Supplementary Material of

A Fusion Method Combining Ground-Level Observations with Chemical Transport Model Predictions Using an Ensemble Deep Learning Framework: An Application in China to Estimate Spatiotemporally-Resolved $PM_{2.5}$ Exposure Fields in 2014-2017

Baolei Lyu[†,*], *Yongtao Hu*[§], *Wenxian Zhang*[⊥,*], *Yunsong Du*[∇], *Bin Luo*[∇], *Xiaoling Sun*[‡], *Zhe Sun*[$], *Zhu Deng*[$], *Xiaojiang Wang*[†], *Jun Liu*[†], *Xuesong Wang*[||], *Armistead G. Russell*[§]

[†] Huayun Sounding Meteorological Technology Co. Ltd., Beijing 100081, China

[§] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

[⊥] Hangzhou AiMa Technologies, Hangzhou, Zhejiang 311121, China

[∇] Sichuan Environmental Monitoring Center, Chengdu, Sichuan 610091, China

[‡] Meteorological Bureau of Shenzhen Municipality, ShenZhen, Guangdong 518040, China

[$] Department of Earth System Science, Tsinghua University, Beijing 100084, China

‖ State Key Joint Laboratory of Environmental Simulation and Pollution Control, College of Environmental Sciences and Engineering, Peking University, Beijing 100871, China

*Corresponding authors: Wenxian Zhang (pkuzhangwx@gmail.com); Baolei Lyu (baoleilv@foxmail.com).

Cover Sheet

Items	Count
Pages	16
Figures	13
Texts	1

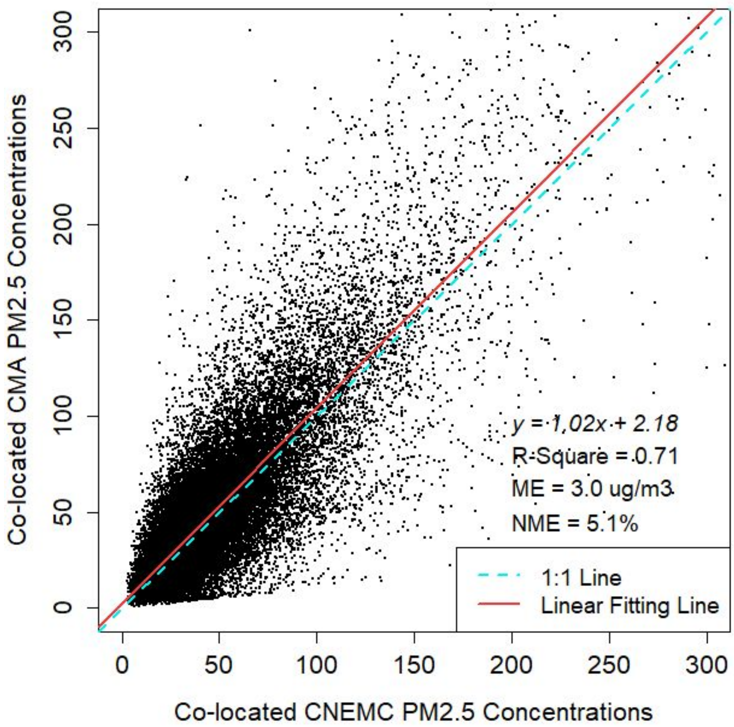
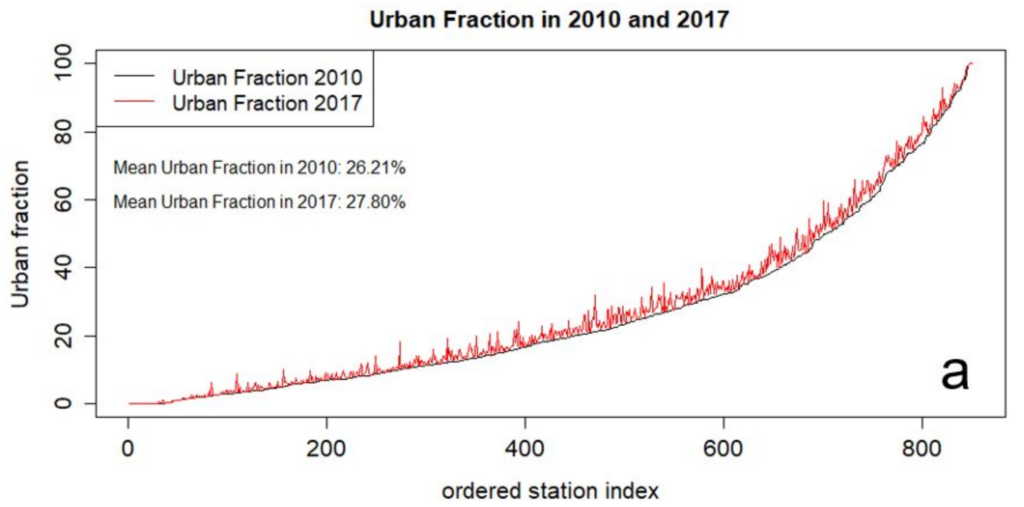


Figure S1 The relationship between the measurements at the CMA and CNEMC monitors that co-located in a same grid cell.



