# Responses to Reviewers' comments

To the esteemed Editor and Reviewers,

We sincerely thank the reviewer for the valuable comments and for the time spent reviewing the manuscript. We have tried our best to revise the manuscript according to the detailed suggestions and addressed each comment as carefully as possible. We believe these revisions improve the completeness and scientific rigor of the manuscript.

Please find the point-to-point responses to the reviewers' comments as follows:

Reviewers' comments are in black.

Author's responses are in blue color.

Changes in the manuscript are in red color.

Yours Sincerely,

Weihua Chen

On behalf of the authors

# Reviewer:1

**Comments to the Author:**

1. This paper develops an explainable classification machine learning model with FNR-divided ozone photochemical regimes as a label to quantify the impact of meteorology and emissions on the ozone formation regimes (VOC-limited, $NO_x$-limited, and transitional regimes). The authors provide a comprehensive assessment of the spatiotemporal evolution, seasonality, and the COVID-19 lockdown response of OFS over China, and reveal an apparent two-stage regime shift during 2005–2024. However, the core methodological logic is not fully convincing. Because the regimes are derived solely from the satellite $HCHO/NO_2$ ratio and a prescribed threshold, they do not explicitly encode meteorological effects. ML attribution therefore quantifies drivers of an FNR-based classification proxy rather than providing a physically grounded diagnosis of OFS. This disconnect weakens the manuscript's ability to address the key gap stated in the Introduction regarding meteorological impacts on OFS. Moreover, FNR thresholds are known to be region-dependent. Applying the uniform national thresholds potentially introduces non-negligible uncertainty, affecting OFS analysis. Therefore, the authors should clarify the conceptual rationale of this framework and demonstrate robustness to threshold/label uncertainty before it can be considered for publication.

Response: We sincerely thank the reviewer for this important and constructive comment. We agree that the original manuscript did not sufficiently distinguish between the satellite-based diagnosis of ozone ($O_3$) formation sensitivity (OFS) and the

subsequent machine learning (ML) attribution analysis. In particular, the original presentation may have given the impression that the ML framework provides an independent physical diagnosis of OFS, whereas its actual role is to explain the statistical associations between FNR-derived OFS states and their meteorological and emission-related covariates. We have substantially revised the manuscript to clarify this conceptual boundary and to address the uncertainty associated with threshold-based regime classification.

1. We clarified the conceptual role of the framework.

In the revised manuscript, we explicitly state that OFS is diagnosed using the satellite $HCHO/NO_2$ ratio (FNR) together with prescribed threshold ranges, whereas the ML model is used only to attribute the relative importance of meteorological and emission-related factors associated with the occurrence and evolution of these FNR-derived OFS states. Therefore, the ML results are interpreted as statistical attribution rather than an independent physically based diagnosis or strict causal inference. This clarification has been incorporated throughout the manuscript, especially in **Section 2.2.4** and **Section 3.4**.

2. We revised the manuscript wording to avoid overstating the role of machine learning.

To better reflect the actual scope of the study, we moderated several claims throughout the manuscript and revised the title to avoid implying that the ML model itself diagnoses OFS. The revised title is:

Long-Term Ozone Formation Sensitivity in China: Spatiotemporal Evolution and

Machine Learning Attribution.

We also expanded the discussion of limitations to emphasize the exploratory nature of the ML framework and the dependence of attribution results on threshold-defined OFS labels (**Section 4, Lines 637–645**).

3. We addressed the reviewer's concern regarding threshold and its uncertainty.

We have clarified that the revised manuscript does not apply a uniform national FNR threshold. Instead, region-specific FNR threshold ranges were derived for different regions of China. In addition, we conducted new sensitivity analyses under multiple threshold scenarios to evaluate the robustness of OFS classification and its derived spatiotemporal patterns to threshold uncertainty. These revisions have been incorporated in **Section 2.2.1** and **Section 3.2**.

**Specific comments:**

2. Title and claimed contribution (OFS diagnosis vs ML attribution): The title is "Explainable Machine Learning diagnosis of Ozone Formation Sensitivity in China", however, OFS is diagnosed using the FNR approach, while the machine learning component is primarily used for driver attribution. The title should therefore be revised to avoid misleading readers about the role of ML in OFS diagnosis. Relatedly, the statement in L39 ("Robust identification of OFS") does not appear to be a contribution of this study. The main contribution is the attribution analysis, not the identification/diagnosis itself.

Response: We sincerely thank the reviewer for this careful and constructive

comment. We agree that the original title could give the impression that ML was used directly to diagnose OFS, whereas in this study OFS is primarily diagnosed using the satellite-based FNR, and the ML is mainly used to attribute the relative roles of meteorological and emission-related factors in the evolution of the FNR-derived OFS states. To avoid this potential misunderstanding, we have revised the title as follows:

Long-Term Ozone Formation Sensitivity in China: Spatiotemporal Evolution and Machine Learning Attribution

We have also revised the Introduction to more clearly distinguish the respective roles of FNR-based OFS classification and ML-based attribution of controlling factors. In particular, we no longer frame "robust identification of OFS" as a central contribution of this study. We have revised the text in Lines 42–43:

Consequently, a reliable and regionally appropriate classification of OFS regimes, together with a clear attribution of their controlling factors, is essential for developing effective, region-specific multi-pollutant control strategies.

We also appreciate the reviewer's comment regarding the statement on "robust identification of OFS." We agree that the main novelty of this study does not lie in proposing a completely new OFS diagnostic method. Instead, the methodological advance of this study is that, rather than directly adopting the widely used fixed threshold range of Duncan et al. (2010), we derived region-specific FNR threshold ranges for different regions in China based on satellite observations and surface $O_3$ data and then used these thresholds to classify OFS regimes. In this sense, the "robust identification" refers to improving the regional appropriateness and reliability of FNR-

based OFS classification in China, not to the development of a fundamentally new OFS diagnostic framework. We have clarified it in Lines 90–97:

Rather than proposing a fundamentally new OFS diagnostic framework, we derive region-specific FNR threshold ranges for different regions of China based on satellite observations and surface $O_3$ data and use these thresholds to classify OFS regimes. We further couple a Random Forest (RF) model with SHapley Additive exPlanations (SHAP) to statistically characterize how meteorological and emission-related factors are associated with the occurrence and long-term evolution of the FNR-derived OFS states. In this framework, OFS is diagnosed using the satellite-based FNR indicator, whereas the machine learning analysis is used for driver attribution rather than to diagnose OFS directly. This framework provides a more regionally appropriate basis for OFS classification in China and offers an interpretable perspective on the relative roles of emissions and meteorology in OFS evolution.

3. A major concern is the conceptual linkage between the FNR-based OFS diagnosis and the subsequent attribution to meteorology and emissions. The OFS derived from FNR is diagnosed purely from satellite-observed precursor ratios ($HCHO/NO_2$) and does not explicitly account for meteorological influences. Yet later, the manuscript attributes this OFS to both meteorology and emissions. While the implementation and some results may appear plausible, the causal/diagnostic logic is unclear: OFS FNR is, by definition, only a function of two precursors plus a chosen threshold, so it is not obvious how this framework can address the stated problem in L70-71 or the key gap

raised in the Introduction (4th paragraph) regarding meteorological impacts on OFS. More broadly, OFS is already a relatively simple regime concept. The current ML framework appears to "complexify" it without a clear practical payoff. The authors should clarify the design rationale of this modeling/analysis framework and explicitly explain how it can diagnose meteorological impacts on OFS if the OFS metric itself is precursor-ratio-based.

Response: We sincerely thank the reviewer for this important and thoughtful comment. We agree that the conceptual linkage between the FNR-based OFS diagnosis and the subsequent attribution to meteorology and emissions needed to be clarified more explicitly in the original manuscript.

Our intention is not to suggest that meteorology is directly encoded in the FNR definition itself. Rather, OFS is first diagnosed using the satellite-based precursor ratio ($HCHO/NO_2$), while the ML analysis is then used to examine which combinations of meteorological and emission-related conditions are statistically associated with the occurrence and transition of these FNR-derived OFS states. In this sense, the ML framework does not redefine OFS but instead treats the diagnosed OFS categories as integrated response states and statistically maps them to their potential controlling factors. We have now clarified this framework explicitly in Lines 90–97 of the revised manuscript:

Rather than proposing a fundamentally new OFS diagnostic framework, we derive region-specific FNR threshold ranges for different regions of China based on satellite observations and surface $O_3$ data and use these thresholds to classify OFS regimes. We

149 further couple a Random Forest (RF) model with SHapley Additive exPlanations
150 (SHAP) to statistically characterize how meteorological and emission-related factors
151 are associated with the occurrence and long-term evolution of the FNR-derived OFS
152 states. In this framework, OFS is diagnosed using the satellite-based FNR indicator,
153 whereas the machine learning analysis is used for driver attribution rather than to
154 diagnose OFS directly. This framework provides a more regionally appropriate basis
155 for OFS classification in China and offers an interpretable perspective on the relative
156 roles of emissions and meteorology in OFS evolution.

157 More specifically, although OFS is defined based on precursor ratios, the
158 spatiotemporal distribution of those ratios is itself influenced by both emissions and
159 meteorological conditions. Meteorological factors can affect precursor emissions,
160 chemical transformation, transport, dilution, and boundary-layer mixing, thereby
161 indirectly modulating the probability of different OFS states (Jin et al., 2013; Vazquez
162 Santiago et al., 2024). We have therefore revised the manuscript to state clearly that
163 meteorological factors affect OFS indirectly—not through the FNR definition itself, but
164 by modulating the occurrence probabilities of different OFS states. We also now
165 explicitly note that the attribution results should be interpreted as statistical associations
166 rather than strict causal relationships in Lines 229–234 of the revised manuscript:

167 It should be noted that, although OFS is diagnosed here using the FNR indicator,
168 the spatiotemporal distribution of FNR-derived OFS states is jointly shaped by
169 precursor emissions, chemical transformation, transport, dilution, and boundary-layer
170 mixing (Kleinman et al., 2005; Liu and Shi, 2021). Meteorological conditions therefore

do not enter the OFS definition directly but can influence the probability of different OFS states by modulating the underlying precursor environment. Based on this rationale, we treat the diagnosed OFS regimes as response classes and use RF to statistically map their relationships with meteorological and emission-related predictors.

In addition, to address the reviewer's concern about the practical value of the framework, we have clarified that the purpose of the ML analysis is not to "complexify" OFS, but to provide a decomposable and interpretable attribution framework for understanding the long-term evolution and regional heterogeneity of OFS across China, especially under changing emission backgrounds and meteorological variability. This has also been emphasized in Lines 622–630 of the revised manuscript:

Although the definition of OFS is based on the relative proportions of precursors, its spatiotemporal patterns actually reflect the combined effects of emission sources and meteorological conditions. Therefore, we treat OFS types as integrated response states and apply ML techniques to statistically model their relationships with meteorological parameters and emission variables. Within this framework, meteorological factors indirectly influence the formation of OFS by modulating the probabilities of different OFS states, rather than directly altering the FNR definition itself. Our focus is on the FNR-derived OFS, rather than absolute $O_3$ concentrations, enabling us to elucidate how meteorology governs the classification and evolution of sensitivity regimes. It should be noted that the attribution analysis presented here is based on statistical associations, and the results only reflect the relative contributions of each variable to OFS states, without implying strict causal relationships.

And Lines 651–655:

Overall, the satellite–indicator (FNR)–interpretable machine learning (RF + SHAP) framework developed here provides a scalable and decomposable framework for interpreting the long-term evolution and associated factors influencing FNR-derived OFS states from national to urban-cluster scales. This framework offers a useful scientific basis for dynamic sensitivity reclassification and the design of region-specific $NO_x$–VOC control strategies, with broader relevance for $O_3$ mitigation in other megacity regions.

Finally, we have added a clearer discussion of the limitations of this framework, including that this is an exploratory application of a classification model to OFS states, and that the attribution analysis is intended to complement, rather than replace, process-based chemical diagnosis. We have revised the expression in Lines 637–645 of the revised manuscript:

This study is subject to several uncertainties and limitations. First, this is an exploratory application of a classification model to OFS states, and the performance of the RF depends on class balance; however, the threshold-based classification leads to class imbalance, and some influencing factors are not included. Second, the thresholds themselves are uncertain, and the coarse spatial resolution of OMI data, along with mismatches with ground observations, may introduce classification uncertainty and affect model performance. The use of higher-resolution data such as TROPOMI, combined with $O_3$ observations, can improve diagnostic accuracy (Ren et al., 2022). Future work integrating ML with observations and chemical transport models may help

bridge this gap (Xiong et al., 2024). Therefore, the present framework should be viewed as a complementary statistical attribution tool, rather than a substitute for process-based OFS diagnosis using chemical transport models or observation-based box models.

4. Use of a fixed national FNR threshold and resulting uncertainty: The authors use a single fixed FNR threshold (1 and 2) to classify ozone formation regimes across China. However, previous studies suggest that the appropriate FNR thresholds are higher in several major Chinese megacities than in other regions. Applying a non-local (uniform) threshold can bias regime classification and may artificially reduce inferred $O_3$ sensitivity to VOCs, especially at the national scale (Ren et al., ACP, 2022, https://doi.org/10.5194/acp-22-15035-2022.). This threshold-related uncertainty is non-negligible. I recommend estimating region-specific thresholds by connecting satellite $HCHO/NO_2$ to ground-based $O_3$ responses (e.g., using observed $O_3$-precursor relationships) and then reassessing the regime classification and subsequent analyses.

Response: We sincerely thank the reviewer for this important and constructive comment. We fully agree that the use of a uniform national FNR threshold may introduce substantial uncertainty into OFS classification, especially in China where precursor emissions, atmospheric environments, and meteorological conditions vary markedly across regions. In response to this comment, we have made a major revision to the manuscript by removing the original fixed threshold approach based on Duncan et al. (2010) and instead deriving region-specific FNR threshold ranges for different regions of China by linking satellite-derived $HCHO/NO_2$ ratios with ground-based

ozone ($O_3$) observations.

