# Reply to Reviewer Comments

The manuscript proposes a machine-learning counterfactual framework to estimate the impact of the Chinese Spring Festival (CSF) on $PM_{2.5}$ in Hangzhou and the "2+26" cities. The study is timely and policy-relevant, with a clear intention to distinguish air-quality changes from emissions, and the manuscript is well organized and clearly presented. However, several aspects of causal ML practice, the temporal validation strategy, and issues of data representativeness need to be strengthened before publication.

Major Comments

**Comment #1:**

The manuscript frames its analysis within a causal framework, treating the Chinese Spring Festival (CSF) as a "treatment" and using the XGBoost model to predict a counterfactual business-as-usual (BAU) scenario. While this is a conceptually appropriate starting point, the current methodology does not yet meet a rigorous causal ML design. The CSF is a composite factor, bundling the effects of fireworks, altered traffic patterns, and changes in industrial/construction activity. This complexity challenges the core identification assumptions required for causal claims.

Furthermore, the analysis does not adequately address potential influence of these assumptions, such as the inconsistent overlap in covariate distributions between festival and non-festival periods. Some features, like the lunar calendar day, are inherently confounded with the treatment, violating conditional independence. The study could be characterized as a causally inspired counterfactual prediction for BAU rather than a causal estimator under verified identification conditions. Hence, the authors may wish to reconsider the title and tone down the causal claims to avoid overstatement.

**Response:**

We sincerely thank the reviewer for this important and constructive comment. We agree that the Chinese Spring Festival (CSF) represents a composite intervention involving multiple concurrent behavioral and emission changes (e.g., fireworks activities, reductions in traffic volume, and modifications in industrial or construction operations). As a result, several key identification assumptions required for strict causal inference—such as conditional independence, adequate covariate overlap between treated and untreated periods, and the independence of certain covariates (e.g., lunar-day indicators) from the intervention—cannot be fully validated in this context. We appreciate the reviewer's clarification, which has helped us refine the conceptual framing of the paper. Following the reviewer's suggestion, we have revised the manuscript in several ways to ensure that the study is presented as a causally informed counterfactual analysis rather than a strict causal estimator:

(1). Revised the title to remove any implication of strong causal identification.
The new title is:
"Impact of the Chinese Spring Festival on $PM_{2.5}$ air quality in the Beijing-Tianjin-Hebei and surrounding region: A machine-learning-based counterfactual modeling approach".

(2). We have modified the methodological description in the Introduction to clearly state that the proposed framework is not intended to identify structural causal effects. The revised text now clarifies that our approach provides a causally informed counterfactual prediction of the BAU scenario rather than a formal causal estimate.

(3). We have added a dedicated clarification paragraph at the end of Section 2.2 (lines 127-137), explicitly acknowledging the composite nature of the CSF intervention, the challenges associated with validating causal identification assumptions, the inherent confounding of calendar-related variables such as the lunar-day index, and the resulting limitations for making strict causal claims.

(4). Adjusted wording throughout the manuscript (e.g., replacing "causal impact" with

"holiday-related effects" or "deviations relative to BAU") to avoid overstating causal interpretation while preserving the inferential value of the counterfactual framework.
These revisions ensure that the manuscript's framing is fully aligned with the reviewer's recommendation. We again thank the reviewer for helping us improve the rigor, clarity, and conceptual precision of the study.

**Comment #2:**

The current modeling approach, which relies on instantaneous covariates, does not account for the temporal auto-correlation inherent in air pollution. The concentration at any given time is also influenced by the emissions and meteorological conditions of previous periods. The choice of a random 80/20 split for model validation may introduce data leakage when evaluating the model performance. A blocked or rolling time-based cross-validation would be more appropriate here.
Separately, uncertainty quantification has been extensively discussed in ML-based atmospheric remote sensing, yet is not addressed in the present manuscript; providing calibrated predictive uncertainty would improve the interpretability of the results.