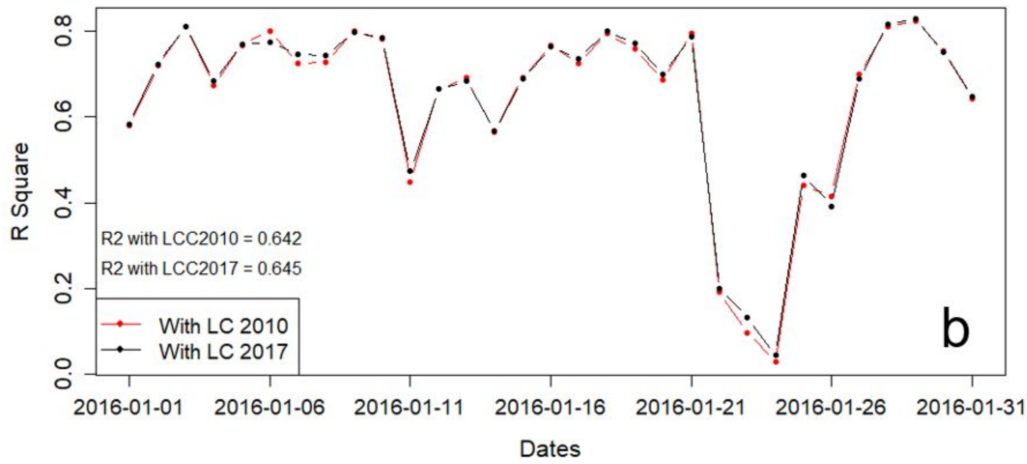


Figure S2 a) The urban fractions from the MODIS land cover data of year 2010 and year 2017. b) The test performances respectively using the MODIS land cover data of year 2010 and year 2017 in the data fusion model.

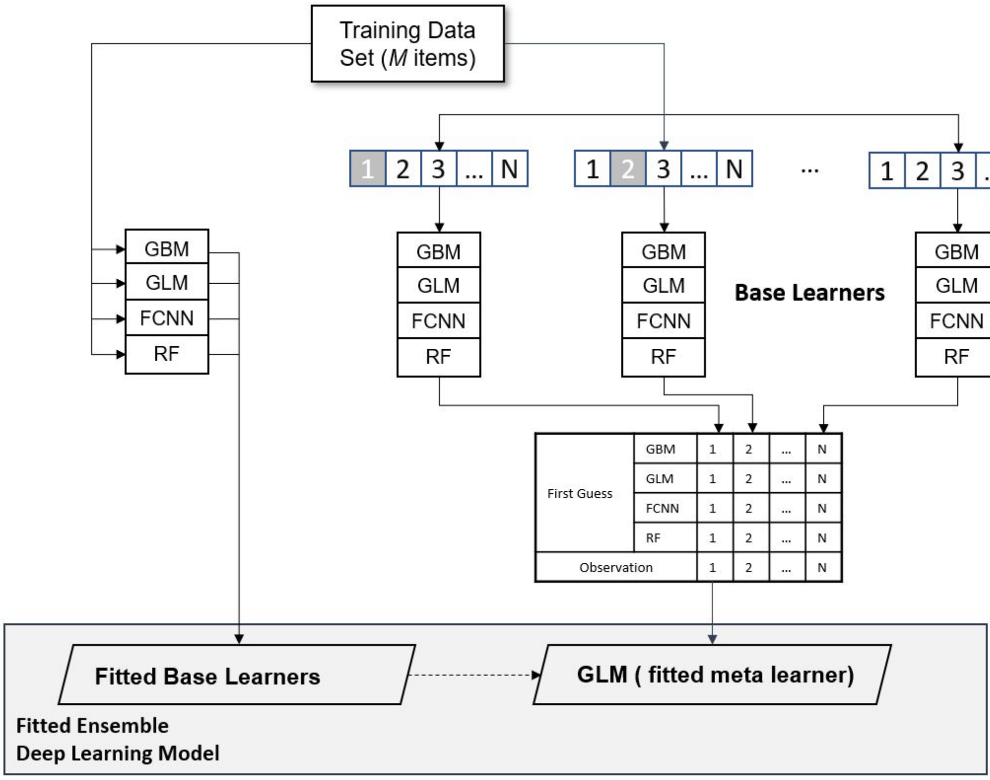


Figure S3 The model fitting procedure of the ensemble deep learning method, where the GBM, GLM, FCNN and RF respectively refer to the models of Gradient Boosting Machine, General Linear Model, Fully Connected Neural Network and Random Forest.

