GRNN = 0.9731 \times AQI$ with $SD = 0.0498$ and $t = 19.54 (p < 2 \times 10^{-16})$ (Fig. 5e), $EMD - GRNN = 0.98523 \times AQI$ with $SD = 0.04182$ and $t = 23.56 (p < 2 \times 10^{-16})$ (Fig. 5f), $EMD{-}IMFs{-}Hybrid = 0.96547 \times AQI$ with $SD = 0.03847$ and $t = 25.1 (p < 2 \times 10^{-16})$ (Fig. 5g) and $EMD{-}SVR{-}Hybrid = 0.94239 \times AQI$ with $SD = 0.03539$ and $t = 26.63 (p < 2 \times 10^{-16})$ (Fig. 5h). The T tests for the four linear models are significant, so the linear models are rational from a statistical viewpoint. Generally, the coefficients for GRNN, Wave-GRNN, EMD-GRNN,

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

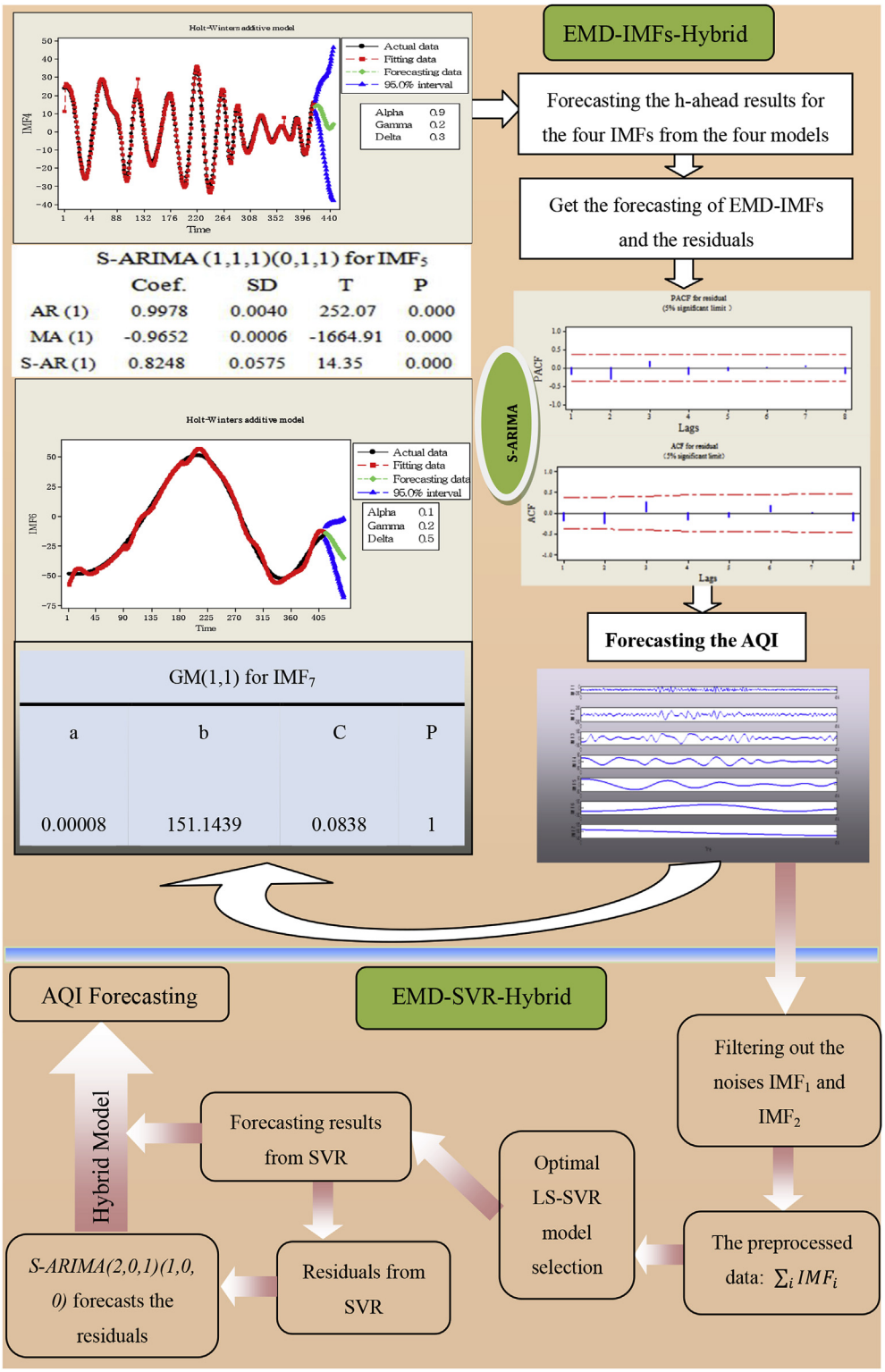

Fig. 4. The models' parameters and tests.

