Supporting Information for "Improving $PM_{2.5}$ Forecasts in China Using an Initial Error Transport Model"

Huangjian Wu,[*,†] *Xiaogu Zheng,*[‡] *Jiang Zhu,*[‡,§] *Wei Lin,*[*,‖] *Haitao Zheng,*[⊥] *Xueshun Chen,*[#,Δ] *Wei Wang,*[○] *Zifa Wang,*[#,§,Δ] *and Song Xi Chen*[†]

[†]Guanghua School of Management and Center for Statistical Science, Peking University, Beijing 100871, China

[‡]CAS-TWAS Center of Excellence for Climate and Environment Sciences, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

[§]University of Chinese Academy of Sciences, Beijing 100049, China

[‖]School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, China

[⊥]Key Laboratory of Environmental Optics and Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei, Anhui 230031, China

[#]State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

[Δ]Center for Excellence in Regional Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, Fujian 361021, China

[○]China National Environmental Monitoring Center, Beijing 100012, China

Corresponding Authors

*Email: wuhuangjian@pku.edu.cn

*Email: weilin@math.pku.edu.cn

14 pages, 5 figures

S1. Descriptions and Configurations of NAQPMS

The CTM adopted in this study is the Nested Air Quality Prediction Modeling System (NAQPMS)[1] developed by the Institute of Atmospheric Physics, Chinese Academy of Sciences, which has been successfully applied to operational air quality forecasting, data assimilation,[2,3] emission inversion,[4] and the study of extreme pollution events.[5,6] The NAQPMS uses a terrain-following coordinate system. The advection and diffusion processes are solved through a mass conservative, peak-preserving, mixing ratio bounded advection scheme developed by Walcek and Aleksic[7] and a diffusion scheme by Byun and Dennis,[8] respectively. The reaction process consists of four modules. The first module is the Carbon Bond Mechanism Z (CBM-Z),[9] which calculates 133 reactions for 53 gas-phase species. The second module is an aerosol thermodynamic model (ISORROPIA v1.7)[10] that simulates inorganic aerosols using an ammonia–sulfate–nitrate–chloride–sodium–water system. The third module simulates secondary organic aerosols from anthropogenic[11] and biogenic[12] precursors. The fourth module simulates 28 heterogeneous reactions for 14 species, including dust, sea salt, sulfate, and black carbon.[13] The deposition process includes dry deposition and wet deposition; the former is implemented through the scheme provided by Wesely,[14] while the latter is calculated by the Regional Acid Deposition Model (RADM).[15] The emission inventory consists of multiple sources. The dust and sea salt emissions were calculated by schemes developed by Luo and Wang[16] and Athanasopoulou et al.[17] Biogenic and biomass burning emissions were estimated from work by Guenther et al.[18] and Cao et al.[19] Anthropogenic emissions were obtained from the MIX inventory[20] for the year 2010.

In this study, meteorological fields for NAQPMS were provided by the Weather Research and Forecasting (WRF) model,[21] with initial and boundary conditions from the National Centers for Environmental Prediction's Global Forecast System (GFS).

Domain configurations for WRF and NAQPMS are shown in Figure 1a. The outer domain (D1) covers East Asia at a 45 km horizontal resolution, and the inner domain (D2) covers central and eastern China at a 15 km horizontal resolution. The vertical space is divided into 20 layers with 8 layers below 2 km.

S2. Descriptions and Configurations of Optimal Interpolation

We applied optimal interpolation (OI)[22] as the data assimilation method to provide the assimilated ICs for our forecasts. Because of its ease of implementation and computational efficiency, OI has been widely used in air quality data analysis.[3,23,24] The method calculates the best linear unbiased estimate of the state vector by

$$\boldsymbol{x}^{\mathrm{a}} = \boldsymbol{x}^{\mathrm{b}} + \boldsymbol{B}\boldsymbol{H}^{\mathrm{T}}\left(\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}^{\mathrm{b}}\right)$$

where $\boldsymbol{x}^{\mathrm{a}}$ and $\boldsymbol{x}^{\mathrm{b}}$ are the analysis and background (forecast) fields, respectively, $\boldsymbol{y}$ is the vector of observations, $\boldsymbol{H}$ is the observation operator, and $\boldsymbol{B}$ and $\boldsymbol{R}$ are the background and observation error covariance matrices, respectively.

