

# Downscaling Winter Daily PM<sub>2.5</sub> Concentrations From Large-Scale Meteorological Dataset -a Case Study for a City in North China

Yonghe Liu (✉ [yonghe\\_hpu@163.com](mailto:yonghe_hpu@163.com))

Henan Polytechnic University <https://orcid.org/0000-0001-7052-5126>

Yang Li

Henan Polytechnic University

Mingshi Wang

Henan Polytechnic University

Hailing Wang

Henan Polytechnic University

Xiyue Wang

Henan Polytechnic University

---

## Research Article

**Keywords:** PM<sub>2.5</sub>, air pollution, North China, statistical downscaling, generalized linear models, optimal grid-box method

**Posted Date:** November 12th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-973715/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

---

# Abstract

Winter air pollution in North China becomes a serious environmental problem in recent years, which has aroused a widespread concern. Estimating PM<sub>2.5</sub> concentration is necessary for the government to take actions in leading times, or to reproduce the historical values. In this study, we attempt to construct statistical downscaling (SD) models based on large-scale meteorological variables, to estimate the PM<sub>2.5</sub> in Jiaozuo, a city in the heavy-pollution area of North China. Predictors were screened from large-scale meteorological variables, by selecting the grid boxes with the values highly correlated to the PM<sub>2.5</sub> in Jiaozuo. Correlation maps show that PM<sub>2.5</sub> is usually related to comparatively low pressure and high relative humidity at the lower atmosphere and comparatively high pressure at the upper atmosphere, high air temperature in all the troposphere and weak winter monsoonal winds. After training the SD models with the winter samples during 2014-2017, the daily PM<sub>2.5</sub> is simulated with correlation coefficients of 0.607 (mean) and 0.548 (maximum) to the observation, by the logarithmic transformed PM<sub>2.5</sub> values. The SD models can roughly reflect the daily variation of PM<sub>2.5</sub>. Compared to the PM<sub>2.5</sub> model based on the local meteorological data, larger correlations are obtained (0.57 versus 0.51-0.53, without logarithmic transformation). In the days of "APEC Blue Sky" and the "SCO Blue Sky" with emissions largely reduced by strict controls on industry and traffic in surrounding areas, low PM<sub>2.5</sub> is indeed observed, however, the SD models without considering any change of emissions also produced low PM<sub>2.5</sub> simulations, implying that the large-scale circulation plays a main role in the daily variation of air pollution. Similar effects were obtained for the traffic-control month and the Chinese Spring festival with massive fireworks burning. Basing on the large-scale reanalysis variables and assuming the same emission level as that in the model training period, the downscaled winter PM<sub>2.5</sub> has a significant decreasing trend during 1979-2017.

## 1. Introduction

The impact of the changing climate on environment is a public concern around the world today. A great amount of future climate predictions for the globe were produced by general climate models (GCM) in many organizations. However, these predictions are in rough resolutions which cannot meet the needs of local impact studies in different areas. Downscaling is useful to resolve such scale discrepancy. There are two categories of downscaling techniques can be used: dynamical downscaling and statistical downscaling(SD). Dynamical downscaling refers to the using of regional climate models to simulate the climate in the interested areas in higher resolutions, driven by the datasets of large-scale climate prediction which is simulated by GCMs. Dynamical downscaling can generate reasonable downscaled result but it is computationally expensive for downscaling long-term climate scenarios. SD refers to the technique that derives local variables from large-scale variables (LSV) in climate scenarios by constructing empirical relation between the large-scale variables and the local variables (Maraun et al., 2010). Compared to dynamical downscaling, SD is computationally very cheap. The statistical downscaling on air temperature, precipitation and wind power is widely used now(Araya-Osses et al., 2020; Maraun et al., 2010; Lucio-Eceiza et al., 2019). Nevertheless, there is no widely applicable SD models which can be directly used for anywhere. In other words, SD models need to be tailored for different sites/areas respectively. The techniques for constructing SD models, including screening predictors, reducing the dimensionality of predictors and selecting mathematical structure of models, were presented in many previous studies(Horton and Bronnimann, 2019; Laura Bettolli and Clorinda Penalba, 2018; Liu et al., 2019).

There have been many SD models for downscaling surface air temperature and precipitation now, but very few studies have been presented for one public concern in China: the concentration of PM<sub>2.5</sub> (suspended particulate matter (PM<sub>2.5</sub>) with aerodynamic diameter less than 2.5 μm). Epidemiological studies have reported that exposure to PM<sub>2.5</sub> has adverse effects on human health (Kunzli et al., 2005; Zheng et al., 2016). Unlike surface air temperature and precipitation which are basically the meteorological variables themselves, the PM<sub>2.5</sub> are the outcome of multiple factors: meteorology(Chen et al., 2018), local emission(Zhang et al., 2019) and air transport(Zhao et al., 2020). The emission and transport of PM<sub>2.5</sub> in local areas are very complex in contrast to the meteorological factor. But according to many previous studies, meteorology is the main factor which influent the daily variations of concentration of PM<sub>2.5</sub> (Chen et al., 2018; Chen et al., 2020; Wang et al., 2020; Zhang et al., 2020). That means, under a given emission level of pollutants, the relative concentration level of PM<sub>2.5</sub> can be estimated/predicted from the meteorological factors. Actually, at a certain area, the human activities generally keep stable during different years. The emission levels can be reduced year by year through the gradual innovation of technologies, but after the reduction reaches a certain limit, a minimum emission will be maintained. Thus, at a presumed constant emission level of pollutants, the climatic variation or trend of PM<sub>2.5</sub> concentration can be estimated from pure meteorological factors.

Estimation of PM<sub>2.5</sub> from pure meteorological factors seems practically meaningless at the daily scale, since there have been many accurate and operational methods for estimating daily PM<sub>2.5</sub> concentration, such as numerical models or machine learning models(Analitis et al., 2020; Chen et al., 2019; Zheng et al., 2016). Most of these advanced high-accuracy models are highly relied on many localized datasets (such as the data of local meteorology, remote sensing and pollutant emissions) (Zhang et al., 2019) and can be used to reconstruct the daily/hourly PM<sub>2.5</sub> concentration in the period with all these available datasets, however, such models is irrelevant

to the climate science, and cannot be used to reflect/predict the relative PM<sub>2.5</sub> levels in the periods with some of the datasets unavailable. The downscaling of PM<sub>2.5</sub> concentration from large-scale meteorological variables is useful in certain aspects. For example, people may want to know the relative concentration level of PM<sub>2.5</sub> which is caused by the meteorological conditions in the last five decades in which most of the time there was no observational PM<sub>2.5</sub> data. It is worth noting that there was no PM<sub>2.5</sub> records in China before 2013, and the remote sensed AOD data is also only available for the last two decades. People may also want to know how the future PM<sub>2.5</sub> concentration will change under the background of climate change. In such a sense, downscaling the relative level of PM<sub>2.5</sub> concentration has the meaning in climate science just as that for the downscaling of precipitation, wind and air temperature. In weather prediction, the PM<sub>2.5</sub> concentration can be generally well predicted by numerical models, but predicting PM<sub>2.5</sub> through statistical downscaling based on large-scale circulation is also valuable, since this method is very simpler and computationally cheaper than numerical models. Till now, few studies have addressed such downscaling problem.

In this study, we made an attempt to construct SD models for downscaling daily PM<sub>2.5</sub> concentration in winter at a site in North China. The aim of this study is to investigate following related questions: whether the large-scale circulation representing variables can be used to get satisfying predictors; what large-scale circulations are related to the site PM<sub>2.5</sub>; what predictors can be identified from different large-scale variables; how the performance can be obtained by such SD models when compared with the models constructed from local ground based meteorological variables; what performance can be obtained in reproducing the daily PM<sub>2.5</sub> concentration in the winters of recent years. As a research on downscaling based on purely meteorological large-scale variables, we do not expect the downscaled outputs be very accurate performance as that achieved by many studies on modelling PM<sub>2.5</sub>.

## 2. Study Site, Data And Methods

### 2.1 Study site

As a case study, we conduct this statistical downscaling of the PM<sub>2.5</sub> in Jiaozuo, a city in North China Plain (NCP). The area in NCP, including the Beijing-Tianjin-Hebei region (BTH) and the North of Henan province, is one of the most populous regions in China, and has been the region having the most severe air pollution before 2014(Zhang et al., 2012). To cope with the severe air pollution problem, the State Council of China established the toughest-ever air pollution prevention and Control Action Plan in 2013. By taking such effective measures, the air quality is improved year by year(Xiao et al., 2020). However, in order to maintain the sustainable economic development to guarantee the necessary employment, most of the industries have to be preserved. Thus, a least emission level will have to be last for a long time in the future before important technological innovations being available, and the air pollution problem cannot be completely resolved. Heavy air pollution is still frequent in winter times now, although the most serious pollution situation has been improved compared to previous years.

As a case study, we use the PM<sub>2.5</sub> concentration in the urban area of Jiaozuo (Fig. 1). Jiaozuo is a medium-sized city (according the standard for the classification of cities in China) in the south part of NCP and belongs to the northwest of Henan Province. Jiaozuo is a neighbor city of Zhengzhou, the metropolis of Henan Province. The administration region of Jiaozuo has an area of 4071 km<sup>2</sup> and a population of 3.6 million. The urban area in the region includes the center urban area of Jiaozuo and other six county-level cities. The whole Jiaozuo administration area and its surrounding areas in the north of Henan province are highly industrialized. The large industry production in this area gives rise to large pollutant emissions, because heavy chemical industry accounts for a large proportion of gross domestic production (GDP) in the north of Henan Province. Due to the shielding effect of Taihang Mountain on atmospheric circulation and the high industrial activities in North China, a long belt of heavy air pollution stretches along the East and south side of Taihang Mountain in the NCP (Fig. 1). The area of Jiaozuo is just in the south of this belt, and Jiaozuo is also famous for being one of the most polluted cities in China in the last two decades. Due to the fact that the central urban area of Jiaozuo just lies at the foot of Taihang mountain, the pollution diffusion condition of the area is worse than the surrounding areas.

### 2.2 Data

In this study, we obtained the hourly PM<sub>2.5</sub> concentration data recorded by three air quality monitoring stations (Huanbaoju, Gaoxinqu and Yingshicheng) in the city of Jiaozuo, which is managed by the Environmental Protection Bureau of Jiaozuo. Considering very few missing values exists in the PM<sub>2.5</sub> observation at station Huangbaoju, we only used the data at this station to construct SD models. The data for stations Gaoxinqu and Yingshicheng were used as reference to compare or validate. The dataset for Huanbaoju covers the period 2014-2018, only the values in the first several months of 2014 are missing. The mean and the maximum of each day were extracted from this hourly dataset. Both the daily mean PM<sub>2.5</sub> and maximum PM<sub>2.5</sub> were used as predictands for the statistical downscaling.

The ERA-Interim Reanalysis dataset (ERA-I) from the European Centre for Medium-Range Weather Forecasts (ECMWF) (Dee et al., 2011) were downloaded. The variables in ERA-I with multiple pressure levels (500hpa, 700hpa, 850hpa and 1000hpa) were used to screen predictors: relative humidity (R), specific humidity(Q), latitudinal direction wind (V), longitudinal direction wind (U), air temperature (T) and geopotential height (Z). Here, the ERA-Interim data forecasted at 0h UTC but for 12h UTC in every day were used to represent the meteorological status of that day. The selected dataset covers the period 1979-2018. This dataset has the domain of 60°E-150°E and 10°N-65°N, which covers the areas surrounding China. To facilitate the data processing, the dataset was degraded into a 1°×1° resolution grid when downloaded. Only the datasets from the winter months (November, December, January and February) were used. In this study, in order to preserve the temporal continuity of the samples in each winter, we take the samples in November and December of each year and those in January and February of the next year into one winter sample set. Therefore, the winter sample set of 2014 should also include the samples in January and February of 2015.

Very few PM<sub>2.5</sub> values are missing (only Jan.1 2015 and Dec. 22-26, 2018), so these days were discarded from the samples, both for the observed PM<sub>2.5</sub> values and the final screened predictors.

Considering there is no published ground-based meteorological data available in the administrative area of Jiaozuo, we also collected surface meteorological data (including air temperature, wind speed, air pressure, relative humidity, pan evaporation and sun shine duration) observed by Zhengzhou meteorological station, which is one of the nearest site (within 100km) to Jiaozuo, and both Jiaozuo and Zhengzhou shares generally the same weather. Accordingly, the PM<sub>2.5</sub> concentration data observed by the Center Air Quality Monitoring Station of Zhengzhou during 2014-2018 was also collected, which is used to construct a simple model for estimating PM<sub>2.5</sub> concentration based on ground based meteorological data, in order to compare the performance with other different models. Here, we do not plan to construct the downscaling model based on the large-scale circulation for Zhengzhou, considering the models for these cities should have theoretically little differences.

## **2.3 Downscaling PM<sub>2.5</sub> based on large-scale circulations**

### **2.3.1 Physical relation between the site PM<sub>2.5</sub> and the large-scale circulations**

Correlation maps were used to reflect the physical relation between the PM<sub>2.5</sub> at Jiaozuo and the large-scale circulations. For the field of a LSV at a certain pressure level, a Pearson's correlation coefficient (CC) can be calculated between the daily LSV time series at each grid box and the PM<sub>2.5</sub> series (logarithmic transformed) in Jiaozuo. The CCs for all the grid boxes can comprise a correlation map for the LSV.

