Aerosol and Air Quality Research

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.210247

Multi-model Evaluation and Bayesian Model Averaging in Quantitative Air Quality Forecasting in Central China

Haixia Qi[1], Shuangliang Ma[2*], Jing Chen[2], Junping Sun[2], Lingling Wang[2], Nan Wang[2], Weisi Wang[2], Xiefei Zhi[3*], Hao Yang[1]

[1] Hubei Key Laboratory for Heavy Rain Monitoring and Warning Research, Institute of Heavy Rain, China Meteorological Administration, Wuhan, China
[2] Henan Key Laboratory of Environmental Monitoring Technology, Henan Ecological Environment Monitoring Center, Henan, China
[3] Key Laboratory for Aerosol-Cloud-Precipitation of China Meteorological Administration, Nanjing University of Information Science and Technology, Nanjing, China

ABSTRACT

There has been much interest in air pollution and the forecasting skill of air quality models in China since winter 2013. Different air quality models use different parameters (e.g., meteorological fields, emission sources and the initial concentrations of pollutants) and therefore their forecast results tend to have large systematic and random errors. We evaluated the concentrations of six pollutants in Henan Province predicted by three air quality models—the China Meteorological Administration Unified Atmospheric Chemistry Environment (CUACE) model, the Nested Air Quality Prediction (NAQP) model and the Community Multiscale Air Quality (CMAQ) model. We then established multi-model ensemble Bayesian model averaging (BMA). The prediction effect for $PM_{2.5}$ and O_3 was ranked as CUACE > CMAQ > NAQP and the prediction effect for SO_2, NO_2 and CO was CMAQ > NAQP > CUACE. All the models systematically underestimated O_3 and heavy $PM_{2.5}$ pollution events. For $PM_{2.5}$ concentrations with a 24-h lead time, the root-mean-square error of BMA decreased by 35, 37, 68 and 50%, respectively, in winter, spring, summer and autumn relative to the CUACE model, whereas the normalized mean bias of BMA decreased by 67, 83, 94 and 55%, respectively, for O_3 in the four seasons. Compared with the CMAQ model, the root-mean-square error of the SO_2, NO_2 and CO forecasts by BMA were reduced by 29, 33 and 39%, respectively. The evolution of the concentrations of the six pollutants during a heavy pollution event predicted by BMA was consistent with the observations.

Keywords: Pollutant concentrations, Pollution episodes, Model evaluation, Bayesian model averaging

OPEN ACCESS

Received: October 6, 2021
Revised: December 22, 2021
Accepted: March 28, 2022

* **Corresponding Authors:**
Shuangliang Ma
573537681@qq.com
Xiefei Zhi
xf_zhi@163.com

Publisher:
Taiwan Association for Aerosol Research
ISSN: 1680-8584 print
ISSN: 2071-1409 online

1 INTRODUCTION

The prediction of air pollution involves a number of different disciplines (e.g., meteorology, physics and chemistry) and different macro- and micro-processes, resulting in a series of complex problems (Tai *et al.*, 2010; Zhang *et al.*, 2015; Zhu *et al.*, 2018; Zhong *et al.*, 2018; Li *et al.*, 2019; Zhai *et al.*, 2019; Chen *et al.*, 2021). Research and the development of technologies to forecast air quality will provide a reference for decision-making and provide effective technical support for governments to release warnings of air pollution events, to provide emergency response and support legislation to reduce emissions.

Numerical methods to predict air quality are important in providing warnings about air pollution. There are three main approaches to forecasting air quality. The first method is to make a numerical prediction based on theories of atmospheric dynamics, physics and chemistry

(Galmarini *et al.*, 2012). The second method is to base predictions on the long-term statistical laws between air pollutants and meteorological elements (Tandon *et al.*, 2013; Singh *et al.*, 2013; Sun and Sun, 2016; Huang *et al.*, 2018; Shishegaran *et al.*, 2020; Gao *et al.*, 2020). The third method is to carry out statistical ensemble analyses based on the results of multiple numerical predictions of pollutant concentrations or to perform spectral decomposition of the multi-model prediction results to reconstruct the model (Monache *et al.*, 2020; Monteiro *et al.*, 2013; Galmarini *et al.*, 2013; Donnelly *et al.*, 2015; Huang *et al.*, 2017).

The Community Multiscale Air Quality (CMAQ) model is a relatively mature and well-known numerical prediction model (Wang *et al.*, 2006). The three air quality models commonly used in China are: (1) the Nested Air Quality Prediction Modelling System (NAQP), which was independently developed by the Institute of Atmospheric Physics, Chinese Academy of Sciences; (2) the China Meteorological Administration (CMA) Unified Atmospheric Chemistry Environment (CUACE) atmospheric chemistry system developed by the CMA; and (3) the Urban Air Quality Numerical Prediction System (CAPPS) developed by the Chinese Academy of Meteorological Sciences. Localized models based on the Weather Research and Forecasting (WRF)-Chem and CMAQ models are also used.

The prediction performances of these models have been evaluated by a number of research groups (Mebust *et al.*, 2003; Emmons, 2010; Rao *et al.*, 2011; Galmarini *et al.*, 2012; Nopmongcol *et al.*, 2012, 2017). Nopmongcol *et al.* (2012) evaluated the Comprehensive Air Quality Model with Extensions (CAMx) and showed that all pollutants, except SO_2, were underestimated in winter and summer. The prediction of NO_x and NO_2 was better in winter than in summer in January and July 2006. The winter O_3 concentrations were low, which was attributed to a low bias in the O_3 boundary conditions. PM_{10} was widely under-predicted in both winter and summer. Zhu *et al.* (2015) evaluated the performance of the NAQP model in terms of the 24-h forecast and 7-day potential forecast for the daily mean concentration of $PM_{2.5}$ during summer 2013 and showed that the statistical indicators for the 24-h forecast satisfied the performance criteria for the mean fractional bias and mean fractional error. The model reproduced the tendency of the $PM_{2.5}$ concentrations in the 7-day potential forecast well. However, there were regular systematic errors among the models for the prediction of pollutant concentrations.

Different models use different parameterization schemes (e.g., meteorological field driving, emission sources and the initial concentrations of pollutants), so there are uncertainties in the model predictions. Significant systematic and random deviations are also found between the model forecasts and observations. To minimize the uncertainty in the model products and to make full use of all types of data to improve the accuracy of the element forecasts, multi-model ensemble technology has been adopted to post-process the multiple forecasting system products (Monache *et al.*, 2020; Monteiro *et al.*, 2013; Galmarini *et al.*, 2013; Donnelly *et al.*, 2015; Huang *et al.*, 2017).

A large number of different approaches have been proposed, from simple averaging of the results, the construction of a median model and the application of weights derived from past skills scores or Bayesian model averaging (BMA) theory (Raftery *et al.*, 2005, 2013; Knutti *et al.*, 2010; Rahman *et al.*, 2015; Sun and Sun, 2016; Huang *et al.*, 2018). Galmarini *et al.* (2012) applied the Kolmogorov-Zurbenko filter to 13 air quality models that had been spectrally decomposed and the observational ozone concentrations over multiple European and North American sub-regions. They found that the composite model built from the best spectral elements outscored all the ensemble members and the ensemble median. Li *et al.* (2020) designed a dynamic integration forecasting air quality index (AQI) model based on the Box-Jenkins autoregressive integrated moving average, an optimized extreme learning machine and a fuzzy time series to forecast the reconstructed series and time-varying parameters. They found that the ensemble forecast effect was superior to the single model.

Henan Province is one of the most polluted regions in China. Studies of air quality forecast technology in this region are important to improve our ability to deal with heavy pollution events. We used the parameterized statistical post-processing BMA method to aggregate the results from different air quality prediction models (Hoeting *et al.*, 1999). Raftery *et al.* (2005) applied the BMA method to several dynamic statistical models to predict temperature and sea-level pressure fields in a normal distribution. Sloughter *et al.* (2007, 2010) extended the BMA method to quantitatively predict precipitation and wind speed in skewed distributions. However, this method has seldom

been used in the prediction of pollutant concentrations and its applicability to different lead times and different pollutant elements needs further study.

We evaluated the prediction performance for pollutant concentrations of the three commonly used models (CUACE, NAQP and CMAQ) in Henan Province, China. We then used the multi-model ensemble BMA method to aggregate the prediction results of the three models and compared the ensemble prediction results with the original single-model outputs and the model mean prediction to provide technical support for improving the accuracy of air quality forecasting.

