

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

Minghao Qiu^{1,*}, Corwin Zigler², and Noelle Selin^{1,3}

¹Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, USA

²Departments of Statistics and Data Science and Women's Health, University of Texas, Austin, USA

³Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, USA

*Current address: Department of Earth System Science, Stanford University, USA (mhqiu@stanford.edu)

Correspondence: Minghao Qiu (mhqiu@mit.edu)

1 **Abstract.** Evaluating the influence of anthropogenic emissions changes on air quality requires accounting for the influence of
2 meteorological variability. Statistical methods such as multiple linear regression (MLR) models with basic meteorological vari-
3 ables are often used to remove meteorological variability and estimate trends in measured pollutant concentrations attributable
4 to emissions changes. However, the ability of these widely-used statistical approaches to correct for meteorological variability
5 remains unknown, limiting their usefulness in the real-world policy evaluations. Here, we quantify the performance of MLR
6 and other quantitative methods using two scenarios simulated by a chemical transport model, GEOS-Chem, as a synthetic
7 dataset. Focusing on the impacts of anthropogenic emissions changes in the US (2011 to 2017) and China (2013 to 2017) on
8 PM_{2.5} and O₃, we show that widely-used regression methods do not perform well in correcting for meteorological variability
9 and identifying long-term trends in ambient pollution related to changes in emissions. The estimation errors, characterized as
10 the differences between meteorology-corrected trends and emission-driven trends under constant meteorology scenarios, can
11 be reduced by 30%-42% using a random forest model that incorporates both local and regional scale meteorological features.
12 We further design a correction method based on GEOS-Chem simulations with constant emission input and quantify the de-
13 gree to which emissions and meteorological influences are inseparable, due to their process-based interactions. We conclude
14 by providing recommendations for evaluating the effectiveness of emissions reduction policies using statistical approaches.

15 1 Introduction

16 Researchers and policy makers have long been interested in understanding the anthropogenic drivers of trends in observed air
17 pollutant concentrations in order to inform air quality policies. Declining trends in pollutant concentrations such as particulate
18 matter with diameter less than 2.5 microns ($\text{PM}_{2.5}$) have been observed in many countries that adopted policies to limit an-
19 thropogenic emissions such as SO_2 and NO_x , including the US (McClure and Jaffe, 2018) and China (Zhang et al., 2019). As
20 information on anthropogenic emissions are often unavailable or very uncertain, researchers and policy makers often rely on
21 the trends in measured air pollutants to assess the effects of polices. Evaluating the effectiveness of air quality policies requires
22 understanding the degree to which changing trends in observed concentrations can be attributed to anthropogenic emissions
23 changes. However, rigorous attribution requires correcting for the influence of changing meteorology, which has become in-
24 creasingly important but challenging in a changing climate (Saari et al., 2019). Numerous papers attempt to use statistical
25 methods to separate impacts of meteorology from emissions changes in evaluating trends in air quality, but the performances
26 of these commonly-used statistical approaches remain unassessed. Further, the impacts of meteorological variability may not
27 even be distinguishable from emissions-driven air quality trends, due to their interactions; the magnitude of this interaction also
28 remains unquantified. In this paper, we devise a model-based experiment for evaluating the performance of different statistical
29 methods used for meteorological corrections. We focus on a case of identifying emissions-driven linear trends in measured
30 concentrations of $\text{PM}_{2.5}$ and ozone (O_3), when information on the anthropogenic emission is not available.

31 Measured pollutant concentrations are often used as the primary basis for evaluating air quality actions. For example in
32 2013, China’s central government established targets that aimed to reduce annual average $\text{PM}_{2.5}$ concentrations of three urban
33 clusters by 15% to 25% between 2012 and 2017 (State Council of the People’s Republic of China, 2013). This later translated
34 into a stringent and binding target of a maximum annual mean $\text{PM}_{2.5}$ concentration of $60 \mu\text{g}/\text{m}^3$ in 2017 for Beijing, which was
35 ultimately reached (the 2017 concentration was $58.5 \mu\text{g m}^{-3}$) (Beijing Municipal Ecology and Environment Bureau, 2013).
36 However, several studies estimated that the concentration would have exceeded this target in Beijing were it not for meteo-
37 rological conditions in winter 2017 that favored pollution reductions (Vu et al., 2019; Chen et al., 2019; Cheng et al., 2019).
38 The European Union and US Environmental Protection Agency (EPA) use a three-year average of the $\text{PM}_{2.5}$ concentration
39 to determine compliance with air quality standards (European Union, 2020; U.S. Environmental Protection Agency, 2019).
40 The US EPA has also proposed to use statistical approaches that aim to correct for the impacts of weather variability on O_3
41 concentrations in the designation processes (Wells et al., 2021).

42 Many studies use multiple linear regression (MLR) models with basic meteorological variables to correct for meteorological
43 variability in order to estimate the impacts of emissions changes on measured air quality (Otero et al., 2018; Zhai et al., 2019;

Li et al., 2018, 2020; Han et al., 2020; Chen et al., 2020). Zhai et al. (2019) and Li et al. (2020) use MLR models to estimate the degree to which trends in $\text{PM}_{2.5}$ and O_3 from 2013 to 2019 in China were driven by anthropogenic emissions changes. They first use MLR to predict the $\text{PM}_{2.5}$ and O_3 concentrations with meteorological variables, and then interpret the residuals of the MLR model as signals resulting from emissions changes. A related approach is to combine MLR with techniques that can decompose time series of observed concentrations into long-term, seasonal, and short-term components (e.g., Kolmogorov-Zurbenko (KZ) filters (Zurbenko, 1994)). Ma et al. (2016) and Chen et al. (2019) use KZ filters to calculate the long-term component of observed $\text{PM}_{2.5}$ and then apply MLR to separate the impacts of long-term meteorological changes on the concentrations. Henneman et al. (2015) apply MLR to the short-term component (identified by KZ filters) of air pollutant concentrations near Atlanta during 2000 to 2012, to separate the impact of short-term meteorological variability, and then estimate the long-term trend in air quality.

Other statistical methods including non-linear regression or machine learning models have also been used to correct for meteorological variability (Holland et al., 1998; Carslaw et al., 2007; Hayn et al., 2009; Vu et al., 2019). One popular method is to use a generalized additive model (GAM) to estimate non-linear smooth functions of each meteorological variable within a given smoothing function family with penalization on non-smoothness. The US EPA uses a GAM model of temperature, wind direction and speed, humidity, pressure, stability, transport trajectories, and synoptic weather to perform weather corrections in assessing long term trends in O_3 (Camalier et al., 2007). An increasing number of studies use machine learning models (Grange et al., 2018; Vu et al., 2019; Zhang et al., 2020; Shi et al., 2021; Qu et al., 2020). Vu et al. (2019) uses a random forest model to predict pollutant concentrations in Beijing with time index and meteorological variables and then calculates the “weather-normalized” concentration for each day with 1000 sets of meteorological fields drawn from the historical meteorological data. They found that the decrease of $\text{PM}_{2.5}$ during 2013 to 2017 was largely driven by emissions reductions, although the magnitude of reduction is smaller when correcting for the meteorological variability.

Despite the large amount of papers which apply various meteorology correction methods, very little is known about whether these methods can effectively correct for meteorological variability and thus reveal the underlying causal impacts of anthropogenic emissions changes. Most studies cite the prediction performance of their statistical models (such as R^2 and/or mean squared errors) to justify their method choice and analysis. However, good prediction performance does not guarantee correct inference of causal effects (Runge et al., 2019). The performance of these meteorology-corrected methods is unable to be assessed using observational data alone, as the underlying emission-driven trends without influence from meteorological variability cannot be derived from data. Runge et al. documents similar challenges with observational data and proposes to use physical models to benchmark causal inference methods in the broader domains of earth sciences (Runge et al., 2019). Further, statistical analyses often assume that the influence of meteorological variability on pollutant concentration can be cleanly sepa-

74 rated from the influence of anthropogenic emissions changes. This is not completely possible, as the impacts of meteorological
75 variability on pollutant concentration will also vary depending on the emissions. The degree to which this interaction affects
76 the ability to calculate emissions-related trends under changing meteorology also remains unknown.

77 Here, we conduct a model experiment to evaluate the performance of widely-used statistical models in correcting for meteo-
78 rological variability and estimating emissions-driven trends in air quality. We focus on the impacts of anthropogenic emissions
79 changes on annual $\text{PM}_{2.5}$ and summer O_3 in the US (2011-2017) and China (2013-2017), two periods well-studied in previous
80 literature. Using a 3-D atmospheric chemical transport model GEOS-Chem, we simulate two sets of scenarios – “observational
81 scenarios” with assimilated meteorological inputs (with interannual variability) and “counterfactual scenarios” with constant
82 meteorological inputs. Using simulated daily concentrations in the observational scenarios, we estimate meteorology-corrected
83 trends for each grid cell using different statistical correction methods. We then compare the derived trends with the emissions-
84 driven trends in the counterfactual scenarios (which are free of meteorological variability by design), calculating the resulting
85 “error” in trend estimation. We further design a correction method based on GEOS-Chem constant emission simulations, and
86 use it to quantify the degree to which attribution to meteorology and emissions separately is possible. Finally, we apply the
87 different statistical correction methods to observational data from surface monitoring networks in the US and China, discussing
88 the variability across different methods. We conclude by providing recommendations for techniques to evaluate air pollution
89 policies under changing meteorological conditions.

90 2 Method

91 2.1 GEOS-Chem

92 GEOS-Chem is a global three-dimensional chemical transport model driven by assimilated meteorological data from the God-
93 dard Earth Observation System (GEOS-5) of the NASA Global Modeling and Assimilation Office (GMAO) (Bey et al. (2001),
94 <http://www.geos-chem.org/>). The simulation of $\text{PM}_{2.5}$ in GEOS-Chem represents an external mixture of secondary inorganic
95 aerosols, carbonaceous aerosols, sea salt, and dust aerosols. GEOS-Chem includes detailed O_3 - NO_x -volatile organic carbon
96 (VOC)-aerosol-Halogen tropospheric chemistry (Travis et al., 2016; Sherwen et al., 2016). The GEOS-Chem model has been
97 previously used to study the changes in $\text{PM}_{2.5}$ and O_3 during our studied periods, and model simulations have been shown to
98 be consistent with the observed concentrations (e.g., see Li et al. (2017a); Xie et al. (2019) for the US, and Li et al. (2018);
99 Lu et al. (2019); Zhai et al. (2021) for China). Studies in both regions show that the GEOS-Chem model is able to reproduce
100 the spatial, seasonal, and interannual variability and the long-term trends in observed pollutant concentrations, despite biases
101 in absolute concentrations in certain species and regions (Heald et al., 2012; Travis et al., 2016; Tian et al., 2021).

102 We use GEOS-Chem version 12.3.0 with a horizontal resolution of $0.5^\circ \times 0.625^\circ$ in North America and Asia (Wang et al.,
103 2004). For each scenario, we first conduct a global run at a horizontal resolution of $4^\circ \times 5^\circ$, with a 12 month spin-up. These
104 global runs are then used as the boundary conditions for nested simulations in US and Asia with finer resolution of $0.5^\circ \times 0.625^\circ$.

105 2.2 GEOS-Chem scenarios

106 Table 1 shows the simulations included in our model experiments. We simulate two sets of scenarios – “observational sce-
107 narios” with interannual variability in meteorology and “counterfactual scenarios” with constant meteorological inputs. Both
108 scenarios use the same emissions inventory as input (see Method 2.3). For each grid cell, we estimate the linear trends in
109 pollutant concentrations from simulated daily $\text{PM}_{2.5}$ and O_3 concentrations. We focus on the daily 24-hour average $\text{PM}_{2.5}$ of
110 all seasons, and the maximum daily average 8-hour (MDA8) O_3 in summer (June, July, August). Our GEOS-Chem simulations
111 use meteorological fields from the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2)
112 (Gelaro et al., 2017). We aggregate the hourly meteorological data for consistency with the pollutant concentrations: a 24-hour
113 average for $\text{PM}_{2.5}$ analysis and the corresponding 8-hour average for O_3 . Meteorological features that are used in the statistical
114 models can be found in 2.4.

115 2.2.1 Observational scenarios

116 Observational scenarios simulate $\text{PM}_{2.5}$ and O_3 under changing emissions and changing meteorological fields. Trends es-
117 timated under the observational scenarios (β^{obs}) are subject to the influences of interannual meteorological variability. Our
118 model experiments were not specifically designed to reproduce observed air quality in these two regions, but rather to provide
119 a realistic test case for our statistical experiments. Nevertheless, as shown in figure S1 and S2, the simulated concentrations in
120 $\text{PM}_{2.5}$ and O_3 largely reproduce the daily variability in observed pollutant concentrations. The linear trends in simulated $\text{PM}_{2.5}$
121 and O_3 concentrations in the observational scenario are largely consistent with trends of the measured concentrations. For ex-
122 ample, the average trend (\pm one standard deviation) in the US is $-0.27 \pm 0.30 \mu\text{g}^{-3}/\text{year}$ (observation) and $-0.39 \pm 0.24 \text{ ppb}/\text{year}$
123 (GEOS-Chem) for $\text{PM}_{2.5}$, and $-0.91 \pm 0.98 \text{ ppb}/\text{year}$ (observation) and $-1.02 \pm 0.83 \text{ ppb}/\text{year}$ (GEOS-Chem) for O_3 . The only
124 exception is that our model cannot reproduce the increasing $\text{PM}_{2.5}$ trends in Northwest US because we do not consider the
125 interannual variability in the biomass burning emissions.

126 2.2.2 Counterfactual scenarios

127 Counterfactual scenarios simulate $\text{PM}_{2.5}$ and O_3 under changing emissions but constant meteorology. All simulation years in
128 the counterfactual scenario use the meteorological fields of the start year (2011 for US, 2013 for China). Trends estimated

129 under the counterfactual scenario (β^{count}) are not subject to interannual meteorological variability; we use this as a proxy for
130 the trends in pollutant concentrations driven by emissions changes alone.

131 **2.2.3 Assumptions for GEOS-Chem experiments**

132 It is important to note that we do not assume our GEOS-Chem simulations perfectly represent the underlying pollutant con-
133 centration in the real world (although the model compares relatively well with the observational data). Rather, our main focus
134 is to evaluate how much different statistical methods can explain the differences between the observational and counterfactual
135 scenarios. The assumption here is that the differences between observational and counterfactual scenarios are useful approxima-
136 tions of the impacts of meteorological variability on pollutant concentrations. The implications of uncertainty in GEOS-Chem
137 for our results can be found in the discussion section.

GEOS-Chem scenarios	Emissions inventory	Meteorological fields	Trend estimates	Meteorological correction
Counterfactual scenarios	Changing 2011-2017 (US) 2013-2017 (China)	Constant 2011 (US) 2013 (China)	β^{count}	No correction needed
Observational scenarios	Changing 2011-2017 (US) 2013-2017 (China)	Changing 2011-2017 (US) 2013-2017 (China)	$\beta^{uncorrected}$ β^{MLR} β^{GAM} β^{RF} $\beta^{LASSO-regional}$ $\beta^{RF-regional}$ β^{gc}	No correction Linear combination of local features GAM using local features RF using local features LASSO using local and regional features RF using local and regional features Use simulations from constant emissions scenarios
Constant emissions scenarios	Constant 2011 (US) 2013 (China)	Changing 2011-2017 (US) 2013-2017 (China)		

Table 1: Overview of GEOS-Chem scenarios and meteorological correction methods.

138 2.3 Emissions inventory

139 For the US, we use the National Emissions Inventory 2011 (NEI 2011) as a baseline emissions inventory and scale the emissions
140 in 2012 to 2017 to match the annual total emissions each year (U.S. Environmental Protection Agency, 2021a). For China, we
141 use the monthly Multi-resolution Emission Inventory for China (MEIC) during 2013 to 2017 (Li et al., 2017b; Zheng et al.,
142 2018). During the studied time periods, US and China experienced dramatic decreases in anthropogenic emissions, particularly
143 in SO₂ and NO_x. In the US, the total anthropogenic emissions of SO₂ decreased by 57% and NO_x emissions decreased by 26%
144 during 2011 to 2017 (see figure S3). In China, anthropogenic SO₂ emissions decreased by 59% and NO_x emissions decreased
145 by 21% during the 2013-2017 period (see figure S4).

