Response to Reviewer #1

[Comment] This paper uses explainable tree-based machine learning models (XML) to explore the role of meteorology in driving ozone variability in China. The approach shows strong performance in reproducing observed ozone variability, separating meteorological from non-meteorological influences, and identifying key meteorological drivers of ozone anomalies. It effectively addresses limitations of traditional statistical methods and CTM simulations, and the framework could serve as a useful reference for future studies in other regions or time periods.

In terms of scientific findings, the study highlights a shift in the dominant drivers of ozone trends around 2019 in China, along with an increase in meteorology-driven positive ozone anomalies between 2013 and 2023.

The paper is well written—clear, concise, and with uncertainties appropriately discussed. I recommend publication in ACP after the following comments are addressed.

[Response] We thank the reviewer for the positive evaluation of our manuscript. We are pleased that the reviewer finds the study well written and scientifically meaningful. We have carefully addressed all comments and revised the manuscript accordingly. Our detailed responses are provided below.

[Comment] Line 135: The number of cities is missing in Figure 1.

[Response] We thank the reviewer for catching this omission. We have now added the number of cities for each region in Figure 1 (NCP: 12 cities; YRD: 19 cities; PRD: 9 cities). The updated figure is included in the revised manuscript.

[Comment] Equation 3: What does the subscript "l" represent? Should it be "i" instead? The same question applies to Equations 4 and 5.

[Response] We thank the reviewer for noting this inconsistency. The subscript should be *i*, representing the index of meteorological variables. We have revised Equations 3 to 6 to uniformly use the subscript *i*.

[Comment] Figure 2: How should the negative SHAP values be interpreted - for example, for T2M below 290 K?

[Response] SHAP values quantify the contribution of each feature to the model prediction relative to the base SHAP value (*C* in equation 1). A negative SHAP value indicates that the feature acts to decrease the predicted ozone concentration relative to the baseline (i.e., the mean model prediction over the training dataset).

For example, in Fig. 2, T2M values below ~290 K generally correspond to negative SHAP values, indicating that relatively low temperatures tend to suppress ozone formation and thus reduce the predicted ozone concentration relative to the model baseline.

We have added a brief clarification in Section 3.1 to help readers interpret the sign of SHAP values:

"Here, positive (negative) SHAP values indicate that the feature value increases (reduces) the predicted ozone concentration relative to the baseline expectation (*C* in Eq. (1))."

[Comment] The ML model is trained on city-level ozone measurements. Are the meteorological inputs also averaged at the city-level (Line 130)? Please clarify. How many ozone monitoring sites are typically included per city in this analysis? Would the model still perform effectively (e.g., achieve high R2) if only a single site were available for a city, given that the meteorological data have a relatively coarse resolution (~50 km)? Does this approach require multiple representative sites per city to be reliable?

[Response] Thank you for raising this important question. We clarify as follows:

- Yes, we averaged meteorological inputs at the city-level too.

- The number of monitoring sites per city varies from 3 to 27, with a median of 8 sites across the three regions.

- We conducted a sensitivity analysis: we retrained the LightGBM model in NCP using only 1 site for each city. The R2 values changed by <0.02, suggesting that the model performance is still robust.

- We note that using multiple monitoring sites provides a more representative estimate of city-level ozone and is therefore preferred. Specifically: (i) it reduces representativeness errors between point measurements and coarse-resolution reanalysis data, and (ii) it helps reduce the measurement noise as some sites have missing data.

These clarifications have been added to Section 2.1.

[Comment] Section 3.3: The authors clearly state that the de-weathering method is intended to identify meteorology-induced ozone anomalies (MOA). However, it may be helpful to explicitly note that this differs from analyzing extreme ozone events. Some extreme events may be mainly driven by non-meteorological factors (reflected in UNIX) but not accounted for in this analysis. Please consider adding a brief clarification for readers.

[Response] We agree with the reviewer that extreme events may be driven by non-meteorological factors. We have added the following clarification in Section 3.3:

"We note that extreme ozone events may also be influenced by non-meteorological factors (e.g., emission pulses or regional transport of precursors), which are captured by the UNIX term in our framework but are not included in the MOA calculation. Therefore, MOA should be interpreted strictly as the meteorologically driven component of ozone anomalies, rather than a comprehensive representation of all extreme ozone events."