2 DATA AND METHODS

2.1 Observational and Air Quality Model Data

The data were provided by Henan Provincial Environmental Monitoring Center and included the daily concentrations of six pollutants ($PM_{2.5}$, PM_{10}, O_3, NO_2, CO and SO_2) collected by meteorological monitoring stations in 18 cities within the National Air Quality Control Network from 2017 to 2019 (Fig. 1). The pollutant concentrations were predicted by three air quality prediction models (CUACE, NAQP and CMAQ; Table 1). The CUACE and NAQP models were run from 1 January 2017 to 31 December 2019 and the CMAQ model was run from 3 November 2018 to 6 December 2019. The meteorological field of the CUACE model is driven by elements of the MM5 mesoscale model with a two-layer nested model (27 km/9 km); we used the 9 km data. The meteorological fields of the NAQP and CMAQ models are driven by the meteorological field of the Weather Research and Forecasting (WRF) mesoscale model, which adopts a three-layer nested model of 45 km/15 km/5 km; we used the 5 km data. All three models adopt the Multi-resolution Emission Inventory for China and the Henan Province local compilation list (2017) gridded by the air quality model of Tsinghua University. The lead time of the pollutant concentration forecasts is 168 h.

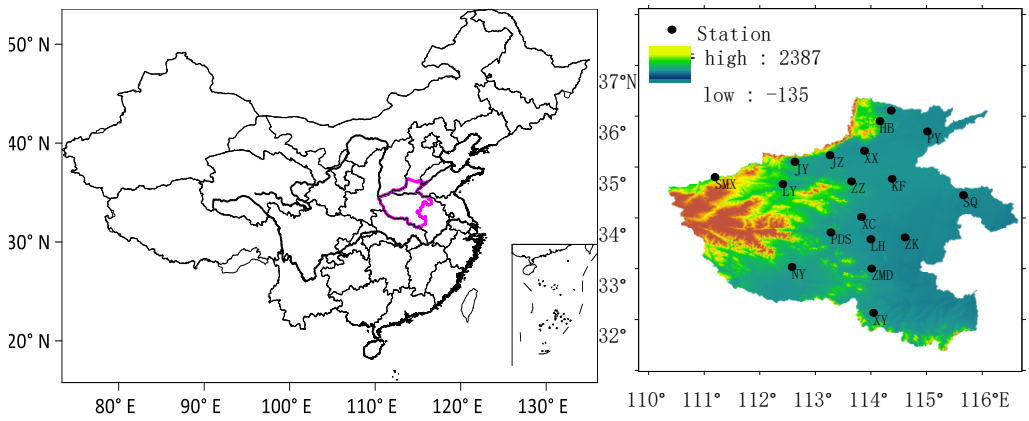

Fig. 1. Location of the study area in China and the 18 meteorological stations (dots) in Henan Province. Topographic map of study area. Shading represents surface elevation (units: m).

Table 1. Description of the three air quality models (CUACE, NAQP and CMAQ).

Model	Driving source of meteorological field	Diffusion scheme (vertical/horizontal)	Chemical mechanisms of gases	Chemical mechanism of aerosols	Emission source inventory	Spatial resolution (km)	Forecast lead time (h)
CUACE	MM5	K-theory/K-theory	RADM2	CAM/ISORROPIA	MEIC + local compilation (2017)	27 × 9	168
NAQP	WRF	K-theory/K-theory	CBM-Z	ISORROPIA	MEIC + local compilation (2017)	45 × 15 × 5	168
CMAQ	WRF	ACM2/eddy diffusion theory	CBM-IV	AE7/ISORROPIA	MEIC + local compilation (2017)	45 × 15 × 5	168

2.2 Introduction to the BMA Method

The BMA method generates a probability density function (PDF) by combining multiple statistical models for ensemble inference and prediction:

$$P\left(y|f_1,...,f_k\right)=\sum_{k=1}^{K}w_k g_k\left[y|\left(f_k,y^T\right)\right] \tag{1}$$

where w_k is the posterior probability of each ensemble member or weight during the training period and $\sum_{k=1}^{K}w_k=1$. $g_k[y|(f_k, y^T)]$ is the PDF of the forecast variable y during the training period y^T when the forecast model f_k is under the best forecasting condition.

Following Sloughter *et al.* (2007), the PDF for the pollutant concentration in Eq. (1) includes two parts. The first part is the probability of a zero concentration of pollutant as a function of the model result using a logistic regression model and the second part is the PDF when the concentration of the pollutant is nonzero.

The first part $P(y = 0|f_k)$ is the probability of a zero concentration of pollutant as a function of f_k using a logistic regression model:

$$\text{logit}\left[P\left(y=0|f_k\right)\right]=\log\frac{P\left(y=0|f_k\right)}{P\left(y>0|f_k\right)}=a_{0k}+a_{1k}f_k+a_{2k}\delta_k \tag{2}$$

where δ_k is an indicator and equals 1 if $f_k = 1$ and equals 0 otherwise. $P(y = 0|f_k)$ is the conditional probability of a zero concentration of pollutant and $P(y > 0|f_k)$ is the probability of a nonzero concentration of pollutant given the forecast f_k. The parameters a_{0k}, a_{1k}, a_{2k} were estimated using the Newton-Raphson iterative method against the training period.

This paper focuses on the second part—namely, the PDF when the amount of pollutant concentration is nonzero. Pollutant concentration is a discontinuous variable so we used a gamma distribution function for fitting (Hamill *et al.*, 2004; Qi *et al.*, 2019):

$$g_k\left(y|f_k\right)=\frac{1}{\beta_k^{\alpha_k}\Gamma\left(\alpha_k\right)}y^{\alpha_k-1}\exp\left(-y/\beta_k\right) \tag{3}$$

$$\mu_k=\alpha_k\beta_k=b_{0k}+b_{1k}f_k^{1/3}\quad \sigma_k^2=\alpha_k\beta_k^2=c_{0k}+c_{1k}f_k \tag{4}$$

where α_k and β_k are the shape and scale parameters of the gamma distribution. μ_k and σ_k^2 are the mean and variance of the gamma distribution. Their relations are $\alpha_k=\mu_k^2/\sigma_k^2$ and $\beta_k=\sigma_k^2/\mu_k$. The values of μ_k and σ_k can be estimated based on f_k through the relationships of Eq. (4). b_{0k} and b_{1k} were estimated by linear regression to observed nonzero cases.

w_k, c_0 and c_1 were estimated by maximizing the log-likelihood function using the expectation-maximization algorithm:

$$l\left(w_1,...,w_K;c_0,c_1\right)=\sum_t\log p\left(y_{st}|f_{1st},...,f_{Kst}\right) \tag{5}$$

where $p(y_{st}|f_{1st}, ..., f_{Kst})$ is calculated using Eq. (1), y_{st} is the observational data, f_{Kst} is forecasting value of each ensemble model and the s and t indexes correspond to space and time in the training period. The PDF of the BMA multi-model ensemble is then:

$$h_k\left(y|f_k\right)=P\left(y>0|f_k\right)g_k\left(y|f_k\right) \tag{6}$$

where y is the observational concentration of the pollutant. The PDF of all ensemble members

obtained in the BMA multi-model ensemble is:

$$PDF\left(y|f_1,\ldots,f_K\right)=\sum_{k=1}^{K}w_k\left[P\left(y>0|f_k\right)g_k\left(y|f_k\right)\right] \tag{7}$$

where w_k is the posterior probability of ensemble member k being the best value.

We used the deterministic forecast result of BMA, which is the median of the BMA predictive PDF. The length of the sliding training period was 30 days.

2.3 Evaluation Methods

The mean bias (MB)refers to the difference between the pollutant concentration predicted by the numerical model and the observational data:

$$MB=\frac{1}{N}\sum_{i=1}^{n}\left(F_i-O_i\right) \tag{8}$$

The normalized mean bias (NMB) is used to standardize the mean bias, which can avoid the problem of over-dispersion of the observed value range:

$$NMB=\frac{\sum_{i=1}^{n}\left(F_i-O_i\right)}{\sum_{i=1}^{n}O_i}\times 100\% \tag{9}$$

The root-mean-square error (RMSE) reflects the deviation between the observed value and the simulated value:

$$RMSE=\left[\frac{1}{N}\sum_{i=1}^{n}\left(F_i-O_i\right)^2\right]^{\frac{1}{2}}\quad\left(i=1\cdots N\right) \tag{10}$$

In Eqs. (8), (9) and (10), O_i is the observed value of sample i, F_i is the numerical simulation forecast value and N is the total number of temporal and spatial samples. The smaller the result, the smaller the error between the predicted and observed values and the better the prediction effect of the model.