In this study, the observations are assumed to be independent, so that the observation error covariance matrix $\boldsymbol{R}$ is diagonal, with the error variances set to 10% of the observed values. The background error covariance matrix $\boldsymbol{B}$ is assumed to be in the Balgovind form,[25] with the covariance between grid points i and j defined by

$$b_{ij} = \left(1 + \frac{d_{ij}}{L}\right) e^{-d_{ij}/L} \sigma_{ij}$$

where d_{ij} is the distance between grid points i and j, L is the characteristic length, and σ_{ij} is the a priori covariance. In accordance with the horizontal resolution of domain (D2) and the distribution of assimilation sites shown in Figure 1, the characteristic length L is set to 80 km. The a prior variance σ_{ij} is the product of the a prior errors at grid points i and j, where the a prior error at each grid point is estimated as 30% of the forecast value at the grid point plus 10% of the average forecast value over the assimilation domain.

S3. Numerical Implementation of Horizontal Advection

By introducing the mixing ratio $Q = e/\rho$, the advection equation for the forecast error e can be written as

$$\frac{\partial(\rho Q)}{\partial t} = -\nabla \cdot (\boldsymbol{v}\rho Q)$$

where ρ is the fluid density and $\boldsymbol{v}$ is the wind vector. We solved this equation by using a mass conservative, peak-preserving, mixing ratio bounded advection scheme,[7] which has been widely adopted for simulating advections in CTMs.[1,26,27] For multi-dimensional flows, the method performs advection calculations by sequentially updating the mixing ratios over all flow dimensions. Applying this method to simulating horizontal advection in our study, the mixing ratio in grid cell (i, j) is approximated by

$$\mathrm{Q}_i^{t+\Delta t} = \left(Q_i^t D_0 - \frac{F_{i+1/2}}{\Delta x_i} + \frac{F_{i-1/2}}{\Delta x_i}\right) \Big/ D_1$$

$$Q_j^{t+\Delta t} = \left(Q_j^t D_1 - \frac{F_{j+1/2}}{\Delta y_j} + \frac{F_{j-1/2}}{\Delta y_j}\right) \Big/ D_2$$

where Δx_i and Δy_j are the grid sizes along the x- and y-directions, respectively, Δt is the time step size, D_k, $k = 0, 1, 2$, are dimensionally dependent fluid densities, and $F_{i \pm 1/2}$ and $F_{j \pm 1/2}$ are the tracer fluxes across the cell faces.

To ensure mass conservation during horizontal advection, the dimensionally dependent fluid densities are specified by

$$D_0 = \rho_{0i}$$

$$D_1 = D_0 - [(\rho_0 v_1)_{i+1/2} - (\rho_0 v_1)_{i-1/2}]\Delta t / \Delta x_i$$

$$D_2 = D_1 - [(\rho_0 v_2)_{j+1/2} - (\rho_0 v_2)_{j-1/2}]\Delta t / \Delta y_j$$

where ρ_{0i} are the initial fluid densities, and $(\rho_0 v_1)_{i \pm 1/2}$ and $(\rho_0 v_2)_{j \pm 1/2}$ are the fluid fluxes across the cell faces. Moreover, the tracer fluxes across the cell faces are defined by

$$F_{i+1/2} = (\rho_0 v_1)_{i+1/2} \Delta t Q_i^{\mathrm{f}}$$

$$F_{j+1/2} = (\rho_0 v_2)_{j+1/2} \Delta t Q_j^{\mathrm{f}}$$

where Q_i^{f} and Q_j^{f} are the outflowing mixing ratios. For details on the calculations of these quantities as well as bounding procedures applied to the mixing ratios, refer to Walcek and Aleksic.[7]

S4. Validation of Assimilated $PM_{2.5}$

We assimilated hourly $PM_{2.5}$ observations into 1-day forecasts for 1003 sites in central and eastern China during January 2018, while reserving observations from the other 323 sites for validation. Spatial distributions of average $PM_{2.5}$ concentrations at the assimilation sites before and after assimilation are shown in SI Figure S2a,b. Before assimilation, $PM_{2.5}$ concentrations over 200 μg m^{-3} are found in the Sichuan Basin, the North China Plain, and the Hubei-Hunan Plain,

while concentrations under 25 μg m^{-3} are found in western and northern China. Compared with observations from the validation sites (circles in SI Figure S2a), significant overestimates are noted in polluted areas, especially in the Sichuan Basin and the Hubei-Hunan Plain. After assimilation, $PM_{2.5}$ concentrations are substantially reduced. Specifically, the average RMSE over all validation sites is reduced by 82%, from 98.9 to 17.5 μg m^{-3}. Accordingly, the bias is reduced from 64.0 to -0.3 μg m^{-3} and the correlation coefficient increases from 0.43 to 0.88.

