Reply to reviewer 1:

General Comments:

The idea of evaluating the performance of statistical and machine learning methods used to correct meteorological variability in emission trends using model results is an interesting one, since there have been many recent papers that have been written that use these methods on real-world data sets, without any real metric of their effectiveness in recovering the true trends in emissions.

Overall, this is a well-written paper with a set of carefully designed experiments to assess the performance of different statistical methods to determine meteorology corrected emission trends. The writing is high-quality, includes proper citations, and the figures are clear and easy to understand. I recommend publishing with a few minor corrections to improve the readability of the manuscript and comprehensibility for researchers without a background in statistics.

Response: We would like to thank the reviewer for the positive assessment of our manuscript and the constructive feedback. We appreciate the opportunity to respond to these thoughtful comments and use these edits to strengthen our paper. Below, we include a detailed response to the reviewer's comments and suggestions. Reviewer comments are in *italics*, our responses are in plain text, and changes to the manuscript are in blue. The line numbers refer to the line numbers in the revised version of the paper.

There were several places where more detail is warranted. Particularly in the description of the application of the different models, there was not a lot of detail and it was difficult to determine how these methods were applied to the data sets. This is important in assessing the conclusions of the paper, that an RF model is preferable to the other statistical methods, as the specific implementation of each method could have a significant impact on its performance. This is particularly true for the machine learning methods.

Response: Thank you for this comment. We have now provided further details in the method and SI section. In the public version, code scripts (in R) to implement these methods will be made available to the readers.

Lines 205-206: More details on the implementation of LASSO and RF can be found in SI.

Lines 462-504: Appendix section Implementation of RF and LASSO

Specific Comments:

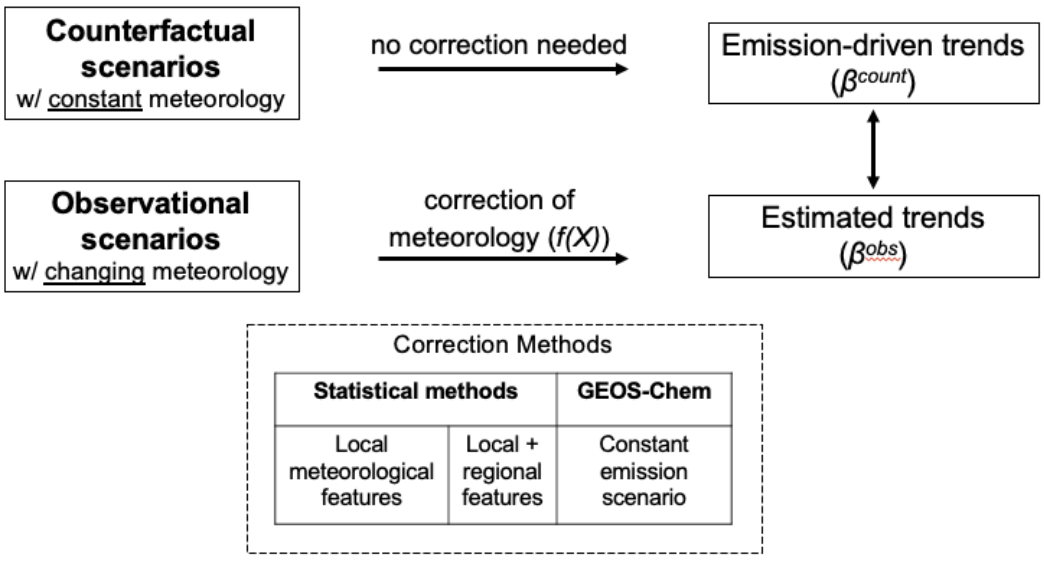
I found the discussion of causal methods in lines 65-76 to be slightly confusing in this paper, since the paper was not focusing on assessing causal links, but rather on testing counterfactuals—this link should have been more directly made clear, especially for the typical reader of this journal who doesn't have a background in statistics/causal inference.

Response: Thank you for raising this question. The target of this method is to correctly estimate the emission-driven trend in air quality (for each location), a parameter with causal interpretation. Therefore, we believe our exercise is related to a broader literature that focuses on estimating the causal impacts of anthropogenic emissions change on air quality, under meteorological variability. Nevertheless, the idea of estimating causal links is tightly related to the idea of estimating counterfactuals. We have now provided further clarification in the introduction section.

On lines 65-69: Despite a large number of papers which apply various meteorology correction methods, very little is known about whether these methods can effectively correct for meteorological variability and thus realistically estimate the counterfactual air quality and reveal the underlying 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 the correct estimation of counterfactuals and causal effects (Runge et al., 2019).

It would have been useful to have some type of overview cartoon for the different experiments and their relationship to the terms in equation 1— Table 1 was useful, but I had to read the paper through twice before the different simulations were clear to me, and it would have been helpful to have some visual aid for this.