146 Natural emissions of multiple chemical species are calculated online in the simulations (rather than prescribed) in the GEOS-
147 Chem model and thus can be influenced by meteorological variability (see Keller et al. (2014) for more details). Impacts of
148 meteorology on PM_{2.5} and O₃ concentrations through changes in the natural emissions are considered here as part of the
149 meteorology - concentration relationship. These emissions include NO_x emissions from lightning and soil processes, sea salt
150 emissions, dust emissions, and biogenic volatile organic carbon (VOC) emissions. However, biomass burning emissions are
151 prescribed in the GEOS-Chem model and we hold them constant at the level of the start year. We make this simplification
152 because the GEOS-Chem model uses prescribed biomass burning emissions from external inventories such as Global Fire
153 Emissions Database (Werf et al., 2017), and it is impossible to distinguish natural fire emissions (part of the meteorological
154 variability) from anthropogenic fire emissions (e.g., from farm residual burning).

155 2.4 Statistical and machine learning models

156 2.4.1 Model with local meteorological variables

157 We assess the performance of statistical and machine learning models to correct for the meteorological variability in the
158 observational scenarios. We evaluate these methods with a commonly-used framework (e.g., used in Li et al. (2018) and
159 Zhai et al. (2019)) which models the air pollutant concentrations of each individual grid cell using an additive form of a
160 trend component, a meteorology component, and time fixed effects (to capture daily and monthly variability not related to
161 meteorology). More specifically, we estimate the following regression equation for each grid cell i :

$$162 \quad y_{it} = \beta_i^{obs} \times t + f_i(X_{it}) + \eta_{it} + \epsilon_{it} \quad (1)$$

where y_{it} denotes the PM_{2.5} or O₃ concentration at grid cell i on day t . t is the time index (e.g., in the US, $t=1$ for January 1st, 2011 and $t=2$ for January 2nd, 2011). X_{it} denotes the local meteorology features (i.e. meteorological variables in grid cell i on day t). η_{it} is the month-of-year \times day-of-month fixed effect to capture daily and monthly variability of pollutant concentrations that are not related to the meteorological variability (e.g., seasonal cycle in O₃ and PM_{2.5}). ϵ_{it} is the normally-distributed error term. β_i^{obs} represents the meteorology-corrected trend in PM_{2.5} or O₃ concentration for grid cell i under a specific method. We use the absolute differences $|\beta_i^{obs} - \beta_i^{count}|$ to evaluate the performance of different methods to correct for meteorological variability for any given grid cell i .

Here, $f_i(X_{it})$ represents the specifications of local meteorological features for grid cell i under different methods. In addition to the commonly-used multiple linear regression (MLR) model, we also evaluate following models with higher flexibility: polynomial regression models (quadratic, cubic), cubic spline models, generalized additive models (GAM, implemented with R package “mgcv” by Wood (2011)), and Random Forest (RF) models. We focus on the methods in table 1 in the main manuscript, and the performance of the other methods can be found in table S1 and S2. Note that the time fixed effects are modelled differently in RF models due to the estimation procedure. More details on the implementation of RF can be found in SI.

We use the following ten variables from MERRA-2 as our selected meteorological features for the statistical analysis: surface temperature, precipitation, humidity, planetary boundary layer height, cloud fraction, surface air pressure, and wind speed (U and V direction, at surface and 850 hpa level). These variables are the most commonly used features in previous studies. We also perform sensitivity analyses that include nine more meteorological features: direct photosynthetically-active radiation, diffuse photosynthetically-active radiation, tropopause pressure, friction velocity, top soil moisture, root soil moisture, snow depth, surface albedo, and surface air density. These features are selected because they are used as primary or intermediate inputs for calculating PM_{2.5} or O₃ concentrations in the GEOS-Chem model and may contain information that help explain variability in pollutant concentrations.

2.4.2 Model with local and regional meteorological variables

We also evaluate models that use both local and regional meteorological features. Regional meteorological features are important for explaining variability in local pollutant concentrations due to 1) pollution transport from neighboring locations, and 2) influences from meteorological systems at synoptic scale (i.e. large scale weather systems that span over 1000 kilometers such as circulation patterns) (Tai et al., 2012; Shen et al., 2015; Zhang et al., 2018; Leung et al., 2018; Han et al., 2020). As the incorporation of both local and regional features can quickly expand the dimensionality of the feature space, here we use the Least Absolute Shrinkage and Selection Operator (LASSO) and the Random Forest (RF) model, two statistical models that

show good prediction performances with high dimensional data inputs. We estimate the following equations:

$$y_{it} = \beta_i^{obs} \times t + g_i(X_{it}, Z_t) + \eta_{it} + \epsilon_{it} \quad (2)$$

where $g_i()$ denotes the functional form fitted by LASSO or RF. X_{it} again denotes the local meteorology features for grid cell i on day t . Z_t denotes the regional scale meteorology features including the meteorological features for every grid cell 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 $10 \times 98 = 980$ regional scale features (Z_t). The coefficient β_i^{obs} is obtained with the double machine learning approach by Chernozhukov et al. (2018). More details on the implementation of LASSO and RF can be found in SI.

2.5 Correction approach using GEOS-Chem constant emissions scenario

We further design and evaluate an approach to correct for meteorology variability with GEOS-Chem simulations (referred to as “constant-emis” approach). The “constant-emis” approach uses GEOS-Chem simulations with constant anthropogenic emissions and changing meteorological fields (“constant emissions scenarios” in table 1). All years in the constant emissions scenario use anthropogenic emissions of the start year (2011 for US, 2013 for China). We estimate the following equations:

$$y_{it} = \beta_i^{gc} \times t + SIM_{it} + \eta_{it} + \epsilon_{it} \quad (3)$$

where SIM_{it} denotes the simulated concentrations on day t in grid cell i in the constant emissions scenarios. SIM_{it} serves a similar purpose as the term “ $f_i(X_{it})$ ” in equation 1, but comes from the GEOS-Chem simulation. Some previous studies have also used model simulations with constant emissions input as a way to characterize meteorological variability (Zhong et al., 2018; Zhao et al., 2020). β_i^{gc} is the estimated meteorology-corrected trend in $PM_{2.5}$ or O_3 concentration using this model-based correction method.

Compared to previous statistical and machine learning approaches, the “constant-emis” approach better captures the meteorological variability as simulated in GEOS-Chem (as SIM_{it} are directly taken from GEOS-Chem). Therefore, the difference between the trend estimates (β^{gc}) and counterfactual trends (β^{count}) provides a conceptual lower bound for estimation errors using the framework of equation 1 to perform meteorological corrections. The commonly-used framework of equation 1 assumes that the impacts of meteorology variability can be separated from the impacts of anthropogenic emissions. In our experiments, this assumption indicates that the differences between the counterfactual scenario and the observational scenario

217 can be solely explained by the meteorological variables. However, the difference in pollutant concentrations between these
218 scenarios is also in part driven by emissions in their interaction with meteorology (despite the fact that our different scenar-
219 ios use the same emissions inventory). We use $|\beta_i^{gc} - \beta_i^{count}|$ to quantify the estimation error associated with ignoring such
220 interactions in this framework.

221 2.6 Air quality observation data

222 We use the surface air quality measurements from the Air Quality Systems administered by the US EPA (U.S. Environmental
223 Protection Agency, 2021b). We use the daily 24-hour average of $PM_{2.5}$ concentrations for all months and the daily maximum
224 8-hour average (MDA8) O_3 concentrations for June, July and August. Figure S1 shows the locations, trends in measured
225 concentrations, and correlations between GEOS-Chem simulations and measured concentrations.

226 The surface air quality measurements in China come from the monitoring network from China’s Ministry of Ecology and
227 Environment China’s Ministry of Ecology and Environment (2021). The monitoring network was launched in 2013 and has
228 expanded to all prefecture level cities in mainland China. We use the daily 24-hour average of $PM_{2.5}$ concentrations and the
229 MDA8 O_3 concentrations for summer. Figure S2 shows the locations, trends in measured concentrations, and correlations
230 between GEOS-Chem simulations and measured concentrations.

231 We use the meteorological variables from MERRA-2 when performing meteorology corrections at these monitoring stations,
232 because the meteorology information is not available for all these variables at the station level. This is consistent with previous
233 analysis estimating the meteorology-corrected trends of the observational air quality data (e.g., Li et al. (2018)).

234 3 Results

235 3.1 Performance of different correction methods: US (2011-2017)

236 Figure 1A and 1C show the trends in $PM_{2.5}$ and O_3 concentrations in the counterfactual scenarios in the US. When holding
237 meteorological fields constant across years, decreasing trends in the simulated $PM_{2.5}$ concentrations across the US result from
238 decreasing anthropogenic emissions. In particular, the counterfactual scenario has substantial declining trends in $PM_{2.5}$ in the
239 East US where SO_2 emissions decreased dramatically. The scenario also has negative linear trends in O_3 concentrations in all
240 but three grid cells in the West. Increases in summer O_3 in these locations result from the non-linear relationship between O_3
241 concentrations and NO_x emissions.

242 Figure 1B shows the degree to which different meteorological correction methods can recover the emissions-driven trends in
243 the counterfactual scenarios. The figure shows the magnitude of estimation error in trend estimates in $PM_{2.5}$ for each grid box

($|\beta^{obs} - \beta^{count}|$). When no correction for meteorology is performed (“uncorrected” in figure 1B), we observe large estimation errors in trend estimates over the Northeast and Southern US by up to $0.25 \mu\text{g m}^{-3}/\text{year}$, an error that is 50% of the trend estimates under the counterfactual scenarios. We find that the widely-used MLR method does not help reduce these errors in $\text{PM}_{2.5}$ trend attribution. MLR has a modest impact on reducing the errors in Northeast US, but it does not decrease the errors over the Southern US and leads to higher errors over Midwest. Nationwide, the average magnitude of errors (relative to the counterfactual scenario) slightly increases with the MLR correction ($0.083 \mu\text{g m}^{-3}/\text{year}$) compared to the uncorrected case ($0.066 \mu\text{g m}^{-3}/\text{year}$). Among the five methods, we find that the RF model using both local and regional scale features (“RF-regional” in figure 1) offers the best performance in recovering the trends in the counterfactual scenarios and is the only method that yields smaller errors than the uncorrected case (the nationwide average error decreased by $0.019 \mu\text{g m}^{-3}/\text{year}$, or 28% less). The RF-regional model also outperforms the RF-local and LASSO-regional models, suggesting the importance of considering non-linearity, interactions between different meteorological features, and regional meteorology information in correctly adjusting for the impacts of meteorology.

Meteorological variability has a substantial influence on the summertime O_3 trends in the US during this period (as shown in figure 1D). Relative to the counterfactual scenario, the uncorrected O_3 trends are biased by over 1-2 ppb/year in large areas of California, Midwest and Southern US (as much as 320% of the counterfactual trends). This is largely driven by the fact that the 2011 and 2012 summer was particularly hot in these regions and led to higher concentrations of O_3 at the beginning of this 7-year period (see figure S6 for the Southern and Midwest US). Therefore, failure to correct for meteorological variability results in much more negative trend estimates in the O_3 concentrations in these areas compared to the counterfactual scenario (see figure S5). Meteorology corrections with MLR or GAM help reduce these estimation errors substantially (nationwide average error is reduced by 51% using MLR or 57% using GAM compared to uncorrected trends), while large errors still persist in the Midwest and South. Similar to the case of $\text{PM}_{2.5}$, the RF-regional model offers the best performance in correcting for meteorological variability (the national average error is further reduced by 42%, compared to MLR), and it is especially helpful in reducing the errors over the Midwest and South (regional average error is reduced by 64% and 44%, respectively, compared to MLR).

3.2 Performance of different correction methods: China (2013-2017)

Figure 2A and 2C show the trends in $\text{PM}_{2.5}$ and O_3 concentrations in the counterfactual scenarios in China. We find a substantial decline in simulated $\text{PM}_{2.5}$ concentration during 2013 to 2017, particularly in eastern and central China. In contrast, there is little change in the simulated $\text{PM}_{2.5}$ concentrations in western China in the counterfactual scenario, where $\text{PM}_{2.5}$ is dominated by dust species largely driven by natural processes (see figure S8). For summer O_3 , there are decreasing trends in

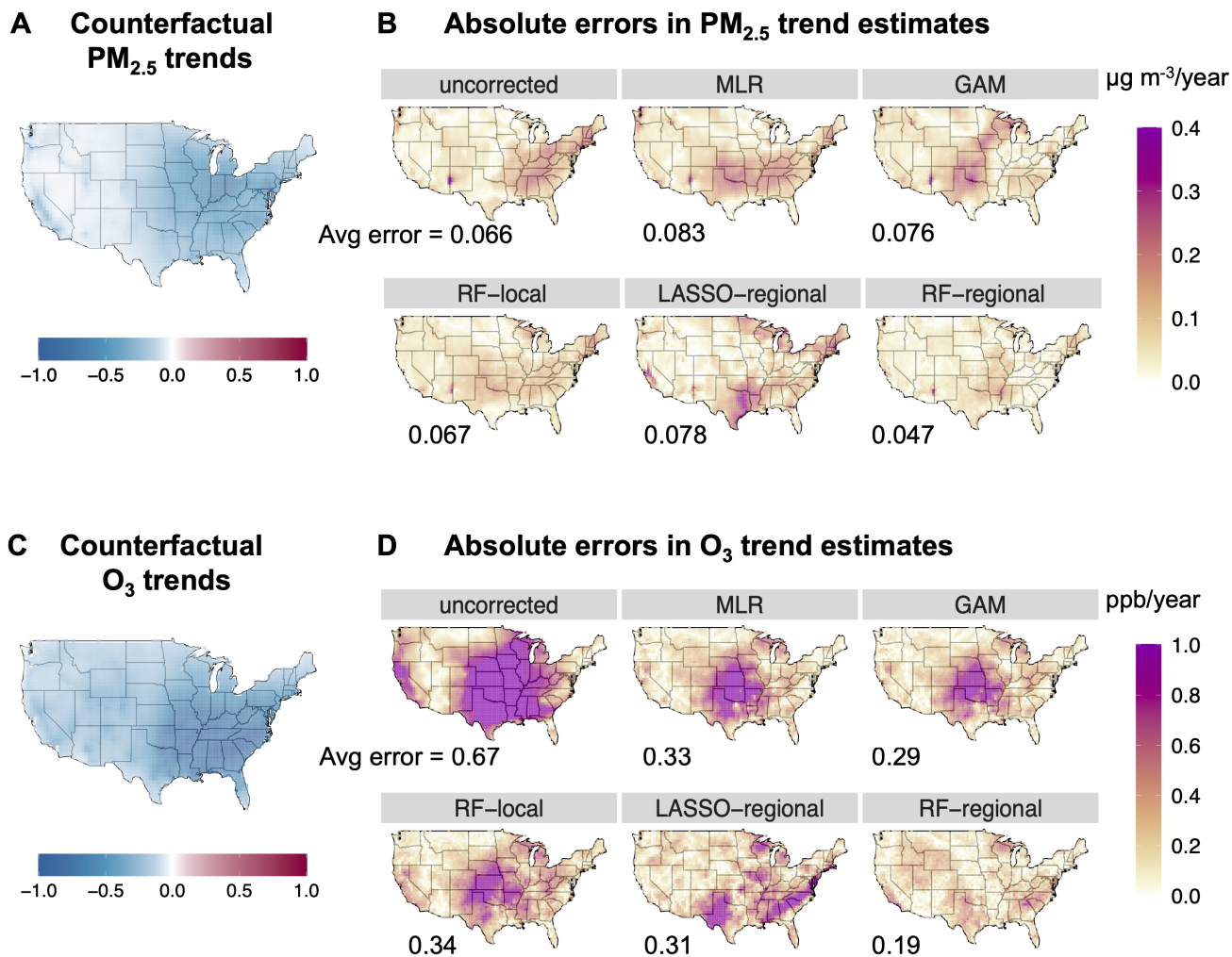


Figure 1. Trend estimates of daily annual PM_{2.5} (Panels A and B) and summer O₃ (C and D) in the US. Panels A and C show trend estimates under the counterfactual scenario (β^{count}). Panels B and D show the absolute magnitude of errors of trend estimates under different correction methods compared with the counterfactual scenarios ($|\beta^{obs} - \beta^{count}|$). The average of the absolute errors for each method is shown in the figure. Unit of trend estimate is $\mu\text{g m}^{-3}/\text{year}$ for PM_{2.5} or ppb/year for O₃.

273 the counterfactual scenario in most parts of China, except for North China and some urban areas. This is largely consistent with
274 previous studies that attempt to attribute emissions-related changes in O_3 concentrations during this period based on modeling
275 or observational data (Li et al., 2018, 2020; Lu et al., 2020).

