Supplement of Atmos. Chem. Phys., 22, 10551–10566, 2022
https://doi.org/10.5194/acp-22-10551-2022-supplement

Atmospheric Chemistry and Physics
Open Access

Supplement of

Statistical and machine learning methods for evaluating trends in air quality under changing meteorological conditions

Minghao Qiu et al.

Correspondence to: Minghao Qiu (mhqiu@stanford.edu)

Supplementary methods

Implementation of LASSO and RF

As the incorporation of both local and regional features can quickly expand the dimensionality of the feature space, we use the Least Absolute Shrinkage and Selection Operator (LASSO) and the Random Forest (RF) model to assess the importance of regional meteorological features. Both methods are commonly-used approaches with good prediction performance with high dimensional data inputs, and are thus appropriate for analysis with a large number of regional meteorological features. For these two methods, we rewrite equation 1 as the following:

$$y_{it} = \beta_i^{obs} \times t + g_i(X_{it}, Z_t, W_t) + \epsilon_{it} \tag{1}$$

where $g_i()$ denotes the functional form fitted by LASSO or RF. X_{it} again denotes the local meteorological features for grid cell i on day t. Z_t denotes the regional scale meteorology features including those for all grid cells in the US on day t (98 cells in 4×5 degrees; we choose a relatively coarse resolution due to computational cost). Meteorological information in each location in the US may help explain the pollutant concentrations in grid cell i. In total, we have 10 local features (X_{it}) and 10×98=980 regional scale features (Z_t). W_t denotes the day and month variable to model the daily and monthly variability in pollutant concentrations that are unrelated to meteorological variability. For LASSO, we use month-of-year×day-of-month fixed effects (same as all the other methods except for RF), and these fixed effects are not penalized in the LASSO regression. For RF, we use the month-of-year variable (from 1 to 12), and day-of-month variable (from 1 to 31), due to the inefficient performance of RF working with a large number of fixed effects. Thus, the difference between RF and the other methods may also come from the different choices in modeling monthly and daily variability.

The coefficient β_i^{obs} is obtained with the following procedure using the double machine learning approach by Chernozhukov et al. (2018):

(1) We first partition the time series of $\{y_{it}, X_{it}, Z_t, W_t\}$ into 4 folds. We use 75% of the data as training data and the remaining 25% for predictions. We train the following two models on the training data:

$$y_{it} = f(X_{it}, Z_t, W_t)$$

$$t = g(X_{it}, Z_t, W_t)$$

(2) We then apply models $f(.)$ and $g(.)$ to the prediction set to get predictions of y_{it} and t for the rest 25% of the data. The above process is repeated four times to derive predictions for the entire time series (predictions denoted as $\widehat{y_{it}}$ and $\widehat{t}$).

(3) We calculate the residuals of each model $\widetilde{y_{it}} = y_{it} - \widehat{y_{it}}$ and $\widetilde{t} = t - \widehat{t}$. The coefficient of interest β_i^{obs} is then calculated as:

$$\beta_i^{obs} = \frac{\sum_t \widetilde{t} \times \widetilde{y_{it}}}{\sum_t \widetilde{t} \times t}$$

This is equivalent to setting up a linear regression of $\widetilde{y_{it}} \sim \widetilde{t}$ and obtaining the slope coefficients (as shown by Chernozhukov et al. (2018)).

The hyper-parameters of RF and LASSO are tuned with 4-fold cross-validation. We also perform two sensitivity analyses: 1) with a different spatial resolution for the regional scale features (2×2.5 degrees instead of 4×5 degrees), and 2) with different numbers of folds to estimate the trend coefficients. Our results are similar across these sensitivity analyses (see figure S16).

The double machine learning framework involves a sample partition procedure (steps (1) and (2) above). This procedure, however, does not fit the purpose of including time fixed effects in the LASSO model (as randomly partitioned training and test sets could have very a unbalanced number of observations from a given month-day pair). Therefore, steps (1) and (2) are only implemented for the RF model, and coefficients of the LASSO model are directly derived from step (3) without sample splitting. This is acceptable for the LASSO model as the risk of "overfitting" has already been eliminated by using the tuned penalizing factor (i.e. the hyper-parameters) derived from 4-fold cross-validation. It is important to note that we quantify the performance of RF and other methods using the differences between "meteorology-corrected" trends (β^{obs}) and the counterfactual trends (β^{count}), instead of their performance in predicting the pollutant concentration. Therefore, if the RF model "overfits the data", it would actually result in a large error, because the overly fit RF model would attribute all variability of $PM_{2.5}$ and O_3 to the meteorological variables and estimate a close-to-zero trend.

SI tables and figures

Model	Annual $PM_{2.5}$ in the US			Summer O_3 in the US		
	average error	median relative error	cells with relative error >50%	average error	median relative error	cells with relative error >50%
No correction	0.066	28%	27%	0.67	154%	84%
MLR (5 features)	0.092	43%	44%	0.38	84%	71%
MLR (10 features)	0.083	40%	40%	0.33	71%	64%
Quadratic	0.088	40%	42%	0.29	60%	58%
Cubic	0.075	39%	41%	0.28	60%	58%
Spline	0.076	40%	41%	0.28	61%	59%
GAM	0.076	40%	43%	0.29	61%	58%
RF-local	0.067	33%	39%	0.34	78%	70%
LASSO-regional	0.078	31%	33%	0.31	68%	65%
RF-regional	0.047	25%	23%	0.19	46%	47%

Table S1. Estimation errors of trend estimates in the US under different correction methods. The average estimation errors, median relative error, and fraction of grid cells with relative error greater than 50% are shown in the table. Relative errors are calculated as the ratio of estimation error to the trend estimate in the counterfactual scenario. MLR (5 features) only use temperature, precipitation, humidity, and surface wind speed (U,V directions) as the meteorological features.

Model	Annual $PM_{2.5}$ in China			Summer O_3 in China		
	average error	median relative error	cells with relative error >50%	average error	median relative error	cells with relative error >50%
No correction	0.89	224%	77%	0.43	95%	74%
MLR (5 features)	1.07	193%	80%	0.42	90%	68%
MLR (10 features)	0.90	159%	79%	0.41	85%	68%
Quadratic	1.00	142%	82%	0.36	76%	62%
Cubic	1.07	143%	82%	0.34	68%	59%
Spline	1.08	140%	84%	0.33	69%	59%
GAM	1.06	139%	82%	0.35	72%	59%
RF-local	0.99	172%	82%	0.31	64%	58%
LASSO-regional	0.83	184%	75%	0.46	98%	73%
RF-regional	0.64	152%	67%	0.28	61%	58%

Table S2. Estimation errors of trend estimates in China under different correction methods. The average estimation errors, median relative error, and fraction of grid cells with relative error greater than 50% are shown in the table. Relative errors are calculated as the ratio of estimation error to the trend estimate in the counterfactual scenario. MLR (5 features) only use temperature, precipitation, humidity, and surface wind speed (U,V directions) as the meteorological features.