Specifically, following Jin et al. (2020) and related studies, we established the relationship between FNR and the probability of high $O_3$ occurrence and used this relationship to derive threshold ranges separately for the five major urban clusters (BTH, FWP, YRD, SCB, and PRD) and for other regions of China (ORC). The revised results show clear regional heterogeneity in the thresholds derived. We have revised the relevant methodology in Lines 160–184:

OFS regimes were then identified according to the corresponding FNR threshold ranges. The threshold ranges were determined mainly following the method of Jin et al. (2020). In this study, region-specific FNR thresholds across China were derived by combining OMI satellite observations with surface $O_3$ measurements. This approach is also consistent with previous studies that derived threshold ranges from fitted $O_3$–FNR relationships (Wang et al., 2021; Ren et al., 2022; Chen et al., 2024). In practice, the transition of OFS from VOC-limited to $NO_x$-limited regimes is gradual rather than abrupt. This transition can be influenced by multiple factors, including topography, meteorological conditions, and the spatial resolution of satellite observations (Jin et al., 2020). Because these influencing factors vary across China, the corresponding FNR threshold ranges are also expected to differ among regions. For this reason, OFS regimes were identified using region-specific FNR threshold ranges. China was divided into six regions, including five major urban clusters with severe $O_3$ pollution, namely BTH, FWP, YRD, SCB, and PRD, and one additional category comprising all remaining areas, referred to as other regions of China (ORC) (Lu et al., 2018). These

six regions were used as the basic units for deriving region-specific FNR threshold ranges, which were then applied to identify OFS regimes in each region.

Because $O_3$ pollution in China occurs mainly during the warm season (April–September) (Lu et al., 2020b), the analysis was restricted to April–September in 2015–2023. To match the OMI overpass time, surface $O_3$ concentrations at 13:00 and 14:00 local time were averaged for each site and day. The resulting daily $O_3$ concentrations were then collocated with OMI FNR values at the corresponding grid cells. For each site, the high-$O_3$ probability was then defined as the fraction of days during the study period on which the 13:00–14:00 mean $O_3$ concentration exceeded 160 μg $m^{-3}$. The resulting paired samples were subsequently pooled within each region for threshold derivation.

To focus on polluted conditions, samples with $NO_2$ vertical column densities below $1.5 \times 10^{15}$ molec $cm^{-2}$ were excluded, following previous studies (Jin et al., 2020; Wang et al., 2021). In addition, high-end outliers (5% or 10%, depending on the region) were removed prior to fitting to reduce the influence of extreme values (Chen et al., 2024). After quality control, paired high-$O_3$ probability and FNR samples in each region were grouped into 100 bins according to FNR and a cubic polynomial was fitted to the binned relationship between high-$O_3$ probability and FNR. The FNR interval corresponding to fitted high-$O_3$ probability values above the $90^{th}$ percentile was then defined as the threshold range for that region (Jin et al., 2020).

And the revised results are presented in Lines 333–351:

**3.2 Determination of region-specific FNR threshold ranges**

To derive region-specific FNR threshold ranges, we first examined the nonlinear relationship between $O_3$ formation and its precursors using satellite-derived HCHO and $NO_2$. Figure S3 shows the distribution of high-$O_3$ probability in the two-dimensional HCHO–$NO_2$ space for the six regions. The overall pattern is similar to the $O_3$ contour structure reported in previous studies (Sillman et al., 1999), indicating that HCHO and NO2 can reasonably serve as proxies for VOCs and $NO_x$, respectively, in characterizing the nonlinear response of $O_3$ formation. Based on this relationship, each region can be broadly divided into three OFS regimes: a $NO_x$-limited regime under relatively high HCHO and low $NO_2$ conditions, a transitional regime under relatively high levels of both HCHO and $NO_2$, and a VOC-limited regime under relatively low HCHO and high $NO_2$ conditions.

Based on the fitted high-$O_3$ probability-FNR relationships, region-specific threshold ranges were derived for the six regions, reflecting the gradual transition among OFS regimes. Figure 3 shows the fitted relationships for the six regions during the warm season (April–September) of 2015–2023. Clear regional differences are found in the derived FNR threshold ranges. Among the six regions, FWP shows the highest threshold range, with a central threshold of 3.42 and a range of [2.26, 4.77], whereas YRD shows the lowest, with a central threshold of 1.59 and a range of [0.90, 2.37]. The corresponding threshold ranges for BTH, SCB, PRD, and ORC are [1.51, 3.30], [1.93, 4.73], [1.66, 3.70], and [1.17, 3.08], respectively. In terms of fitting performance, the correlation coefficient is 0.58 in FWP and exceeds 0.60 in the other regions, reaching 0.96 in ORC, with all p-values below 0.05. These results indicated that the fitted curves

generally capture the variation in high-$O_3$ probability well. Comparison with previous studies (Table. S1) further shows that FNR thresholds vary across regions, reflecting differences in atmospheric environment, precursor emissions, and local conditions (Liu and Shi, 2021).

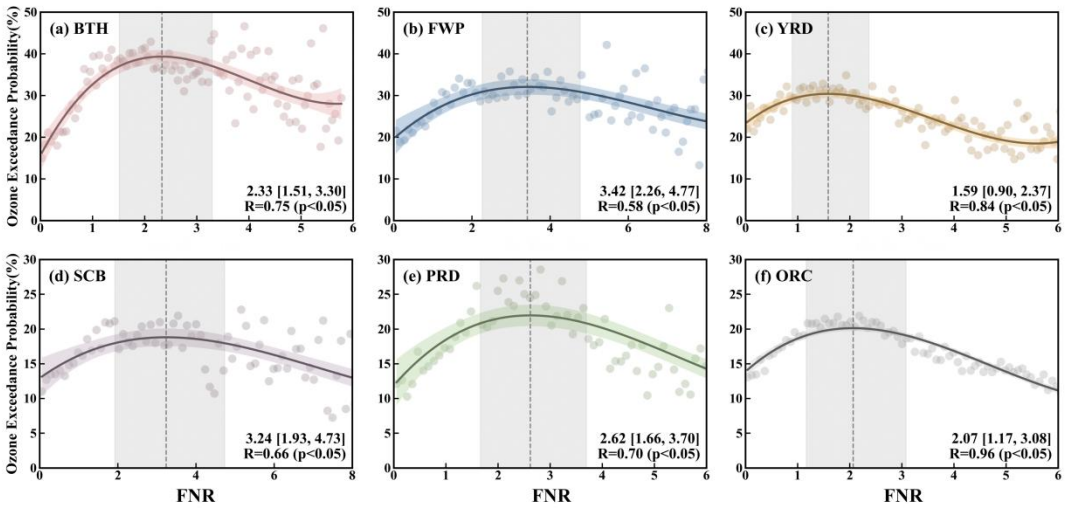



**Figure 3. Relationship between daily high-$O_3$ probability and FNR in the six regions during the warm season (April–September) of 2015–2023. Solid lines show the fitted cubic polynomial curves, and the shaded areas indicate the 95% confidence intervals. Vertical dashed lines mark the peak values of the fitted curves, while the gray shaded bands denote the FNR ranges corresponding to the upper 10% of the fitted curves, which were used to define the threshold ranges.**

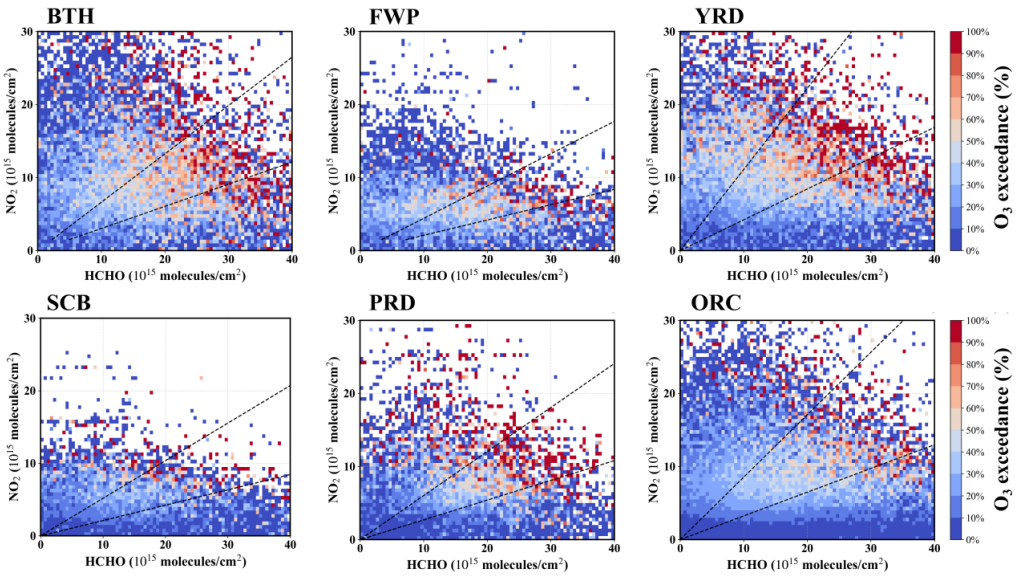



**Figure S3. $O_3$ exceedance probability as a function of OMI HCHO and $NO_2$ for the six regions during the warm season (April–September) from 2015 to 2023. The black dashed lines indicate the OMI HCHO/$NO_2$ values corresponding to the derived FNR threshold ranges used to classify $NO_x$-limited, transitional, and**

**VOC-limited regimes.**

In addition, to directly address the reviewer's concern regarding threshold-related uncertainty, we further added a threshold-sensitivity analysis in the revised manuscript. Three threshold scenarios were designed: the region-specific thresholds derived in this study, literature-based minimum–maximum ranges, and literature-based average ranges. The results show that OFS classification is indeed sensitive to threshold choice, and the magnitude of this sensitivity varies substantially across regions. This confirms that threshold uncertainty is non-negligible and that applying literature-based or spatially uniform thresholds without regional constraints may introduce considerable bias into OFS classification. We have added more information in Lines 352–375:

Given these pronounced regional differences, we further assessed how threshold uncertainty affects OFS classification. Three threshold-range scenarios were designed. Specifically, Scenario 1 (S1) used the threshold ranges derived in this study for the warm season of 2015–2023, Scenario 2 (S2) adopted the minimum–maximum threshold ranges reported in previous studies, and Scenario 3 (S3) used threshold ranges calculated from the average of the corresponding maximum and minimum values. The threshold ranges for the six regions under S1-S3 scenarios are listed in Table S1. These three scenarios were then used to evaluate the sensitivity of OFS classification to different FNR threshold selections in the five major city clusters during the warm season over 2005–2023 (Fig. S4).

The results show that OFS classification is strongly affected by threshold choice, although the magnitude of this effect varies markedly across regions. Among the five

regions, FWP shows the largest sensitivity to threshold variation, with changes ranging from −53.8 % to 51.6 % and an average value of about 22.0 %. SCB also exhibits substantial sensitivity, with corresponding changes of −48.8 % to 52.0 %. In contrast, PRD is relatively less sensitive, with changes ranging from −36.3 % to 21.1 % and an average value of about 17.0 %. BTH and YRD show intermediate responses, with ranges of −31.0 % to 37.2 % and −31.5 % to 31.5 %, respectively. These results indicate that the uncertainty introduced by threshold selection is not spatially uniform but depends on regional photochemical characteristics and the width of the transitional range.

Figure S4 further shows that the influence of threshold choice is often asymmetric, with some scenarios producing much larger positive or negative deviations than others. This indicates that OFS classification does not respond linearly to threshold perturbations. In particular, threshold ranges compiled from previous studies can produce substantial departures from the classification results obtained using the region-specific thresholds derived here. Even the average-threshold scenario still leads to noticeable deviations in several regions. These findings suggest that threshold transferability across regions is limited and that applying literature-based thresholds without regional constraints may introduce considerable bias into OFS diagnosis.

Overall, these results demonstrate that uncertainty in FNR thresholds can substantially alter the estimated proportions of different OFS regimes. Region-specific thresholds derived from regional observational relationships therefore provide a more robust basis for OFS classification than fixed or literature-averaged thresholds.

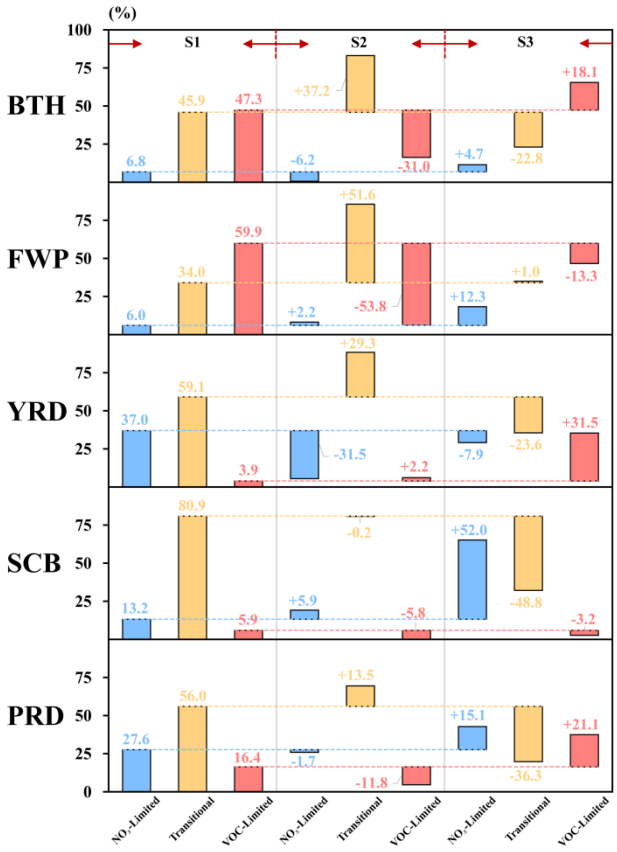



**Figure S4. Sensitivity of OFS regime area fractions to different FNR threshold scenarios in the five major city clusters during the warm season (April–September) over 2005–2023. S1 denotes the region-specific thresholds derived in this study, S2 denotes the minimum–maximum threshold ranges compiled from previous studies, and S3 denotes the averaged threshold ranges based on the corresponding minimum and maximum values. Bars represent deviations in regime area fractions under S2 and S3 relative to S1.**

5. In L21, "non-methane volatile organic compounds" appears only once in the abstract; the abbreviation is unnecessary and can be removed.

Response: We thank the reviewer for this careful comment. We agree that the original expression was unnecessary in this context. Since the abstract refers to VOCs in a general sense rather than specifically to non-methane VOCs throughout the manuscript, we have revised "non-methane volatile organic compounds" to "volatile organic compounds (VOCs)" in the Abstract (Line 16).

6. In L79, please specify the criteria used to define the “five ozone-prone urban clusters” and provide justification for this particular regional delineation (why these clusters).

Response: We thank the reviewer for this helpful comment. The five urban clusters were selected because they are the most consistently $O_3$-polluted and frequently reported $O_3$ hotspot regions in China. This regional delineation was based on two main considerations. First, according to recent national assessments, these clusters have persistently elevated surface $O_3$ concentrations and are among the most severely affected regions in China in terms of $O_3$ pollution (Ozone Pollution Control Committee of Chinese Society of Environmental Sciences, 2024). Second, they have also been widely identified in previous studies as key regions for $O_3$ formation and control, owing to their high anthropogenic emissions, dense population, and strong contrasts in meteorological and topographic conditions (Wang et al., 2017; Zhang et al., 2025). Focusing on these five clusters therefore allows us to capture the main regional heterogeneity of OFS in China while maintaining consistency with previous $O_3$ studies and national air-pollution assessments.