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

Table 1
Optimal individual models descriptions.

Hybrid models	Data	Optimal model descriptions
EMD-IMFs-Hybrid	IMF4	Holt-Winters additive model is employed for modeling IMF_4 data, where the parameters are $\alpha = 0.9$, $\beta = 0.2$ and $\delta = 0.3$, the fitting data and the forecasting data are showed in Fig. 4 (Fig. 4). From the figure we can see that the Holt-Winters model presents well fitting for the IMF_4 data series.
	IMF_5	*S-ARIMA(1,1,1)(0,1,1)* without constant is employed to fit the IMF_5 data, where the parameters are showed in Fig. 4 (Fig. 4). The *T* test proved *S-ARIMA(1,1,1)(0,1,1)* is significant. The parameter for the AR is *0.9978* with stand deviation (*SD*) $SD = 0.0040$ and $P = 0$, the parameter for the MA is *-0.9652* with $SD = 0.0006$ and $P = 0$, while the S-MA is *0.8248* with $SD = 0.0575$ and $P = 0$. Therefore, the *S-ARIMA(1,1,1)(0,1,1)* is effective to forecast the IMF_5.
	IMF_6	According to the characters of IMF_6, we employ the Holt-Winters additive model to fit it, where the parameters respective are $\alpha = 0.1$, $\beta = 0.2$ and $\delta = 0.5$, the fitting data and the forecasting data are showed in the Fig. 4(Fig. 4). From Fig. 4, we can see that the Holt-Winters model presents well fitting to IMF_6.
	IMF_7	The *GM(1,1)* is applied to fit the IMF_7 which shows declining trend. According to the forecasting results and the posterior-variance-test of the *GM(1,1)*, we can say loudly that *GM(1,1)* is suitable for forecasting the IMF_7. Furthermore, $a = 0.00008$, $b = 151.1439$, $C = 0.0838$ and $P = 1$ for the established *GM(1,1)*, so it belongs to the first class in model classifications, the GM classifications referred in Appendix.
	Residual	After carefully research, *S-ARIMA(1,1,0)(1,0,0)* is selected for analyzing the residual data, the ACF(auto-correlation function) and PACF(partial correlation function) are presented in Fig. 4(Fig. 4). Obviously, the model *S-ARIMA(1,1,0)(1,0,0)* with parameters *AR(1) -0.468* and *S-AR(1) -0.9176* is reasonable for the residual data. Furthermore, the *T* tests with $P = 0.011$ and $P = 0.00$ for the *AR(1)* and *S-AR(1)* are significant for amending the residuals.
EMD-SVR-Hybrid	D_t	After removing the noise, we model the $\sum_i IMF_i$ with LS-SVR theory. For the selected LS-SVR, hyper-parameter $\gamma = 96.6$.
	Residual	*S-ARIMA(2,0,1)(1,0,0)* is applied to revise the forecasting results of the EMD-SVR, where the parameters for *AR(1)* and *AR(2)* respective are *0.6482* and *0.0435*, *S-AR* is *-0.8351* and *MA* is *0.5495*.

Table 2
Actual and forecasting results for AQI.