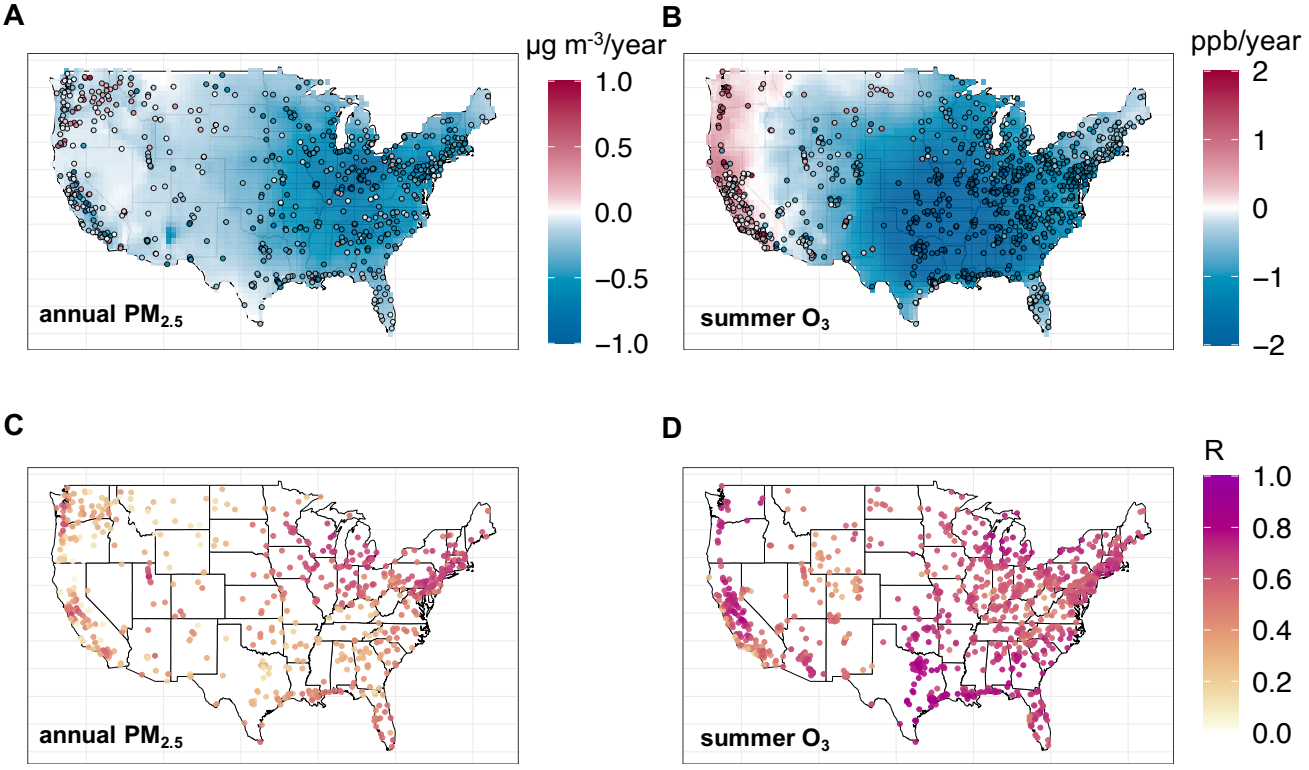

Figure S1. Comparison between the annual $PM_{2.5}$ (Panels A and C) and summer O_3 (Panels B and D) concentrations measured by the monitoring network and GEOS-Chem simulations in the US (2011-2017). Panels A and B show the trends in monitored concentrations (dots) and trends in the observational scenarios in GEOS-Chem simulations (background) without meteorology corrections. Panels C and D show the Pearson correlation coefficient (R) between the daily measured concentrations and simulated concentrations. Observational air quality data is derived from U.S. Environmental Protection Agency (2021b).

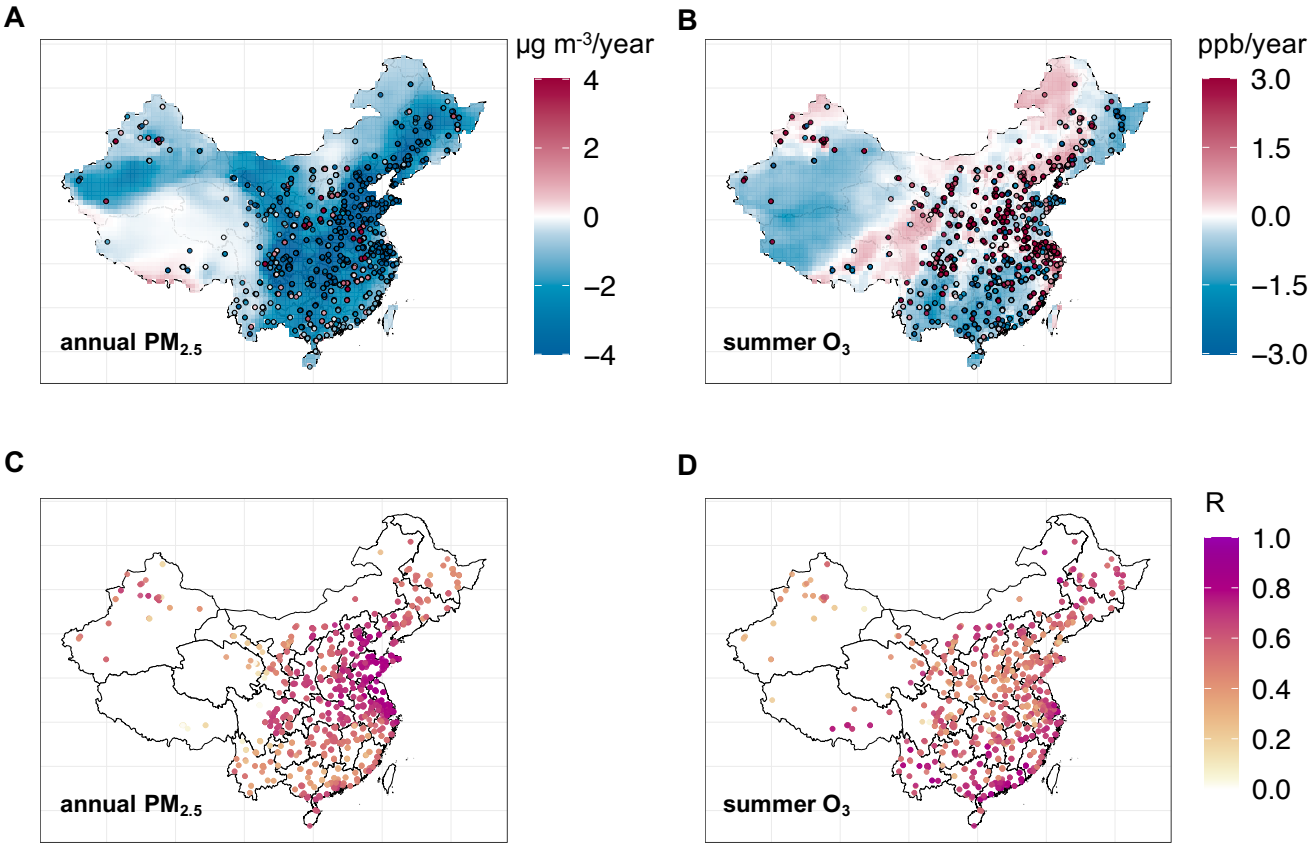

Figure S2. Comparison between the annual $PM_{2.5}$ (Panels A and C) and summer O_3 (Panels B and D) concentrations measured by the surface monitoring network and GEOS-Chem simulations in China (2014-2017). Panels A and B show the trends in monitored concentrations (dots) and trends in the observational scenarios in GEOS-Chem simulations (background) without meteorology corrections. Panels C and D show the Pearson correlation coefficient (R) between the daily measured concentrations and simulated concentrations. Observational air quality data is derived from China's Ministry of Ecology and Environment (2021).

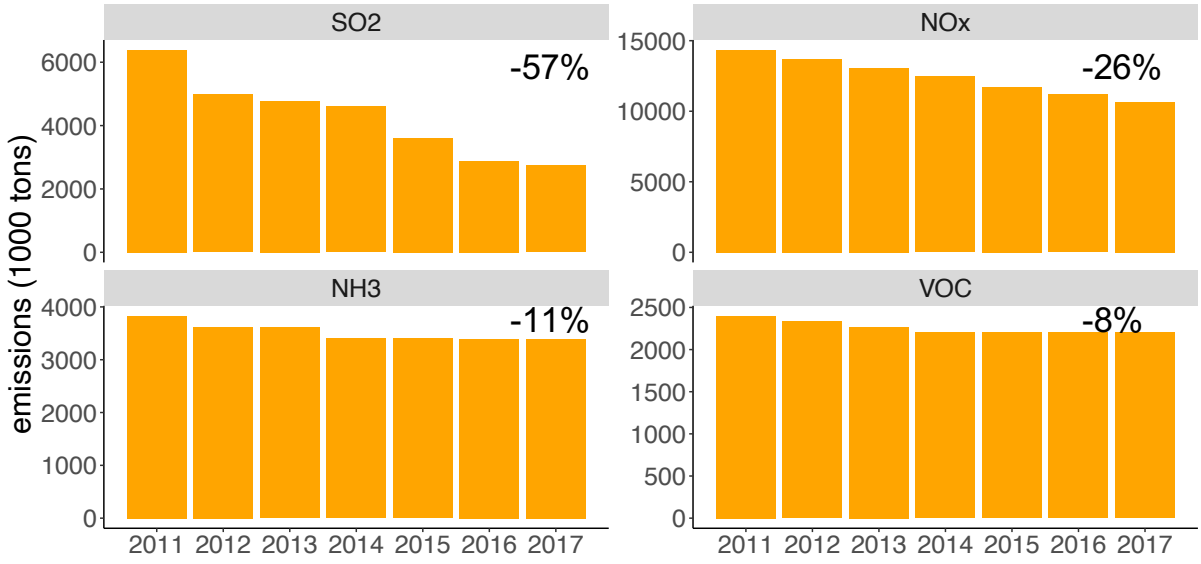

Figure S3. National total anthropogenic emissions in the US (2011- 2017). The emissions data is derived from the national total emissions of criteria air pollutants reported by the US EPA Air Emissions Inventory (U.S. Environmental Protection Agency, 2021a). Changes in emissions between 2011 and 2017 as percentages of the emissions in 2011 are presented in the figure.