The correlation coefficient (COOR) is:

$$R\left(F,O\right)=\frac{Cov\left(F,O\right)}{\sqrt{Var\left[F\right]\cdot Var\left[O\right]}} \tag{11}$$

when $R>0$, this indicates that the two variables are positively correlated; $R<0$ indicates that the two variables are negatively correlated. When $|R|\leq 1$, the closer $|R|$ is to 1, the greater the degree of correlation and the closer $|R|$ is to 0, the smaller the degree of correlation. O is the observed value of the samples and F is the value of the numerical simulation forecast.

3 RESULTS

3.1 Evaluation of Model Simulation Forecasts

The concentrations of pollutants in China show significant regional and seasonal variations. Heavy pollution processes, mainly caused by $PM_{2.5}$, often occur in autumn, winter and early spring. The peak of O_3 pollution generally occurs in summer. We selected the winter data from the three models for comparison. Fig. S1 compares the model-predicted 24 h concentrations of

five pollutants in winter from December 2018 to February 2019 and the concentrations of O_3 in summer (June–August) 2019 with their observed concentrations (OBS). The forecasts of $PM_{2.5}$ concentrations (Fig. S1(a)) by all three models were consistent with the observed trend, with the values varying in the range 50–250 μg m^{-3}. The daily variation trend of the observed concentration of PM_{10} (Fig. S1(b)) was similar to that of $PM_{2.5}$. The CUACE model gave a better forecast, whereas the NAQP forecast was unstable and the CMAQ forecast tended to underestimate the concentrations of particulate matter.

The prediction results of the three models had similar systematic biases for the concentrations of SO_2 and CO (Figs. S1(c) and S1(e)). The predicted values of the CUACE and NAQP models showed a significant positive system bias, whereas the CMAQ model had negative system errors in the forecast values of the two gaseous pollutants. The results of the CUACE model were more biased than those of the NAQP model. The predictions of NO_2 concentrations (Fig. S1(d)) by all the three models showed higher positive systematic deviations, of which the positive deviation of the CUACE model was the largest. For prediction of O_3 concentrations (Fig. S1(f)), the CMAQ model gave lower concentrations than the observations, whereas the forecast values of the other two models were closer to the observations.

The three models performed well in predicting the trends and values of $PM_{2.5}$ and PM_{10} concentrations in the period of frequent pollution, with the CUACE giving the best results. The value of PM_{10} predicted by the CMAQ model was generally lower than the observed value. For the two gaseous pollutants (SO_2 and CO), the CUACE and NAQP models both showed significant positive system deviations, whereas the CMAQ model showed a negative deviation. The prediction results of all three models showed a positive systematic bias for NO_2, with the CUACE model showing the largest deviation. For O_3, the predicted value from the CMAQ model was lower than the observations.

Air quality is closely correlated with the meteorological conditions, which vary greatly in different seasons. How does the prediction effect of the models change with the season? The period covered by the data can affect the prediction and evaluation of the models. The data for the prediction of CMAQ cover about one year, whereas the data for the NAQP and CUACE models cover about three years. We therefore only present the results of the CUACE and NAQP models to accurately compare and evaluate the prediction levels of different models in each month over a long period of time.

Fig. S2 shows that peak pollutant concentrations were distributed in different months. $PM_{2.5}$ had the largest concentration in winter and the largest forecast error (Figs. S2(a1), S2(a2)). PM_{10} was strongly affected by sand dust and the forecast error was largest in April and May when there were high levels of dust (Figs. S2(b1), S2(b2)). The O_3 concentrations peaked in summer (Figs. S2(f1), S2(f2)). The forecast errors of the two models began to increase and the trends of the two models diverged in the transition seasons of spring to summer (May), autumn to winter (September–October) or the peak season of pollutant concentrations. Except for O_3, which had a peak RMSE in summer, the maximum prediction errors for the pollutants occurred in winter and spring.

There were regular systematic errors among the models for the prediction of pollutant concentrations. We therefore removed these systematic deviations and reduced the uncertainties in the model products using multi-model ensemble technology.

3.2 Weight of Each Ensemble Model

BMA had a different weight allocation for each model in a particular study area as a result of inconsistencies in the initial field driving the meteorological background of the three models and the differences in their forecasting performance. We selected the observational and forecast data of pollutants from 18 stations in Henan Province from 10 November 2018 to 5 December 2019 as the research objects. Based on previous studies of the length of the sliding training period for pressure, temperature and precipitation, the optimum sliding training period for the BMA model to optimize the prediction effect was about 30–60 days (Qi *et al.*, 2019, 2020). We therefore used 30 days as the length of the sliding training period and the trained BMA parameters were applied to the BMA model forecast of the next day. The BMA model of each station in the study area was established dynamically every day—that is, the first sliding training period was from 10 November to 9 December 2018 and the forecast date was 10 December 2018. The second sliding training

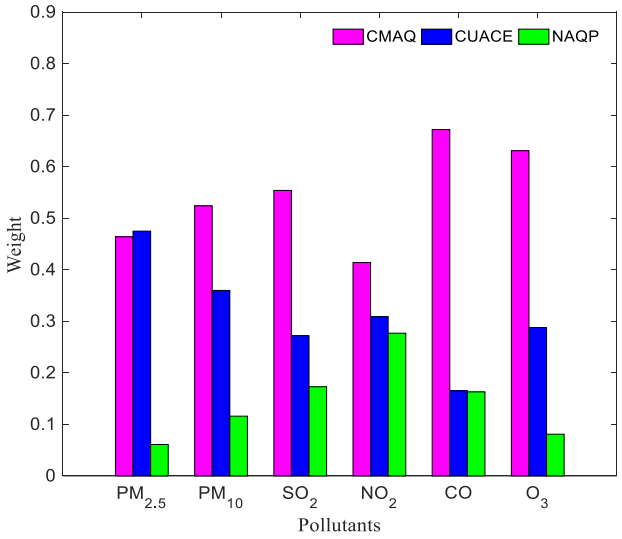

Fig. 2. Average weight of each ensemble model during the sliding training period from 10 November 2018 to 4 December 2019.

period was from 11 November to 10 December 2018 and the forecast date was 11 December 2018. A total of 361 days from 10 December 2018 to 5 December 2019 were chosen as the verification period.

Fig. 2 shows the average weight of the six pollutants for each ensemble model during the sliding training period. The CMAQ model had a larger weight for the six pollutant concentration forecasts and the NAQP model had the smallest weight. For $PM_{2.5}$ and PM_{10}, the weight of the CMAQ model was equivalent to that of the CUACE model (about 0.45). However, the CMAQ showed better forecasting ability for the gaseous pollutants SO_2, NO_2, CO and O_3. The CMAQ weight for CO reached 0.67, whereas the weights of the other two models were both about 0.16.

3.3 Evaluation of the Prediction of $PM_{2.5}$ and O_3 Concentrations Using BMA

Two typical pollutant concentrations of $PM_{2.5}$ and O_3 for 2019 were selected to analyze the forecast capability of BMA for pollutant concentrations. The CUACE model (Fig. 3(a)) predicted the peak concentration of $PM_{2.5}$ more accurately for the peak periods in winter. The forecast value of the CMAQ model was slightly lower than the observations and the NAQP model was less stable in forecasting the peak concentration of $PM_{2.5}$. In general, the three models all performed fairly well in predicting the peak concentration of $PM_{2.5}$ and the CUACE model gave the best prediction. After error correction using BMA, the overall $PM_{2.5}$ concentration forecast was even closer to the observations and more stable, but there was little room for improvement. The forecast results of the three models for O_3 concentrations in summer (Fig. 3(b)) all showed a degree of systematic underestimation, although the CUACE model was better than the other two models. After correction by BMA, the forecast of O_3 concentrations was improved and the overall forecast result was much closer to the observations, although it was still slightly low when the concentration was > 160 μg m^{-3}. This is because the post-processing technology of the multi-mode ensemble was dependent on the results of each single-model forecast, which eliminates the inter-model and single-model systematic errors.