To clarify this point, we have added the relevant explanation in the revised manuscript (Lines 103–107) as follows:

These regions were selected because they have persistently high $O_3$ levels and are widely recognized as major $O_3$ pollution hotspots in China (Ozone Pollution Control Committee of Chinese Society of Environmental Sciences, 2024). They also represent distinct combinations of emission characteristics, meteorological conditions, and topographic settings, making them suitable for examining the regional heterogeneity of

OFS (Wang et al., 2017; Zhang et al., 2025b).

7. In L97, the threshold defining "anomalously large HCHO columns" (> $1.0 \times 10^{17}$ molec $cm^{-2}$) should be justified. Even if this follows prior work, the manuscript should briefly explain the reasoning/physical basis for this cutoff.

Response: We sincerely thank the reviewer for this helpful comment. We agree that the upper threshold used to identify anomalously large HCHO columns should be more clearly justified in the manuscript.

In the revised manuscript, we have clarified that the cutoff of $1.0 \times 10^{17}$ molec $cm^{-2}$ was not arbitrarily chosen but follows the quality-control approach adopted in previous studies using the OMI HCHO product, particularly Zhu et al. (2017). This threshold is intended to exclude unrealistic extreme values and retrieval outliers that are far above the typical range of OMI-retrieved tropospheric HCHO columns. As now stated in the manuscript, the OMI HCHO columns retrieved with the Smithsonian Astrophysical Observatory (SAO) algorithm generally fall within the range of $4 \times 10^{15}$ to $4 \times 10^{16}$ molec $cm^{-2}$, with a detection limit of approximately $1.0 \times 10^{16}$ molec $cm^{-2}$ (González Abad et al., 2015). Therefore, values exceeding $1.0 \times 10^{17}$ molec $cm^{-2}$ are substantially above the normal observational range and are treated as anomalous retrievals rather than physically representative HCHO columns.

We have accordingly revised the relevant text in the manuscript to make this rationale explicit and to show that this screening step is a data-quality control measure based on both the retrieval characteristics of the OMI product and prior literature, rather

than an arbitrary threshold choice (Lines 123–129):

HCHO data were obtained from the Level-3 OMHCHOd product with a spatial resolution of 0.1° × 0.1° (Chance, 2019). This product is generated based on the Level-2 OMHCHO retrievals after applying quality control filters to exclude pixels with cloud fractions > 0.3, solar zenith angles > 70°, or those affected by the OMI row anomaly. The HCHO columns were retrieved using the Smithsonian Astrophysical Observatory (SAO) algorithm, with typical values ranging from $4 \times 10^{15}$ to $4 \times 10^{16}$ molec $cm^{-2}$ and with a detection limit of approximately $1.0 \times 10^{16}$ molec $cm^{-2}$ (González Abad et al., 2015). Following Zhu et al. (2017), only values between $-0.5 \times 10^{16}$ and $1.0 \times 10^{17}$ molec $cm^{-2}$ were retained, while values outside this range were treated as outliers and removed.

8. L154: For model evaluation, I suggest training the model using 2005-2020 data and performing an independent validation using 2021-2023. This split would more accurately demonstrate model generalization.

Response: We sincerely thank the reviewer for this valuable suggestion. We agree that a training-validation split of 2005–2020 for training and 2021–2023 for independent validation is a rigorous strategy for evaluating temporal generalization, especially in prediction-oriented studies.

However, the primary objective of this study is not to predict OFS in future years, but to systematically investigate the statistical relationships between OFS regimes and their associated environmental drivers over the full observational period (2005–2023),

and to further conduct an explainable machine learning analysis.

Within this framework, the SHAP method is used to decompose model predictions into interpretable feature attributions, which can be consistently compared across different samples and temporal subsets (Lundberg and Lee, 2017). Therefore, adopting a strict temporal hold-out strategy (e.g., training the model on 2005–2020 and independently validating it on 2021–2023) would result in the model being interpreted only under a late-period distribution, thereby limiting SHAP-based explanations from fully covering key stages such as policy implementation periods. This would consequently restrict a comprehensive characterization of regime changes across the entire study period.

To address this, a stratified random sampling strategy was adopted, in which samples were split into 80% for training and 20% for testing (Özüpak et al., 2025). To ensure that both the training and testing sets include samples from all years, a year-wise stratification procedure was further implemented. This design maintains overall distributional consistency while preserving interannual variability associated with key periods such as policy interventions and the COVID-19 pandemic, enabling the model to learn more representative overall distribution patterns. In addition, SHAP, when applied to a fixed trained model, can be used to compare feature attributions across different temporal subsets (e.g., Zhang et al., 2024). Therefore, this sampling strategy is also designed to support subsequent SHAP-based cross-temporal interpretation and comparisons between different stages, including pre- and post-policy periods.

Furthermore, given the presence of class imbalance in the dataset, a strict temporal

hold-out split may lead to an unbalanced class distribution in the validation set, thereby introducing potential sampling bias and compromising the reliability of model evaluation. Therefore, stratified sampling helps preserve the intrinsic structural characteristics of the original dataset while ensuring a more balanced representation of majority and minority classes (Sadaiyandi et al., 2023).

To address the reviewer's concern, we have clarified in the revised manuscript the present study focuses on cross-temporal mechanistic interpretation and feature attribution analysis.

We have provided a more detailed explanation of this in Lines 235–248 of the revised manuscript:

Model training and evaluation were based on annual mean gridded data for the warm seasons (April–September) from 2005 to 2023. Because the primary aim of this study was not to predict OFS in future years, but to characterize the relationships between OFS regimes and their environmental drivers over the full study period, a year-wise stratified random sampling strategy was adopted. In studies aimed at predicting future years, a common practice is to train the model on earlier-period data and validate it on later-period data in order to assess temporal extrapolation skill (Zhu et al., 2019; Wang et al., 2023). Such a strategy was not adopted here because it would bias model evaluation and subsequent SHAP interpretation toward the late-period data distribution. Instead, samples from each year were randomly divided into 80% for training and 20% for validation (Özüpak et al., 2025), so that both subsets contained data from all years. This design preserves interannual variability associated with key periods, including

major policy interventions and the COVID-19 pandemic, while maintaining comparable sample distributions across the training and validation sets. In addition, given the class imbalance in the dataset, this stratified sampling strategy helps maintain a more balanced class distribution in both subsets, thereby improving the reliability of model evaluation (Sadaiyandi et al., 2023). It also provides a more suitable basis for subsequent SHAP-based interpretation across different temporal subsets, thereby facilitating comparisons of OFS regimes before and after policy implementation.

9. In the Results, the unit formatting "molec.cm$^{-2}$" is unusual. Please revise to a standard format with a space, e.g., "molec cm$^{-2}$" consistently throughout.

Response: Following the reviewer's suggestion, we have revised the unit formatting throughout the manuscript and standardized all $NO_2$ and HCHO column densities to "molec cm$^{-2}$".

10. In L212, the FNR trend over the Sichuan Basin (SCB) does not appear statistically significant; describing it as "a modest upward trend" seems inaccurate. It would be more appropriate to state that no significant trend is observed. Additionally, SCB does not seem to be analyzed with the same two-stage trend framework used elsewhere. Visual inspection suggests a decrease during ~2005-2011 and an increase during ~2013-2023, with 2012 resembling an outlier/break point. The authors should justify the staging choice for SCB and apply a consistent approach. Finally, since SCB is discussed first in the text, it would be clearer to label it as panel (a) rather than panel (d).

Response: We sincerely thank the reviewer for this careful and constructive comment. We agree that our original wording was not sufficiently precise for SCB. In the revised manuscript, we have replaced “a modest upward trend” with “no significant overall trend”, which is more consistent with the statistical result for SCB.

Regarding the staging issue, we appreciate the reviewer’s suggestion and have now clarified the analytical basis more explicitly. In the revised manuscript, we first applied the Pettitt test to all five regions using a consistent procedure to identify possible change points in the FNR time series. The results show that the regions do not follow a fully uniform pattern. In particular, SCB and PRD differ from BTH, FWP, and YRD, which show clearer phase-like transitions around the policy period. For SCB, the Pettitt test does not indicate a statistically significant change point, and its overall FNR trend is also not significant. Therefore, although visual inspection may suggest some fluctuation around 2012, we do not consider it appropriate to interpret SCB using the same two-stage framework applied to regions with clearer and statistically supported phase shifts. Instead, we now describe SCB as showing no significant long-term trend, with OFS remaining largely within the transitional regime. We have clarified it in Lines 396–400:

In contrast, SCB shows a relatively high annual mean FNR of 3.34 but no significant change point or clear long-term trend. In addition, SCB did not show a statistically significant change point in the Pettitt test, which further supports the decision not to interpret its FNR evolution using the same two-stage framework. Together with the relatively small changes in both $NO_2$ and HCHO, this suggests that the precursor structure in SCB remained comparatively stable over the study period,

and OFS stayed largely within the transitional regime.

For PRD, the Pettitt test and long-term trend analysis indicate yet another pattern. Unlike SCB, PRD exhibits a significant long-term increase in FNR, but its evolution is not best described as the same two-stage behavior seen in BTH, FWP, and YRD. We therefore treat PRD separately in the revised text and clarify that it represents a distinct regional evolution pattern rather than being grouped into the common two-stage framework. We have clarified it in Lines 401–404:

PRD differs from the other two patterns by exhibiting a significant long-term increase in FNR, at a rate of 0.06 per year ($p < 0.05$). This trend indicates a shift in OFS from VOC-limited or transitional regimes toward more $NO_x$-limited regimes, likely associated with reductions in $NO_x$-rich sources in the region, particularly from road traffic and industrial emissions, as suggested by previous studies (Bian et al., 2019; Li et al., 2025).

Accordingly, the revised manuscript now applies to a consistent initial change-point detection method across regions, while allowing for region-specific interpretation when the statistical characteristics differ. This revision avoids implying that all five regions follow the same two-stage evolution and provides a more accurate interpretation of SCB and PRD.

With respect to the figure presentation, we have adjusted the order of discussion in the text to better align with the order shown in the figure, which we believe improves readability without requiring changes to the panel labels in the formatted multi-panel figure.

552

553 11. In L216, the statement “driven primarily by substantial $NO_x$ reductions from road
554 traffic and industrial sources” appears to lack direct supporting evidence in the
555 manuscript (data or references).

556 Response: We agree that the original wording was too strong, as the manuscript
557 did not provide direct evidence within the main text to support such a specific source
558 attribution. In the revised manuscript, we have therefore moderated the expression and
559 added supporting references to make the interpretation better grounded.

560 Specifically, instead of stating that the change was “driven primarily by substantial
561 $NO_x$ reductions from road traffic and industrial sources,” we now describe it more
562 cautiously as likely associated with reductions in NOx-rich sources in the region,
563 particularly from road traffic and industrial emissions, as suggested by previous studies.
564 We have revised it in Lines 401–404:

565 PRD differs from the other two patterns by exhibiting a significant long-term
566 increase in FNR, at a rate of 0.06 per year ($p < 0.05$). This trend indicates a shift in OFS
567 from VOC-limited or transitional regimes toward more $NO_x$-limited regimes, likely
568 associated with reductions in $NO_x$-rich sources in the region, particularly from road
569 traffic and industrial emissions, as suggested by previous studies (Bian et al., 2019; Li
570 et al., 2025).

571

572 12. Given the threshold uncertainty raised in Comment #3, the conclusions in Section
573 3.2 should be treated as provisional. A reanalysis after adopting region-specific (or
574 otherwise justified) thresholds is necessary to confirm the robustness of the reported

regime shifts.

Response: We sincerely thank the reviewer for this important comment. We fully agree that the conclusions regarding OFS regime shifts should not rely on a fixed and potentially non-representative national threshold, and that their robustness needs to be reassessed using more regionally appropriate threshold settings.

In response, we have substantially revised the analysis by replacing the original fixed threshold approach with region-specific FNR threshold ranges derived from the relationships between satellite-based $HCHO/NO_2$ and ground-based $O_3$ observations for different regions of China. Based on these updated thresholds, we reanalyzed OFS classifications and their long-term evolution for each major urban cluster.

The revised results confirm that the proportions of VOC-limited, transitional, and $NO_x$-limited regimes do change to some extent after adopting region-specific thresholds, indicating that OFS classification is indeed sensitive to threshold choice. However, the overall regional contrasts and long-term regime evolution remain interpretable under the revised framework. To further assess robustness, we also added a threshold-sensitivity analysis under multiple threshold scenarios, which explicitly quantifies the uncertainty associated with threshold selection.

Accordingly, the conclusions in the revised Section 3.2 and the subsequent OFS analysis are now based on the region-specific threshold framework, rather than the original uniform threshold assumption. The detailed derivation and justification of these region-specific thresholds are provided in our response to Comment #4.

13. In Section 3.3, the ML validation results in Table S2 should be explicitly mentioned and cited in the main text; otherwise the reader cannot easily assess model performance and credibility.

Response: We have revised the manuscript by explicitly citing the model evaluation table in the main text and adding a concise summary of the validation results in Section 3.4.1 (Lines 477–490):

**3.4.1 Performance of RF models for OFS classification**

Table 1 presents a comparison of the five model configurations. Overall, all models demonstrate relatively high classification performance, with AS, GMS, and ROC-AUC all exceeding 0.70. Among the regions, the models for the YRD and PRD perform the best, with all category-specific metrics generally above 0.70.

In contrast, the BTH and FWP models show relatively low performance for minority classes, particularly the $NO_x$-limited category, with F1 scores of 0.60 and 0.57, respectively, while other categories generally achieve metrics above 0.60, indicating moderate performance in identifying this category. In the SCB model, the NOx-limited category performs the worst, with PS, RS, and F1 all below 0.50, reflecting the difficulty for the model in effectively capturing its characteristics.

Overall, after sampling and hyperparameter optimization, the models achieve robust performance on the majority classes, but classification performance for minority classes remains limited. Ahmed et al. (2025) noted that SMOTE can effectively improve classification performance by addressing data imbalance, but it may also lead to decreased accuracy for majority classes, highlighting the trade-offs involved when

handling imbalanced datasets.