276 Figure 2B shows the magnitude of estimation errors in the trend estimates of annual $PM_{2.5}$ in China under different correction
277 methods. We find the underlying meteorological variability has a substantial impact on $PM_{2.5}$ trends in China during this period.
278 We observe large differences between the uncorrected and counterfactual trends in simulated $PM_{2.5}$ concentrations, particularly
279 in Northwest and Northeast China. Similar to the model experiments in the US, we find that MLR and GAM methods fail to
280 correct for this underlying meteorological variability and lead to further increases in estimation errors in many locations.
281 Relative to the counterfactual scenario, the nationwide average error increases to $0.90 \mu g m^{-3}/year$ with MLR and $1.06 \mu g$
282 $m^{-3}/year$ with GAM (compared to $0.89 \mu g m^{-3}/year$ with no correction). We find that the RF-regional model recovers the
283 counterfactual trends better than other methods (nationwide average error: $0.64 \mu g m^{-3}/year$; an improvement by 30% relative
284 to MLR), but it is still not able to correct for the persistent estimation errors over Northwest China. We further analyze the
285 performance of correction methods for the different component species of $PM_{2.5}$. As shown in figure S9 and S10, the MLR
286 model is particularly unable to correct for the impacts of meteorological variability on nitrate and dust species. Compared
287 with MLR, the RF-regional model better corrects for the impacts of meteorology on secondary organic aerosol species in
288 South and Central China and ammonium in Northeast, but only yields modest improvement in correcting for the errors in dust
289 concentrations over Northwest China (see figure S11). In a sensitivity analysis, we use an approach that first fits RF-regional
290 models of each individual $PM_{2.5}$ species, and then combine predictions to each species to derive trend estimates. The results
291 are largely similar to the main approach that fits models to the total $PM_{2.5}$ concentration (see figure S12).

292 Figure 2D shows the magnitude of errors in the trend estimates for summer O_3 under different correction methods in
293 China. We find that the MLR model only modestly reduces the estimation errors compared to the uncorrected cases, and
294 the RF-regional model offers the best overall performance. The nationwide average error is reduced to $0.28 ppb/year$ using
295 the RF-regional model (relative to $0.43 ppb/year$ uncorrected and $0.41 ppb/year$ with MLR). Similar to the evaluation of
296 summer time O_3 in the US, we find the non-linear models (GAM, RF-local) perform better than MLR, but are not as good as
297 the RF-regional model. Surprisingly, the LASSO-regional model performs the worst in recovering the counterfactual trends.
298 This suggests the importance of considering non-linearity and regional meteorological features in understanding the O_3 –
299 meteorology relationships. Compared to the US case, we find the impacts of meteorological variability on O_3 and the method
300 performances are much more spatially heterogeneous (see figure S5, S7), which may be partially due to the more heterogeneous
301 O_3 regimes in China during this period.

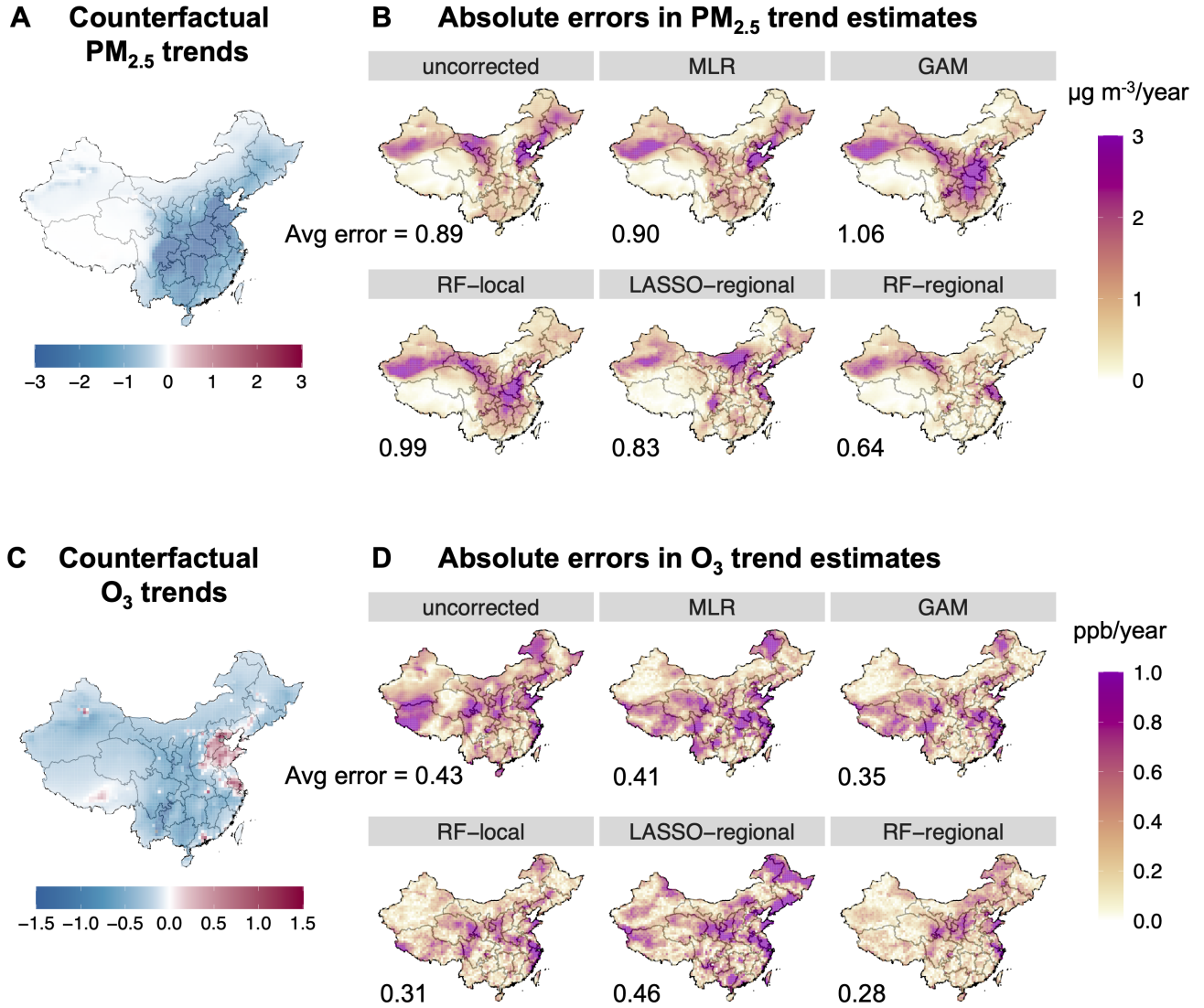


Figure 2. Trend estimates of daily annual PM_{2.5} (Panels A and B) and summer O₃ (C and D) in China. Panels A and C show trend estimates under the counterfactual scenario (β^{count}). Panels B and D show the absolute magnitude of errors of trend estimates under different correction methods compared with the counterfactual scenarios ($|\beta^{obs} - \beta^{count}|$). The average of the absolute errors for each method is shown in the figure. The unit of the trend estimate is $\mu\text{g m}^{-3}/\text{year}$ for PM_{2.5} or ppb/year for O₃.

302 3.3 Limitations in separating meteorological and emissions influence: quantified with constant emission scenarios

303 In our model experiments in both US and China, we find large differences remain between the trends evaluated with statistical
304 models (even the best-performed RF-regional model) and counterfactual trends. The remaining differences could result from
305 two different factors: 1) the statistical model cannot capture the complex relationship between meteorology and pollutant
306 concentrations, and/or 2) the differences between the observational scenarios and counterfactual scenarios depend not only
307 on the meteorological variability but also the anthropogenic emissions in their interaction with meteorology (i.e. impacts of
308 meteorology on air quality also depends on the level of emissions).

309 We quantify the potential magnitude of this second factor using our constant-emis approach. As the constant-emis approach
310 captures the exact relationship between meteorology and pollutant concentrations in GEOS-Chem, the error of the constant-
311 emis approach is only associated with the second factor above and thus provides a conceptual lower bound of the estimation
312 errors that can be achievable by any statistical approaches. Figure 3 shows the estimation errors of trend estimates using the
313 constant emissions scenarios simulated by GEOS-Chem. We focus on the trends in summer O_3 in the US and annual $PM_{2.5}$ in
314 China, for which we see the largest impacts of meteorological variability on the pollutant trends and the largest improvements
315 in reducing estimation errors from the correction methods. Compared to the statistical models (e.g., MLR and RF-regional
316 in figure 3A and 3C), trends evaluated using the constant-emis approach are very similar to the trends in the counterfactual
317 scenarios. The national average error of trend estimates is only 0.04 ppb/year for the O_3 trends in the US (relative to 0.33
318 ppb/year under MLR or 0.19 ppb/year under RF-regional), and only 0.08 $\mu g m^{-3}/year$ for the $PM_{2.5}$ trends in China (relative
319 to 0.91 $\mu g m^{-3}/year$ under MLR or 0.64 $\mu g m^{-3}/year$ under RF-regional).

320 However, the estimation errors calculated above are non-negligible and can be large in certain regions. As shown in Figure
321 3B and 3D, the constant-emis approach generally yields trend estimates biased by 10% relative to the counterfactual trends, but
322 the errors can be up to 40% in certain areas. This error term is the result of ignoring how emissions could potentially influence
323 the impacts of meteorology on the pollutant concentrations – that is, the impacts of the same meteorological variability on
324 concentrations may be different in the start year (with high emissions) compared to the end year (with low emissions).

325 3.4 Application to observational data

326 Figure 4 shows the regional trends in O_3 in the US and trends in $PM_{2.5}$ in China estimated from the GEOS-Chem simulations
327 and the measured concentrations from surface monitoring networks (only grid cells that overlap with monitor locations are
328 shown here). As shown in figure 4A, how to correct for meteorological variability is important for attributing summer O_3
329 trends to emissions reductions in the US. Based on measured concentrations, the regional average uncorrected O_3 trend is

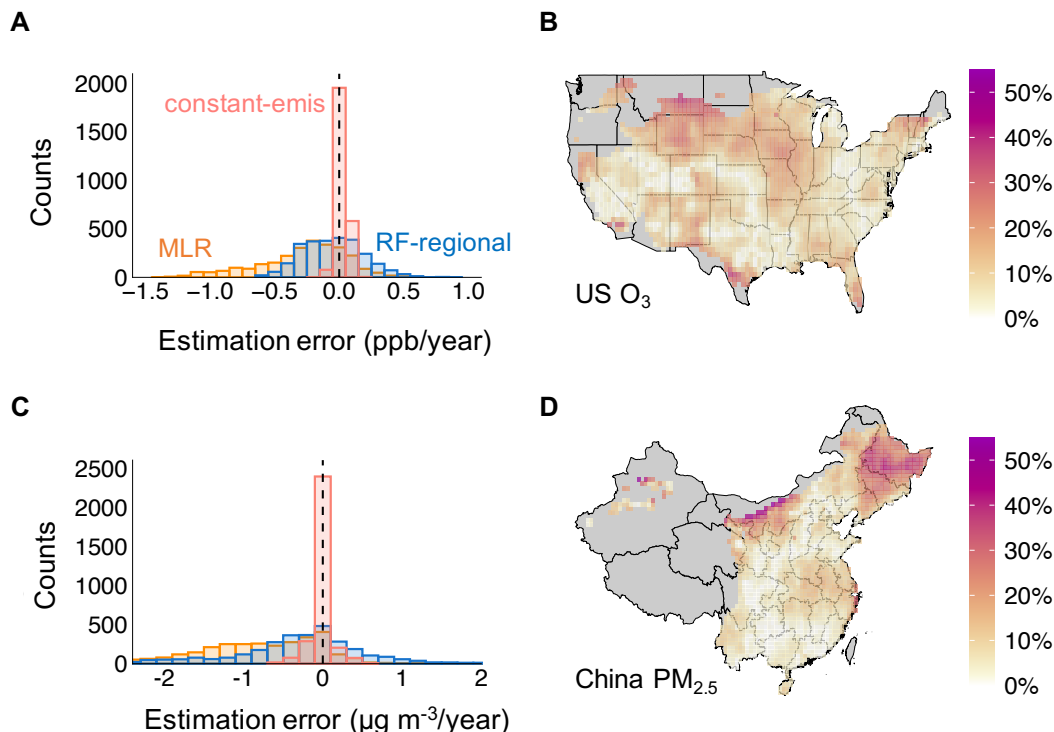


Figure 3. Panels A and C show the histogram of estimation errors in trend estimates assessed using MLR, RF-regional and constant-emis. Panels B and D show the percentage of the errors assessed with the constant-emis method relative to the trends in the counterfactual scenario ($|\beta^{gc} - \beta^{count}|/|\beta^{count}|$). Panels B and D only show grid cells with a trend in the counterfactual scenarios >0.2 ppb/year or $>0.2 \mu\text{g m}^{-3}/\text{year}$; remaining grid cells are shown in gray. Panels A and B illustrate the summer O₃ trends in the US. Panels C and D illustrate the annual PM_{2.5} trends in China.

330 -1.49 ppb/year and -1.15 ppb/year in Midwest and Southern US, respectively, which overestimates the reductions in concen-
 331 trations attributable to anthropogenic emissions changes. Correcting for the meteorological variability with MLR model yields
 332 regional average trend at -0.54 ppb/year in Midwest (a decrease by 53% in magnitude relative to uncorrected trends) and -0.71
 333 ppb/year in the Southern US (a decrease by 52%). RF-regional model further reduces the absolute magnitude of the declines
 334 in O₃ attributable to emissions reductions to -0.02 ppb/year for Midwest and -0.40 ppb/year for the Southern US. Importantly,
 335 these patterns are consistent with the results from our model experiments in these regions. For example in the GEOS-Chem
 336 simulation, the RF-regional model also estimates a much less negative emissions-driven trend in the Southern US compared to
 337 the uncorrected case and MLR estimates. For the GEOS-Chem simulations, RF-regional estimates are 39% smaller than MLR
 338 estimates, and this is comparable to the magnitude changes for the observational data (RF-regional estimates are 44% smaller
 339 than MLR). As the RF-regional model performs the best in recovering counterfactual trends in the GEOS-Chem simulations,

340 this suggests RF-regional may also perform the best in recovering the underlying emission-driven trends when applying to the
 341 observational data.

342 Figure 4B shows the trends in $\text{PM}_{2.5}$ concentrations estimated from the GEOS-Chem simulation and the observational data
 343 from China's surface monitoring network using different correction methods. Based on the observational data, our analysis
 344 reveals that the choice of methods for meteorological correction can yield very different results for certain regions. Much
 345 smaller reduction of $\text{PM}_{2.5}$ concentrations is attributed to anthropogenic emissions changes in the North, Northeast and East of
 346 China using the RF-regional model, relative to the MLR estimates. For example, the average emissions-driven trend estimated
 347 from the observational data is $-4.9 \mu\text{g m}^{-3}/\text{year}$ in Beijing under the RF-regional model, compared with $-9.6 \mu\text{g m}^{-3}/\text{year}$
 348 under the MLR model. These patterns are consistent with the patterns of the trend estimates estimated from our GEOS-Chem
 349 simulations with different statistical methods.

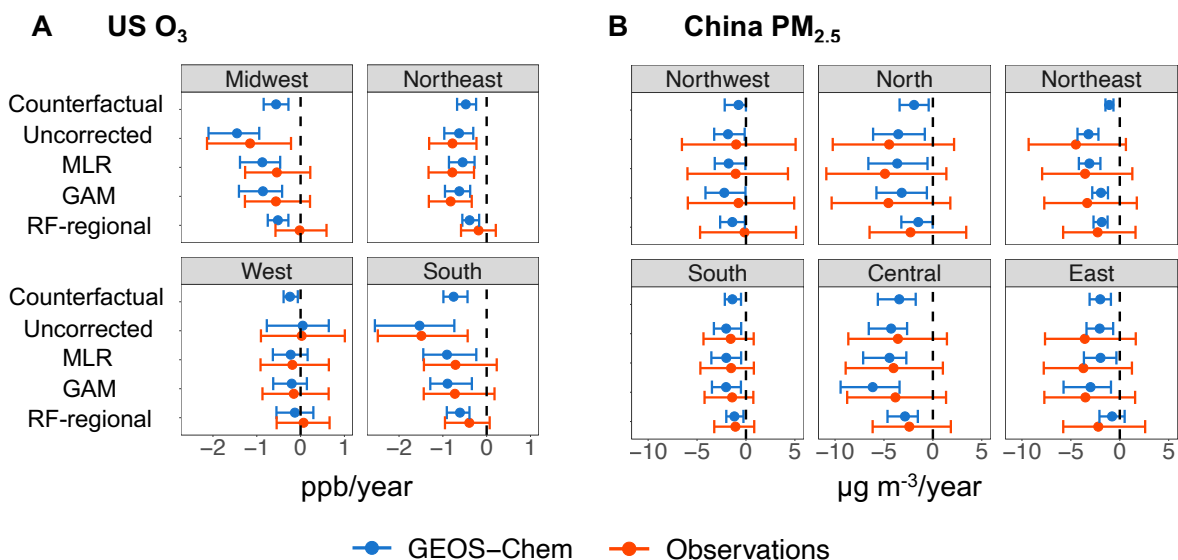


Figure 4. Trends in O_3 in the US (panel A) and $\text{PM}_{2.5}$ in China (panel B) estimated from the observational data (red) and GEOS-Chem simulations (blue) under different correction methods. Trends in pollutant concentrations are estimated at the monitor level (for the observational data) or at the grid cell level (for GEOS-Chem simulations). The point indicates the average value of the assessed trends of all monitors (or grid cells) within a region. The error bars show the 10th and 90th percentile of the assessed trends of all monitors/grid cells within a region. Panel A illustrates the summer O_3 trends in the US (unit: ppb/year). Panel B illustrates the annual $\text{PM}_{2.5}$ trends in China (unit: $\mu\text{g}/\text{m}^3/\text{year}$). We classify the US states into four regions according to the US Census Bureau and classify China's provinces into six regions based on the structure of China's subnational electric grid.