The correlation maps can actually reflect the relation of the LSV's anomalies to the site PM<sub>2.5</sub>, and cannot reflect the true circulation over the area. In order to compare different circulations at high PM<sub>2.5</sub> days and low PM<sub>2.5</sub> days, we divide the winter daily samples into a high-concentration group and a low-concentration group by the threshold of PM<sub>2.5</sub>=38.5mg/m<sup>3</sup>, the medium PM<sub>2.5</sub> value of the samples. The geopotential height, U-wind and V-wind in each group were averaged over all the samples in the group, respectively.

### **2.3.2 Statistical downscaling Method**

The statistical downscaling comprises three steps: screening predictors, training statistical models and assessing the model performance.

Predictor screening is vital to SD models, and the most commonly used LSVs for downscaling precipitation are related to air pressure, moisture and air circulation. For PM<sub>2.5</sub>, theoretically, the above variables are also applicable and were also used in this study. The pressure levels and the grid boxes of the LSVs were considered. In this study, we use a correlation map based predictor screening method. In this method, the correlation maps should be created by relating each candidate variables with the predictand, then the grid boxes with the highest CCs are identified as the final predictors(Liu et al., 2019; Fu et al., 2018; Zhang and Yan, 2015). Here, we call this method the Optimal Grid Box (OGB) method. We identify the high correlation centers manually in the correlation maps. The final screened predictors by OGB are the raw values of the large-scale variables, thus the physical relation to the predictand is easier to explain when compared to the Principle Components Analysis (PCA)(Hertig et al., 2019; Pryor and Schoof, 2019) based screening.

Generalized linear models (GLM) (Pryor and Schoof, 2019) were used as the mathematical form of the downscaling models in this study. GLMs performs comparatively good(Gutierrez et al., 2019) among different models.

The LASSO based GLM fitting was used to estimate the GLM parameters. The predictors screened from the multiple variables are also usually interior correlated, thus using all the predictors can inevitably cause overfitting problem. However, there is no strict rules to use or discard each predictor by considering the correlation coefficient values among different predictors. In a standard overfitting problem, the

regression coefficient/parameter of a predictor has a wrong sign which is different to the sign of the correlation coefficient of this predictor to the predictand. For example, one predictor screened from the air pressure has positive correlation to the predictand, however, during the model fitting process, this predictor gets a negative regression coefficient. This effect is caused by the high correlation of this air pressure based predictor to another predictor extracted from the air temperature (or air humidity): the signal of the air pressure based predictor is substituted by the signal of the air temperature based predictor, while a meaningless regression coefficient with wrong sign is obtained; the regression coefficients for both these two predictors can give rise to a wrong output when new samples other than the training samples are fed to the model. To avoid this overfitting problem, the predictor based on air pressure should be discarded, and the functionality of the discarded air pressure predictor is played by the air temperature predictor. The Least Absolute Shrinkage and Selection Operator (LASSO) algorithm is such a method (Hammami et al., 2012). By using a threshold, LASSO can automatically determine which predictor to be discarded in the model fitting process. If overfitting problem still exists after a LASSO based model fitting, then the threshold in LASSO should be increased to a larger value and refit the model.

### 2.3.3 Model validation

One-winter-leave-out cross validation was used to validate the models. There are only four years (2014-2017) of observed PM2.5 available, therefore, in each training-validation step, three years were used for training, and one year were used for validation. Rooted mean squared error (RMSE) and the Spearman's correlation coefficients were used to evaluate the model performance.

During the winter days of 2014-2017, there are some specific periods in which significant emission reduction or increase was caused by some social activities. These specific periods can be regarded as ideal scientific experiments to study the mechanism of air pollution. The detailed information for these specific periods is introduced here:

1) "APEC Blue Sky" (Nov.7-12, 2014). The 22nd APEC Economic Leaders meeting of Asia-Pacific Economic Cooperation (APEC) in 2014 was held in Beijing. Very strong emission restrictions over North China (Beijing, Tianjin, Hebei, Shanxi, Inner Mongolia and Shandong) were executed during this period to ensure the high quality of air in Beijing. During this period, the air quality in Beijing and almost the entire area of North China was comparatively good. Perhaps influenced by this emission reduction, good air quality was also observed in Jiaozuo as well as in other surrounding areas. Nevertheless, most of the related studies on "APEC Blue Sky" were focused on the air quality in Beijing(Lin et al., 2017; Wang and Dai, 2016).

2) "SCO Blue Sky" (Dec. 14-15, 2015). An important leaders meeting of the Shanghai Cooperation Organization (SCO) was held in Zhengzhou, Henan Province, during December 14-15, 2015. To ensure the good air quality, strong restrictions were executed in Henan province. As a city neighboring to Zhengzhou, Jiaozuo also got several days of good air quality.

3) Traffic control with even and odd-numbered license plates in multiple cities of Henan province. To avoid the happening of bad air quality in the last month of 2017, in each day, half of the cars were restricted to travel in the main urban areas of multiple cities (Jiaozuo, Zhengzhou, Kaifeng and Anyang) in December 4-31, 2017.

4) The Chinese Spring Festival. The Spring Festival (the first day of each year in the Chinese Lunar Calendar) is well known as the very important traditional festival in China, and burning fireworks are always a necessary activity in the New Year's Eve (the night before the Spring Festival). In recent years, more fireworks are burning than in the past because the residents become richer than ever. This activity is bound to create severe pollution and every one can smell the strong sulfide in air during this festival. Nevertheless, since the Spring Festival in 2016, burning fireworks were banned in the main urban area of Jiaozuo to reduce air pollution, but no restriction was required in county-level urban areas and the wide rural areas. Similar restrictions were also executed in other cities in the surrounding areas of Jiaozuo.

### 2.3.4 Schemes for comparison

In order to analyze whether the GLM outputs are sensitive to different choices on using predictors, two different schemes for comparison are considered. The first scheme is to analyze whether the predictors extracted from wind field has large influence on the output. The design of this scheme is due to the fact that multiple predictors are usually obtained from the wind field, including U-wind and V-wind, which makes the process for screening predictors complex. The second comparison scheme is to use the single-grid box value at the high correlation center or the average value in multiple grid boxes of the high correlation area to get predictor values. Here, in order to obtain multiple grid box values, we determine each window through manual inspection on each correlation map. The sizes for the windows are different for different high-correlation areas.

Four combinations by the above two schemes were designed for comparison: 1) use the value of the single grid box at each high-correlation center, and no wind based predictors used; 2) use the average value of multiple grid boxes in a certain window, and no wind based predictors used; 3) use the single grid box value at each high-correlation center, with wind based predictors used; 4) use the average value over the multiple grid boxes, with wind based predictor used.

### 2.3.5 Variance inflation

In statistical downscaling, the models usually produce outputs with smaller variance than the observed, because the random variability is ignored by the models. Such underestimation of variance is not suitable for risk warning. Variance inflation techniques (Maraun et al., 2010) can be used to enlarge the variance. In this study, we use a simple variance inflation:

$$T_2 = \bar{T}_1 + (T_1 - \bar{T}_1) \frac{VAR^{1/2}(T)}{VAR^{1/2}(T_1)} \quad (1)$$

Here,  $T$  is the observed variable,  $T_1$  is the simulation of the same variable,  $\bar{T}_1$  is the mean

value of  $T_1$ ,  $T_2$  is the simulation with the variance inflated.  $VAR(.)$  represent the variance.

## 2.4 Estimating PM2.5 based on local meteorological data

Estimating PM2.5 from local ground based meteorological data is a common way. In this study, we also constructed PM2.5 models based on the local meteorological data. In following paragraphs, such models are represented as the 'local models', while the above models based on LSVs in ERAI are represented as the 'downscaling models'. Local models were constructed here is to figure out which model has the advantage. Local models based on GLMs were constructed to estimate the PM2.5 concentrations both at Jiaozuo (only Huanbaoju station) and Zhengzhou, based on the same meteorological data observed in Zhengzhou. The parameters were estimated by the LASSO-GLM method. Similarly, the one-winter-leave-out cross validation was also used here to obtain the Pearson's correlation coefficient (CC) between the model's output and the observed PM2.5, as a metric of the model's performance.

## 3. Result And Discussion

### 3.1 Statistics of PM2.5 concentration in Jiaozuo

Histograms for the daily mean and maximum PM2.5 in Jiaozuo show skewed distributions(Fig), indicating the Pearson's coefficient and common linear regression are not suitable to be used for PM2.5. Transformation is usually used to convert such skewed-distribution variables into more Gaussian distribution variables(Qian et al., 2015). Here, we use a logarithmic transformation.

$$\begin{aligned} PM'_{mean} &= \log(PM_{mean}) \\ PM'_{max} &= \log(PM_{max}) \end{aligned} \quad (2)$$

The logarithmic transformed data seem to be more Gaussian like. All the following modelling process is based on these logarithmic transformed data. Although common linear regression can be used here, but we still prefer GLM considering that GLM is also suitable for the data of normal distribution.

The Frequencies of the daily maximum PM2.5 in each hour (Fig. 2) show that daily peaks can happen in every hour, but mostly the peaks happen in two periods: 9:00-12:00 and 20:00-24:00. The frequencies of maximum at night (20:00-24:00) is much higher than those in 9:00-12:00. Similarly, the average PM2.5 for each hour shows two peaks in 10:00-11:00 and 20:00-21:00 and the latter is usually higher than the former.

### 3.2 Correlation maps

The CCs were calculated between the mean/max daily PM2.5 of Jiaozuo and the time series extracted at each grid box from the LSVs, then the correlation maps were produced. According to these maps, the high-correlation areas are not necessarily close to Jiaozuo or in its surrounding area.

### 3.2.1 Geopotential height

The correlation maps between the PM<sub>2.5</sub> and the geopotential height show that heavily polluted days correspond to the negative-correlation area at the Mongolian Plateau at the 500 hPa layer and the positive-correlation area stretching from the east of North China to the Korean Peninsula (Fig. 3). The negative-correlation area covering the south part of China and the positive-correlation area covering the Northwest Pacific and the Indian peninsula on the 850 hPa and 1000 hPa layers. This shows that the main circulation in heavy air pollution days of Jiaozuo is characterized by the relative low-pressure troposphere on the Mongolian Plateau and the relative high pressure from the east of North China to the Korean Peninsula, indicating that the cold air is weakly transported from the Mongolian Plateau to the south. The low pressure in the South Central China appears in the lower troposphere. At the location of Jiaozuo, the relative high pressure appears at the top of troposphere and the relative low pressure appears at the bottom of troposphere. The existence of relative high pressure over the Indian subcontinent and the Pacific Ocean indicates that the low-layer atmosphere tends to convergent circulation, thus, under such conditions, the pollution diffusion in eastern China is weak.

According to the correlation maps, there are one negative-correlation center (Mongolia-Siberia) and one positive-correlation center (North China-Korean Peninsula) at the 500hPa layer, while there is one negative correlation center (the south-eastern part of China) and two positive-correlation centers (Indian subcontinent and the Northwest Pacific). The geopotential at these centers were used as predictors to construct SD models. The correlation centers at 700hPa layer and 850hPa layer were not used here because these two layers can be regarded as the transition from the 500hPa layer to the 1000hPa layer and are not typical enough.

### 3.2.2 Relative humidity and specific humidity

The correlation maps calculated from relative humidity and specific humidity at the four pressure levels are generally similar and show a broken structure of high-correlation areas which can pass the significant tests at the 0.05 level. There are many such small patches of high or low correlation areas. Nevertheless, the broken structure obtained from specific humidity is not apparent as that from relative humidity.

From the correlation map of relative humidity at 500hPa, there is a negative-correlation belt stretching from the Indian subcontinent to Southwestern China and a tiny high-correlation region in the lower Yellow River Basin. At the 700hPa level, there are negative-correlation areas in India, Northeastern China, Northwest Pacific, a large positive-correlation area at the Mongolia-Siberia region and a small positive-correlation area between the lower reaches of the Huai River and the Yangtze River (the Jiang-Huai region). The specific humidity at 700hPa has some similar patterns as the relative humidity at 700hPa, except that there is no apparent positive-correlation area at the Mongolia-Siberia region, and no negative-correlation area in Northeastern China. At the 700hPa level, the positive-correlation region obtained by specific humidity at the Jiang-Huai region has a larger area than that obtained by the relative humidity. At the 850hPa/1000hPa level, the positive-correlation area covering the Jiang-Huai region becomes larger and stretches to the south of China. The negative correlation center is also apparent in the west Pacific Ocean. The above patterns indicate that heavy air pollution in Jiaozuo is related to high humidity in the southeast of China at the bottom of the Troposphere. From lower levels to upper levels of the troposphere, the high-correlation region surrounding Jiaozuo gradually moves northward and becomes smaller.

We select the three high-correlation centers from the relative humidity at 700hPa as model predictors: the negative-correlation area at India subcontinent, the negative-correlation area in Northeastern China and the positive-correlation area in the Jiang-Huai region. Two correlation centers from the relative humidity at 1000hPa are also selected: the negative-correlation center at the Northwest Pacific, and the positive-correlation center at the eastern part of China.