**Table 1. Performance metrics of the optimized model on the validation dataset.**

| Models | Class | PS | RS | F1 | AS | GMS | ROC-AUC |
|---|---|---|---|---|---|---|---|
| BTH | $NO_x$-limited[1] | 0.51 | 0.73 | 0.60 | 0.94 | 0.83 | 0.96 |
| | VOC-limited | 0.88 | 0.88 | 0.90 | 0.88 | 0.89 | 0.95 |
| | Transitional | 0.83 | 0.78 | 0.81 | 0.83 | 0.82 | 0.90 |
| FWP | $NO_x$-limited[1] | 0.51 | 0.64 | 0.57 | 0.96 | 0.79 | 0.97 |
| | VOC-limited | 0.88 | 0.88 | 0.88 | 0.86 | 0.85 | 0.93 |
| | Transitional | 0.75 | 0.72 | 0.74 | 0.82 | 0.79 | 0.89 |
| YRD | $NO_x$-limited | 0.84 | 0.87 | 0.85 | 0.89 | 0.88 | 0.95 |
| | VOC-limited[1] | 0.69 | 0.75 | 0.72 | 0.98 | 0.86 | 0.99 |
| | Transitional | 0.90 | 0.87 | 0.88 | 0.87 | 0.86 | 0.93 |
| SCB | $NO_x$-limited[1] | 0.32 | 0.48 | 0.39 | 0.90 | 0.67 | 0.87 |
| | VOC-limited[1] | 0.51 | 0.68 | 0.58 | 0.94 | 0.81 | 0.92 |
| | Transitional | 0.94 | 0.89 | 0.91 | 0.85 | 0.72 | 0.81 |
| PRD | $NO_x$-limited | 0.85 | 0.82 | 0.84 | 0.92 | 0.88 | 0.96 |
| | VOC-limited[1] | 0.66 | 0.80 | 0.72 | 0.90 | 0.86 | 0.94 |
| | Transitional | 0.86 | 0.82 | 0.84 | 0.81 | 0.81 | 0.88 |

[1]**Minorities classes (categories with fewer samples).**

14. For Fig. 7, please provide a more detailed explanation of what is shown: which class/output the SHAP values correspond to (especially when it comes to classification); whether the plotted values are the mean of SHAP values or mean(|SHAP|); and how the aggregation is performed across samples.

Response: We thank the reviewer for the valuable comments. We have clarified that Fig. 7 does not present raw SHAP values for a single class. Instead, it shows the relative contribution of each feature to the overall OFS classification in each regional model. Specifically, SHAP values were first calculated for each sample and for each of the three OFS classes. Then, for each class, the mean absolute SHAP value, mean(|SHAP|), was computed for each feature across all samples. These class-specific

mean(|SHAP|) values were subsequently averaged across the three OFS classes to obtain the overall mean(|SHAP|) for each feature in a given regional model. Finally, these overall mean(|SHAP|) values were normalized to express the relative contribution percentage of each feature, which facilitates comparison among regions.

We have added a more explicit explanation in the revised manuscript to clarify (i) which SHAP output is used, (ii) that the plotted values are based on mean(|SHAP|) rather than signed SHAP values, and (iii) how the aggregation is performed across classes and samples (Lines 492–496):

For each regional model, SHAP values were first calculated for each sample and for each OFS class. For each class, mean(|SHAP|) was then computed for each feature across all samples, and these class-specific mean(|SHAP|) values were subsequently averaged across the three OFS classes to represent the overall importance of each feature in that regional model. To facilitate comparison among regions, the resulting overall mean(|SHAP|) values were further normalized and expressed as relative contribution percentages.

15. In L323 (“gradual shifts in climate, e.g., warming, increased solar radiation, and changes in stability”) and L328 (“the complex composition of residential heating, solvent use, and industrial emissions”), I could not find direct evidence or datasets in the manuscript supporting these claims. Please provide supporting analyses/references, or revise these statements to align with the evidence presented.

Response: We agree that the original statements were not sufficiently supported by

direct evidence or datasets presented in the manuscript. In the revised version, we have therefore removed these unsupported statements to ensure that all interpretations remain closely aligned with the analyses and evidence provided in the study.

16. In L358, SHAP values decompose contributions relative to an expected output (base value). Because the manuscript fits separate models for different regions (each with its own baseline and data distribution), the absolute magnitudes of SHAP values are not directly comparable across regions. Please revise this part of the discussion and, if cross-region comparisons are intended, adopt a method that ensures comparability (e.g., reporting relative importance within each region).

Response: We fully agree that, because SHAP values are defined relative to the model-specific expected output and are affected by the underlying data distribution, the absolute magnitudes of SHAP values are not directly comparable across separately trained regional models. In response, we have revised both the analysis and the discussion to avoid overstating such comparisons.

(1) To reduce structural differences among regional models, we retrained all regional RF models using a unified set of hyperparameter settings. While this does not eliminate differences in baseline prediction and data distribution, it helps ensure that the regional models are constructed under a consistent modeling framework. We have clarified it in Lines 249–257:

To account for class imbalance and optimize model performance, the Synthetic Minority Over-sampling Technique (SMOTE; Fernandez et al., 2018) was applied to

the training data. To reduce the structural differences of the model and facilitate consistent comparison across the five regions, a unified set of model parameters was adopted for all regional models. Although this does not eliminate differences in model baseline and data distribution across regions, it helps ensure that the regional models are constructed under a consistent modeling framework. Hyperparameter optimization was performed using GridSearchCV (Hastie et al., 2009; Ahmed et al., 2025) over combinations of candidate parameter values. Based on validation performance across all regions, a single parameter set was selected as the optimal configuration. The final model configuration consisted of 200 trees (n_estimators = 200), a maximum tree depth of 25 (max_depth = 25), and a minimum of 1 sample per leaf (min_samples_leaf = 1).

(2) For cross-region comparison, we no longer interpret absolute SHAP magnitudes directly. Instead, within each regional model, we calculated the overall mean absolute SHAP value [mean(|SHAP|)] for each feature and then normalized these values to obtain the relative contribution percentage of each variable within that region. The regional comparison in the revised manuscript is therefore based on relative importance within each model, rather than raw SHAP magnitudes across models. We have clarified this explicitly in the Methods (Lines 274–279) sections:

For each regional model, the mean absolute SHAP value [mean(|SHAP|)] was calculated for each feature. Because SHAP values are defined relative to the model-specific expected output and are influenced by the underlying data distribution, the absolute magnitudes of SHAP values are not directly comparable across different regional models. Therefore, to facilitate cross-region comparison, these overall

mean(|SHAP|) values were further normalized to obtain the relative contribution of each variable within each regional model. It should be noted that this metric reflects the statistical contribution of individual features to model predictions rather than their direct causal effects in atmospheric physical or chemical processes.

(3) In the revised discussion, we have avoided direct statements comparing the absolute sizes of SHAP values across regions. Instead, we focus on the relative contribution structure within each region and the directional interpretation of individual predictors in the SHAP summary plots (Lines 491–523):

**3.4.2 Relative contributions and regional heterogeneity of OFS influencing factors**

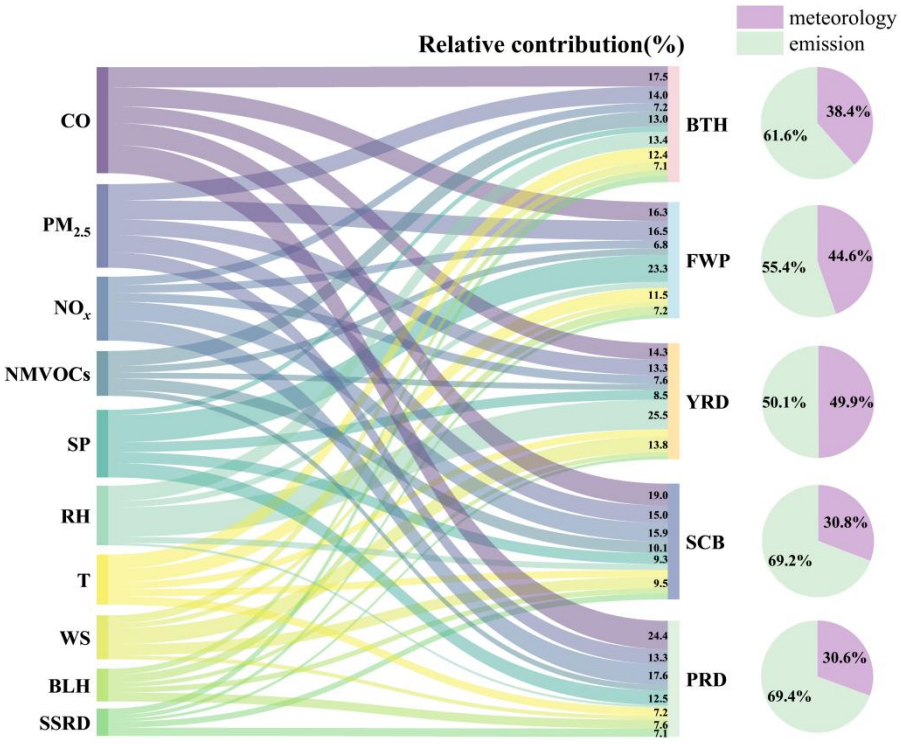
For each regional model, SHAP values were first calculated for each sample and for each OFS class. For each class, mean(|SHAP|) was then computed for each feature across all samples, and these class-specific mean(|SHAP|) values were subsequently averaged across the three OFS classes to represent the overall importance of each feature in that regional model. To facilitate comparison among regions, the resulting overall mean(|SHAP|) values were further normalized and expressed as relative contribution percentages. Figure 7 shows the relative contribution of each feature to OFS classification in the five city clusters. Overall, emission-related variables contribute more than meteorological variables, with their total contribution ranging from 50.1% in YRD to 72.2% in PRD. This result indicates that OFS classification is primarily constrained by emission-related factors in most city clusters. In YRD, meteorological and emission-related variables contribute almost equally to OFS

721 classification, accounting for 49.9% and 50.1%, respectively.

722 A closer examination of individual emission predictors reveals substantial regional
723 heterogeneity. In BTH, FWP, and YRD, CO and $PM_{2.5}$ make the largest contributions
724 (17.5% and 14.0% in BTH; 16.3% and 16.5% in FWP; 14.3% and 13.3% in YRD).
725 NMVOCs also show a relatively large contribution in BTH, reaching 13.0%. By
726 contrast, SCB and PRD are characterized by high contributions from CO and $NO_x$ (19.0%
727 and 15.9% in SCB; 24.4% and 17.6% in PRD). The strong model importance of CO
728 and $PM_{2.5}$ likely reflects both the influence of their significant long-term trends in the
729 model and their broader association with regional photochemical environments. First,
730 both variables show a two-stage trend in most urban clusters from 2005 to 2023, with
731 an initial period of non-significant change followed by a significant decline (Fig. S5),
732 a pattern that is similar to the phase-like changes in OFS identified in this study (Fig.
733 S6). Second, CO can enhance hydroperoxy radical ($HO_2$) production through the
734 reaction CO + OH → $HO_2$, thereby increasing $O_3$ production efficiency under
735 conditions of declining NOx (Ren et al., 2013; Seinfeld and Pandis, 2016). Third, a
736 reduction in $PM_{2.5}$ may weaken its heterogeneous uptake of hydroxyl radicals ($HO_x$)
737 and $NO_x$ and thus allow more radicals and reactive nitrogen to remain available for
738 photochemical $O_3$ reactions (Li et al., 2019). Therefore, these results suggest that CO
739 and $PM_{2.5}$ function not simply as co-emitted pollutants, but also as informative
740 predictors of broader changes in oxidation capacity and regional pollution conditions.

741 Meteorological predictors also show robust and regionally distinct contributions.
742 Surface pressure (SP), relative humidity (RH), and temperature (T) are consistently

743 important across regions, highlighting the roles of weather system stability, moisture
744 conditions and thermal conditions in modulating photochemistry and pollutant
745 accumulation. In the YRD, the contribution of RH is particularly notable (25.5%), while
746 in the FWP, SP contributes more substantially (23.3%), indicating that OFS in these
747 regions is also strongly influenced by meteorological factors. More broadly,
748 meteorological factors can influence atmospheric oxidation capacity, boundary-layer
749 structure, and pollutant dispersion, thereby shaping both the spatial distribution and
750 temporal evolution of OFS. Therefore, the relatively high contributions of certain
751 meteorological factors in the YRD and FWP may reflect a greater sensitivity of OFS
752 classification to regional humidity and atmospheric stability conditions.




753
754 **Figure 7. Relative contribution of emission-related and meteorological factors to overall OFS**
755 **classification.**

756 And Lines 567–599:

### 3.4.4 SHAP responses of individual predictors across OFS regimes

To further interpret how individual predictors influence OFS classification, we examined the SHAP value distributions of different variables for each urban cluster (Fig. 9). The results reveal both common features and regional differences in the directional and nonlinear effects of the predictors. In general, many variables show broadly opposite SHAP responses between VOC-limited and $NO_x$-limited regimes, indicating that the same predictor can exert contrasting effects under different chemical sensitivity conditions.

For emission-related variables, samples classified as $NO_x$-limited are generally associated with relatively low emission levels. By contrast, under VOC-limited regimes, higher emission values are more consistently associated with positive SHAP values, indicating a greater probability of being classified as VOC-limited. This pattern is broadly consistent with the spatial distribution of OFS, in which NOx-limited regimes are more common in suburban or peripheral areas, whereas VOC-limited regimes are concentrated in urban cores with stronger anthropogenic emissions (Xue et al., 2014; Wang et al., 2021; Johnson et al., 2024). Overall, the regional differences in the SHAP responses of emission variables likely reflect differences in local emission intensity and chemical sensitivity.

Meteorological variables also show substantial but regionally distinct effects on OFS classification. SP, RH, and T emerge as influential predictors in several regions, but their directional effects vary geographically. In BTH and YRD, higher temperatures are more closely associated with $NO_x$-limited classification, which may reflect

enhanced photochemical activity and BVOCs emissions under warm conditions (Pétron et al., 2001; Duncan et al., 2009). In several other regions, however, higher temperature is more often linked to VOC-limited classification, which may also be related to the spatial distribution of urban heat islands (Zhang et al., 2026; Zhou et al., 2015). SP influences pollutant accumulation and vertical mixing through atmospheric stability and subsidence; in general, higher SP favors VOC-limited classification, whereas lower SP is more closely associated with $NO_x$-limited regimes, with particularly asymmetric SHAP responses in YRD. RH is especially important in BTH and YRD, where high humidity tends to promote VOC-limited classification. Higher relative humidity may promote aerosol growth, thereby reducing the radiation reaching the surface and limiting the photochemical production of HCHO (Ye et al., 2016).