Comparisons of $PM_{2.5}$ time series at three validation sites in Beijing, Shanghai, and Guangzhou are shown in SI Figure S2c–e. At the Aoti Beijing site, pollution episodes are captured by the forecasts before assimilation. However, the forecasts are significantly higher than the observations, especially during the heavily polluted periods. After assimilation, the RMSE of $PM_{2.5}$ in Beijing is reduced by 93%, from 145.9 to 10.2 μg m^{-3}. Meanwhile, the bias is reduced from 91.8 to 2.8 μg m^{-3} and the correlation coefficient increases from 0.61 to 0.97. Similar improvements by data assimilation are noted at the Jingan Shanghai and Business School Guangzhou sites, with RMSEs reduced by 88% and 82%, biases reduced by 90.0% and 93%, and correlation coefficients increased to 0.98 and 0.96, respectively.

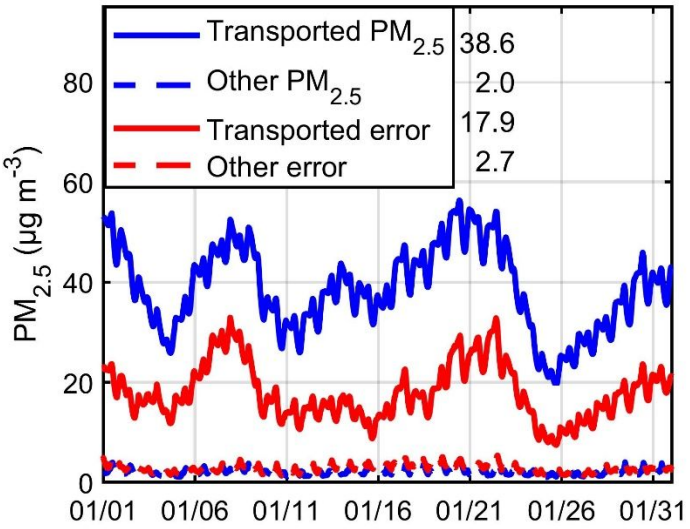


Figure S1. Time series of the transported and other $PM_{2.5}$ and forecast errors, averaged over the assimilation domain. The transported error refers to the part that is transported from an hour ago; the transported $PM_{2.5}$ is similarly defined. Averages during the period are shown to the right of the legend.