Response: We agree with the reviewer and have now provided a schematic flow chart as a visual overview of our methodology (as the current figure 1).



Lines 146. What fraction of the meteorology-concentration relationship is due to changes in natural emissions?

Response: Natural emissions are often calculated online in GEOS-Chem and are therefore hard to separate from the direct effect of meteorology on air quality. To at least partially answer this question, here we focus on soil NOx emissions and biogenic secondary organic aerosol emissions in the US as two examples to understand the variability of natural emissions across different meteorological years. We find that the annual total soil NOx emission (in the US) varies by as much as 38% across years (total soil NOx emissions in 2012 are 38% larger than 2014), and biogenic SOA emission varies by as much as 25%. The variability in soil NOx and biogenic SOA is equivalent to 7% of the anthropogenic NO emissions, and 8% of the anthropogenic SOA emissions. Variability in soil NOx and biogenic SOA is slightly larger during the summer season. Nevertheless, our results suggest that the contribution of natural emission changes to the meteorology-concentration relationship is relatively small compared to the direct influence of meteorological variables on air quality. This is consistent with previous analyses; for example, Porter and Heald find that the soil NOx emission changes can explain ~10% of the ozone-temperature relationship in the US. We should note that the changes in biomass burning emissions due to climate variability could be potentially large, especially in regions like the western US, but they are held constant in the current model experiments for simplicity. We have now added a paragraph in the discussion section on the role of natural emissions in the meteorology-concentration relationship.

On lines 390-401: Changes in natural emissions due to meteorological variability play an important role in the air quality-meteorology relationship. Our model experiment considers natural emission changes that can be simulated online with assimilated meteorological fields in GEOS-Chem, including soil NOx emissions, biogenic VOC emissions, and dust emissions. We find that the statistical models perform notably worse in correcting for the variability in dust-related $PM_{2.5}$ (see figure A12 for results using RF-regional), likely because dust$PM_{2.5}$ is extremely variable, with zero concentration on most non-dust days but extremely high concentration during the occasional dust storms. Our findings can potentially shed light on another important source of natural emissions, wildfire emissions, which are also quite variable but have become an increasingly important contributor to $PM_{2.5}$ and O_3 in certain regions (e.g., western US) (Burke et al., 2021). While emissions from biomass burning are held constant in our model experiments as the wildfire emissions are prescribed in GEOS-Chem, wildfire emissions are significantly influenced by climatic variability (Abatzoglou et al., 2016, Xie et al., 2022) and will likely be a substantial challenge for any meteorological correction method in the future that attempts to separate changes in anthropogenic emissions from the variability in climate and associated natural emissions.

Is there a clear separation between the training and test data for the Random Forest? As I understand this was in part the point of the double-machine learning method, but this should be more clearly spelled out. It would obviously be problematic if both the training and test data are used to evaluate the performance of the RF method to recover the emission-driven trends, as this would give an artificially good performance for this method.

Response: Thank you for raising this important question. With the double machine learning method, we implement a sample splitting step while implementing the RF model. We first randomly partition each dataset (timeseries) into 4 folds. We use 75% of the data as training data and the remaining 25% for predictions. This step is then repeated four times to derive the predictions of the full timeseries, to avoid the risk of overfitting. We have now provided further details in the method and appendix sections.

On lines 204-206: In particular, the hyper-parameters and coefficients of LASSO and RF are selected and fitted using 4-fold cross-validation to avoid the "overfitting risk". More details on the implementation of LASSO and RF can be found in SI.

It is also important to note that we quantify the performance of RF and other methods using the differences between "meteorology-corrected" trends (β^{obs}) and the counterfactual trends (β^{count}), instead of their performance in predicting the pollutant concentration. Therefore, if the RF model "overfits the data", it would actually result in a large error and a poor performance when compared with the counterfactual trends. This is because the overly fit RF model would attribute all variability of PM and O3 to the meteorological variables, and therefore estimate a close-to-zero trend in air quality driven by emissions. In fact, if we implement the RF method without the sample-splitting step, we would have an average error of over 0.15 ug/m^3/year, larger than any evaluated methods (e.g., 0.047 ug/m^3/year from a properly implemented RF model).

Lines 500-504 in the appendix.

Lines 170-176. Can you spell out what you mean by the "uncorrected" method here? Does that mean the term $f_i(X_{it})$ is neglected in equation 1?

Response: Yes, this is correct. "Uncorrected" means simply fitting a trend without the term $f_i(X_{it})$ (with only the time fixed effects included). We have further clarified this in the method section.

On lines 178-180: We refer to the trend estimates estimated without $f_i(X_{it})$ as "uncorrected".