### 3.2.3 Temperature

The correlation maps obtained from air temperature show a high-positive center in the east of China and a negative region in the Mongolia-Siberia region, indicating that heavy PM<sub>2.5</sub> in Jiaozuo is related to the relatively warmer air temperature in the surrounding area of Jiaozuo and the relatively colder air temperature in the Mongolia-Siberia region, implying a stationary and weak diffusion condition for air pollution. Usually, in winter, when cold waves move through China, the air temperature in eastern China becomes colder, and the heavy air pollution can be blown away in a relatively short time. This circulation for cold-wave days is contrary to the above patterns of air temperature. From lower levels to upper levels in the troposphere, the high-correlation region surrounding Jiaozuo gradually moves northward, indicating the warm air mass moves to Eastern China earlier in the upper-level troposphere than in the lower-level troposphere. The positive-area of air temperature at 1000hPa in Eastern China seems broken or becomes weaker in a belt along the Taihang Mountain, and this weaker belt stretches southward to the Yangtze River basin. Considering all the above patterns, we selected both the centers in Mongolia-Siberia region and the eastern China as model predictors.

## 3.2.4 Wind

From the correlation maps from U-wind, a long belt of negative correlation stretches along the longitudinal direction, from North China to Japan, in 500hPa and 700hPa levels, indicating a relatively westward wind anomaly. In lower level troposphere (850hPa and 1000hPa), the negative area is connected with a large region in the Northwest Pacific, implying that the heavy air pollution in Jiaozuo is related to the westward wind anomaly widely existing in East Asia. Meanwhile, in the upper levels (500hPa and 700hPa), there is also a large and long positive-correlation area in the north of Asia continent, covering all the Central Asia, Mongolia, Northeastern China and Siberia, indicating a strong west wind anomaly in the upper-level troposphere. This pattern is not shown in lower levels, implying that it is related to the stable westerly jet.

In the upper levels of the U-wind, there is a large positive-correlation area from the whole area of East China to the Mongolia Plateau and Siberia, indicating a strong northward wind anomaly. Meanwhile a negative-correlation area exists in the Northwest Pacific, indicating a southward wind anomaly there. In the lower levels, this positive-correlation area becomes smaller and weaker, and the negative-correlation area seems moves westward and covers Northeastern China.

All the above patterns show the heavy air pollution is related to a low-pressure system with weakened wind blowing from the Mongolia-Siberia region to the surrounding areas in Eastern China. Under this atmospheric condition, the strong westerly in the subarctic area can prevent cold air mass from moving to mid-latitude East Asia. For wind variables, we select the negative center (South China) and the positive center (North China) of U850, V500 and V850 as predictors.

## 3.2.5 Main circulations in high/low PM2.5 days

By comparing the geopotential height at 500hPa in low-PM2.5 days and high-PM2.5 days, low-PM2.5 days reflect comparatively low pressure and strong northwestern wind over the southeast of China. In both high PM2.5 and low PM2.5 days, west wind dominates the area north to 35°N, indicating the stable westerly jet. Comparatively, in low-PM2.5 days, the wind in Southeastern China become more southerly.

At 850hpa, both in high-PM2.5 and low-PM2.5 days, there is a high pressure belt stretching from southwest to northeast in China, covering the Sichuan basin, Loess Plateau, North China and Northeastern China. A low-pressure area also exists in the Northwest Pacific. In the high-pressure belt, the area over North China in high-PM2.5 days poses a weak and divergent air flow to the surrounding areas.

Comparatively, in high-PM2.5 days, the pressure in this belt is not as high as that in low-PM2.5 days and the pressure in the Northwest Pacific is also not as low as that in low-PM2.5 days. Therefore, the land-sea pressure gradient during low-PM2.5 days are larger than that in high-PM2.5 days, producing a strong wind which is conducive to the dissipation of PM2.5 in the eastern part China.

Combining the geopotential heights at the upper level (500hpa) and the lower level(850hpa), it is obvious that high PM2.5 concentration usually corresponds to comparatively high pressure in upper air and low pressure in lower air over North China, which implies a static stratification of atmosphere. In low PM2.5 days, on the contrary, comparatively low pressure in upper air and high pressure in lower air imply a strong downward air flow, which is usually related to the cold wave coming from the north pole.

Table 1  
Screened predictors based on correlation maps. All these predictors are anomalies relative to the interannual mean during the winter months of 1979-2018.

| predictor       | variable | Correlation sign | area                           | longitude | latitude |
|-----------------|----------|------------------|--------------------------------|-----------|----------|
| X <sub>1</sub>  | Z500     | negative         | Mongolia-Siberia               | 73°E      | 61°N     |
| X <sub>2</sub>  | Z500     | positive         | North China, Korean Peninsular | 119°E     | 36°N     |
| X <sub>3</sub>  | Z1000    | negative         | Southern China                 | 113°E     | 29°N     |
| X <sub>4</sub>  | Z1000    | positive         | India subcontinent             | 75°E      | 19°N     |
| X <sub>5</sub>  | Z1000    | positive         | Northwest Pacific              | 137°E     | 33°N     |
| X <sub>6</sub>  | RH500    | negative         | India subcontinent             | 84°E      | 13°N     |
| X <sub>7</sub>  | RH700    | negative         | Northeast China                | 125°E     | 45°N     |
| X <sub>8</sub>  | RH700    | negative         | India subcontinent             | 84°E      | 15°N     |
| X <sub>9</sub>  | RH700    | positive         | Yangtze River                  | 115°E     | 32°N     |
| X <sub>10</sub> | RH1000   | positive         | North China                    | 115°E     | 32°N     |
| X <sub>11</sub> | RH1000   | negative         | Northwest Pacific              | 142°E     | 30°N     |
| X <sub>12</sub> | T700     | negative         | Mongolia-Siberia               | 73°E      | 61°N     |
| X <sub>13</sub> | T700     | positive         | North China                    | 119°E     | 36°N     |
| X <sub>14</sub> | T850     | negative         | Mongolia-Siberia               | 73°E      | 61°N     |
| X <sub>15</sub> | T850     | positive         | North China                    | 120°E     | 30°N     |
| X <sub>16</sub> | U850     | negative         | North China                    | 113°E     | 35°N     |
| X <sub>17</sub> | U850     | positive         | South China                    | 111°E     | 26°N     |
| X <sub>18</sub> | V500     | positive         | Loess Plateau                  | 107°E     | 36°N     |
| X <sub>19</sub> | V500     | negative         | Northwest Pacific              | 140°E     | 38°N     |
| X <sub>20</sub> | V850     | positive         | South China                    | 116°E     | 29°N     |
| X <sub>21</sub> | V850     | negative         | Northwest Pacific              | 143°E     | 23°N     |

### 3.3 Performance of the downscaling models

#### 3.3.1 Parameter estimation and the final models

The LASSO algorithm was used during the GLM training (parameter estimation), and some of the predictor parameters/coefficients were replaced by zeroes. The parameters were estimated based on four different sample sets, respectively and the RMSE and the CCs (based on the logarithmic transformed PM<sub>2.5</sub>) (Fig. 2-3). The predictor set including wind (15 predictor values) get better model performance than the predictor set without using wind variable (21 predictor values). Generally, the predictors extracted from averaged multiple grid boxes has poorer model performance than the single grid box based predictors. The differences between the simulated PM<sub>2.5</sub> under the four combinations of the comparison schemes (the single grid box combining 15 predictors, the average of multiple grid boxes combining 15-predictor, the single grid box combining 21 predictors, the average of multiple grid boxes combining 21-predictor) are generally small and share the similar variations (Fig. 6).

Table 2

Pearson's Correlation coefficients between the GLM simulated logarithmic daily PM2.5 and the observed ones in Jiaozuo, during the winter months of 2014-2018.

| predictand    | Validation period | Predictors from single grid box |            |               |            | Predictors from area-averaged multiple grid boxes |            |               |            |
|---------------|-------------------|---------------------------------|------------|---------------|------------|---|------------|---------------|------------|
|               |                   | 15 predictors                   |            | 21 predictors |            | 15 predictors                                     |            | 21 predictors |            |
|               |                   | calibration                     | validation | calibration   | validation | calibration                                       | validation | calibration   | validation |
| Mean PM2.5    | 2014              | 0.6156                          | 0.6490     | 0.6344        | 0.6756     | 0.6067  | 0.6441     | 0.6360        | 0.6731     |
|               | 2015              | 0.6370                          | 0.4286     | 0.6567        | 0.4756     | 0.6529  | 0.4803     | 0.6605        | 0.4709     |
|               | 2016              | 0.6271                          | 0.5257     | 0.6405        | 0.5768     | 0.6298  | 0.5373     | 0.6320        | 0.5678     |
|               | 2017              | 0.6227                          | 0.6572     | 0.6490        | 0.6445     | 0.6293  | 0.6258     | 0.6515        | 0.6204     |
| Maximum PM2.5 | 2014              | 0.5415                          | 0.5924     | 0.5709        | 0.6171     | 0.5610  | 0.6150     | 0.5679        | 0.6048     |
|               | 2015              | 0.5857                          | 0.4120     | 0.5981        | 0.4453     | 0.5955  | 0.4303     | 0.5987        | 0.4216     |
|               | 2016              | 0.5647                          | 0.4852     | 0.5637        | 0.5277     | 0.5619  | 0.4845     | 0.5578        | 0.4986     |
|               | 2017              | 0.5722                          | 0.5772     | 0.5842        | 0.5688     | 0.5729  | 0.5357     | 0.5833        | 0.5305     |

Table 3

RMSE obtained between the GLM simulated logarithmic daily PM2.5 and the observed ones in Jiaozuo, during the winter months of 2014-2018.

| predictand    | Validation period | Predictors from single grid box |            |               |            | Predictors from area-averaged multiple grid boxes |            |               |            |
|---------------|-------------------|---------------------------------|------------|---------------|------------|---|------------|---------------|------------|
|               |                   | 15 predictors                   |            | 21 predictors |            | 15 predictors                                     |            | 21 predictors |            |
|               |                   | calibration                     | validation | calibration   | validation | calibration                                       | validation | calibration   | validation |
| Mean PM2.5    | 2014              | 4.4775                          | 4.5896     | 4.4695        | 4.5722     | 4.4747  | 4.5679     | 4.4729        | 4.5609     |
|               | 2015              | 4.5305                          | 4.3888     | 4.5228        | 4.3934     | 4.5197  | 4.4044     | 4.5201        | 4.4007     |
|               | 2016              | 4.4312                          | 4.7066     | 4.4353        | 4.6845     | 4.4379  | 4.6796     | 4.4359        | 4.6859     |
|               | 2017              | 4.5331                          | 4.3628     | 4.5201        | 4.3936     | 4.5265  | 4.3833     | 4.5161        | 4.4097     |
| Maximum PM2.5 | 2014              | 4.9451                          | 5.0767     | 4.9442        | 5.0714     | 4.9471  | 5.0635     | 4.9472        | 5.0631     |
|               | 2015              | 4.9834                          | 4.9432     | 4.9820        | 4.9420     | 4.9796  | 4.9525     | 4.9821        | 4.9440     |
|               | 2016              | 4.9152                          | 5.1689     | 4.9124        | 5.1772     | 4.9206  | 5.1491     | 4.9146        | 5.1727     |
|               | 2017              | 5.0233                          | 4.8221     | 5.0138        | 4.8280     | 5.0150  | 4.8286     | 5.0101        | 4.8419     |

According to the CCs between the simulated and the observed, the models trained with samples from 2015-2017 get the largest CCs, both for the mean PM2.5 and the maximum PM2.5, with an averaged CC of 0.642 and 0.584, respectively. The models trained with the samples of 2015-2017 are as follows:

$$\ln(\text{PM}_{2.5,\text{mean}}) = 4.4483 - 0.0878X_1 + 0.0698X_4 + 0.0036X_5 - 0.0648X_6 - 0.0195X_8 + 0.2354X_{10}$$

$$- 0.1X_{11} - 0.0478X_{14} + 0.0083X_{15} + 0.0345X_{17} + 0.049X_{18} - 0.0182X_{19}$$

$$\ln(\text{PM}_{2.5,\text{max}}) = 4.9244 - 0.0568X_1 + 0.0679X_4 - 0.0239X_6 - 0.0397X_8 + 0.1914X_{10}$$

$$- 0.0649X_{11} - 0.0439X_{14} + 0.0041X_{15} + 0.0318X_{17} + 0.054X_{18} - 0.0176X_{19}$$

Here, the predictors  $X_1$ - $X_{19}$  are listed in Table 1.

We use these models based on samples in 2015-2017 to produce results for further analysis, since relatively good model performance was obtained based on these samples. Two additional PM2.5 series observed at other two sites (Gaoxinqu and Yingshicheng) in Jiaozuo were also used to validate the models.

For the models for daily mean PM<sub>2.5</sub>, the CCs between the observations and the modeled results are 0.52 for both the sites (Gaoxinqu and Yingshicheng). The average of daily mean PM<sub>2.5</sub> over the three sites in Jiaozuo has a CC of 0.55 to the modeled result. All the above CCs are lower than the CC obtained at Huanbaoju (0.56).