Text S1

The cross-validation technique is used to prevent overfitting and make full use of the training set to fit the machine learning model. For n-fold cross validation, the training data set is equally and randomly divided into n groups and is used to fit the model in an n-time iteration. For each of the n iterations, one group of training data set is held as the validation data set, and the remaining n-1 groups are used as the training set to fit the model. In this study, 10 folds are used and the selection process can be found in Figure S2. After repeating such fitting-prediction procedure for n times, four independent first-guess prediction data sets were generated with the same length as the training data set. The final prediction results were then obtained using the meta-learner as linear combination of the output of the four base learners.

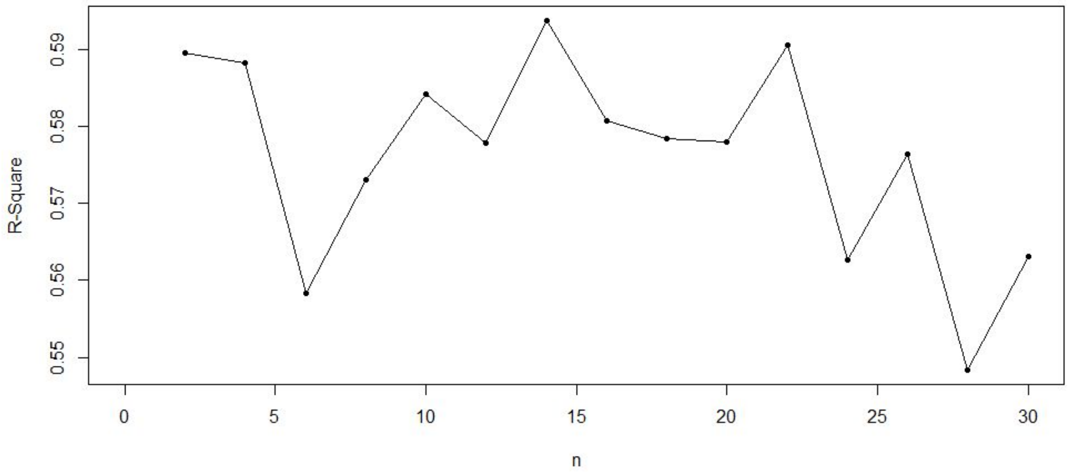


Figure S4 The ensemble deep-learning model performance obtained using different values of *n*, the number of folds in cross-validation. The tests were conducted for January 1, 2016. The performance of other days exhibited similar pattern.

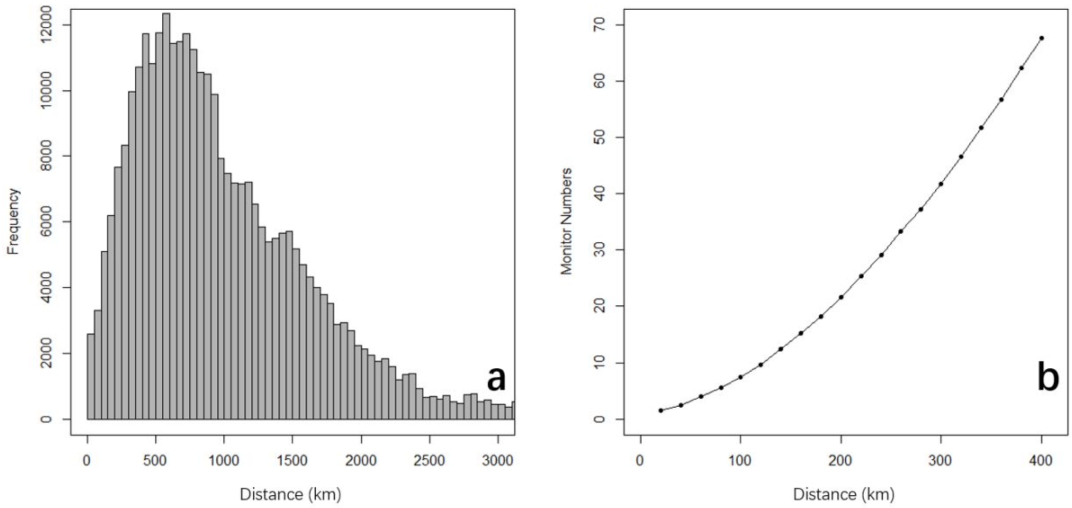


Figure S5 (a) The histogram of distances between monitors; (b) the mean number of monitors with different radius distance.

