

Dear Dr. Stefano Galmarini and Reviewer #1,

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 #1

This manuscript applies LightGBM-based meteorological normalization and SHAP analysis to investigate warm- and cold-season ozone variability across China during 2015–2024. The national-scale dataset and direct comparison between the two seasons are valuable, and the results may provide useful information for ozone management. However, the main quantitative interpretation is still not sufficiently supported by the analysis results. I think major revision is required before the manuscript can be considered for publication.

Specific comments:

1. Chapter 2: Details about the methodology are required. Please clarify the variables used to calculate the KNN interpolation (maximum gap length, proportion of imputed

observations by site and year,...), MDA8 O₃ (the minimum number of valid hourly values, ...), etc.... Also, the detailed setup of hyperparameters for LightGBM and for SHAP should be described.

Response:

We thank the reviewer for this suggestion. We have expanded Sect. 2.1 to describe KNN interpolation and MDA8 O₃ calculation, Sect. 2.2 and Table S4 to detail LightGBM training and parameter selection, and Sect. 2.3 to clarify the SHAP setup and aggregation procedure.

For KNN interpolation, Euclidean distances were calculated from standardized CO, PM_{2.5}, NO₂, SO₂, hour, day of year, and day of week. Candidate donors were restricted to ± 2 h of the target hour and ± 45 d of its day of year, with at least 20 eligible donors required. Missing values were estimated from the 10 nearest donors using inverse-distance weighting. Only consecutive gaps of up to 24 h were filled; longer gaps and those with insufficient donors remained missing. Across the 1,142 retained sites, imputed values accounted for 4.58% of all expected hourly O₃ records during 2015–2024, pooled across sites and years.

Daily MDA8 O₃ was the maximum 8 h rolling mean with window end times between 08:00 and 24:00 local time (UTC+8). Each window required at least six valid hourly observations, and a daily value was retained only when at least 14 valid windows were available.

LightGBM models were trained separately for each site and season. Data from 2015–2021 were used for training, with hyperparameters selected by year-ordered validation within that period using the lowest mean RMSE; 2022 was used for validation and 2023–2024 for independent testing. Table S4 lists the candidate values for num_leaves (16, 24, 31) and n_estimators (100, 200, 300), together with the fixed settings: objective = "regression", metric = "rmse", learning_rate = 0.01, min_child_samples = 30, colsample_bytree = 0.8, reg_alpha = 0.1, reg_lambda = 0.1, and random_state = 42. The selected combination could differ among sites and seasons.

For SHAP analysis, a separate LightGBM model used only the nine meteorological predictors, without temporal variables, and adopted the hyperparameters selected for the corresponding site and season in Sect. 2.2. SHAP values were referenced to the

explainer's expected model output. Mean absolute values were used to rank variable importance, while signed values described the direction of contributions to model predictions. Daily signed values were averaged for each site, season, and year, then across sites within cities and equally weighted cities within regions. The black lines in Figs. 6b and 7b show the sums of the nine regional annual mean signed SHAP values.

2. Line 128-: Unix time was included as a predictor to represent long-term changes and was kept fixed during meteorological resampling. However, long-term changes in temperature, radiation, humidity, or other meteorological variables can also be correlated with Unix time. Therefore, the model may assign part of the meteorological trend to the temporal predictor, and the normalized component is not necessarily a uniquely identified non-meteorological signal. I think sensitivity tests excluding Unix time, or using alternative representations of the long-term trend, are necessary before concluding that non-meteorological changes were dominant.

Response:

We thank the reviewer for raising this concern. We agree that Unix time may covary with long-term changes in meteorological conditions, and keeping it fixed during resampling does not ensure that the normalized component represents a uniquely identified non-meteorological signal.

To examine the influence of this predictor, we repeated the meteorological-normalization analysis at site 1001A with and without Unix time. Both configurations used the same seasonal data and paired meteorological samples. Hyperparameters were selected separately for the two configurations using the same parameter grid and year-ordered validation procedure within the training period. The comparison is shown in Fig. S2, with the experimental setup and detailed results described in Supplementary Text 2.