Date	AQI	ARIMA		SVR		Wavelet-SVR		GRNN		Wavelet-GRNN		EMD-GRNN		EMD-IMFs-Hybrid		EMD-SVR-Hybrid	
		Forecast	ARE	Forecast	ARE	Forecast	ARE	Forecast	ARE	Forecast	ARE	Forecast	ARE	Forecast	ARE	Forecast	ARE
07–24	139	210	51.1	126	9.4	100	28.4	167	20.2	135	3.2	153	9.7	139	0	136	2.2
07–25	133	230	72.9	127	4.5	99	25.4	153	15.0	155	16.6	147	10.2	133	0	133	0
07–26	**176**	**242**	**37.5**	**96**	**45.5**	**138**	**21.7**	121	**31.0**	**149**	15.4	**141**	20.0	**170**	**3.4**	**163**	**7.4**
07–27	130	243	86.9	106	18.5	156	20.2	133	2.1	170	30.7	136	4.8	134	3.1	132	1.5
07–28	137	250	82.5	141	2.9	143	4.3	142	4.0	142	3.8	133	2.6	139	1.5	136	0.7
07–29	102	257	152	157	53.9	103	1.0	137	34.2	140	36.9	131	28.7	107	4.9	110	7.8
07–30	118	266	125.4	130	10.2	98	17.2	103	12.7	120	1.7	130	9.8	118	0	119	0.8
07–31	70	273	290	89	27.1	110	57.7	112	60.3	129	84.7	128	82.6	77	10	83	18.6
08–1	82	279	240.2	93	13.4	113	37.2	90	10.1	114	38.5	125	52.6	82	0	89	8.5
08–2	**192**	**284**	**47.9**	**112**	**41.7**	**109**	**43.3**	109	**43.4**	**85**	56.0	**121**	37.0	**175**	**8.9**	**163**	**15.1**
08–3	69	291	321.7	107	55.1	117	69.9	183	164.7	114	64.7	115	66.5	75	8.7	78	13
08–4	63	297	371.4	101	60.3	119	88.1	108	71.5	107	69.6	107	70.4	82	30.2	90	42.9
08–5	85	303	256.5	114	34.1	98	15.3	122	43.8	97	14.7	101	18.3	60	29.4	78	8.2
08–6	75	309	312	115	53.3	83	11.1	107	42.1	104	38.4	96	28.2	74	1.3	83	10.7
08–7	107	315	194.4	105	1.9	93	13.0	103	4.1	100	6.3	95	11.0	39	63.6	47	56.1
08–8	124	321	158.9	98	21	91	26.4	117	6.0	106	14.3	98	20.6	112	9.7	93	25
08–9	127	327	157.5	98	22.8	94	25.7	123	3.4	109	13.8	106	16.9	125	1.6	93	26.8
08–10	134	334	149.3	87	35.1	98	27.0	121	9.4	113	15.4	117	12.4	150	11.9	126	6
08–11	133	340	155.6	88	33.8	80	40.2	119	10.3	121	9.2	132	0.8	127	4.5	113	15
08–12	**162**	**346**	**113.6**	**80**	**50.6**	**67**	**58.9**	**120**	**25.9**	**124**	**23.3**	**140**	13.3	**166**	**2.5**	**153**	**5.6**
08–13	145	352	142.8	65	55.2	74	49.2	125	13.5	137	5.3	140	3.5	153	5.5	142	2.1
08–14	109	358	228.4	64	41.3	83	23.5	145	33.4	146	33.7	135	23.6	47	56.9	48	56
08–15	117	365	212	76	35	85	27.3	133	13.9	140	20.1	129	10.0	158	35	148	26.5
08–16	104	371	256.7	89	14.4	90	13.5	110	5.4	129	23.7	123	18.2	160	53.8	152	46.2
08–17	149	377	153	92	38.3	78	47.8	114	23.4	115	22.8	119	19.9	137	8.1	134	10.1
08–18	125	384	207.2	97	22.4	78	37.4	125	0.2	116	7.3	119	4.6	144	15.2	143	14.4
08–19	95	390	310.5	83	12.6	85	10.7	118	24.5	113	19.1	121	27.0	112	17.9	118	24.2
08–20	134	397	196.3	81	39.6	86	35.5	105	21.6	113	15.5	121	9.3	94	29.9	106	20.9
08–21	104	403	287.5	85	18.3	88	15.5	117	12.5	118	13.3	120	15.8	90	13.5	106	1.9
08–22	128	410	220.3	82	35.9	84	34.3	109	14.5	116	9.3	118	7.5	82	35.9	104	18.8
MAPE			186.4		30.3		30.9		25.9		24.2		21.86		**15.6**		16.4

EMD-IMFs-Hybrid and EMD-SVR-Hybrid are bigger than *0.9*, which demonstrate the above five models are rational for AQI forecasting. Specially, GRNN shows its potential application in AQI forecasting.

6. Conclusion

Forecasting air pollution is significant for preventing contamination or maximally reducing pollution incident harm to forecast air pollution, and it plays a vital role in air pollution warnings and controlling. Pollution indexes series are non-stationary and chaotic, so it difficult to attain accurate forecasting for air pollution indexes. Faced with such a difficult problem, this paper proposes two hybrid models, EMD-IMFs-Hybrid and EMD-SVR-Hybrid, for AQI forecasting.