Section 3.4. How are the observations corrected? Are these using the meteorological correction models as determined from the GEOS-Chem model?

Response: For the observational data of $PM_{2.5}$ and O_3 from the surface monitors, they are corrected with the same kind of statistical models such as MLR and RF (as in the GEOS-Chem analysis), but with different numerical coefficients. For example, we also use a MLR with the ten meteorological variables to perform meteorological corrections for the observational air quality data, but the regression coefficients are determined directly from the observational data, and are thus different from the model used for GEOS-Chem. We have provided further clarification in the method section.

On lines 332-333: Here, to correct for the meteorology variability in observational data, we implement the same set of statistical methods as shown in Table1, but with different numerical coefficients directly derived from the observational data.

Reply to reviewer 2:

Overall, I think this is a novel approach to the characterization of meteorological adjustments to air pollutant concentration trends and it is worth publishing. The use of air quality models to create a synthetic dataset by which meteorological adjustment methods can be evaluated has not been previously published that I am aware of. The selection of air quality model, statistical adjustment methods, and regions are appropriate and the presentation as a whole is clear and well thought-out. I have several minor concerns listed below that I think could be addressed through minor revisions.

Response: We thank the reviewer for the positive assessment of the paper, and the thoughtful and constructive feedback. We appreciate the opportunity to respond to these detailed comments to strengthen our paper. Below, we include a detailed response to the reviewer's comments and suggestions. Reviewer comments are in *italics*, our responses are in plain text, and changes to the manuscript are in blue. The line numbers refer to the line numbers in the revised version of the paper.

- From the manuscript, it isn't clear how the linear trend estimates are calculated. Are they calculated using linear regression or some other method (e.g. Theil-Sen)?

Response: The linear trend estimates are calculated with the standard ordinary least square approach from regression models with different forms of the meteorology correction methods (for example, linear or non-linear combinations of the meteorological variables). We further clarified this in the method section.

On lines 172-173: β^{obs} represents the meteorology-corrected trend in $PM_{2.5}$ or O_3 concentration for grid cell *i* estimated with the standard ordinary least square method.

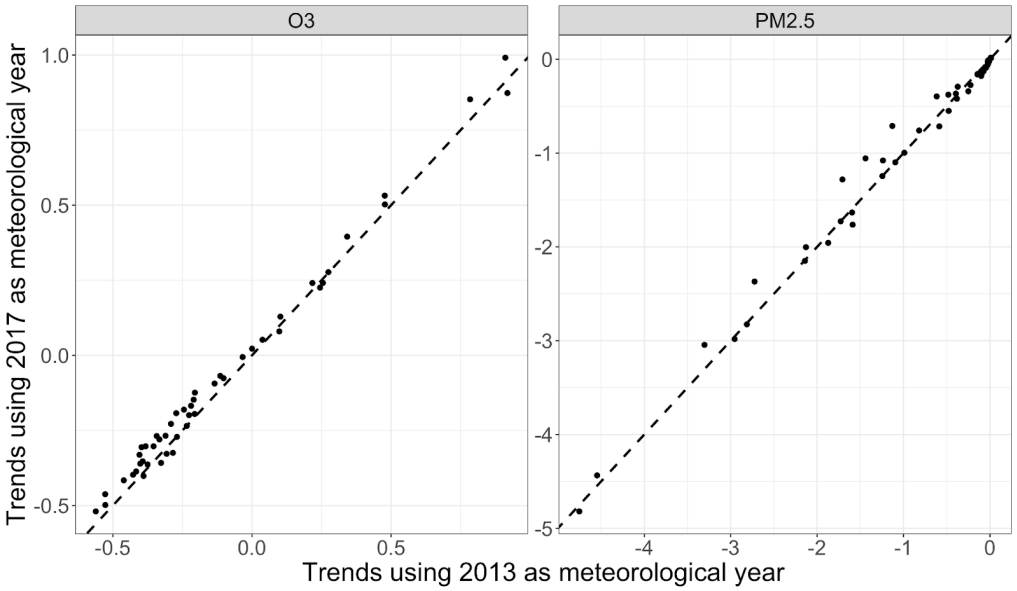
- One of my biggest concerns is that the time period (7 years for U.S., 5 years for China) is too short to calculate meaningful trend estimates, and because of this, the trend estimates themselves may be a source of additional uncertainty. I fully understand the time and resource constraints of running chemical transport models, therefore, I'm not suggesting that this must be done, but merely that this is addressed as a limitation in the discussion. As a potential future research application, one could expand this type of analysis to a set of model runs evaluated over a longer time period, such as EPA's EQUATES series for 2002-2017 (https://www.epa.gov/cmaq/equates). The emissions and meteorology inputs for the EQUATES model runs are available from this website.