Compared with VOC-limited and $NO_x$-limited regimes, the transitional regime shows less consistent directional responses across variables and regions. This suggests that transitional regimes are governed by a more complex combination of emissions and meteorological influences, with stronger nonlinearity and greater classification uncertainty. Such complexity implies that transitional regimes may be the most difficult to manage using simple precursor-reduction strategies alone and may require more detailed process-based analysis to support effective control design.

Overall, the SHAP analysis highlights that OFS classification is shaped by both shared and region-specific predictor responses. Emission-related variables remain the dominant contributors in most regions, whereas meteorological variables modulate the OFS transitions in a region-dependent manner. These results provide a more process-

informed interpretation of the regional heterogeneity and temporal evolution of OFS across China during the warm season.

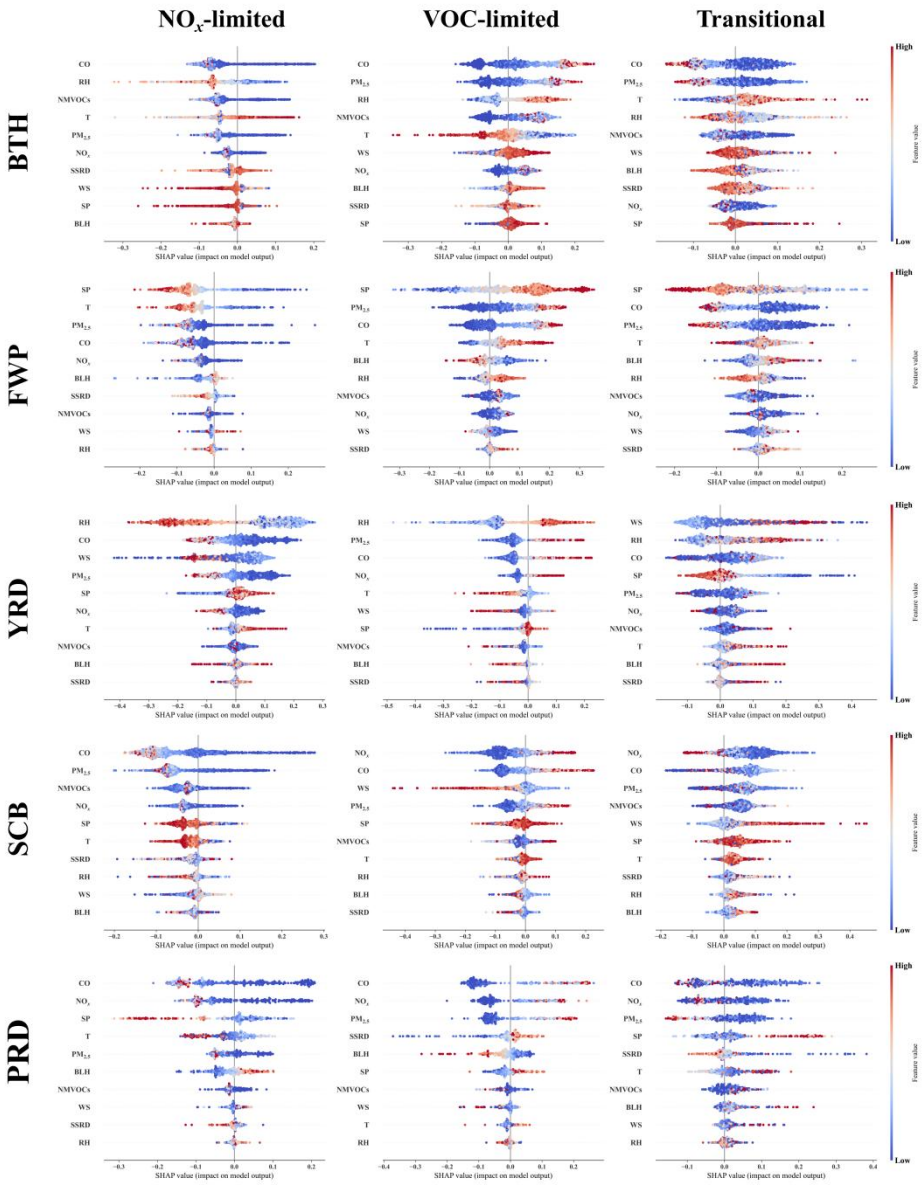



**Figure 9. SHAP value distributions of individual predictors in the RF models for the five urban clusters in China.**

17. The conclusion in L382 ("dominance of slowly varying meteorological drivers") seems inconsistent with the narrative across Sections 3.2 and 3.3. Section 3.2 emphasizes anthropogenic precursor reductions driving a shift from VOC-limited to

$NO_x$-limited regimes, whereas Section 3.3 claims meteorology contributes >60% and dominates OFS changes. Moreover, Section 3.3 appears to provide only an overall (time-aggregated) attribution. To reconcile these results, the authors should add time-resolved attribution analyses that track meteorology vs emissions contributions over time, in a way that directly corresponds to the trend/stage analyses in Section 3.2.

Response: We agree that the original conclusion was not sufficiently consistent with the narrative in Sections 3.2 and 3.3, particularly in overstating the role of meteorology relative to emissions. In response, we have substantially revised the analysis and the related text to improve temporal consistency and to better reconcile the trend analysis with the attribution results.

(1) Following the revision of the OFS framework, all analyses were refocused on the warm season (April–September) on a yearly basis and based on the newly derived region-specific FNR thresholds. The RF models were retrained using these updated OFS categories, which led to some changes in the attribution results and their interpretation. The revised manuscript now shows that emission-related variables generally remain the dominant predictors of OFS classification across most urban clusters, while meteorological contributions become more evident only under certain regional and temporal conditions.

We have modified the expression in Lines 497–500:

Overall, emission-related variables contribute more than meteorological variables, with their total contribution ranging from 50.1% in YRD to 72.2% in PRD. This result indicates that OFS classification is primarily constrained by emission-related factors in

most city clusters. In YRD, meteorological and emission-related variables contribute almost equally to OFS classification, accounting for 49.9% and 50.1%, respectively.

(2) To directly address the reviewer's concern about temporal consistency, we added a time-resolved attribution analysis by comparing two periods, 2005–2012 and 2013–2023, in a way that corresponds to the stage-based analysis in Section 3.2. The revised results show that emission-related variables remained dominant in most regions and OFS categories during both periods, generally accounting for more than 50% of the total contribution. At the same time, the relative contribution of meteorological factors increased modestly after 2013 in several regions and categories, indicating that meteorological regulation became more pronounced under a cleaner emission background, but did not replace emissions as the primary driver overall. We have added the relevant information in Lines 526–548:

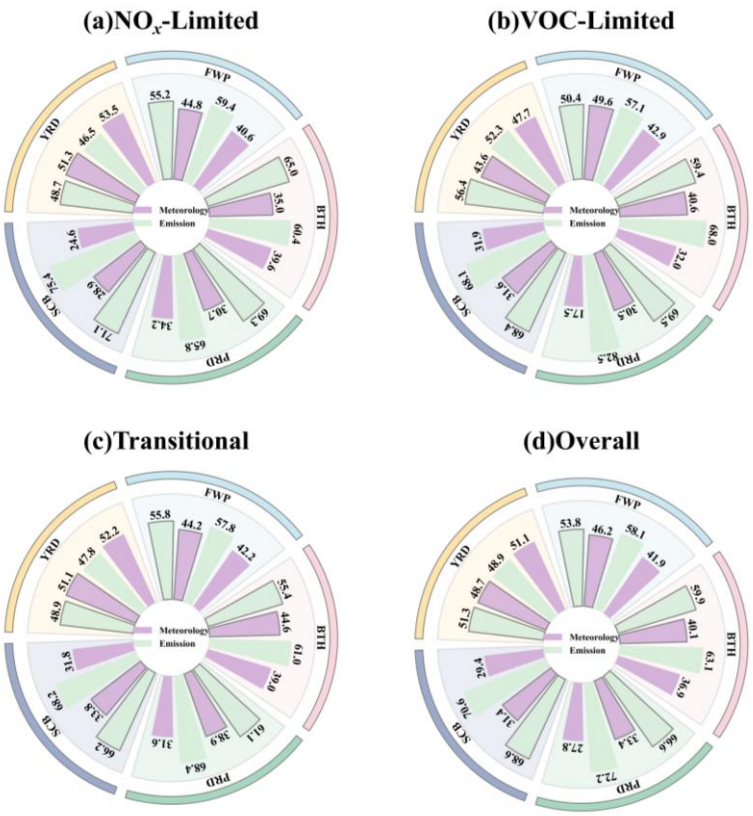
**3.4.3 Regional and temporal shifts in OFS-related factor contributions**

To further characterize the temporal evolution of the factors influencing OFS, we compared the relative contributions of emission-related and meteorological variables between the warm-season periods of 2005–2012 and 2013–2023 (Fig. 8). Overall, emission-related variables remained the dominant predictors across most urban clusters and OFS categories in both periods, with their contributions generally exceeding 50%. This suggests that the long-term evolution of OFS in China was primarily governed by changes in precursor emissions, although the relative role of meteorology varied by region and OFS category.

For the $NO_x$-limited regime, the contributions of meteorological and emission-

related factors in YRD were broadly comparable, whereas emission-related contributions in BTH, YRD, and PRD increased slightly after 2013 (+4.6%, +2.2%, and +3.5%, respectively). By contrast, meteorological contributions increased moderately in FWP and SCB (+4.2% and +4.3%, respectively). In the VOC-limited and transitional regimes, emission-related factors also remained dominant overall. However, meteorological contributions to the VOC-limited regime increased markedly after 2013 in BTH, FWP, and PRD, by +8.6%, +6.7%, and +13.0%, respectively. For the transitional regime, meteorological contributions increased in all urban clusters except YRD, accompanied by corresponding declines in emission-related contributions of approximately 2.0%–7.3%.

From an overall perspective (Fig. 8d), the relative contribution of emission-related factors decreased slightly after 2013 in BTH, FWP, SCB, and PRD, by about 2.0%–5.6%. In YRD, the relative contributions of meteorological and emission-related factors remained broadly comparable, although the balance shifted slightly toward emission-related influences after 2013. Taken together, these results indicate that, while OFS classification remained predominantly emission-driven, the regulatory role of meteorology became more pronounced after 2013. This pattern suggests a modest transition from a regime controlled mainly by emissions toward one jointly modulated by emissions and meteorological conditions. Such a shift is consistent with the substantial reduction in precursor emissions after 2013 and implies that, under a cleaner emission background (Fig. S5), OFS classification may become increasingly sensitive to meteorological variability.




876
877 **Figure 8. Relative contributions (%) of emission-related and meteorological factors to OFS classification in**
878 **five urban clusters during the warm season (April–September) for the periods of 2005–2012 and 2013–2023,**
879 **shown separately for (a) $NO_x$–limited, (b) VOC-limited, (c) transitional, and (d) overall OFS states. Gray**
880 **outlined bars denote 2013–2023, whereas bars without outlines denote 2005–2012.**

881

882

883

884

885

886

887

888

889

# Reviewer:2

**Comments to the Author:**

1. The manuscript tackles a central issue in China's "post-$PM_{2.5}$ era": how ozone formation sensitivity (OFS) evolves under the combined influence of emission controls and climate variability. The long-term, national-scale framework based on satellite precursors (OMI $NO_2$ and HCHO, 2005–2023), together with the integration of an indicator approach (FNR = $HCHO/NO_2$) and explainable machine learning (RF–SHAP), is compelling and potentially valuable for informing differentiated ozone-control strategies. Overall, the study is clearly structured and well-written. Methods are technically robust. The topic and findings are highly relevant to atmospheric chemistry and broadly align with the scope and scientific standards of Atmospheric Chemistry and Physics. Nevertheless, several issues should be addressed before the manuscript can be considered suitable for publication in ACP.

Response: We sincerely thank the reviewer for the positive evaluation of our study. We have revised the manuscript carefully in response to the comments, which were highly constructive and have helped us further improve the quality of the paper.

**Specific comments:**

2. Using policy issuance/implementation (e.g., 2013) as the breakpoint for phase division is reasonable. However, incorporating a formal change-point analysis (e.g., Pettitt test or a Bayesian change-point method) would substantially strengthen the argument by demonstrating whether $NO_2$, FNR, and/or the regime area fractions exhibit

statistically significant structural shifts around 2013. This would make the "policy-driven phase reversal" claim more robust, more publishable, and less open to challenge.

Response: We sincerely thank the reviewer for this valuable suggestion. We agree that defining the phase division solely based on the timing of policy implementation is not sufficient, and that a formal change-point analysis is needed to better support the stage-based framework.

In the revised manuscript, we have therefore added a Pettitt test to detect structural shifts in the warm-season annual series of $NO_2$, HCHO, and FNR for the five major urban clusters and the other regions of China during 2005–2023. The corresponding method, equations, and description have been added to the revised Methods section, and the results are summarized in Table S2.

The Pettitt test results show that the detected change points are generally concentrated around 2013–2017, although the exact timing varies among variables and regions. In particular, significant change points in FNR were identified around 2013–2014 in most regions, which provides quantitative support for using 2013 as the common dividing year in the subsequent analyses, because these structural shifts broadly coincide with the implementation of major clean-air policies in China.

At the same time, the test results also indicate that the phase behavior is not completely uniform across regions. For example, SCB did not show a statistically significant FNR change point, whereas PRD exhibited a distinct long-term increasing pattern rather than the clearer two-stage transition seen in BTH, FWP, and YRD. We have therefore revised the manuscript to make the interpretation more cautious and

region-specific, rather than implying that all regions follow the same policy-driven phase reversal.

Accordingly, the revised manuscript now uses 2013 as a common reference year supported by formal change-point analysis, while also explicitly acknowledging regional differences in the timing and form of structural shifts. This revision makes the staging framework more robust and less dependent on policy timing alone.

We have clarified it in Lines 201–215:

To further identify structural shifts in the long-term evolution of $O_3$ precursors and OFS, and to provide an objective basis for subsequent phase-based analyses, we applied the Pettitt test to the annual time series of $NO_2$, HCHO, and FNR in each region. The Pettitt test is a non-parametric method for detecting a single abrupt change point in a time series and does not require the data to follow a normal distribution (Pettitt, 1979). It has been widely used in environmental studies to detect statistically significant shifts in long-term observations (Baruah et al., 2021). The test statistic $U_{t,n}$ is defined as:

$$U_{t,n}=\sum_{i=1}^{t}\sum_{i=t+1}^{n}\mathrm{sgn}(x_j-x_i)\ (t=1,2,...,n)\ , \quad (5)$$

where $x_i$ and $x_j$ are the values of the time series, and sgn (·) denotes the sign function. The change point is identified as the time corresponding to the maximum absolute value of $U_{t,n}$, and its significance is evaluated as:

$$p = 2\exp\left(-\frac{6k_t^2}{n^3+n^2}\right)\ , \quad (6)$$

where $k_t=\max_{1\le t<n}|U_{t,n}|$. A p-value below 0.05 was considered statistically significant.