**Response:**

We appreciate the reviewer's suggestions regarding the treatment of temporal dependence, model validation strategy, and predictive uncertainty.

(1) Temporal autocorrelation and validation scheme.

We agree that air pollution data exhibit temporal autocorrelation and that random splitting may overestimate performance if used for forecasting tasks. However, the purpose of this study is not to predict future concentrations, but to estimate counterfactual business-as-usual (BAU) concentrations under the same meteorological conditions during the Chinese Spring Festival (CSF) period. The model therefore serves as a nonlinear regression tool to capture the contemporaneous relationships between $PM_{2.5}$ and its covariates, rather than the temporal evolution of pollution.

To examine the reviewer's concern, we additionally tested a blocked time-based cross-validation (five-fold TimeSeriesSplit) using only instantaneous covariates. Under this scheme, the cross-validated $R^2$ dropped to -0.50 and the test-set $R^2$ to 0.09 (RMSE ≈ 20 μg/m$^3$), indicating that instantaneous features alone cannot represent temporal dependence.

We then introduced simple lagged predictors (1-3 h lags of $PM_{2.5}$, temperature, and wind speed), after which the time-block cross-validated $R^2$ increased to 0.69 and the test-set $R^2$ to 0.94 (RMSE ≈ 5.4 μg/m$^3$). These results confirm that explicit short-term history improves robustness under sequential validation. Nevertheless, because lagged $PM_{2.5}$ values could leak information from neighboring time steps and obscure the interpretation of "instantaneous" BAU conditions, the main analysis retains the model without lagged terms. The lagged-feature experiment was used only to verify the reviewer's concern regarding temporal dependence, and its results support our choice of focusing on contemporaneous covariates in the counterfactual analysis.

(2) Predictive uncertainty.

Following the reviewer's recommendation, we implemented an ensemble-based bootstrap approach to quantify predictive uncertainty (Text S3). Fifty XGBoost models with identical hyperparameters were trained on bootstrap-resampled training sets, producing an ensemble of counterfactual BAU predictions. The ensemble mean was taken as the point estimate, and the 2.5$^{th}$-97.5$^{th}$ percentile range as the 95% prediction interval. The ensemble-mean predictions achieved an $R^2$ ≈ 0.95 on the test set, slightly lower than the single model but providing calibrated uncertainty estimates. The new Figure S2 in the revision visualizes observed $PM_{2.5}$, ensemble-mean predictions, and their 95 % prediction intervals, along with the deviation (Δ = Observed - Predicted).

These additions address the reviewer's concerns by (a) verifying temporal dependence through lag-feature tests, (b) clarifying the rationale for using a random split in a counterfactual, non-forecasting context, and (c) incorporating calibrated uncertainty quantification to improve interpretability.

We sincerely thank the reviewer once again for these helpful suggestions.

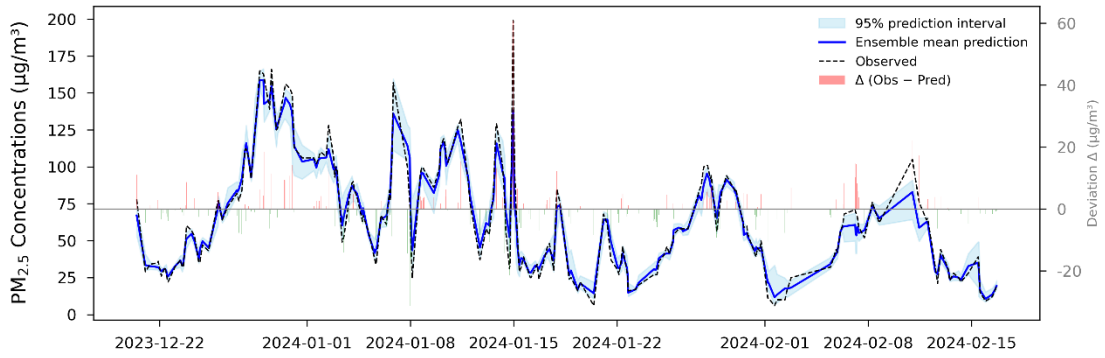


**Figure S2.** Observed $PM_{2.5}$ concentrations and ensemble-based counterfactual predictions from the XGBoost model, together with their 95% prediction intervals and the corresponding deviations (Δ = observed - predicted).