Response to Reviewer #2

[Comment] The authors employed five machine learning algorithms, LightGBM, XGBoost, CatBoost, Random Forest, and Extra Trees to quantify the influences of meteorological factors on ozone variabilities across the three major city clusters in eastern China. The predictors include 14 meteorological variables from MERRA-2, together with two temporal features, the day of year and the UNIX time variable defined as a continuous, non-repeating count of days since 1-Jan 2013. Shapley Additive Explanations (SHAP) were used to quantify the overall feature importance as well as their influence on meteorology-induced ozone anomaly (MOA) during specific time period (e.g., Figure 4b).

This manuscript is generally well written and suited to the scope of ACP. It is a good practice to test and employ multiple machine learning algorithms. Such an approach is less prone to the uncertainties derived from the sole reliance of single algorithm. The authors demonstrate a solid understanding of machine learning (ML) application and potential challenges such as multicollinearity. The justification for retaining correlated predictors is reasonable. (Line 158). I therefore recommend minor revision prior to the acceptance of this manuscript.

[Response] We thank the reviewer for recognizing the strengths of the manuscript. We have carefully revised the manuscript following the suggestions. Our detailed responses are provided below.

Specific comments:

[Comment] The abstract is well written and informative. But for lines 31 to 32: I'm not sure the meaning of "indicating a strengthening meteorological amplification of pollution episodes rather than more frequent events"? Does it mean meteorology plays a more dominant role during ozone episodes compared to non-episode days?

[Response] We thank the reviewer for highlighting this ambiguity. Our intended meaning is that, over 2013–2023, the magnitude of positive meteorology-induced ozone anomalies (MOA) has increased significantly, while the frequency and duration of such events show no significant trends. In other words, when unfavorable meteorological conditions occur, they now tend to amplify ozone levels more strongly than in the past, but such conditions are not occurring more often. We have revised the sentence for clarity in the abstract:

"...indicating that meteorological conditions increasingly amplify the intensity of ozone pollution episodes, without a corresponding increase in their frequency or duration."

[Comment] Line 46: Ozone has been increasing since 2013. But policy published in 2013 is cited. It seems this 2013 policy is not relevant to the post-2013 observed ozone increases?

[Response] We appreciate this careful reading. To clarify the relationship between policy implementation and ozone trends, we now state:

"The Chinese government has implemented a series of clean air actions since 2013 to reduce anthropogenic emissions of air pollutants (State Council of the People's Republic of China, 2013; 2018). Early emission control efforts (2013−2017) primarily targeted $PM_{2.5}$ and NO_x reductions, which may have contributed to ozone increases under the VOC-limited photochemical conditions prevalent in eastern Chinese cities. In the subsequent phase (2018–2020) integrated VOC and NO_x controls were implemented to address ozone pollution (Liu et al., 2023; Wang et al., 2023)."

[Comment] Line 82: The description of "global explainability" may be slightly imprecise. global explainability is more about how ML model learns and interprets the data. Thus, it is more about the data as a whole but not ML model itself. Just as the authors describe the local explainability as "the data-point level" (Line 86).

[Response] We agree with the reviewer's point. We have revised the description to more precisely define global vs. local explainability as follows:

"XML information can be broadly categorized into global and local explainability (Flora et al., 2024). Global explainability characterizes how input features influence model predictions across the dataset, often by ranking input features in terms of importance..."

[Comment] Line 117: One might raise the skepticism that 14 variables are too many. It is best to include a short justification of why this is not an issue here. e.g., the amount of data for training is sufficient. Tree-based algos such as random forest can address collinearity through procedure such as bagging (trained on subset of features/samples and aggregation).

[Response] We thank the reviewer for this suggestion. We have added a brief justification in Section 2.2:

"While 14 meteorological predictors may appear numerous, the large sample sizes (~35,000–74,000 samples per region) provide sufficient statistical power to support model training without overfitting. In addition, tree-based ensemble methods inherently mitigate multicollinearity through bootstrap aggregation and random feature subsampling (Breiman, 2001, Chen and Guestrin, 2016)."

[Comment] Line 127: There might be a discrepancy in the variable of solar radiation between MERRA2 and ERA5. (See the TOAR paper by Lu et al.; https://doi.org/10.5194/acp-25-7991-2025, e.g., Figure S12). Modeling with MERRA2, solar

radiation might be estimated less important overall. This might partially explain why SR is quantified to be less important than T over YRD here (Figure 2b). There is nothing wrong in using MERRA2. But this potential discrepancy between these two reanalysis products is worth noting.

