

Dear Dr. Stefano Galmarini and Reviewer #2,

We sincerely appreciate the opportunity to revise our manuscript entitled “Meteorological normalization of surface ozone variability across warm and cold seasons in China during 2015–2024” (egusphere-2026-2846). We thank the Editor and both reviewers for their careful evaluation and constructive comments. These suggestions have helped us clarify the methodology, strengthen the sensitivity analyses, and improve the interpretation and presentation of our results.

In response to the reviewers’ comments, we have revised the manuscript and Supplement as follows.

First, we have expanded the descriptions of the quality control criteria, KNN interpolation procedure, model settings, and SHAP aggregation procedure. We have also clarified the calculation of maximum daily 8 h average O₃ concentrations (MDA8 O₃). We replaced the original random training-testing split with a chronological partition for training, validation, and testing. Hyperparameters were selected using time-ordered validation within the training period. We also clarified the treatment of November–December 2014 in the first cold season and assessed the sensitivity of the national cold-season trend to alternative treatments of these boundary months.

Second, we have strengthened the sensitivity analyses. We examined the influence of the Unix time predictor on meteorologically normalized trends at site 1001A and added a Random Forest comparison. RF used the same data, seasonal definitions, temporal partition, predictors, and normalization procedure as LightGBM. This comparison evaluates the consistency of regional and seasonal normalized trends, annual normalized concentrations, and SHAP results across the two learning algorithms. We report both their common patterns and their differences, including differences in trend significance, variable rankings, and annual magnitudes.

Third, we have added monthly analyses of meteorologically normalized MDA8 O₃ trends during the cold season and compared the November–February and March–April periods. These analyses address whether the cold-season increase is concentrated in spring. The results show that the increase extends beyond March–April, although its rate varies among months and regions.

Finally, we have revised the interpretation of the normalization and SHAP results. The meteorologically normalized long-term component is described as a model-based composite signal. Our analysis cannot independently identify the individual physical processes underlying this signal. We also distinguish the annual SHAP total from the meteorological contribution defined as the difference between observed and meteorologically normalized O₃ concentrations. Their regional annual patterns are compared qualitatively. Explanations involving humidity, radiation, wind, and transport are presented as possible atmospheric contexts for the modeled associations. We explicitly acknowledge the limitations arising from correlated predictors, the lack of long-term and spatially consistent precursor observations, and the absence of dedicated process diagnostics.

We have also reorganized the Supplement to improve clarity and readability. The detailed descriptions of the sensitivity analyses and model comparisons have been moved from the figure and table captions into five dedicated Supplementary Text sections. These sections cover the treatment of November–December 2014, sensitivity to the Unix-time predictor, monthly cold-season trends, comparison between RF and LightGBM, and sensitivity of SHAP results to the learning algorithm. The supplementary figure and table captions have been kept concise, and the corresponding references have been updated throughout the responses.

Below, we provide point-by-point responses to the reviewers' comments. The reviewers' comments are presented in black, and our responses are shown in blue.

Reviewer #2

This manuscript uses site-specific LightGBM models to analyse daily MDA8 O₃ observations from 1173 monitoring sites across China during 2015–2024. Meteorological normalisation is applied by retaining the temporal variables while resampling meteorological conditions, and SHAP is then used to explore regional and seasonal differences in the meteorological factors associated with O₃ variability. The national-scale comparison of warm- and cold-season O₃ is useful, especially the emphasis on the less-studied cold-season increase and its regional differences. These results are relevant to the interpretation of recent O₃ trends and may have implications for season-specific and region-specific control strategies. I think the study has the potential to make a useful contribution, but several aspects of the analysis need to be

clarified or tested more carefully. In particular, I am concerned about how the “non-meteorological component” is interpreted, the use of random train–test splitting for a long-term trend analysis, the uncertainty in the quantitative attribution, and the strength of some of the SHAP-based interpretations.

Specific comments:

1. The normalisation keeps Unix time, DOY and DOW unchanged while meteorological variables are resampled. The resulting normalised series therefore represents the long-term signal retained by the model after averaging over the selected meteorological distribution. I do not think this can be treated as an independently identified “non-meteorological contribution” in a strong sense. This is particularly important for O₃ because the model contains no NO_x, VOC, aerosol or other precursor-related variables. Changes in precursor emissions, NO titration, chemical sensitivity, aerosol effects on photolysis, background O₃ and regional transport may all be partly represented through the time variable or remain mixed within the residual. The authors acknowledge some of this in the uncertainty Sect., but elsewhere the manuscript gives rather precise values for the “non-meteorological contribution” (e.g. 16.1 and 10.2 $\mu\text{g m}^{-3}$), which reads as a stronger attribution than the model can support. Recent work on weather normalisation has also highlighted the potential “proxy trap” between temporal and meteorological predictors, particularly for secondary pollutants such as O₃ (Dai et al., 2025). I suggest that the authors distinguish more clearly between a meteorologically normalised long-term component and a physically identified non-meteorological contribution.