**Comments #3:**

The abstract opens with acute short-term health risks from extremely high $PM_{2.5}$, but the regional result emphasizes an average decrease of 19.0 ± 17.5 μg/m$^3$ over the extended holiday period. These two statements are not contradictory but currently feel weakly connected. Besides that, Section 3.2 (Hangzhou) explicitly reports large concurrent source changes (e.g., vehicles -31%; dust +2790%), yet Section 3.5 ("2+26") estimates fireworks' contribution "under the assumption that emissions from other sources remained unchanged." The authors need to address this inconsistency or provide sensitivity analysis under alternative assumptions.

**Response:**

We thank the reviewer for this insightful comment. We address both aspects raised in this comment below.

(1) Connection between acute $PM_{2.5}$ risks in the abstract and the regional mean decrease. We agree that these two points were initially presented in a way that appeared only loosely connected. Following the reviewer's suggestion, we have revised the abstract to explicitly clarify that the regional mean decrease (-19.0 ± 17.5 μg/m$^3$ across the "2

+ 26" cities) reflects the multi-day reduction in anthropogenic activities during the extended holiday period, whereas the acute short-term health risks refer specifically to the sharp, short-lived $PM_{2.5}$ spikes caused by concentrated fireworks during the peak window on New Year's Eve. These two patterns are therefore not contradictory but operate on different temporal scales. The revised abstract now clearly links the regional baseline reduction with the episodic, fireworks-driven $PM_{2.5}$ peaks.

(2) Consistency between Sections 3.2 and 3.5 regarding the "unchanged sources" assumption. We thank the reviewer for highlighting the apparent inconsistency between the detailed source changes reported for Hangzhou (Section 3.2) and the simplifying assumption used in the regional "2 + 26" analysis (Section 3.5). We agree that assuming "other emission sources remained unchanged" is a strong simplification. In the regional analysis, this assumption was intentionally used as part of the counterfactual BAU framework to isolate the incremental effect of fireworks.

Importantly, as shown in Section 3.2 for Hangzhou, concurrent source changes did occur. Traffic activity decreased substantially (vehicle-related emissions $\approx$ -63%), while local dust emissions increased due to fireworks fallout. Regional gas-phase tracers further confirm this pattern: $NO_2$, a traffic indicator, decreased by about 12%, whereas CO, a combustion tracer, increased by $\approx$ 18% on New Year's Eve. These observations demonstrate that non-fireworks sources were not constant but generally weakened.

Therefore, the "unchanged-source" assumption does not overstate the fireworks-related contribution; instead, it yields a conservative lower-bound estimate. If other anthropogenic emissions declined, the fraction of observed $PM_{2.5}$ attributable to fireworks would, in reality, be even higher. This interpretation is now explicitly stated in Section 3.5, supported by independent evidence from $NO_2$ and CO behavior.

In the revised manuscript, we have:

(a) clarified this rationale in Section 3.5 (lines 400-408),

(b) softened the original wording regarding "unchanged sources" and

(c) explicitly emphasized that the resulting fireworks contribution represents a conservative BAU-based lower-bound estimate.

We sincerely thank the reviewer once again for this important comment, which significantly improved the clarity and consistency of the manuscript.

**Comment #4:**

Section 2.1 requires several clarifications. First, key details for the ERA5 dataset, including its temporal/spatial resolution and a reference link, should be provided in the manuscript or SI (Text S1/Table S1). To address the potential for reanalysis data to smooth over urban-scale extremes, a brief comparison of ERA5 variables against ground-station data would strengthen the analysis. Additionally, the usage of total precipitation (TP) needs to be explained; since it is an accumulated value, please describe any transformation performed to make it suitable for an hourly model. The specific parameters or a reference for Emanuel's saturated vapor pressure formula should be included.