For the maximum PM<sub>2.5</sub>, the CCs obtained are 0.50, 0.48, 0.52 and 0.54, for Huanbaoju, Gaoxinqu and Yingshicheng and the average over the three stations, respectively. It is very interesting that the model was trained based on the data of station Huanbaoju, theoretically, the highest CC should be obtained for this station, but smaller CC was obtained instead than that for the three-station average and the station Yingshicheng. This demonstrates that the single station of Huanbaoju has some local variabilities which is related to some local conditions.

The simulated daily PM<sub>2.5</sub> mean/maximum values has variations roughly corresponding to the observed values (Fig. 6). The simulated peaks and valleys usually have one day leading or lagging behind the observed ones. This inconsistency maybe related to the local atmospheric condition which affects the pollution transportation. It is common that the heavy PM<sub>2.5</sub> episodes usually lasts for several days, and the simulated PM<sub>2.5</sub> in this study can reflect such multi-day variations.

### 3.3.2 Variance inflation

The simulation underestimated the daily variances of both mean PM<sub>2.5</sub> and maximum PM<sub>2.5</sub> (Fig. 7). Some specific peak values were not produced by the models. In order to match the simulated variances with the observed, variance inflation was used. The simulated standard deviation is 0.43 of the observed standard deviation. Considering the emission levels are different for each year, we use a different inflation factor for each year.

After this inflating, the magnitude of the variations can match the observed, however, very few peaks are produced by the simulation. The mean PM<sub>2.5</sub> are simulated with better performance than the maximum PM<sub>2.5</sub>, and the underestimation of maximum PM<sub>2.5</sub> is also more significant than that of the mean PM<sub>2.5</sub>. In contrast, the models perform the poorest in winter 2015 and perform the best in winter 2014. Similar good performance is obtained in 2017 as that in 2014.

After inflating the variance, the increase of RMSE is smaller than 0.01 µg/m<sup>3</sup>, while the relative biases are reallocated for the winter of each year, and the average absolute biases are reduced from 0.64 to 0.59 and from 0.72 to 0.57 respectively, for the mean PM<sub>2.5</sub> and maximum PM<sub>2.5</sub>. The variance inflation does not change CCs, and causes very small increase of biases, but can improve the risk warning on peak PM<sub>2.5</sub> days.

### 3.3.3 Outputs in special periods

In the special periods of the APEC Blue sky and SCO Blue sky, the simulated PM<sub>2.5</sub> at these days are also comparatively low, indicating that these blue-sky days were mainly caused by good atmospheric circulation (Fig. 8). Therefore, as for whether the emission control measures have a significant impact on reducing PM<sub>2.5</sub> concentrations in the two special periods, this study cannot give a conclusion. According to the experiments through numerical transport simulation, the emission control measures explains a PM<sub>2.5</sub> reduction of -21.8% in Beijing (Gao et al., 2017).

Similarly, in the December 2017, the even-odd-number traffic control in multiple urban areas shows no significant impact on PM<sub>2.5</sub>, since the observed PM<sub>2.5</sub> also has several peak days corresponding to the downscaled results.

In the three spring festivals of 2015 (in winter 2014), 2016 (in winter 2015) and 2017 (in winter 2016), the PM<sub>2.5</sub> values are relatively small. However, according to the downscaling results, there should be a peak value during the spring festival of 2015 and 2016. During the Spring Festival of 2016, there is actually a peak observed, but it is lower than the corresponding peak simulated by the downscaling model (Fig. 9). During the Spring Festival of 2017, although the fireworks were banned in the main urban areas, however, burning fireworks in other wide areas were still allowed and the emissions certainly has important effects on PM<sub>2.5</sub>. This effect is also not significant by comparing with the simulated PM<sub>2.5</sub>.

### 3.3.4 Downscaled annual PM<sub>2.5</sub> during 1979-2017

In this study, the historical simulation for the winters of 1979-2017 was produced purely based on large-scale atmospheric predictors without considering the annual changes of emissions. The mean daily PM<sub>2.5</sub> and maximum PM<sub>2.5</sub> are shown Fig. 10. Mann-Kendall tests (Kendall, 1955; Mann, 1945) shown that both the simulated average PM<sub>2.5</sub> mean and maximum have decreasing trends which can pass test at the 95% significance level, during 1979-2017. This result implies that under the same assumed anthropogenic and natural emission level, due to the large-scale atmospheric factors, the PM<sub>2.5</sub> concentrations is generally dropping. This result is contrary to the

recent observed trend which is largely caused by the increased pollutant emission. The dropping trend is mainly resulted from the lower values after 2005.

Theoretically, the relation between pan evaporation and PM2.5 is very complex. Nevertheless, the pan evaporation in high PM2.5 days is comparatively low because in hazy days, the high relative humidity and weak air flow can reduce water evaporation. In this study, the CC between the observed EVP (represented by the data observed at Zhengzhou station) and observed PM2.5 during winters of 2014-2017 is -0.44, while that between the observed EVP at Zhengzhou and the simulated PM2.5 is -0.42 during the same period. (R=-0.39 during 1979-2018).

### 3.3.5 Comparison between the downscaling models and the local meteorological data based models

Table 4  
Correlation coefficients (CC) and rooted mean squared error (RMSE) obtained by the PM2.5 (daily mean) concentration models based on local meteorological data

|      | Calibration period  | Zhengzhou   |            | Jiaozuo     |            |
|------|---------------------|-------------|------------|-------------|------------|
|      |                     | Calibration | Validation | Calibration | Validation |
| R    | Nov.,2014-Feb.,2015 | 0.613       | 0.638      | 0.627       | 0.669      |
|      | Nov.,2015-Feb.,2016 | 0.615       | 0.518      | 0.666       | 0.487      |
|      | Nov.,2016-Feb.,2017 | 0.614       | 0.536      | 0.624       | 0.623      |
|      | Nov.,2017-Feb.,2018 | 0.565       | 0.720      | 0.413       | 0.705      |
|      | Nov.,2018-Feb.,2019 | 0.613       | 0.553      | 0.644       | 0.531      |
| RMSE | Nov.,2014-Feb.,2015 | 4.491       | 4.783      | 4.446       | 4.653      |
|      | Nov.,2015-Feb.,2016 | 4.545       | 4.551      | 4.518       | 4.330      |
|      | Nov.,2016-Feb.,2017 | 4.526       | 4.634      | 4.426       | 4.734      |
|      | Nov.,2017-Feb.,2018 | 4.573       | 4.425      | 4.400       | 4.361      |
|      | Nov.,2018-Feb.,2019 | 4.556       | 4.496      | 4.488       | 4.462      |

The local meteorological data based models for daily mean PM2.5 concentrations were trained by the LASSO-GLM algorithm. In all cases based on different sample sets of the cross validation, the parameters for air temperature and pan evaporation were set to zeros by the LASSO algorithm, indicating that the useful signals carried by these two variables can be represented by the signals in other variables. The PM2.5 models for Zhengzhou and Jiaozuo (based on the same meteorological data observed in Zhengzhou) are as follows,

Zhengzhou:  $\text{Log}(\text{PM}_{2.5}) = 4.4785 - 0.0374\text{Prs} + 0.1857\text{RH} - 0.1997\text{wind} - 0.0847\text{SSD}$

Jiaozuo:  $\text{Log}(\text{PM}_{2.5}) = 4.4410 - 0.0086\text{Prs} + 0.2657\text{RH} - 0.1493\text{wind} - 0.0562\text{SSD}$

The parameters of the above two models were obtained by the sample set in winter months of 2015-2018, which generally obtained the largest average CCs in calibration and validation periods, as fitted by different sample sets of cross validation (Table 4). The values of R and RMSE obtained by these local meteorological based models are close to that obtained by the downscaling models.

The outputs of the downscaling models and the local meteorological data based models were shown in Fig. 11(a). It is clear that the observed daily PM2.5 values in Jiaozuo and Zhengzhou are generally consistent to each other, indicating that the air quality in these two cities are highly controlled by the same weather system and that using the meteorological data observed in Zhengzhou to represent the weather in Jiaozuo is reasonable. The outputs of the local model and the downscaling model produces similar variations although they are not completely the same, and both of them underestimated the observed values. Based on the comparison between the modeled outputs and the observations, the downscaling models for Jiaozuo get a larger R value than the PM2.5 model based on local meteorological data, and the R value obtained by the local model for Jiaozuo is slightly smaller than that for Zhengzhou.

## 4. Conclusion

In this study, a statistical downscaling model of the winter daily PM<sub>2.5</sub> in a city (Jiaozuo, Henan Province) in North China was constructed based on large-scale meteorological variables. The predictors were screened based on the correlation maps between the large-scale variables in ERA-Interim reanalysis and the observed daily mean and maximum PM<sub>2.5</sub> concentrations. Just as some statistical downscaling studies for precipitation, the high-correlation areas reflecting the large-scale circulations around the site is usually not coincident with the location of the site. Within the same large-scale variable, the high-correlation areas are different among multiple pressure levels. The high-correlation areas covering the city Jiaozuo indicate that high PM<sub>2.5</sub> concentration is usually related to comparatively low pressure in the lower atmosphere and high pressure in the upper atmosphere, high relative humidity in the lower atmosphere, high temperature in all the troposphere and weak winter monsoon. Besides this, some areas far from the Jiaozuo location also have high correlations to the PM<sub>2.5</sub> of Jiaozuo: the low geopotential height and low air temperature in the upper troposphere in the west Mongolia and Siberia, the high geopotential height in the lower troposphere around Japan and the Northwest Pacific, the low humidity at the India subcontinent and the Northwest Pacific, the west wind anomaly in the subarctic areas and the southward wind anomaly in the Northwest Pacific. This demonstrate that predictors for estimating PM<sub>2.5</sub> can be selected from some positions far from the location of the observed predictand.

Among the multiple high-correlation areas, we selected some values of the large-scale variables as predictors, then trained GLMs using the LASSO-GLM algorithm to eliminate the overfitting problem which is caused by the collinearity in the multiple predictors. According the cross validation, the predictor extracted from the multiple grid boxes has no advantage over the predictor extracted from the single grid box, especially for the daily maximum PM<sub>2.5</sub>. Meanwhile, the incorporation of predictors from wind variables can give rise to better performance than the predictors without using wind variables. Nevertheless, such different choices result in very small difference in the model outputs.

The standard deviation of the modeled PM<sub>2.5</sub> explains 0.43 of the observed counterpart. The CCs between the model outputs and the observed PM<sub>2.5</sub> (on the logarithmic transformed version) is 0.607 and 0.548, for the mean daily PM<sub>2.5</sub> and the maximum daily PM<sub>2.5</sub>, respectively. Since the modeled outputs underestimated the variance, variance inflation was used to the outputs, then the relative bias of the modeled outputs was reduced significantly with a very small increasing of RMSE. Compared to the PM<sub>2.5</sub> estimating models based on the local ground observed meteorological data, the downscaling models based on large scale circulations have a better performance.

During the periods of "APEC Blue Sky" and "SCO Blue Sky", the statistical downscaling model also reproduced the low PM<sub>2.5</sub> levels, implying that the large-scale circulation plays a main role in these blue-sky days. The effect of emission control measures during these periods is difficult to quantify by comparing the observed PM<sub>2.5</sub> with the simulation of such statistical models. Similarly, the restriction on traffic control in December, 2017 and the emission from fireworks for celebrating Chinese Spring Festival also produce no significant effects on daily PM<sub>2.5</sub> variations when compared with the effects of atmosphere circulation conditions. In some cases, theoretically, burning fireworks can aggravate air pollution, but in the Spring Festivals of 2015 and 2016, the observed air quality is very good while the statistical model simulated medium peaks.

Without considering the changes of pollutant emissions, the simulated winter PM<sub>2.5</sub> with the statistical downscaling models driven by the large-scale circulations during 1979-2017 has a slight but statistically significant decreasing trend, which is contrary to the observed increasing trend or that felt by the public in recent years due to the increased pollutant emissions. Similar to the observed daily PM<sub>2.5</sub>, the simulated daily PM<sub>2.5</sub> values in 1979-2017 have a weak correlation to pan evaporation.

Overall, the CCs obtained by the statistical downscaling models are not ideally large, but it definitely demonstrates that the SD models can reflect the important impacts of the large-scale circulation on the PM<sub>2.5</sub>. It is promising that there is some space to improve the model performance by combining more predictors of local meteorological factors, local emissions or pollutant transportation with these large-scale predictors. Meanwhile, in this study, not all the useful signals in the high-correlation areas were used and only the signal at some of the high-correlation centers were used, therefore, the model performance perhaps may be improved by utilizing more signals, which need more efficient data-mining methods. Moreover, such SD models can be used to derive future projections of PM<sub>2.5</sub> for local sites, based on the rough-resolution outputs of global climate models. Although only the PM<sub>2.5</sub> at a site of Jiaozuo is used to construct the downscaling models, the high correlation areas of each large-scale variable to different sites in surrounding areas may be similar, since the PM<sub>2.5</sub> concentration values among the neighboring sites are highly correlated to each other.

## Declarations

### Declaration of Competing Interest

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.

### **CRedit authorship contribution statement**

**Yonghe Liu:** Writing - review & editing. **Yang Li:** Data processing, software. **Mingshi Wang:** Methodology. **Hailing Wang:** Methodology. **Xiyue Wang:** Visualization.

### **Acknowledgement**

This work is funded jointly by the Special fund for preventing air pollution in Jiaozuo (Grant No. H21-169) and the Key Scientific Research Projects of Universities in Henan Province (Grant No. 19A170007).