Response:

We thank the reviewer for highlighting the distinction between a meteorologically normalized long-term component and a physically identified non-meteorological contribution. We agree that keeping temporal predictors fixed during meteorological resampling does not identify the physical processes behind the remaining long-term variation. Temporal and meteorological predictors may covary. The retained signal therefore cannot be assumed to represent only non-meteorological influences. We also agree that discussing this limitation only in the limitations Sect. was insufficient while stronger attribution language remained elsewhere.

We have revised the relevant text to refer to the normalized series as a “meteorologically

normalized long-term component”. This component is the mean model-predicted O₃ concentration under resampled meteorological conditions, with temporal predictors held fixed. It may reflect changes in precursor emissions, chemical sensitivity, natural emissions, background O₃, regional transport, and other processes not represented in the model. Because precursor concentrations and emissions were not included as direct predictors, we cannot separate or quantify these influences. We therefore describe the normalized trends as changes that remain evident after meteorological normalization. We do not interpret them as estimates of specific emission changes or independently identified physical contributions.

This interpretation is reflected in the revised Methods, Results, and Conclusions. Sect. 2.2 explains the composite nature of the normalized component and states that it is not a physically isolated estimate of anthropogenic emission changes. Sect. 3.3 describes changes in the normalized series across seasons and regions. The Conclusions emphasize that the O₃ increase remains evident after meteorological normalization. Sect. 3.5 discusses the lack of long-term, spatially consistent precursor observations and explains why temporal predictors cannot distinguish the individual processes. These revisions keep the interpretation within the scope of the normalization framework.

2. The current sensitivity analysis considers the sampling window and the number of resampling iterations, but the main attribution results still depend on one machine-learning algorithm, one meteorological dataset and one normalisation framework. Recent studies of Chinese O₃ have shown that estimated meteorological contributions can vary considerably across datasets and methods (Wang et al., 2025). Earlier machine-learning studies have also produced different estimates of the meteorological contribution over partly overlapping periods. These studies are not directly comparable because the periods and seasonal definitions differ, but they do show that the quantitative attribution is method-sensitive. Before values such as 16.1 $\mu\text{g m}^{-3}$ versus 3.1 $\mu\text{g m}^{-3}$ are treated as robust, some comparison with an alternative model or attribution approach would be helpful. The manuscript should also state clearly whether the meteorological variables are resampled together as a daily meteorological vector or independently. Independent resampling would break the covariance among temperature, humidity, radiation, cloud, precipitation and boundary-layer conditions and could generate physically unrealistic combinations.

Response:

We thank the reviewer for noting that the normalized results may depend on the model and methodological setup. Our original analysis did not assess sensitivity to other meteorological datasets or normalization frameworks. To assess sensitivity to the learning algorithm, we added a Random Forest (RF) analysis. RF used the same station data, seasonal definitions, temporal partition, predictors, and normalization procedure as LightGBM. Only the algorithm and model-specific settings differed (Table S4). This comparison therefore tests algorithm choice, not the choice of meteorological dataset or normalization framework.

Table S4. Hyperparameter settings for the LightGBM and RF models.

Model	Parameters
LightGBM	objective = "regression"; metric = "rmse"; num_leaves $\in$ {16, 24, 31}; n_estimators $\in$ {100, 200, 300}; learning_rate = 0.01; min_child_samples = 30; colsample_bytree = 0.8; reg_alpha = 0.1; reg_lambda = 0.1; random_state = 42
RF	n_estimators = 300; max_features = 3; min_samples_leaf = 3; bootstrap = True; max_depth = None; random_state = 42

RF performed comparably to LightGBM. Its mean site-level testing R^2 was 0.62 in the warm season and 0.70 in the cold season, compared with 0.62 and 0.68 for LightGBM. In the warm season, pooled RMSE and MAE were 21.1 and 16.6 $\mu\text{g m}^{-3}$ for RF, versus 20.9 and 16.4 $\mu\text{g m}^{-3}$ for LightGBM. The corresponding cold-season values were 18.0 and 13.7 $\mu\text{g m}^{-3}$ for RF, versus 18.3 and 14.0 $\mu\text{g m}^{-3}$ for LightGBM (Fig. S5). RF thus provides an alternative model for assessing the sensitivity of the normalized patterns to algorithm choice.