The forecasting ability of the models varied with different levels pollution. We therefore need to evaluate the prediction effect of BMA on different pollution levels to provide a reference for future applications. Based on the Ministry of Environmental Protection's Environmental AQI Technical Regulations (HJ 633-2012), the mass concentration of $PM_{2.5}$ was divided into three levels: $PM_{2.5} < 75$ μg m^{-3} (cleaning conditions), 75 μg $m^{-3} < PM_{2.5} < 150$ μg m^{-3} (light to moderate pollution) and ≥ 150 μg m^{-3} (heavy pollution). The concentration of O_3 was classified into two levels: < 160 and ≥ 160 μg m^{-3}. For $PM_{2.5} < 75$ μg m^{-3}, the mass bias scores ranged from 0 to 20 μg m^{-3}, which was a small positive deviation (Fig. 4). The mass bias of BMA was 3 μg m^{-3}, which was the smallest error. For 75 μg $m^{-3} \leq PM_{2.5} < 150$ μg m^{-3}, the mass bias was the smallest

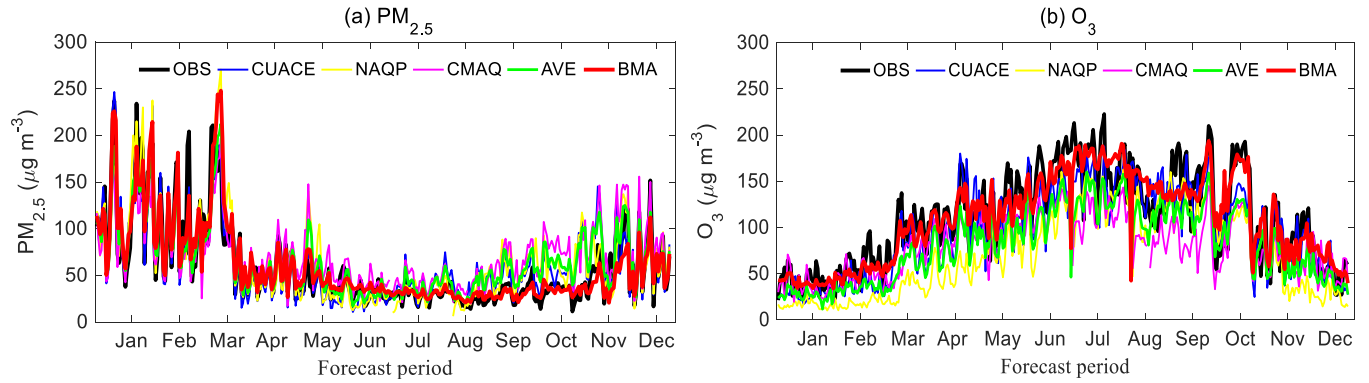

Fig. 3. Comparison of the 24-h forecasts of $PM_{2.5}$ and O_3 concentrations during the forecast period from 10 December 2018 to 5 December 2019 predicted by BMA, AVE and the three single models.

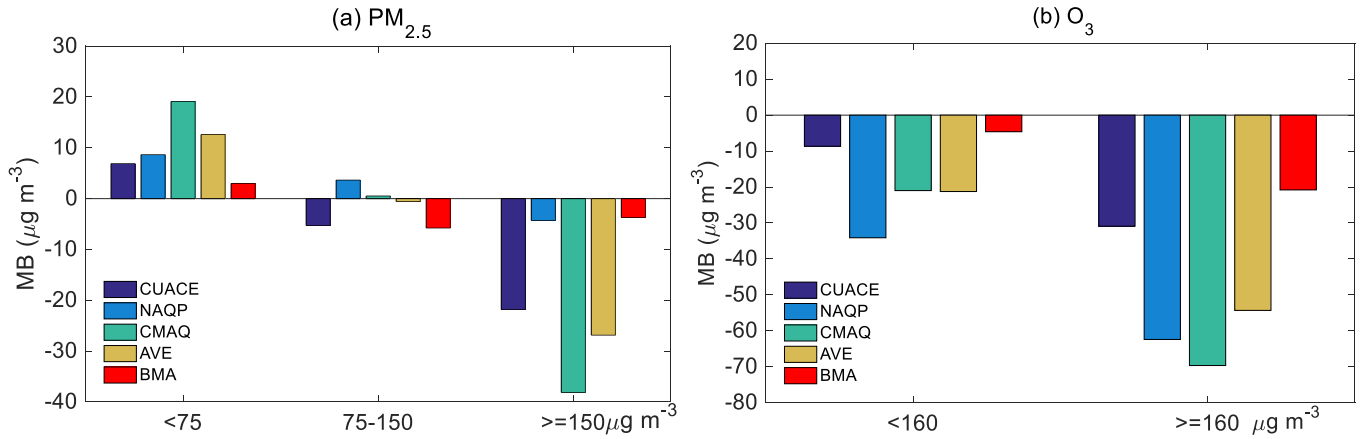

Fig. 4. Mean bias (MB) of (a) $PM_{2.5}$ and (b) O_3 concentrations for different levels of pollution at all stations during the forecast period with a 24-h lead time by BMA, AVE and the three single models.

of the three levels—that is, the models had the best prediction effect and the BMA prediction effect did not exceed the optimum single model. For heavy pollution ($PM_{2.5} \geq 150$ μg m^{-3}), the mass bias was negative for all three models and ranged from –38 to –3 μg m^{-3}. This means that the forecasts of all three models underestimated the $PM_{2.5}$ concentration. BMA gave a significant improvement compared with the best single model (NAQP), with the error reduced by 13.6%. The mass bias of the prediction of O_3 concentrations was < 0 for the two levels of pollutant concentrations, showing the systematic underestimation. The BMA prediction effect gave the best results, with the mass bias for the two concentration levels reduced by 46 and 32%, respectively, compared with the optimum CUACE method.

We analyzed the effects of different seasons on the typical concentrations of $PM_{2.5}$ and O_3. Fig. 5 compares the RMSE and NMB of the seasonal forecasts of $PM_{2.5}$ and O_3 with a 24-h lead time obtained by the three models and BMA during the forecast period. The RMSE for $PM_{2.5}$ of all models (20–50 μg m^{-3}) was larger in autumn and winter than in spring and summer (Fig. 5(a)). This shows that the large RMSE was caused by the large concentrations of pollutants in autumn and winter rather than a decrease in the forecasting ability of the models. Compared with the optimum single-model CUACE, the RMSE of the BMA forecast decreased by 37, 35, 68 and 50% in winter, spring, summer and autumn, respectively. The NMB of $PM_{2.5}$ (Fig. 5(b)) was mainly positive and the error was larger in winter and spring, so the forecast value was overestimated in these two seasons. The forecast was better with smaller errors in summer and autumn. The RMSE for O_3 (Fig. 5(c)) was similar in the four seasons. The maximum error in summer for the three models was about 40–63 μg m^{-3}. The RMSE of the BMA forecast was the lowest in all four seasons, whereas that of the NAQP forecast was the highest, except in summer. Unlike the $PM_{2.5}$

