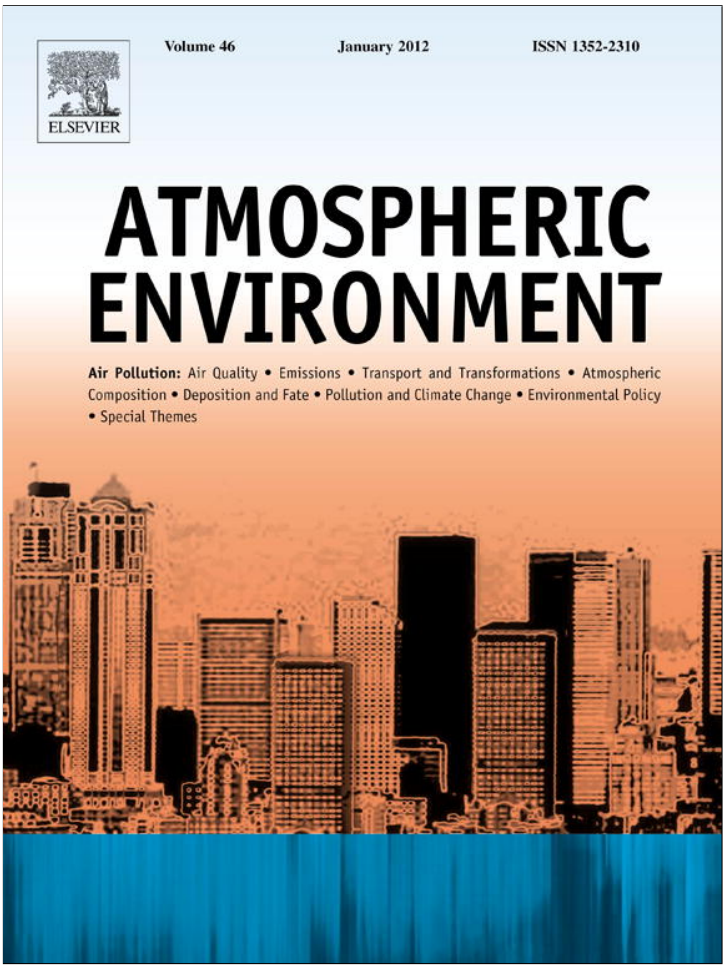

Atmospheric Environment 74 (2013) 346–359

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

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

Ensemble and enhanced PM_{10} concentration forecast model based on stepwise regression and wavelet analysis

Yuanyuan Chen [a,b], Runhe Shi [a,b,*], Shijie Shu [a,b], Wei Gao [a,b,c]

[a] Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai 200062, China
[b] Joint Laboratory for Environmental Remote Sensing and Data Assimilation, ECNU and CEODE, Shanghai 200062, China
[c] USDA UV-B Monitoring and Research Program, Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO, USA

HIGHLIGHTS

- Suitable empirical forecast model was established for cities in eastern China.
- Enhanced forecast model was established for PM_{10} concentration.
- Ensemble PM_{10} concentration forecast model was established for eastern China.
- The model performance metrics are evaluated and verified.
- A comparatively high accuracy and precision were gotten with established forecast method.

ARTICLE INFO

Article history:
Received 5 September 2012
Received in revised form
25 February 2013
Accepted 2 April 2013

Keywords:
PM_{10} concentration
Wavelet analysis
Stepwise regression
Enhanced model
Ensemble model

ABSTRACT

An ensemble and enhanced PM_{10} (particulate matter with a diameter less than 10 μm) concentration forecast model was established in eastern China based on data from 2005 to 2009. The enhanced model consists of a single stepwise regression forecast model and a combined forecast model based on wavelet decomposition and stepwise regression. Six individual forecast results were obtained with a combined model that can predict PM_{10} concentrations at multiple scales. By decomposing variables into detailed and approximated components in six scales and with the application of stepwise regression, the best-fitted forecast models were established in each component of the different scales. Then, the predicted results of the detail and approximation components were reconstructed in each scale as the enhanced prediction. A regional model was established for eastern China. The accuracy rate of each forecasted result by the regional model was calculated using testing data from 2010 based on the needs of operational forecasting. Precision evaluations were also performed. A comparatively higher accuracy was obtained by the combined model. The advantage of predicting the PM_{10} concentration with the combined model had wide spatial and temporal suitability. An enhanced forecast model was established for each city of eastern China with improvements, where all the predicted results in each city were evaluated by the accuracy rate and precision validation. In each city, the best-fitted model with the highest precision was selected and combined in an ensemble. The ensemble and enhanced forecast model had a significant improvement in accuracy rate and the highest precision of PM_{10} concentration forecasting in eastern China.

1. Introduction

Air quality is an important issue that is attracting increasingly more attention around the world (Kurt and Oktay, 2010). Air pollutants such as sulfuric dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and particulate matter with a diameter less than 10 μm (PM_{10}) have been reported to the public using the API (air pollution index) in many countries, especially in metropolitan areas (Jiang et al., 2004). Particulate matter can cross the nasal passages during inhalation, arriving at the throat and even the lungs. With long-term exposure to PM_{10}, the incidence of associated diseases (e.g., respiratory, cardiovascular disease, reduced lung function, heart attacks) increases in human beings (Künzli et al., 2000; Bravo and Bell, 2011). Thus, it is necessary to monitor air quality in real time and to predict its trends.

* Corresponding author. Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai 200062, China. Tel.: +86 021 54341232.
E-mail addresses: yychen0701@gmail.com (Y. Chen), shirunhe@gmail.com, rhshi2012@gmail.com (R. Shi).

1352-2310/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2013.04.002

The PM_{10} concentration in the air is affected by human activities as well as meteorological factors (Saliba et al., 2010). Anthropogenic sources such as automobile exhaust, industrial discharges, fossil fuel combustion and coal burning are the main sources of PM_{10} (Querol et al., 2002; Contini et al., 2010; Salvador et al., 2011). Furthermore, PM_{10} can be transported in the air and deposited in other places through atmospheric motion and precipitation (Singh, 1995; Senaratne et al., 2005). Therefore, many real-time monitoring stations should be established in different cities, and even in different functional areas within a city, to ensure area representation. However, such a project would be complex, costly and limited in terms of spatio-temporal coverage (Qu et al., 2010). Satellite-derived data are an important source and can help compensate for this limitation (Mishchenko et al., 2007; Kharol et al., 2011).

Based on prior knowledge, many studies have focused on forecasting PM_{10}. Two types of methods (deterministic and statistical) are generally used. Deterministic method employs meteorological, emission and chemistry models (Zhang, 2004; Bruckman, 1993; Coats, 1996; Lurmann, 2000; Jeong et al., 2011), which can simulate the discharge of a pollutant, the transfer and diffusion process, the removal process with a limited number of monitoring stations in terms of animated figures (Baklanov et al., 2008; Kim et al., 2010). However, the simulation results suffer from low precision (Vautard et al., 2007; Stern et al., 2008). Therefore, statistical methods are more appropriate for air quality forecasting (Manders et al., 2009).

Statistical methods such as artificial neural networks, nonlinear regression, and multiple linear regression are widely used for particulate matter forecasting (Hooyberghs et al., 2005; Hoi et al., 2009; Li et al., 2011). When using an artificial neural network, the precision of the simulation greatly depends on the experience of the model designer. And there is an inherent conflict between training and predictive ability, which can induce local extrema and overfitting. Moreover, the simulation results are different even with the same model and parameters. As a more stable statistical method, multiple linear regression can explain the variation in the dependent variable as a function of multiple independent variables that are widely used for PM_{10} forecasting. However, redundant independent variables can introduce collinearity. Thus, stepwise regression has an advantage in avoid the collinearity, however, extremum events can be neglected. Forecasting air quality with single stepwise regression is insufficient for high-accuracy prediction.

Wavelet analysis is an effective method for spatio-temporal characteristic analysis and forecasting (Kim et al., 2002; Murtagh et al., 2004; Qiu et al., 2011). With an adjustment window, wavelet analysis can stretch or translate the data and can focus on each detailed part of the long time series data. It is rare that combined wavelet analysis and stepwise regression forecast models are used for air quality forecasting. Moreover, most studies have attempted to simulate air quality using a uniform model, without attempting to integrate different forecast models as an ensemble.

The objective of this study was to establish an ensemble and enhanced method to forecast PM_{10} concentrations with higher precision in eastern China. First, based on wavelet analysis, the PM_{10}, meteorological parameters and satellite-derived AOD (aerosol optical depth) were decomposed into detail and approximation components in six scales with wavelet transformation. With stepwise regression, a PM_{10} forecasting model of detail and approximation components at each scale was constructed. Then, as the final forecast, the predicted results of the detail and approximation components at each scale were reconstructed based on the theory of wavelet decomposition. Second, single stepwise regression and a combined forecast model were used as an enhanced regional air quality forecast method for PM_{10} concentration forecasting in eastern China. The performance of the regional model was evaluated, and its spatial and temporal suitability were analyzed. Third, the enhanced model was used to determine the PM_{10} concentration in each city, and the best-fitted forecast model for each city was selected and integrated as an ensemble for forecasting the PM_{10} concentration in eastern China. The accuracy rate and precision of the ensemble forecast model were evaluated and compared with the regional model.

2. Data and methods

2.1. Study area

Locating in the lower and middle reaches of the Yangtze River, Eastern China extends from 113°E to 123°E and 23°N to 39°N while containing one municipality (Shanghai) and six provinces (Shandong, Jiangsu, Zhejiang, Anhui, Jiangxi and Fujian) (Fig. 1). The climate here is characterized by the East Asian monsoon with typical seasonal changes, as the weather is dry and cold in winter with high temperatures and abundant rainfall in summer. As one of the most competitive and dynamic economic regions, Eastern China has suffered natural environmental stress. Various types of industries (e.g., light industry, machinery, electronics) are scattered throughout this area. To ensure the development of the economy, much labor and widely accessible transportation are needed. Therefore, more attention should be paid to the environment, especially to air quality in eastern China.

2.2. Air quality data

The air pollution index (API) of 23 cities in Eastern China from 2005 to 2010 was downloaded from the Ministry of Environmental Protection of the People's Republic of China. The API was selected when particulate matter was the primary pollutant in each city. The PM_{10} concentrations were calculated with the exchange formula of API and PM_{10} concentration (See Part 1 in Supplement Information).