In this study, the Pettitt test was applied to the annual warm-season (April–

September) series of $NO_2$, HCHO, and FNR for each region during 2005–2023. For each variable, the year corresponding to the maximum absolute test statistic was taken as the detected change point. Because the Pettitt test identifies only one abrupt change point, it may not fully capture series with multiple shifts or several equally strong candidates (Zhang and Song, 2015). Therefore, the detected change points were interpreted together with the temporal evolution of each variable and the broader policy context.

**Table S2. Mann–Kendall trend test and Sen's slope estimate for HCHO, $NO_2$, and FNR in the five major urban agglomerations in China during the warm season (April–September) from 2005 to 2023.**

| Variable | Region | Sen's slope | MK p-value | Pettitt's change point | Pettitt's p-value |
|---|---|---|---|---|---|
| $NO_2$[a] | BTH | -0.23 | < 0.05 | 2015 | < 0.05 |
| | FWP | -0.09 | < 0.05 | 2014 | < 0.05 |
| | YRD | -0.14 | < 0.05 | 2014 | < 0.05 |
| | SCB | -0.01 | > 0.05 | 2008 | > 0.05 |
| | PRD | -0.20 | < 0.05 | 2014 | < 0.05 |
| HCHO[a] | BTH | 0.15 | < 0.05 | 2013 | < 0.05 |
| | FWP | 0.12 | < 0.05 | 2015 | < 0.05 |
| | YRD | 0.12 | < 0.05 | 2015 | < 0.05 |
| | SCB | 0.09 | < 0.05 | 2017 | < 0.05 |
| | PRD | 0.06 | < 0.05 | 2015 | > 0.05 |
| FNR[b] | BTH | 0.06 | < 0.05 | 2014 | < 0.05 |
| | FWP | 0.06 | < 0.05 | 2013 | < 0.05 |
| | YRD | 0.04 | < 0.05 | 2013 | < 0.05 |
| | SCB | 0.00 | < 0.05 | 2017 | > 0.05 |
| | PRD | 0.06 | > 0.05 | 2014 | < 0.05 |

[a] HCHO and $NO_2$ are reported in units of $\times 10^{15}$ molec $cm^{-2}$. [b] FNR is dimensionless.

And Lines 282–287:

Before examining the spatial and temporal evolution of $O_3$ precursors and FNR, we first assessed whether their long-term changes were characterized by structural shifts. Pettitt test results (Table S2) indicate that significant change points in warm-season $NO_2$, HCHO, and FNR were generally clustered around 2013–2017, although the exact timing varied among variables and regions. Because these shifts broadly coincided with the implementation of the Air Pollution Prevention and Control Action Plan (2013–2017), 2013 was adopted as the common dividing year for the subsequent analyses. The following sections therefore compare the two periods 2005–2012 and 2013–2023.

3. FNR thresholds (VOC-limited < 1; $NO_x$-limited > 2) may vary across seasons, regions, and chemical environments. I recommend adding a clearer statement and discussion, preferably in the Conclusions, on the uncertainty and potential variability of these thresholds, and how such variability might influence regime classification and inferred trends.

Response: We sincerely thank the reviewer for this constructive comment. We fully agree that FNR thresholds are not universal and may vary with season, region, and chemical environment, and that such variability can directly affect OFS classification and the interpretation of long-term regime changes. In response, we have substantially revised the manuscript to make this uncertainty more explicit and to assess its implications more systematically.

(1) Instead of relying on fixed national thresholds such as VOC-limited < 1 and

$NO_x$-limited > 2, we derived region-specific FNR threshold ranges for six regions in China based on the fitted relationships between satellite-based FNR and ground-observed high-$O_3$ probability during the warm season. The revised results show clear regional heterogeneity in these thresholds, confirming that a uniform national threshold is insufficient for OFS diagnosis in China. We have clarified it in Lines 160–184:

OFS regimes were then identified according to the corresponding FNR threshold ranges. The threshold ranges were determined mainly following the method of Jin et al. (2020). In this study, region-specific FNR thresholds across China were derived by combining OMI satellite observations with surface $O_3$ measurements. This approach is also consistent with previous studies that derived threshold ranges from fitted $O_3$–FNR relationships (Wang et al., 2021; Ren et al., 2022; Chen et al., 2024). In practice, the transition of OFS from VOC-limited to $NO_x$-limited regimes is gradual rather than abrupt. This transition can be influenced by multiple factors, including topography, meteorological conditions, and the spatial resolution of satellite observations (Jin et al., 2020). Because these influencing factors vary across China, the corresponding FNR threshold ranges are also expected to differ among regions. For this reason, OFS regimes were identified using region-specific FNR threshold ranges. China was divided into six regions, including five major urban clusters with severe $O_3$ pollution, namely BTH, FWP, YRD, SCB, and PRD, and one additional category comprising all remaining areas, referred to as other regions of China (ORC) (Lu et al., 2018). These six regions were used as the basic units for deriving region-specific FNR threshold ranges, which were then applied to identify OFS regimes in each region.

Because $O_3$ pollution in China occurs mainly during the warm season (April–September) (Lu et al., 2020b), the analysis was restricted to April–September in 2015–2023. To match the OMI overpass time, surface $O_3$ concentrations at 13:00 and 14:00 local time were averaged for each site and day. The resulting daily $O_3$ concentrations were then collocated with OMI FNR values at the corresponding grid cells. For each site, the high-$O_3$ probability was then defined as the fraction of days during the study period on which the 13:00–14:00 mean $O_3$ concentration exceeded 160 μg $m^{-3}$. The resulting paired samples were subsequently pooled within each region for threshold derivation.

To focus on polluted conditions, samples with $NO_2$ vertical column densities below $1.5 \times 10^{15}$ molec $cm^{-2}$ were excluded, following previous studies (Jin et al., 2020; Wang et al., 2021). In addition, high-end outliers (5% or 10%, depending on the region) were removed prior to fitting to reduce the influence of extreme values (Chen et al., 2024). After quality control, paired high-$O_3$ probability and FNR samples in each region were grouped into 100 bins according to FNR and a cubic polynomial was fitted to the binned relationship between high-$O_3$ probability and FNR. The FNR interval corresponding to fitted high-$O_3$ probability values above the 90th percentile was then defined as the threshold range for that region (Jin et al., 2020).

(2) To directly address the uncertainty raised by threshold variability, we added a dedicated threshold-sensitivity analysis using three threshold scenarios: S1, the warm-season thresholds derived in this study; S2, the minimum–maximum ranges compiled from previous studies; and S3, the average of the corresponding minimum and

maximum values from the literature. Based on these three schemes, we reassessed OFS classification in the five major urban clusters during the warm season from 2005 to 2023. The results show that threshold choice can substantially alter the estimated proportions of VOC-limited, transitional, and $NO_x$-limited regimes, and that the magnitude of this effect is strongly region dependent. We have added relevant information in Lines 352–375:

Given these pronounced regional differences, we further assessed how threshold uncertainty affects OFS classification. Three threshold-range scenarios were designed. Specifically, Scenario 1 (S1) used the threshold ranges derived in this study for the warm season of 2015–2023, Scenario 2 (S2) adopted the minimum–maximum threshold ranges reported in previous studies, and Scenario 3 (S3) used threshold ranges calculated from the average of the corresponding maximum and minimum values. The threshold ranges for the six regions under S1-S3 scenarios are listed in Table S1. These three scenarios were then used to evaluate the sensitivity of OFS classification to different FNR threshold selections in the five major city clusters during the warm season over 2005–2023 (Fig. S4).

The results show that OFS classification is strongly affected by threshold choice, although the magnitude of this effect varies markedly across regions. Among the five regions, FWP shows the largest sensitivity to threshold variation, with changes ranging from −53.8 % to 51.6 % and an average value of about 22.0 %. SCB also exhibits substantial sensitivity, with corresponding changes of −48.8 % to 52.0 %. In contrast, PRD is relatively less sensitive, with changes ranging from −36.3 % to 21.1 % and an

average value of about 17.0 %. BTH and YRD show intermediate responses, with ranges of −31.0 % to 37.2 % and −31.5 % to 31.5 %, respectively. These results indicate that the uncertainty introduced by threshold selection is not spatially uniform but depends on regional photochemical characteristics and the width of the transitional range.

Figure S4 further shows that the influence of threshold choice is often asymmetric, with some scenarios producing much larger positive or negative deviations than others. This indicates that OFS classification does not respond linearly to threshold perturbations. In particular, threshold ranges compiled from previous studies can produce substantial departures from the classification results obtained using the region-specific thresholds derived here. Even the average-threshold scenario still leads to noticeable deviations in several regions. These findings suggest that threshold transferability across regions is limited and that applying literature-based thresholds without regional constraints may introduce considerable bias into OFS diagnosis.

Overall, these results demonstrate that uncertainty in FNR thresholds can substantially alter the estimated proportions of different OFS regimes. Region-specific thresholds derived from regional observational relationships therefore provide a more robust basis for OFS classification than fixed or literature-averaged thresholds.

**Table S1. Comparison of the FNR threshold ranges for the five city clusters during the warm season (April–September, 2005–2023) derived in this study and reported in previous studies. Values in brackets indicate the threshold ranges.**

| Region | FNR threshold ranges | | | Threshold Range | Studied period | Reference | Spatial Coverage |
|---|---|---|---|---|---|---|---|
| | Scenarios 1 (S1[1]) | Scenario 2 (S2[2]) | Scenario 3 (S3[3]) | | | | |
| BTH | 2.33[1.51, 3.30] | [1.00, 5.10] | [1.92, 2.85] | [2.0, 3.1] | Apr.–Sep. 2019–2022 | BTH | Song et al., 2023 |
| | | | | [4.0, 5.1] | May–Oct. 2018–2022 | Jing-Jin-Ji-Lu-Yu (JJJLY) | Li et al., 2024 |
| | | | | [3.0, 3.8] | Apr.–Sep. 2019–2021 | Beijing-Tianjin-Hebei-Shandong | Ren et al., 2022 |
| | | | | [1.142, 2.268] | May–Nov. 2016–2018 | North China (NC) | Du et al., 2022 |
| | | | | [1.2, 2.0] | Apr.–Oct. 2013–2014 | | |
| | | | | [1.1, 1.9] | Apr.–Oct. 2015–2016 | BTH | Zhang et al., 2024 |
| | | | | [1.0, 1.8] | Apr.–Oct. 2017–2019 | | |
| FWP | 3.42[2.26, 4.77] | [1.14, 4.30] | [1.96, 3.19] | [1.142, 2.268] | May–Nov. 2016–2018 | North China (NC) | Du et al., 2022 |
| | | | | [1.524, 3.001] | May–Nov. 2016–2018 | Northwest China (NWC) | Du et al., 2022 |
| | | | | [3.2, 4.3] | Apr.–Sep. | Shanxi- | Ren et al., 2022 |

| | | | | | |
|---|---|---|---|---|---|
| | | | 2019–2021 | Shaanxi-Henan | |
| | | [2.9, 4.0] | May–Oct. 2018–2022 | Jiang-Zhe-Hu-Wan (JZHW) | Li et al., 2024 |
| | | [2.2, 3.3] | Apr.–Sep. 2019–2021 | Shanghai-Jiangsu-Zhejiang-Anhui | Ren et al., 2022 |
| YRD | 1.59[0.90, 2.37] | [1.00, 4.00] | [1.72, 2.64] | | |
| | | [1.3, 2.1] | Apr.–Oct. 2013–2014 | | |
| | | [1.2, 2.0] | Apr.–Oct. 2015–2016 | YRD | Zhang et al., 2024 |
| | | [1.0, 1.8] | Apr.–Oct. 2017–2019 | | |
| | | [3.1, 4.2] | May–Oct. 2018–2022 | Chuan-Yu (CY) | Li et al., 2024 |
| | | [0.887, 2.479] | May–Nov. 2016–2018 | Southwest China (SWC) | Du et al., 2022 |
| SCB | 3.24[1.93, 4.73] | [0.89, 4.30] | [1.70, 2.85] | | |
| | | [2.5, 4.3] | Apr.–Sep. 2019–2021 | Sichuan-Chongqing-Guizhou-Yunnan | Ren et al., 2022 |
| | | [1.4, 2,2] | Apr.–Oct. 2013–2014 | | |
| | | [1.2, 2.0] | Apr.–Oct. 2015–2016 | Chuan-Yu (CY) | Zhang et al., 2024 |
| | | [1.1, 1.9] | Apr.–Oct. | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | 2017–2019 | | |
| | | | | [2.3, 3.4] | May–Oct. 2018–2022 | Guangdong-Hong Kong-Macao-Guangxi-Hainan | Li et al., 2024 |
| | | | | [1.041, 3.068] | May–Nov. 2016–2018 | South China (SC) | Du et al., 2022 |
| PRD | 2.62[1.66, 3.70] | [1.04, 3.80] | [2.02, 3.13] | [2.7, 3.8] | Apr.–Sep. 2019–2021 | Guangdong-Hong Kong-Macao-Guangxi-Hainan | Ren et al., 2022 |
| | | | | [2.4, 3.2] | Apr.–Oct. 2013–2014 | | |
| | | | | [2.0, 2.8] | Apr.–Oct. 2015–2016 | Guangdong | Zhang et al., 2024 |
| | | | | [1.7, 2.5] | Apr.–Oct. 2017–2019 | | |
| | | | | [2.3, 4.2] | May–Oct. 2016–2019 | China | Wang et al., 2021 |
| | | | | [2.2, 3.2] | Apr.–Sep. 2019–2021 | China | Ren et al., 2022 |
| China | 2.02[1.20, 2.97] | [1.65, 4.20] | [2.03, 3.51] | [1.92, 3.77] | Jun.–Aug. 2023 | China | Fan et al., 2025 |
| | | | | [2.10, 3.87] | Sep.–Nov. 2023 | China | Fan et al., 2025 |
| | | | | [1.64, 2.52] | May–Nov. 2016–2018 | China | Du et al., 2022 |
| ORC | 2.07[1.17, 3.08] | / | / | / | / | / | / |

[1] S1 denotes the region-specific thresholds derived in this study. [2] S2 denotes the minimum–maximum threshold ranges compiled from previous studies. [3] S3 denotes the averaged threshold ranges based on the corresponding minimum and maximum values.