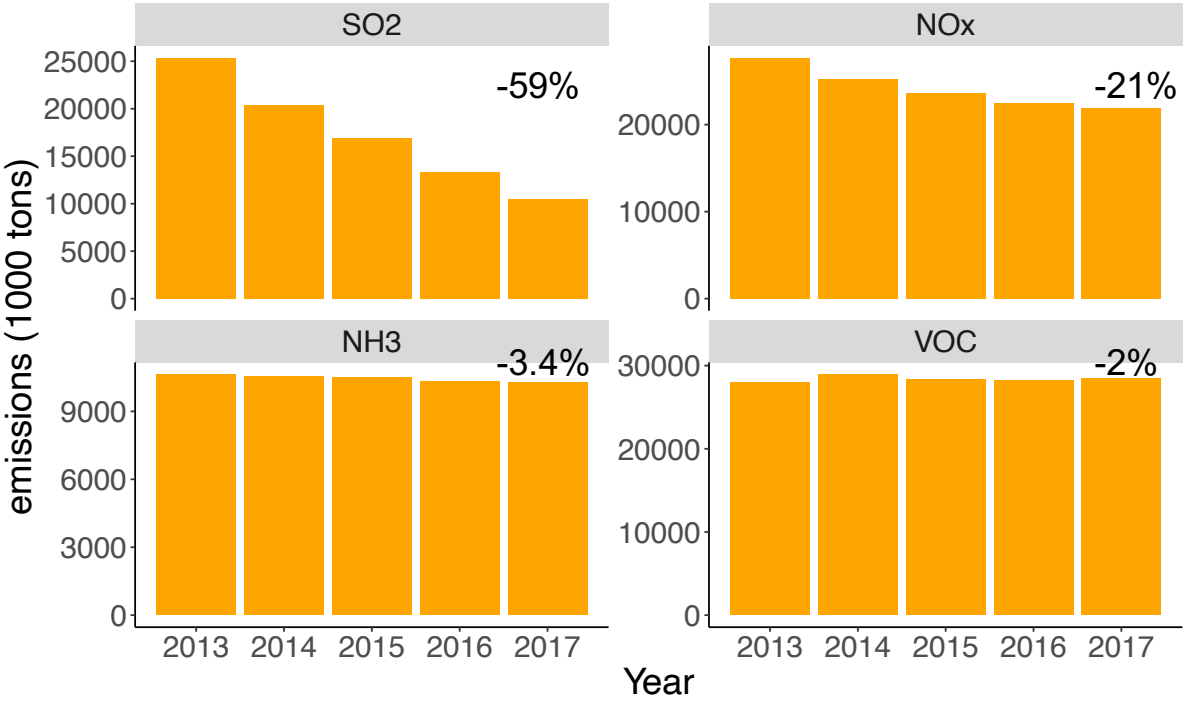

Figure S4. National total anthropogenic emissions in China (2013- 2017). The emissions data is derived from the Multi-resolution Emission Inventory (MEIC) (Li et al., 2017). Changes in emissions between 2013 and 2017 as percentages of the emissions in 2013 are presented in the figure.

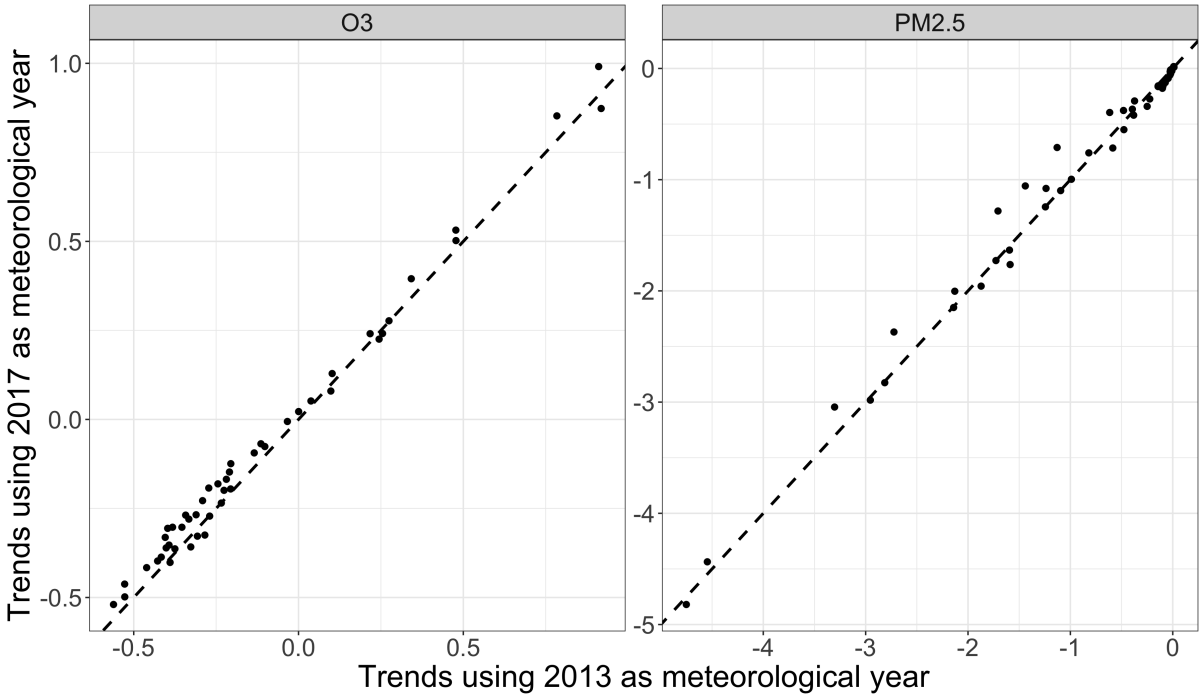

Figure S5. Counterfactual trends in O_3 (unit: ppb/yr) and $PM_{2.5}$ (unit: μg m^{-3}/year) in China with different meteorological years. Each dot represents one grid cell in China. The x-axis shows the trends in air quality in the counterfactual scenario using the meteorological field in 2013, and the y-axis shows the trends in air quality in the counterfactual scenario using the meteorological field in 2017. Results here are derived from simulation at 4×5 degrees.