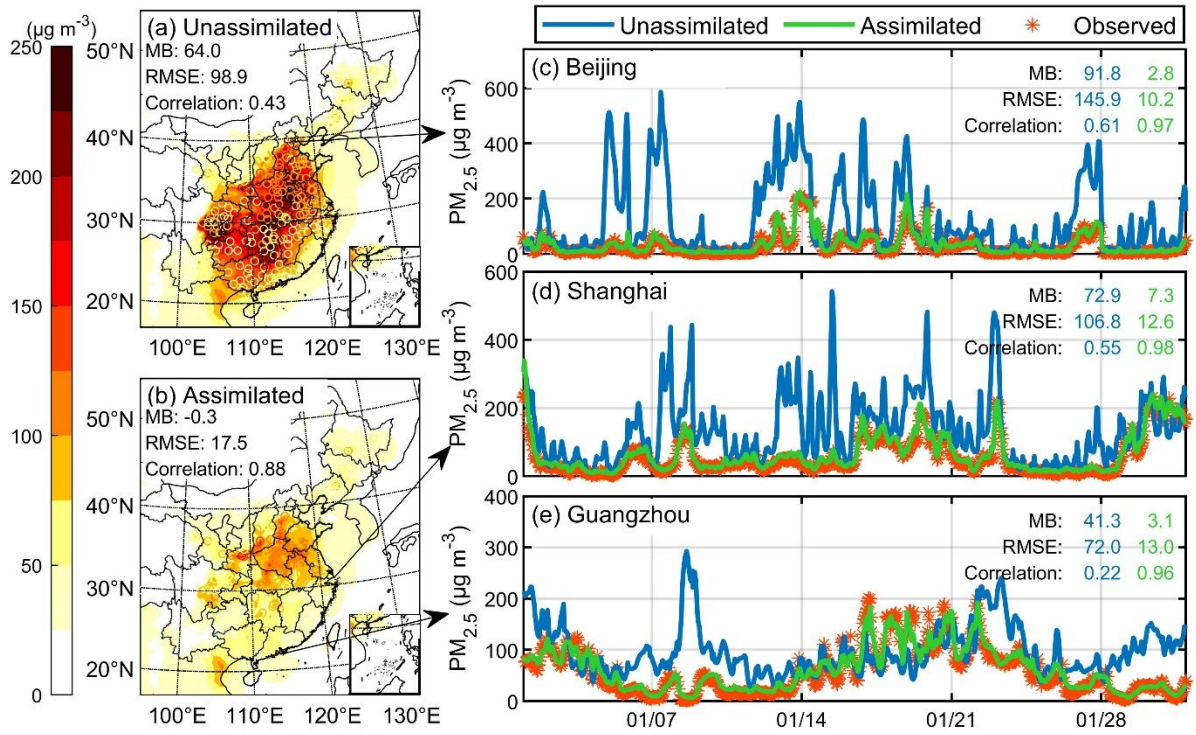


Figure S2. Validation of assimilated $PM_{2.5}$. (a,b) Average concentrations of the unassimilated (a) and assimilated (b) $PM_{2.5}$ over January 2018. Colors of circles indicate average $PM_{2.5}$ observations at individual validation sites. The mean bias (MB), root-mean-square error (RMSE), and correlation coefficient over all validation sites are shown in the upper-left corner of each panel. (c–e) Time series of the unassimilated, assimilated, and observed $PM_{2.5}$ at the Aoti Beijing (c), Jingan Shanghai (d), and Business School Guangzhou (e) sites. The MBs, RMSEs, and correlation coefficients before and after assimilation are shown in the upper-right corner of each panel.

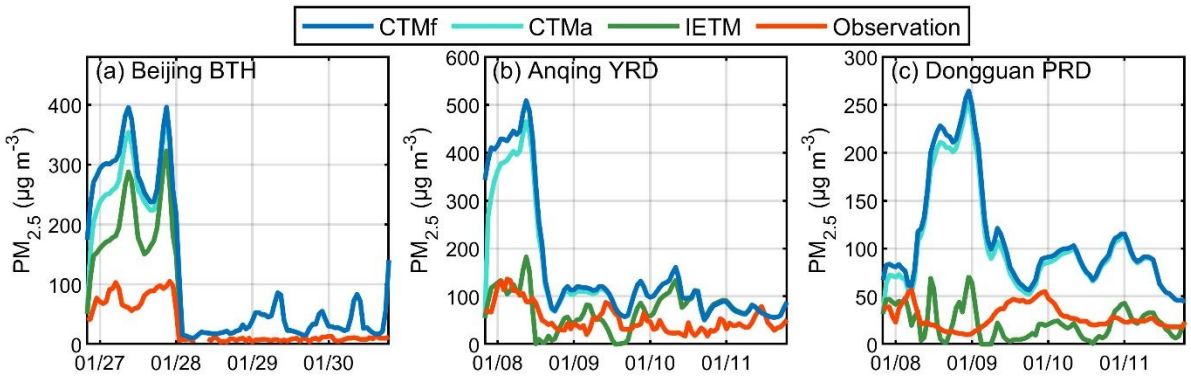


Figure S3. Time series of the CTMf, CTMa, and IETM forecasts at the Beijing (a), Anqing (b), and Dongguan (c) sites, which are located in the Beijing-Tianjin-Hebei (BTH), Yangtze River Delta (YRD), and Pearl River Delta (PRD) regions, respectively. The CTMf and CTMa methods refer to CTM forecasting with unassimilated and assimilated ICs, respectively, and IETM refers to CTMf corrected by the IETM output.