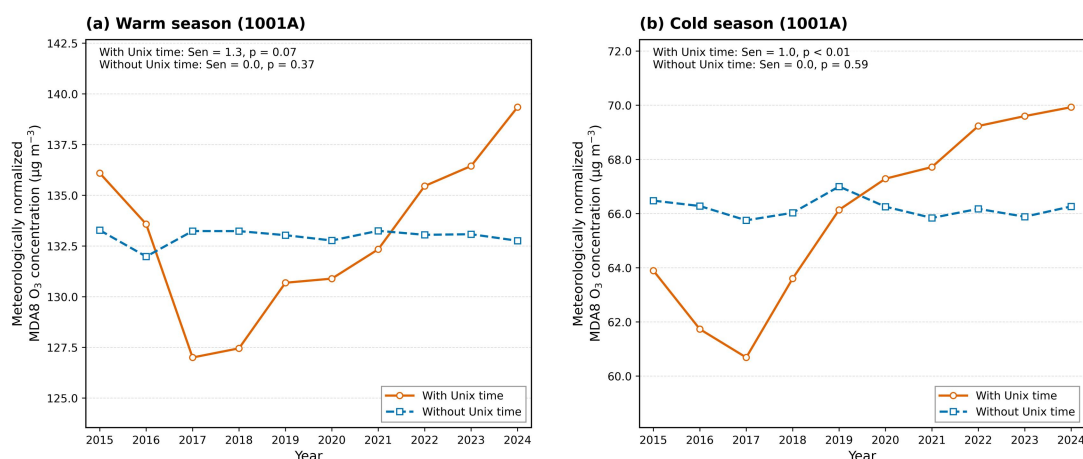


Figure S2. Sensitivity of meteorologically normalized MDA8 O₃ concentrations to the Unix-time predictor at site 1001A. Panels (a) and (b) show the annual series for the warm and cold seasons, respectively.

With Unix time retained, the warm-season Theil–Sen slope was $1.3 \mu\text{g m}^{-3} \text{ yr}^{-1}$ ($p = 0.07$), and the cold-season slope was $1.0 \mu\text{g m}^{-3} \text{ yr}^{-1}$ ($p < 0.01$). Without Unix time, both slopes were nearly zero, with $p = 0.37$ in the warm season and $p = 0.59$ in the cold season. The marked reduction shows that the estimated normalized trend is sensitive to the temporal predictor. This comparison illustrates the role of the temporal representation in the analysis, while the interpretation of the normalized component remains subject to covariation between time and meteorology.

The revised manuscript therefore treats the normalized series as a meteorologically normalized long-term component. Sect. 2.2 refers to the sensitivity test and states that Unix time cannot be interpreted as a uniquely non-meteorological signal. The same Sect. explains that the normalized component may reflect changes in precursor emissions, background O₃, regional transport, natural emissions, and other processes not explicitly represented in the model. Sect. 3.5 also discusses this limitation.

3. Line 133-: The daily dataset was randomly divided into training and testing subsets at an 80:20 ratio. However, ozone and meteorological variables have strong temporal autocorrelation, so adjacent days or the same pollution episode can be included in both datasets. This may overestimate the model's performance. I recommend using cross-validation based on selected time series blocks (not randomly).

Response:

We thank the reviewer for highlighting the potential effect of temporal autocorrelation on model evaluation. We agree that the original 80:20 random split could place adjacent observations, or observations from the same pollution episode, in both the training and testing subsets and thereby overestimate predictive performance.

To address other reviewer comments concerning the robustness of the results, we also evaluated a random forest (RF) model. We replaced the random split with the same chronological, year-blocked partition for LightGBM and RF in both seasons: data from 2015–2021 were used for training, 2022 for validation, and observations assigned to the 2023–2024 seasonal test period were held out for testing. Within the training period, LightGBM hyperparameters were selected separately for each site and season through forward-chaining validation. Each validation year was evaluated using a model trained only on preceding years, and the combination with the lowest mean RMSE was retained. Sect. 2.2 describes the temporal partition and parameter-selection procedure; Table S4 lists the candidate hyperparameters and fixed model settings.

Under the revised partition, the mean site-level testing R^2 values for LightGBM were 0.62 and 0.68 in the warm and cold seasons, respectively. Across all available test observations, the corresponding RMSE values were 20.9 and 18.3 $\mu\text{g m}^{-3}$, and the MAE values were 16.4 and 14.0 $\mu\text{g m}^{-3}$. Fig. 1 shows the spatial distributions of site-level testing R^2 and the observed–predicted relationships. This evaluation replaces the results based on the original random split.

4. Lines 270–: The cold season is defined as November–April, which combines winter with the spring. Because ozone can increase rapidly during March and April, especially in southern China, the reported cold-season trend may be mainly driven by springtime changes. Please provide monthly normalized trends or at least separate November–February and March–April analyses.

Response:

We thank the reviewer for pointing out that the November–April definition combines months with different ozone characteristics and may obscure differences between winter and spring. We agree that the cold-season mean alone cannot establish whether the reported increase occurs across the season or is concentrated in March–April, particularly in southern China.