2.3. Independent data

The MODIS sensor onboard the polar-orbiting Terra and Aqua spacecraft provides high spatial resolution observations of ocean, land, aerosols and clouds. With the development of the MODIS algorithms, AOD determination has greatly improved. The latest algorithm (C005) is a significant improvement over its predecessor, C004. In this study, the AOD at 0.55 μm from the C005 Level-2 aerosol products (Terra, MOD04; Aqua; MYD04) was obtained from 2005/1/1 to 2010/6/30 (http://modis.gsfc.nasa.gov/), with a spatial resolution of a 10 × 10 km pixel. The residual data were downloaded from Eastern China Normal University (ECNU, http://dbps.ecnu.edu.cn/data/terra/). The satellite receiving system of ECNU was built in May 2010. The collection 005 algorithm was also run, and the products were processed in real time.

Four-times daily, meteorological factors, including surface temperature, potential temperature, precipitable water, pressure, relative humidity, sea level pressure, u-wind, v-wind, specific humidity, and total cloud cover in eastern China from 2005 to 2010 were downloaded from the NCEP/NCAR Reanalysis datasets The coordinate meteorological data of 23 cities in eastern China were derived using the corresponding centered latitude and longitude. The correlations between PM_{10} and each meteorological factor at each of the 4 times were analyzed, and the fitted temporal meteorological factors were chosen.

2.4. Stepwise regression

Stepwise regression is a type of multiple linear regression that can select the best-fitted combination of independent variables for

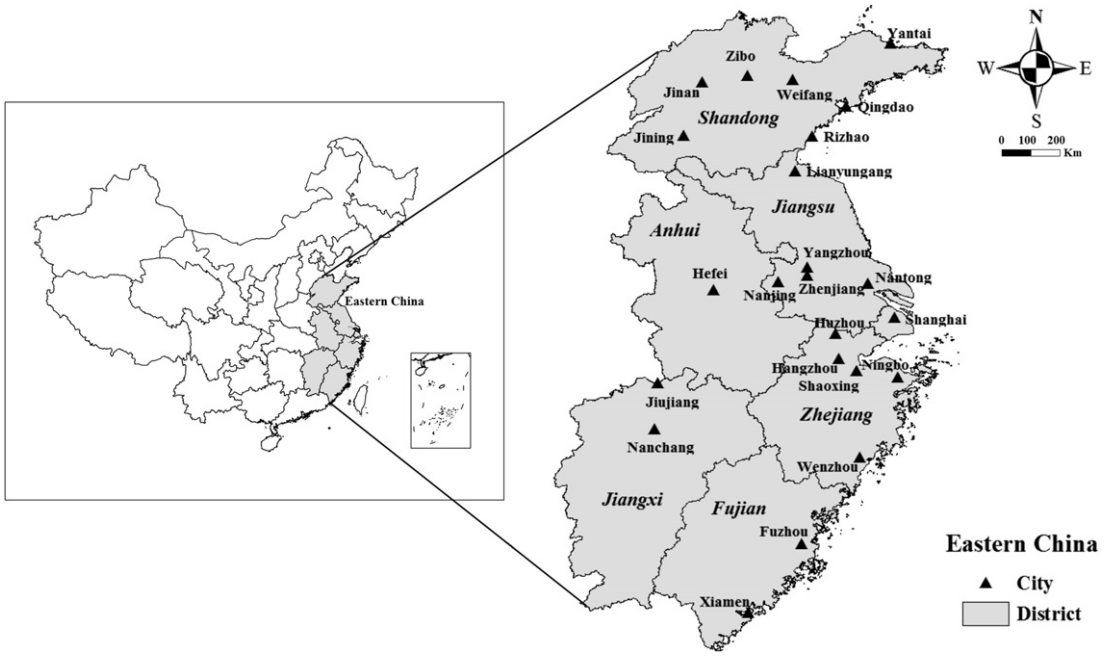

Fig. 1. Study area and sampling cities in eastern China.

dependent variable prediction with forward-adding and backward-deleting variables. The stepping procedure begins as an initial model definition, with a stepped forward addition of a variable to the previous model. The critical F value is then used to check the eligibility of the added variable. With a new variable added, the previous variables in the model may lose their predictive ability. Thus, stepping criteria are used to check the significance of all the included variables. If the variable is insignificant, then the backward method is used to delete it. Forward adding and backward deleting are repeated until no variable is added or removed. The stepping procedure is eliminated when the optimized model is established.

2.5. Wavelet analysis

Wavelet analysis is a useful mathematical method for data or signals in a time series and frequencies (Torrence and Compo, 1998; Durka, 2003). Wavelet analysis can be used to reveal every detail of the signal with shifting or dilating. A continuous wavelet transform (CWT) can be defined as follows (Bruce et al., 2001; Mallat, 1999).

Given a mother wavelet function $\psi(t)$, it belongs to L2(R) (two-dimensional space). With shifting or dilating, the mother wavelet produces a group of continuous wavelets as:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)$$

The wavelet transformation coefficient can be expressed as:

$$\Phi_f(a,b) = |a|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} f(t)\psi^*_{a,b}\left(\frac{t-b}{a}\right)\mathrm{d}t$$

where $\Phi_f(a,b)$ is the wavelet transform coefficient; $\psi^*_{a,b}$ is the conjugate function of $\psi_{a,b}$; a is the frequency resolution, which refers to the periodicity and indicates the width of the wavelet; and b is the time parameter, which refers to shifting in the time series.

Wavelet multi-resolution analysis is a typical CWT that can decompose a signal into separate components (detail and approximation) and different resolutions with mother wavelets and scaling (Mallat, 1999; Bruce et al., 2002). The scale of the signal is usually decomposed as an a level ($a = 2, 4, 8, \ldots, 2^n$). The detail component is also called a high-frequency signal, and it represents obvious and rapid changes. The approximation component is called a low-frequency signal, and it represents coarse changes. The Sym8 wavelet was used in this study. Sym8 has been demonstrated to be a useful mother wavelet and is more appropriate for signal compression (Eriksson et al., 2000; Zhang et al., 2012).

2.6. Model performance evaluations

The performances of each model were evaluated by the following metrics.

Accuracy, vacancy ratio and missing rate were calculated based on the operational forecast,

$$\text{Accuracy rate} = \frac{N_{\text{right}}}{N} \times 100\%$$

$$\text{Vacancy rate} = \frac{N_{\text{higher}}}{N} \times 100\%$$

$$\text{Missing rate} = \frac{N_{\text{lower}}}{N} \times 100\%$$

The index of agreement (IA)

$$\text{IA} = 1 - \frac{\sum_i^N (\text{O}_i - \text{P}_i)^2}{\sum_i^N \left(|\text{P}_i - \overline{\text{O}}| + |\text{O}_i - \overline{\text{O}}|\right)^2}$$

The root mean squared error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_i^N (\text{O}_i - \text{P}_i)^2}{N}}$$

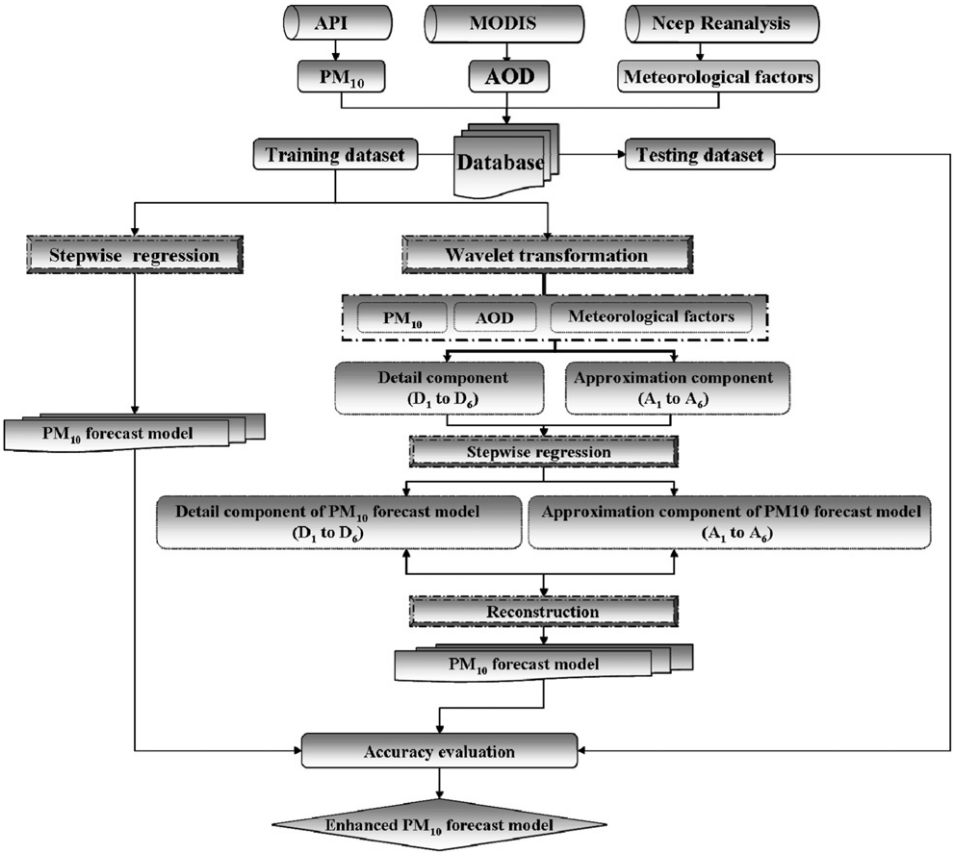

Fig. 2. The process of the enhanced PM_{10} concentration forecast model.

The mean absolute error (MAE)

$$\mathrm{MAE} = \frac{\sum_{i}^{N}|\mathrm{O}_i - \mathrm{P}_i|}{N}$$

The mean relative error (MRE)

$$\mathrm{MRE} = \frac{1}{N}\sum_{i}^{N}\frac{|\mathrm{O}_i - \mathrm{P}_i|}{\mathrm{O}_i}$$

where $O_i$is the observed value, P_i is the predicted value, $\overline{O}$is the average of the observed value, N is the total number of the testing dataset, N_{right} is the number of the predicted pollution level in accordance with the observed pollution level, N_{higher} is the number of the predicted pollution level that is higher than the observed pollution level, and N_{lower} is the number of the predicted pollution level that is lower than the observed pollution level. The pollution level division was according to the API (SeeTable A1 in Supplement Information).