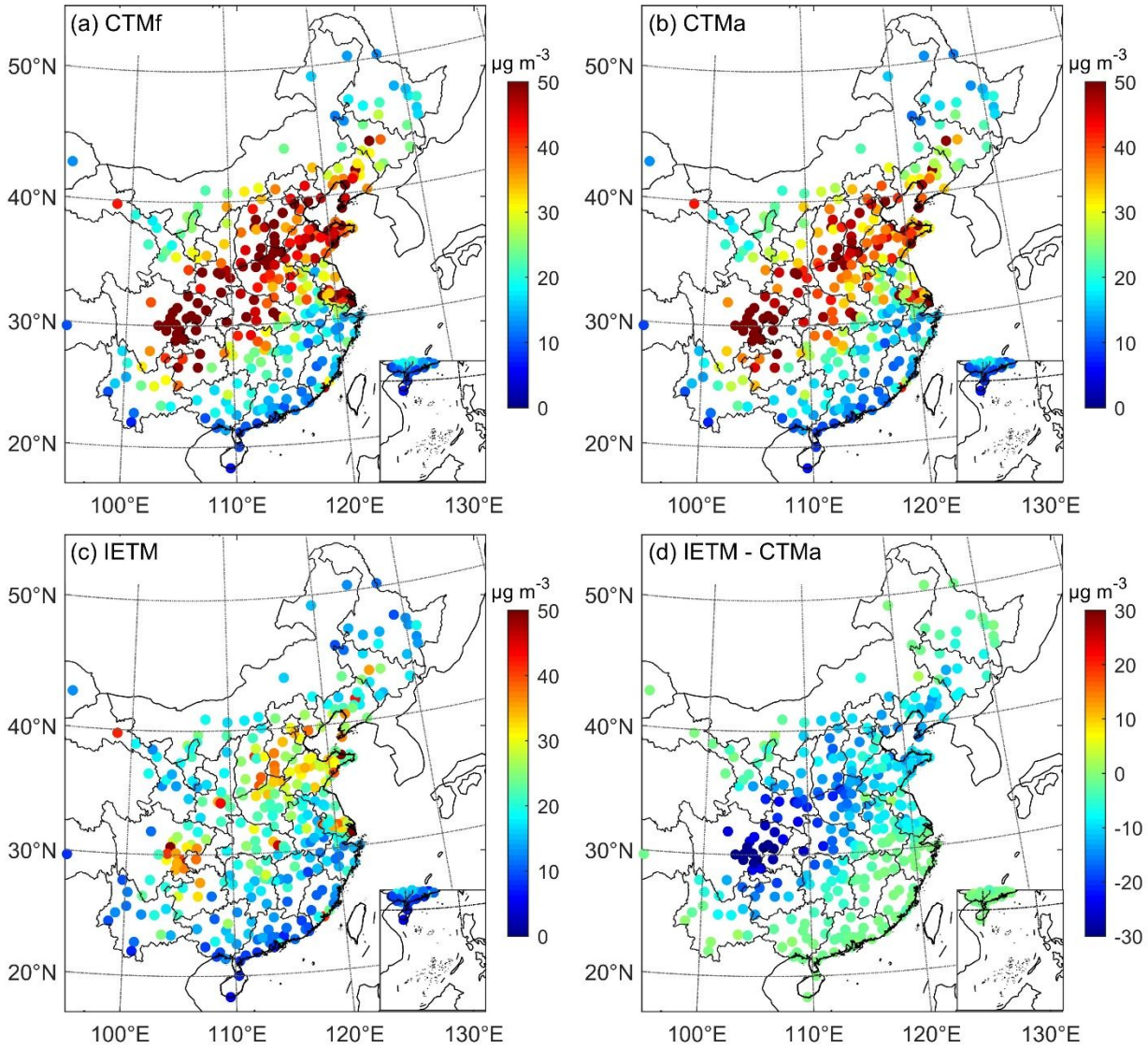
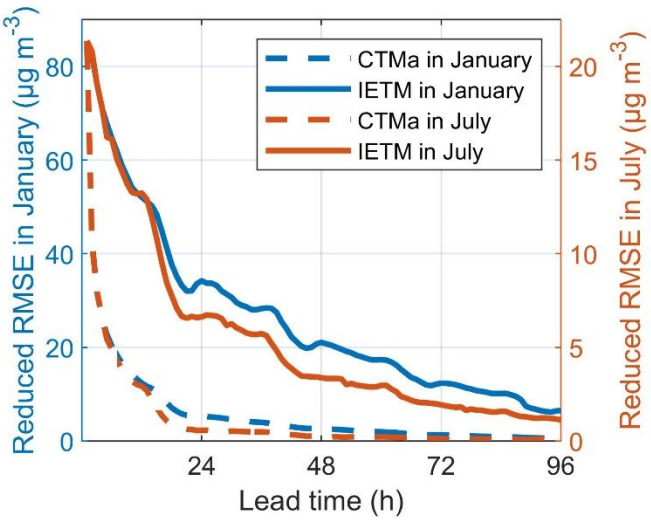


Figure S4. Maps of RMSE at validation sites for 1-day forecasts during July 2017. The RMSEs of the CTMf, CTMa, and IETM methods are shown in (a)–(c), respectively, and differences between the RMSEs of IETM and CTMa are shown in (d).