### **Code availability**

No special code is used.

### **Data availability**

Data will be made available on reasonable request from the corresponding author.

### **Ethics approval**

The authors have agreed for authorship, read and approved the manuscript, and given consent for submission and subsequent publication of the manuscript.

### **Consent to participate**

All authors consent to participate of the present study.

### **Consent for publication**

All authors read the manuscript and agree for the publication.

### **Conflicts of interest**

The authors declare that there is no conflict of interest.

## **References**

1. Analitis, A. et al., 2020. Prediction of PM<sub>2.5</sub> concentrations at the locations of monitoring sites measuring PM<sub>10</sub> and NO<sub>x</sub>, using generalized additive models and machine learning methods: A case study in London. *ATMOSPHERIC ENVIRONMENT*, 240(117757).
2. Araya-Osses, D., Casanueva, A., Roman-Figueroa, C., Manuel Uribe, J. and Paneque, M., 2020. Climate change projections of temperature and precipitation in Chile based on statistical downscaling. *CLIMATE DYNAMICS*, 54(9-10): 4309-4330.
3. Chen, J. et al., 2019. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *ENVIRONMENT INTERNATIONAL*, 130(104934).
4. Chen, Z. et al., 2018. Understanding meteorological influences on PM<sub>2.5</sub> concentrations across China: a temporal and spatial perspective. *ATMOSPHERIC CHEMISTRY AND PHYSICS*, 18(8): 5343-5358.
5. Chen, Z. et al., 2020. Influence of meteorological conditions on PM<sub>2.5</sub> concentrations across China: A review of methodology and mechanism. *ENVIRONMENT INTERNATIONAL*, 139(UNSP 105558).
6. Dee, D.P. et al., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY*, 137(656A): 553-597.
7. Fu, G., Charles, S.P., Chiew, F.H.S., Ekstrom, M. and Potter, N.J., 2018. Uncertainties of statistical downscaling from predictor selection: Equifinality and transferability. *ATMOSPHERIC RESEARCH*, 203: 130-140.
8. Gao, M. et al., 2017. Distinguishing the roles of meteorology, emission control measures, regional transport, and co-benefits of reduced aerosol feedbacks in "APEC Blue". *ATMOSPHERIC ENVIRONMENT*, 167: 476-486.

9. Gutierrez, J.M. et al., 2019. An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. *INTERNATIONAL JOURNAL OF CLIMATOLOGY*, 39(9SI): 3750-3785.
10. Hammami, D., Lee, T.S., Ouarda, T.B.M.J. and Lee, J., 2012. Predictor selection for downscaling GCM data with LASSO. *JOURNAL OF GEOPHYSICAL RESEARCH-ATMOSPHERES*, 117(D17116).
11. Hertig, E., Trambly, Y., Romberg, K., Kaspar-Ott, I. and Merckenschlager, C., 2019. The impact of soil moisture on precipitation downscaling in the Euro-Mediterranean area. *CLIMATE DYNAMICS*, 52(5-6): 2869-2884.
12. Horton, P. and Bronnimann, S., 2019. Impact of global atmospheric reanalyses on statistical precipitation downscaling. *CLIMATE DYNAMICS*, 52(9-10): 5189-5211.
13. Kendall, M.G., 1955. Rank Correlation Methods. Charles Griffin and Company, 196 pp.
14. Kunzli, N. et al., 2005. Ambient air pollution and atherosclerosis in Los Angeles. *ENVIRONMENTAL HEALTH PERSPECTIVES*, 113(2): 201-206.
15. Laura Bettolli, M. and Clorinda Penalba, O., 2018. Statistical downscaling of daily precipitation and temperatures in southern La Plata Basin. *INTERNATIONAL JOURNAL OF CLIMATOLOGY*, 38(9): 3705-3722.
16. Lin, H. et al., 2017. Mortality benefits of vigorous air quality improvement interventions during the periods of APEC Blue and Parade Blue in Beijing, China. *ENVIRONMENTAL POLLUTION*, 220(A): 222-227.
17. Liu, Y., Feng, J., Shao, Y. and Li, J., 2019. Identify optimal predictors of statistical downscaling of summer daily precipitation in China from three-dimensional large-scale variables. *ATMOSPHERIC RESEARCH*, 224: 99-113.
18. Lucio-Eceiza, E.E., Fidel Gonzalez-Rouco, J., Garcia-Bustamante, E., Navarro, J. and Beltrami, H., 2019. Multidecadal to centennial surface wintertime wind variability over Northeastern North America via statistical downscaling. *CLIMATE DYNAMICS*, 53(1-2): 41-66.
19. Mann, H.B., 1945. Nonparametric tests against trend. *Econometrica*(13): 245-259.
20. Maraun, D. et al., 2010. PRECIPITATION DOWNSCALING UNDER CLIMATE CHANGE: RECENT DEVELOPMENTS TO BRIDGE THE GAP BETWEEN DYNAMICAL MODELS AND THE END USER. *REVIEWS OF GEOPHYSICS*, 48(RG3003).
21. Pryor, S.C. and Schoof, J.T., 2019. A hierarchical analysis of the impact of methodological decisions on statistical downscaling of daily precipitation and air temperatures. *INTERNATIONAL JOURNAL OF CLIMATOLOGY*, 39(6): 2880-2900.
22. Qian, C., Zhou, W., Fong, S.K. and Leong, K.C., 2015. Two Approaches for Statistical Prediction of Non-Gaussian Climate Extremes: A Case Study of Macao Hot Extremes during 1912-2012. *JOURNAL OF CLIMATE*, 28(2): 623-636.
23. Wang, P. and Dai, X., 2016. "APEC Blue" association with emission control and meteorological conditions detected by multi-scale statistics. *ATMOSPHERIC RESEARCH*, 178: 497-505.
24. Wang, Y. et al., 2020. Acute and chronic health impacts of PM<sub>2.5</sub> in China and the influence of interannual meteorological variability. *ATMOSPHERIC ENVIRONMENT*, 229(117397).
25. Xiao, Q. et al., 2020. Changes in spatial patterns of PM<sub>2.5</sub> pollution in China 2000-2018: Impact of clean air policies. *ENVIRONMENT INTERNATIONAL*, 141(UNSP 105776).
26. Zhang, L. et al., 2020. Spatiotemporal variations and influencing factors of PM<sub>2.5</sub> concentrations in Beijing, China. *ENVIRONMENTAL POLLUTION*, 262(114276).
27. Zhang, Q. et al., 2019. Drivers of improved PM<sub>2.5</sub> air quality in China from 2013 to 2017. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 116(49): 24463-24469.
28. Zhang, Q., He, K. and Huo, H., 2012. Cleaning China's air. *NATURE*, 484(7393): 161-162.
29. Zhang, X. and Yan, X., 2015. A new statistical precipitation downscaling method with Bayesian model averaging: a case study in China. *CLIMATE DYNAMICS*, 45(9-10): 2541-2555.
30. Zhao, H. et al., 2020. How aerosol transport from the North China plain contributes to air quality in northeast China. *SCIENCE OF THE TOTAL ENVIRONMENT*, 738(139555).
31. Zheng, Y., Zhang, Q., Liu, Y., Geng, G. and He, K., 2016. Estimating ground-level PM<sub>2.5</sub> concentrations over three megalopolises in China using satellite-derived aerosol optical depth measurements. *ATMOSPHERIC ENVIRONMENT*, 124(B): 232-242.

## Figures

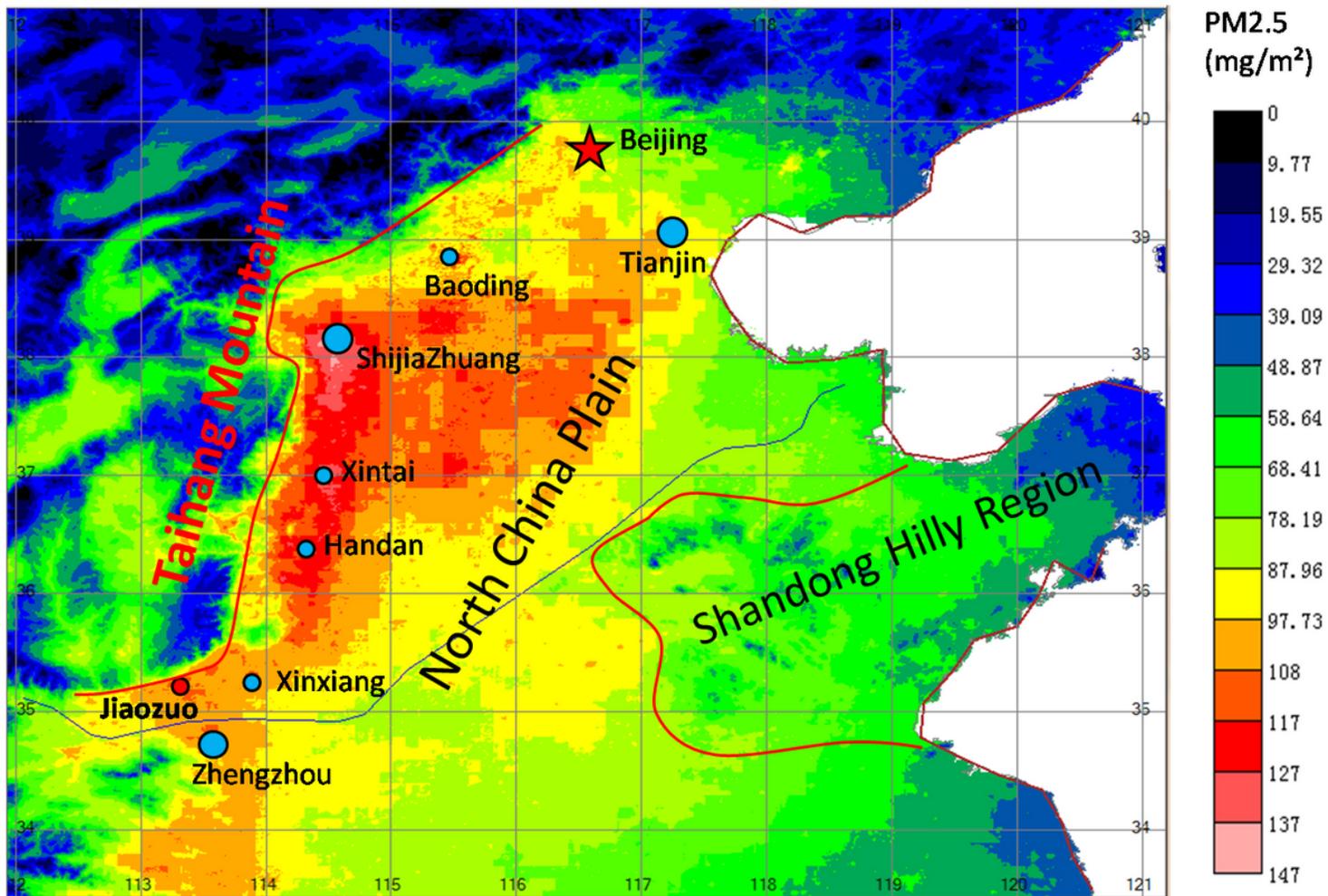


Figure 1

Spatial pattern (represented by the background colors) of average PM2.5 in North China, 2013. The red line on the left represent the boundary between the Taihang Mountain and the North China Plain. Only some cities in the highly polluted areas along the foot of the Taihang Mountain are sketched here. The different sized round dots represent the cities in different administrative levels.