350 4 Discussion

351 We designed a model experiment that enables us to directly quantify the performance of different statistical models to evaluate
352 the causal trends in pollutant concentrations driven by anthropogenic emissions changes. Based on our evaluations of either
353 $\text{PM}_{2.5}$ or O_3 trends across US and China during periods of recent emission declines, our analysis shows that widely-used MLR
354 and GAM methods do not perform well in correcting for the meteorological variability and recovering simulated emissions-
355 driven trends. We propose a random forest model that uses both local and regional meteorological features, which offers the
356 best overall performance in recovering the emissions-driven trends across both species and countries. Applying this model to
357 observational data suggests that estimates based on MLR or similar methods may overestimate the impacts of anthropogenic
358 emissions changes on the decline of pollutant concentrations in certain regions in the US and China. However, the RF-regional
359 method does not outperform all the other approaches in every location despite its better overall performance (see figures S13
360 and S14). This suggests that using multiple statistical approaches may be necessary to derive robust conclusions for attributing
361 pollutant trends to emission changes.

362 With our model experiments, we also quantify the estimation errors in assuming the emission impacts can be perfectly
363 separated from the meteorological variability. These errors likely bound the estimation errors that can be achieved by any
364 statistical corrections of meteorological variability with this assumption. In the future, more complex statistical and machine
365 learning methods could be applied to distinguish emissions- and meteorologically-driven changes, but attribution solely based
366 on observed concentrations and meteorology will be limited by physical interactions between emissions and meteorology. We
367 find that the estimation errors resulting from these interactions are overall much smaller compared to the estimation errors
368 of the existing statistical methods, but can still be important for certain regions at certain times. Furthermore, the intertwined
369 relationships between emissions and meteorology are also much more complex in reality compared to our model experiments.
370 For example, meteorology can also directly influence anthropogenic emissions (e.g., increased electricity consumption during
371 extreme weather conditions (U.S. Energy Information Agency, 2019; He et al., 2020)). Therefore, the estimation errors that can
372 be achieved by more flexible statistical models can potentially be even bigger than the errors quantified with our constant-emis
373 approach.

374 While the GEOS-Chem model provides us with a framework for causal experiments to test statistical methods, its use
375 in our model experiments introduces some uncertainty and limitations. Specifically, our experiments assess the performance
376 of statistical methods in correcting for the meteorology-pollution relationships encoded in GEOS-Chem, which may differ
377 from the complex relationships observed in the observational data. Several studies have shown that GEOS-Chem and similar
378 models do not capture certain meteorology-pollution relationships in the observational data (e.g., temperature - O_3 relationship

379 (Porter and Heald (2019)) and influence of regional meteorological patterns (Fiore et al. (2009))). The relationships encoded
380 in GEOS-Chem may be different from the underlying meteorology-pollution relationships in the following three ways: (1)
381 parameters in GEOS-Chem that describe these relationships are uncertain; (2) the relationships in GEOS-Chem are incorrect
382 or incomplete; and (3) the relationships in GEOS-Chem are deterministic compared to the potential stochastic underlying
383 processes. While the parameterization schemes of the model may have little impact on our assessment of the statistical methods
384 if the functional forms are correct, different functional forms may affect the relative performance of various statistical methods.
385 The performance of any individual statistical method is likely to be worse in the real world compared to its ability to reproduce
386 a deterministic meteorology-pollution relationship encoded in GEOS-Chem. Further model-based experiments could apply our
387 methods to different atmospheric models in order to test if these conclusions differ by different models.

388 Our research reveals multiple directions for future research to enhance our understanding of the usage of statistical models
389 to evaluate trends in pollutant concentrations under changing meteorological conditions. One key but challenging question
390 is to better understand the estimation errors of these existing approaches, e.g. why the MLR model is able to correct for
391 the meteorological variability in some locations but not others. In this paper, we only test a selection of methods based on
392 their popularity in the existing literature and propose a simple-to-use model (RF-regional). More complex models (such as
393 convolutional neural networks) may offer better performance, but the estimation error will likely be bounded by the errors of the
394 constant-emis approach. Our work only evaluates the statistical and machine learning models in expressions 1 and 2, which only
395 represent one (popular) set of evaluations that performs location-specific trend estimation with adjustments for meteorology and
396 secular trends. However, other statistical model specifications specifically targeted to questions of meteorological interaction
397 or that permit borrowing information across locations may generate different results. A deeper investigation of the estimation
398 error due to assuming perfect separation between meteorology and emission is also essential for understanding how we should
399 interpret studies that use these statistical methods. For example, further work could explore how these errors will vary by
400 the magnitude of emissions reductions and the chemistry regimes. Our analysis suggests the relative performance of different
401 methods is largely similar in monitoring data and the GEOS-Chem experiments (at least for certain regions). It is interesting to
402 further explore how the patterns of performance might differ across different types of monitor locations and conditions.

403 **5 Recommendations for attributing trends to emissions changes**

404 Using statistical methods to causally infer relationships between simulated air pollutant concentrations and anthropogenic
405 emissions is challenging, not to mention understanding the drivers of observed air pollutants in the real world. Understanding
406 the uncertainty of statistical models in characterizing the meteorology-pollution relationship is essential to evaluating the

effectiveness of policy interventions with observational data. Here, we make several recommendations to researchers and policy makers based on our analysis.

For those who aim to infer causal effects of emissions changes on air quality based on observational data on concentrations and meteorology, we recommend using multiple statistical methods to correct for the meteorological variability when evaluating the impacts of policies or interventions on air quality. From our two case studies, we find a relatively large variability between the trend parameters estimated by different statistical methods (especially at the grid cell or monitor level). Some methods perform better in certain locations but not in others (though RF-regional is the best-performing method overall). Using multiple approaches (linear/non-linear and at local/regional scale) may help to quantify uncertainty related to meteorology corrections. These findings also suggest that empirical analyses may benefit from considering the impacts of meteorological variability on air quality separately for each region or even for each monitor location (if data permits), instead of attempting to determine a general relationship between meteorological variability and air pollution over a large spatial domain. Finally, analysts should be particularly cautious when using statistical methods to estimate impacts of anthropogenic emissions on air quality in regions where pollution variability is dominated by meteorologically-influenced environmental processes such as dust emissions, as we consistently show that typical statistical methods (in combination with the standard set of meteorological variables) do not work well in those regions.

Due to the non-negligible estimation errors in recovering the counterfactual trends even with the best-performed statistical approach we test, we believe these statistical analyses are most useful in understanding the patterns of anthropogenic emissions on air quality when aggregated across larger spatial areas, rather than providing specific trends for individual monitor locations. There is a higher degree of consistency among the trend estimates across different methods when aggregated at regional level, but assessment at local level is more sensitive to method choices. The absolute magnitude of monitor-level trends need to be interpreted with caution, considering both the uncertainty from the statistical methods and also the limit of meteorological correction due to ignoring the interactions between meteorology and emissions.

Because measured pollutant concentrations are subject to the influence of underlying meteorological variability, many efforts have attempted to correct for the impacts of meteorological variability and use “meteorology-corrected” concentrations and trends to assist in evaluating the effectiveness of air quality policies. Our study evaluates existing methods that aim to correct for the meteorological variability and finds many of these methods do not perform well. This raises potential concerns about the use of “meteorology-corrected” concentrations as targets for policy evaluation. Meteorology-corrected concentrations and trends remain useful metrics to quantify the influence of emissions. However, a more comprehensive evaluation of the effectiveness of policy requires interpreting measurements with all available tools, ideally including both statistical analyses and physical models.

437 *Code and data availability.* The GEOS-Chem simulation of different scenarios, code for different statistical methods and monitor-level trend
438 estimates will be made available to readers in a public repository.

439 *Author contributions.* M.Q. and N.E.S. designed the research. M.Q. performed the statistical analysis and GEOS-Chem modeling simula-
440 tions. All authors interpreted the results and wrote the paper.

441 *Competing interests.* The authors declare no competing interests.

442 *Acknowledgements.* We thank Colette Heald and Valerie Karplus for helpful comments and discussions. We thank Yixuan Zheng for assis-
443 tance with the MEIC emissions inventory. We thank Ke Li for sharing code of step-wise MLR analysis. This publication was supported by
444 US EPA grant RD-835872-01. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of
445 the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

446 Appendix: Supplementary methods

447 Implementation of LASSO and RF

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

$$453 \quad y_{it} = \beta_i^{obs} \times t + g_i(X_{it}, Z_t, W_t) + \epsilon_{it} \quad (1)$$

454 where $g_i()$ denotes the functional form fitted by LASSO or RF. X_{it} again denotes the local meteorology features for grid cell
455 i on day t . Z_t denotes the regional scale meteorology features including the meteorological features for all grid cells in the
456 US on day t (98 cells in 4×5 degrees; we choose a relatively coarse resolution due to computational cost). Meteorological
457 information in each location in the US may help explain the pollutant concentrations in grid cell i . In total, we have 10 local
458 features (X_{it}) and $10 \times 98 = 980$ regional scale features (Z_t). W_t denotes the day and month variable to model the daily and
459 monthly variability in pollutant that are unrelated to meteorological variability. For LASSO, we use month-of-year \times day-of-
460 month fixed effect (same as all the other methods except for RF), and these fixed effects are not penalized in the LASSO
461 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
462 inefficient performance of RF working with large number of fixed effects. Thus, the difference between RF and the other
463 methods may also come from the different choice of modeling monthly and daily variability.

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

466 (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
467 remaining 25% for predictions. We train the following two models on the training data:

$$468 \quad y_{it} = f(X_{it}, Z_t, W_t)$$

$$469 \quad t = g(X_{it}, Z_t, W_t)$$

470 (2) We then apply models $f(\cdot)$ and $g(\cdot)$ to the prediction set to get predictions of y_{it} and t for the rest 25% of the data. The
 471 above process is repeated four times to derive predictions for the entire time series (predictions denoted as \widehat{y}_{it} and \widehat{t}).
 472 (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
 473 as:

$$474 \quad \beta_i^{obs} = \frac{\sum_t \widetilde{t} \widetilde{y}_{it}}{\sum_t \widetilde{t} \widetilde{t}}$$

475 this is equivalent to setting up a linear regression of $\widetilde{y}_{it} \sim \widetilde{t}$ and obtain the slope coefficients (as shown by Chernozhukov et al.
 476 (2018)).

477 The hyper-parameters of RF and LASSO are tuned with 4-fold cross validation. We also perform two sensitivity analyses: 1)
 478 with a different spatial resolution of the regional scale features (2×2.5 degrees instead of 4×5 degrees), and 2) with different
 479 numbers of folds to estimate the trend coefficients. Our results are similar across these sensitivity analyses (see figure S15).

480 The double machine learning framework involves a sample partition procedure (steps (1) and (2) above). This procedure,
 481 however, does not fit the purpose of including time fixed effects in the LASSO model (as randomly partitioned training and
 482 test sets could have very unbalanced number of observations from a given month-day pair). Therefore, step (1) and (2) are only
 483 implemented for the RF model, and coefficients of the LASSO model is directly derived from step (3) without sample splitting.
 484 This is okay for the LASSO model as the risk of “overfitting” has already been eliminated by using the tuned penalizing factor
 485 (i.e. the hyper-parameters) derived from a 4-fold cross-validation.

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.

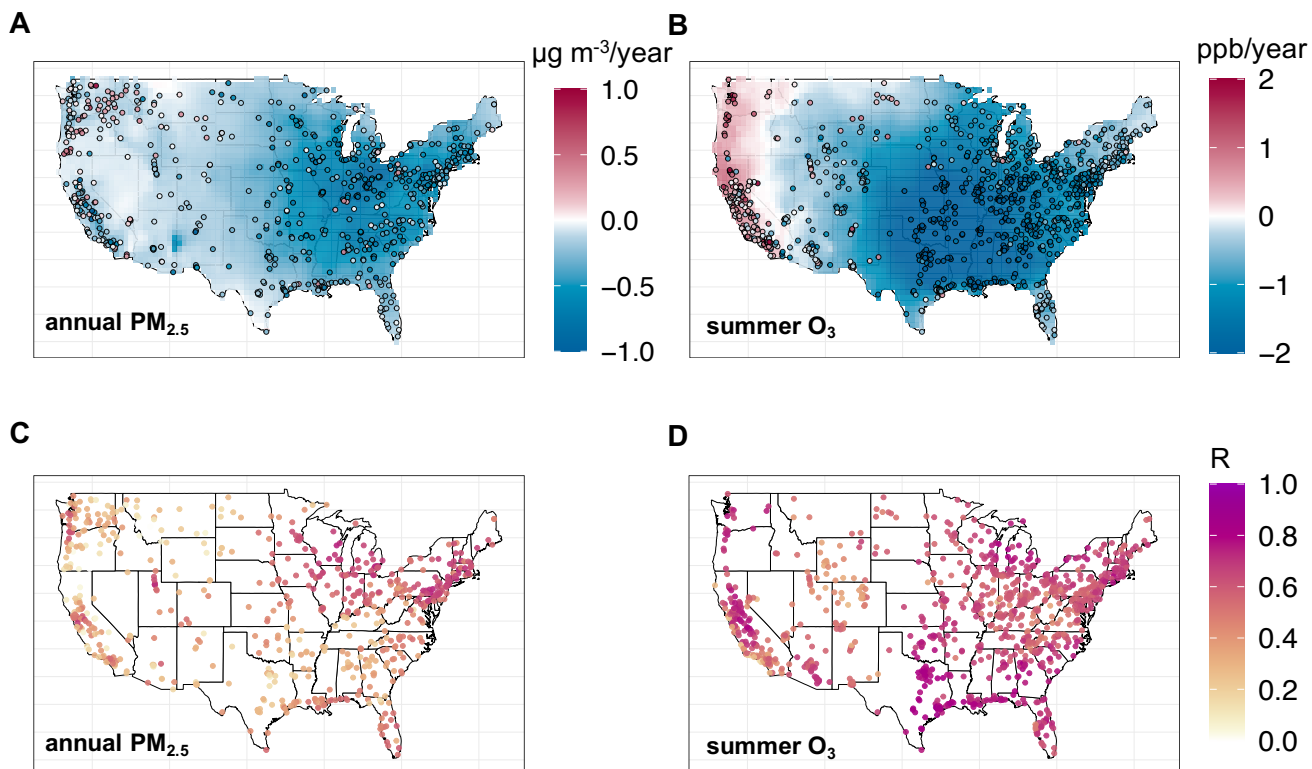


Figure S1. Comparison between the annual $\text{PM}_{2.5}$ (Panels A and C) and summer O_3 (Panels B and D) concentrations measured by the monitoring network and GEOS-Chem simulations in the US (2011-2017). Panels A and B show the trends in monitored concentrations (dots) and trends in the observational 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.