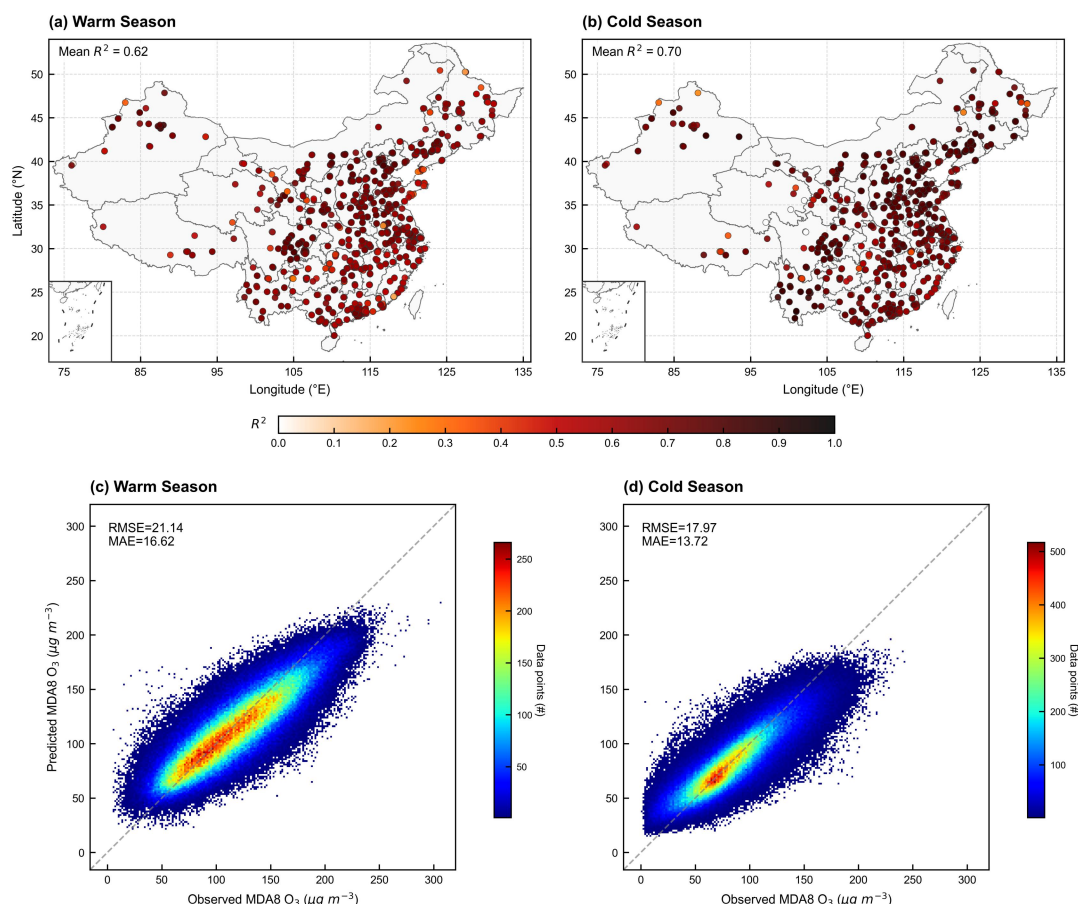


Figure S5. Performance of the RF model. (a, b) Spatial distributions of site-level testing R^2 during the warm and cold seasons, respectively. (c, d) Density scatter plots of observed versus predicted daily MDA8 O_3 during the warm and cold seasons, respectively.

Using the LightGBM results in Fig. 5 as a reference, RF reproduced the main regional and seasonal patterns, as shown in Table S5 and Fig. S7. A detailed comparison of the regional trends and annual normalized concentrations between the two models is provided in Supplementary Text 4. Both models estimated the largest normalized trends in the FWP in both seasons and the weakest cold-season trend in the SCB. Cold-season slopes exceeded warm-season slopes in the NCP, YRD, PRD, and FWP; the SCB showed the opposite pattern. The main difference in statistical significance was in the PRD warm season. The RF trend was significant at $0.7 \mu g^{-3} yr^{-1}$ ($p < 0.001$), whereas the similar LightGBM trend of $0.8 \mu g^{-3} yr^{-1}$ was not ($p = 0.074$).

Table S5. Regional meteorologically normalized MDA8 O_3 Theil–Sen trends estimated with RF and LightGBM.

Region	Season	RF slope ($\mu\text{g m}^{-3} \text{ yr}^{-1}$)	LightGBM slope ($\mu\text{g m}^{-3} \text{ yr}^{-1}$)
NCP	warm	0.8 ($p<0.001$)	1.0 ($p<0.001$)
NCP	cold	1.4 ($p<0.001$)	1.5 ($p<0.001$)
YRD	warm	0.7 ($p<0.001$)	0.9 ($p<0.001$)
YRD	cold	1.2 ($p<0.001$)	1.3 ($p<0.001$)
PRD	warm	0.7 ($p<0.001$)	0.8 ($p=0.074$)
PRD	cold	1.2 ($p<0.001$)	1.3 ($p<0.001$)
SCB	warm	1.0 ($p<0.001$)	1.3 ($p<0.001$)
SCB	cold	0.6 ($p=0.001$)	0.6 ($p=0.002$)
FWP	warm	1.1 ($p<0.001$)	1.5 ($p<0.001$)
FWP	cold	1.8 ($p<0.001$)	2.0 ($p<0.001$)

Annual normalized concentrations also differed between the models, as illustrated in Fig. S7 and further discussed in Supplementary Text 4. Averaged across the five regions and both seasons, RF estimates were $0.92 \mu\text{g m}^{-3}$ higher than LightGBM estimates in 2015–2017. They were $0.14 \mu\text{g m}^{-3}$ lower in 2018–2021 and $0.63 \mu\text{g m}^{-3}$ lower in 2022–2024. The largest single-year absolute difference was $3.28 \mu\text{g m}^{-3}$ in the FWP warm season in 2015. The RF results support the broad regional and seasonal patterns, but trend magnitudes, annual concentrations, and the significance of the PRD warm-season trend depend to some extent on the model. We therefore use this comparison to assess algorithm sensitivity, not to claim that exact quantitative estimates are unchanged across methods.

