

Response to Referee #1

Atmospheric Chemistry and Physics

Manuscript egusphere-2026-3211

Nonlinear and regime-dependent precursor-PM_{2.5} responses during wintertime calm episodes in Beijing

Honglin Li, Xinwen Liu, and Song Xi Chen

21 September 2026

The responses below represent our current assessment and revision plan. Some details may be refined as additional checks and sensitivity analyses are being completed in preparation of the revised manuscript. The referee's comments are reproduced verbatim in italics, followed by our responses.

This study investigates the nonlinear and regime-dependent effects of SO₂ and NO₂ on PM_{2.5} accumulation in Beijing during the winter heating seasons of 2016–2023. The authors develop an episode-based causal inference framework that restricts the analysis to calm episodes, use CO and PM_{2.5} as proxies for unresolved local emissions, and combine proxy calibration with GAM/SCAM models to characterize precursor responses and effect modification. The study further examines spatial heterogeneity, joint SO₂-NO₂ response surfaces, and year-to-year precursor-associated PM_{2.5} changes. The topic is important and potentially suitable for publication. However, there are substantial concerns about whether the current data and assumptions are sufficient to support the causal inferences proposed in this paper. Several substantial issues concerning the credibility of causal identification assumptions, model specification and presentation of results. I recommend major revision before reconsideration for publication.

Reply: We appreciate you for the careful and largely encouraging assessment. We recognize your concerns regarding the inferential scope, identification assumptions, model specification, and presentation of the results. Our planned revisions will address these issues by clarifying the distinction between conditional exposure–response relationships and causal interpretations, strengthening the relevant diagnostic and sensitivity analyses, and moderating mechanistic and policy interpretations where the available evidence is insufficient.

The manuscript repeatedly refers to the estimated SO₂ and NO₂ relationships as causal effects, while the identification strategy depends on strong assumptions S1–S7, including error orthogonality, exclusion restrictions, a control-function representation, and approximation assumptions for the rotated proxies. I think you should clearly distinguish among a causal exposure effect, a conditional exposure-response relationship, and a statistical association. You should state which parts of the identification strategy are empirically supported and which rely entirely on assumptions.

Reply: Thank you for the insightful comments. The causal relationship among the major players that lead to PM_{2.5} is summarized in the directed acyclic graph (DAG) in Fig. 2, which is based on the understanding of atmospheric physics and chemistry. Within this structure,

SO₂ and NO₂ are specified a priori as precursor exposures because of their established roles in sulfur- and nitrogen-containing secondary aerosol formation, rather than inferred from statistical associations from observed data.

We refer to the proposed method within the DAG as a “causal inference framework” because the specified DAG is translated into structural equations and combined with a proxy-based control-function approach for latent emission confounding. The phrase “causal effect” was used only in the theoretical identification discussion, where it refers to the causal estimand whose identification and consistent estimation are conditional on the assumptions stated in Sect. S2. In the Results and Discussion, terms such as “effect” and “marginal effect” were generally used to describe changes in the fitted conditional precursor–PM_{2.5} response rather than to assert causality. In the revision, we will describe the fitted relationships as “precursor responses” or “conditional exposure–response relationships” to avoid confusion.

We were not clear on the roles of Assumptions S1–S7. They are not to define the causal structure, but to formalize conditions under which the specified structural system can be identified from observational data and under which the proxy-control estimator can consistently recover the target structural estimand. Assumptions S4 and S6 are partly motivated by prior knowledge that CO and PM_{2.5–10} contain source-related information on unresolved emission variability. However, because the latent emission factor U is unobserved, the adequacy of the control-function representation and the preservation of the relevant information on U cannot be directly verified empirically. The remaining assumptions mainly serve as identification and statistical conditions, rather than empirically testable statements about the atmospheric system. In the revision, we will make these distinctions explicit.

In Assumption S4 (Control Function Representation): is the relationship between emissions and concentrations potentially nonlinear? There may have interactions between emission intensity and dispersion conditions. Please give the interactions between emission intensity and dispersion conditions. Additional diagnostic evidence is necessary to justify the validity of this linear approximation.