The IA was used to measure the deviation of the predicted value from the observed value and the observed value on average in magnitude (Kang et al., 2005). RMSE was used to measure the sensitivity and extremum effect of the predicted value. MAE was used to evaluate the absolute error range of the predicted value on average. MRE was used to reflect the specification of the predicted value on average.

2.7. Establishment of enhanced forecast model

The data from 2005 to 2009 were selected as the training dataset, and those from 2010 were selected as the testing dataset. The training data were simulated with the enhanced PM_{10} concentration forecast model as shown in Fig. 2. Based on wavelet multi-resolution analysis, sym8 was used, and the a level was set as 2^1, 2^2, 2^3, 2^4, 2^5 and 2^6. Then, a dataset with six detail components and six approximation components was obtained in MATLAB. All 13 parameters (PM_{10}, AOD and meteorological factors) were decomposed into six scales ($a = 2^1, 2^2, 2^3, 2^4, 2^5, 2^6$) with detail components (D1 to D6) and approximation components (A1 to A6) in each scale. Thus, 12 training datasets were obtained. Then, a PM_{10} forecast model in each dataset was established with the corresponding AOD and meteorological factors on the same scale and with the same components based on stepwise regression. A total of 12 PM_{10} forecast models were built comprising detail components and approximation components. According to the principle of wavelet decomposition, the PM_{10} prediction was reconstructed in six scales

Table 1
Descriptive statistics of observed values and simulated values based on the regional enhanced model.

PM_{10}	Minimum (mg m^{-3})	Maximum (mg m^{-3})	Mean (mg m^{-3})	Standard deviation	Kurtosis	Skewness
Observed value	0.012	0.600	0.095	0.056	10.488	2.259
SSR	0.153	0.281	0.196	0.020	−0.145	0.306
C1	0.047	0.311	0.153	0.030	1.156	0.424
C2	0	0.252	0.055	0.025	5.067	1.424
C3	0.053	0.292	0.156	0.028	0.650	0.540
C4	0.107	0.358	0.221	0.031	1.094	0.674
C5	0.003	0.429	0.149	0.080	0.236	0.689
C6	0.003	0.432	0.158	0.081	0.097	0.542

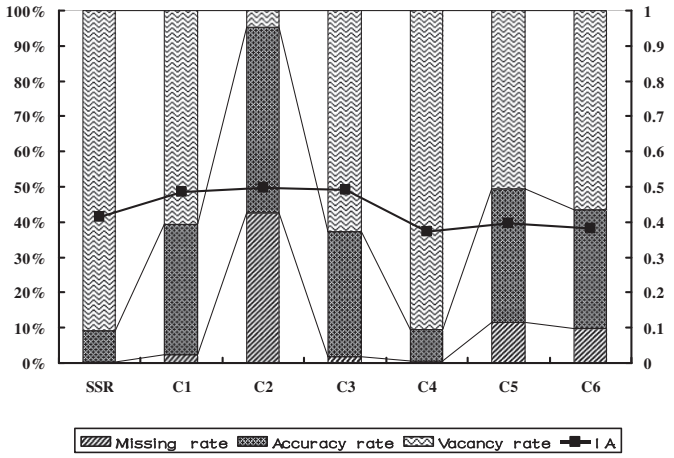

Fig. 3. Performance evaluation of the regional model based on the needs of operational forecasting and the index of agreement of each dataset of PM_{10} concentration predictions.

with combined detail and approximation components. Then, a PM_{10} forecast model based on combined wavelet analysis and stepwise regression was established. Comparing the accuracies of the two models' performances (PM_{10} forecast model with single stepwise regression & PM_{10} forecast model based on combined stepwise regression and wavelet analysis), the best performance model for PM_{10} concentration forecasting was selected as an enhanced PM_{10} forecast model.

3. Results and discussion

3.1. Regional model and performance evaluation

The combined model takes all the independent variable into consideration and six datasets of PM_{10} predictions were obtained (shorted as C1, C2, C3, C4, C5 and C6). On the contrary, signal stepwise regression (abbreviated as SSR) discard surface temperature, potential temperature and pressure for two reasons. First, wavelet analysis must consider detailed components that variables with seldom contribution could also bring obvious effect to them. Second, the stepwise regression has ability in eliminating the redundant variables so it's inefficient to consider those small contributors to PM_{10} concentration, including surface temperature, potential temperature and pressure.

3.1.1. Characteristics of the predicted values based on a regional model

The characteristics of the seven sets of predicted results are shown in Table 1 with descriptive statistics.

The observed PM_{10} concentrations ranged from 0.012 mg m^{-3} to 0.6 mg m^{-3}, with a comparatively low average of 0.095 mg m^{-3}, which imply that most of the data were accumulated with low values. The standard deviation of the observed values is 0.054. The kurtosis and skewness of the observed values were the highest in the eight datasets. Thus, the observed data had a leptokurtic distribution and were right-skewed, and many extreme data appeared at the right side. The dataset of the PM_{10} prediction by SSR showed that the range of the dataset was small and that the mean value was comparatively high, which implies that the

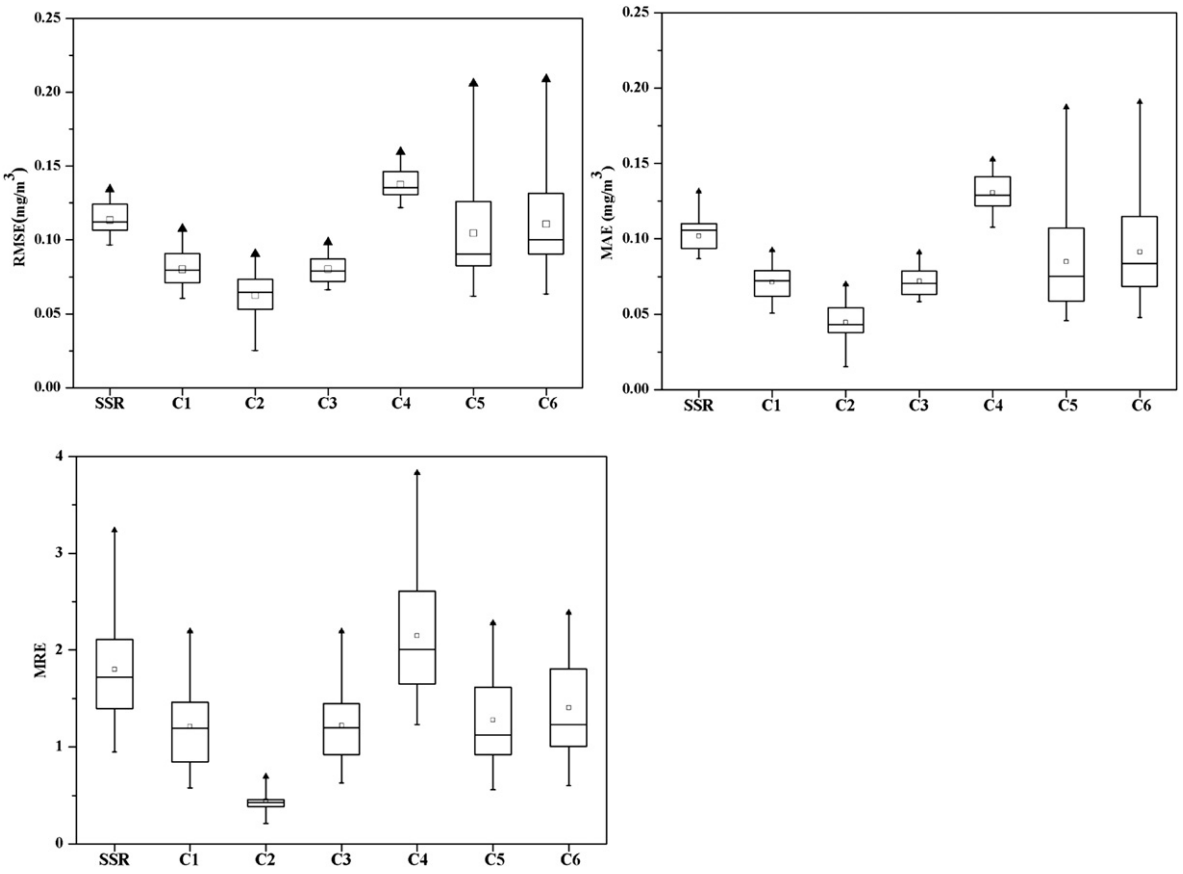

Fig. 4. The precision of the enhanced regional model.

predicted PM_{10} concentrations were slightly higher in total. The lowest standard deviation implied that slight differences existed among the predicted values, and the negative kurtosis implied that the distribution of the predicted PM_{10} concentrations was smooth and positive. The smallest skewness indicated that the predicted values were right-skewed with few extreme values. When comparing the observed values with the six results based on enhanced forecast model, C2 is remarkable, with the lowest average value. The kurtosis and skewness were higher than the other predictions and more similar to the observed values. The PM_{10} concentration predicted by C2 was very low with a few extremum values, and the distribution of the whole prediction dataset showed a leptokurtic distribution and was right-skewed. The highest mean value appeared in C4. The predicted PM_{10} concentrations by C4 ranged from 0.107 to 0.358 mg m^{-3}, with a standard deviation of 0.031; this finding implies that there were slight differences in the predicted values in this dataset and that most of the data were accumulated at a higher value. The PM_{10} concentration prediction in C5 and C6 accumulated at high values, and the standard deviations were higher than those of other datasets, which implies that the data in the two datasets had obvious discrepancies between each other.