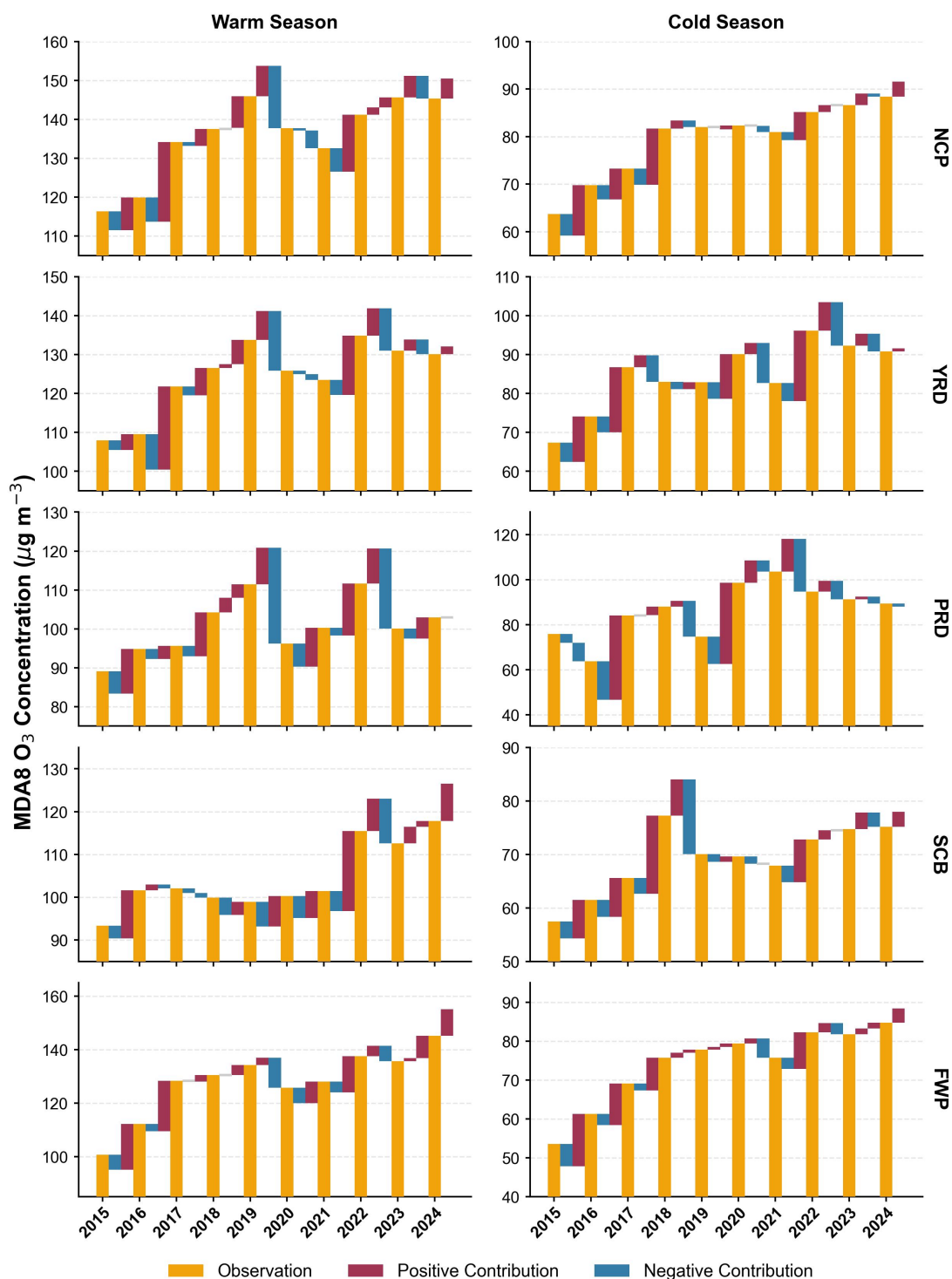


Figure S7. Regional mean MDA8 O₃ concentrations and their year-to-year decomposition in the five key regions during 2015–2024, based on the RF model. The legend is the same as in Fig. 5.

The nine meteorological predictors were sampled jointly as a daily vector. In each draw, all nine values came from the same historical observation day within the circular ± 14 d window centered on the target day of year. The target date's temporal predictors

remained fixed. Joint sampling preserves observed combinations of temperature, humidity, radiation, cloud cover, precipitation, boundary-layer conditions, pressure, and wind. It avoids combining values independently drawn from different dates. We have clarified this procedure in Sect. 2.2. Sensitivity to other meteorological datasets and normalization frameworks remains a limitation.