Reply: Thank you for raising this important point. Assumption S4 is introduced for the benchmark linear model and is used to obtain a transparent identification and consistency result. If the relationship between the latent emission factor and its proxies is nonlinear or depends on dispersion conditions, for example $U = g(W, D) + \zeta$, where D denotes dispersion-related meteorological conditions, the control-function argument can still be generalized provided that the latent-emission contribution enters the outcome model additively and that the residual component ζ does not retain exposure-related confounding information. We will add this nonlinear extension and the corresponding derivation to the revised SI.

Because U is unobserved, its functional relationship with the proxies cannot be tested empirically. In our SCAM framework, the empirical outcome model does not impose linear effects of the proxy indices in that e_1 and e_2 enter the model as smooth terms. As shown in Fig. 3, however, their estimated contributions are approximately linear over most of the empirical support, providing an indirect observable diagnostic that is broadly consistent with the linear approximation used in the benchmark derivation.

We will further assess sensitivity to possible interactions with dispersion conditions using parsimonious interaction models, allowing the observed proxy indices to interact with key dispersion variables, including wind speed and boundary-layer height. We will examine whether the inclusion of these interaction terms materially changes the estimated precursor responses.

Line 106-110: Low wind speed does not exclude regional transport, particularly under shallow boundary-layer conditions. Additional evidence such as back-trajectory or regional transport analysis, would strengthen the interpretation of local PM_{2.5} formation.

Reply: Thank you for raising the point. We recognize that limiting the surface wind speed cannot completely exclude regional transport. Our episode design therefore imposes multiple restrictions intended to minimize regional influence. Each calm episode begins after a preceding northerly flushing period that reduces the pre-existing regional pollution burden, and the subsequent calm period is selected to represent an atmospheric regime characterized by reduced horizontal transport and limited vertical dilution. This characterization is defined by the PDI-based selection criterion and supported by the meteorological characteristics of the selected episodes shown in Figs. S1–S2. The initial PM_{2.5} level is also included in the model to account for the remaining pollution state at episode onset.

The prolonged-episode sensitivity analysis in Sect. S2.5 further shows poorer model performance and shifts in the estimated precursor responses when each episode is extended by 5 h, indicating that the fitted precursor–PM_{2.5} relationships are less stable once the analysis is extended beyond the selected calm period.

Given the short duration and limited spatial scale of these episodes, back-trajectory estimates might be sensitive to the resolution and representation of near-surface meteorological fields. We seek your understanding and pardon for not conducting it. We have instead used transport diagnostics derived directly from the observed hourly wind vectors, consistent with the spatial and temporal scales of our observation-based analysis. The net and maximum horizontal displacements during the selected calm episodes, together with the paired comparison of equal-length 5-h windows immediately before and after episode termination will be reported in the revised Sect. S2.5 of the SI.

Section 3.2: AIC improvement scores are calculated from 500 bootstrap replications but averaged across site clusters without accounting for cluster sample size differences. If some clusters have small sample sizes, their AIC estimates have larger uncertainty, and simple averaging may underweight their influence.

Given the sample size and model complexity, AIC selection may be prone to overfitting. The authors should present the distribution of AIC improvements across candidate pairs, not just mean values, to allow readers to assess selection uncertainty.

Reply: Thank you for the insightful comments. The sample sizes were 714, 1,354, and 1,192 for the Suburban, UrbanNW, and UrbanSE clusters, respectively. We agree that the sample-size disparity can affect the magnitude of the AIC improvement, potentially giving larger-sample clusters greater influence on the pooled mean. We therefore revised the pooled criterion to use the standardized AIC improvement, $\Delta\text{AIC}/N$, corresponding to expressing AIC on a per-observation scale as AIC/N . This standardization does not change the ranking of candidate models within each cluster, but places the AIC improvements on a comparable per-observation scale before averaging across site clusters. In our reanalysis, the selected effect-modification structure remained unchanged under this standardized criterion. To make the uncertainty in modifier selection more transparent, we will also report the full bootstrap distributions of $\Delta\text{AIC}/N$ for all candidate modifier pairs within each site cluster in the revised main text.

