Responses to Reviewers' Comments on "Multidecadal ozone trends in China and implications for human health and crop yields: A hybrid approach combining chemical transport model and machine learning" by Mao et al. (MS No.: acp-2023-1052)

We would like to thank the reviewers for the thoughtful and insightful comments. The manuscript has been revised accordingly, and our point-by-point responses are provided below. The reviewers' comments are *italicized*, our replies are in black font, and our new/modified text cited below is highlighted in **bold**.

Response to Referee #1

We thank the reviewer for the very helpful comments. The paper has been revised accordingly to address the reviewer's concerns point by point, and all changes are cited and discussed in the responses below.

Specific comments:

The authors addressed all comments satisfactorily but did not correct the plot scales in Figures 4, 6, S2, and S7-S9 so it is difficult to compare the results between the subplots, e.g., (a) vs (c). A single scale should be used for each figure so that readers can make visual comparisons between the subplots. The study is recommended for acceptance after the plot scales are corrected.

We thank the reviewer for the comments. The scale in each mentioned figure has been revised to a single scale accordingly.

Response to Referee #2

First, the observations are not only used for validation but also for training. Thus, the response "it is worth noting that the observations were only used for the purpose of model evaluation to assess the accuracy and robustness of the model" is problematic. Specifically, the scale mismatch issue is not resolved. The authors should recognize that a site and a grid of 0.25 deg are at different spatial scales, at which observations are not directly comparable.

We are sorry for the problematic statement. Yes, the observations were used both for training and validation. We intended to explain in our response that during both training and evaluation phases, observational data were not used as predictor variables and were solely utilized as the response variables and for comparison with model results. We also fully acknowledge that spatial scale mismatch is a common problem for model-observation comparison, and now discuss so more fully. We emphasize that the whole purpose of using machine learning (ML) here and in other similar studies is to minimize the biases of model output, whereby the biases can arise from incomplete model physics, input and parameter errors, numerical errors, coding errors, as well as representation errors (i.e., mismatch in spatial scales between model output and observations, as the reviewer pointed out), so that the output of the ML-enhanced hybrid model can be the closest to the observations for more accurate impact evaluation. That is, the biases arising from the spatial scale issue has indeed been considered and inherently addressed in our bias reduction approach. See below for our modified text:

P5 L176: "**The primary purpose of utilizing ML here was to minimize the biases of model output as compared with observations, whereby the biases could arise from incomplete model physics, input and parameter errors, numerical errors, coding errors, as well as representation errors (i.e., mismatch in spatial scales between model**

grid cells and site observations), so that the output of the hybrid model could have the closest values to the observations and enable more accurate impact evaluation. In this study, we used the LightGBM ML algorithm to integrate GEOS-Chem-simulated O_3 at a lower resolution with higher-resolution multi-source data to produce higher-resolution hourly O_3 and MDA8-O_3 fields. … (P5 L187) **The training and evaluation processes are both performed at the site level in accordance with the observations, whereby the predictor variables and model responses were first sampled at the same locations using the bilinear interpolation approach (Accadia et al., 2003). This approach of handling spatial scale mismatch between model grid cells and site observations has been commonly used in previous studies (e.g., Li et al., 2021). When predicting the gridded O_3 concentrations with the trained model, predictor variables at different spatial resolutions were all regridded to the same resolution of 0.25°×0.25° consistent with the ERA5 meteorological fields.** …"

Second, as shown in Fig. R2, it is doubtful whether it is necessary to use ERA5 to downscale simulated O3 concentrations. The results based on ERA5 showed no significant improvements over those based on MERRA2.

We thank the reviewer for the comments. To respond to the reviewer's concern, we have extensively conducted additional analysis using two separate meteorological datasets (ERA5 & MERRA2) for model training to investigate into whether using a higher-resolution meteorological fields truly brings benefits to the outcomes in terms of accuracy. We have shown the comparison of the performance between the MERRA2 and ERA5 datasets in the supplementary materials. Ultimately, we selected the results using the ERA5 dataset for further analysis due to its moderately higher accuracy in terms of the considered statistical metrics (e.g., as shown in **Fig. S3**, for MDA8 O_3, R^2 increases from 0.69 to 0.72, and RMSE decreases from 25.21 to 23.76 μg m^{-3}) as well as the inclusion of more refined spatial details within the original GEOS-Chem grid cells, because the primary objective of the bias correction process to achieve the highest-possible level of accuracy in ozone concentration estimation for further impact analysis. We now elaborate these points further in the revised manuscript:

P5 L191: "… **When predicting the gridded O_3 concentrations with the trained model, predictor variables at different spatial resolutions were all regridded to the same resolution of 0.25°×0.25° consistent with the ERA5 meteorological fields. By taking the advantage of these higher-resolution datasets, the hybrid approach can not only correct the biases of the GEOS-Chem-simulated O_3, but also refine them into a finer resolution. To evaluate if the hybrid approach truly benefits from using a higher-resolution meteorological fields, we also repeated the whole training exercise with the input meteorology of GEOS-Chem (MERRA2 at 2.0°×2.5°) instead of ERA5.**"

P8 L272: "… **To test if using the higher-resolution meteorological data offers better prediction accuracy compared with the original input meteorology of GEOS-Chem, the MERRA2 dataset driving GEOS-Chem was also used to train the model. We found that the higher-resolution ERA5 dataset performed better in reproducing observed O_3 concentrations with moderately smaller RMSE and larger R^2 (Fig. S3), demonstrating the level to which a higher-resolution meteorological dataset, despite not being strictly consistent with the input meteorology for the CTM, can help enhance the performance of the hybrid approach and help resolve finer spatial details within the original CTM grid cells.** In summary, …"

Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A., and Speranza, A.: Sensitivity of Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple Nearest-Neighbor Average Method on High-Resolution Verification Grids,

Weather and Forecasting, 18, 918-932, https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2, 2003.

Li, K., Jacob, D. J., Liao, H., Qiu, Y., Shen, L., Zhai, S., Bates, K. H., Sulprizio, M. P., Song, S., Lu, X., Zhang, Q., Zheng, B., Zhang, Y., Zhang, J., Lee, H. C., and Kuk, S. K.: Ozone pollution in the North China Plain spreading into the late-winter haze season, Proceedings of the National Academy of Sciences, 118, e2015797118, https://doi.org/10.1073/pnas.2015797118, 2021.