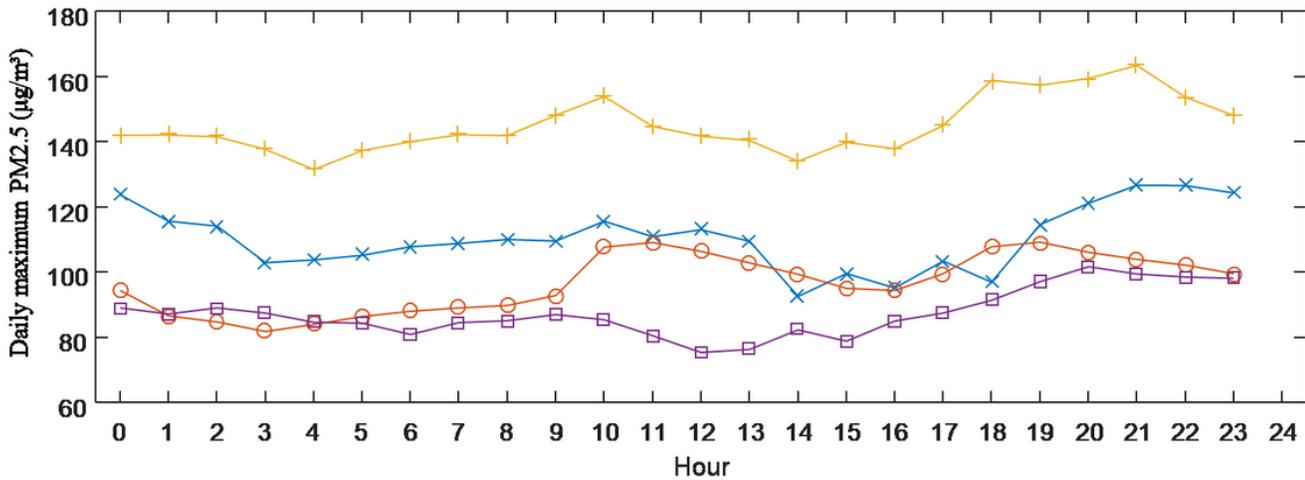
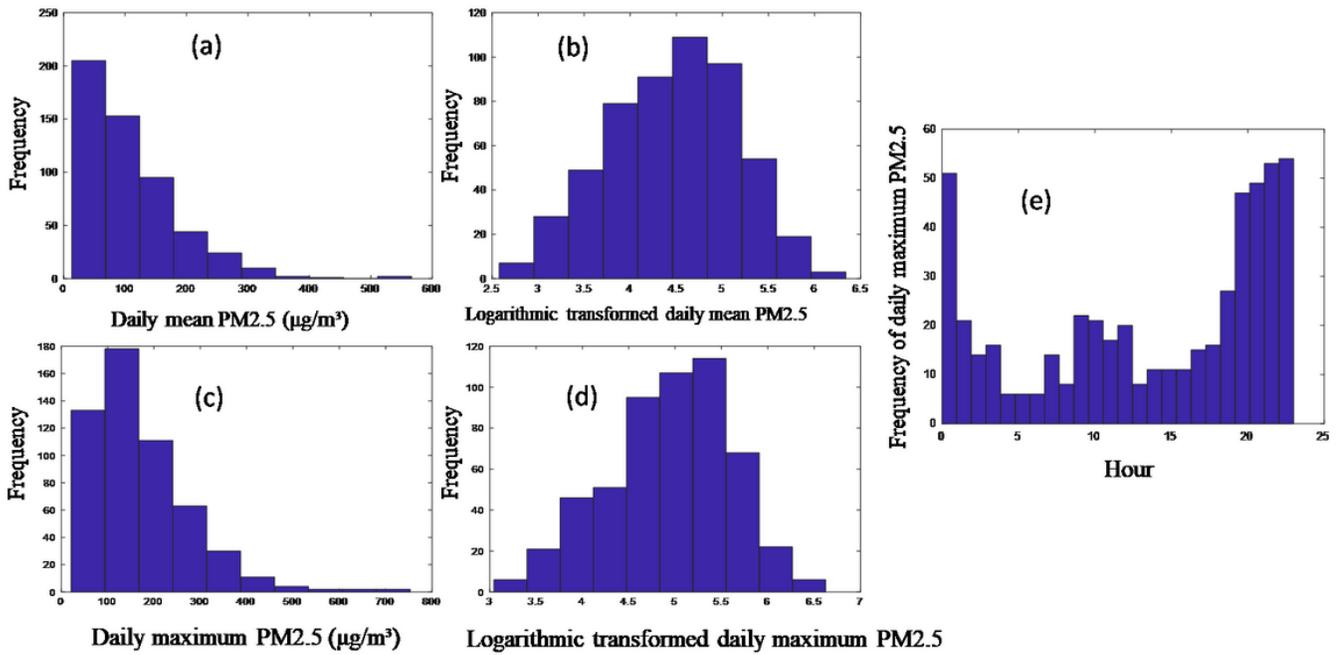


Figure 2

Distributions of the PM2.5 concentrations: Histograms for mean (a) and maximum (c) daily PM2.5 concentrations, the logarithmic transformed mean (b) and maximum(d), frequencies of the daily maximum in each hour (e), and the mean of daily maximum PM2.5 concentration (µg/m<sup>3</sup>) in each hour (f).

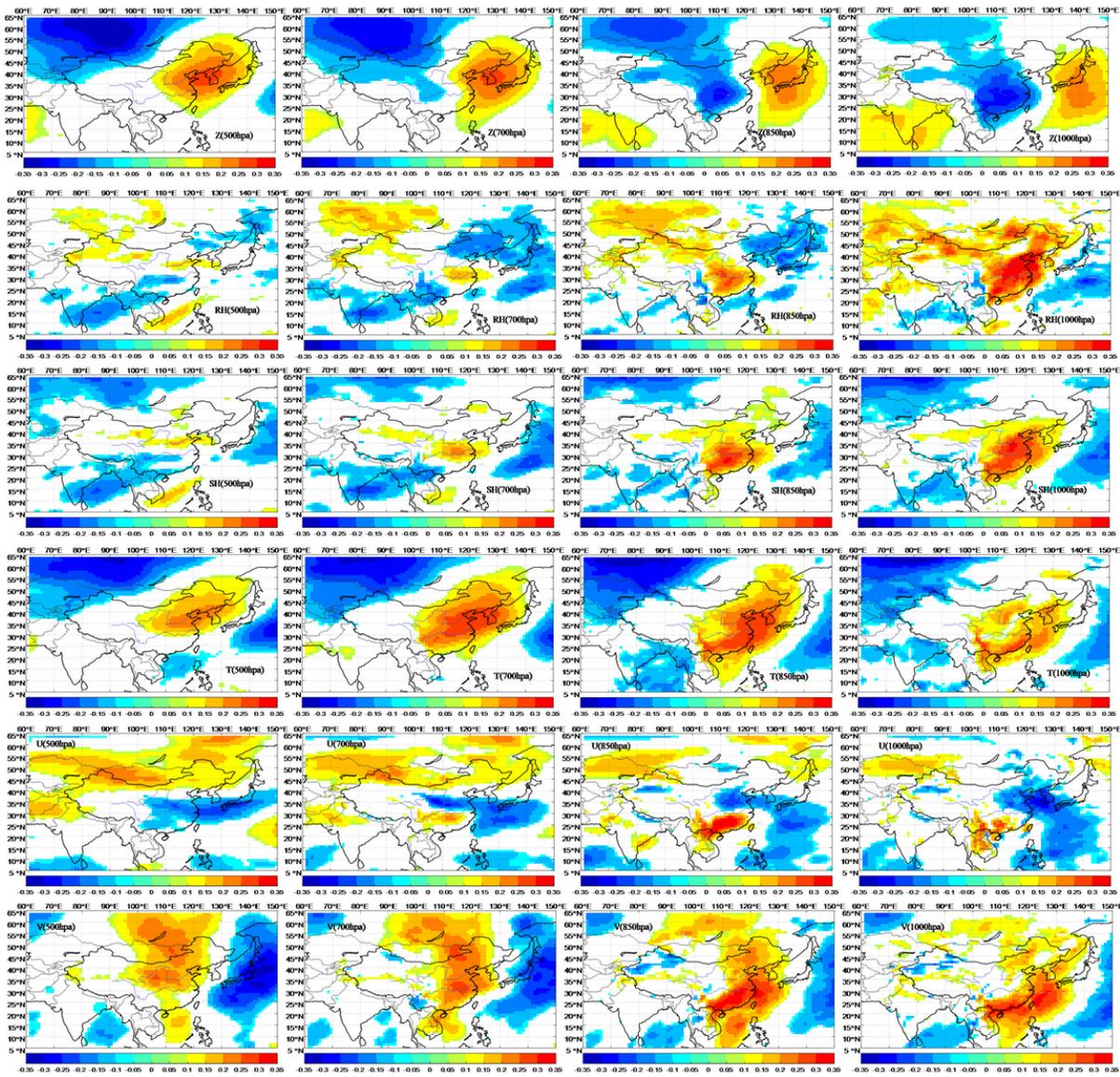
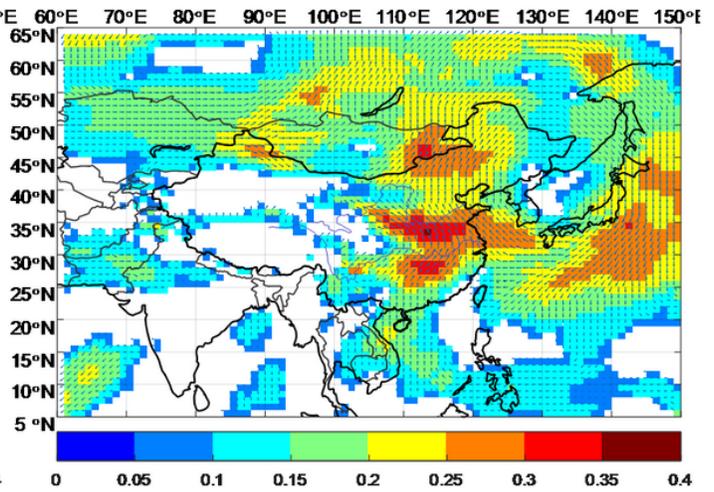
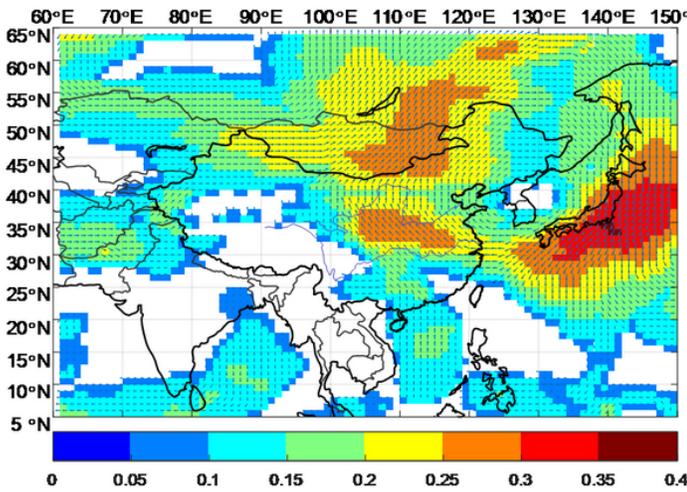


Figure 3

Correlation maps calculated between time series of the large-scale variables at each grid box and the observed mean daily PM2.5 in Jiaozuo

500hpa

700hpa



850hpa

1000hpa

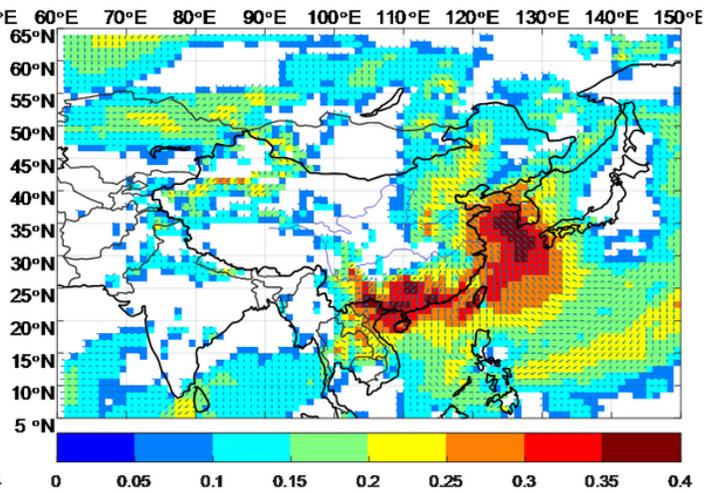
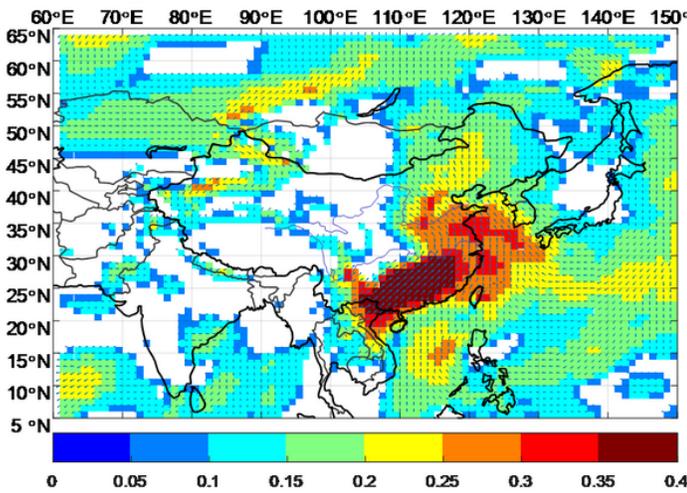


Figure 4

Combination of the correlations from the anomalies of U-wind and V-wind to the observed mean daily PM2.5 in Jiaozuo. The different colors represent the absolute values of the UV vectors comprised by the two correlations.