**Response:**

We thank the reviewer for these detailed and constructive suggestions. All points raised have now been fully addressed in the revised manuscript, as summarized below.

(1) ERA5 dataset details

Section 2.1 (lines 98-100) and Table S1 have been revised to explicitly state that hourly single-level ERA5 reanalysis data were used, with a horizontal resolution of 0.25° × 0.25° and hourly temporal resolution. A reference link to the Copernicus Climate Data Store has also been added to ensure full reproducibility.

(2) Evaluation of ERA5 representativeness

To assess the representativeness of the ERA5 variables used in this study (20 December 2023-16 February 2024), we compared hourly ERA5 data at the grid cell centered over downtown Hangzhou with observations from the Hangzhou Xiaoshan International Airport station (~20 km away). The comparison reveals that ERA5 exhibits very small mean biases for near-surface temperature (≈ 0.36 °C) and wind speed (≈ 0.40 m/s), demonstrating its high fidelity in capturing regional meteorological states. While inherent discrepancies exist between a 0.25° grid average and point measurements due to spatial smoothing and local micro-terrain, ERA5's reliability in representing synoptic-scale transitions and atmospheric dynamics is well documented (Hersbach et

al., 2020). Importantly, the objective of this work is to represent regional and synoptic-scale forcing to drive the counterfactual $PM_{2.5}$ prediction. This requires key predictors such as boundary-layer height and solar radiation that are typically unavailable from surface stations. ERA5 provides a spatially and temporally consistent dataset that avoids the localization and discontinuity, such as 3-hour intervals or missing records, often found in ground observations. Given the study's focus on regional-scale meteorology, ERA5 offers a more appropriate and robust basis for the machine-learning-based modeling framework.

3. Processing of total precipitation (TP)

ERA5 total precipitation (TP) represents liquid and frozen water accumulated over the previous hour and is expressed in meters of water equivalent. Since the hourly dataset already reflects accumulation over a one-hour period, no additional differencing or temporal transformation was required prior to its use in the model.

4. Emanuel's saturated vapor pressure formula

The specific formulation and reference have been added to Text S1 in the Supplementary Information. Saturation vapor pressure (E, Pa) is computed using Emanuel's empirical equation (Emanuel, 1994):

$$In\ E = 53.67957 - \frac{6743.769}{T} - 4.8451\ In\ T$$

where T is absolute temperature (K). The equation is applied to 2-m air temperature (T2) and dew-point temperature (D2) to obtain saturation ($E_S$) and actual vapor pressure ($E_a$), respectively, with relative humidity calculated as $RH = (E_a/E_S) \times 100\%$.

Together, these revisions substantially improve the transparency, rigor, and reproducibility of the meteorological data processing procedures in the manuscript. We sincerely appreciate the reviewer's helpful suggestions.

**Ref**:

Hersbach, H., et al. The ERA5 global reanalysis, Quarterly journal of the royal meteorological society, 146, 1999-2049, https://doi.org/10.1002/qj.3803, 2020.

**Comment #5:**

There is the spatial representativeness mismatch between the machine learning model, which uses a 14-site city average, and the DN-PMF analysis, which uses chemical data from a single site. This difference could introduce a bias, particularly for localized sources like fireworks. The authors could discuss this limitation and its potential impact on their findings. Given the team's related work (e.g., Journal of Environmental Sciences), a brief comparison of the methodological advantages and efficiency gains relative to prior work would also help position the contribution.

**Response:**

We thank the reviewer for this important comment, which was also raised by Reviewer #1. Please see our responses to reviewer #1 for details. Below we also provide a concise clarification regarding (1) the spatial representativeness mismatch and (2) the methodological contribution relative to prior work.