To address this concern, we analysed monthly trends in meteorologically normalized MDA8 O₃ for the national series and five key regions during 2015–2024. For each calendar month within the cold season, available daily normalized concentrations were averaged within each cold-season year to obtain a 10-year series. Trends were estimated using Theil–Sen slopes, and their significance was assessed using the Mann–Kendall test. We also examined the November–February and March–April means separately. The monthly slopes are presented in Fig. S6, and the detailed analysis, including the comparison between November–February and March–April, is provided in Supplementary Text 3.

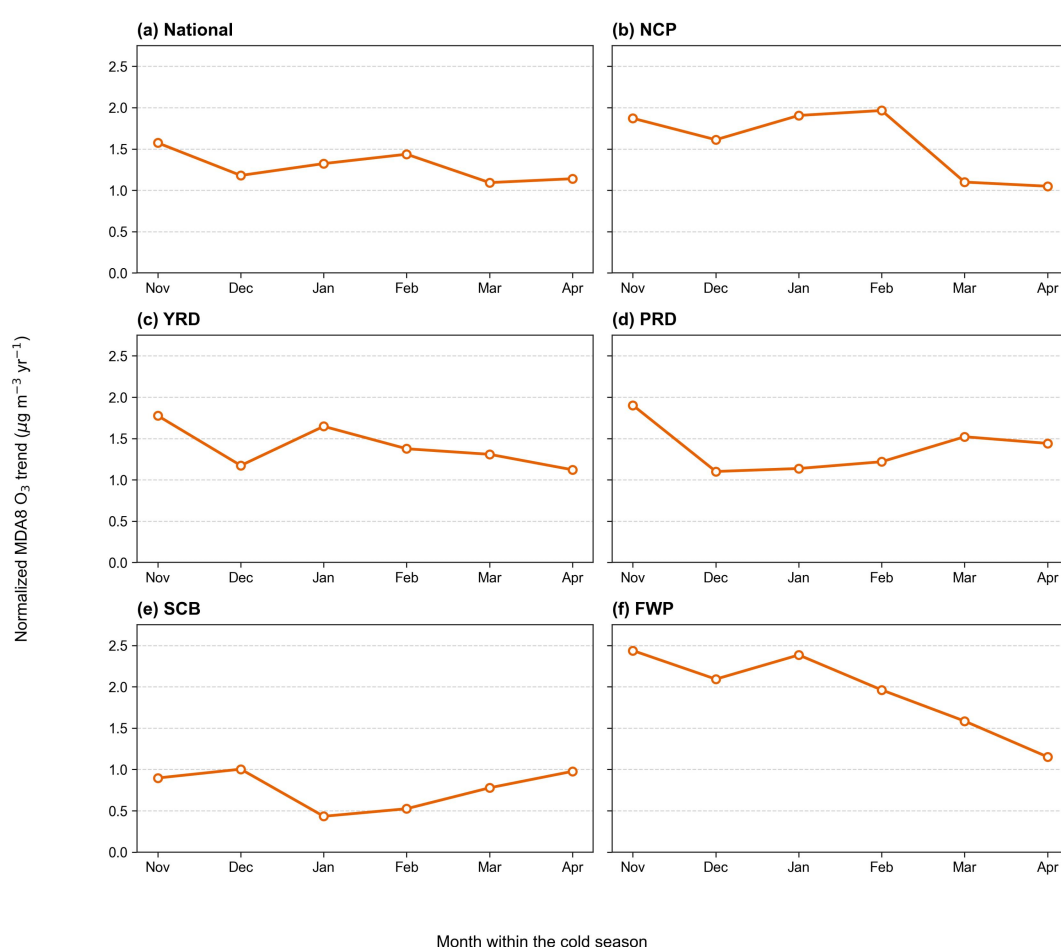


Figure S6. Monthly Theil–Sen slopes of meteorologically normalized MDA8 O₃ for the national series and five key regions. Points show month-specific slopes estimated from the 2015–2024 cold-season series; lines only connect the monthly estimates.

Meteorologically normalized MDA8 O₃ increased significantly in every cold-season month nationally and in all five regions ($p < 0.05$). National monthly slopes ranged

from $1.1 \mu\text{g m}^{-3} \text{ yr}^{-1}$ in March to $1.6 \mu\text{g m}^{-3} \text{ yr}^{-1}$ in November. The national November–February slope was $1.4 \mu\text{g m}^{-3} \text{ yr}^{-1}$, compared with $1.1 \mu\text{g m}^{-3} \text{ yr}^{-1}$ for March–April, and both trends were significant ($p < 0.001$). Thus, the national increase was already evident during the four months preceding March, with a larger estimated slope for November–February than for March–April. This comparison does not support interpreting the national cold-season increase as a trend concentrated in the spring months.