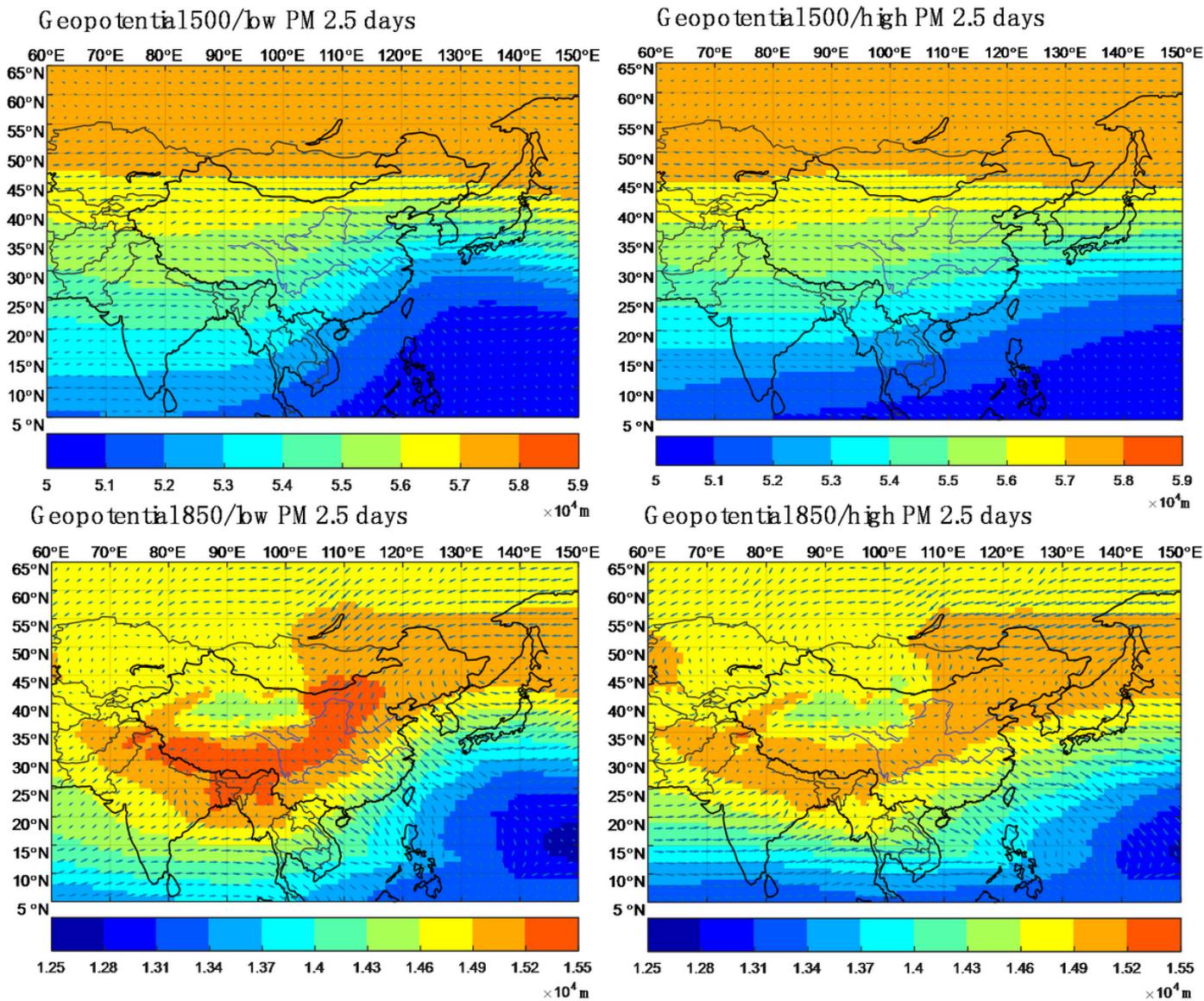


Figure 5

Averaged wind and the geopotential height (m) at 500hPa and 850hPa for low PM2.5 days and high PM2.5 days

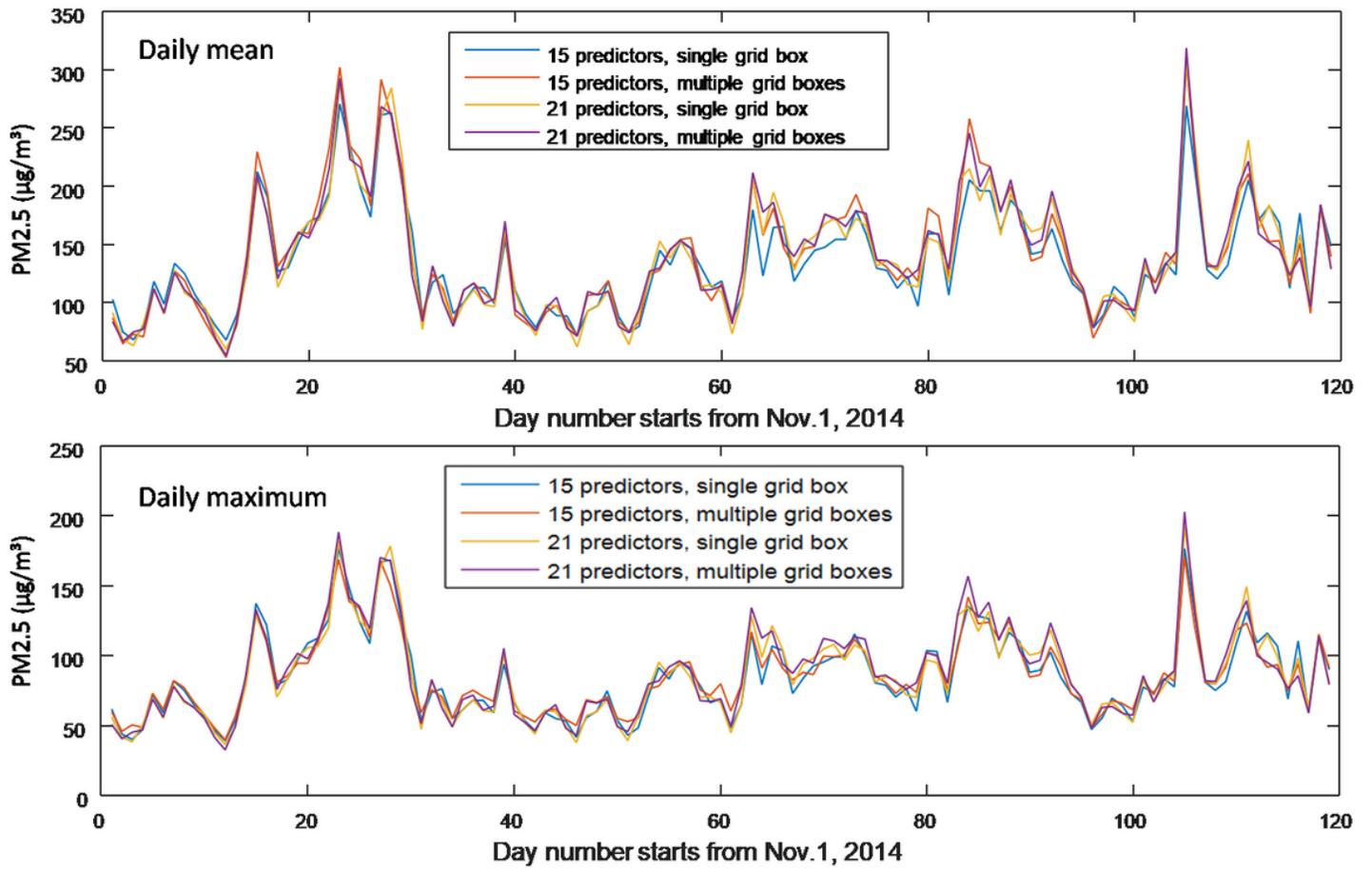
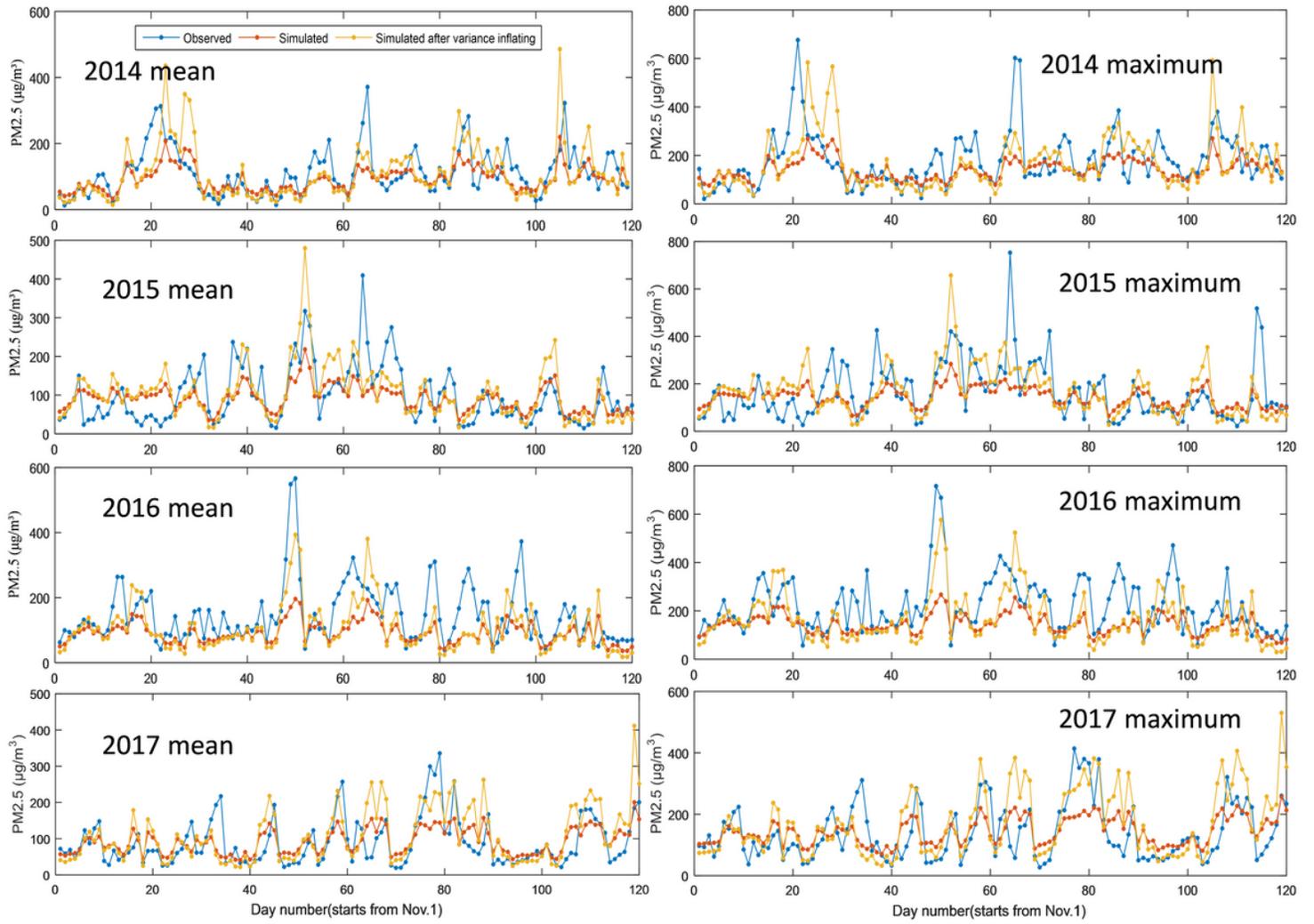


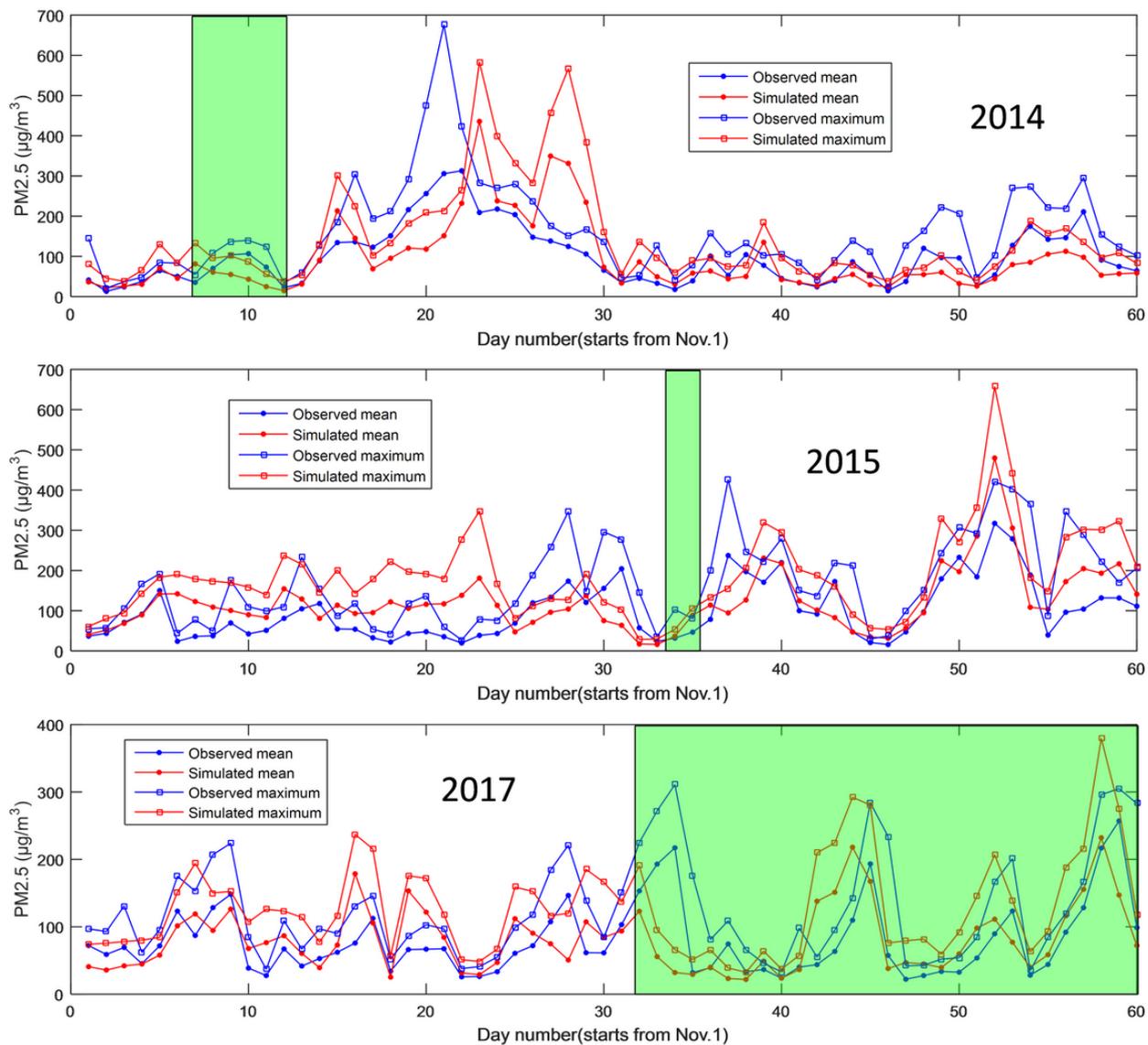
Figure 6

Simulated variation of mean values and extreme values in 2014 by different models