AQI data from Xingtai, China collected from June 2014 to August 2015 are utilized to test the effectiveness of the proposed hybrid models. In terms of the forecasting error, the two proposed

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

Table 3
Forecasting precision indexes.

Model	MAE	AI_{MAE} (%)	RMSE	AI_{RMSE} (%)	MAPE (%)	AI_{MAPE} (%)	IA
ARIMA	198.545	−996.900	208.6240	−752.900	186.400	−1035.300	0.7761
SVR	35.812	−97.900	42.7060	−74.600	30.200	−84.100	0.9740
Wavelet-SVR	36.2368	−100.190	42.7388	−74.710	30.890	−88.350	0.9743
GRNN	26.7732	−47.9101	36.0930	−47.54722	25.91	−57.9878	0.9797
Wavelet-GRNN	25.9814	−43.53572	33.0731	−35.20195	24.24	−47.80488	0.9832
EMD-GRNN	22.3637	−23.54953	27.6400	−12.99158	21.86	−33.29268	0.9880
EMD-IMFs-Hybrid	**17.240**	**4.800**	25.767	−5.300	**15.600**	**4.800**	0.9910
EMD-SVR-Hybrid	18.101		**24.462**		16.400		**0.9915**

Table 4
Contingency table for the forecasted and observed AQI levels.

Models	Forecasted	Observed					
		I and II	III	IV	VandVI	Total	F%
ARIMA	I and II	**0**	0	0	0	0	–
	III	0	**0**	0	0	0	–
	IV	0	0	**0**	0	0	–
	VandVI	7	20	3	**0**	30	0%
	Total	7	20	3	0	30	
	T%	0%	0%	0%	–		**0%**
SVR	I and II	**3**	13	2	0	18	16.67%
	III	4	**6**	1	0	11	54.55%
	IV	0	1	**0**	0	1	0%
	VandVI	0	0	0	**0**	0	–
	Total	7	20	3	0	30	
	T %	42.86%	30%	0	–		**30%**
Wavelet-SVR	I and II	**3**	17	1	0	21	14.29%
	III	4	**2**	2	0	8	25%
	IV	0	1	**0**	0	1	0%
	VandVI	0	0	0	**0**	0	–
	Total	7	20	3	0	30	
	T %	42.86%	10%	0	–		**16.67%**
GRNN	I and II	**1**	0	0	0	1	100%
	III	5	**18**	3	0	26	69.23%
	IV	1	2	**0**	0	3	0%
	VandVI	0	0	0	**0**	0	–
	Total	7	20	3	0	30	
	T %	14.29%	90%	0	–		**63.33%**
Wavelet-GRNN	I and II	**1**	1	1	0	3	33.33%
	III	6	**17**	2	0	25	68%
	IV	0	2	**0**	0	2	0%
	VandVI	0	0	0	**0**	0	–
	Total	7	20	3	0	30	
	T %	14.29%	85%	0%	–		**60%**
EMD-GRNN	I and II	**2**	2	0	0	4	50%
	III	5	**17**	3	0	25	68%
	IV	0	1	**0**	0	1	0%
	VandVI	0	0	0	**0**	0	–
	Total	7	20	3	0	30	
	T %	28.57%	85%	0%	–		**63.33%**
EMD-IMFs-Hybrid	I and II	**6**	5	0	0	11	54.55%
	III	1	**12**	0	0	13	92.31%
	IV	0	3	**3**	0	6	50%
	VandVI	0	0	0	**0**	0	–
	Total	7	20	3	0	30	
	T %	85.71%	60%	100%	–		**70%**
EMD-SVR-Hybrid	I and II	**6**	4	0	0	10	60%
	III	1	**15**	0	0	16	93.75%
	IV	0	1	**3**	0	4	**75%**
	VandVI	0	0	0	**0**	0	–
	Total	7	20	3	0	30	
	T %	85.71%	75%	100%	–		**80%**

hybrid models have higher forecasting precision compared with ARIMA, SVR, EMD-GRNN, GRNN, Wavelet-GRNN and Wavelet-SVR. Therefore, the proposed hybrid models can be used as effective and simple tools for air pollution warnings and management. In particular, in terms of MAPE, from small to large, the ranking of the eight models is EMD-IMFs-Hybrid, EMD-SVR-Hybrid, EMD-GRNN, Wavelet-GRNN, GRNN, SVR, Wavelet-SVR and ARIMA. Furthermore, the hybrid models with EMD and residual revision are more effective than individual models and hybrid models without residual revision for AQI forecasting, such as ARIMA, SVR, EMD-GRNN, GRNN, Wavelet-GRNN and Wavelet-SVR. In particular, EMD is more effective than Wavelet for AQI