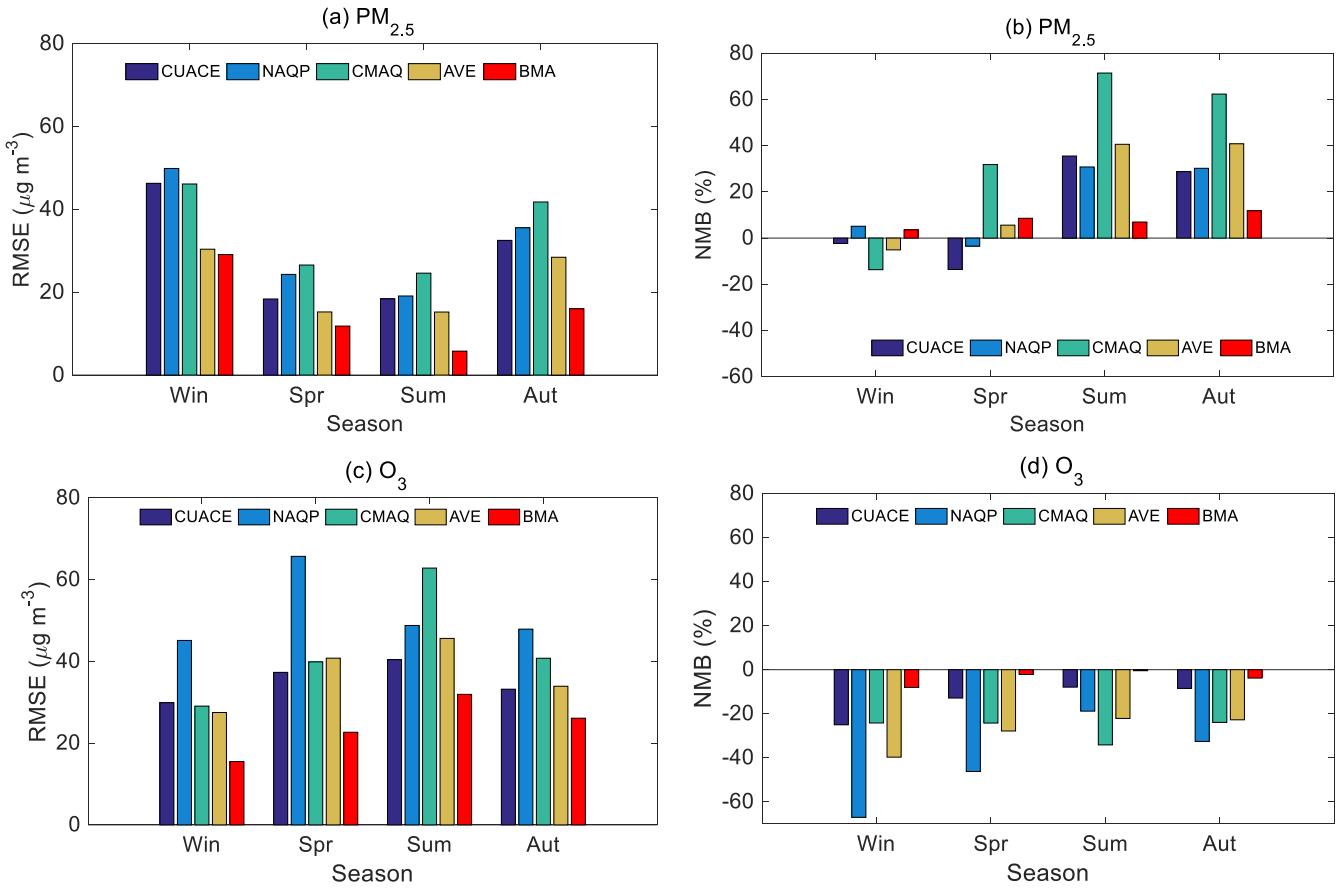

Fig. 5. Comparison of the RMSE and NMB of the $PM_{2.5}$ and O_3 forecasts with a 24-h lead time in four seasons by BMA, AVE and the three single models during the forecast period.

concentration, the NMB of all three models (Fig. 5(d)) was less than zero in all four seasons, which means that the models underestimated the O_3 concentrations. Compared with the CUACE model, the NMB of BMA decreased by 67, 83, 94 and 55% in winter, spring, summer and autumn, respectively.

The models therefore simulate moderate or light $PM_{2.5}$ pollution well, but underestimate heavy pollution. The models systematically underestimate O_3 concentrations, which provides a basis for future improvements. BMA can make the corrected forecasts of peak pollutant concentrations more consistent with the observations and therefore provides good technical support for operational forecasting.

3.4 Evaluation of the Prediction of the Concentrations of Six Pollutants by BMA

Fig. 6 compares the RMSEs of the forecasts for the concentrations of the six pollutants with lead times of 1–7 days for the BMA, the ensemble average of the three models (AVE), and the NAQP, CMAQ and CUACE models at all stations during the forecast periods. The RMSE index was selected to evaluate the forecast performance of the three models. A total of 361 days from 10 December 2018 to 5 December 2019 were chosen as the verification period. Fig. 6 shows that, with the extension of the forecast lead time, the single- and multi-model ensemble forecasting had very similar forecast errors for the six pollutants, indicating that the forecasting of pollutants was similar to the forecasting of temperature, with a certain degree of sustainability. The RMSE after BMA was smaller than those of the single models and AVE, so the BMA ensemble forecast method had a better effect.

Taking the 24-h forecast lead time as an example (Figs. 6(a) and 6(b)), the RMSE of the BMA forecast for $PM_{2.5}$ and PM_{10} was reduced by about 24% relative to that of the CUACE model, which

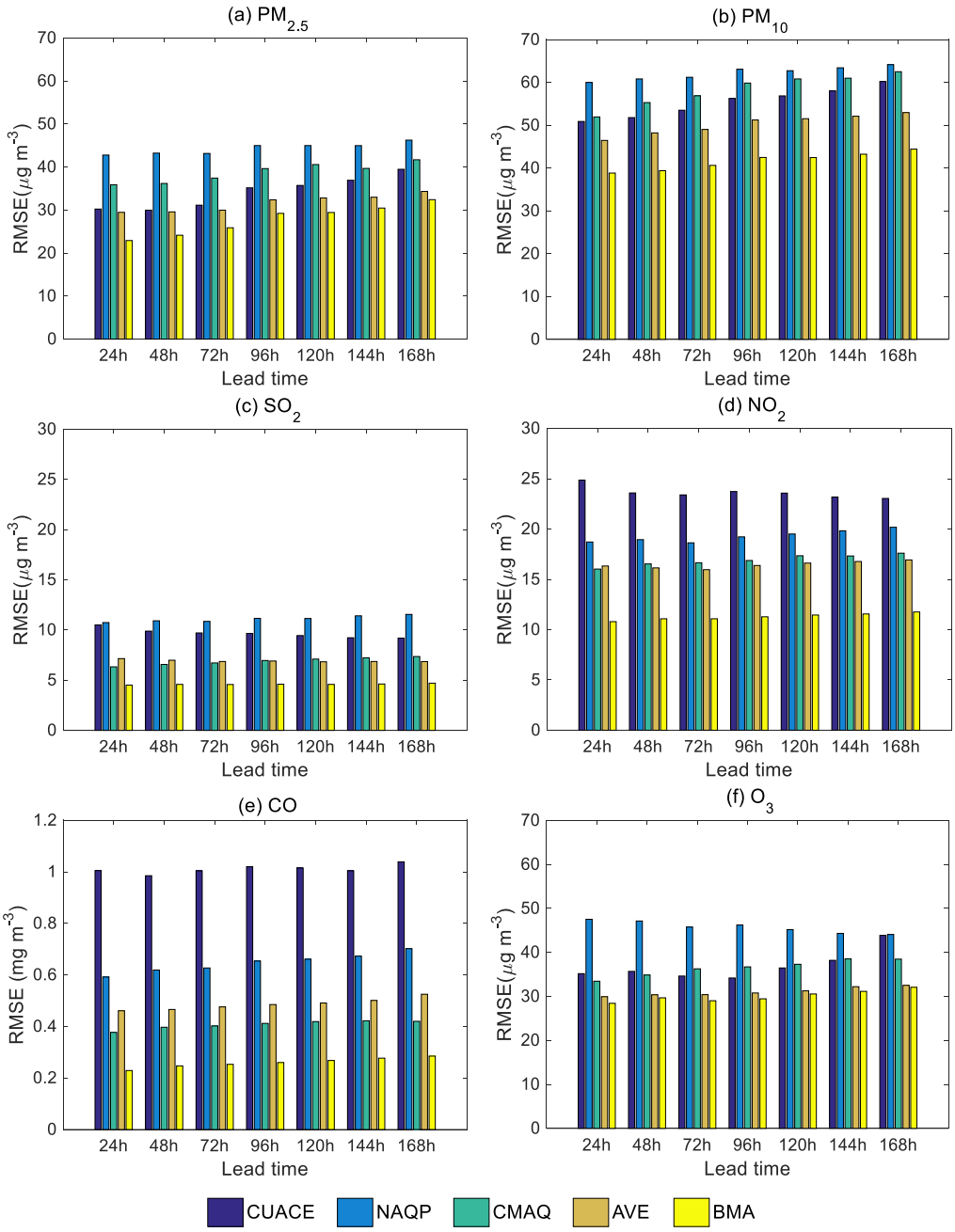

Fig. 6. RMSEs of the forecasts of the six pollutant concentrations with 1–7-day lead times by BMA, AVE, NAQP, CMAQ and CUACE at all stations during the forecast period.

had a better prediction effect. The RMSEs of the $PM_{2.5}$ and PM_{10} forecasts were decreased by about 22 and 16%, respectively, compared with the AVE. The CMAQ model had a relatively good prediction effect for the gaseous pollutants SO_2, NO_2 and CO (Figs. 6(c), 6(d) and 6(e)) and its error was smaller than that of the other two models. The prediction effect of the AVE was close to that of the CMAQ method and the forecast effect of the concentration of pollutants was not significantly improved. The RMSEs of AVE with a lead time of 1–7 days were all larger than the optimum single-model CMAQ result, particularly for CO forecasting, which shows that the three models had greater dispersions and higher uncertainties for the forecast of CO. However, BMA showed a more obvious improvement effect for SO_2, NO_2 and CO than CMAQ for the 1–7-day forecast, with the RMSE reduced by about 29–36, 33 and 32–39%, respectively. The best BMA forecast was still for O_3 (Fig. 6(f)), but the improvement was relatively small (about 1–5%) compared with the AVE.