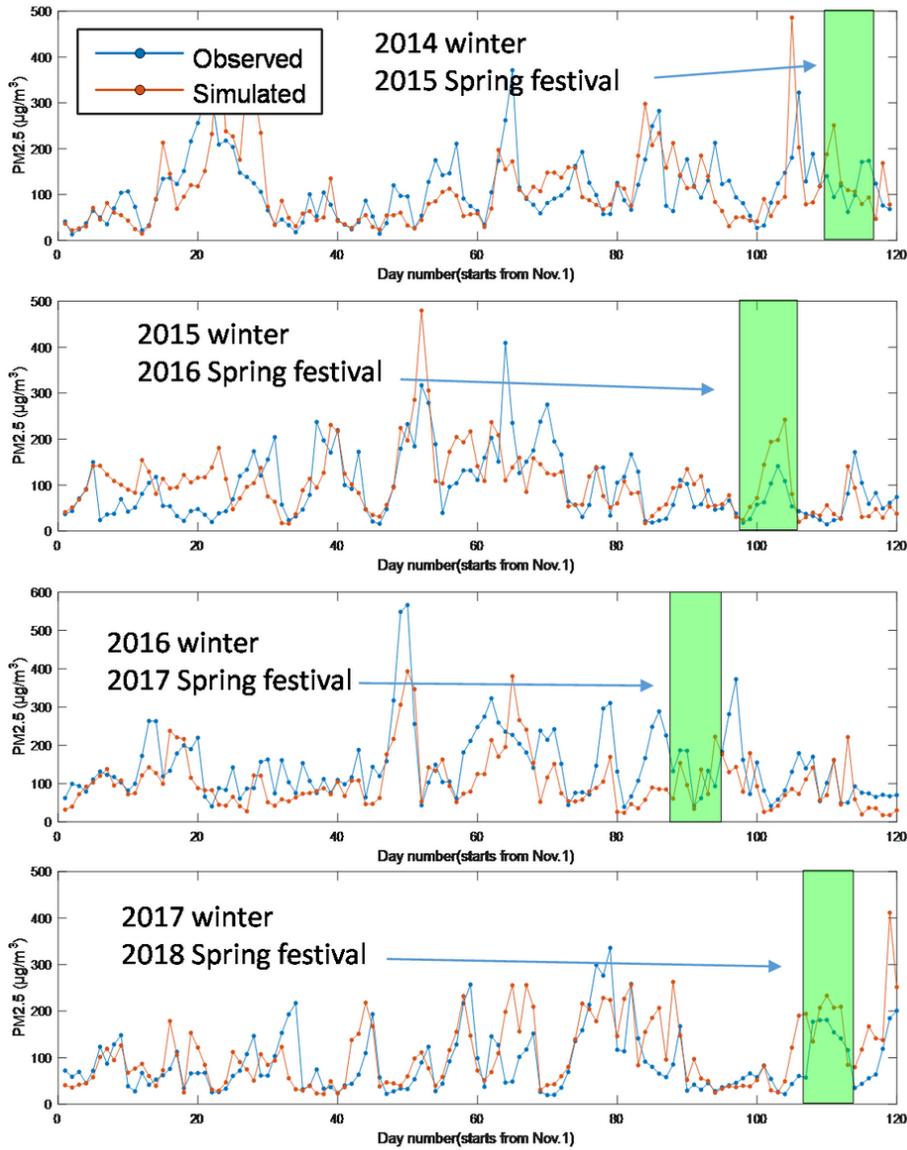
**Figure 7**

Variance inflation on the simulated daily mean PM2.5(a-d) and daily maximum PM2.5(e-h)



**Figure 8**

Simulated PM2.5 and the observed PM2.5 in Jiaozuo in three special periods: the “APEC Blue” (Nov.7-12, 2014), the “SCO Blue” (Dec. 14-15, 2015) and the traffic control with even and odd-numbered license plates (Dec.4-31, 2017)



**Figure 9**

Simulated and the observed PM2.5 concentration in the winter months during 2014-2017 (The periods marked in green are in Spring's Festival)

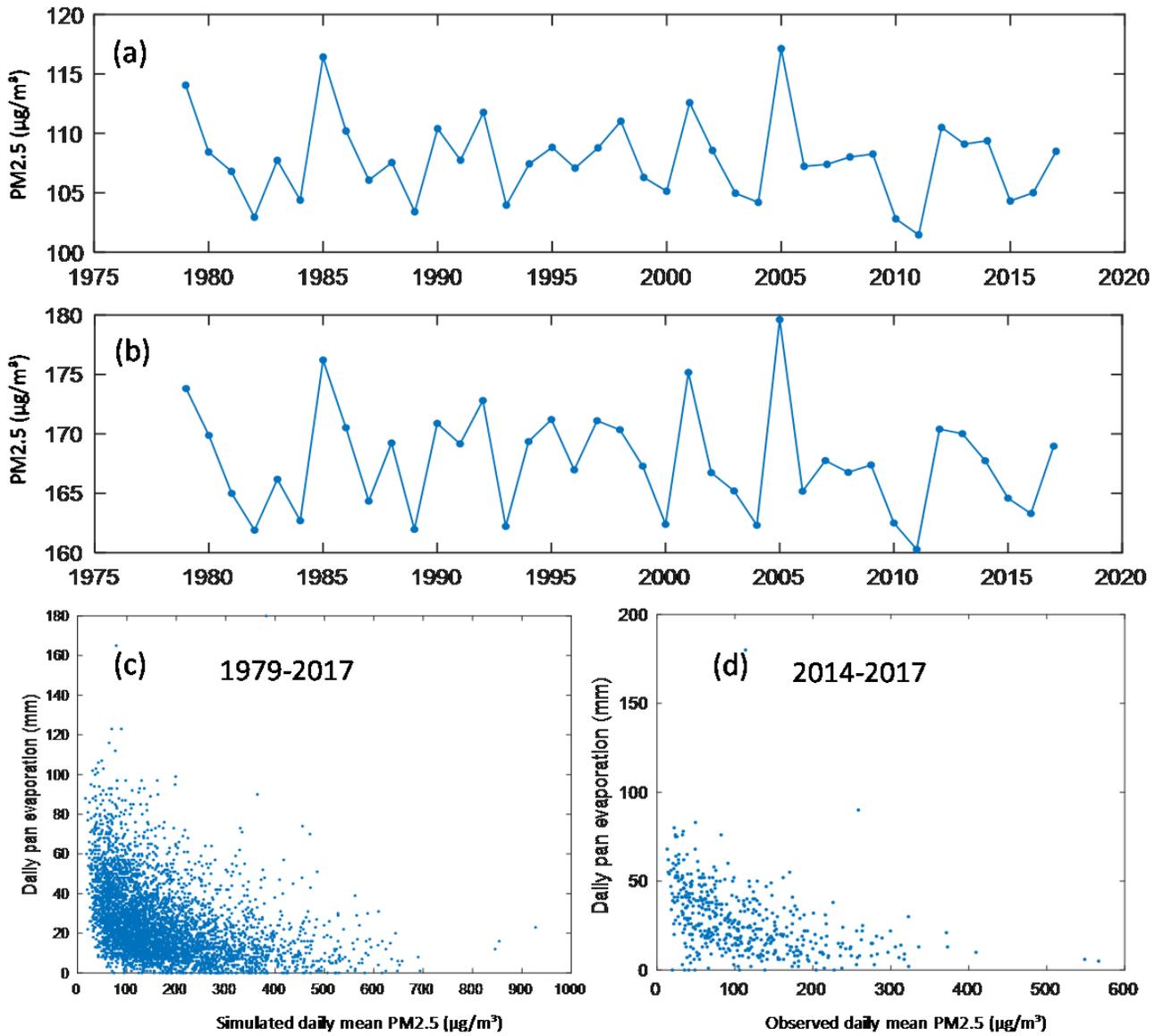


Figure 10

Simulated annual average of daily Mean (a) and daily maximum (b), and the relation between daily PM2.5 and the observed pan evaporation: (c) the simulated daily mean PM2.5 versus pan evaporation; (d) the observed daily mean PM2.5 versus pan evaporation.

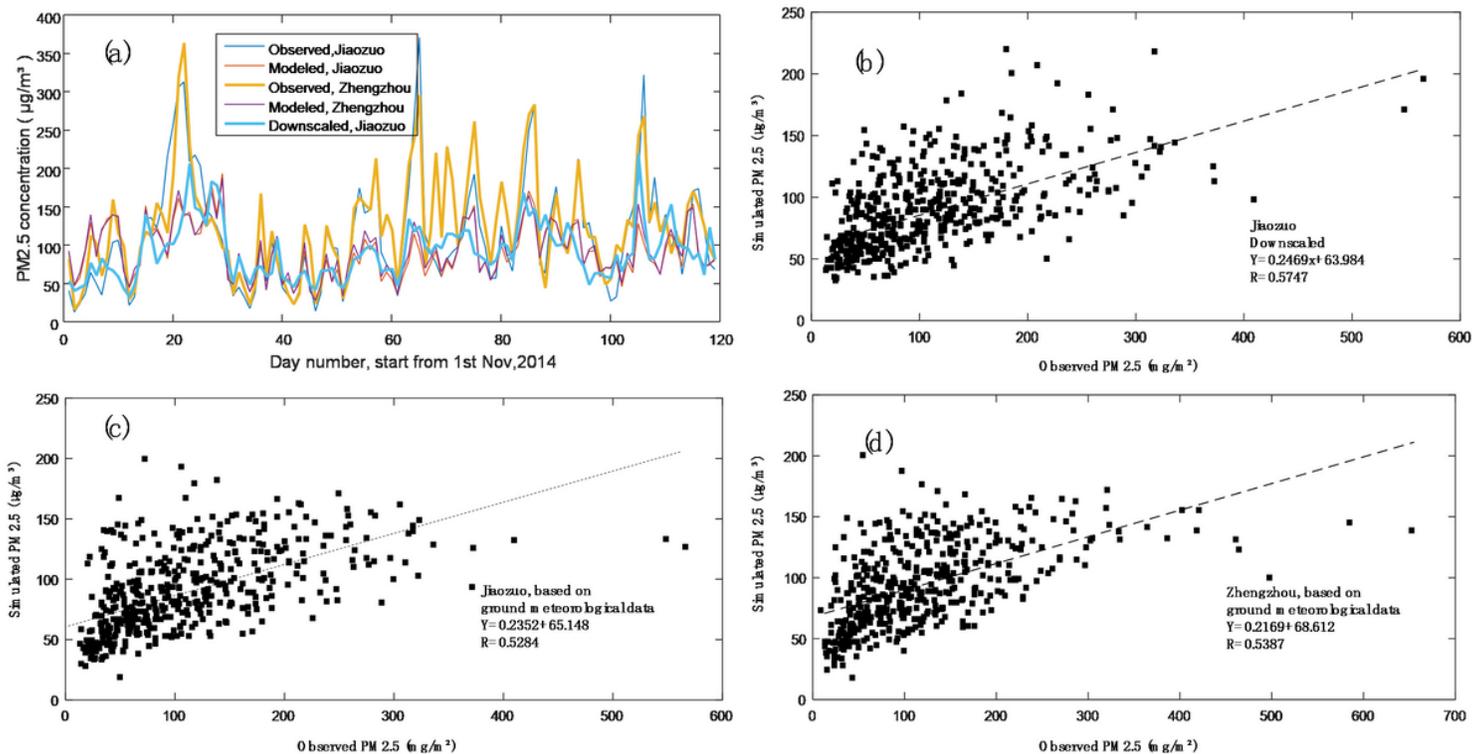


Figure 11

Comparison of performances between the large-scale circulation based downscaling models and the local meteorological data based models (a. modeled daily outputs of different models and the observed daily PM<sub>2.5</sub> concentrations, from November 1, 2014 to February 28, 2015. b. downscaled output versus the observed in Jiaozuo, c. the output of the local model versus the observed, for the PM<sub>2.5</sub> in Jiaozuo, d. the output of the local model versus the observed, for the PM<sub>2.5</sub> in Zhengzhou). The metrological data for the local models of Jiaozuo and Zhengzhou are the same, which were observed at the Zhengzhou meteorological station. In a-c, the dots represent all the samples in the winter months of 2014-2018.