3 Many of the predictors used here are strongly correlated, particularly temperature, solar radiation, relative humidity, cloud cover, precipitation and boundary-layer height. This may not greatly affect predictive performance, but it can affect how SHAP importance is distributed among predictors. For this reason, statements such as SSRD being the dominant factor at 80.48 % of cold-season sites should be interpreted as model attribution rather than evidence that solar radiation is physically the uniquely dominant control at those sites. This issue becomes more important because the manuscript builds several regional process interpretations around the top-ranked SHAP variable. Some assessment of the stability of these rankings, for example across alternative models, variable groupings or importance metrics, would strengthen this part of the paper. It would also be useful to clarify whether the annual sum of signed meteorological SHAP values is expected to agree quantitatively with the independently calculated meteorological contribution (O-N). If not, the black “net meteorological contribution” lines in Figs. 6 and 7 represent a different quantity from that used in Sect. 3.3 and this distinction should be made clear.

Response:

We thank the reviewer for highlighting the influence of correlated predictors on SHAP interpretation and the distinction between the annual SHAP total and the normalization-based meteorological contribution. Correlations among predictors can affect how SHAP importance is distributed among variables. We therefore interpret a “dominant” variable as the predictor with the highest mean absolute SHAP value at a given site and in a given season. The revised cold-season result identifies SSRD as the highest-ranked variable at 84.5% of sites. This describes its importance in the fitted models rather than establishing it as the uniquely dominant physical control on O₃. The ranking metric is described in Sect. 2.3, and the influence of predictor correlations is acknowledged in Sect. 3.5. The atmospheric processes discussed in Sect. 3.4 are presented as possible explanations of the modeled associations.

To assess the sensitivity of the rankings to algorithm choice, we compared LightGBM with RF using the same nine meteorological predictors and excluding temporal variables. Both models ranked variables by their mean absolute SHAP values. Across sites with results from both models, first-ranked variable agreement was 92.0% in the warm season and 95.4% in the cold season; second-ranked variable agreement was 76.2% and 89.0%, respectively. Both models most frequently ranked T2M first in the NCP and SSRD first in the YRD, SCB, and FWP during the warm season. The main difference occurred in the PRD, where LightGBM most often ranked RH first and RF ranked V10 first. During the cold season, both models most frequently ranked SSRD first in all five regions (Figs. 7a and S9a). The detailed comparison of variable rankings between LightGBM and RF for both seasons is provided in Supplementary Text 5..

The regional annual SHAP totals also showed similar sign patterns between models, although their magnitudes differed (Figs. 6b, 7b, S8b, and S9b). These differences, together with the consistency of the variable rankings, are examined in detail in Supplementary Text 5. For example, the FWP warm-season total in 2015 was $-2.2 \mu\text{g m}^{-3}$ with LightGBM and $-9.8 \mu\text{g m}^{-3}$ with RF. The comparison supports the main seasonal and regional patterns while showing sensitivity in some variable rankings and annual contribution magnitudes. Agreement between algorithms does not eliminate the influence of correlated predictors or establish their independent physical effects.

We have also clarified that the annual SHAP total and $M_i = O_i - N_i$ are distinct quantities. The SHAP total summarizes signed contributions relative to the expected output of the separate model containing only meteorological predictors. In Figs. 6b and 7b, the black lines show the sum of the nine regional annual mean signed SHAP values, while the purple dashed lines show M_i . Their reference baselines and calculation procedures differ, so their magnitudes are not expected to agree quantitatively. We compare their regional annual patterns qualitatively and discuss differences in magnitude and sign in Sect. 3.4. This distinction is explained in Sect. 2.3 and in the captions of Figs. 6 and 7.