146

147 **Figure S5.** Reduced RMSEs over all validation sites as functions of lead time for $PM_{2.5}$ forecasts

148 during July 2017 and January 2018.

REFERENCES

(1) Wang, Z.; Maeda, T.; Hayashi, M.; Hsiao, L.-F.; Liu, K.-Y. A Nested Air Quality Prediction Modeling System for Urban and Regional Scales: Application for High-Ozone Episode in Taiwan. *Water Air Soil Pollut.* **2001**, *130* (1-4), 391-396.

(2) Tang, X.; Zhu, J.; Wang, Z. F.; Gbaguidi, A. Improvement of ozone forecast over Beijing based on ensemble Kalman filter with simultaneous adjustment of initial conditions and emissions. *Atmos. Chem. Phys.* **2011**, *11* (24), 12901-12916.

(3) Zheng, H.; Liu, J.; Tang, X.; Wang, Z.; Wu, H.; Yan, P.; Wang, W. Improvement of the Real-time $PM_{2.5}$ Forecast over the Beijing-Tianjin-Hebei Region using an Optimal Interpolation Data Assimilation Method. *Aerosol Air Qual. Res.* **2018**, *18* (5), 1305-1316.

(4) Kong, L.; Tang, X.; Zhu, J.; Wang, Z.; Pan, Y.; Wu, H.; Wu, L.; Wu, Q.; He, Y.; Tian, S.; Xie, Y.; Liu, Z.; Sui, W.; Han, L.; Carmichael, G. Improved Inversion of Monthly Ammonia Emissions in China Based on the Chinese Ammonia Monitoring Network and Ensemble Kalman Filter. *Environ. Sci. Technol.* **2019**, *53* (21), 12529-12538.

(5) Wang, Z.; Li, J.; Wang, X.; Pochanart, P.; Akimoto, H. Modeling of Regional High Ozone Episode Observed at Two Mountain Sites (Mt. Tai and Huang) in East China. *J. Atmos. Chem.* **2006**, *55* (3), 253-272.

(6) Lu, M.; Tang, X.; Wang, Z.; Gbaguidi, A.; Liang, S.; Hu, K.; Wu, L.; Wu, H.; Huang, Z.; Shen, L. Source tagging modeling study of heavy haze episodes under complex regional transport processes over Wuhan megacity, Central China. *Environ. Pollut.* **2017**, *231*, 612-621.

(7) Walcek, C. J.; Aleksic, N. M. A simple but accurate mass conservative, peak-preserving, mixing ratio bounded advection algorithm with Fortran code. *Atmos. Environ.* **1998**, *32* (22), 3863-3880.

(8) Byun, D. W.; Dennis, R. Design artifacts in Eulerian air quality models: Evaluation of the effects of layer thickness and vertical profile correction on surface ozone concentrations. *Atmos. Environ.* **1995**, *29* (1), 105-126.

(9) Zaveri, R. A.; Peters, L. K. A new lumped structure photochemical mechanism for large-scale applications. *J. Geophys. Res.* **1999**, *104* (D23), 30387-30415.

(10) Nenes, A.; Pandis, S. N.; Pilinis, C. ISORROPIA: A New Thermodynamic Equilibrium Model for Multiphase Multicomponent Inorganic Aerosols. *Aquat. Geochem.* **1998**, *4* (1), 123-152.

(11) Odum, J. R.; Jungkamp, T. P. W.; Griffin, R. J.; Flagan, R. C.; Seinfeld, J. H. The Atmospheric Aerosol-Forming Potential of Whole Gasoline Vapor. *Science* **1997**, *276* (5309), 96-99.

(12) Pandis, S. N.; Harley, R. A.; Cass, G. R.; Seinfeld, J. H. Secondary organic aerosol formation and transport. *Atmos. Environ., Part A* **1992**, *26* (13), 2269-2282.

(13) Li, J.; Wang, Z.; Zhuang, G.; Luo, G.; Sun, Y.; Wang, Q. Mixing of Asian mineral dust with anthropogenic pollutants over East Asia: a model case study of a super-duststorm in March 2010. *Atmos. Chem. Phys.* **2012**, *12* (16), 7591-7607.