[Response] We thank the reviewer for this valuable observation. We have added a discussion of reanalysis uncertainty in Section 4:

"It is also noted that systematic differences in meteorological reanalysis products for certain variables may influence the relative importance assigned to predictors. For instance, Lu et al. (2025) reported discrepancies in surface shortwave radiation between MERRA-2 and ERA5 over eastern China, potentially affecting radiation-related attributions. While our conclusions regarding the dominant role of temperature remain robust, future studies could benefit from multi-reanalysis ensembles to quantify this uncertainty."

[Comment] Section 2.2: Could the authors provide more details regarding the setups for the five machine learning algorithms (e.g., tree number, learning rate, etc.)? Perhaps put them in supplementary?

[Response] Hyperparameters are now listed in Table S1:

Table S1. Hyperparameter configurations for the five machine learning models with full dataset across three regions (NCP, YRD, PRD)

Hyperparameter	NCP	YRD	PRD
LightGBM			
n_estimators	224	48	422
num_leaves	2098	1217	164
min_child_samples	16	19	17
learning_rate	1.26e-1	1.76e-1	1.16e-1
log_max_bin	7	9	9
colsample_bytree	7.83e-1	1.0	9.27e-1
reg_alpha	3.81e-2	3.60e-3	1.47
reg_lambda	56.1	6.15e-2	3.53e-2
XGBoost			
n_estimators	163	182	230
max_leaves	45	507	1859
min_child_samples	2.87e-1	3.46e-2	9.19e-3
learning_rate	8.18e-2	9.05e-2	8.74e-2
subsample	8.79e-1	8.79e-1	8.96e-1
colsample_bylevel	9.54e-1	7.37e-1	5.80e-1
colsample_bytree	9.83e-1	9.54e-1	9.45e-1
reg_alpha	7.07e-2	9.35e-2	4.50e-2
reg_lambda	12.6	3.60	35.6
CatBoost			
early_stopping_rounds	12	13	12
learning_rate	9.03d-2	7.55e-2	9.03e-2
n_estimators	655	1074	861
Random Forest			
n_estimators	211	201	432
max_features	4.52e-1	7.81e-1	5.18e-1
max_leaves	5000	5000	5000
Extra Trees			
n_estimators	247	54	389
max_features	1.0	1.0	1.0
max_leaves	5000	5000	5000

[Comment] Caption of figure 1: Could the authors explain what numbers in brackets here? Also in the right panel, what do the shaded areas along the line represent?

[Response] Thank you for pointing this out. We forgot to include brackets in Figure 1, and have updated the figure. The shaded areas represent ±1 standard deviation across cities within each region. We have updated Figure 1 and its caption as follows:

"Figure 1. Locations of ozone monitoring cities and regional averaged monthly MDA8 ozone concentrations analyzed in this study. The three regions are North China Plain (NCP, blue), Yangtze River Delta (YRD, Red), and Pearl River Delta (PRD, yellow). Numbers in brackets denote the number of cities. The right panels show monthly mean ozone concentrations in the three regions over January 2013 to September 2023. Shaded areas represent ±1 standard deviation across cities within each region."

[Comment] L161: There is confusion about how the ML actually being trained. For "Each region", does it mean all data within the region of city clusters are all learned as a whole? If so, in L171, shouldn't the predicted MDA8 ozone be on the d-th day in year y from a particular site? That is, the dimension of site should be included here?

[Response] All data within the region of city clusters are learned as a whole. In Line 171, the predicted MDA8 ozone is for a particular region, not for a particular site, therefore, the dimension of site is not needed in Equation (1). We have updated the description to clarify this point.

"For a given ML model, the predicted MDA8 ozone concentration on the d-th day in year y in a specific region ($O_3ori_{y,d}$) can be expressed as, ..."

[Comment] Regarding the attributable de-weathering approach in section 2.3, this is a relatively novel and interesting idea. It may be helpful, if the authors could consider providing a more detailed explanation of this part and clarify my following comment:

Shapley is primarily about assigning the contribution of each "player" (i.e., each meteorological variable), to each instance of ozone value. Therefore, the shapely values from all these variables are meaningful for that particular instance (i.e., ozone in a particular day). The authors "*compute the climatological mean SHAP values for each meteorological variable by averaging their daily SHAP values across an 11-year period within a 15-day moving window (Eq. (3))*". I am slightly unclear about the interpretation of directly averaging SHAP values across different instances.