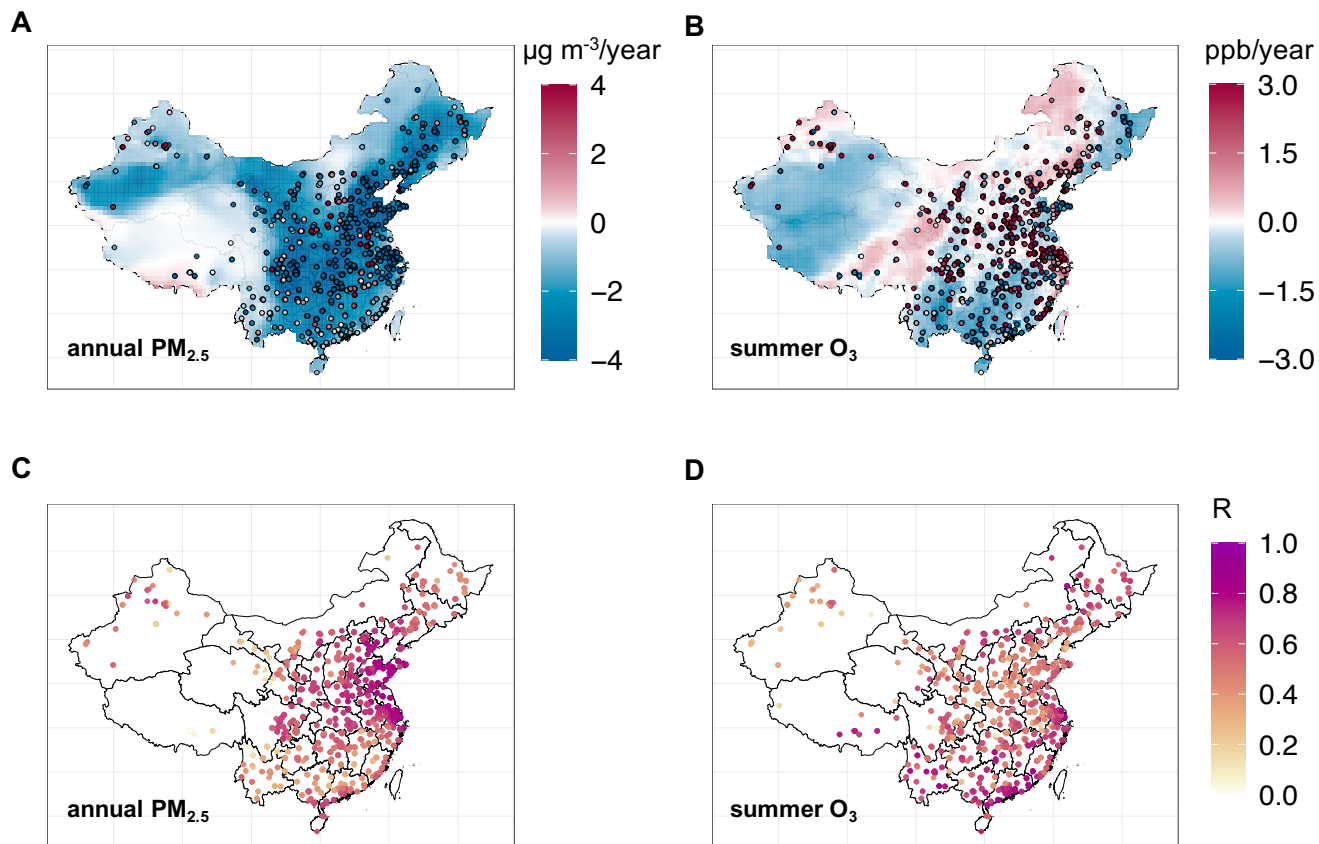


Figure S2. Comparison between the annual $\text{PM}_{2.5}$ (Panels A and C) and summer O_3 (Panels B and D) concentrations measured by the surface monitoring network and GEOS-Chem simulations in China (2014-2017). Panels A and B show the trends in monitored concentrations (dots) and trends in the observational 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.

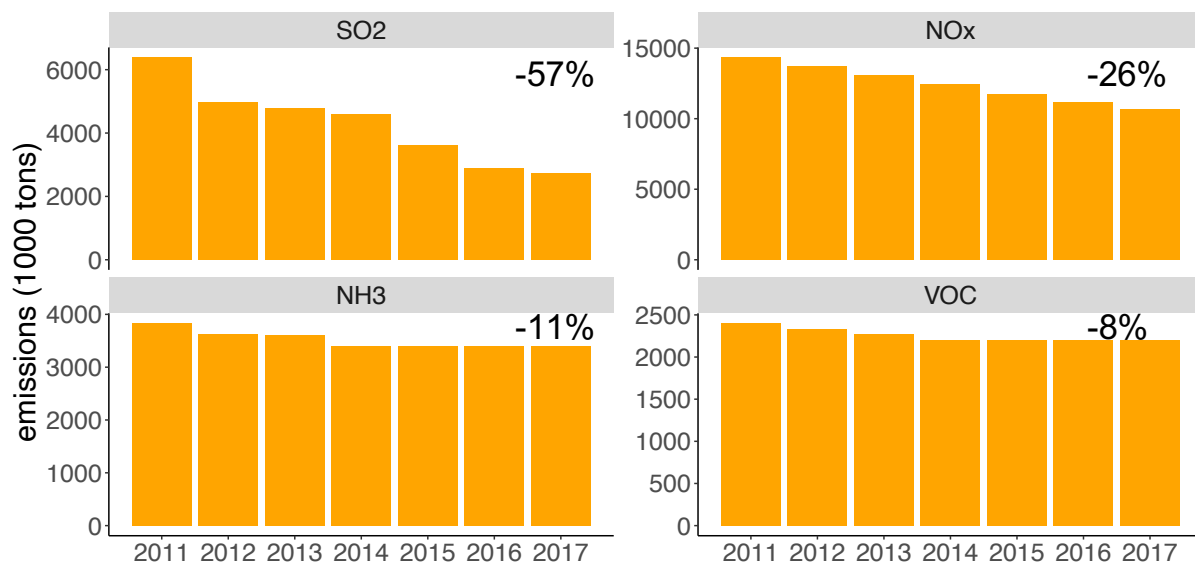


Figure S3. National total anthropogenic emissions in the US (2011- 2017). The emissions data is derived from the national total emissions of criterion air pollutants reported by the US EPA Air Emissions Inventory.

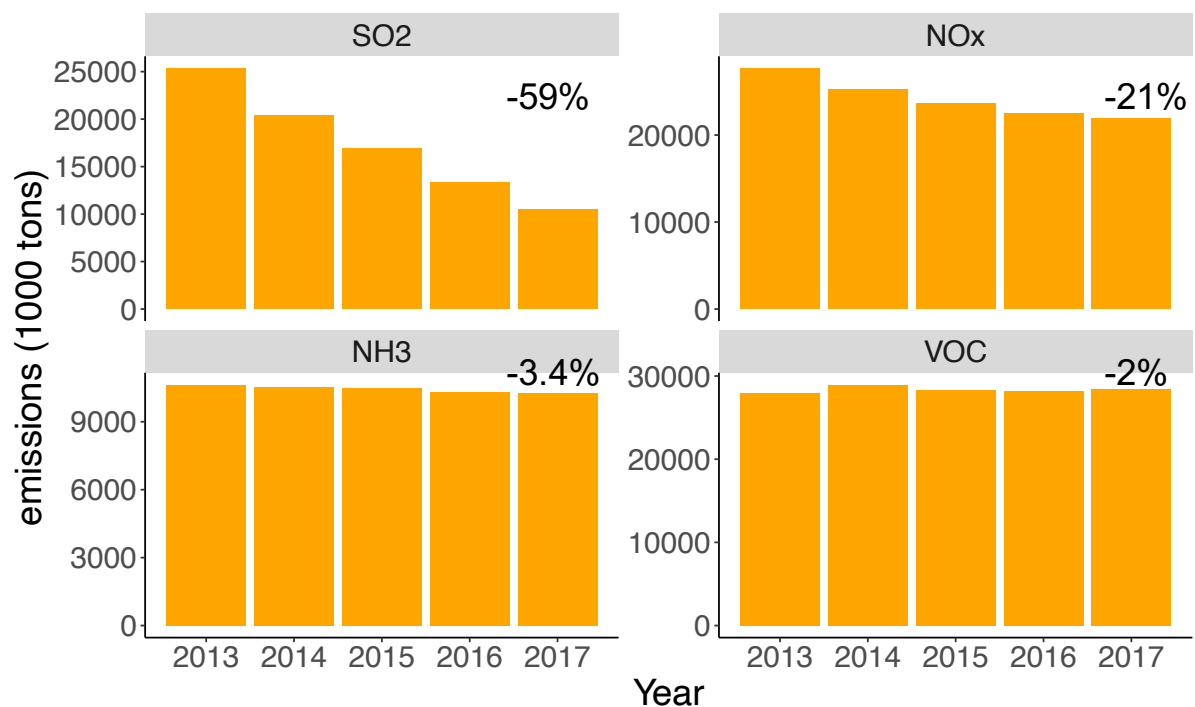
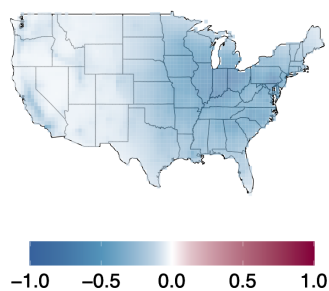
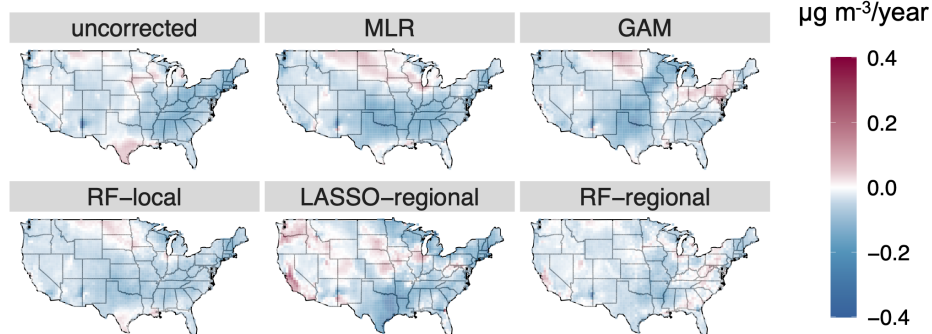


Figure S4. National total anthropogenic emissions in China (2013- 2017). The emissions data is derived from the Multi-resolution Emission Inventory (MEIC).

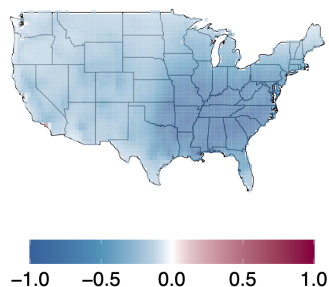
A Counterfactual PM_{2.5} trends



B Errors in PM_{2.5} trend estimates



C Counterfactual O₃ trends



D Errors in O₃ trend estimates

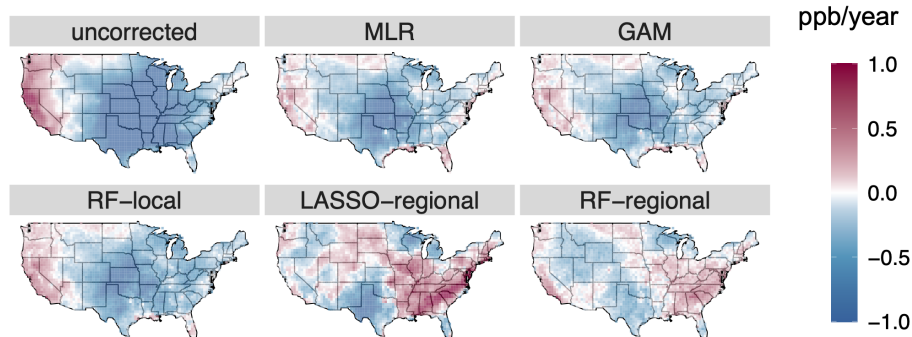


Figure S5. Trend estimates of daily annual PM_{2.5} (Panels A and B) and summer O₃ (C and D) in the US. Panels A and C show trend estimates under the counterfactual scenario (β^{count}). Panels B and D show the estimation errors of trend estimates under different correction methods compared with the 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 $\mu\text{g m}^{-3}/\text{year}$ for PM_{2.5} or ppb/year for O₃.

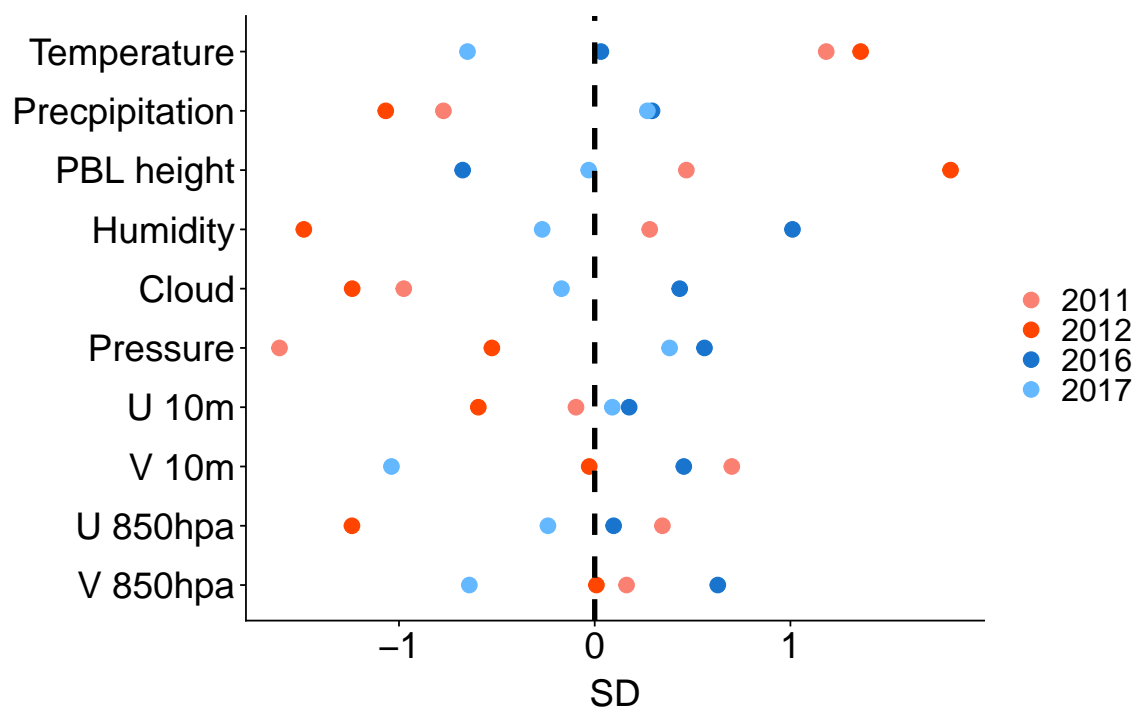
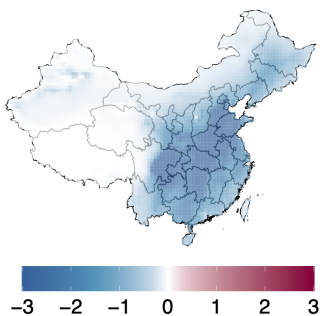
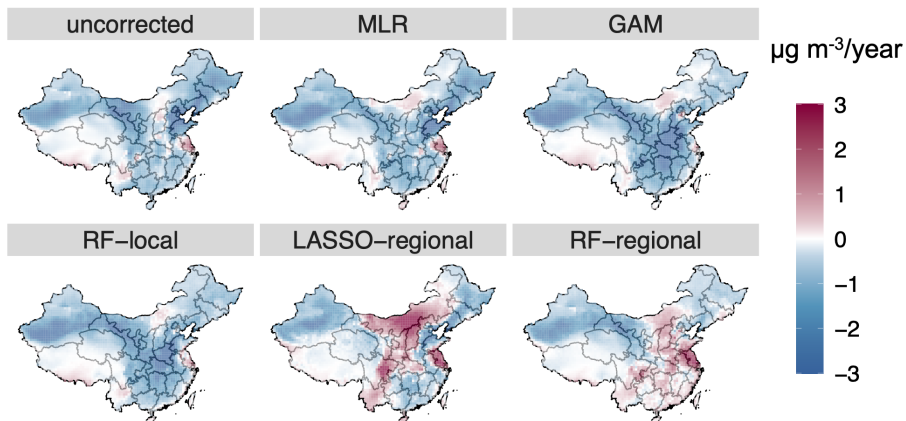


Figure S6. Deviations of meteorological features from the 7-year average in the US (South and Midwest). The deviation is quantified in the units of standard deviation (SD) across the 7-year period. Zero indicates the 7-year average. This plot shows the summer time average of daily MDA8 meteorological variables for each year aggregated over South and Midwest US.