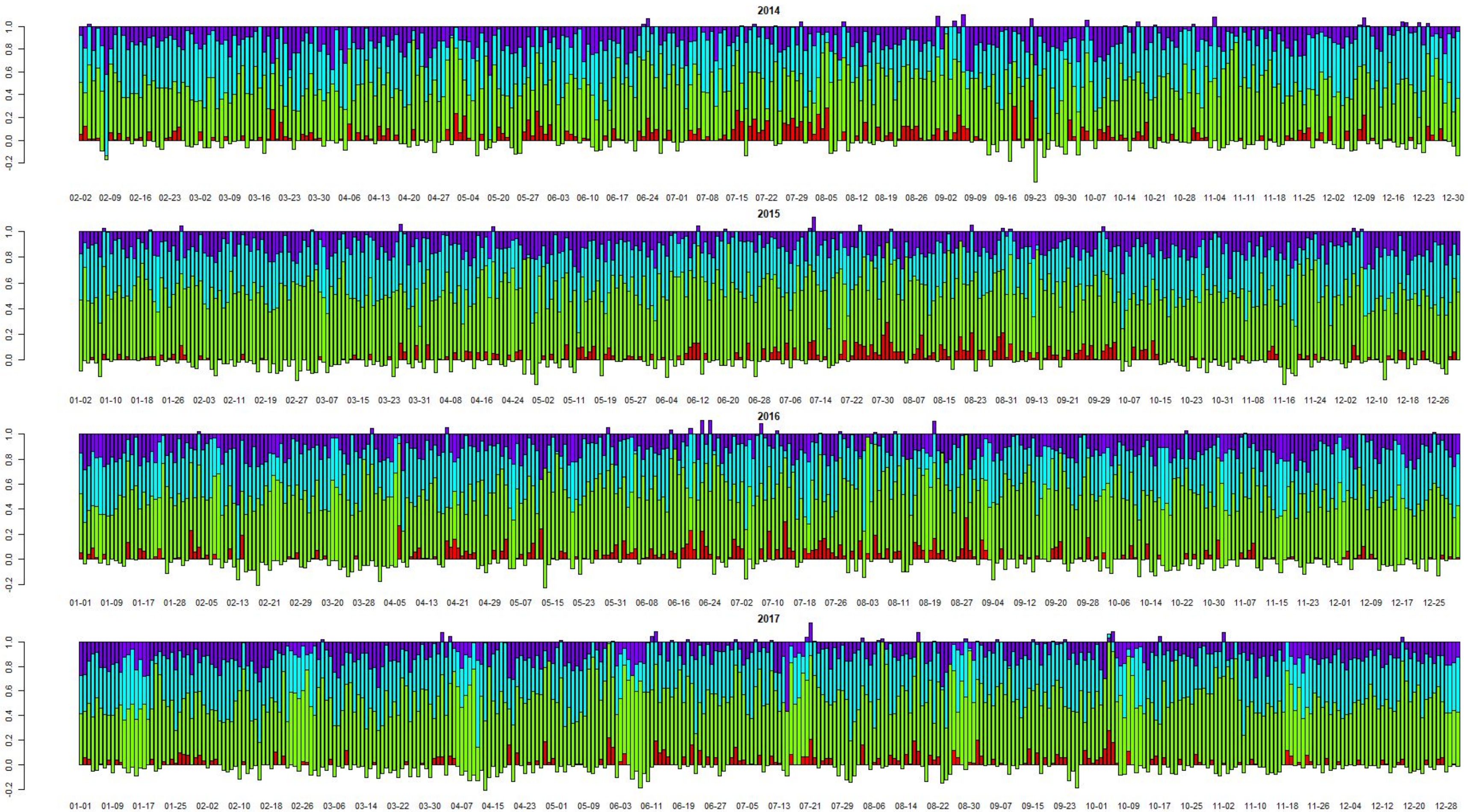


Figure S6 The fitted coefficients of each base learners in the meta-learner. The blue, cyan, green and red bars respectively represent the base learners of FCCN, GBM, RF and GLM.