Regarding potential overfitting, the SCAMs are regularized by smoothing penalties and shape constraints. The final fitted models use approximately 54–59 effective degrees of freedom for 714–1,354 observations across the three clusters. Together with the existing leave-one-episode-

out validation, these results would not indicate substantial overfitting of the fitted nonlinear models. We will add these discussion in the revision.

Figure 5: it reveals substantial differences in precursor responses across the three site clusters. UrbanNW shows the strongest NO₂ response but weakest SO₂ response, interpreted as “reactive nitrogen chemistry dominance”. Why would nitrogen chemistry be more active there? It may need more detailed consideration of local emission sources, site characteristics, and meteorological factors.

Reply: Thank you for highlighting this issue. The estimated response represents the conditional sensitivity of PM_{2.5} to changes in NO₂ after adjustment for the other modeled factors. Our original interpretation of this spatial heterogeneity as “reactive nitrogen chemistry dominance” was partly motivated by the effect-modification results in Fig. 6, where the O₃ modification was strongest in UrbanNW. Because O₃ reflects aspects of the prevailing oxidative and photochemical environment while being chemically coupled with NO_x, the combination of a stronger NO₂ response and O₃ modification provides a plausible mechanistic context for the observed UrbanNW response.

We will therefore moderate the mechanistic wording and further examine site characteristics and local source context, together with the observed O₃ modification, to assess whether the spatial heterogeneity can be supported by additional evidence.

Section 3.3.3: The annual attribution shows increased NO₂-related PM₅ contributions after 2020. However, 2020-2022 includes the COVID-19 pandemic period, during which NO_x emissions were anomalously low, potentially altering the statistical relationship between NO₂ and PM_{2.5}. Does inclusion of these “anomalous years” affect trend inference?

Reply: Thank you for raising this point. We recognize that 2020–2022 represent an unusual emission period and that the COVID-related emission reductions have provided an important natural experiment in atmospheric-chemistry research. We therefore understand your concern regarding their influence on the annual analysis.

Our annual analysis is based on a unified nonlinear exposure–response function estimated from the full 2016–2023 sample. Because the response function is nonlinear, the marginal NO₂ response is allowed to vary across concentration levels. Therefore, a lower-emission or lower-NO₂ environment during the COVID period does not by itself invalidate the identification of the exposure–response relationship; such observations are accommodated within the lower-concentration part of the same response surface.

To examine whether the COVID period produced an anomalously low NO₂ concentration regime within our analytical sample, we additionally compared precursor concentrations during 2020–2021 with those in the remaining years using Welch’s two-sample *t*-tests. In our annual definition, 2022 refers to the 2022–2023 heating season, which spans the transition out of the COVID-related restriction period. For NO₂, concentrations during 2020–2021 were not significantly lower in any of the three site clusters (adjusted $p \geq 0.159$).

We will revise Sect. 3.3.3 to make this unified nonlinear-response assumption and the interpretation of the annual standardized estimates more explicit.

Line 356 – 361: The higher R² values and lower RMSE values indicate improved model fit, but can these results validate the plausibility of its causal interpretation?

Reply: Thank you for this important comment. The higher R^2 and lower RMSE values were intended as measures of model performance and specification adequacy. Good model fit cannot establish causal validity, whereas poor fit would indicate that the proposed model structure is empirically inadequate and may cast doubt on the casual relationship.

The improved R^2 and RMSE values show that specifications incorporating nonlinear responses, effect modification, and proxy adjustment are more consistent with the observed data than simpler alternatives. These improvements are therefore supportive of the adequacy of the proposed model structure. The leave-one-episode-out validation further indicates that the improved model fit is not simply a consequence of overfitting the estimation sample.

We will revise the manuscript to make these more explicit.

Line 84: Why is a five-point weighted moving average applied? The five-point weighted moving average is applied, Could this smoothing alter the timing of the precursor-PM_{2.5} relationships, particularly for the short 2-3 h exposure windows?