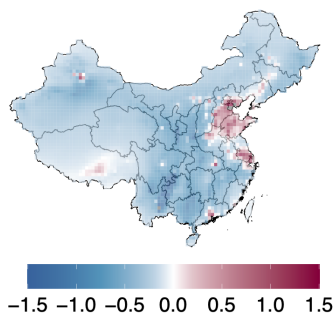
A Counterfactual PM_{2.5} trends



B Errors in PM_{2.5} trend estimates



C Counterfactual O₃ trends



D Errors in O₃ trend estimates

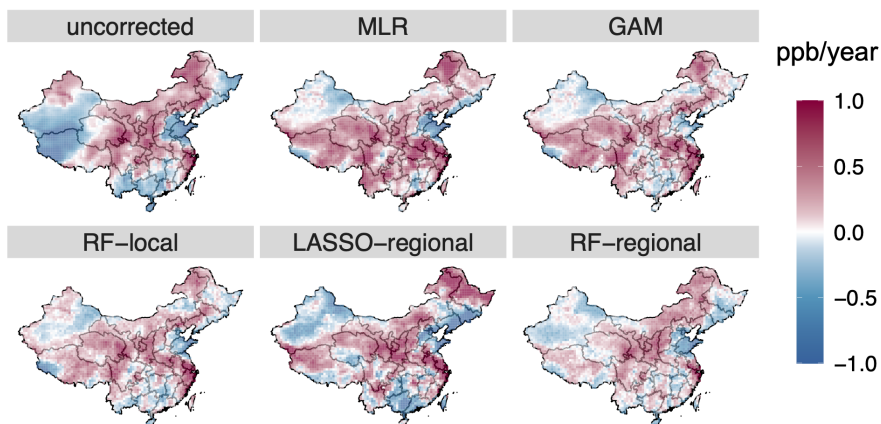


Figure S7. Trend estimates of daily annual PM_{2.5} (Panels A and B) and summer O₃ (C and D) in China. Panels A and C show trend estimates under the counterfactual scenario (β^{count}). Panels B and D show the estimation errors of trend estimates under different correction methods compared with the counterfactual scenarios ($\beta^{obs} - \beta^{count}$). The average of the absolute error for each method is shown in the figure. Unit of trend estimate is $\mu\text{g m}^{-3}/\text{year}$ for PM_{2.5} or ppb/year for O₃.

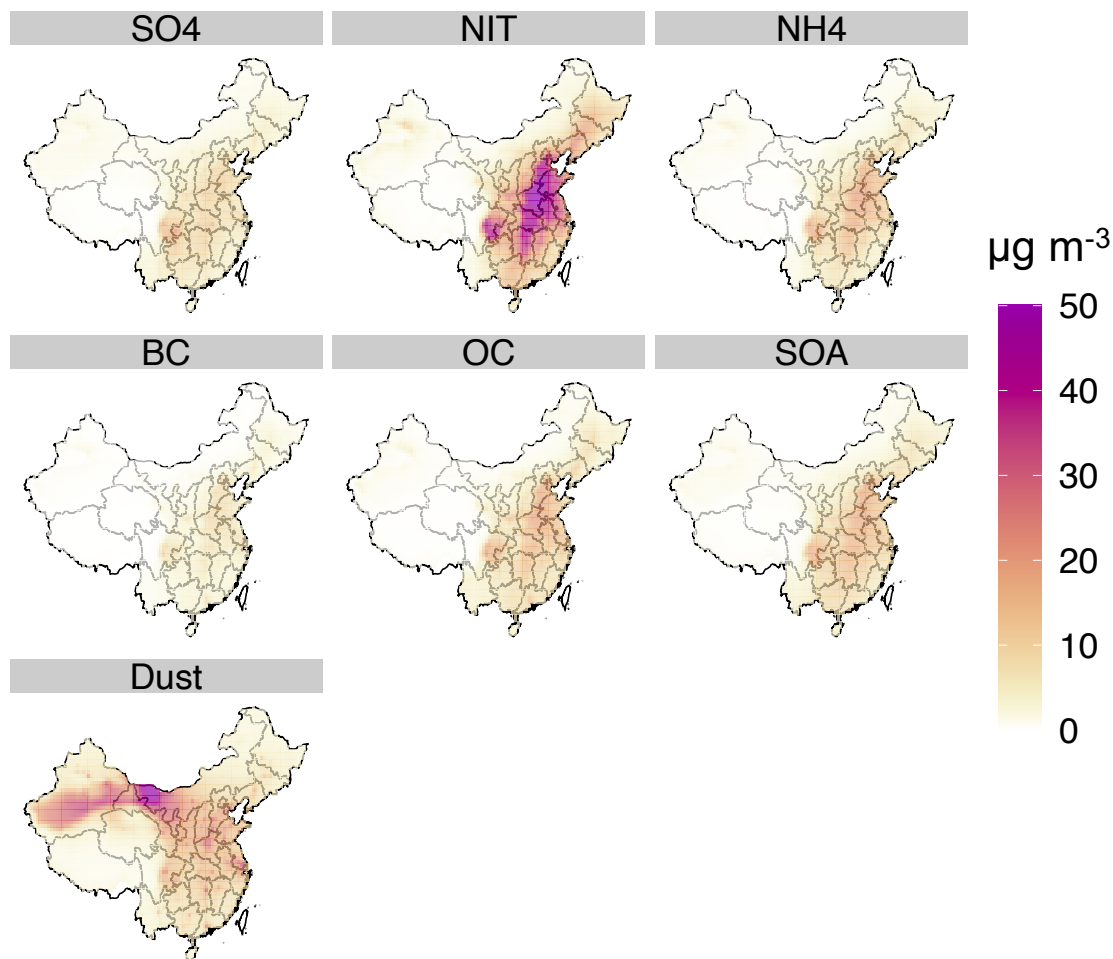


Figure S8. Concentrations of component species of PM_{2.5} in China (average across 2013-2017). The figure shows concentrations of sulfate (SO₄), nitrate (NIT), ammonium (NH₄), black carbon (BC), organic carbon (OC), secondary organic aerosol (SOA), and dust.

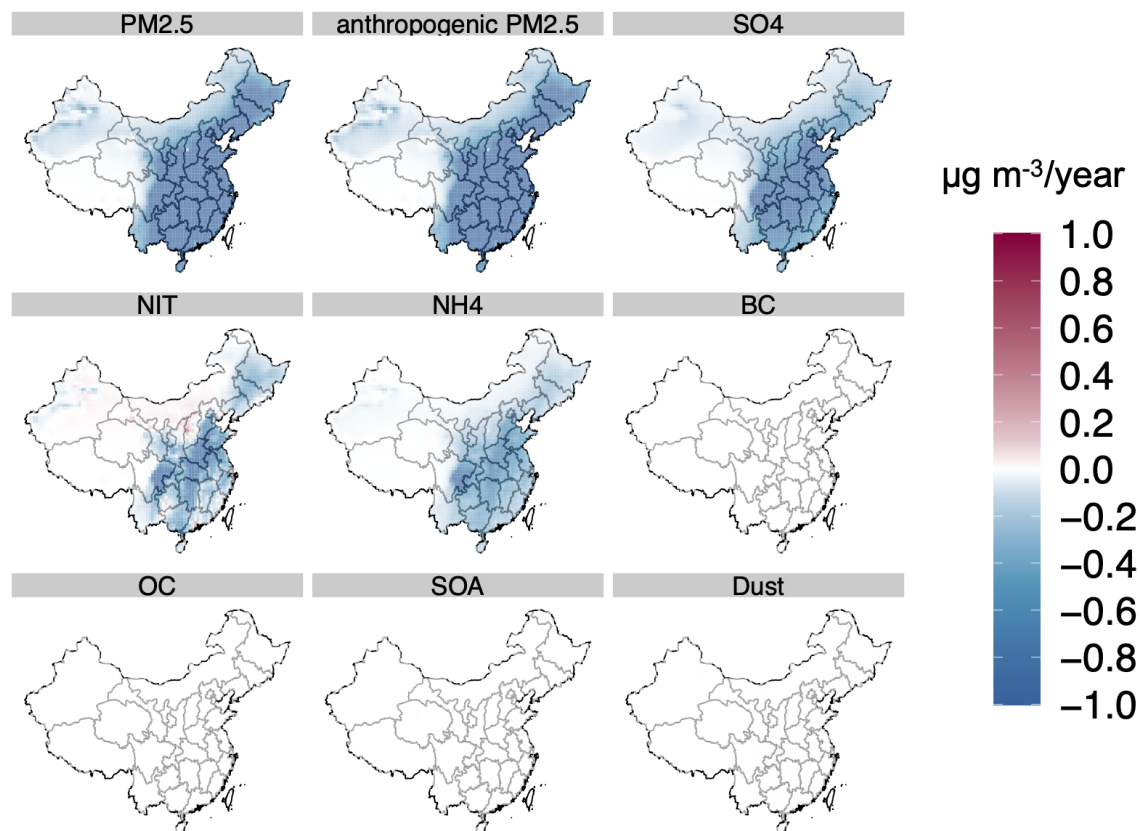


Figure S9. Counterfactual trends of component species of $\text{PM}_{2.5}$ in China. The figure shows counterfactual trends of total $\text{PM}_{2.5}$, anthropogenic $\text{PM}_{2.5}$ (total $\text{PM}_{2.5}$ excluding dust and sea salt), sulfate (SO_4), nitrate (NIT), ammonium (NH_4), black carbon (BC), organic carbon (OC), secondary organic aerosol (SOA), and dust.

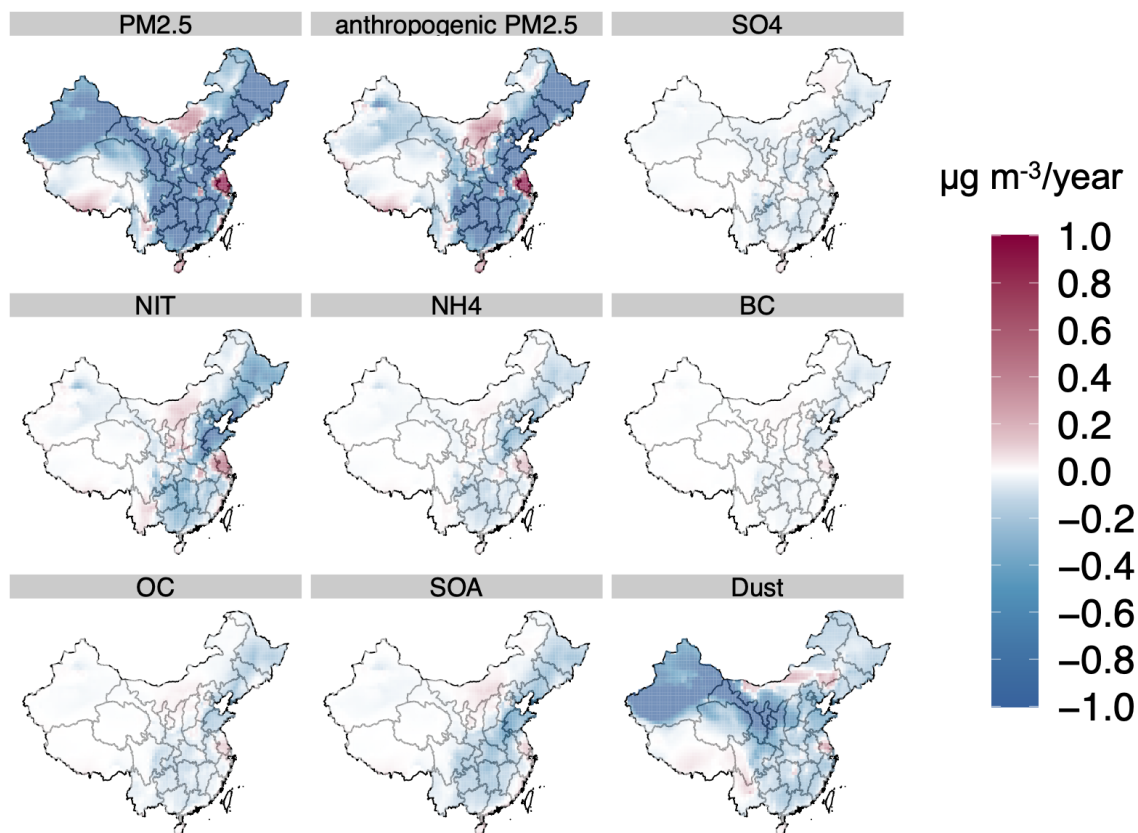


Figure S10. Differences between counterfactual trends and trends evaluated under MLR ($\beta^{MLR} - \beta^{count}$) of component species of PM_{2.5} in China. The figure shows estimation errors of total PM_{2.5}, anthropogenic PM_{2.5} (total PM_{2.5} excluding dust and sea salt), sulfate (SO₄), nitrate (NIT), ammonium (NH₄), black carbon (BC), organic carbon (OC), secondary organic aerosol (SOA) and dust.

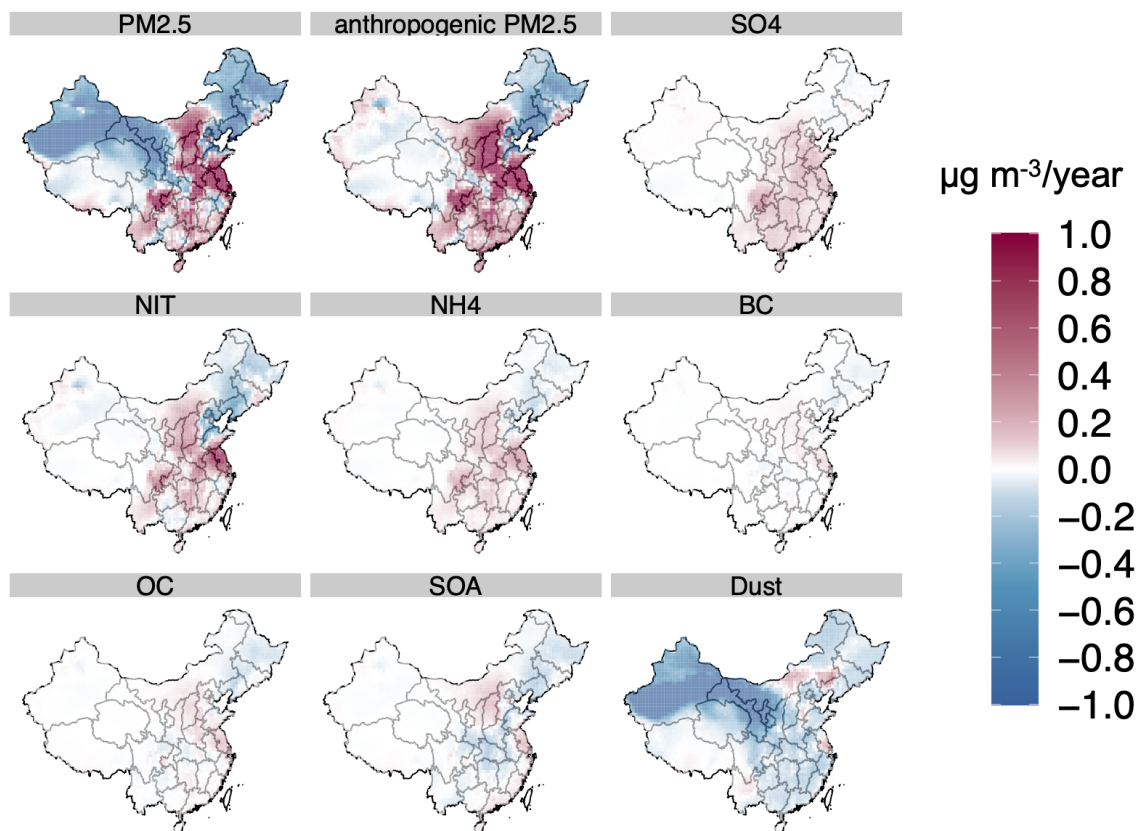


Figure S11. Differences between counterfactual trends and trends evaluated under RF-regional ($\beta^{RF-regional} - \beta^{count}$) of component species of PM_{2.5} in China. The figure shows estimation errors of total PM_{2.5}, anthropogenic PM_{2.5} (total PM_{2.5} excluding dust and sea salt), sulfate (SO₄), nitrate (NIT), ammonium (NH₄), black carbon (BC), organic carbon (OC), secondary organic aerosol (SOA) and dust.