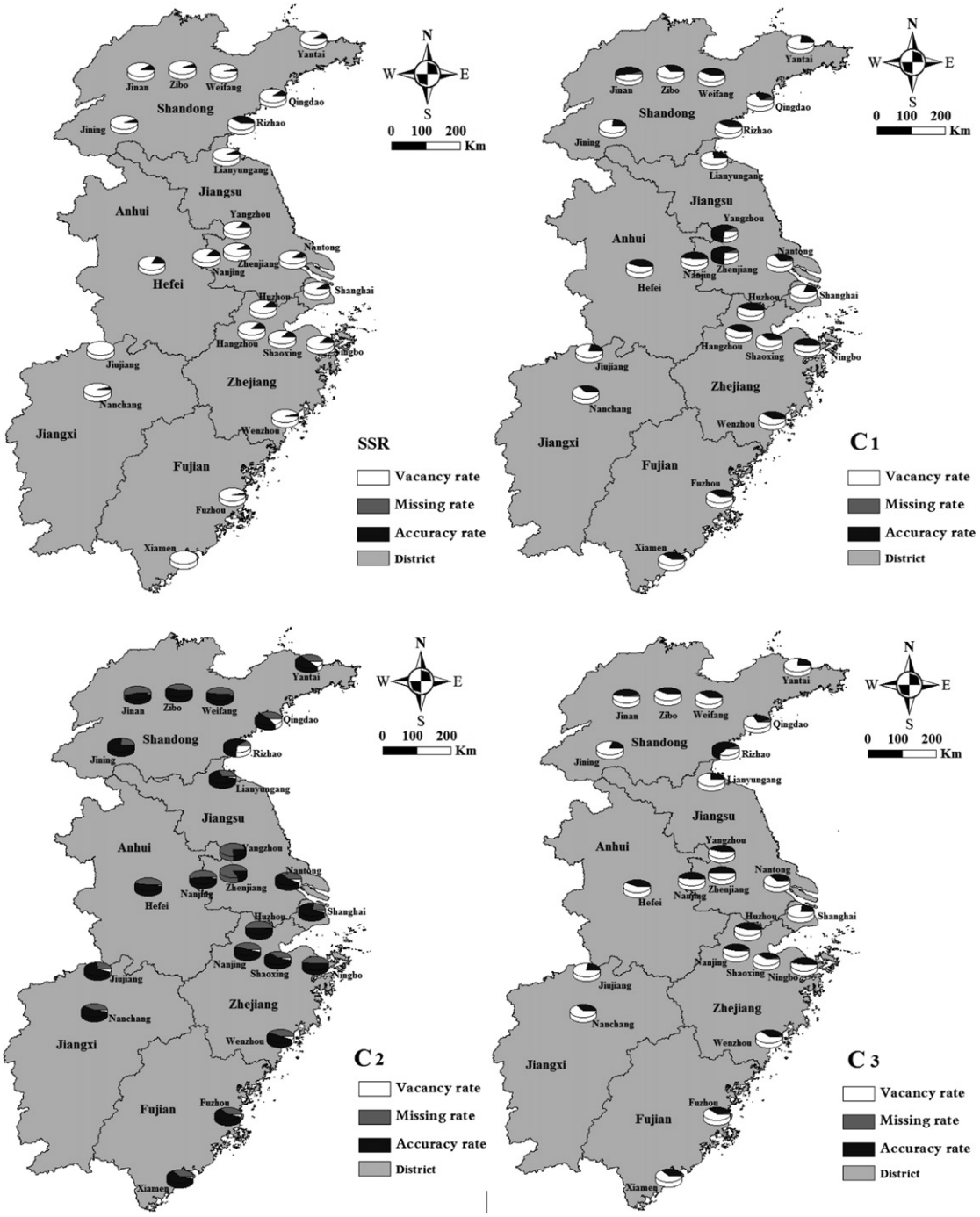

Fig. 5. The spatial distributions of the accuracy rate, missing rate, vacancy rate and IA with the enhanced regional forecast model.

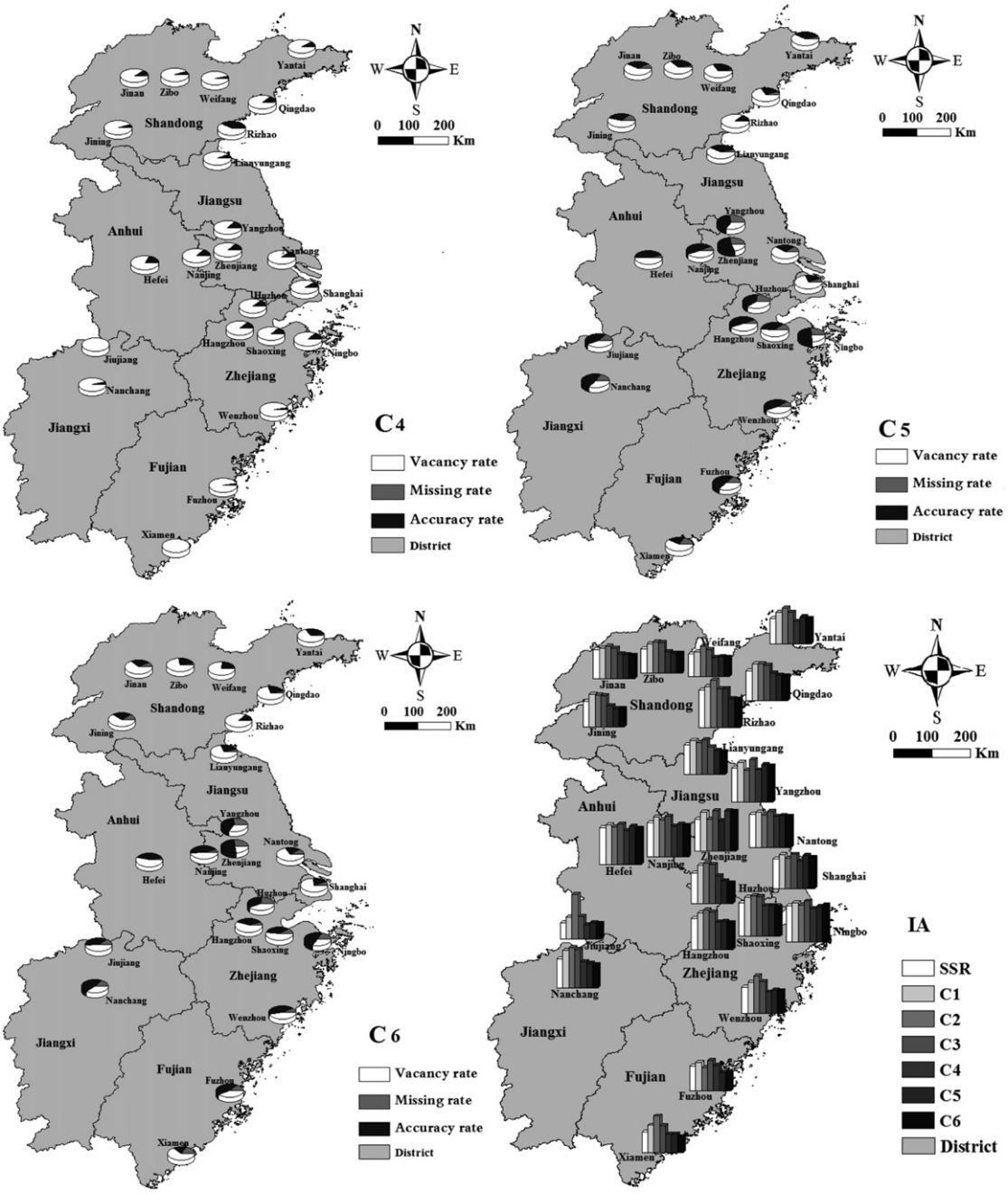

Fig. 5. (*continued*).

3.1.2. Performance evaluation of the regional model

To evaluate the PM_{10} concentration prediction for operational forecasting, the predicted PM_{10} concentrations were converted into a pollution index and classified into five pollution levels. The accuracy rate, vacancy rate, missing rate and IA of each set of PM_{10} concentration predictions were calculated (Fig. 3).

The accuracy rates of the seven sets of PM_{10} concentrations predicted based on the regional model were ordered as followed: C2 (52.64%) > C5 (37.73%) > C1 (37.18%) > C3 (35.56%) > C6 (33.61%) > C4 (9.04%) > SSR (8.74%). There were two extreme values, to which more attention should be paid in Fig. 3: one was the missing rate of C2, and the other was the vacancy rate of SSR. The IA of each PM_{10} concentration prediction was ordered as follows: C2 > C3 > C1 > SSR > C5 > C6 > C4. The highest accuracy rate appeared in C2, the same as IA. Prediction by single stepwise regression was not well performed with the lowest accuracy rate. The reason for this result is that the PM_{10} concentrations predicted by C2 were accumulated in a lower grade and the SSR-predicted values were accumulated in a higher grade. However, the air quality in most of the year in eastern China had a clean grade. Therefore, because of the higher prediction made by SSR, it compensated for the disadvantage in predicting higher values, and the IA of this prediction result was improved. Thus, by comparison, we can conclude that C2 for PM_{10} concentration prediction was the most suitable forecast model for eastern China in general.

To further evaluate the precision of each prediction, the RMSE, MAE and MRE were also analyzed, as shown in Fig. 4. The results show that the precision of the forecast was greatly improved by C2, and this improvement was significant comparing with the other six predictions. Thus, C2 was the best-fitted model in general.

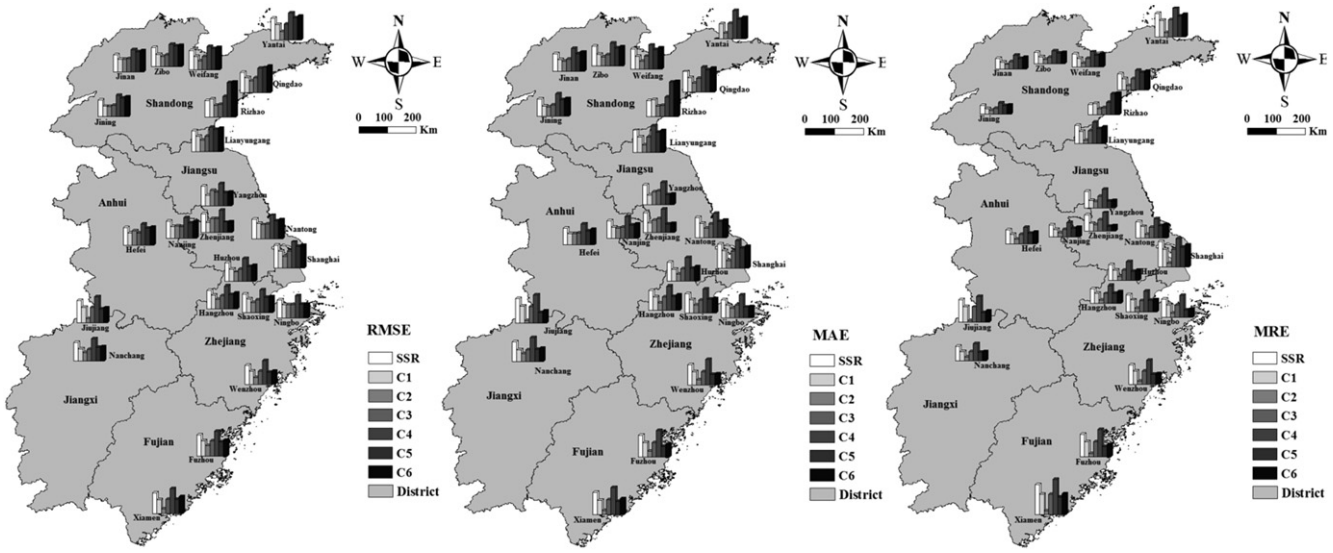

Fig. 6. Distributions and ranges of RMSE, MAE and MRE of each simulated result based on the enhanced regional model.