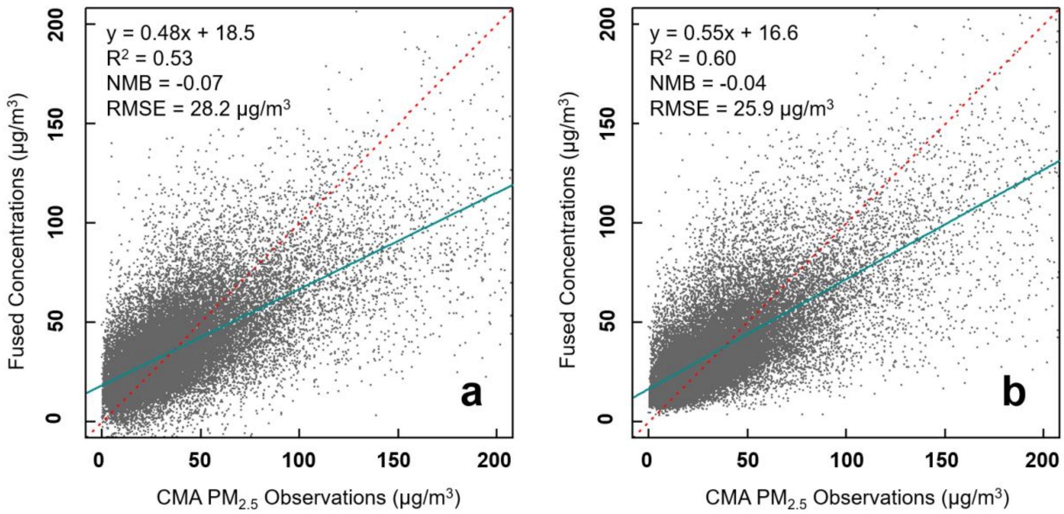


Figure S7 The performance of a) the linear regression model combined with residual kriging, and b) the ensemble deep learning model without residual kriging, evaluated against $PM_{2.5}$ observations in 2016 at the CMA monitors. The green line reflects the linear regression of predictions against observations; the dashed red line is the one-to-one line indicating perfect agreement.