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

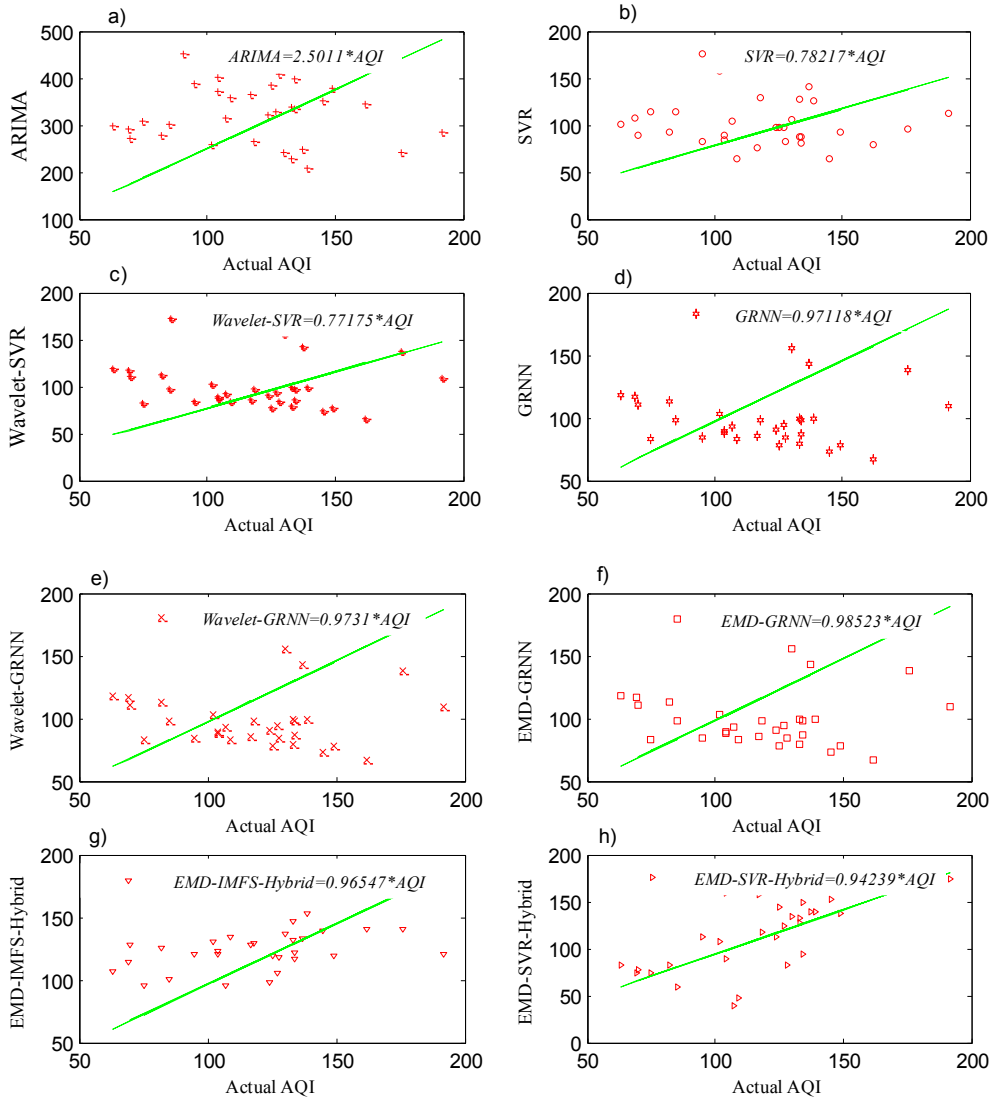

Fig. 5. Actual AQI vs forecasting values of eight models, the green line is based on the least squares method (The solid line with a slope of *1*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

data decomposition; for example, EMD-GRNN is better than Wavelet-GRNN. The proposed hybrid models can be applied to forecast other pollution indexes, such as disease burden analyses, economic time series and so on.