3.1.3. Spatial applicability validation

The analysis on the spatial distribution of the accuracy rate, missing rate, vacancy rate and IA in each city are shown in Fig. 5. Seven simulated datasets can be classified into four categories according to their performance. SSR and C4 were classified into the first category, both of them having a high vacancy rate, low accuracy rate and very low missing rate. The second category is represented by C1, C3, C5 and C6. The characteristics of this category include a vacancy rate that was slightly higher, especially for C1, C3 and north of the study area in C5 and C6. Most of the higher accuracy rate cities appeared in the center of the study area. The third category is represented by C2. Most of the cities had the highest accuracy rate, and the accuracy rate in most of the cities exceeded half the total samples. Almost no vacancy rate appeared in most of the cities. In all the cities, the highest IA appeared in the four scales, C1, C2, C3 and C5, and the frequencies of appearance of these scales were 7%, 57%, 32% and 4%, respectively; these results indicate that 57% of the cities had the highest IA with the C2 model, coinciding

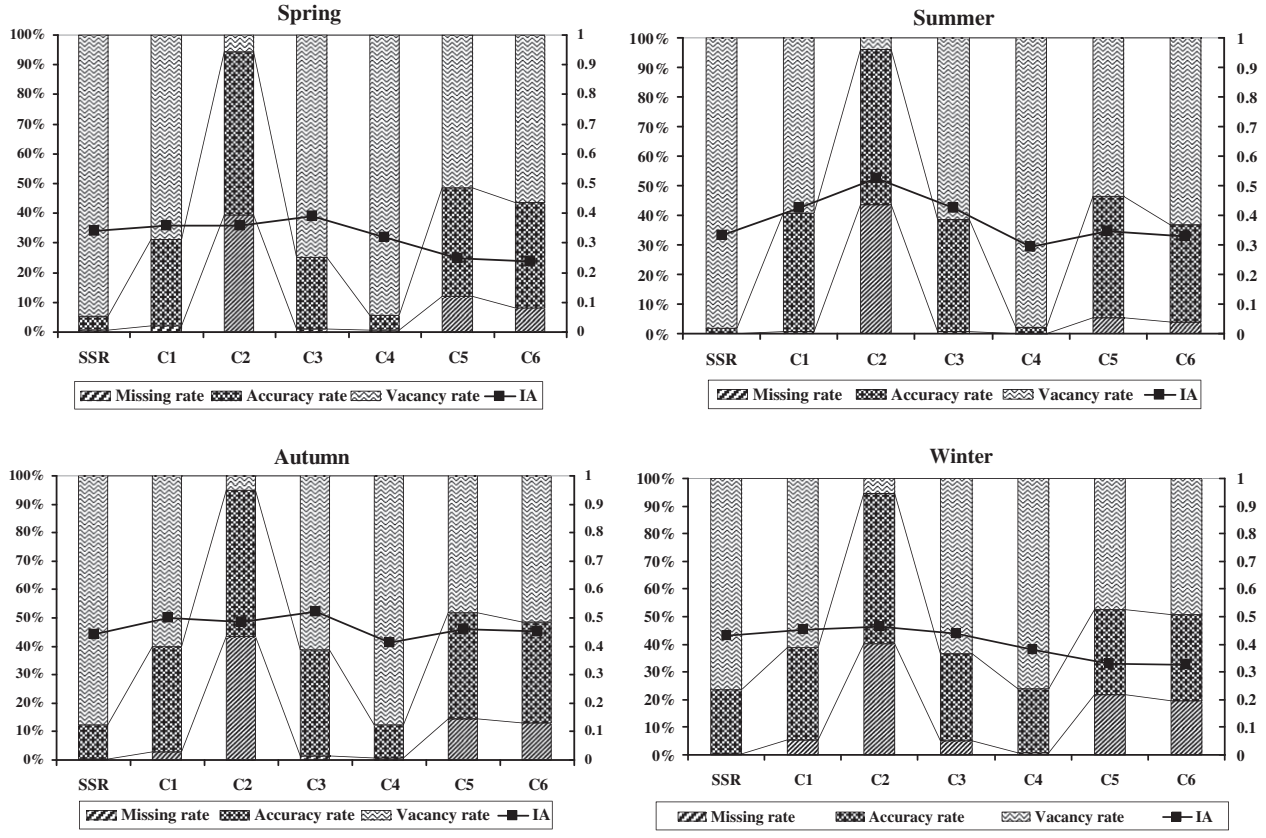

Fig. 7. The seasonal variations of accuracy rate, missing rate, vacancy rate and IA for the seven sets of PM_{10} concentration predictions by the enhanced regional model.

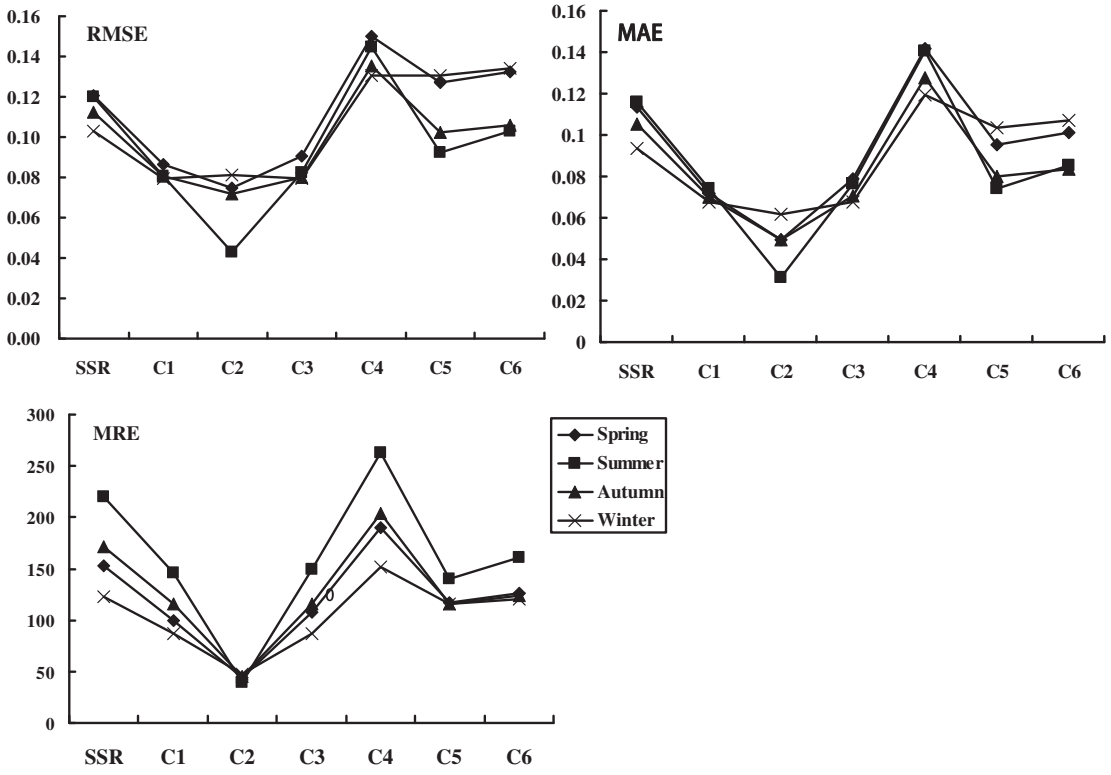

Fig. 8. The seasonal variations of RMSE, MAE and MRE for the seven sets of PM_{10} concentration predictions.

with the accuracy rate distribution. Thus, C2 had the advantage in predicting PM_{10} concentrations in eastern China with the highest accuracy rate and IA.

The distributions of RMSE, MAE and MRE with the regional model are shown in Fig. 6. The lowest RMSE in C2 was found in 82% of the cities. The lowest MAE in C2 was found in 89% of the cities in eastern China. Moreover, the MRE calculated from C2 was the lowest for all the cities. It can be found that C2 had the highest precision.

After reviewing all the performance evaluations above, C2 was found to be the most suitable scale for eastern China, as it had the lowest error. The advantage of the simulated result by C2, based on the regional model, was universal for eastern China.

3.1.4. Temporal applicability validation

One year was divided into four seasons as spring (March, April and May), summer (June, July and August), autumn (September, October and November) and winter (December, January and February). The accuracy rate, missing rate and vacancy rate calculated by the seven simulated results in each season are shown in Fig. 7. The highest accuracy rates in the four seasons were all appeared in C2. With the actual requirements of operational forecasting, a higher accuracy rate is preferred. C2 had the advantage of predicting PM_{10} concentrations at any time of the year with slight difference. The RMSE, MAE and MRE calculations also show that C2 had the highest precision (Fig. 8). Thus, PM_{10} concentration predicted by C2 based on the enhanced regional model was universal and temporally stable.

Wavelet decomposition had an advantage in decomposing non-stationary into stationary and regular signals with two components (detail and approximation components), the higher the decomposition level selected, the smoother and more sensitive the detailed and approximation component signals could be generated. However, errors appeared during decomposition, and higher decomposition levels always accompanied higher accumulated errors.

Table 2
Summarization of the independent variables used in each model of the 23 cities.