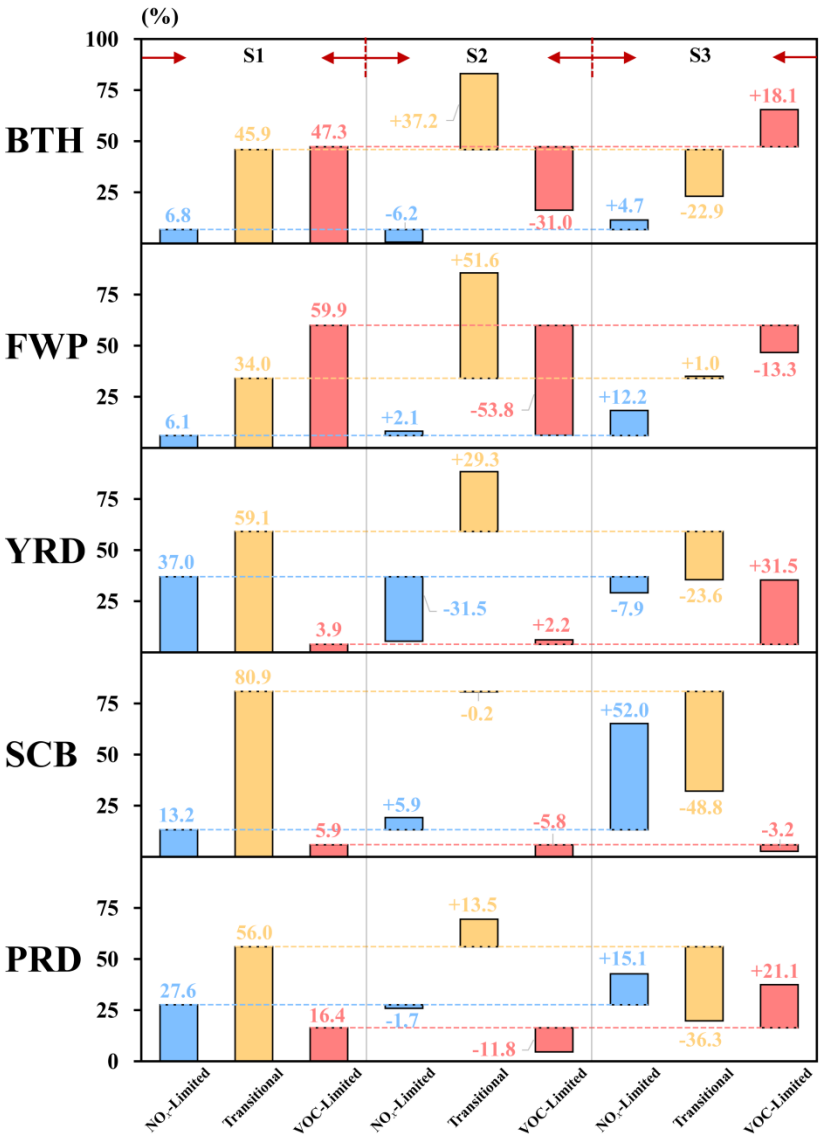



**Figure S4. Sensitivity of OFS regime area fractions to different FNR threshold scenarios in the five major city clusters during the warm season (April–September) over 2005–2023. S1 denotes the region-specific thresholds derived in this study, S2 denotes the minimum–maximum threshold ranges compiled from previous studies, and S3 denotes the averaged threshold ranges based on the corresponding minimum and maximum values. Bars represent deviations in regime area fractions under S2 and S3 relative to S1.**

4. While the manuscript cites relevant literature, a more explicit comparison with previous findings would better contextualize the novelty and contribution of this work.

For example, how do the identified meteorological drivers and their relative importance compare with those reported in other regions with similar climatic conditions or under comparable emission-control trajectories?

Response: We fully agree that a more explicit comparison with previous studies helps better position the novelty and contribution of this work. In response, we have expanded the discussion in the revised manuscript to place our findings in a broader context, especially regarding the identified meteorological drivers and their relative importance under different regional environments and emission-control backgrounds.

Specifically, we now compare our results with previous studies showing that meteorological factors such as temperature, relative humidity, solar radiation, and boundary-layer or stability-related variables play important roles in modulating $O_3$ formation and sensitivity in China and other regions (Chen et al., 2020; Liang et al., 2024; Xu et al., 2025). In line with earlier studies, our results confirm that meteorological influences are regionally heterogeneous and become more evident where OFS is more sensitive to humidity, atmospheric stability, or thermal conditions. For example, the relatively strong contributions of RH in YRD and SP in FWP are now discussed in the context of their regional climatic and atmospheric characteristics.

At the same time, our results differ from many previous studies in that we focus not on $O_3$ concentration itself, but on FNR-derived OFS states, which allows us to show how meteorological factors influence the classification and evolution of sensitivity regimes, rather than simply $O_3$ levels. We have clarified it in Lines 626–628:

Our focus is on the FNR-derived OFS, rather than absolute $O_3$ concentrations,

enabling us to elucidate how meteorology governs the classification and evolution of sensitivity regimes.

We have also clarified that, despite the important role of meteorology, our revised results show that emission-related variables remain the dominant predictors in most regions, while meteorological regulation becomes more pronounced under cleaner-emission conditions after 2013. This staged interpretation is now discussed more explicitly and compared with previous findings on $O_3$ responses under changing emission-control trajectories. In this way, the revised manuscript better highlights both the consistency of our results with the broader literature and the novelty of our framework in combining long-term OFS diagnosis with interpretable attribution analysis. We have added more relevant discussion in Lines 497–500:

Overall, emission-related variables contribute more than meteorological variables, with their total contribution ranging from 50.1% in YRD to 72.2% in PRD. This result indicates that OFS classification is primarily constrained by emission-related factors in most city clusters. In YRD, meteorological and emission-related variables contribute almost equally to OFS classification, accounting for 49.9% and 50.1%, respectively.

And Lines 526–561:

**3.4.3 Regional and temporal shifts in OFS-related factor contributions**

To further characterize the temporal evolution of the factors influencing OFS, we compared the relative contributions of emission-related and meteorological variables between the warm-season periods of 2005–2012 and 2013–2023 (Fig. 8). Overall, emission-related variables remained the dominant predictors across most urban clusters

and OFS categories in both periods, with their contributions generally exceeding 50%. This suggests that the long-term evolution of OFS in China was primarily governed by changes in precursor emissions, although the relative role of meteorology varied by region and OFS category.

For the $NO_x$-limited regime, the contributions of meteorological and emission-related factors in YRD were broadly comparable, whereas emission-related contributions in BTH, YRD, and PRD increased slightly after 2013 (+4.6%, +2.2%, and +3.5%, respectively). By contrast, meteorological contributions increased moderately in FWP and SCB (+4.2% and +4.3%, respectively). In the VOC-limited and transitional regimes, emission-related factors also remained dominant overall. However, meteorological contributions to the VOC-limited regime increased markedly after 2013 in BTH, FWP, and PRD, by +8.6%, +6.7%, and +13.0%, respectively. For the transitional regime, meteorological contributions increased in all urban clusters except YRD, accompanied by corresponding declines in emission-related contributions of approximately 2.0%–7.3%.

From an overall perspective (Fig. 8d), the relative contribution of emission-related factors decreased slightly after 2013 in BTH, FWP, SCB, and PRD, by about 2.0%–5.6%. In YRD, the relative contributions of meteorological and emission-related factors remained broadly comparable, although the balance shifted slightly toward emission-related influences after 2013. Taken together, these results indicate that, while OFS classification remained predominantly emission-driven, the regulatory role of meteorology became more pronounced after 2013. This pattern suggests a modest

transition from a regime controlled mainly by emissions toward one jointly modulated by emissions and meteorological conditions. Such a shift is consistent with the substantial reduction in precursor emissions after 2013 and implies that, under a cleaner emission background (Fig. S5), OFS classification may become increasingly sensitive to meteorological variability.

These findings are broadly consistent with previous regional studies, while also highlighting the strong spatial heterogeneity of OFS influencing factors across China. In the Guangdong–Hong Kong–Macao Greater Bay Area, FNR has been reported to be mainly influenced by temperature, shortwave radiation, total column water, and surface pressure (Chen et al., 2020), whereas in Guangdong Province it is primarily associated with surface solar radiation, relative humidity, and temperature (Liang et al., 2024b), although emission-related variables were not explicitly considered. By contrast, in the Chengdu–Chongqing region, summer OFS has been shown to be mainly driven by CO, $PM_{2.5}$, NMVOCs, and $NO_x$ emissions, with increasing pollution shifting OFS toward VOC-limited regimes (Xu et al., 2025). In the North China Plain, the decline in $PM_{2.5}$ concentrations reduces the loss of $HO_2$, enhances their reaction with NO, thereby strengthening $NO_x$ sensitivity and leading to increased $O_3$ levels (Li et al., 2019). Together, these studies support the view that OFS evolution reflects the combined influence of emissions and meteorology, but that their relative roles differ substantially among regions. Compared with these regional analyses, our results provide a unified national-scale framework showing that emission-related factors generally dominate OFS classification, while meteorological modulation becomes more evident under a

cleaner emission background.

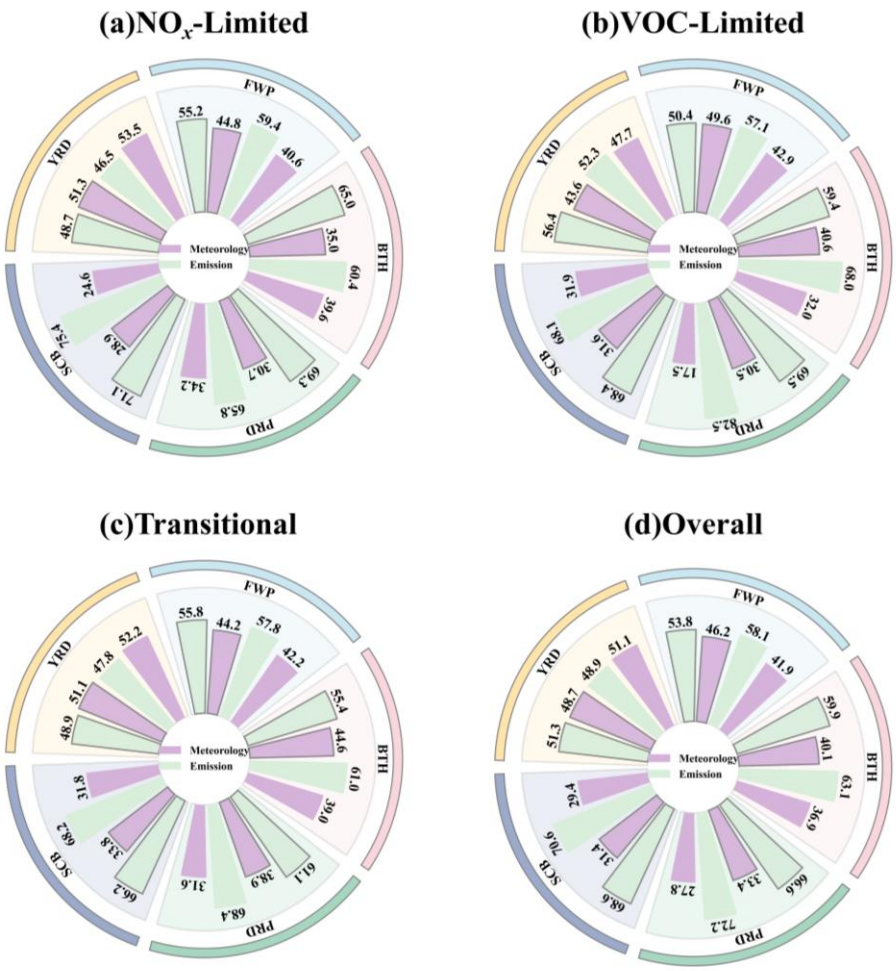



**Figure 8. Relative contributions (%) of emission-related and meteorological factors to OFS classification in five urban clusters during the warm season (April–September) for the periods of 2005–2012 and 2013–2023, shown separately for (a) $NO_x$-limited, (b) VOC-limited, (c) transitional, and (d) overall OFS states. Gray outlined bars denote 2013–2023, whereas bars without outlines denote 2005–2012.**

**Technical comments:**

1. The manuscript describes filtering HCHO pixels with cloud fraction > 30% and anomalous values (> $1.0 \times 10^{17}$ molec $cm^{-2}$), followed by a 3 × 3 moving average. However, it does not specify how cloud fraction is obtained (e.g., from an OMI cloud

mask/cloud product) or how “anomalous values” are defined (e.g., statistical outliers versus physically implausible retrievals). In addition, resampling HCHO (0.1° × 0.1°) to the $NO_2$ grid (0.25° × 0.25°) may introduce spatial-averaging biases; please clarify the resampling method (e.g., nearest neighbor, bilinear interpolation, area-weighted averaging). These details are important for reproducibility and for interpreting the robustness of the derived trends and regimes.

Response: We sincerely thank the reviewer for this careful and constructive comment. We agree that these processing details are important for reproducibility and for evaluating the robustness of the derived HCHO fields, FNR trends, and OFS classification.

In the revised manuscript, we have clarified three aspects of the HCHO preprocessing. First, we now explicitly state that the cloud fraction information was taken directly from the OMI Level-3 OMHCHOd product, which is generated from the OMI Level-2 OMHCHO retrievals after applying standard quality-control screening, including the exclusion of pixels with cloud fraction > 0.3, solar zenith angle > 70°, or those affected by the OMI row anomaly. Second, we clarified that the so-called “anomalous values” refer not to statistical outliers identified from our study period, but to physically implausible or retrieval-abnormal values outside the retained range of $-0.5 \times 10^{16}$ to $1.0 \times 10^{17}$ molec $cm^{-2}$, following Zhu et al. (2017) and subsequent studies. Third, we now explicitly state that the HCHO data were resampled from 0.1° × 0.1° to 0.25° × 0.25° using bilinear interpolation in order to match the $NO_2$ grid.

We have clarified it in Lines 123–131:

HCHO data were obtained from the Level-3 OMHCHOd product with a spatial resolution of 0.1° × 0.1° (Chance, 2019). This product is generated based on the Level-2 OMHCHO retrievals after applying quality control filters to exclude pixels with cloud fractions > 0.3, solar zenith angles > 70°, or those affected by the OMI row anomaly. The HCHO columns were retrieved using the Smithsonian Astrophysical Observatory (SAO) algorithm, with typical values ranging from $4 \times 10^{15}$ to $4 \times 10^{16}$ molec $cm^{-2}$ and with a detection limit of approximately $1.0 \times 10^{16}$ molec $cm^{-2}$ (González Abad et al., 2015). Following Zhu et al. (2017), only values between $-0.5 \times 10^{16}$ and $1.0 \times 10^{17}$ molec $cm^{-2}$ were retained, while values outside this range were treated as outliers and removed. To reduce random noise, a 3 × 3 moving average based on the eight-neighborhood was applied to the HCHO fields. The smoothed HCHO data were then resampled to a 0.25° × 0.25° grid using bilinear interpolation to match the $NO_2$ product, following an approach similar to that of Koukouli et al. (2016) for $SO_2$ data.