The forecast effect for the six pollutant concentrations with a 1–7-day lead time after correction by the multi-model ensemble BMA was better than those of the AVE and the single-model forecast. The 24-h forecast showed that, compared with AVE, the prediction RMSEs of $PM_{2.5}$ and PM_{10} by BMA decreased by about 22 and 16%, respectively. Compared with the optimum single-model CMAQ method, the prediction RMSEs of SO_2, NO_2 and CO by BMA were reduced by about 29, 33 and 39%, respectively. BMA gave a smaller improvement (about 5%) in the RMSE of the forecast of O_3 than AVE. The forecast was therefore improved after BMA to calibrate forecast ensembles and BMA can provide technical support for operational forecasting.

3.5 Spatial Distribution of the Forecast Errors of BMA

Fig. 7 shows the RMSE distribution of the 24-h concentration forecasts of six pollutants by the single models and BMA in Henan Province during the verification period. For $PM_{2.5}$ (Figs. 7(a1–a4)), the prediction error of the CUACE model was high in the northwest and low in the southeast. The maximum error of the RMSE was in the range 34–37 μg m^{-3}, which was the smallest forecast error of all the models. The CMAQ model showed that the forecast error in the north-central region was larger than that in the south; the forecast error was largest (40 μg m^{-3}) in region ZZ. The prediction error of the NAQP model was high in east and low in west Henan, in contrast with that of the CUACE model, and the RMSE was the largest among the three models, with a maximum of 50 μg m^{-3}. The error distribution of the BMA forecast was similar to the optimum single-model CUACE forecast, but the improvement in the error was not obvious. The spatial distributions of the CUACE and CMAQ forecast errors were the same for PM_{10} (Figs. 7(b1–b4)), especially in northwest Henan, where the maximum error was as high as 55–65 μg m^{-3}. The error of the NAQP model was high in the north and low in the south of the province. The second largest error in the RMSE (60–77 μg m^{-3}) was near region ZZ. The BMA forecast error was improved compared with the single models, with the RMSE reduced to 45–53 μg m^{-3} in northwest Henan.

The large errors in the single models for SO_2 (Figs. 7(c1–c4)) were concentrated in north-central Henan. The RMSEs of the CUACE and NAQP models had a larger error range > 9 μg m^{-3} and the CMAQ forecast effect was slightly better. The BMA forecast was also high in the north and low in the south of the province, but the prediction error values and ranges were smaller with the best RMSE of 6–10 μg m^{-3}.

The error distributions for NO_2 and CO (Figs. 7(d1–d4) and 7(e1–e4), respectively) in the three models and BMA were similar to that of SO_2, with the largest prediction error for the CUACE model. The error distributions for O_3 (Figs. 7(f1–f4)) in the CUACE and CMAQ models were higher in north and lower in south Henan, with the largest simulation error in region ZZ. The NAQP model showed a larger error in northwest than southwest Henan and the error was the largest of the three models. BMA showed a greater improvements for the errors in NO_2, CO and O_3 concentrations.

The prediction errors of the three models for the concentrations of the six pollutants in Henan were all high in the north and low in the south. The model prediction effect was ranked as CUACE > CMAQ > NAQP for $PM_{2.5}$, PM_{10} and O_3, whereas the prediction effects for SO_2, NO_2 and CO were in the order CMAQ > NAQP > CUACE. The BMA prediction error was similar to that of the optimum single-model CUACE for $PM_{2.5}$, but greatly reduced the prediction errors for the other five pollutants compared with the three single models.

3.6 Analysis of a Heavy Pollution Process

Henan Province was affected by persistent heavy pollution events from 20 to 26 February 2019, especially at Puyang (PY), Anyang (AY) and Hebi (HB) stations, where the pollution lasted for 7 days, with the worst conditions on 20 February. The $PM_{2.5}$ concentration (Fig. 8(a)) at PY, AY and HB stations on 20 February reached 460, 373 and 274 μg m^{-3}, respectively, and the PM_{10} concentration (Fig. 8(b)) reached 460, 404 and 305 μg m^{-3}, respectively. The model forecasts underestimated the air quality in terms of particulate matter. The BMA forecasts for the stations with concentrations > 150 μg m^{-3} were much closer to the observations than the single models and AVE forecasts. Although the BMA forecast effect for severe $PM_{2.5}$ pollution was still lower, it was much improved. BMA and the NAQP model had similar forecasting effects for PM_{10} (Fig. 8(b)), with little improvement. The three models all overestimated SO_2 (Fig. 8(c)) and NO_2 (Fig. 8(d)) pollution, although the CMAQ model had a relatively good effect. By contrast, BMA had a good

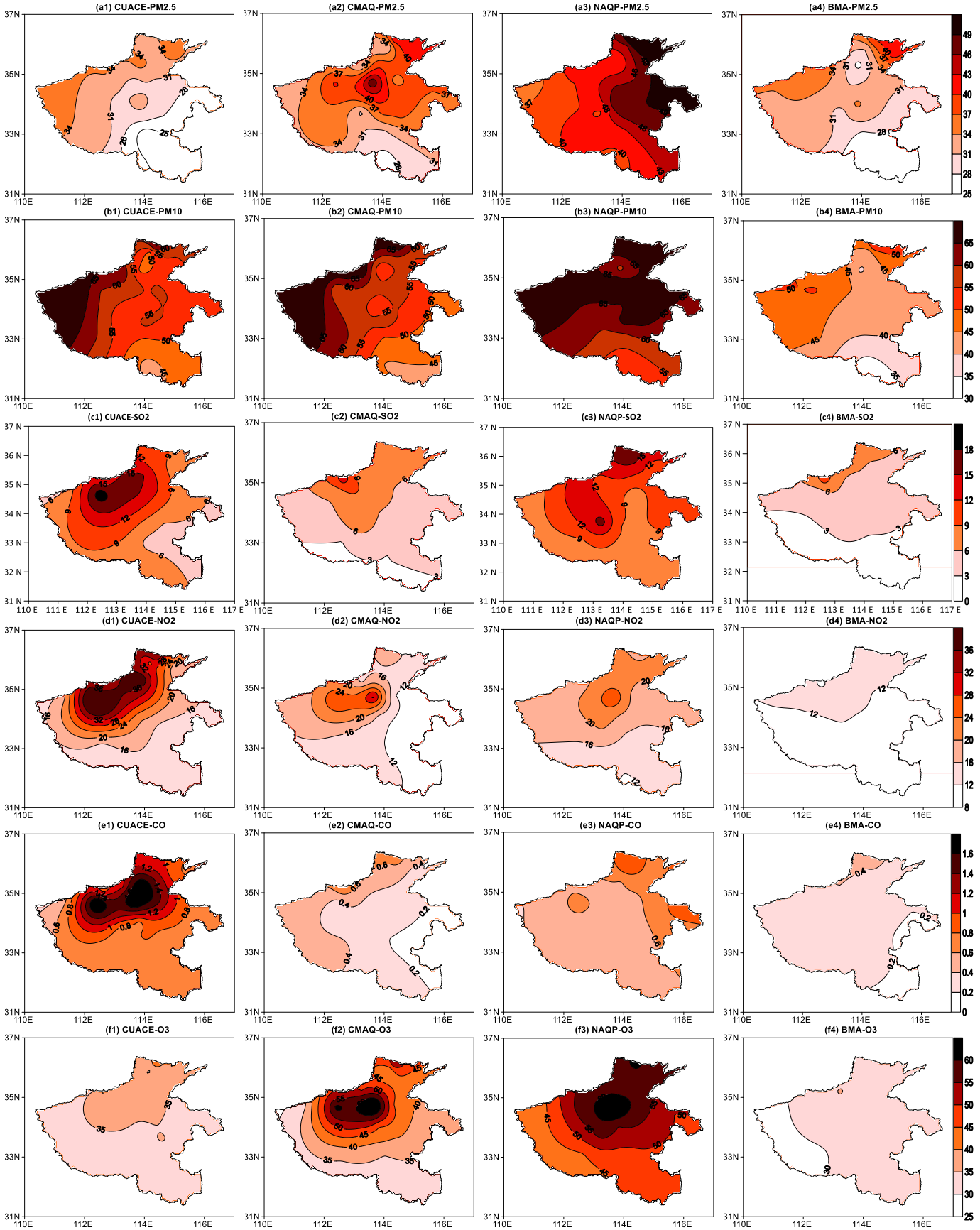

Fig. 7. RMSE distribution of the 24-h forecasts of the concentrations of six pollutants by the three single models and BMA in Henan during the forecast period.