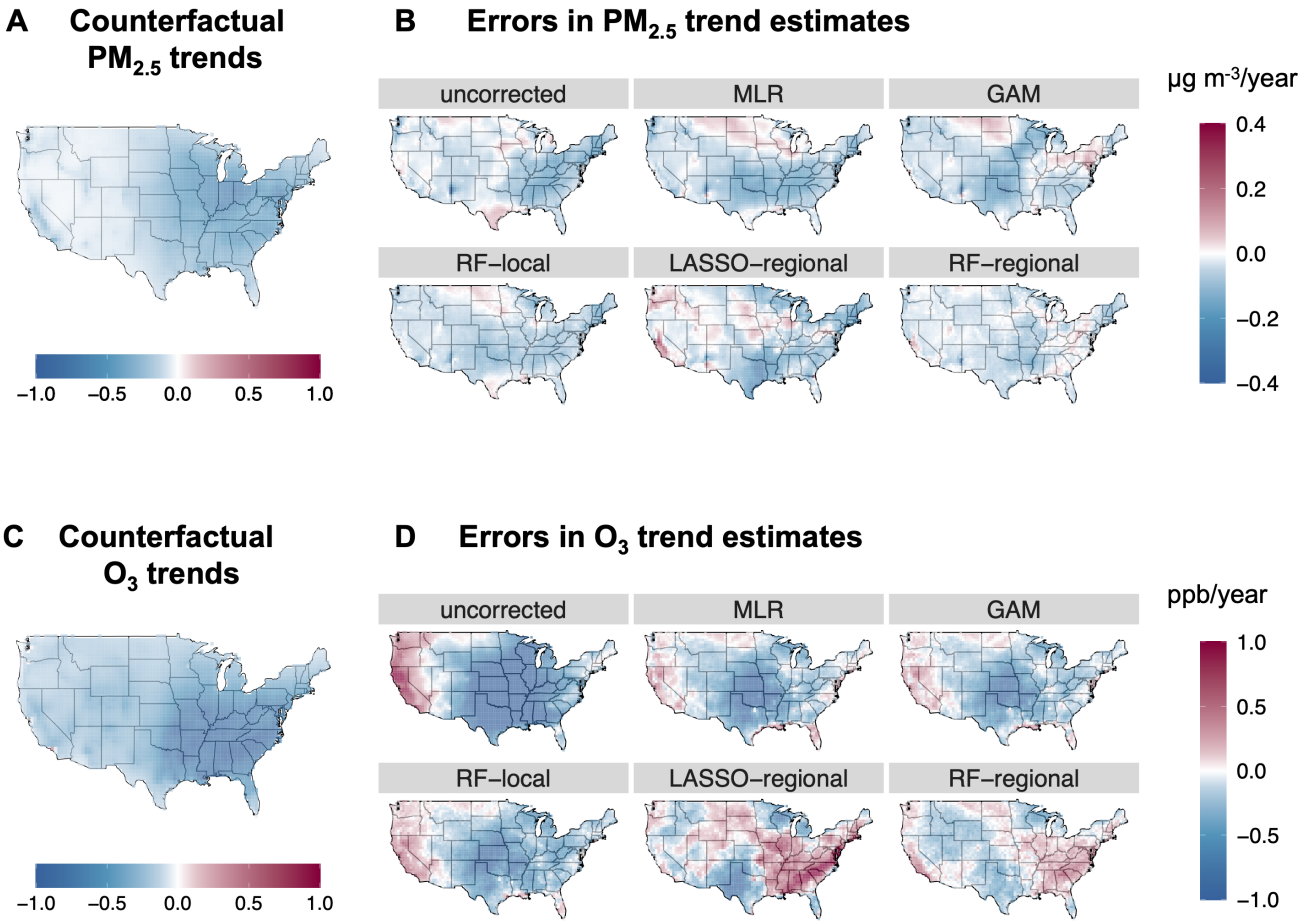

Figure S6. Trend estimates of daily annual $PM_{2.5}$ (Panels A and B) and summer O_3 (C and D) in the US. Panels A and C show trend estimates under the counterfactual scenario (β^{count}). Panels B and D show the estimation errors of trend estimates under different correction methods compared with the counterfactual scenarios ($\beta^{obs} - \beta^{count}$). The average of the absolute error for each method is shown in the figure. Unit of trend estimate is μg m^{-3}/year for $PM_{2.5}$ or ppb/year for O_3.

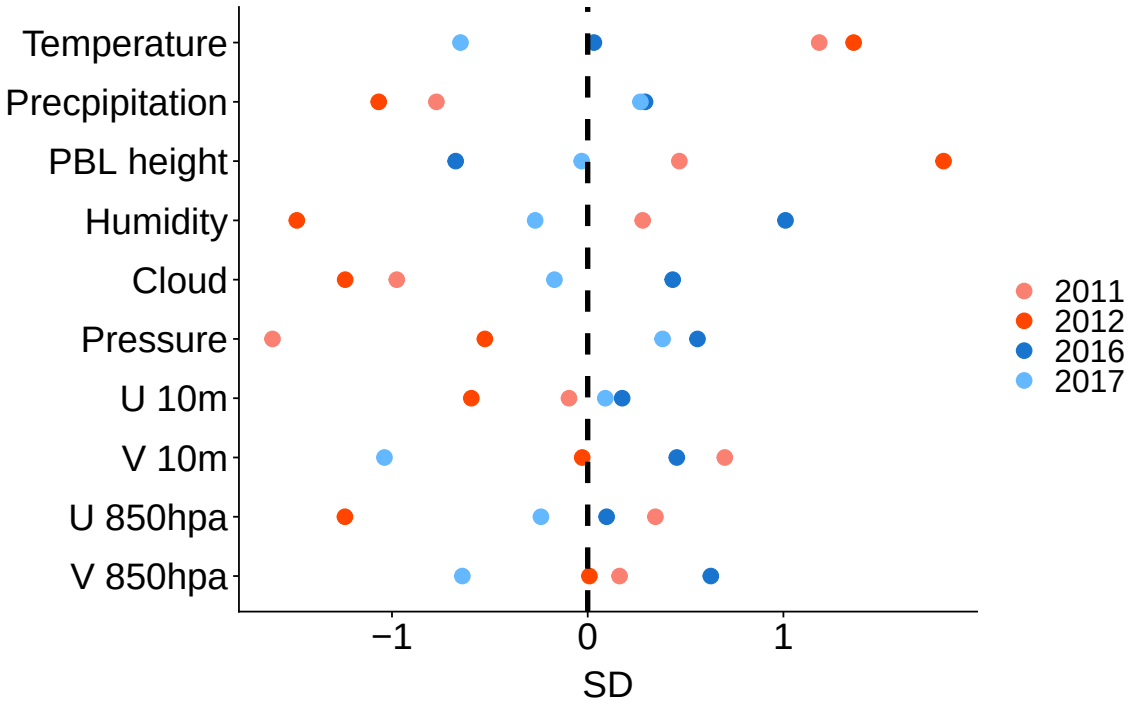

Figure S7. Deviations of meteorological features from the 7-year average in the US (South and Midwest). The deviation is quantified in the units of standard deviation (SD) across the 7-year period. Zero indicates the 7-year average. This plot shows the summer time average of daily MDA8 meteorological variables for each year aggregated over South and Midwest US.