Response: We agree with the reviewer's comment regarding the length of the simulations. Following this suggestion, we have now added discussion regarding the length of the simulation, and we further suggest potential future studies in the discussion section. The EPA's EQUATES series seems like a promising starting point for a future study on that front, but as the reviewer points out, would require additional simulations of the counterfactual scenario (e.g., with constant meteorology) to apply our approach.

On lines 409-412: Constrained by computational resources and availability of emission inventories, our simulation only covers a relatively short time period which could result in additional uncertainty in the linear trend estimates. When possible, future studies could evaluate performances of the statistical models with longer simulations and alternative trend estimates (such as the Theil-Sen estimator)

- Another potential source of uncertainty in this application is the choice of meteorological year for the set of model runs where the meteorology is held constant. For example, as the authors describe, the 2011 year which was held constant for the U.S. was unusually hot and dry throughout the central region of the country. As a sensitivity analysis, I think it would be useful to see how much the predictions change by holding the meteorology constant using other years. For example, 2013 and 2014 were cooler and wetter than average for much of the U.S. For the current manuscript, I think it would be sufficient for the authors to address this point in the discussion.

Response: Thank you for raising this issue. We agree with the reviewer that it would be important to better understand how the choice of the meteorological year could influence our results. In a sensitivity run we previously did, we simulated the counterfactual scenario (constant-meteorology) for China using the meteorological fields at the end year 2017 (at 4x5 degree resolution). We find that the linear trend in $PM_{2.5}$ and O_3 for each grid cell is highly consistent in the counterfactual scenarios regardless of the choice of the meteorological years (see figure below). As the meteorological conditions are quite different between 2013 and 2017 in China (e.g., the winter of 2013 was particularly bad for air quality in Northern China due to several stagnation episodes, while the meteorological conditions in winter 2017 favored the dispersion of air pollutants), the findings from this sensitivity analysis potentially also apply to the US case. This suggests the potential influence due to the choice of meteorological year is relatively small here in our cases. We have now discussed this sensitivity analysis in the method section and added the following figure in the SI.



Lines 131-134: In a sensitivity analysis, we also simulate the counterfactual scenario for China using the meteorological fields at the end year 2017 (at 4x5 degree resolution, due to computational constraints). We find that the linear trend in $PM_{2.5}$ and O_3 for each grid cell is highly consistent in the counterfactual scenarios across the choice of the meteorological years (see figure A5).

- The choice of the June-August period for ozone does not capture the period of maximum ozone concentrations for all regions of the U.S. For example, the southeast U.S. typically sees its highest ozone concentrations in April or May, while California may experience peak ozone in September or October. I think a period of April to October would be sufficient to capture the peak ozone concentrations in all regions of the U.S. Again, nothing needs to be redone, but it would be useful to discuss this in the manuscript.

Response: Thank you for this suggestion. We have now noted this limitation in our method section.

On lines 109-111: Our focus on the three summer months is consistent with many previous studies (e.g., Shen et al., 2015), although this may not capture the peak ozone season for certain regions of the US and China.

- As far as the interaction between meteorological and emissions-based effects, I agree that this is both a concern and a major challenge for any meteorological adjustment approach, and ultimately, it may not be possible to estimate the magnitude of these interactions. One major source of these interactions, especially in recent years, is wildfires: dry meteorological conditions contribute to more wildfires, and more wildfires contribute additional emissions. Wildfires are especially difficult to capture in a chemical transport model due to their unpredictability and the difficulty of characterizing their emissions. As wildfires can be an especially large contributor to PM2.5 concentrations, it would be useful to see them discussed in the context of their contribution to met/emissions interactions and overall uncertainty.

Response: Thank you for raising this very important issue. We completely agree with the reviewer that wildfires (or PM/O3 caused by natural emissions in general) are likely major sources of meteorological-emissions interaction. While our focus is largely on the interactions between meteorology and anthropogenic emissions and the challenge of diagnosing the effects of anthropogenic emissions, we have now provided further discussion on this issue.

On lines 395-401: Our findings can potentially shed light on another important source of natural emissions, wildfire emissions, which are also quite variable but have become an increasingly important contributor to $PM_{2.5}$ and O_3 in certain regions (e.g., western US) (Burke et al., 2021). While emissions from biomass burning are held constant in our model experiments as the wildfire emissions are prescribed in GEOS-Chem, wildfire emissions are significantly influenced by climatic variability (Abatzoglou et al., 2016, Xie et al., 2022) and will likely be a substantial challenge for any meteorological correction method in the future that attempts to separate changes in anthropogenic emissions from the variability in climate and associated natural emissions.

- While I don't plan to list them all here, I noticed several typos and minor grammatical errors in the manuscript while reading it.

Response: Thank you! We have now proofread the manuscript from the beginning to the end and corrected typos and grammatical errors.