In future work, we will try to model other air pollution indexes and build a more comprehensive air pollution forecasting theory. For example, the testing sample size selection for the seasonal effect analysis. For the daily AQI forecasting, the testing sample size is usually 30 (Zhou et al., 2014). Air pollution mainly comes from industry, stoves, heating boilers and transportations. To protect visibility and people's health, this study suggests the following measures: i) the industrial layout should be reasonable, i.e., factories should not be excessively concentrated in one region; ii) regional heating and central heating should replace the thousands of stoves and heating boilers; iii) traffic pollution should be reduced by improving the combustion design of transportation tools and encouraging people use public transportation instead of driving; iv) new energy research and the ratio of new energy sources in total energy generation, such as solar energy, hydrogen fuel and wind energy should be increased; and v) afforestation should be increased. The rough leaves can absorb a large amount of floating dust and purify the air.

Conflict of interests

The authors declare that there are no conflicts of interests for the publication of this study.

Fundings

Fundamental Research Funds for the Central Universities (Grant No. lzujbky-2015-186) and National Social Science Foundation (Grant No. 12CTJ012).

Acknowledgments

The authors would like to thank Professor Jianzhou Wang for providing us constructive suggestions for this research and sharing his insights on related topics. The work is supported by Fundamental Research Funds for the Central Universities (Grant No. lzujbky-2015-186) and National Social Science Foundation (Grant No. 12CTJ012).

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

Appendix

1. Posterior-variance-test

After building the gray forecasting model, the posterior-variance-test is commonly used to test it. The posterior-variance-test is based on the error. According to the size of the residual data in each period, it investigates the probability of the smaller residuals and the size of the indicators about the standard deviation of prediction error. The specific approach: assuming that the original sequence and the residual sequence are $\{x_t^0\}$ and $\{e_t\}$, the corresponding variance estimation are, respectively, s_x^2 and s_e^2. P is the posteriori error ratio, and C is small error probability, defined, respectively, as

$$C = \frac{S_e}{S_x} \text{ and.}$$

In the actual data analysis, P is the difference between the residual error and the average residual which is less than a given proportion of $0.6745S_x$.

The predictive ability of the model is described by C and P. The smaller C means that the residual series is not discrete even though the historical data are discrete. The greater P indicates there are more residual points in the critical interval defined by $0.6745\ S_x$. According to C and P two indexes, which can be used to evaluate the predictive ability of the forecasted model, the evaluation criteria see Table F1.

Table F1
Model classification

Model level	C	P
Level 1 (Good)	$C \leq 0.35$	$p \geq 0.95$
Level 2 (Qualified)	$0.35 < C \leq 0.5$	$0.80 \leq p < 0.95$
Level 3 (Barely qualified)	$0.5 < C \leq 0.65$	$0.70 \leq p < 0.80$
Level 4 (Unqualified)	$C > 0.65$	$p < 0.7$

2. The definition of AQI

For one pollutant po ($PM_{2.5}$, PM_{10}, CO, SO_2, NO_2 and O_3), the air quality index (AQI) is calculated by

$$AQI_{po} = \frac{AQI_{Hi} - AQI_{Lo}}{BP_{Hi} - BP_{Lo}}(C_{po} - BP_{Lo}) + AQI_{Lo},$$

where AQI_{po} is the AQI value for pollutant po, C_{po} is the concentration of pollutant po, BP_{Hi} and BP_{Lo} respective are the high breakpoint value and the low breakpoint value for C_{po}, AQI_{Hi} and AQI_{Lo} are AQI indexes corresponding to BP_{Hi} and BP_{Lo} which can be refereed in terms of Table 3 in Yuan and Liu (2014). After calculating the AQI_{po} for $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2 and O_3, the AQI is calculated by $AQI = \max\{AQI_{PM_{2.5}}, AQI_{PM_{10}}, AQI_{CO}, \ldots, AQI_{O_3}\}$. Specially, the 24-hourly AQI (or daily AQI) is calculated according to the AQI_{po} values of 24-hourly average data of $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2, average of hourly maximum O_3 values and average of 8-hourly maximum O_3 values.