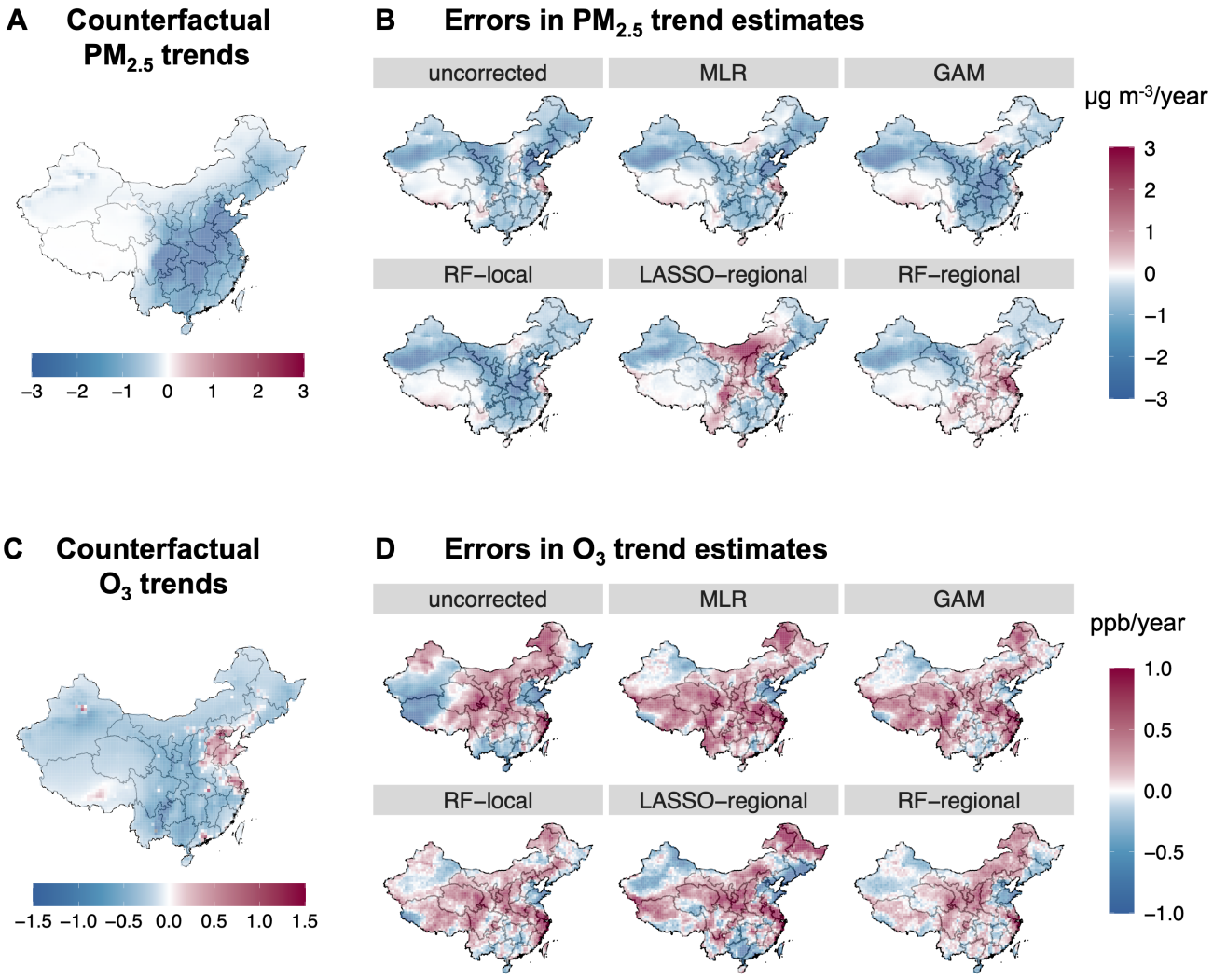

Figure S8. Trend estimates of daily annual $PM_{2.5}$ (Panels A and B) and summer O_3 (C and D) in China. Panels A and C show trend estimates under the counterfactual scenario (β^{count}). Panels B and D show the estimation errors of trend estimates under different correction methods compared with the counterfactual scenarios ($\beta^{obs} - \beta^{count}$). The average of the absolute error for each method is shown in the figure. Unit of trend estimate is μg m^{-3}/year for $PM_{2.5}$ or ppb/year for O_3.

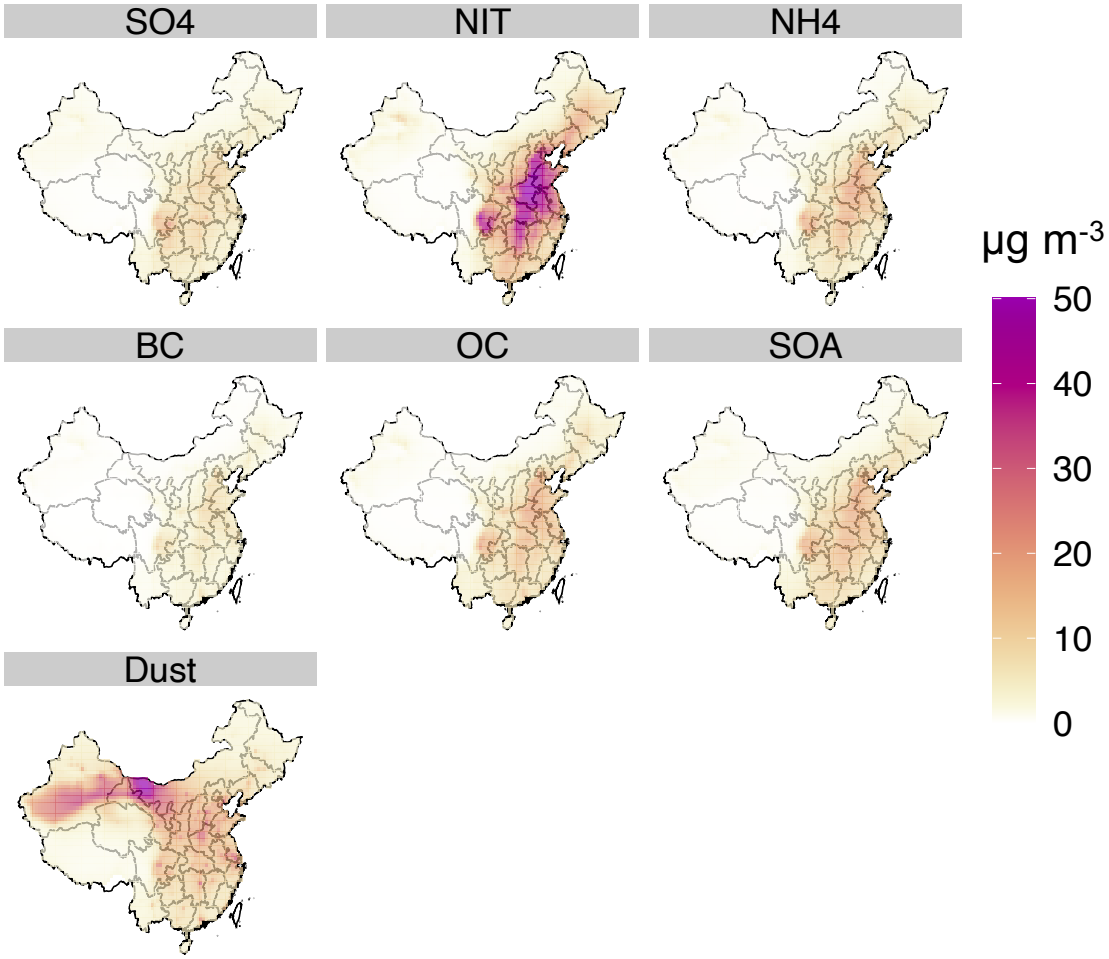

Figure S9. Concentrations of component species of $PM_{2.5}$ in China (average across 2013-2017). The figure shows concentrations of sulfate (SO_4), nitrate (NIT), ammonium (NH_4), black carbon (BC), organic carbon (OC), secondary organic aerosol (SOA), and dust.