Variable	Stepwise regression	C1	C2	C3	C4	C5	C6	Best fitted scale
AOD	87.0%	91.6%	95.8%	95.8%	95.8%	100%	100%	100%
Surface temperature	17.4%	54.2%	54.2%	58.3%	87.5%	87.5%	70.8%	45.8%
Potential temperature	17.4%	58.3%	62.5%	58.3%	75.0%	87.5%	79.2%	58.3%
Precipitable water	34.8%	58.3%	79.2%	91.7%	87.5%	95.8%	100%	75%
Pressure	0%	29.2%	50.0%	62.5%	75.5%	79.2%	70.8%	37.5%
Relative humidity	56.5%	83.3%	83.3%	87.5%	100%	95.8%	87.5%	83.3%
Sea level pressure	17.4%	45.8%	62.5%	91.7%	70.8%	83.3%	91.7%	54.2%
u-wind	34.8%	75%	83.3%	87.5%	91.7%	95.8%	91.7%	62.5%
v-wind	34.8%	50%	66.7%	87.5%	87.5%	95.8%	95.8%	66.7%
Specific humidity	17.4%	58.3%	75.0%	79.2%	91.6%	100%	87.5%	66.7%
Total cloud cover	21.7%	62.5%	79.2%	79.2%	95.8%	100%	95.8%	70.8%

Table 3
Descriptive statistics of PM_{10} concentrations predicted by the city enhanced model.

PM_{10}	Minimum (mg m^{-3})	Maximum (mg m^{-3})	Mean (mg m^{-3})	Standard deviation	Kurtosis	Skewness
Observed	0.012	0.600	0.095	0.056	10.488	2.259
SSR	0.027	0.281	0.106	0.041	0.641	0.878
C1	0.008	0.380	0.090	0.038	2.728	0.823
C2	0.001	0.640	0.095	0.059	10.625	2.245
C3	0.001	0.681	0.112	0.074	4.780	1.599
C4	0.006	0.643	0.117	0.067	5.061	1.681
C5	0.000	0.623	0.116	0.074	2.343	1.092
C6	0.001	0.641	0.117	0.072	2.452	1.363

Thus, the selection of decomposition level was important in application. In this study, the optimum decomposition level must give consideration to both sensitivity and lowest error, while the $a = 2^2$ level of decomposition was the appropriate level to meet the application of the PM_{10} concentration forecast for the whole region.

3.2. Ensemble and enhanced PM_{10} concentration forecast model

PM_{10} concentrations in different cities may be affected by different meteorological factors. Thus, a specified and enhanced PM_{10} concentration prediction model was established for cities in eastern China based on local weather characteristics.

With the enhanced forecast model described in Section 2, seven sets of PM_{10} concentration predictions were obtained at each city.

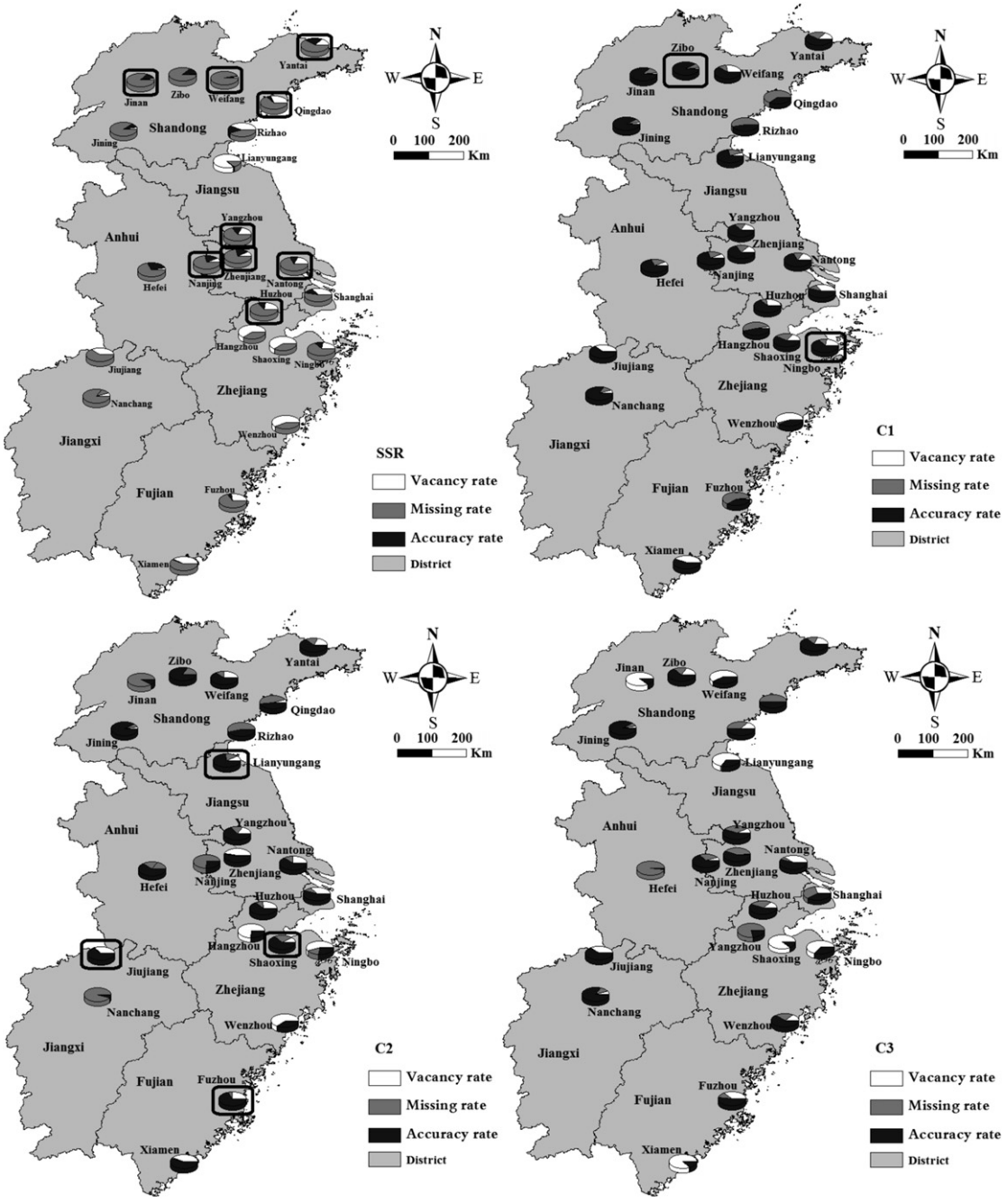

Fig. 9. The distribution of accuracy rate, missing rate, vacancy rate and IA of PM_{10} concentrations predicted by the enhanced city model.

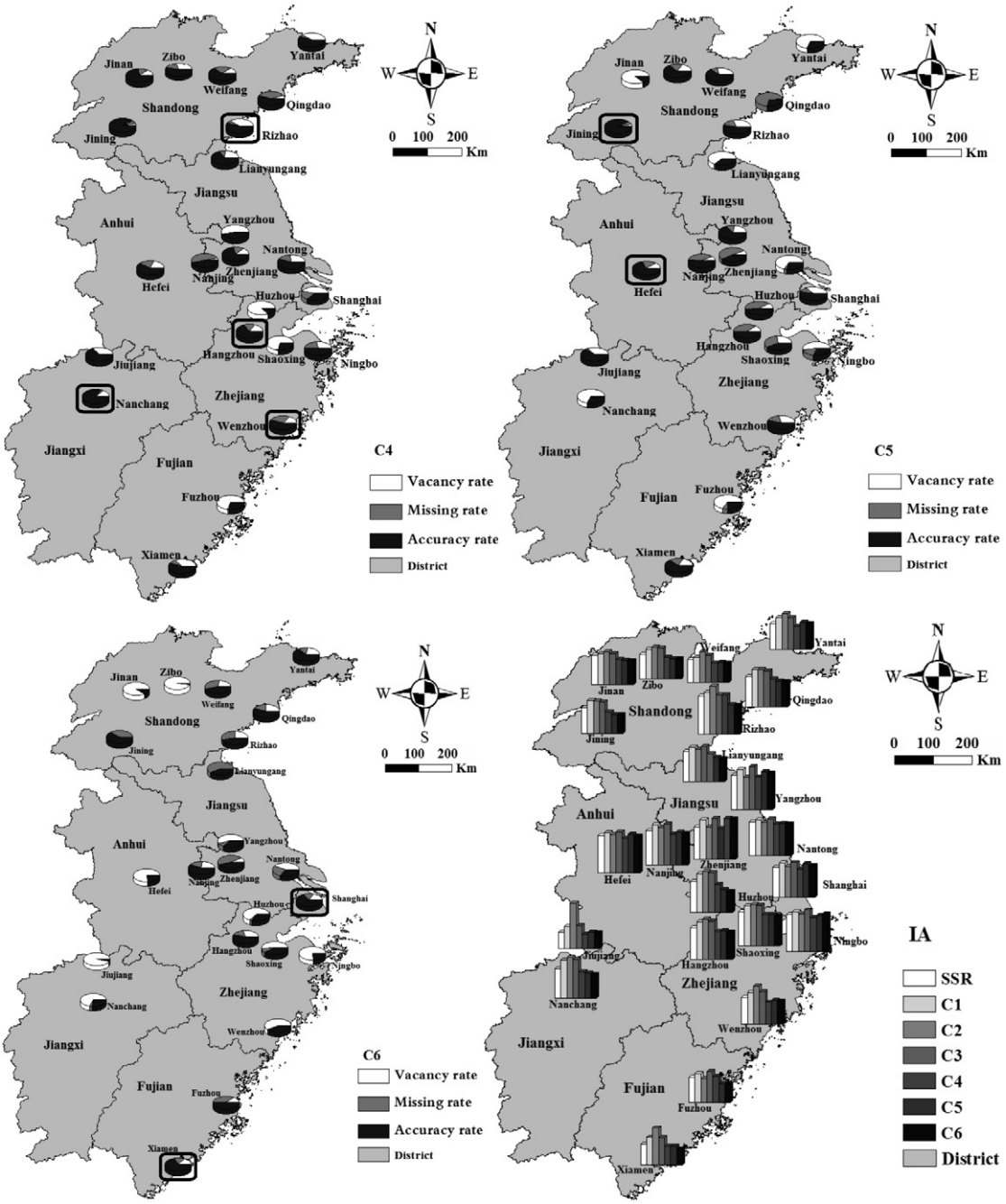

Fig. 9. (*continued*).

The independent variables used in the seven predicted model of each city were analyzed and summarized. Then, the accuracy rates and precision of each set of the PM_{10} concentration prediction were compared among the cities. Finally, the optimum model was selected for each city, and the final independent variables used in the optimum model of each city were analyzed. The best-fitted model in each city was combined in an ensemble to test the integrity of PM_{10} concentration forecasting in eastern China. Then, the improvements were analyzed with the application of the ensemble model.