The monthly pattern nevertheless differed among regions. In the FWP, the November, December, and January slopes were 2.4 , 2.1 , and $2.4 \mu\text{g m}^{-3} \text{ yr}^{-1}$, respectively, and were the largest among the five regions in those months. To address the reviewer’s concern about southern China, we also examined the PRD results. Its March and April slopes were 1.5 and $1.4 \mu\text{g m}^{-3} \text{ yr}^{-1}$, respectively, with the April slope being the largest among the five regions. However, the PRD also showed significant positive trends in November, December, January, and February. The springtime increase in this region therefore occurred alongside increases earlier in the cold season. These regional differences show why the monthly results are needed to interpret the seasonal mean.

We have revised Sect. 3.3 to state that the normalized O_3 increase extends beyond March–April, while its magnitude varies among months and regions. Figure S6 presents the monthly trends for the national series and five key regions, while Supplementary Text 3 provides the detailed results and the comparison between November–February and March–April. Together, these results show that the increase is present throughout the cold season, although a single seasonal mean cannot fully describe the differences in its monthly and regional rates.

5. Lines 307-: The meteorological contribution defined by the normalization method, $\text{O}-\text{N}$, and the annual sum of meteorological SHAP values are not necessarily the same quantity because they are calculated relative to different baselines. I think these two estimates should be directly compared for each year and region. Without this comparison, the net meteorological contributions in Figs. 6b and 7b may be interpreted as equivalent to the decomposition in Fig. 5, although this equivalence is not guaranteed.

Response:

We thank the reviewer for this point. The normalization-based meteorological contribution $M_i = \text{O}_i - \text{N}_i$ and the annual SHAP total use different reference baselines. M_i

is the difference between observed and meteorologically normalized annual mean MDA8 O₃, whereas the SHAP total sums the contributions of nine meteorological predictors relative to the expected output of a separate model. The two quantities should therefore not be treated as equivalent.

We now show both for each region and year in Figs. 6b and 7b: the purple dashed line represents M_i , and the black line represents the SHAP total. To obtain the latter, daily signed SHAP values were averaged by site, season, and year, then within cities and across equally weighted cities within each region before summing the nine variables. The two series show broadly similar annual patterns, but their magnitudes and some of their signs differ. We therefore compare their patterns qualitatively and clarify in Sect. 2.3 and 3.4 and the figure captions that the SHAP total is not equivalent to the normalization-based decomposition in Fig. 5.

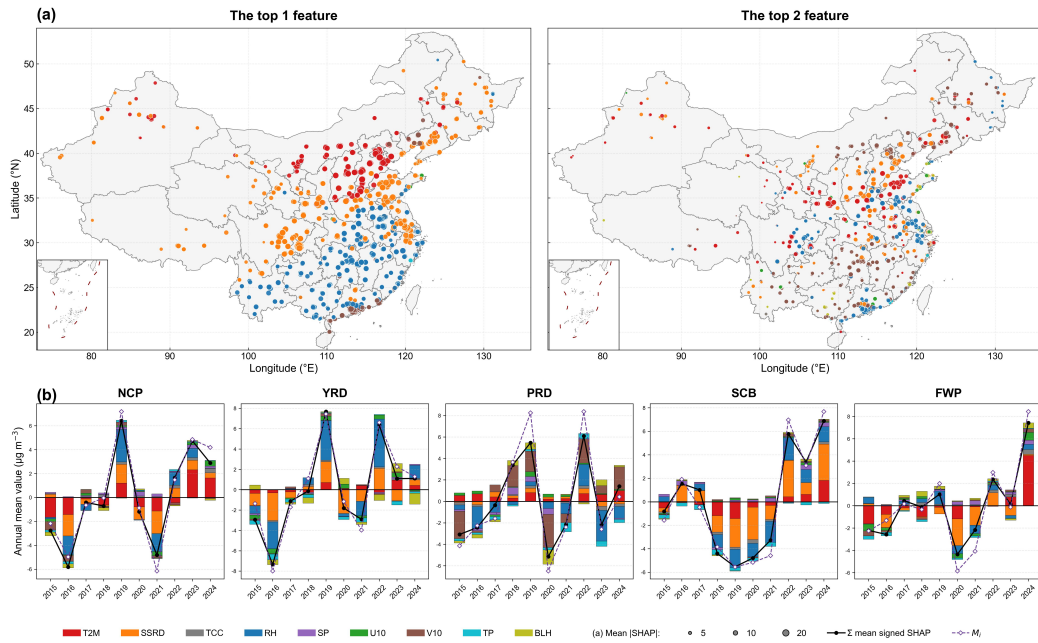


Figure 6. SHAP-based interpretation of warm-season meteorological influences on MDA8 O₃ during 2015–2024. (a) Spatial distributions of the two leading meteorological variables at monitoring sites. Colors identify variables and marker sizes show mean absolute SHAP values. (b) Regional annual mean signed SHAP values for the nine meteorological predictors. Colored bars represent individual variables; black lines show their sum, and purple dashed lines show $M_i = O_i - N_i$ from meteorological normalization for qualitative comparison.