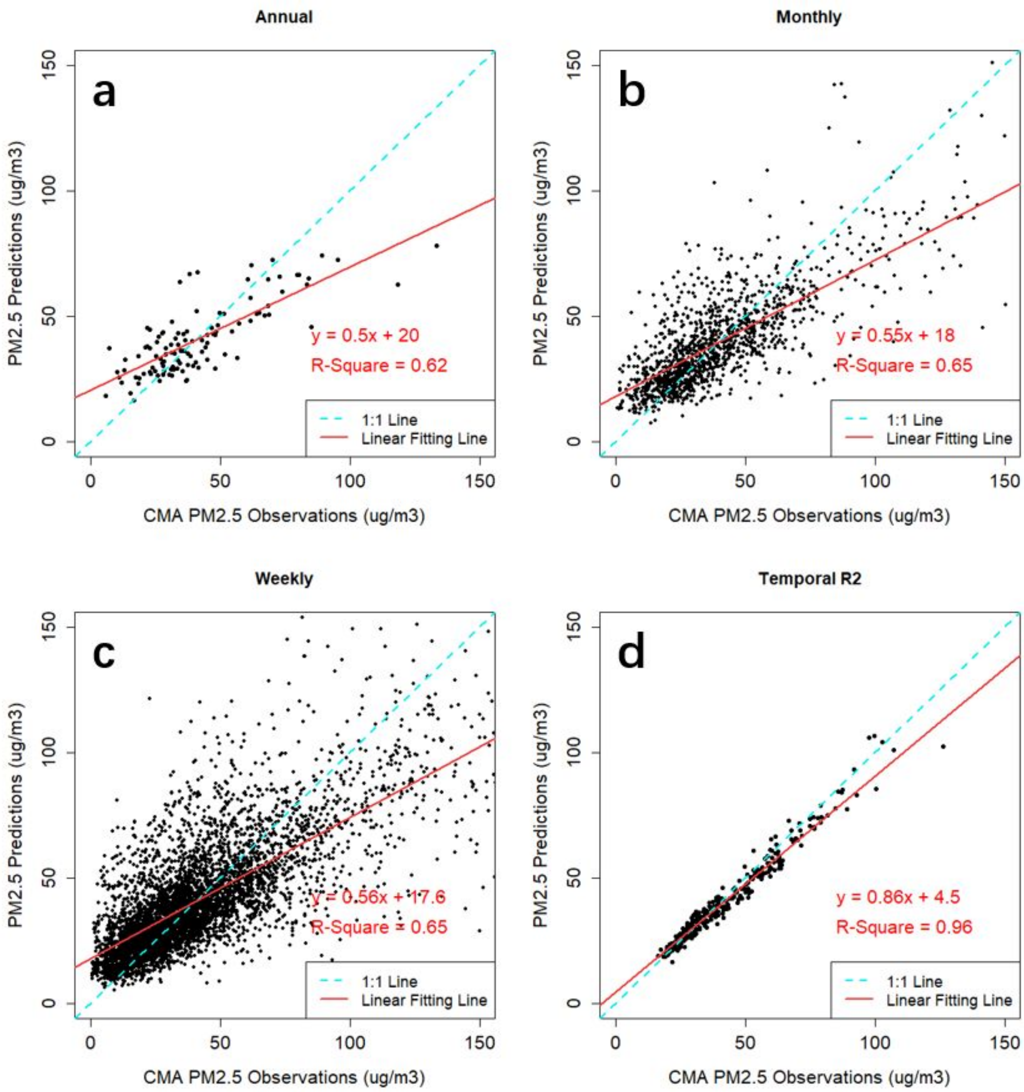


Figure S8 The scatterplot of fused $PM_{2.5}$ concentrations against observed $PM_{2.5}$ concentrations in 2016 evaluated at the independent CMA monitors. The dashed red line reflects the linear regression of the fused data against the observations; the black line is the one-to-one line indicating the perfect agreement. The panels a), b) and c) refer to the evaluation performance, respectively, at the annual, monthly and weekly scales. The panel d) is the temporal R^2 by daily averages with all the data at all evaluated CMA monitors.

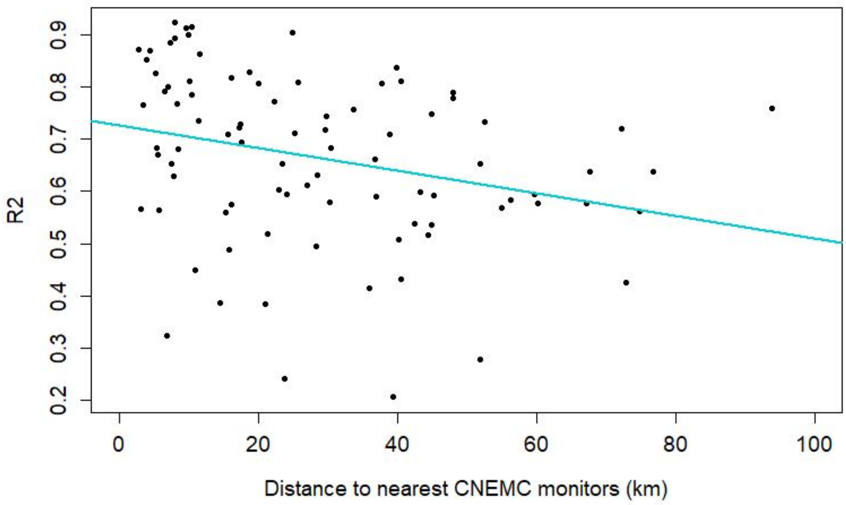


Figure S9 The relationship between R^2 and the distance between a CMA monitor and its nearest neighboring CNEMC monitor.