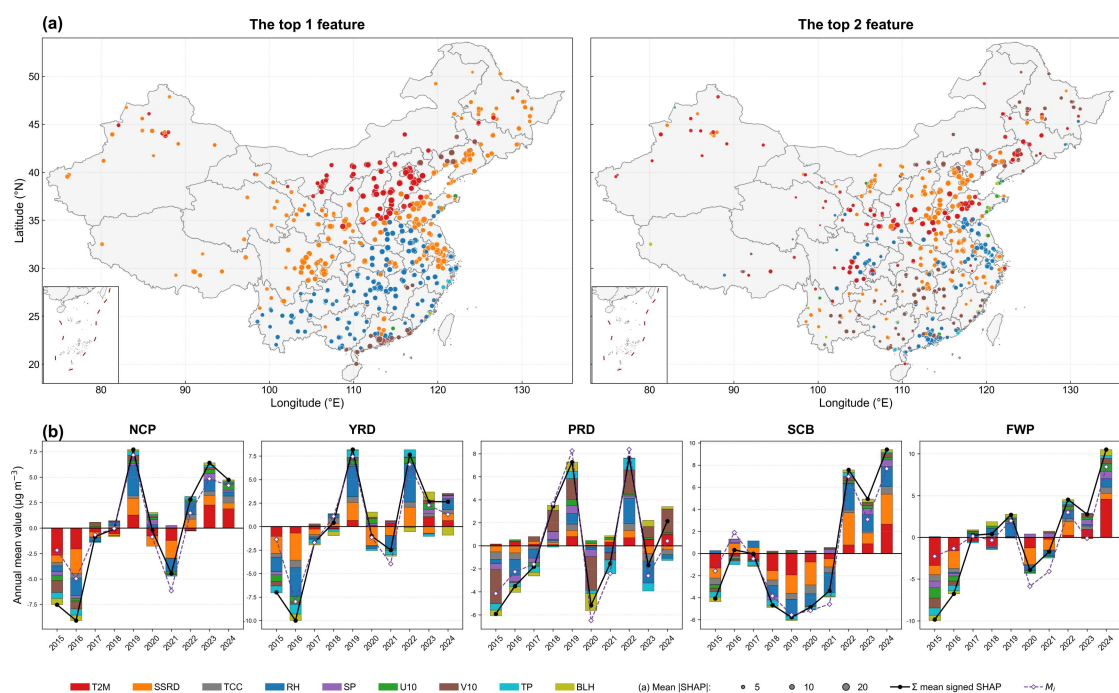


Figure S8. SHAP-based interpretation of warm-season meteorological influences on MDA8 O₃ during 2015–2024 using RF. (a) The two leading meteorological variables at each site, ranked by mean absolute SHAP values. (b) Regional annual mean signed SHAP values. The black line is their sum, and the purple dashed line shows M_i from the normalization analysis for qualitative comparison.

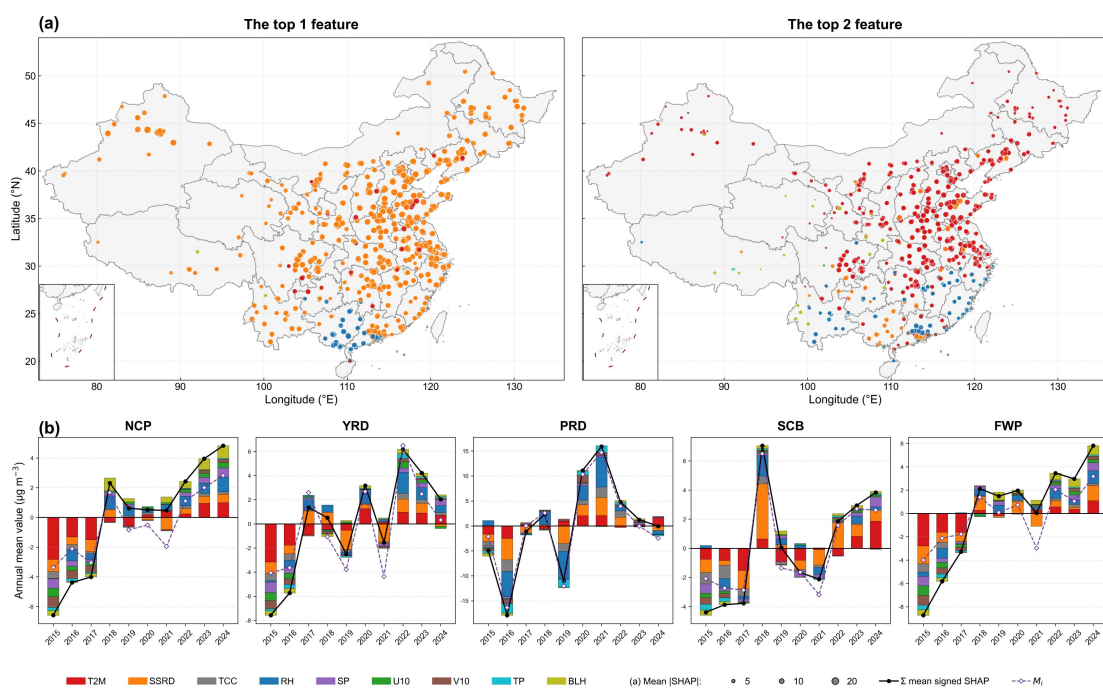


Figure S9. As in Fig. S8, but for the cold season.

4. The observations are described as covering 2015–2024, while the cold season for a given year is defined as November–December of the previous year plus January–April of the named year. Under this definition, the 2015 cold season requires data from November–December 2014. However, the manuscript reports a 2015 cold-season value and uses 2015–2024 to calculate the cold-season trend. It is not clear whether 2014 data were actually used but not described, or whether the 2015 cold season contains only January–April. If the latter is the case, the first cold season is not directly comparable with the later years and this may affect the estimated trend.

Response:

We thank the reviewer for identifying the need to explain the first cold season. We agree that the definition of the 2015 cold season requires November–December 2014 as well as January–April 2015, and that the treatment of the boundary months can affect the estimated trend. We have clarified the treatment of November–December 2014 in Sect. 2.1 of the revised manuscript. The three boundary-period treatments are summarized in Table S1, their effects on the national cold-season trends are shown in Fig. S1, and the detailed sensitivity analysis is provided in Supplementary Text 1.

Of the 1,142 retained monitoring sites, 690 had matched 2014 historical files, while the remaining 452 lacked a matched file. Available station observations were retained. Where daily MDA8 O₃ observations were missing during November–December 2014, the supplemented configuration used estimates from the nearest ChinaHighO₃ Version 3 grid cell when valid grid values were available. Thus, the supplementing procedure was restricted to eligible gaps in the boundary months.