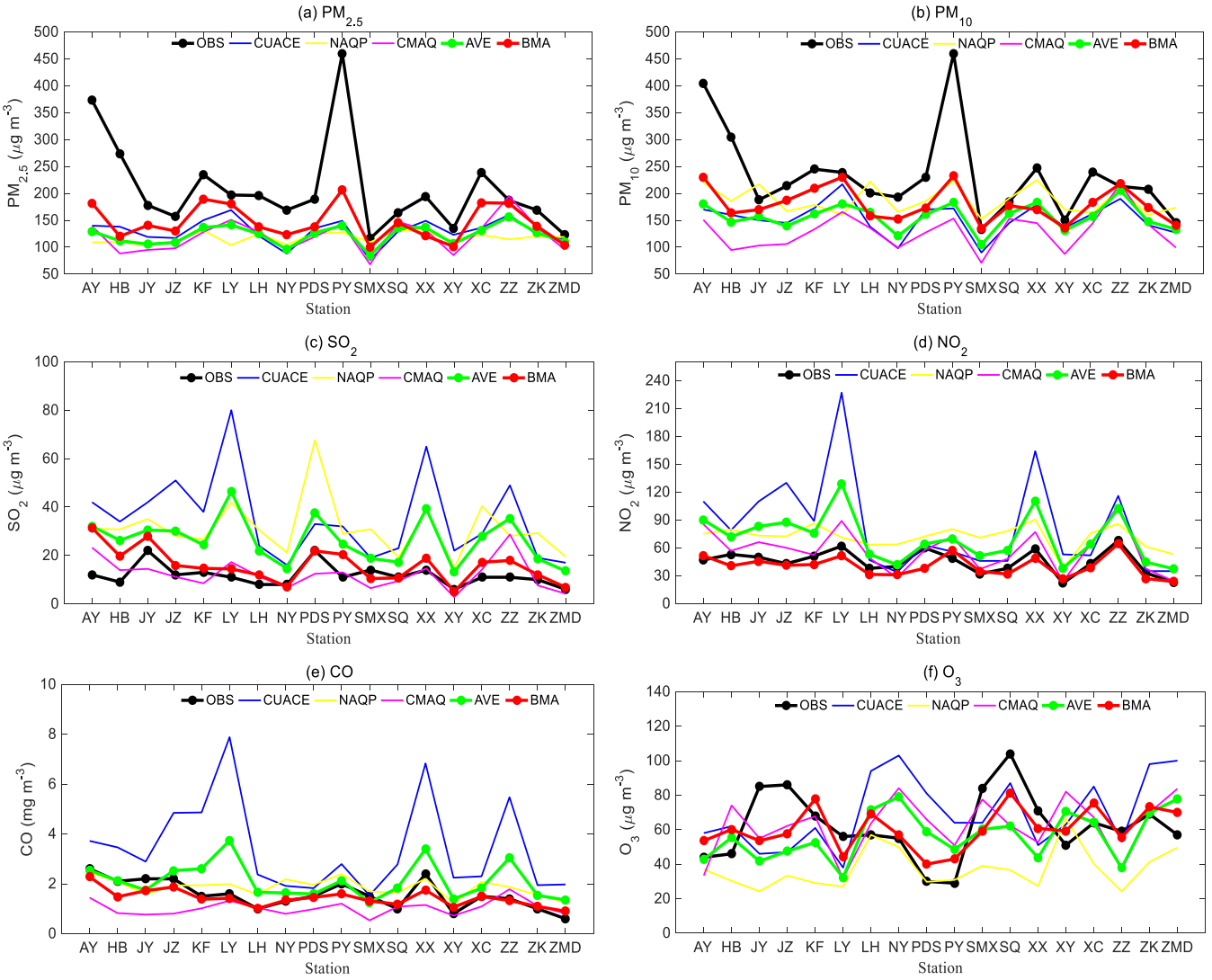

Fig. 8. Comparison of the 24-h forecasts of the concentrations of the six pollutants by BMA, AVE and the three single models on 20 February 2019.

correction effect on the positive bias of the model. The NAQP forecast was better for CO (Fig. 8(e)), but the BMA forecast was nearer the observations. The three models were unstable in predicting the trend for O_3 (Fig. 8(f)), but the BMA forecast was stable with a small bias relative to the observations.

4 CONCLUSIONS

The prediction errors in the concentrations of six pollutants in Henan region were high in the north and low in the south for all models in winter (December to the following February) when pollution frequently occurs (the peak period for O_3 is summer). The overall ranking of the models for the prediction effects for $PM_{2.5}$, PM_{10} and O_3 concentrations was CUACE > CMAQ > NAQP, whereas the order for the SO_2, NO_2 and CO concentrations was CMAQ > NAQP > CUACE. BMA reduced the RMSEs of the 24-h forecasts for $PM_{2.5}$ and PM_{10} by about 16% relative to the AVE forecasts. BMA reduced the RMSEs for SO_2, NO_2, and CO by about 29, 33 and 39%, respectively, compared with the optimum single-model CMAQ. The improvement with BMA (about 5%) was less than that of AVE for O_3. The NMB of BMA with a 24-h lead time decreased by 67, 83, 94 and 55% compared with the CUACE model for O_3. All the models presented a systematic underestimation for O_3 and heavy pollution by $PM_{2.5}$.

The systematic deviations of the three models in different seasons and their correlations with the observations were inconsistent. For example, the three models had higher correlation coefficients in winter and spring for $PM_{2.5}$, but were more likely to have low values. By contrast, the correlation coefficients were lower in summer and autumn and the values were easily overestimated. For PM_{10}, the three models generally underestimated the forecasts in all four seasons, especially in winter and spring. In terms of the O_3 forecasts, the CUACE model underestimated pollution in all four seasons, whereas the CMAQ model overestimated the concentrations in all the seasons except summer.

The BMA model was good for forecasting the peak periods of the two main pollutants ($PM_{2.5}$ and O_3). After error correction, the overall forecast result of the $PM_{2.5}$ concentration for BMA approached the real situation and was more stable, leaving little room for improvement. For O_3, all three models showed a systematic underestimation, which was corrected by BMA. However, the forecast was still a little less than the observations when the concentration was > 160 μg m^{-3}. This is because the post-processing technology of the multi-model ensemble depends on the single-model forecasts and therefore the inter-model and single-model systematic errors are removed.

Future research will use seasonal data in the BMA model to conduct training and forecasting. The BMA model adopted in this paper did not strictly test the distribution of the PDF for all six pollutant concentrations and there is still a lack of model-matching, which may be one of the reasons for the small improvement in the ensemble forecasting results.

ACKNOWLEDGMENTS

The authors thank TianLiang Zhao, Lin Liu, Ting Peng and Luying Ji and the anonymous reviewers for their suggestions and comments on this paper. This study was supported in part by the National Key R&D Program of China (2017YFC0212405). The authors are grateful for support from the General and Key Research Project of Hubei Meteorological Bureau (2022Y06 and 2020Z03).

DISCLAIMER

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

SUPPLEMENTARY MATERIAL

Supplementary material for this article can be found in the online version at https://doi.org/10.4209/aaqr.210247

REFERENCES

Chen, Y., Jiang, N., Zhu, L.L. (2021). Impact of anomalous high pressure over the North Pacific on $PM_{2.5}$ pollution in Beijing-Tianjin-Hebei Region. Torrential Rain Disasters 40, 608–616 (in Chinese)

Donnelly, A., Misstear, B., Broderick, B. (2015). Real time air quality forecasting using integrated parametric and non-parametric regression techniques. Atmos. Environ. 103, 53–65. https://doi.org/10.1016/j.atmosenv.2014.12.011

Emmons, L.K., Walters, S., Hess, P.G., Lamarque, J.F., Pfister, G.G., Fillmore, D., Granier, C., Guenther, A., Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S.L., Kloster, S. (2010). Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4). Geosci. Model Dev. 3, 43–67. https://doi.org/10.5194/gmd-3-43-2010

Galmarini, S., Bianconi, R., Appel, W., Solazzo, E., Mosca, S., Grossi, P., Moran, M., Schere, K., Rao, S.T. (2012). ENSEMBLE and AMET: Two systems and approaches to a harmonized, simplified

and efficient facility for air quality models development and evaluation. Atmos. Environ. 53, 51–59. https://doi.org/10.1016/j.atmosenv.2011.08.076

Galmarini, S., Kioutsioukis, I., Solazzo, E. (2013). *E pluribus unum**: Ensemble air quality predictions. Atmos. Chem. Phys. 13, 7153–7182. https://doi.org/10.5194/acp-13-7153-2013