Reply: Thank you for requesting clarification of this preprocessing step. The five-point filter used in this study is a centered symmetric weighted moving average with weights 0.1:0.2:0.4:0.2:0.1, applied identically to all pollutant variables. The filter was introduced to reduce high-frequency observational variability and short-term noise in the routine monitoring series, and similar moving-average smoothing has been applied in atmospheric studies (Luo et al., 2022; Li et al., 2020; Zhong et al., 2021).

Because the same centered symmetric filter is applied to both the precursor and PM_{2.5} series, it does not introduce a systematic relative time shift between them under an ideal linear distributed-lag representation. Smoothing may nevertheless make neighboring lag orders more difficult to distinguish in finite samples. The episode-based bootstrap procedure therefore provides an empirical assessment of the resulting lag-selection uncertainty. We will clarify the rationale and temporal implications of the symmetric filter in the revised SI.

Line 234: “These two quite stable effect modifiers reveal that...” is somewhat overstated. The results suggest an association with oxidative capacity and thermodynamic conditions, but may not by themselves establish the underlying chemical drivers.

Reply: Thank you for drawing our attention to the inappropriate wording in this statement. By “quite stable effect modifiers”, we originally intended to describe the statistical stability of the modifier selection in the bootstrap AIC comparison, rather than to establish TEMP and O₃ as the underlying chemical drivers.

We will therefore remove this wording and revise the corresponding text to distinguish more clearly between the statistical evidence for effect modification and its possible chemical interpretation.

Line 248: The interpretation of O₃ as an oxidative-capacity indicator should be discussed more carefully because O₃ and NO₂ are strongly chemically coupled.

Reply: Thank you for this helpful comment. We also considered whether the strong chemical coupling between O₃ and NO₂ should be represented explicitly in the structural equations. At the analysis scale considered here, NO₂ and O₃ are treated jointly as characterizing the prevailing NO_x–O₃ chemical state. O₃ is therefore used as a conditioning variable for the NO₂ response rather than as a standalone measure of oxidative capacity. Variation in O₃ across episodes then characterizes how the NO₂–PM_{2.5} relationship varies across photochemical regimes. We will revise the corresponding text to make this interpretation more explicit and cautious.

Line 303: “support-wide average slope” is used without definition, Please clarify it.

Reply: Thank you for pointing this out. We will define this quantity explicitly as the average local iso-response slope across empirically supported SO₂–NO₂ combinations, with the local trade-off calculated from the ratio of the corresponding partial derivatives. The “±” value summarizes the standard variation of these local slopes across the supported response surface.

Figure S6 uses “m/s” in the axis labels, whereas the text and equations use “m s⁻¹”. Please use a consistent unit format throughout the manuscript and figures.

Figure S8 uses “Colour” whereas the main text uses “color”. Please check and use one spelling convention consistently throughout the manuscript.

Table S1: Should “All correlations are significant at 0.001” be “All correlations are significant at $p < 0.001$ ”?

Reply: Thank you for noting these inconsistencies. We will correct them throughout the Supplementary Information and conduct an additional consistency check of the remaining text, tables, and figures for similar issues.

We again thank the referee for the constructive comments, which have helped us clarify the scope, assumptions, and interpretation of our study.

Sincerely,

Song Xi Chen
on behalf of all authors

References

- Li, Q., Wu, B., Liu, J., Zhang, H., Cai, X., and Song, Y.: Characteristics of the atmospheric boundary layer and its relation with PM_{2.5} during haze episodes in winter in the North China Plain, *Atmospheric Environment*, 223, 117 265, <https://doi.org/10.1016/j.atmosenv.2020.117265>, 2020.
- Luo, S., Zhu, Y., and Chen, S. X.: Episode based air quality assessment, *Atmospheric Environment*, 285, 119 242, <https://doi.org/10.1016/j.atmosenv.2022.119242>, 2022.
- Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., Zhang, L., Zhang, Y., Sun, J., and Zhang, W.: Robust prediction of hourly PM_{2.5} from meteorological data using LightGBM, *National Science Review*, 8, nwaa307, <https://doi.org/10.1093/nsr/nwaa307>, 2021.