We compared three treatments that differed only in their handling of November–December 2014. Scenario 0 retained available station observations and left missing values unfilled. Scenario 1 retained station observations and supplemented eligible gaps using ChinaHighO₃. Scenario 2 excluded November–December 2014 and retained January–April 2015 for the first cold season. The national cold-season Theil–Sen slopes for Scenarios 0, 1, and 2 were 2.1, 2.2, and 1.6 $\mu\text{g m}^{-3} \text{ yr}^{-1}$, respectively, and all three trends were significant ($p < 0.01$).

The Supplement also compares ChinaHighO₃ estimates with station daily values during November–December 2014. The grid estimates averaged 4.4 $\mu\text{g m}^{-3}$ below station values, with an RMSE of 6.2 $\mu\text{g m}^{-3}$ and an MAE of 4.7 $\mu\text{g m}^{-3}$. The supplemented

values therefore carry uncertainty and do not fully reproduce station-level concentrations. Nevertheless, the national cold-season increase remained positive and statistically significant under all three treatments. These results clarify the boundary-period handling and show that the estimated trend magnitude depends on that handling, while its direction and significance are retained.

Table S1. Treatment of November–December 2014 in the three scenarios.

Scenario	Treatment
Scenario 0	Retain available station MDA8 O ₃ ; leave missing values unfilled.
Scenario 1	Retain station values and fill eligible gaps using the nearest ChinaHighO ₃ grid cell.
Scenario 2	Exclude November–December 2014; retain January–April 2015.

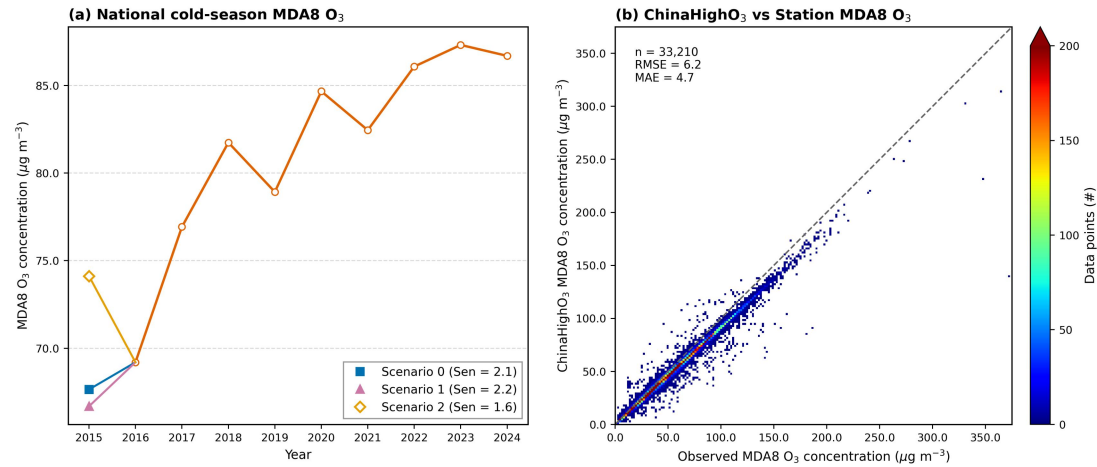


Figure S1. (a) Annual national cold-season MDA8 O₃ during 2015–2024 under the three boundary-period treatments. (b) ChinaHighO₃ estimates versus station daily MDA8 O₃ during November–December 2014.

Minor comments:

1. The definition of a “short gap” in the KNN imputation procedure is not given.

Response:

We thank the reviewer for pointing out that the definition of a “short gap” was unclear.

In this study, a short gap refers to a continuous interval of missing hourly O₃ observations lasting no more than 24 h. KNN interpolation was applied only to gaps within this limit and with sufficient donor support, requiring at least 20 eligible donor samples. Gaps longer than 24 h or lacking sufficient donor support remained missing.

Sect. 2.1 now specifies the 24 h threshold and the donor requirement, clarifying which missing hourly observations were eligible for reconstruction.

2. It is unclear which variables were used to calculate KNN distance and whether they were standardised before applying Euclidean distance.

Response:

We thank the reviewer for requesting this clarification. The features used to calculate KNN distances were CO, PM_{2.5}, NO₂, SO₂, hour of day, day of year, and day of week. The numerical features were standardized before Euclidean distances were calculated, so that differences in their units and scales would not determine the distance calculation.

Reconstruction required at least 20 eligible donor samples. The 10 nearest donors were then selected, and their O₃ observations were combined using inverse-distance weighting to estimate the missing value. Sect. 2.1 now specifies the input features, standardization, distance metric, and neighbor selection and weighting settings, making the KNN reconstruction procedure explicit.