References

Bai, Y., Li, Y., Wang, X., Xie, J., Li, C., 2016. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmos. Pollut. Res. 7 (3), 557–566.
Box, G.E., Jenkins, G.M., 1976. Time series analysis: forecasting and control. Holden Day, San Francisco, 3226 (3228), 10.
Chaudhuri, S., Dutta, D., 2014. Mann–Kendall trend of pollutants, temperature and humidity over an urban station of India with forecast verification using different ARIMA models. Environ. Monit. Assess. 186 (8), 4719–4742.
Chen, W., Tang, H., Zhao, H., 2016. Urban air quality evaluations under two versions of the national ambient air quality standards of China. Atmos. Pollut. Res. 7 (1), 49–57.
Cheng, C.H., Wei, L.Y., 2014. A novel time-series model based on empirical mode decomposition for forecasting TAIEX. Econ. Modell. 36, 36–141.
Cobourn, W.G., 2010. An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. Atmos. Environ. 44 (25), 3015–3023.
de Mattos Neto, P.S., Madeiro, F., Ferreira, T.A., Cavalcanti, G.D., 2014. Hybrid intelligent system for air quality forecasting using phase adjustment. Eng. Appl. Artif. Intel. 32, 185–191.
Díaz-Robles, L.A., Ortega, J.C., Fu, J.S., Reed, G.D., Chow, J.C., Watson, J.G., Moncada-Herrera, J.A., 2008. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. Atmos. Environ. 42 (35), 8331–8340.
Elangasinghe, M.A., Singhal, N., Dirks, K.N., Salmond, J.A., Samarasinghe, S., 2014. Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. Atmos. Environ. 94, 106–116.
Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J., 2015. Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos. Environ. 107, 118–128.
Fernando, H.J.S., Mammarella, M.C., Grandoni, G., Fedele, P., Di Marco, R., Dimitrova, R., Hyde, P., 2012. Forecasting PM 10 in metropolitan areas: efficacy of neural networks. Environ. Pollut. 163, 62–67.
Gardner, M.W., Dorling, S.R., 1999. Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmos. Environ. 33 (5), 709–719.
Goyal, P., Chan, A.T., Jaiswal, N., 2006. Statistical models for the prediction of respirable suspended particulate matter in urban cities. Atmos. Environ. 40 (11), 2068–2077.
Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O., 2005. A neural network forecast for daily average PM10 concentrations in Belgium. Atmos. Environ. 39 (18), 3279–3289.
Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.-C., Tung, C.C., Liu, H.H., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. P. Roy. Soc. Lond. A. Mat. 454 (1971), 903–995.
Huang, N.E., Wu, M.L., Qu, W., Long, S.R., Shen, S.S., Zhang, J.E., 2003. Applications of Hilbert-Huang transform to non-stationary financial time series analysis. Appl. Stoch. Model. Bus. 19 (3), 245–268.
Huang, N.E., Wu, Z., 2008. A review on Hilbert-Huang transform: method and its applications to geophysical studies. Rev. Geophys. 46 (2).
Jian, L., Zhao, Y., Zhu, Y.P., Zhang, M.B., Bertolatti, D., 2012. An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China. Sci. Total Environ. 426, 336–345.
Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J., Shao, D., 2004. Progress in developing an ANN model for air pollution index forecast. Atmos. Environ. 38 (40), 7055–7064.
Kolehmainen, M., Martikainen, H., Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. Atmos. Environ. 35 (5), 815–825.
Konovalov, I.B., Beekmann, M., Meleux, F., Dutot, A., Foret, G., 2009. Combining deterministic and statistical approaches for PM10 forecasting in Europe. Atmos. Environ. 43 (40), 6425–6434.
Kumar, A., Goyal, P., 2011. Forecasting of daily air quality index in Delhi. Sci. Total Environ. 409 (24), 5517–5523.
Kurt, A., Oktay, A.B., 2010. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert. Syst. Appl. 37 (12), 7986–7992.
Lu, W.Z., Wang, D., 2008. Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. Sci. Total Environ. 395 (2), 109–116.
Mindell, J., Joffe, M., 2004. Predicted health impacts of urban air quality management. J. Epidemiol. Commun. H. 58 (2), 103–113.
Niu, M., Wang, Y., Sun, S., Li, Y., 2016. A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting. Atmos. Environ. 134, 168–180.
Ordieres, J.B., Vergara, E.P., Capuz, R.S., Salazar, R.E., 2005. Neural network prediction model for fine particulate matter (PM2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environ. Model. Softw. 20 (5), 547–559.
Ortiz-García, E.G., Salcedo-Sanz, S., Pérez-Bellido, Á.M., Portilla-Figueras, J.A., Prieto, L., 2010. Prediction of hourly O3 concentrations using support vector regression algorithms. Atmos. Environ. 35 (44), 4481–4488.
Pérez, P., Trier, A., Reyes, J., 2000. Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ. 34 (8), 1189–1196.
Qin, L.T., Liu, S.S., Liu, H.L., Zhang, Y.H., 2010. Support vector regression and least squares support vector regression for hormetic dose–response curves fitting. Chemosphere 78 (3), 327–334.
Reikard, G., 2012. Forecasting volcanic air pollution in Hawaii: tests of time series models. Atmos. Environ. 60, 593–600.