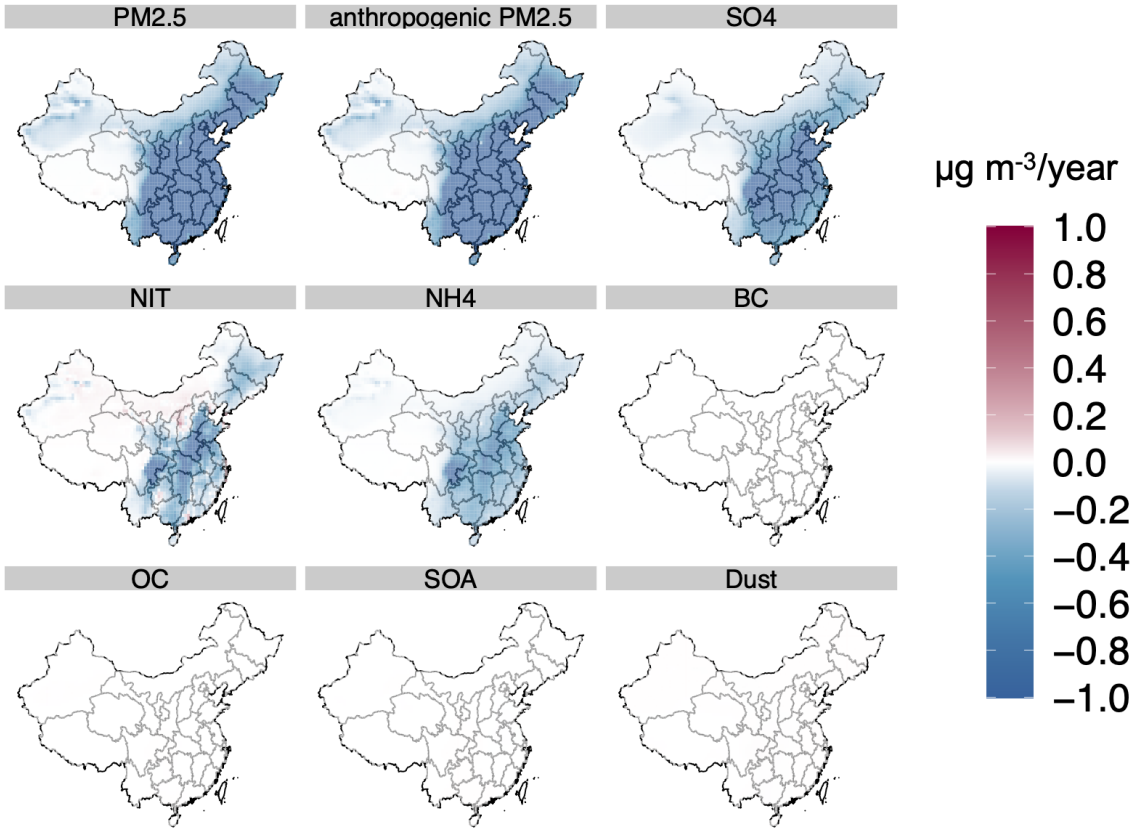

Figure S10. Counterfactual trends of component species of $PM_{2.5}$ in China. The figure shows counterfactual trends of total $PM_{2.5}$, anthropogenic $PM_{2.5}$ (total $PM_{2.5}$ excluding dust and sea salt), sulfate (SO_4), nitrate (NIT), ammonium (NH_4), black carbon (BC), organic carbon (OC), secondary organic aerosol (SOA), and dust.

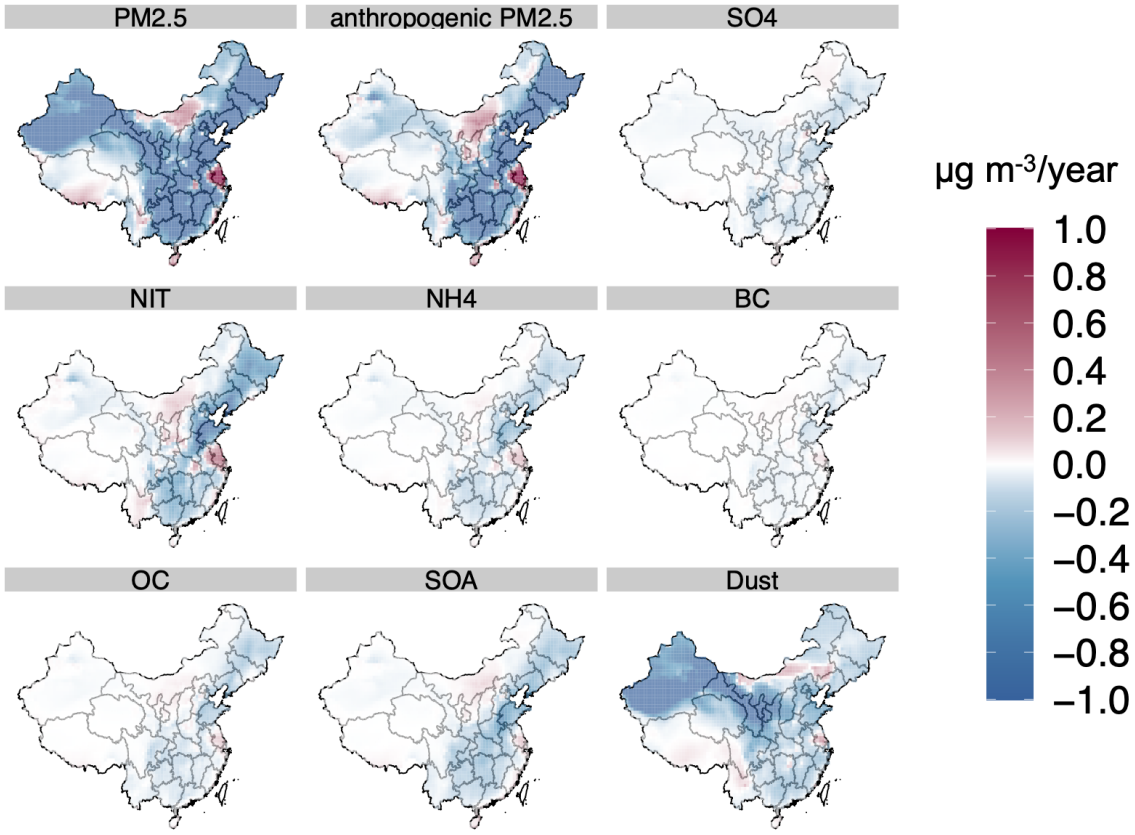

Figure S11. Differences between counterfactual trends and trends evaluated under MLR ($\beta^{MLR} - \beta^{count}$) of component species of $PM_{2.5}$ in China. The figure shows estimation errors of total $PM_{2.5}$, anthropogenic $PM_{2.5}$ (total $PM_{2.5}$ excluding dust and sea salt), sulfate (SO_4), nitrate (NIT), ammonium (NH_4), black carbon (BC), organic carbon (OC), secondary organic aerosol (SOA) and dust.

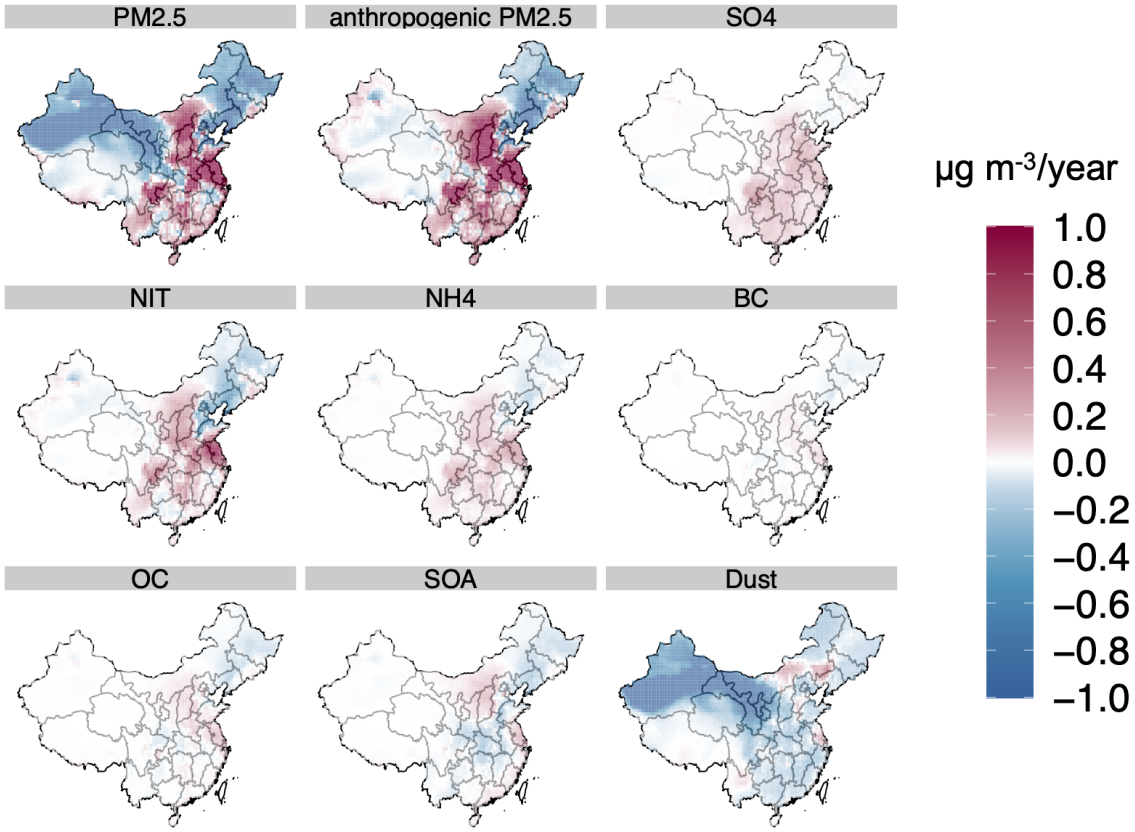

Figure S12. Differences between counterfactual trends and trends evaluated under RF-regional ($\beta^{RF-regional} - \beta^{count}$) of component species of $PM_{2.5}$ in China. The figure shows estimation errors of total $PM_{2.5}$, anthropogenic $PM_{2.5}$ (total $PM_{2.5}$ excluding dust and sea salt), sulfate (SO_4), nitrate (NIT), ammonium (NH_4), black carbon (BC), organic carbon (OC), secondary organic aerosol (SOA) and dust.