(14) Wesely, M. L. Parameterization of surface resistances to gaseous dry deposition in regional scale numerical models. *Atmos. Environ.* **1989**, *23* (6), 1293-1304.

(15) Stockwell, W. R.; Middleton, P.; Chang, J. S.; Tang, X. The second generation regional acid deposition model chemical mechanism for regional air quality modeling. *J. Geophys. Res.* **1990**, *95* (D10), 16343-16367.
(16) Luo, G.; Wang, Z.-F. A Global Environmental Atmospheric Transport Model (GEATM): Model Description and Validation (in Chinese). *Chin. J. Atmos. Sci.* **2006**, *30* (3), 504-518.
(17) Athanasopoulou, E.; Tombrou, M.; Pandis, S. N.; Russell, A. G. The role of sea-salt emissions and heterogeneous chemistry in the air quality of polluted coastal areas. *Atmos. Chem. Phys.* **2008**, *8* (19), 5755-5769.
(18) Guenther, A.; Karl, T.; Harley, P.; Wiedinmyer, C.; Palmer, P. I.; Geron, C. Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature). *Atmos. Chem. Phys.* **2006**, *6* (11), 3181-3210.
(19) Cao, G.-L.; Zhang, X.-Y.; Wang, D.; Zheng, F.-C. Inventory of atmospheric pollutants discharged from biomass burning in China continent (in Chinese). *China Environ. Sci.* **2005**, *25* (4), 389-393.
(20) Li, M.; Zhang, Q.; Kurokawa, J.-i.; Woo, J.-H.; He, K.; Lu, Z.; Ohara, T.; Song, Y.; Streets, D. G.; Carmichael, G. R.; Cheng, Y.; Hong, C.; Huo, H. J., Xujia; Kang, S.; Liu, F.; Su, H.; Zheng, B. MIX: a mosaic Asian anthropogenic emission inventory under the international collaboration framework of the MICS-Asia and HTAP. *Atmos. Chem. Phys.* **2017**, *17* (2), 935-963.
(21) Powers, J. G.; Klemp, J. B.; Skamarock, W. C.; Davis, C. A.; Dudhia, J.; Gill, D. O.; Coen, J. L.; Gochis, D. J.; Ahmadov, R.; Peckham, S. E.; Grell, G. A.; Michalakes, J.; Trahan, S.; Benjamin, S. G.; Alexander, C. R.; Dimego, G. J.; Wang, W.; Schwartz, C. S.; Romine, G. S.; Liu, Z.; Snyder, C.; Chen, F.; Barlage, M. J.; Yu, W.; Duda, M. G. The Weather Research and Forecasting Model: Overview, System Efforts, and Future Directions. *Bull. Amer. Meteor. Soc.* **2017**, *98* (8), 1717-1737.
(22) Daley, R., *Atmospheric Data Analysis*. Cambridge University Press: Cambridge, U.K., 1991.
(23) Candiani, G.; Carnevale, C.; Finzi, G.; Pisoni, E.; Volta, M. A comparison of reanalysis techniques: Applying optimal interpolation and Ensemble Kalman Filtering to improve air quality monitoring at mesoscale. *Sci. Total Environ.* **2013**, *458-460*, 7-14.
(24) Lee, E.-H.; Ha, J.-C.; Lee, S.-S.; Chun, Y. PM_{10} data assimilation over south Korea to Asian dust forecasting model with the optimal interpolation method. *Asia-Pac. J. Atmos. Sci.* **2013**, *49* (1), 73-85.
(25) Balgovind, R.; Dalcher, A.; Ghil, M.; Kalnay, E. A Stochastic-Dynamic Model for the Spatial Structure of Forecast Error Statistics. *Mon. Weather Rev.* **1983**, *111* (4), 701-722.
(26) Jacobson, M. Z. GATOR-GCMM: A global- through urban-scale air pollution and weather forecast model 1. Model design and treatment of subgrid soil, vegetation, roads, rooftops, water, sea ice, and snow. *J. Geophys. Res.* **2001**, *106* (D6), 5385-5401.
(27) Kadowaki, M.; Katata, G.; Terada, H.; Nagai, H. Development of the Eulerian atmospheric transport model GEARN-FDM: Validation against the European tracer experiment. *Atmos. Pollut. Res.* **2017**, *8* (2), 394-402.