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069

Rojas, A., Górriz, J.M., Ramírez, J., Illán, I.A., Martínez-Murcia, F.J., Ortiz, A., Gómez Río, M., Moreno-Caballero, M., 2013. Application of empirical mode decomposition (EMD) on DaTSCAN SPECT images to explore Parkinson disease. Expert, Syst. Appl. 40 (7), 2756–2766.

Sheng, N., Tang, U.W., 2015. The frist offical ranking city by air quality in China- A review and analysis. Cities 51, 139–149.

Slini, T., Karatzas, K., Moussiopoulos, N., 2002. Statistical analysis of environmental data as the basis of forecasting: an air quality application. Sci. Total Environ. 288 (3), 227–237.

Song, Y., Qin, S., Qu, J., Liu, F., 2015. The forecasting research of early warning systems for atmospheric pollutants: a case in Yangtze River Delta region. Atmos. Environ. 118, 58–69.

Sowlat, M.H., Gharibi, H., Yunesian, M., Mahmoudi, M.T., Lotfi, S., 2011. A novel, fuzzy-based air quality index (FAQI) for air quality assessment. Atmos. Environ. 45 (12), 2050–2059.

Tratar, L.F., Strmčnik, E., 2016. The comparison of Holt–Winters method and Multiple regression method: a case study. Energy 109, 266–276.

Tien, T.L., 2009. A new grey prediction model FGM(1,1). Math. Comput. Model 49 (7), 1416–1426.

Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., Kolehmainen, M., 2011. Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci. Total Environ. 409 (7), 1266–1276.

Wang, D., Wei, S., Luo, H., Yue, C., Grunder, O., 2017a. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Sci. Total Environ. 580, 719–733.

Wang, P., Zhang, H., Qin, Z., Zhang, G., 2017b. A novel hybrid-Garch model based on ARIMA and SVM for PM2.5 concentrations forecasting. Atmos. Pollut. Res. 5 (8), 850–860.

Wang, J.Z., Ma, Z.X., Li, L., 2005. Detection, mining and forecasting of impact load in power load forecasting. Appl. Math. Comput. 168 (1), 29–39.

Wang, P., Liu, Y., Qin, Z., Zhang, G., 2015. A novel hybrid forecasting model for PM10 and SO2 daily concentrations. Sci. Total Environ. 505, 1202–1212.

Yang, Z., Wang, J., 2017. A new air quality monitoring and early warning system: air quality assessment and air pollutant concentration prediction. Environ. Res. 158, 105–117.

Yahya, K., Zhang, Y., Vukovich, J.M., 2014. Real-time air quality forecasting over the southeastern United States using WRF/Chem-MADRID: multiple-year assessment and sensitivity studies. Atmos. Environ. 92, 318–338.

Yeganeh, B., Shafie Pour Motlagh, M., Rashidi, Y., Kamalan, H., 2012. Prediction of CO concentrations based on a hybrid partial least square and support vector machine model. Atmos. Environ. 55, 357–365.

Yi, J., Prybutok, V.R., 1996. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environ. Pollut. 92 (3), 349–357.

Yuan, Y., Liu, M.Y., 2014. Differences of air quality index (AQI) and air pollution index (API). G. Z. Chem. Ind. 42 (12), 164–166.

Zhang, J., Wang, R., Bai, F., Zheng, J., 2011. A quasi-MQ EMD method for similarity analysis of DNA sequences. Appl. Math. Lett. 24 (12), 2052–2058.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012. Real-time air quality forecasting, part I: history, techniques, and current status. Atmos. Environ. 60, 632–655.

Zhou, Q., Jiang, H., Wang, J., Zhou, J., 2014. A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Sci. Total Environ. 496, 264–274.

Zhu, S., Wang, J., Zhao, W., Wang, J., 2011. A seasonal hybrid procedure for electricity demand forecasting in China. Appl. Energ 88 (11), 3807–3815.

Please cite this article in press as: Zhu, S., et al., Daily air quality index forecasting with hybrid models: A case in China, Environmental Pollution (2017), http://dx.doi.org/10.1016/j.envpol.2017.08.069