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] Department of Ecosystem Science and Sustainability, 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