Gao, H., Yang, W., Wang, J., Zheng, X. (2020). Analysis of the effectiveness of air pollution control policies based on historical evaluation and deep learning forecast: A case study of Chengdu-Chongqing Region in China. Sustainability 13, 206. https://doi.org/10.3390/su13010206

Hamill, T.M., Whitaker, J.S., Xue, W. (2004). Ensemble Reforecasting: Improving medium-range forecast skill using retrospective forecasts. Mon. Weather Rev. 132, 1434–1447. https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T. (1999). Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. Statist. Sci. 14, 382–417. https://doi.org/10.1214/ss/1009212519

Huang, J., Mcqueen, J., Wilczak, J., Djalalova, I., Monache, L.D. (2017). Improving NOAA NAQFC $PM_{2.5}$ predictions with a bias correction approach. Weather Forecasting 32, 407–421. https://doi.org/10.1175/WAF-D-16-0118.1

Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., Gu, D., Liu, Y. (2018). Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China Plain. Environ. Pollut. 242, 675–683. https://doi.org/10.1016/j.envpol.2018.07.016

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., Meehl, G. (2010). Challenges in combining projections from multiple climate models. J. Clim. 23, 2739–2758. https://doi.org/10.1175/2009jcli3361.1

Li, H.M., Wang, J.Z., Yang, H.F. (2020). A novel dynamic ensemble air quality index forecasting system. Atmos. Pollut. Res. 11, 1258–1270. https://doi.org/10.1016/j.apr.2020.04.010

Li, M., Yuan, K., Hu, K., Liu, M.B. (2019). Trigger mechanism and main factors of urban heavy pollution processes in Wuhan. Torrential Rain Disasters 38, 624–631. https://doi.org/10.3969/j.issn.1004-9045.2019.06.007 (in Chinese)

Mebust, M.R., Eder, B.K., Binkowski, F.S., Roselle, S.J. (2003). Models-3 community multiscale air quality (CMAQ) model aerosol component 2. Model evaluation. J. Geophys. Res. 108, 4184–4202. https://doi.org/10.1029/2001JD001410

Monache, L.D., Alessandrini, S., Djalalova, I., Wilczak, J., Knievel, J.C., Kumar, R. (2020). Improving Air Quality Predictions over the United States with an Analog Ensemble. Weather Forecasting 35, 2145–2162. https://doi.org/10.1175/WAF-D-19-0148.1

Monteiro, A., Ribeiro, I., Tchepel, O., Carvalho, A., Martins, H., Sá, E., Ferreira, J., Martins, V., Galmarini, S., Miranda, A.I., Borrego, C. (2013). Ensemble techniques to improve air quality assessment: Focus on O_3 and PM. Environ. Model Assess. 18, 249–257. https://doi.org/10.1007/s10666-012-9344-0

Nopmongcol, U., Koo, B., Tai, E., Jung, J., Piyachaturawat, P., Emery, C., Yarwood, G., Pirovano, G., Mitsakou, C., Kallos, G. (2012). Modeling Europe with CAMx for the Air Quality Model Evaluation International Initiative (AQMEII). Atmos. Environ. 53, 177–185. https://doi.org/10.1016/j.atmosenv.2011.11.023

Nopmongcol, U., Liu, Z., Stoeckenius, T., Yarwood, G. (2017). Modeling intercontinental transport of ozone in North America with CAMx for the Air Quality Model Evaluation International Initiative (AQMEII) phase 3. Atmos. Chem. Phys. 17, 9931–9943. https://doi.org/10.5194/acp-17-9931-2017

Qi, H.X., Peng, T., Lin, C.Z., Peng, T., Ji, L.Y., Li, L., Meng, C. (2020). Probabilistic forecasting of the precipitation over the Qingjiang River basin using BMA multimodel ensemble technique. Meteorol. Mon. 46, 108–118. https://doi.org/10.7519/j.issn.1000-0526.2020.01.011 (in Chinese)

Qi, H.X., Zhi, X.F., Peng, T., Bai, Y.Q., Lin, C.Z. (2019). Comparative study on probabilistic forecasts of heavy rainfall in mountainous areas of the Wujiang River basin in China based on TIGGE Data. Atmosphere 10, 608. https://doi.org/10.3390/atmos10100608

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M. (2005). Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather. Rev. 133, 1155–1174. https://doi.org/10.1175/MWR2906.1

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., Petelin, D., Grancharova, A., Ribeiro (2013). Evolving Gaussian process models for prediction of ozone concentration in the air.

Simul. Modell. Pract. Theory 33, 68–80. https://doi.org/10.1016/j.simpat.2012.04.005

Rahman, N.H.A., Lee, M.H., Suhartono, Latif, M.T. (2015). Artificial neural networks and fuzzy time series forecasting: An application to air quality. Qual. Quant. 49, 2633–2647. https://doi.org/10.1007/s11135-014-0132-6

Rao, S.T., Galmarini, S., Puckett, K. (2011). Air quality model evaluation International Initiative (AQMEII): Advancing the state of the science in regional photochemical modeling and its applications. Bull. Am. Meteorol. Soc. 92, 23–30. https://doi.org/10.1175/2010BAMS3069.1

Shishegaran, A., Saeedi, M., Kumar, A., Ghiasinejad, H. (2020). Prediction of air quality in Tehran by developing the nonlinear ensemble model. J. Cleaner Prod. 259, 1–16. https://doi.org/10.1016/j.jclepro.2020.120825

Singh, K.P., Gupta, S., Rai, P. (2013). Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos. Environ. 80, 426–437. https://doi.org/10.1016/j.atmosenv.2013.08.023

Sloughter, J.M., Gneiting, T., Raftery, A.E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. J. Am. Stat. Assoc. 105, 25–35. https://doi.org/10.1198/jasa.2009.ap08615

Sloughter, J.M., Raftery, A.E., Gneiting, T., Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Mon. Weather Rev. 135, 3209–3220. https://doi.org/10.1175/MWR3441.1

Sun, W., Sun, J. (2016). Daily $PM_{2.5}$ concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. J. Environ. Manage. 188, 144–152. https://doi.org/10.1016/j.jenvman.2016.12.011

Tai, A.P.K., Mickley, L.J., Jacob, D.J. (2010). Correlations between fine particulate matter ($PM_{2.5}$) and meteorological variables in the United States: Implications for the sensitivity of $PM_{2.5}$ to climate change. Atmos. Environ. 44, 3976–3984. https://doi.org/10.1016/j.atmosenv.2010.06.060

Tandon, A., Yadav, S., Attri, A.K. (2013). Non-u4linear analysis of short term variations in ambient visibility. Atmos. Pollut. Res. 4, 199–207. https://doi.org/10.5094/APR.2013.020

Wang, Z.F., Xie, F.Y., Wang, X.Q., An, J.L., Zhu, J. (2006). Development and application of nested air quality prediction modeling system. J. Atmos. Sci. 31, 778–790. (in Chinese)

Zhai, S., Jacob, D.J., Wang, X., Shen, L., Li, K., Zhang, Y., Gui, K., Zhao, T., Liao, H. (2019). Fine particulate matter ($PM_{2.5}$) trends in China, 2013–2018: Separating contributions from anthropogenic emissions and meteorology. Atmos. Chem. Phys. 19, 11031–11041. https://doi.org/10.5194/acp-19-11031-2019

Zhang, X.Y., Wang, J.Z., Wang, Y.Q., Liu, H.L., Sun, J.Y., Zhang Y.M. (2015). Changes in chemical components of aerosol particles in different haze regions in China from 2006 to 2013 and contribution of meteorological factors. Atmos. Chem. Phys. 15, 12935–12952. https://doi.org/10.5194/acp-15-12935-2015

Zhong, J., Zhang, X., Wang, Y., Liu, C., Dong, Y. (2018). Heavy aerosol pollution episodes in winter Beijing enhanced by radiative cooling effects of aerosols. Atmos. Res. 209, 59–64. https://doi.org/10.1016/j.atmosres.2018.03.011

Zhu, L.L., Yan, P.Z., Wang, Z.F., Li, J., Zhang, X.Z., Tang, L.L., Li, J.J, Liu, B. (2015). An operational evaluation of the regional air quality forecast modeling system in Jiangsu. Environ. Monit. China. 31, 17–23. https://doi.org/10.19316/j.issn.1002-6002.2015.02.004 (in Chinese)

Zhu, W.H., Xu, X.D., Zheng, J., Yan, P., Wang, Y.J., Cai, W.Y. (2018). The characteristics of abnormal wintertime pollution events in the Jing-Jin-Ji region and its relationships with meteorological factors. Sci. Total Environ. 626, 887–898. https://doi.org/10.1016/j.scitotenv. 2018.01.083