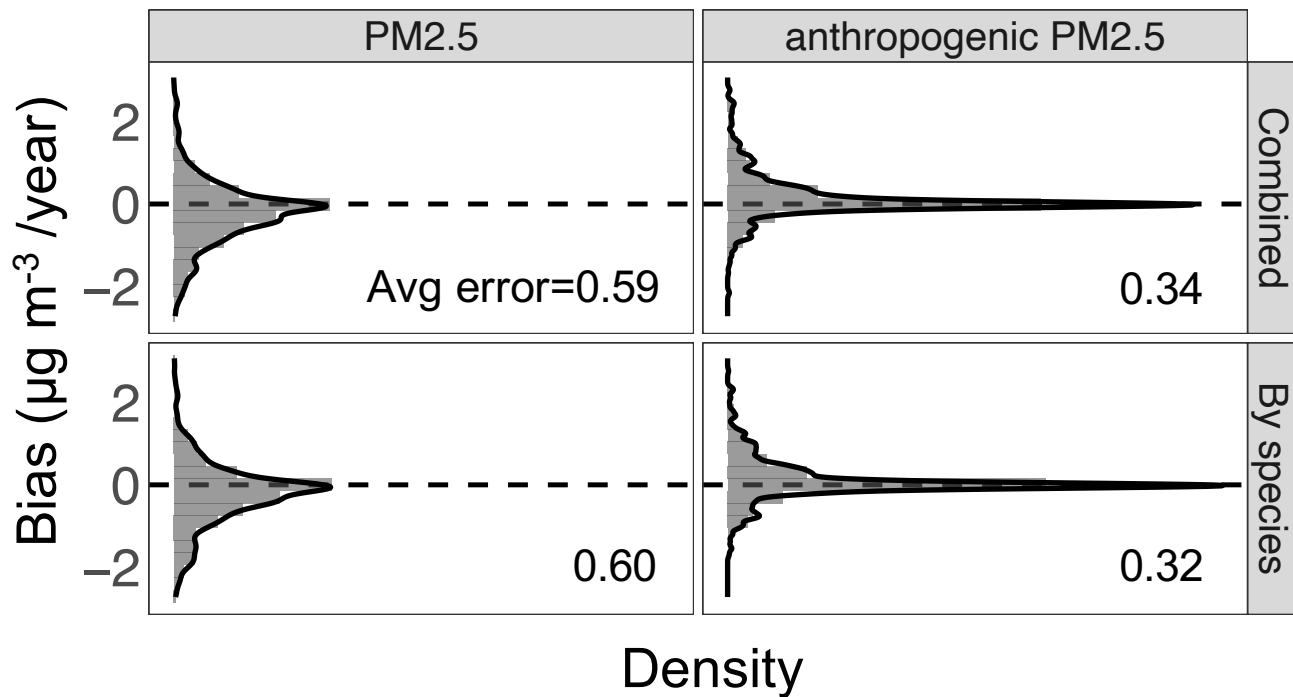
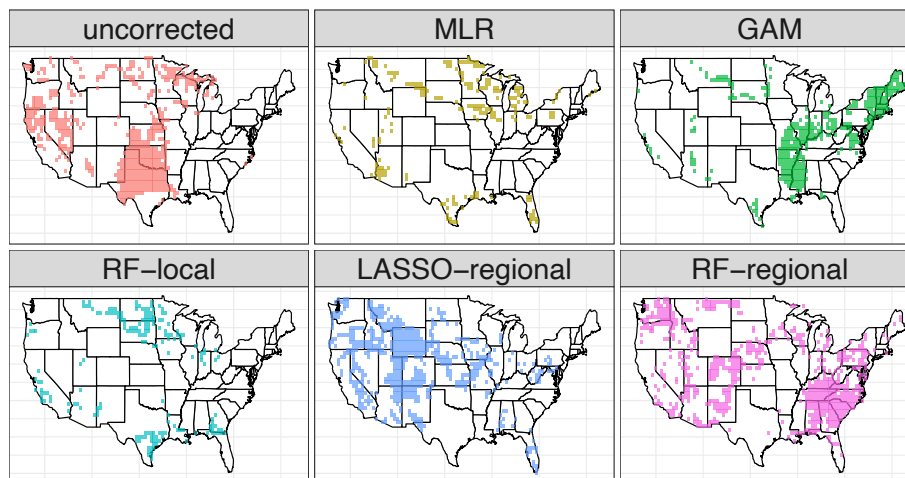


Figure S12. Histograms of estimation errors of trend estimates under two implementations of the *RF-regional* method (China PM_{2.5}). The upper panels (Combined) show results of fitting RF models to the combined concentrations of PM_{2.5} to directly estimate trends (the main results). The lower panels (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 for each implementation is shown in the figure.

A annual $\text{PM}_{2.5}$



B summer O_3

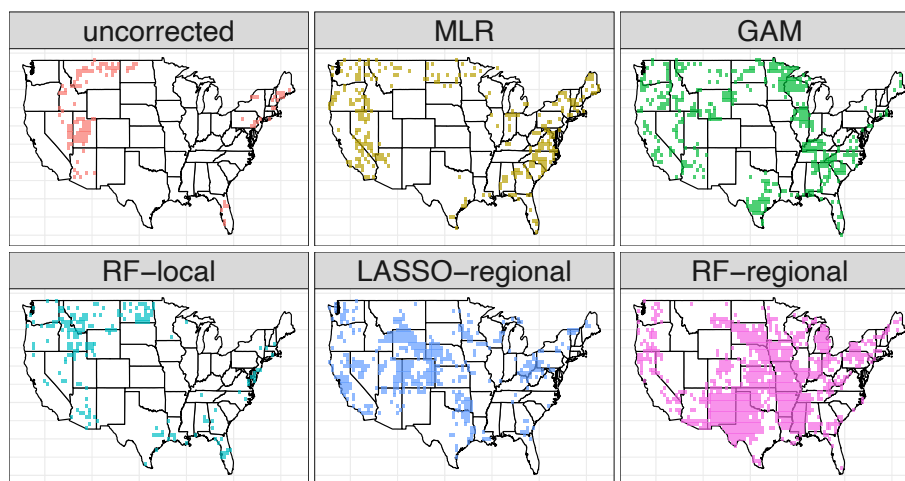
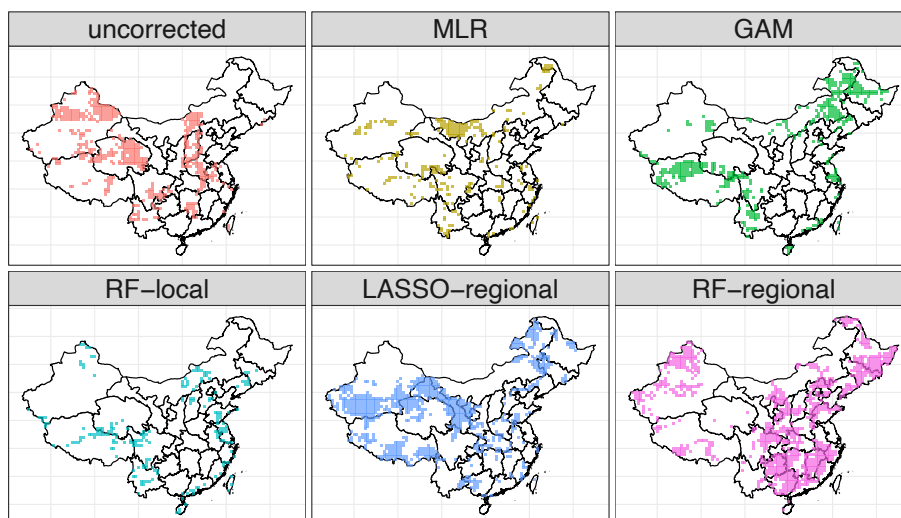


Figure S13. Best-performing correction method for each grid cell (US). For each method, the figure shows the grid cells at which the trend estimate has the smallest estimation error (i.e. closest to the trend in the counterfactual scenario) among the tested methods.

A annual $\text{PM}_{2.5}$



B summer O_3

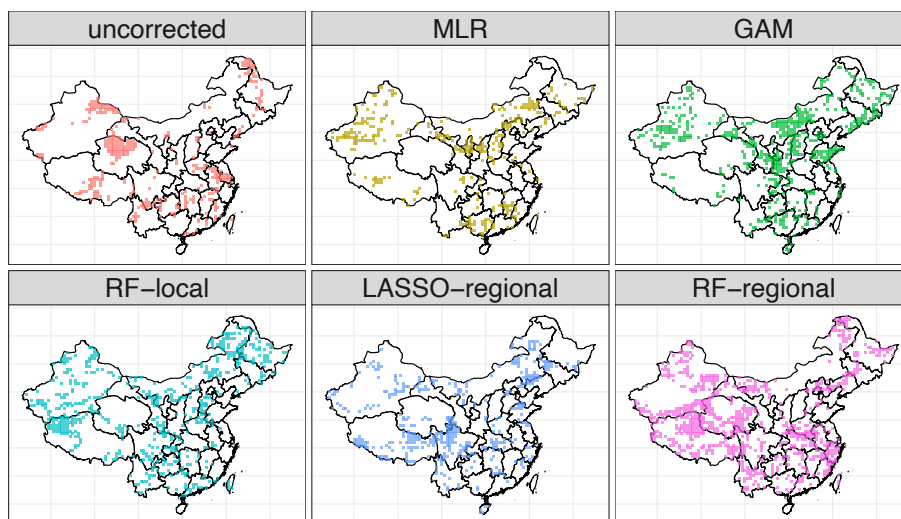


Figure S14. Best-performing correction method for each grid cell (China). For each method, the figure shows the grid cells at which the trend estimate has the smallest estimation error (i.e. closest to the trend in the counterfactual scenario) among the tested methods.

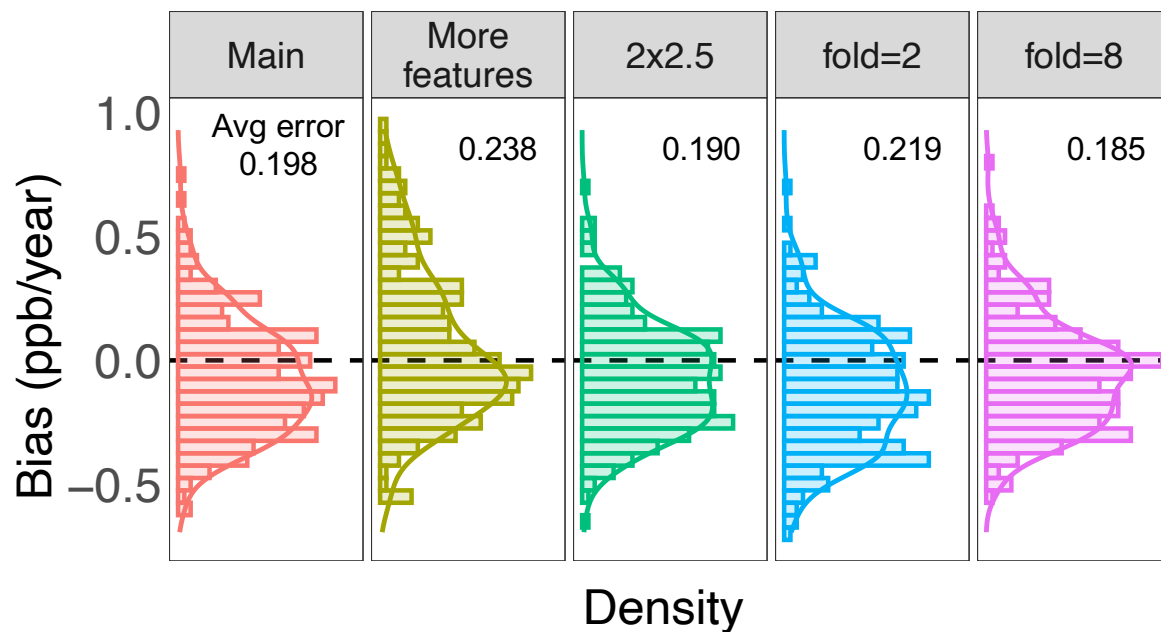


Figure S15. Histograms of estimation errors of trend estimates under different implementations of the *RF-regional* method (US O–3). From left to right: Main (the main results), More features (including 9 extra meteorological features), 2x2.5 (using regional features with spatial resolution of $2 \times 2.5^\circ$, instead of $4 \times 5^\circ$), 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.

487 **References**

- 488 Beijing Municipal Ecology and Environment Bureau: Beijing Clean Air Action Plan (2013–2017), 744, 140 837, [http://sthjj.beijing.gov.cn/](http://sthjj.beijing.gov.cn/bjhrb/index/xxgk69/sthjlyzwg/wrygl/603133/index.html)
 489 [bjhrb/index/xxgk69/sthjlyzwg/wrygl/603133/index.html](http://sthjj.beijing.gov.cn/bjhrb/index/xxgk69/sthjlyzwg/wrygl/603133/index.html), 2013.
- 490 Bey, I., Jacob, D. J., Yantosca, R. M., Logan, J. A., Field, B. D., Fiore, A. M., Li, Q., Liu, H. Y., Mickley, L. J., and Schultz, M. G.: Global
 491 modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation, *Journal of Geophysical Research*
 492 *Atmospheres*, 106, 23 073–23 095, <https://doi.org/10.1029/2001JD000807>, 2001.
- 493 Camalier, L., Cox, W., and Dolwick, P.: The effects of meteorology on ozone in urban areas and their use in assessing ozone trends, *Atmo-*
 494 *spheric Environment*, 41, 7127–7137, <https://doi.org/10.1016/j.atmosenv.2007.04.061>, 2007.
- 495 Carslaw, D. C., Beevers, S. D., and Tate, J. E.: Modelling and assessing trends in traffic-related emissions using a generalised additive
 496 modelling approach, *Atmospheric Environment*, 41, 5289–5299, <https://doi.org/10.1016/j.atmosenv.2007.02.032>, 2007.
- 497 Chen, L., Zhu, J., Liao, H., Yang, Y., and Yue, X.: Meteorological influences on PM_{2.5} and O₃ trends and associated health burden since
 498 China’s clean air actions, *Science of The Total Environment*, 744, 140 837, 2020.
- 499 Chen, Z., Chen, D., Kwan, M. P., Chen, B., Gao, B., Zhuang, Y., Li, R., and Xu, B.: The control of anthropogenic emissions contributed
 500 to 80% of the decrease in PM_{2.5} concentrations in Beijing from 2013 to 2017, *Atmospheric Chemistry and Physics*, 19, 13 519–13 533,
 501 <https://doi.org/10.5194/acp-19-13519-2019>, 2019.
- 502 Cheng, J., Su, J., Cui, T., Li, X., Dong, X., Sun, F., Yang, Y., Tong, D., Zheng, Y., Li, Y., Li, J., Zhang, Q., and He, K.: Dominant role of
 503 emission reduction in PM_{2.5} air quality improvement in Beijing during 2013–2017: A model-based decomposition analysis, *Atmospheric*
 504 *Chemistry and Physics*, 19, 6125–6146, <https://doi.org/10.5194/acp-19-6125-2019>, 2019.
- 505 Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J.: Double/debiased machine learning for
 506 treatment and structural parameters, 2018.
- 507 China’s Ministry of Ecology and Environment: National Air Quality Monitoring Data, <https://quotsoft.net/air/>, 2021.
- 508 European Union: Air Quality Standards in the European Union, 744, 140 837, <https://ec.europa.eu/environment/air/quality/standards.htm>,
 509 2020.
- 510 Fiore, A. M., Dentener, F., Wild, O., Cuvelier, C., Schultz, M., Hess, P., Textor, C., Schulz, M., Doherty, R., Horowitz, L., et al.: Multimodel
 511 estimates of intercontinental source-receptor relationships for ozone pollution, *Journal of Geophysical Research: Atmospheres*, 114, 2009.
- 512 Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R.,
 513 et al.: The modern-era retrospective analysis for research and applications, version 2 (MERRA-2), *Journal of climate*, 30, 5419–5454,
 514 2017.
- 515 Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., and Hueglin, C.: Random forest meteorological normalisation models for Swiss
 516 PM₁₀ trend analysis, *Atmospheric Chemistry and Physics*, 18, 6223–6239, <https://doi.org/10.5194/acp-18-6223-2018>, 2018.
- 517 Han, H., Liu, J., Shu, L., Wang, T., and Yuan, H.: Local and synoptic meteorological influences on daily variability in summertime surface
 518 ozone in eastern China, *Atmospheric Chemistry and Physics*, 20, 203–222, <https://doi.org/10.5194/acp-20-203-2020>, 2020.

519 Hayn, M., Beirle, S., Hamprecht, F. A., Platt, U., Menze, B. H., and Wagner, T.: Analysing spatio-temporal patterns of the global NO₂-
520 distribution retrieved from GOME satellite observations using a generalized additive model, *Atmospheric Chemistry and Physics*, 9,
521 6459–6477, <https://doi.org/10.5194/acp-9-6459-2009>, 2009.