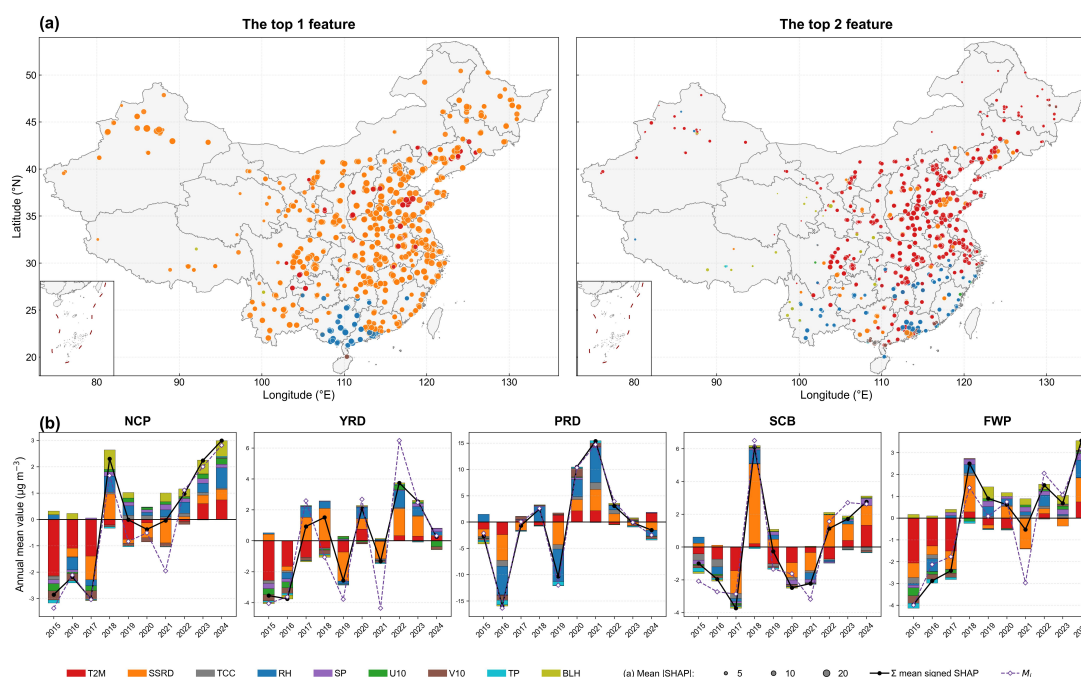


Figure 7. As in Fig. 6, but for the cold season

6. Lines 362-: Mechanistic interpretations are stronger than the presented analysis can support. Direct NO_x/VOC observations and transport or chemistry diagnostics were not included. I think these should be clearly presented as hypotheses unless they are supported with additional chemical observations or process-based modeling.

Response:

We thank the reviewer for this comment. We agree that these mechanistic explanations remain hypotheses because the study lacks long-term and spatially consistent NO_x and VOC observations as well as dedicated transport and chemistry diagnostics. SHAP values describe associations between meteorological predictors and model-predicted MDA8 O₃; they do not establish physical causality.

We have revised Sect. 3.3 and 3.4 to distinguish the analytical findings from possible mechanistic explanations. Precursor changes, reduced NO titration, background O₃, and regional transport are discussed as potential influences on the long-term trends, rather than drivers identified by this study. For the warm-season coastal regions, the relationships of RH and V10 with predicted O₃ are described as model-based associations; cloud cover, radiation, mixing, and ventilation are discussed only as potentially covarying conditions. In the cold-season PRD, interannual changes in

signed RH SHAP values indicate changes in the modeled humidity association, not evidence of a specific chemical or transport pathway.

Sect. 2.3 and 4 clarify these limits of SHAP interpretation and acknowledge that the present analysis cannot resolve individual chemical mechanisms or source contributions. Longer precursor records and chemical transport modeling would be needed to evaluate the proposed explanations.