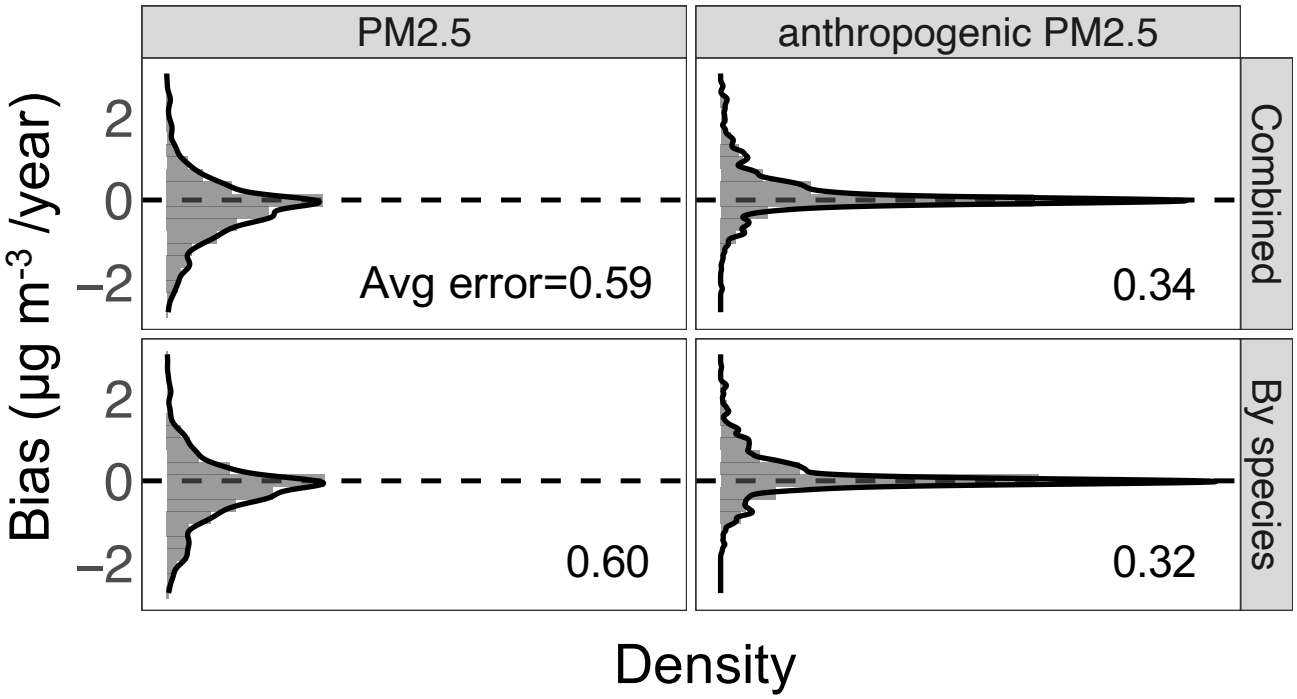

Figure S13. Histograms of estimation errors of trend estimates in $PM_{2.5}$ under two implementations of the *RF-regional* method (applied to China). The upper panels (*Combined*) show results of fitting RF models to the total concentrations of $PM_{2.5}$ to directly estimate trends (the main results). The lower panels (*By species*) show results of fitting RF models to individual $PM_{2.5}$ species and then combine predictions to estimate trends. The left panels show results for total $PM_{2.5}$ and right panels show results for anthropogenic $PM_{2.5}$ (total $PM_{2.5}$ excluding dust and sea salt). Average of the estimation errors across all grid cells is shown in the figure.

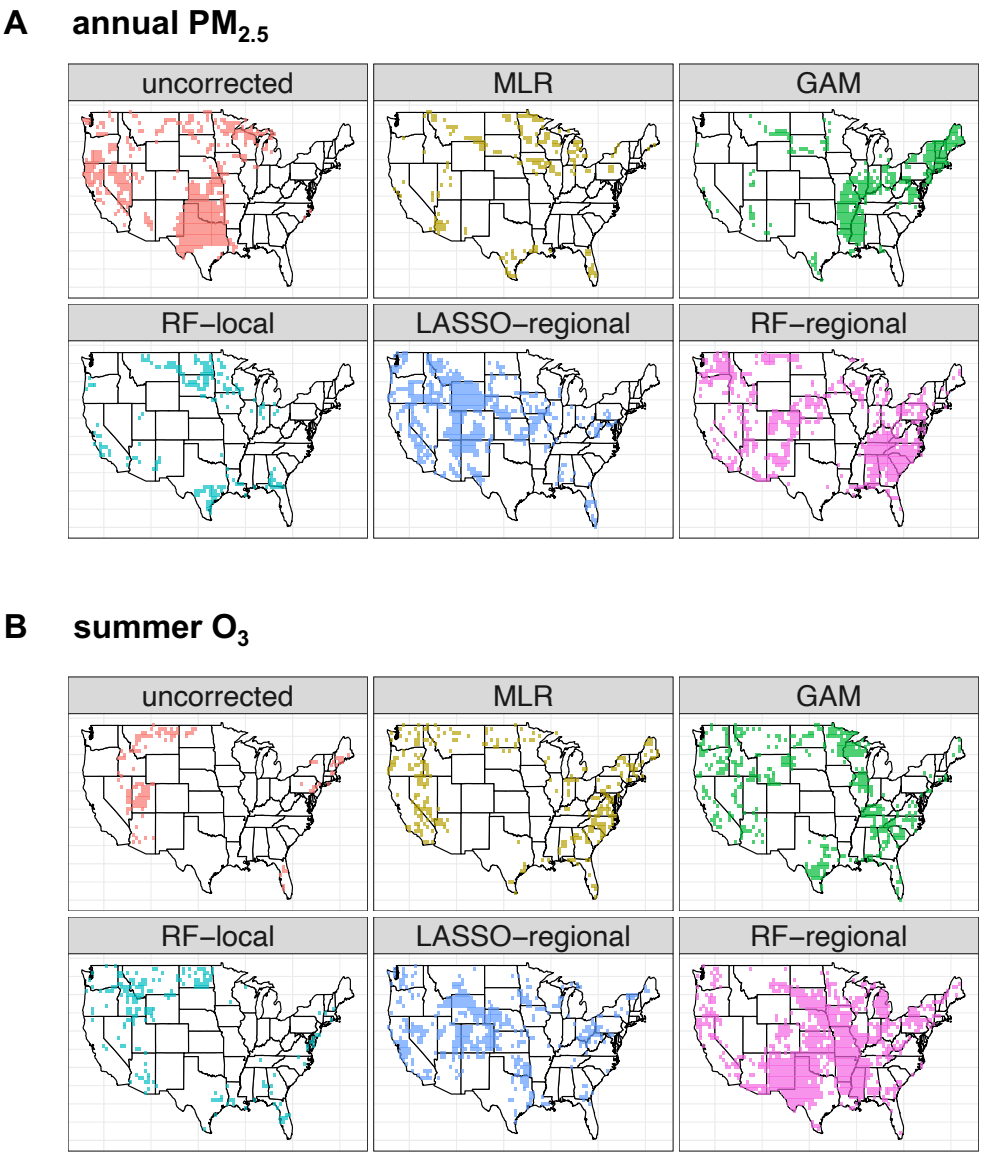

Figure S14. Best-performing correction method for each grid cell (US). For each method, the figure shows the grid cells at which the trend estimate has the smallest estimation error (i.e. closest to the trend in the counterfactual scenario) among the tested methods.