3.2.1. Independent variables used in each city model

Enhance models were built and seven sets of PM_{10} concentration predictions were acquired for each city. The independent variables used in each model of each city were analyzed as shown in Table 2. As for stepwise regression, 87% of the cities were use AOD as an important variable, and 56.5% cities in eastern China were use relative humidity for predicting the PM_{10} concentration. Precipitable water, u-wind and v-wind were also used in most of the cities. As for other six prediction models, we can found that importance of AOD used for PM_{10} prediction was universal. Similarly with stepwise regression model, in most cities the importance of using relative humidity, precipitable water, u-wind and v-wind for PM_{10} prediction were also reflected in the six models. However, the important role of potential temperature, specific humidity and total cloud cover in predicting PM_{10} concentration was also shown with enhanced model. As potential temperature is an important parameter of temperature, it is conservatively during the dry

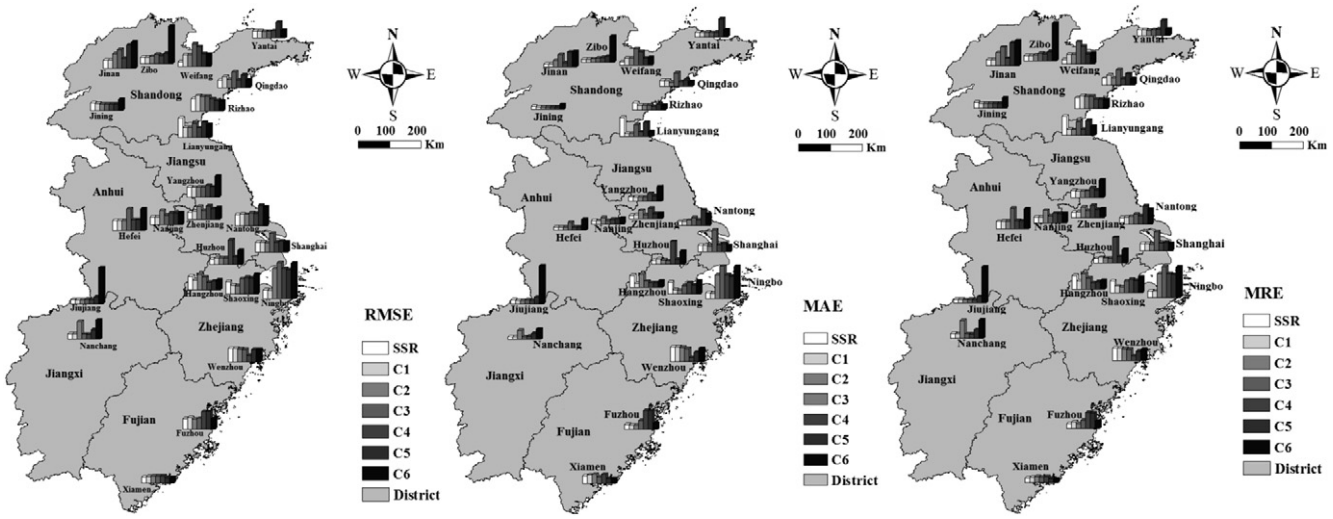

Fig. 10. Distributions and ranges of RMSE, MAE and MRE of each simulated result based on the enhanced city model.

adiabatic process. However, it is no longer conservative when the liquid water content in the atmosphere changed (Ertel, 1942). This change mainly caused by the release of latent heat in the wet precipitation process (Robison, 1989). Specific humidity reflected the liquid water content in the atmospheric. Higher total cloud cover usually appeared in rainy and cloudy day. All these were indirectly affected the PM_{10} concentration in terms of humidity and wet deposition. The humidity in the atmospheric will affect the concentration of PM_{10}. The meteorological factors which indirectly affect the PM_{10} concentration can be reflected with the filtering effect by wavelet analysis.

Overall, independent variables which had a great correlation with PM_{10} concentration, such as AOD, relative humidity, precipitable water, u-wind, and v-wind, also play an important role in PM_{10} concentration prediction in eastern China. It is accordance with other studies (Li et al., 2011). Meteorological factors which indirectly affect the PM_{10} concentration also play an important role in eastern China, such as potential temperature, specific humidity and total cloud cover.

3.2.2. *Characteristics of the PM_{10} concentrations predicted with the city model*

The descriptive statistics of the 161 prediction datasets with the enhanced model are shown in Table 3 and are compared with the observed dataset. The range of PM_{10} concentration predictions by SSR was comparatively small. The standard deviation was 0.041, which was slightly lower than the observed dataset and implies that the PM_{10} concentrations predicted by SSR had small variations. The dataset distribution predicted by SSR was a slope with a kurtosis of 0.641. Thus, the PM_{10} concentrations predicted by SSR were intensive, and it was difficult to forecast the abnormally high value with signal stepwise regression. For the PM_{10} concentration

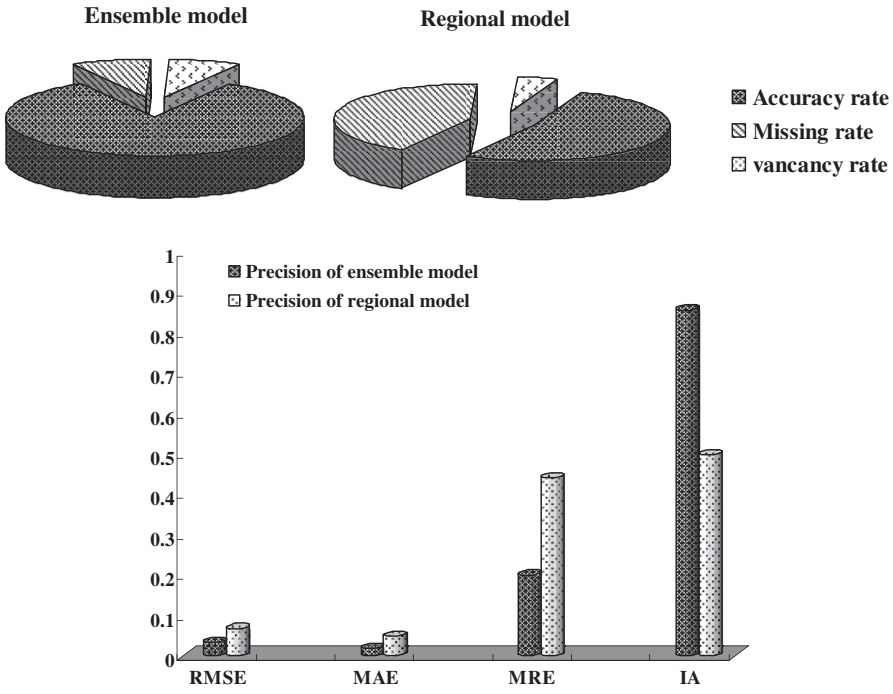

Fig. 11. Comparison of the precision of the ensemble and enhanced PM_{10} concentration forecast model-based city model with the precision of the best-fitted regional model.

predicted by C2, the range and value were similar to the observed dataset. Moreover, the distributions of the results predicted by C2 could reflect the distribution of the observed values on the whole, and their datasets also had similar variations. The PM_{10} concentrations predicted by C1 were slightly intensive with a small range, and the mean value was low, which implies that the distribution of the data was stable and did not corresponding to the actual air quality in the nature. The predicted results by C3, C4, C5 and C6 could be summarized as one category. These predicted results all had wide ranges, a higher mean value and wide variation. However, the kurtosis and skewness were smaller than the observed dataset, which implies that the number of abnormal data was more than that in the observed dataset and that most of them were not right-skewed. Thus, PM_{10} forecasting by combined wavelet analysis and stepwise regression has an advantage in predicting the abnormal value, especially for C2.

3.2.3. The ensemble and enhanced PM_{10} concentration forecast model

The accuracy rate, missing rate, vacancy rate and IA of the seven sets of PM_{10} concentration predictions in each city were calculated (Fig. 9).

The accuracy rate in 65.2% of the cities observed was improved by the combined model. The prediction accuracy rates for Zibo and Jining reached 90.8% and 90.8%, respectively. Based on the combined model, the accuracy rate of predictions for Lianyungang, Shaoxing, Hangzhou, Wenzhou, Xiamen and Shanghai had an increase of 60.8%, 25.4%, 27.3%, 20.6%, 10.5% and 11.5%, respectively. The results show that the enhanced model was able to improve the accuracy rate of the PM_{10} forecast and that the improvements in most of the cities were significant. When comparing the accuracy rate and IA of the seven predicted models, it was found that highest IA of each city was in accordance with the highest accuracy rate.

The precision evaluations by RMSE, MAE and MRE were also calculated and are shown in Fig. 10. The scale distributions of the lowest RMSE, MAR and MRE were coincident with each other with highest accuracy rate. Then, the best fitted scales were selected for each city by comparing the precision of each prediction model, and they were integrated as an ensemble model. In total, the mainly used variables in best fitted model were AOD, relative humidity, precipitable water, total cloud cover, specific humidity, v-wind and u-wind (Table 2). The performances of the ensemble model were compared with the regional model (Fig. 11). The accuracy of the ensemble model was 83.5%. The accuracy rate had a 31% aggregation ratio by the ensemble model when compared with the PM_{10} concentrations predicted by C2 based on the regional model. When comparing the IA, it was obvious that the IA was greatly improved, as the IA increased to 0.855 in the ensemble model. Meanwhile, the RMSE, MAE and MRE sharply decreased to 0.033, 0.018 and 0.2, respectively. It can be inferred that the ensemble and enhanced PM_{10} concentration forecast models for eastern China were useful for precision improvement and that the improvement was very significant.

4. Conclusion

In this study, an ensemble and enhanced PM_{10} concentration forecast model was formed based on stepwise regression and wavelet analysis in eastern China.

Based on the regional scale, seven sets of PM_{10} concentration predictions were obtained by the regional model. The calculation of the accuracy rate based on the requirements of operational forecasting confirmed that the combined forecast model had an advantage in PM_{10} concentration forecasting in eastern China. Precision evaluations for the seven predicted results also showed that the precision obtained by the combined model, especially in scale 2, increased significantly. The advantage of C2 based on the regional model was spatially and temporally universal.