1. Spatial representativeness mismatch

We agree that the ML model (city-wide average) and the DN-PMF analysis (single-site composition) operate at different spatial scales. To directly assess this issue, we performed a linear regression between hourly $PM_{2.5}$ at the Wolongqiao site and the 14-site city average over the full study period. The strong correlation ($R^2 = 0.78$, slope = 1.11) demonstrates that Wolongqiao reliably captures the same temporal pollution dynamics as the broader urban area, particularly during winter episodes and the Spring Festival period.

Although absolute levels differ slightly, fireworks emissions during New Year's Eve form a city-wide, highly synchronized plume. Their chemical signatures (e.g., sharp $K^+$ and EC enhancements) are consequently spatially coherent, making the single-site DN-PMF analysis suitable for identifying and tracking this dominant source. We have added this clarification as a discussed limitation in the revised manuscript (Section 3.4, lines 357-370).

2. Methodological advantages relative to prior work

We have also clarified the methodological contribution of this study. Compared with traditional receptor modeling approaches, our ML-based counterfactual framework provides a direct "no-holiday" baseline, enabling cleaner attribution of fireworks impacts. The method requires only routine monitoring and reanalysis data, is highly scalable (as demonstrated in the "2+26" region), and can be applied rapidly to large city networks. This represents a clear efficiency gain relative to chemical-speciation-based source apportionment. The consistency between the ML results and the DN-PMF source apportionment further supports the robustness of the approach.

We have added these points to better position the contribution in the revised manuscript.

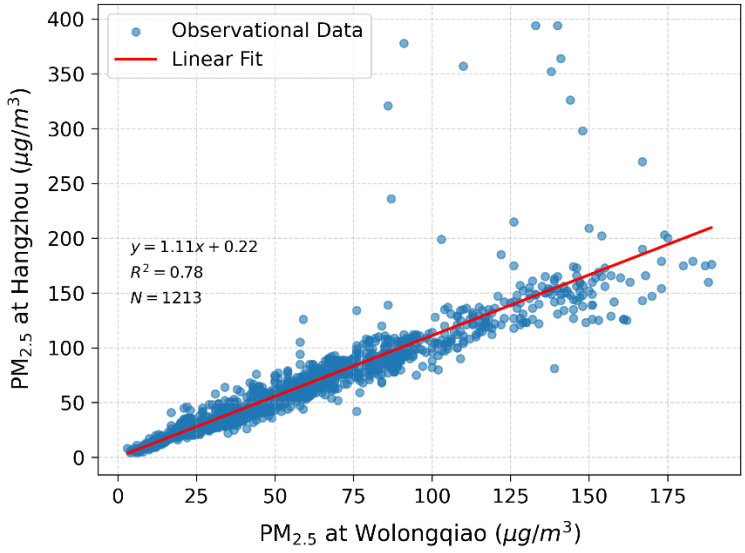


**Figure S9.** Linear regression analysis of hourly $PM_{2.5}$ concentrations between the Wolongqiao site (x-axis) and the Hangzhou city-wide average (14 sites, y-axis) during the study period (2023/12/20 - 2024/2/20). The strong correlation ($R^2 = 0.78$) supports the representativeness of the Wolongqiao site in capturing city-scale pollution trends during the observation period.

**Minor corrections:**

- Line 22: twenty-eight -> 28
- Line 119: meterorology -> meteorology
- Line 164: A -> An

- Line 207: in the midnight of the New Year Eve -> at midnight on New Year's Eve
- Line 244: Please add units for RMSE and MAE.
- Line 260: reliablity -> reliability
- Line 261: techique -> technique
- Line 323: deterioriation -> deterioration
- Table S1: Please use Pa (not pa) for pressure unit.

**Response:**

All minor corrections listed (typos, unit formatting, and grammatical adjustments) have been implemented in the revised manuscript, including those on Lines 22, 119, 164, 207, 244, 260, 261, 323, and in Table S1.