522 He, P., Liang, J., Qiu, Y. L., Li, Q., and Xing, B.: Increase in domestic electricity consumption from particulate air pollution, *Nature Energy*,
523 5, 985–995, 2020.

524 Heald, C. L., Collett, J. L., Lee, T., Benedict, K. B., Schwandner, F. M., Li, Y., Clarisse, L., Hurtmans, D. R., Van Damme, M., Clerbaux, C.,
525 Coheur, P. F., Philip, S., Martin, R. V., and Pye, H. O.: Atmospheric ammonia and particulate inorganic nitrogen over the United States,
526 *Atmospheric Chemistry and Physics*, 12, 10 295–10 312, <https://doi.org/10.5194/acp-12-10295-2012>, 2012.

527 Henneman, L. R., Holmes, H. A., Mulholland, J. A., and Russell, A. G.: Meteorological detrending of primary and secondary pollutant
528 concentrations: Method application and evaluation using long-term (2000-2012) data in Atlanta, *Atmospheric Environment*, 119, 201–
529 210, <https://doi.org/10.1016/j.atmosenv.2015.08.007>, 2015.

530 Holland, D. M., Principe, P. P., and Sickles, J. E.: Trends in atmospheric sulfur and nitrogen species in the eastern United States for 1989-
531 1995, *Atmospheric Environment*, 33, 37–49, [https://doi.org/10.1016/S1352-2310\(98\)00123-X](https://doi.org/10.1016/S1352-2310(98)00123-X), 1998.

532 Keller, C. A., Long, M. S., Yantosca, R. M., Da Silva, A., Pawson, S., and Jacob, D. J.: HEMCO v1. 0: a versatile, ESMF-compliant
533 component for calculating emissions in atmospheric models, *Geoscientific Model Development*, 7, 1409–1417, 2014.

534 Leung, D. M., Tai, A. P., Mickley, L. J., Moch, J. M., Van Donkelaar, A., Shen, L., and Martin, R. V.: Synoptic meteorological modes of
535 variability for fine particulate matter (PM_{2.5}) air quality in major metropolitan regions of China, *Atmospheric Chemistry and Physics*, 18,
536 6733–6748, <https://doi.org/10.5194/acp-18-6733-2018>, 2018.

537 Li, C., Martin, R. V., Van Donkelaar, A., Boys, B. L., Hammer, M. S., Xu, J. W., Marais, E. A., Reff, A., Strum, M., Ridley, D. A., Crippa,
538 M., Brauer, M., and Zhang, Q.: Trends in Chemical Composition of Global and Regional Population-Weighted Fine Particulate Matter
539 Estimated for 25 Years, *Environmental Science and Technology*, 51, 11 185–11 195, <https://doi.org/10.1021/acs.est.7b02530>, 2017a.

540 Li, K., Jacob, D. J., Liao, H., Shen, L., Zhang, Q., and Bates, K. H.: Anthropogenic drivers of 2013-2017 trends in summer
541 surface ozone in China., *Proceedings of the National Academy of Sciences of the United States of America*, 116, 422–427,
542 <https://doi.org/10.1073/pnas.1812168116>, 2018.

543 Li, K., Jacob, D. J., Shen, L., Lu, X., De Smedt, I., and Liao, H.: Increases in surface ozone pollution in China from 2013 to 2019: anthro-
544 pogenic and meteorological influences, *Atmospheric Chemistry and Physics*, 20, 11 423–11 433, 2020.

545 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
546 inventories in China: A review, *National Science Review*, 4, 834–866, <https://doi.org/10.1093/nsr/nwx150>, 2017b.

547 Lu, X., Zhang, L., Chen, Y., Zhou, M., Zheng, B., Li, K., Liu, Y., Lin, J., Fu, T. M., and Zhang, Q.: Exploring 2016-2017 surface
548 ozone pollution over China: Source contributions and meteorological influences, *Atmospheric Chemistry and Physics*, 19, 8339–8361,
549 <https://doi.org/10.5194/acp-19-8339-2019>, 2019.

Lu, X., Zhang, L., Wang, X., Gao, M., Li, K., Zhang, Y., Yue, X., and Zhang, Y.: Rapid increases in warm-season surface ozone and resulting health impact in China since 2013, *Environmental Science & Technology Letters*, 7, 240–247, 2020.

Ma, Z., Xu, J., Quan, W., Zhang, Z., Lin, W., and Xu, X.: Significant increase of surface ozone at a rural site, north of eastern China, *Atmospheric Chemistry and Physics*, 16, 3969–3977, <https://doi.org/10.5194/acp-16-3969-2016>, 2016.

McClure, C. D. and Jaffe, D. A.: US particulate matter air quality improves except in wildfire-prone areas, *Proceedings of the National Academy of Sciences of the United States of America*, 115, 7901–7906, <https://doi.org/10.1073/pnas.1804353115>, 2018.

Otero, N., Sillmann, J., Mar, K. A., Rust, H. W., Solberg, S., Andersson, C., Engardt, M., Bergström, R., Bessagnet, B., Colette, A., Couvidat, F., Cuvelier, C., Tsyro, S., Fagerli, H., Schaap, M., Manders, A., Mircea, M., Briganti, G., Cappelletti, A., Adani, M., D’Isidoro, M., Pay, M. T., Theobald, M., Vivanco, M. G., Wind, P., Ojha, N., Raffort, V., and Butler, T.: A multi-model comparison of meteorological drivers of surface ozone over Europe, *Atmospheric Chemistry and Physics*, 18, 12 269–12 288, <https://doi.org/10.5194/acp-18-12269-2018>, 2018.

Porter, W. C. and Heald, C. L.: The mechanisms and meteorological drivers of the ozone–temperature relationship, *Atmospheric Chemistry and Physics Discussions*, pp. 1–26, <https://doi.org/10.5194/acp-2019-140>, 2019.

Qu, L., Liu, S., Ma, L., Zhang, Z., Du, J., Zhou, Y., and Meng, F.: Evaluating the meteorological normalized PM_{2.5} trend (2014–2019) in the “2+26” region of China using an ensemble learning technique, *Environmental Pollution*, 266, 115 346, <https://doi.org/10.1016/j.envpol.2020.115346>, 2020.

Runge, J., Bathiany, S., Boltt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J.: Inferring causation from time series in Earth system sciences, *Nature Communications*, 10, 1–13, <https://doi.org/10.1038/s41467-019-10105-3>, 2019.

Saari, R., Mei, Y., Monier, E., and Garcia-Menendez, F.: Effect of Health-related Uncertainty and Natural Variability on Health Impacts and Co-Benefits of Climate Policy, *Environmental Science and Technology*, 53, 1098–1108, <https://doi.org/10.1021/acs.est.8b05094>, 2019.

Shen, L., Mickley, L. J., and Tai, A. P.: Influence of synoptic patterns on surface ozone variability over the eastern United States from 1980 to 2012, *Atmospheric Chemistry and Physics*, 15, 10 925–10 938, <https://doi.org/10.5194/acp-15-10925-2015>, 2015.

Sherwen, T., Schmidt, J. A., Evans, M. J., Carpenter, L. J., Großmann, K., Eastham, S. D., Jacob, D. J., Dix, B., Koenig, T. K., Sinreich, R., Ortega, I., Volkamer, R., Saiz-Lopez, A., Prados-Roman, C., Mahajan, A. S., and Ordóñez, C.: Global impacts of tropospheric halogens (Cl, Br, I) on oxidants and composition in GEOS-Chem, *Atmospheric Chemistry and Physics*, 16, 12 239–12 271, <https://doi.org/10.5194/acp-16-12239-2016>, 2016.

Shi, Z., Song, C., Liu, B., Lu, G., Xu, J., Van Vu, T., Elliott, R. J., Li, W., Bloss, W. J., and Harrison, R. M.: Abrupt but smaller than expected changes in surface air quality attributable to COVID-19 lockdowns, *Science Advances*, 7, <https://doi.org/10.1126/sciadv.abd6696>, 2021.

State Council of the People’s Republic of China: The Air Pollution Prevention and Control Action Plan (2013–2017), 744, 140 837, http://www.gov.cn/zwggk/2013-09/12/content_2486773.htm, 2013.

581 Tai, A. P., Mickley, L. J., Jacob, D. J., Leibensperger, E. M., Zhang, L., Fisher, J. A., and Pye, H. O.: Meteorological modes of variability
 582 for fine particulate matter (PM_{2.5}) air quality in the United States: Implications for PM_{2.5} sensitivity to climate change, *Atmospheric*
 583 *Chemistry and Physics*, 12, 3131–3145, <https://doi.org/10.5194/acp-12-3131-2012>, 2012.

584 Tian, R., Ma, X., and Zhao, J.: A revised mineral dust emission scheme in GEOS-Chem: Improvements in dust simulations over China,
 585 *Atmospheric Chemistry and Physics*, 21, 4319–4337, <https://doi.org/10.5194/acp-21-4319-2021>, 2021.

586 Travis, K. R., Jacob, D. J., Fisher, J. A., Kim, P. S., Marais, E. A., Zhu, L., Yu, K., Miller, C. C., Yantosca, R. M., Sulprizio, M. P., Thompson,
 587 A. M., Wennberg, P. O., Crounse, J. D., St Clair, J. M., Cohen, R. C., Laughner, J. L., Dibb, J. E., Hall, S. R., Ullmann, K., Wolfe, G. M.,
 588 Pollack, I. B., Peischl, J., Neuman, J. A., and Zhou, X.: Why do models overestimate surface ozone in the Southeast United States?,
 589 *Atmospheric Chemistry and Physics*, 16, 13 561–13 577, <https://doi.org/10.5194/acp-16-13561-2016>, 2016.

590 U.S. Energy Information Agency: Heat wave results in highest U.S. electricity demand since 2017, 744, 140 837, [https://www.eia.gov/](https://www.eia.gov/todayinenergy/detail.php?id=40253)
 591 [todayinenergy/detail.php?id=40253](https://www.eia.gov/todayinenergy/detail.php?id=40253), 2019.

592 U.S. Environmental Protection Agency: NATIONAL PRIMARY AND SECONDARY AMBIENT AIR QUALITY STANDARDS, 744,
 593 140 837, <https://ecfr.federalregister.gov/current/title-40/chapter-I/subchapter-C/part-50>, 2019.

594 U.S. Environmental Protection Agency: Criteria pollutants National Tier 1 for 1970 - 2020, 744, 140 837, [https://www.epa.gov/](https://www.epa.gov/air-emissions-inventories/air-pollutant-emissions-trends-data)
 595 [air-emissions-inventories/air-pollutant-emissions-trends-data](https://www.epa.gov/air-emissions-inventories/air-pollutant-emissions-trends-data), 2021a.

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

598 Vu, T., Shi, Z., Cheng, J., Zhang, Q., He, K., Wang, S., and Harrison, R.: Assessing the impact of Clean Air Action Plan on Air Quality Trends
 599 in Beijing Megacity using a machine learning technique, *Atmospheric Chemistry and Physics*, pp. 1–18, [https://doi.org/10.5194/acp-2019-](https://doi.org/10.5194/acp-2019-173)
 600 173, 2019.

601 Wang, Y. X., McElroy, M. B., Jacob, D. J., and Yantosca, R. M.: A nested grid formulation for chemical transport over Asia: Applications to
 602 CO, *Journal of Geophysical Research: Atmospheres*, 109, <https://doi.org/10.1029/2004JD005237>, 2004.

603 Wells, B., Dolwick, P., Eder, B., Evangelista, M., Foley, K., Mannshardt, E., Misenis, C., and Weishampel, A.: Improved estimation of trends
 604 in US ozone concentrations adjusted for interannual variability in meteorological conditions, *Atmospheric Environment*, 248, 118 234,
 605 2021.

606 Werf, G. R., Randerson, J. T., Giglio, L., Leeuwen, T. T. v., Chen, Y., Rogers, B. M., Mu, M., Van Marle, M. J., Morton, D. C., Collatz, G. J.,
 607 et al.: Global fire emissions estimates during 1997–2016, *Earth System Science Data*, 9, 697–720, 2017.

608 Wood, S. N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models,
 609 *Journal of the Royal Statistical Society (B)*, 73, 3–36, 2011.

610 Xie, Y., Wang, Y., Dong, W., Wright, J. S., Shen, L., and Zhao, Z.: Evaluating the Response of Summertime Surface Sulfate to Hydroclimate
 611 Variations in the Continental United States: Role of Meteorological Inputs in the GEOS-Chem Model, *Journal of Geophysical Research:*
 612 *Atmospheres*, 124, 1662–1679, <https://doi.org/10.1029/2018JD029693>, 2019.

613 Zhai, S., Jacob, D. J., Wang, X., Shen, L., Li, K., Zhang, Y., Gui, K., Zhao, T., and Liao, H.: Fine particulate matter (PM 2.5) trends in
 614 China, 2013–2018: Separating contributions from anthropogenic emissions and meteorology, *Atmospheric Chemistry and Physics*, 19,
 615 11 031–11 041, 2019.

616 Zhai, S., Jacob, D. J., Wang, X., Liu, Z., Wen, T., Shah, V., Li, K., Moch, J. M., Bates, K. H., Song, S., Shen, L., Zhang, Y., Luo, G., Yu, F.,
 617 Sun, Y., Wang, L., Qi, M., Tao, J., Gui, K., Xu, H., Zhang, Q., Zhao, T., Wang, Y., Lee, H. C., Choi, H., and Liao, H.: Control of particulate
 618 nitrate air pollution in China, *Nature Geoscience*, 14, 389–395, <https://doi.org/10.1038/s41561-021-00726-z>, 2021.

619 Zhang, H., Yuan, H., Liu, X., Yu, J., and Jiao, Y.: Impact of synoptic weather patterns on 24h-average PM_{2.5} concentrations in the North
 620 China Plain during 2013–2017, *Science of the Total Environment*, 627, 200–210, <https://doi.org/10.1016/j.scitotenv.2018.01.248>, 2018.

621 Zhang, Q., Zheng, Y., Tong, D., Shao, M., Wang, S., Zhang, Y., Xu, X., Wang, J., He, H., Liu, W., Ding, Y., Lei, Y., Li, J., Wang, Z., Zhang,
 622 X., Wang, Y., Cheng, J., Liu, Y., Shi, Q., Yan, L., Geng, G., Hong, C., Li, M., Liu, F., Zheng, B., Cao, J., Ding, A., Gao, J., Fu, Q., Huo,
 623 J., Liu, B., Liu, Z., Yang, F., He, K., and Hao, J.: Drivers of improved PM_{2.5} air quality in China from 2013 to 2017, *Proceedings of the*
 624 *National Academy of Sciences of the United States of America*, pp. 1–7, <https://doi.org/10.1073/pnas.1907956116>, 2019.

625 Zhang, Y., Vu, T. V., Sun, J., He, J., Shen, X., Lin, W., Zhang, X., Zhong, J., Gao, W., Wang, Y., Fu, T. M., Ma, Y., Li, W., and Shi, Z.:
 626 Significant Changes in Chemistry of Fine Particles in Wintertime Beijing from 2007 to 2017: Impact of Clean Air Actions, *Environmental*
 627 *Science and Technology*, 54, 1344–1352, <https://doi.org/10.1021/acs.est.9b04678>, 2020.

628 Zhao, Y., Zhang, K., Xu, X., Shen, H., Zhu, X., Zhang, Y., Hu, Y., and Shen, G.: Substantial Changes in Nitrate Oxide and Ozone after
 629 Excluding Meteorological Impacts during the COVID-19 Outbreak in Mainland China, *Environmental Science Technology Letters*,
 630 <https://doi.org/10.1021/acs.estlett.0c00304>, 2020.

631 Zheng, B., Tong, D., Li, M., Liu, F., Hong, C., Geng, G., Li, H., Li, X., Peng, L., Qi, J., et al.: Trends in China’s anthropogenic emissions
 632 since 2010 as the consequence of clean air actions, *Atmospheric Chemistry and Physics*, 18, 14 095–14 111, 2018.

633 Zhong, Q., Ma, J., Shen, G., Shen, H., Zhu, X., Yun, X., Meng, W., Cheng, H., Liu, J., Li, B., Wang, X., Zeng, E. Y., Guan, D., and Tao, S.:
 634 Distinguishing Emission-Associated Ambient Air PM_{2.5} Concentrations and Meteorological Factor-Induced Fluctuations, *Environmental*
 635 *Science and Technology*, 52, 10 416–10 425, <https://doi.org/10.1021/acs.est.8b02685>, 2018.

636 Zurbenko, I. G.: Detecting and tracking changes in ozone air quality, *Air and Waste*, 44, 1089–1092,
 637 <https://doi.org/10.1080/10473289.1994.10467303>, 1994.