2. Figure 2d shows high HCHO values over the western China, including the Qinghai-Tibet Plateau. Is it artificial from satellite retrieval? If not, what's the possible source? In addition, the contour color of the small panel is not consistent with the main figure. Please clarify.

Response: We sincerely thank the reviewer for this careful and important comment. Using the retrieval-uncertainty information provided in the OMI OMHCHOd product, we found that western China, particularly the Qinghai–Tibet Plateau, exhibits substantially larger HCHO retrieval uncertainties than eastern China. This indicates that

the high HCHO values in this region should be interpreted with caution and may be affected, at least in part, by retrieval uncertainty under complex terrain and surface conditions. At the same time, we do not exclude the possibility that some of the elevated values may also reflect real regional influences, including topographic confinement and locally enhanced VOC accumulation in mountainous areas, as suggested by previous studies. We have therefore revised the text to present this issue more carefully and to avoid over-interpreting the western HCHO enhancement as purely physical or purely artificial. We have clarified it in Lines 307–312:

However, they may also be influenced by the relatively large retrieval uncertainties in western China, especially over the Qinghai–Tibet Plateau and surrounding areas, where complex terrain and surface conditions can affect satellite retrieval accuracy (Fig. S2; Xia et al., 2024). Therefore, the elevated HCHO columns in these regions likely result from a combination of real atmospheric signals and retrieval uncertainties and should be interpreted with caution.

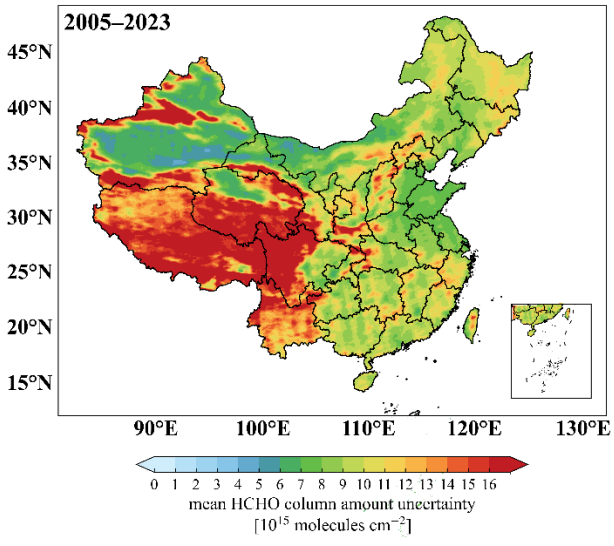



**Figure S2. Uncertainty of the mean HCHO column over China during the warm season (April–September) from 2005 to 2023.**

In addition, we have corrected the figure as follows:

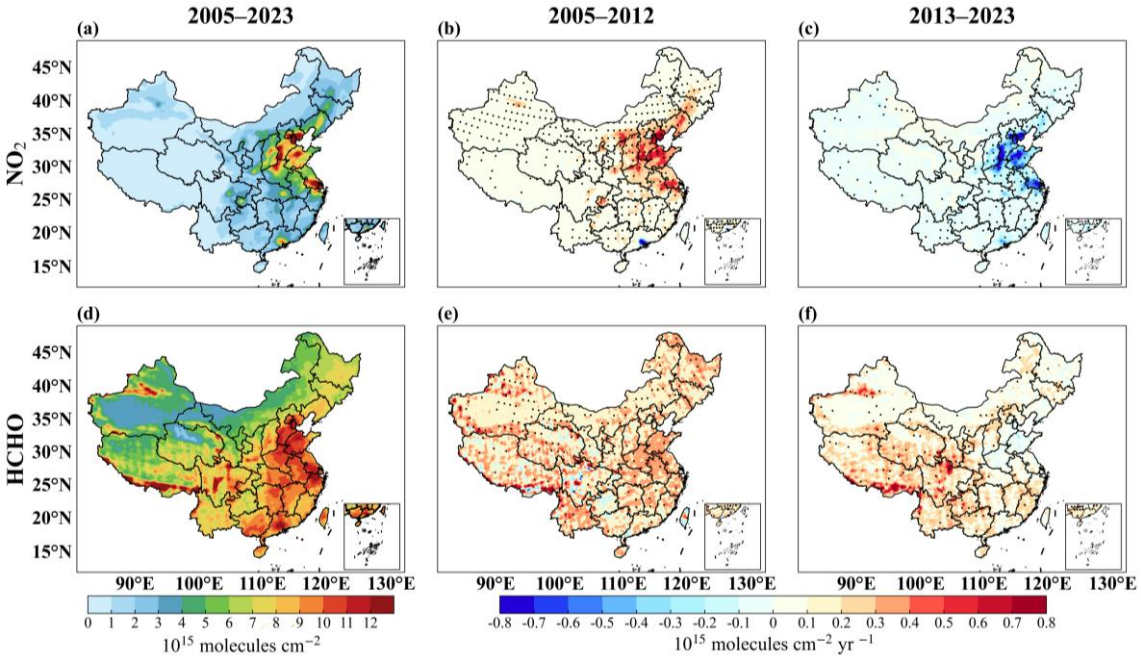



**Figure 2. Spatiotemporal variations of tropospheric $NO_2$ and HCHO columns over China during the warm season (April–September) from 2005 to 2023. (a, d) Distribution of annual average columns for 2005–2023; spatial distributions of the Sen's slope and Mann-Kendall (MK) test results for annual means during 2005–2012 (b, e) and 2013–2023 (c, f). Black dots in panels (b, c, e, f) mark grids that pass the MK significance test ($p < 0.05$).**

3. The caption of Figure 2 states 2013–2020, whereas the analysis elsewhere emphasizes 2013–2023. Please revise to ensure consistency and confirm that the stated MK–Sen time window matches the actual analysis period.

Response: We sincerely thank the reviewer for carefully pointing out this inconsistency. We confirm that the overall study period in this manuscript is 2005–2023, and that 2013 was treated as the turning point for the subsequent phase-based trend analysis. Accordingly, the MK–Sen trend analysis was conducted separately for the two periods, 2005–2012 and 2013–2023, rather than for 2013–2020. The previous statement in the caption of Fig. 2 was therefore incorrect and has been revised for consistency with the actual analysis as follows:

Figure 2. Spatiotemporal variations of tropospheric $NO_2$ and HCHO columns over China during the warm season (April–September) from 2005 to 2023. (a, d) Distribution of annual average columns for 2005–2023; spatial distributions of the Sen's slope and Mann-Kendall (MK) test results for annual means during 2005–2012 (b, e) and 2013–2023 (c, f). Black dots in panels (b, c, e, f) mark grids that pass the MK significance test ($p < 0.05$).

4. Please standardize the unit notation for $NO_2$ and HCHO columns (e.g., "molec $cm^{-2}$" or "molecules $cm^{-2}$") throughout the manuscript and clarify at first mention that FNR is dimensionless.

Response: We have standardized the units of $NO_2$ and HCHO column densities throughout the manuscript as "molec $cm^{-2}$" to ensure consistency. In addition, we have clarified at the first mention of FNR that, because HCHO and $NO_2$ have the same units, FNR is dimensionless. These revisions have been incorporated in the revised manuscript (Lines 158–159):

Because the two quantities have the same units, FNR is dimensionless.

5. Providing "last access" information (e.g., "last access: 6 September 2025") is helpful. I suggest also reporting the product version and/or DOI wherever applicable and ensuring consistent formatting and completeness across the manuscript (Methods, Data availability, and any Supplement).

Response: Modified as suggested.

## Reference


Chen, Y., Yan, H., Yao, Y., Zeng, C., Gao, P., Zhuang, L., Fan, L., and Ye, D.: Relationships of ozone formation sensitivity with precursors emissions, meteorology and land use types, in Guangdong-Hong Kong-Macao Greater Bay Area, China, J. Environ. Sci., 94, 1–13, https://doi.org/10.1016/j.jes.2020.04.005, 2020.

Chen, N., Yang, Y., Wang, D., You, J., Gao, Y., Zhang, L., Zeng, Z., and Hu, B.: Changing ozone sensitivity in Fujian Province, China, during 2012–2021: Importance of controlling VOC emissions, Environ. Pollut., 359, 124757, https://doi.org/10.1016/j.envpol.2024.124757, 2024.

Duncan, B. N., Yoshida, Y., Olson, J. R., Sillman, S., Martin, R. V., Lamsal, L., Hu, Y., Pickering, K. E., Retscher, C., Allen, D. J., and Crawford, J. H.: Application of OMI observations to a space-based indicator of $NO_x$ and VOC controls on surface ozone formation, Atmos. Environ., 44, 2213–2223, https://doi.org/10.1016/j.atmosenv.2010.03.010, 2010.

González Abad, G., Liu, X., Chance, K., Wang, H., Kurosu, T. P., and Suleiman, R.: Updated Smithsonian Astrophysical Observatory Ozone Monitoring Instrument (SAO OMI) formaldehyde retrieval, Atmos. Meas. Tech., 8, 19–32, https://doi.org/10.5194/amt-8-19-2015, 2015.

Jin, L., Loisy, A., and Brown, N. J.: Role of meteorological processes in ozone responses to emission controls in California's San Joaquin Valley, J. Geophys. Res. Atmos., 118, 8010–8022, https://doi.org/10.1002/jgrd.50559, 2013.

Jin, X., Fiore, A., Boersma, K. F., Smedt, I. D., and Valin, L.: Inferring Changes in Summertime Surface Ozone–$NO_x$–VOC Chemistry over U.S. Urban Areas from Two Decades of Satellite and Ground-Based Observations, Environ. Sci. Technol., 54, 6518–6529, https://doi.org/10.1021/acs.est.9b07785, 2020.

Liang, Y. N., Wang, X. H., Qi, S. M., Xu, J. M., and Liu, R.: Sensitivity Analysis of Ozone Generation in Guangdong Province Based on OMI Satellite and Ground Observation Data, Environ. Sci., 45, 6248–6254, https://link.cnki.net/doi/10.13227/j.hjkx.202311261, 2024 (in Chinese).

Lu, X., Hong, J., Zhang, L., Cooper, O. R., Schultz, M. G., Xu, X., Wang, T., Gao, M., Zhao, Y., and Zhang, Y.: Severe surface ozone pollution in China: A global perspective, Environ. Sci. Technol. Lett., 5, 487–494, https://doi.org/10.1021/acs.estlett.8b00366, 2018.

Lu, X., Zhang, L., Wang, X., Gao, M., Li, K., Zhang, Y., Yue, X., and Zhang, Y.: Rapid Increases in Warm-Season Surface Ozone and Resulting Health Impact in China Since 2013, Environ. Sci. Technol. Lett., 7, 240–247, https://doi.org/10.1021/acs.estlett.0c00171, 2020b.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA2017, https://doi.org/10.48550/arXiv.1705.07874, 2017.

Özüpak, Y., Alpsalaz, F., and Aslan, E.: Air quality forecasting using machine learning: Comparative analysis and ensemble strategies for enhanced prediction, Water Air Soil Pollut., 236, 464, https://doi.org/10.1007/s11270-025-08122-8, 2025.

Ozone Pollution Control Committee of Chinese Society of Environmental Sciences: China blue book on prevention and control of atmospheric ozone pollution (2023), Science Press, Chinese mainland, 164 pp., ISBN 9787030781840, 2024.

Ren, J., Guo, F., and Xie, S.: Diagnosing ozone–$NO_x$–VOC sensitivity and revealing causes of ozone increases in China based on 2013–2021 satellite retrievals, Atmos. Chem. Phys., 22, 15035–15047, https://doi.org/10.5194/acp-22-15035-2022, 2022.

Sadaiyandi, J., Arumugam, P., Sangaiah, A. K., and Zhang, C.: Stratified sampling-based deep learning approach to increase prediction accuracy of unbalanced dataset, Electronics, 12, 4423, https://doi.org/10.3390/electronics12214423, 2023.

Vazquez Santiago, J., Hata, H., Martinez-Noriega, E. J., and Inoue, K.: Ozone trends and their sensitivity in global megacities under the warming climate, Nat. Commun., 15, 10236, https://doi.org/10.1038/s41467-024-54490-w, 2024.

Wang, T., Xue, L., Brimblecombe, P., Lam, Y. F., Li, L., and Zhang, L.: Ozone pollution in China: A review of concentrations, meteorological influences, chemical precursors, and effects, Sci. Total Environ., 575, 1582–1596, https://doi.org/10.1016/j.scitotenv.2016.10.081, 2017.

Wang, W., van der A, R., Ding, J., van Weele, M., and Cheng, T.: Spatial and temporal changes of the ozone sensitivity in China based on satellite and ground-based observations, Atmos. Chem. Phys., 21, 7253–7269, https://doi.org/10.5194/acp-21-7253-2021, 2021.

Xu, W., Yang, H., He, M., Yang, Z., Zhang, Y., Liu, Z. H., and He, Y. M.: Spatial and Temporal Characteristics of Ozone Formation Sensitivity and Its Influencing Factors in Chengdu-Chongqing

Area, Environ. Sci., 46, 736–745, https://link.cnki.net/doi/10.13227/j.hjkx.202312065, 2025 (in Chinese).

Zhang, C., Xie, Y., Shao, M., and Wang, Q. g.: Application of machine learning to analyze ozone sensitivity to influencing factors: A case study in Nanjing, China, Sci. Total Environ., 929, 172544, https://doi.org/10.1016/j.scitotenv.2024.172544, 2024.

Zhang, S., Yuan, Z., Zheng, Z., and Zhao, K.: Quantitative impacts of dominant large-scale circulation systems on surface ozone pollution in China, J. Environ. Sci., 156, 42–55, https://doi.org/10.1016/j.jes.2024.07.015, 2025.

Zhu, L., Mickley, L. J., Jacob, D. J., Marais, E. A., Sheng, J., Hu, L., Abad, G. G., and Chance, K.: Long-term (2005–2014) trends in formaldehyde (HCHO) columns across North America as seen by the OMI satellite instrument: Evidence of changing emissions of volatile organic compounds, Geophys. Res. Lett., 44, 7079–7086, https://doi.org/10.1002/2017GL073859, 2017.