For example, if we focus one variable, its SHAP value might be -5 for one instance and +5 for the next. Can these two values be averaged? While I understand that this is somewhat analogous to deriving mean feature importance from SHAP values. But in the case of quantifying feature importance, their absolute values in each instance should be computed first before averaging? I'm not implying SHAP values should not be averaged at all. But more explanations should be given.

[Response] We thank the reviewer for this insightful comment. The reviewer's understanding of the SHAP value is correct. It assigns the contribution of each meteorological variable to each ozone level instance.

In our approach, however, SHAP values are not averaged across heterogeneous or unrelated samples. Instead, we perform conditional averaging across years for the same calendar period (i.e., within a 15-day moving window centered on a given day of the year). This procedure aims to estimate the typical (climatological) contribution of each meteorological variable under comparable seasonal conditions. Therefore, the averaging is performed across temporally aligned samples (e.g., similar days across different years), rather than across fundamentally different meteorological regimes. This is conceptually analogous to calculating

a climatological mean and differs from calculating global feature importance, which requires absolute SHAP values.

We have clarified this point in Section 2.3 as follows:

"...by averaging their daily SHAP values across an 11-year period within a 15-day moving window (Eq. (3)). This averaging is performed across years for the same calendar period, aiming to quantify how a meteorological variable affects ozone levels under typical conditions for that time of year."

[Comment] Lines 251 to Lines 253: this sentence might cause confusion. the feature importance appears broadly consistent across the different algorithms, with similar rankings of the key meteorological predictors. The mean |SHAP value| are indeed slightly different, but they all point that T2M is the leading predictor in NCP. Therefore, I suggest authors to rephrase this sentence.

[Response] We agree with the reviewer and rephrased the sentence as followed:

"We observe considerable discrepancies in feature attribution across the five ML models. Although the ranking of key predictors is broadly consistent (e.g., T2M is the leading predictor in NCP), the magnitudes of SHAP values vary substantially across models, leading to uncertainty in their quantitative importance."

[Comment]Line 334: A very minor point. I suggest consistent indexing for figures/tables. i.e., should it be "Figure 4b" but not "Fig. 4b"? Could the authors check through their manuscript?

[Response] We have revised the manuscript accordingly.

References

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, 2001.

Chen, T. and Guestrin, C.: XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), Association for Computing Machinery, New York, NY, USA, 785–794, 2016.

Flora, M. L., Potvin, C. K., McGovern, A., and Handler, S.: A machine learning explainability tutorial for atmospheric sciences, Artif. Intell. Earth Syst., 3, e230018, 2024.

Liu, Y., Geng, G., Cheng, J., Liu, Y., Xiao, Q., Liu, L., Shi, Q., Tong, D., He, K., and Zhang, Q.: Drivers of increasing ozone during the two phases of clean air actions in China 2013–2020, Environ. Sci. Technol., 57, 8954–8964, 2023.

Lu, X., Liu, Y., Su, J., Weng, X., Ansari, T., Zhang, Y., He, G., Zhu, Y., Wang, H., Zeng, G., Li J., He, C., Li S., Amnuaylojaroen, T., Butler, T., Fan Q., Fan S., Forster, G. L., Gao, M., Hu, J., Kanaya, Y., Latif, M. T., Lu, K., Nédélec, P., Nowack, P., Sauvage, B., Xu X., Zhang, L., Li, K., Koo, J., and Nagashima T.: Tropospheric ozone trends and attributions over East and Southeast Asia in 1995–2019: an integrated assessment using statistical methods, machine learning models, and multiple chemical transport models[J]. Atmos. Chem. and Phys., 25, 7991–8028, 2025.

State Council of the People's Republic of China: Notice of the General Office of the State Council on issuing the Air Pollution Prevention and Control Action Plan (in Chinese), available at: https://www.gov.cn/zwgk/2013-09/12/content_2486773.htm (last accessed: 8 March 2026), 2013.

Wang, Y., Zhao, Y, Liu, Y., Jiang, Y., Zheng, B., Xing, J., Liu, Y., Wang, S., and Nielsen, C. P.: Sustained emission reductions have restrained the ozone pollution over China, Nat. Geosci., 16, 967–974, 2023.