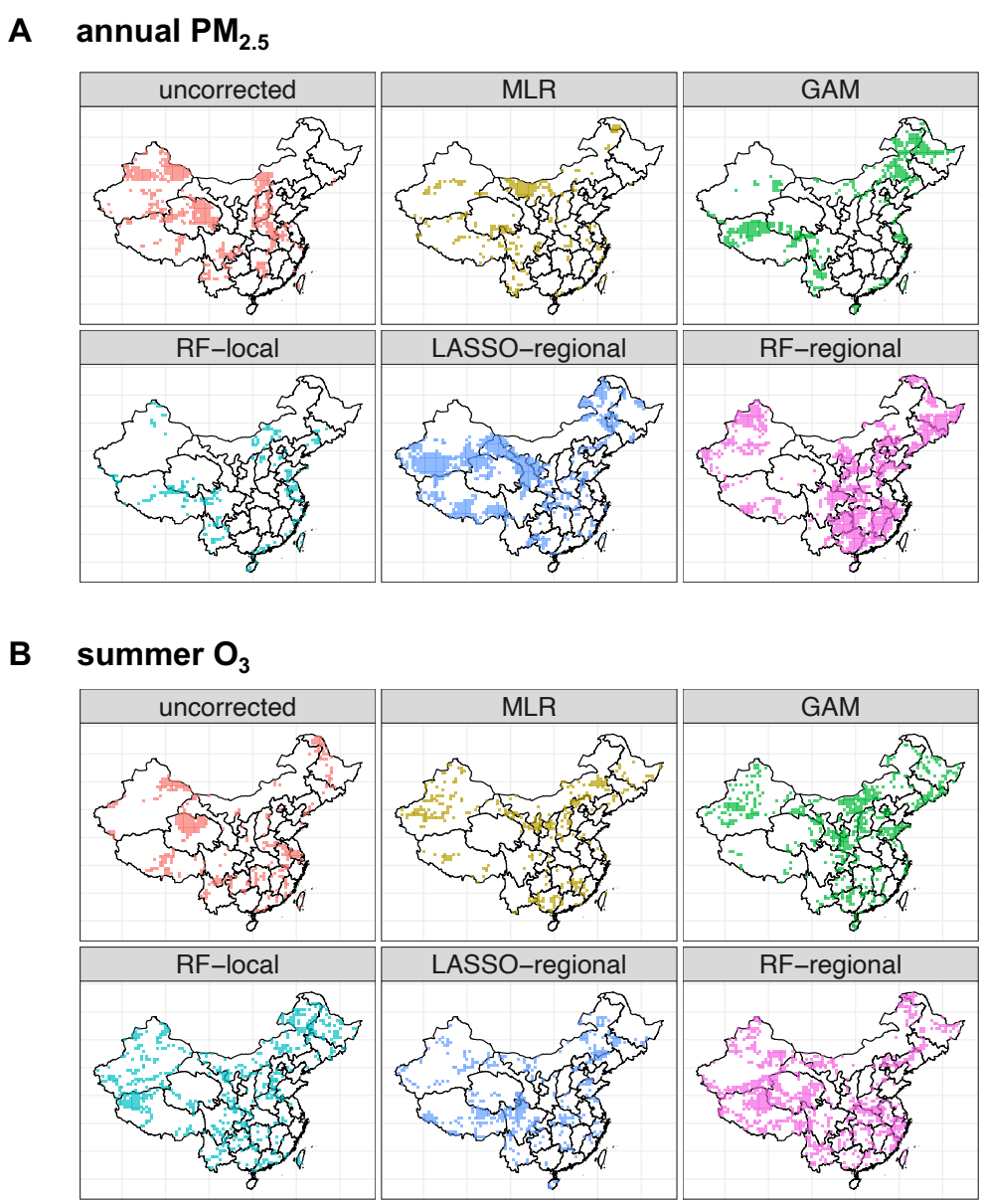

Figure S15. Best-performing correction method for each grid cell (China). For each method, the figure shows the grid cells at which the trend estimate has the smallest estimation error (i.e. closest to the trend in the counterfactual scenario) among the tested methods.

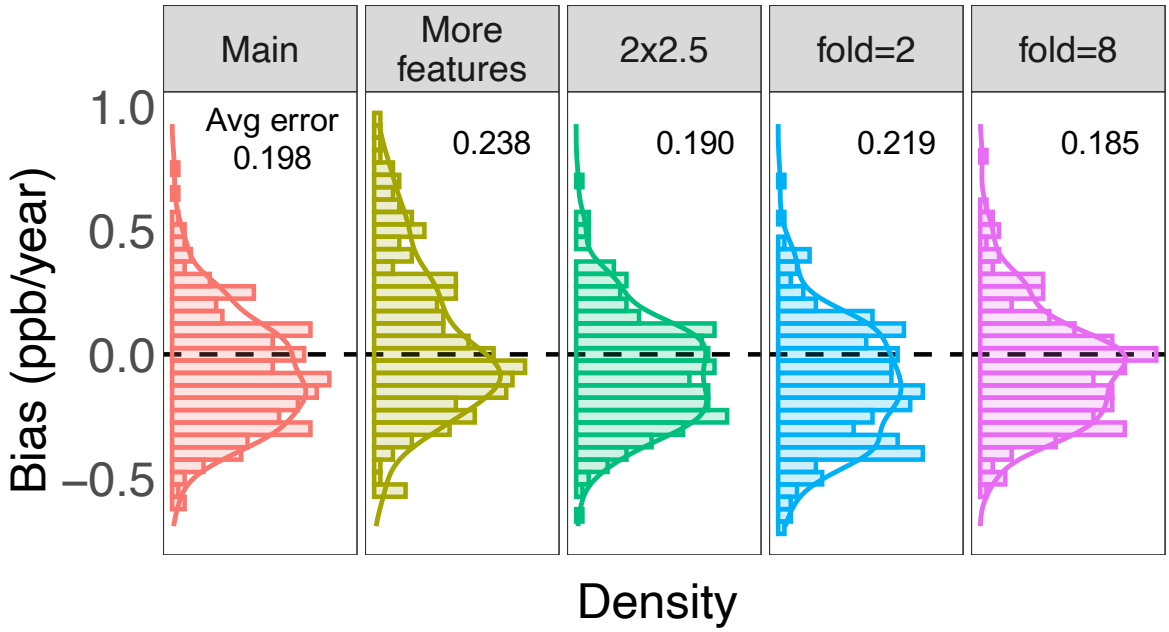

Figure S16. Histograms of estimation errors of trend estimates in O_3 under different implementations of the *RF-regional* method (applied to the US). From left to right: Main (the main results), More features (including 9 extra meteorological features), 2x2.5 (using regional features with spatial resolution at $2\times2.5^\circ$, instead of $4\times5^\circ$), fold=2 (using 2 folds for data-splitting and cross-fitting), fold=8 (using 8 folds for data-splitting and cross-fitting). Average of the absolute error for each implementation is shown in the figure. Here we only use a random subset of all the grid cells in the US due to high computational cost.

References

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J.: Double/debiased machine learning for treatment and structural parameters, 2018.

China's Ministry of Ecology and Environment: National Air Quality Monitoring Data, https://quotsoft.net/air/, 2021.

Li, M., Liu, H., Geng, G., Hong, C., Liu, F., Song, Y., Tong, D., Zheng, B., Cui, H., Man, H., Zhang, Q., and He, K.: Anthropogenic emission inventories in China: A review, National Science Review, 4, 834–866, https://doi.org/10.1093/nsr/nwx150, 2017.

U.S. Environmental Protection Agency: National Emissions Inventory 2017: Technical Support Document, https://www.epa.gov/sites/production/files/2021-02/documents/nei2017_tsd_full_jan2021.pdf, 2021a.

U.S. Environmental Protection Agency: Air Data: Air Quality Data Collected at Outdoor Monitors Across the US, https://aqs.epa.gov/aqsweb/airdata/download_files.html#Meta/, 2021b.