3. Please clarify whether all 1173 sites have sufficiently continuous coverage over the full 2015–2024 period.

Response:

We thank the reviewer for requesting clarification of the temporal coverage. The current analysis uses 1,142 monitoring sites. Sect. 2.1 specifies that sites with more than 10% missing hourly O₃ records over the 2015–2024 study period were excluded, including sites whose prolonged consecutive gaps caused their missing-record fraction to exceed this threshold. The retained sites therefore met a study-period data-coverage criterion, although this does not imply uninterrupted observations at every site or equal completeness in every year.

For the retained sites, KNN reconstruction was restricted to consecutive gaps of no more than 24 h and was performed only when sufficient donor support was available. Longer gaps and gaps without sufficient donor support remained missing. Daily MDA8

O₃ values were subsequently subject to the validity requirements specified in Sect. 2.1: each 8 h window required at least six valid hourly observations, and a daily value was retained only when at least 14 valid windows were available within the specified period.

The November–December 2014 observations required for the first cold season were treated separately from the 2015–2024 coverage assessment. The treatment of these boundary months is clarified in Sect. 2.1 of the revised manuscript, with the three scenarios summarized in Table S1 and the corresponding sensitivity analysis presented in Fig. S1 and Supplementary Text 1. These clarifications specify the final site cohort, the coverage criterion, and the handling of remaining gaps without implying that all retained records were complete.

4. The statement that approximately 89 % of sites showed increasing trends “in both seasons” is ambiguous.

Response:

We thank the reviewer for pointing out this ambiguity. The phrase “in both seasons” could be read as referring to the same subset of sites showing increases in both the warm and cold seasons. We have revised Sect. 3.2 to report the proportions separately for each season.

Among the 1,142 monitoring sites, 88.6% showed increasing trends in observed MDA8 O₃ during the warm season, and 94.3% during the cold season. Statistically significant increases ($p < 0.05$) occurred at 38.7% and 52.0% of all sites, respectively.

The revised wording distinguishes the two seasonal statistics and separates increasing trends from statistically significant increases. It does not report the proportion of sites showing increases simultaneously in both seasons.

5. The relationship among “observation”, positive contribution, negative contribution and normalised concentration in Fig. 5 is difficult to follow from the current caption and legend.

Response:

We thank the reviewer for noting that the relationships among the observed concentrations and the positive and negative contributions in Fig. 5 were difficult to follow. The revised caption identifies the orange bars as the observed annual mean

concentrations, O_i . The meteorological contribution is defined as $M_i = O_i - N_i$, where N_i is the meteorologically normalized annual mean concentration. For the transition from year i to year $i+1$, $\Delta O = O_{i+1} - O_i$, and, following Sect. 2.2, $\Delta N_i = \Delta O - M_i$. Thus, M_i and ΔN_i are the two terms in our defined decomposition and sum to ΔO . The contribution bars are ordered as M_i followed by ΔN_i ; red and blue indicate positive and negative values, respectively. This clarification makes clear what each bar represents and how the signed contributions relate to the observed change.

6. There are several minor typographical and citation-formatting issues, including “variablesthat”, “severalsources”, “asstatistical”, inconsistent spacing around year ranges, and the incomplete citation “Wang, L. et al.” in the Introduction.

Response:

We thank the reviewer for identifying these typographical and citation-formatting issues. We corrected the missing spaces in “variables that,” “several sources,” and “as statistical,” and standardized the presentation of year ranges. We also completed the Introduction citation previously given as “Wang, L. et al.”; it now reads “Wang L. et al. (2025),” with the full bibliographic entry provided in the References. These revisions address the examples raised by the reviewer and improve consistency in the text and references.

7. “China National Environmental Monitoring Centre” and “China National Environmental Centre” are used inconsistently.

Response:

We thank the reviewer for noting the inconsistent name of the monitoring centre. We have standardized the name as “China National Environmental Monitoring Centre” throughout the manuscript and introduced the abbreviation CNEMC at its first mention. We checked the manuscript for other occurrences to ensure consistent naming.

8. Some of the proposed physical explanations, particularly those related to RH and wind, are phrased more strongly than would be supported by a statistical association alone.

Response:

We thank the reviewer for noting that some of our physical explanations, particularly

those involving relative humidity and wind, could be read as stronger than the evidence supports. We agree that SHAP results describe model-attributed associations and do not, by themselves, establish direct physical or causal effects. In Sect. 3.4, we therefore frame the RH- and wind-related results as associations identified by the fitted model. For example, we state that RH may covary with cloud cover, solar radiation, and other meteorological conditions, and that the association of the meridional wind component (V10) may relate to regional transport and ventilation. These are presented as possible atmospheric contexts, not as uniquely identified mechanisms. The cited studies provide context for these interpretations, rather than independent validation of the mechanisms.

Sect. 3.5 further clarifies that SHAP values should not be interpreted as direct physical or causal effects and that correlations among meteorological predictors can affect their relative SHAP values. Figures S8 and S9 present the RF-based SHAP results for the warm and cold seasons, respectively. Supplementary Text 5 provides a detailed comparison with LightGBM, including the agreement in variable rankings and the similarities and differences in regional annual signed SHAP contributions. These revisions retain the identified seasonal and regional SHAP patterns while keeping their interpretation within the scope of the analysis.