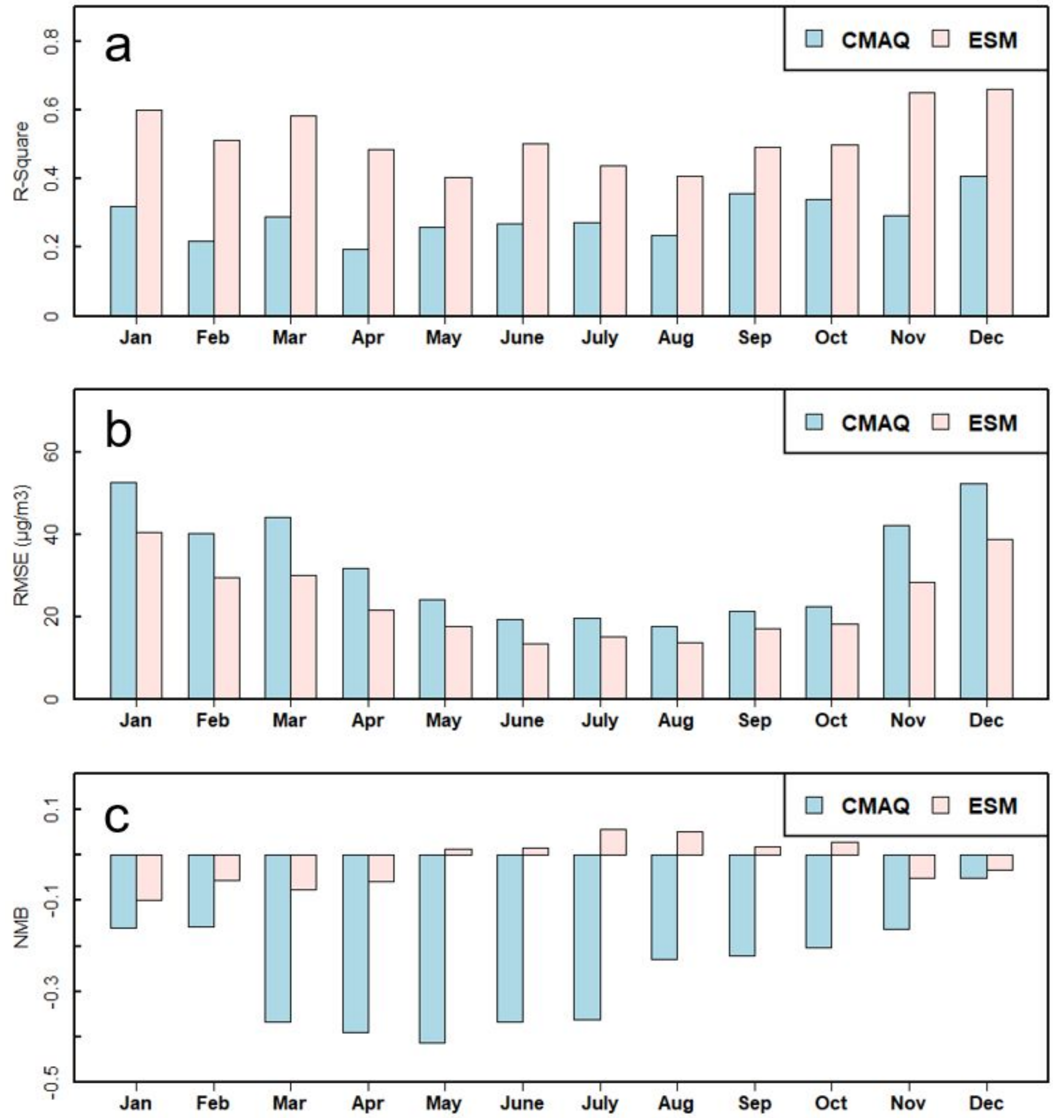


Figure S10 Monthly performance of the raw CMAQ simulated and fused $PM_{2.5}$, evaluated with a) R^2, b) RMSE and c) NMB values, against the CMA observations at independent monitors in 2016.

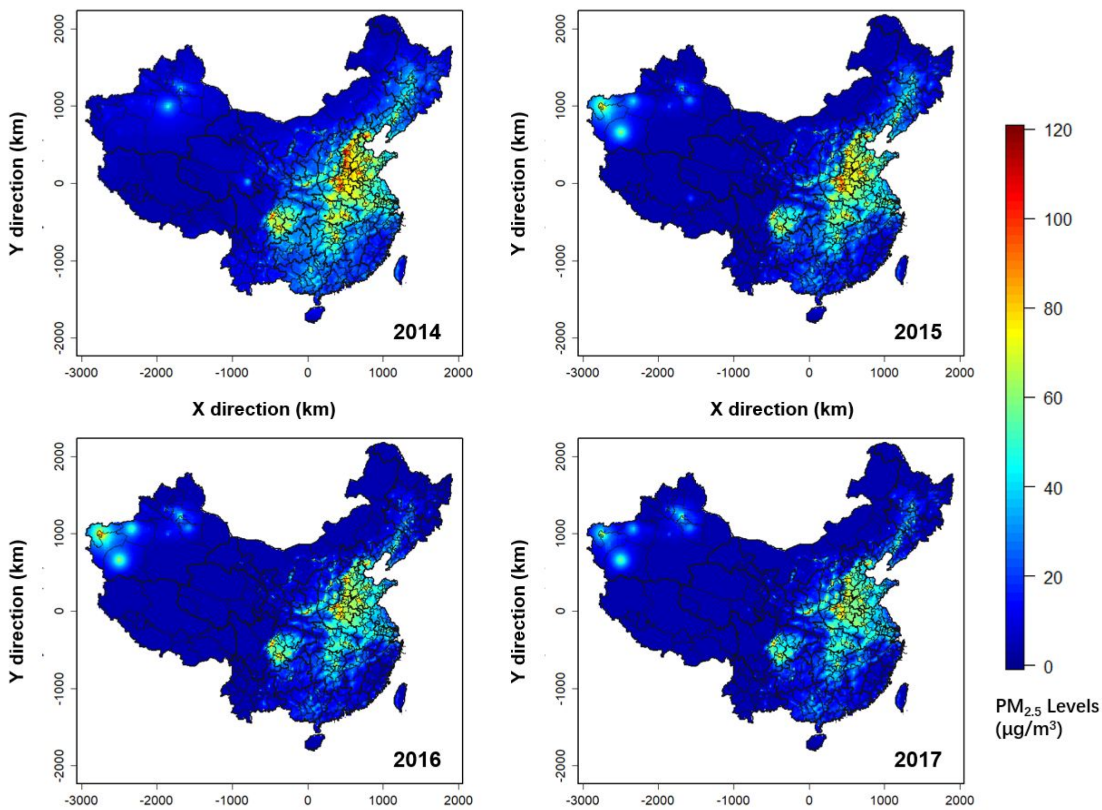


Figure S11 The annual mean of fused $PM_{2.5}$ concentrations in China from 2014 to 2017.