Based on the city scale, the enhanced model for each city was established in eastern China. Overall, the characteristics and distributions of the observed data can be reflected by C2 in each city. The predicted data obtained by single stepwise regression were intensively concentrated, which cannot reflect the actual air pollution. The accuracy rate in 65.2% of the cities was improved with the enhanced model.

An obvious improvement was achieved by the best-fitted model, which was selected for each city. The ensemble of the PM_{10} concentration forecast model with the highest accuracy rate had the best precision. The ensemble and enhanced PM_{10} concentration forecast model proved to be a new and effective model with significant accuracy enhancement and precision improvement in eastern China. In eastern China, AOD, relative humidity, precipitable water, total cloud cover, specific humidity, v-wind and u-wind were played an important role in PM_{10} concentration prediction in most of the cities in eastern China.

Acknowledgments

We would like to thank Earth System Research Library of NOAA for the use of the meteorological data was downloaded from the website. We would like to thank NASA Data Distribution and Archive Center for processing and providing MODIS AOD products. This research was funded by the Shanghai Science and Technology Committee (Grant No. 10DZ0581600), the National Basic Research Program of China (Grant No. 2010CB951603), and the National Science Foundation of China (Grant No. 41201358). We also thank the anonymous reviewers who help us improve the manuscript.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.atmosenv.2013.04.002.

References

Baklanov, A., Mestayer, P.G., Clappier, A., Zilitinkevich, S., Joffre, S., Mahura, A., Nielsen, N.W., 2008. Towards improving the simulation of meteorological fields in urban areas through updated/advanced surface fluxes description. Atmospheric Chemistry and Physics 8, 523–543.

Bravo, M.A., Bell, M.L., 2011. Association spatial heterogeneity of PM_{10} and O_3 in São Paulo, Brazil, and implications for human health studies. Journal of the Air & Waste Management 61, 69–77.

Bruce, L.M., Morgan, C., Larsen, S., 2001. Automated detection of subpixel hyperspectral targets with continuous and discrete wavelet transforms. IEEE Transactions on Geoscience and Remote Sensing 39, 2217–2226.

Bruce, L.M., Koger, C.H., Jiang, L., 2002. Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. Geoscience and Remote Sensing, IEEE Transactions 40, 2331–2338.

Bruckman, L., 1993. Overview of the enhanced geocoded emissions modeling and projection (Enhanced GEMAP) system. In: Proceeding of the Air & Waste Management Association's Regional Photochemical Measurements and Modeling Studies Conference, p. 562. San Diego, CA.

Coats, C.J., 1996. High performance algorithms in the sparse matrix operator kernel emissions modeling system. In: Proceedings of the Ninth Joint Conference on Applications of Air Pollution Meteorology of the American Meteorological Society and the Air and Waste Management Association. Atlanta, GA.

Contini, D., Genga, A., Cesari, D., Siciliano, M., Donateo, A., Bove, M.C., Guascito, M.R., 2010. Characterisation and source apportionment of PM_{10} in an urban background site in Lecce. Atmospheric Research 95, 40–54.

Durka, P.J., 2003. From wavelets to adaptive approximations: time–frequency parametrization of EEG. BioMedical Engineering OnLine 2, 1. http://dx.doi.org/10.1186/1475-925X-2-1.

Eriksson, L., Trygg, J., Johansson, E., Bro, R., Wold, S., 2000. Orthogonal signal correction, wavelet analysis, and multivariate calibration of complicated process fluorescence data. Analytica Chimica Acta 420, 181–195.

Ertel, H., 1942. Ein neuer hydrodynamischer Erhaltungssatz. Die Naturwissenschaften 36, 543–544.

Hoi, K.I., Yuen, K.V., Mok, K.M., 2009. Prediction of daily averaged PM_{10} concentrations by statistical time-varying model. Atmospheric Environment 43, 2579–2581.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O., 2005. A neural network forecast for daily average PM_{10} concentrations in Belgium. Atmospheric Environment 39, 3279–3289.

Jeong, J.I., Park, R.J., Woo, J.H., Han, Y.J., Yi, S.M., 2011. Source contributions to carbonaceous aerosol concentrations in Korea. Atmospheric Environment 45, 1116–1125. http://dx.doi.org/10.1016/j.atmosenv.2010.11.031.

Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J., Shao, D., 2004. Progress in developing an ANN model for air pollution index forecast. Atmospheric Environment 38, 7055–7064.

Kang, D., Eder, B.K., Stein, A.F., Grell, G.A., Peckham, S.E., McHenry, J., 2005. The New England air quality forecasting pilot program: development of an evaluation protocol and performance benchmark. Journal of Air and Waste Management Association 55, 1782–1796.

Kharol, S.K., Badarinath, K.V.S., Sharma, A.R., Kaskaoutis, D.G., Kambezidis, H.D., 2011. Multiyear analysis of Terra/Aqua MODIS aerosol optical depth and ground observations over tropical urban region of Hyderabad, India. Atmospheric Environment 45, 1532–1542.

Kim, C., Yu, I., Song, Y.H., 2002. Prediction of system marginal price of electricity using wavelet transform analysis. Energy Conversion and Management 13, 1839–1851.

Kim, Y., Fu, J.S., Miller, T.L., 2010. Improving ozone modeling in complex terrain at a fine grid resolution: part I – examination of analysis nudging and all PBL schemes associated with LSMs in meteorological model. Atmospheric Environment 44, 523–532.

Künzli, N., Kaiser, R., Medina, S., Studnicka, M., Chanel, O., Filliger, P., Herry, M., Horak Jr., F., Puybonnieux-Texier, V., Quénel, P., Schneider, J., Seethaler, R., Vergnaud, J.-C., Sommer, H., 2000. Public-health impact of outdoor and traffic-related air pollution: a European assessment. The Lancet 356, 795–801.

Kurt, A., Oktay, A.B., 2010. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Systems with Applications 37, 7986–7992.

Li, C., Hsu, N.C., Tsay, S.-C., 2011. A study on the potential applications of satellite data in air quality monitoring and forecasting. Atmospheric Environment 45, 3663–3675.

Lurmann, F.W., 2000. Simplification of the UAMAERO Model for Seasonal and Annual Modeling: the UAMAERO-LT Model. Report prepared for South Coast Air Quality Management District, Diamond Bar, CA by Sonoma Technology, Inc., Petaluma, CA, STI-999420-1996-FR, August.

Mallat, S., 1999. A Wavelet Tour of Signal Processing, second ed. Academic Press, London.

Manders, A.M.M., Schaap, M., Hoogerbrugge, R., 2009. Testing the capability of the chemistry transport model LOTOS-EUROS to forecast PM_{10} levels in the Netherlands. Atmospheric Environment 43, 4050–4059.

Mishchenko, M.I., Geogdzhayev, I.V., Cairns, B., Carlson, B.E., Chowdhary, J., Lacis, A.A., Liu, L., Rossow, W.B., Travis, L.D., 2007. Past, present, and future of global aerosol climatologies derived from satellite observations: a perspective. Journal of Quantitative Spectroscopy and Radiative Transfer 106, 325–347.

Murtagh, J., Starch, L., Renaud, O., 2004. On neuron-wavelet modeling. Decision Support Systems 37, 475–484.

Qiu, J., Sun, X., Suo, S., Shi, S., Huang, S., Liang, R., Zhang, L., 2011. Predicting homo-oligomers and hetero-oligomers by pseudo-amino acid composition: an approach from discrete wavelet transformation. Biochimie 93, 1132–1138.

Qu, W.J., Arimoto, R., Zhang, X.Y., Zhao, C.H., Wang, Y.Q., Sheng, L.F., 2010. Spatial distribution and interannual variation of surface PM_{10} concentrations over eighty-six Chinese cities. Atmospheric Chemistry and Physics 10, 5641–5662.

Querol, X., Alastuey, A., de la Rosa, J., Sanchez-de-la-Campa, A., Plana, F., Ruiz, C.R., 2002. Source apportionment analysis of atmospheric particulates in an industrialised urban site in southwestern Spain. Atmospheric Environment 36, 3113–3125.

Robison, W.A., 1989. On the structure of potential vorticity in baroclinic instability. Tellus 41A, 275–284.

Saliba, N.A., Jam, F.E., Tayar, G.E., Obeid, W., Roumie, M., 2010. Origin and variability of particulate matter (PM_{10} and $PM_{2.5}$) mass concentrations over an Eastern Mediterranean city. Atmospheric Research 97, 106–114.

Salvador, P., Artinano, B., Viana, M.M., Querol, X., Alastuey, A., Gonzalez-Fernandez, I., Alonso, R., 2011. Spatial and temporal variations in PM_{10} and $PM_{2.5}$ across Madrid metropolitan area in 1999-2008. Procedia Environmental Sciences 4, 198–208.

Senaratne, I., Kelliher, F.M., Triggs, C.M., 2005. Source apportionment of PM_{10} during cold, calm weather in Christchurch, New Zealand: preliminary results from a receptor model. Clean Air and Environmental Quality 39, 47–54.

Singh, H.B., 1995. Composition, Chemistry, and Climate of the Atmosphere. Van Nostrand Reinhold, New York, ISBN 0-442-01264-0.

Stern, R., Builtjes, P., Schaap, M., Timmermans, R., Vautard, R., Hodzic, A., Memmesheimer, M., Feldmann, H., Renner, E., Wolke, R., Kerschbaumer, A., 2008. A model inter-comparison study focussing on episodes with elevated PM_{10} concentrations. Atmospheric Environment 42, 4567–4588.

Torrence, C., Compo, G.P., 1998. A practical guide to wavelet analysis. Bulletin of the American Meteorological Society 71, 61–78.

Vautard, R., Builtjes, P.H.J., Thunis, P., Cuvelier, C., Bedogni, M., Bessagnet, B., Honoré, C., Moussiopoulos, N., Pirovano, G., Schaap, M., Stern, R., Tarrason, L., Wind, P., 2007. Evaluation and intercomparison of Ozone and PM_{10} simulations by several chemistry transport models over four European cities within the CityDelta project. Atmospheric Environment 41, 173–188.

Zhang, G.C., 2004. Progress of weather research and forecast (WRF) model and application in the United States. Meteorological 30, 27–31 (in Chinese).

Zhang, Z., Lu, C., Zhang, F., Ren, Y., Yang, K., Su, Z., 2012. A Novel method for non-contact measuring diameter parameters of wheelset based on wavelet analysis. Optik 123, 433–438.