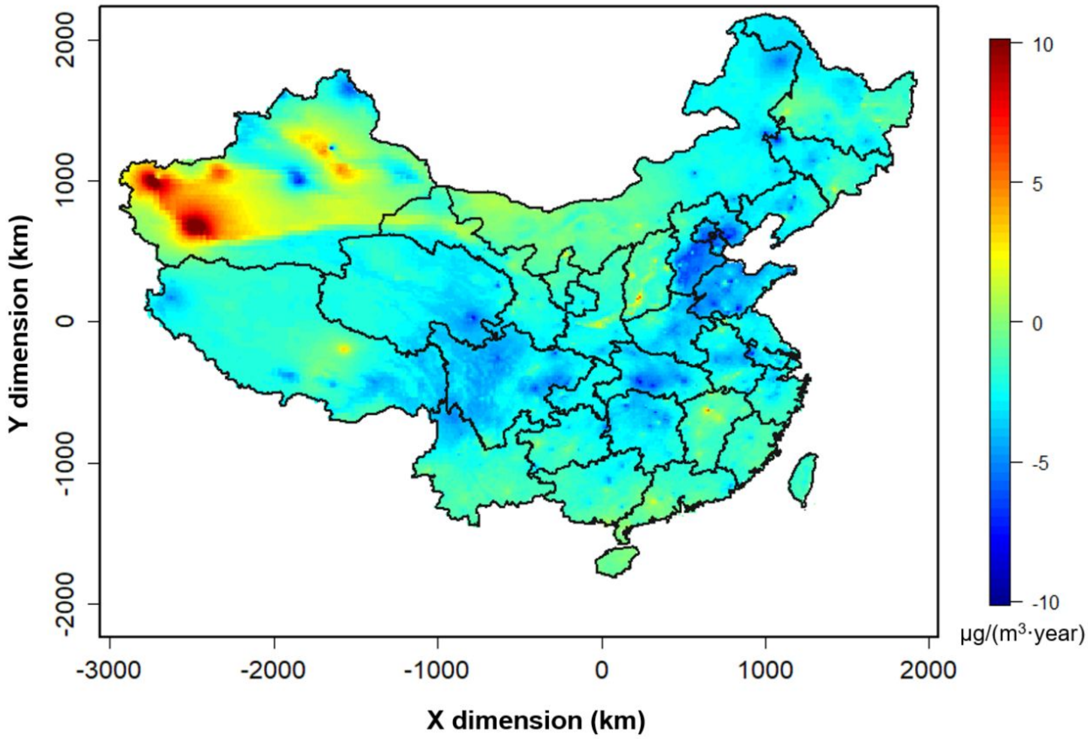


Figure S12. The annual $PM_{2.5}$ concentration changes in China from 2014 to 2017, obtained by averaging the changes in annual mean $PM_{2.5}$ levels between the four years.

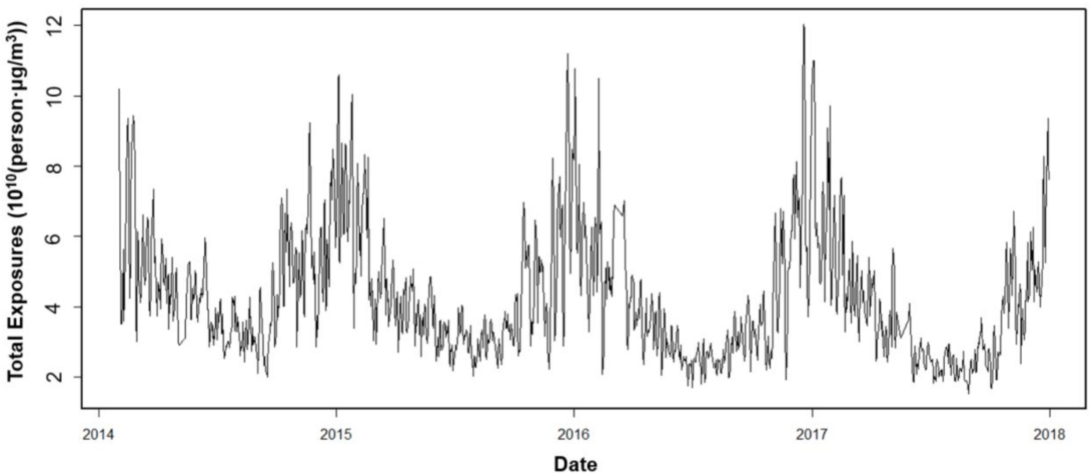


Figure S13 The daily accumulated exposures estimated using the fused $PM_{2.5}$ concentration fields in China.