This study tried to compare the CTM-based and ML-based approaches to estimate surface sulfur dioxide concentrations from satellite vertical column density from OMI over eastern China. However, my strongest concern is about the novelty of this paper that simply combined results from two cases that were rather basic in each approach and the analyses and conclusions did not offer new insights and almost identical as the combination from prior studies using only CTM-based or ML-based approach. In addition, for each approach, the treatment was oversimplified. For example, the CTM-based approach was conducted using a rather coarse-resolution simulation over only 4 months in a single year to compare with observations of annual means in 4 years, with poor performance compared to recent studies indicated by the correlation coefficients. The ML-based approach was conducted using only 5 input variables with poor performance compared to recent studies as well. The paper would benefit from better identifying the novelty of this manuscript and more science-based refinements on each approach.

General Comments

1. Justification of the study region over eastern China. This study was for eastern China only, and should be reflected in the title. What is the rational of restricting the study area to eastern China only?
2. For CTM-based approach, it is recommended to conduct full-year simulations over each year to be more temporally representative.
3. Resolution mismatch between CTM and OMI satellite observations. It is recommended to conduct nested simulations over eastern China to have a finer resolution to be consistent with finer OMI observations.
4. Why XGBoot ML method specifically? How about other ML techniques?

Specific Comments

1. Line 23-24: what is the supporting evidence from this study for the statement that CTM-based approach is better than the ML-based approach over areas without monitoring sites.
2. Line 121-122: is one-month spin-up enough for CTM simulations? The ground-based observations are for annual means spanning 4 years. It is recommended to conduct full-year simulations over each year. Also, only one month in each season may not be representative to support seasonality analyses in Figures 6 and 7.
3. Line 129: it is suggested from this paper that OMI overpassing time over eastern China is 12:15 pm to 2:45 pm, while here it is suggested simulation outputs are only at 2 pm local time. It is recommended to sample simulation output from 12 pm – 3 pm to be consistent with OMI overpassing time.

4. Line 139: what are the uncertainties for using observations from OMI at 0.25 degree resolution to represent subgrid variability for simulations at 2 degree? At least a test using nested simulation at 0.25 degree over eastern China should be added.
5. Line 145-146: the statement is not consistent with Figure S1. From Figure S1, it shows large variability of the correlation coefficients over different years. For example, for results in January, the correlation coefficient is 0.76 in the year of 2015, compared to 0.33 in the year of 2018. Actually, Figure S1 supports that there is large variability over years, and thus simulations over each year should be conducted instead of using simulations over only 4 months in 2015.
6. Line 148-149: I view the scatter plot shows different concentration ranges in the year of 2015 and 2018 and thus there could be large spatial variability of sulfur dioxide concentrations. The more comparable slopes only indicate general regional convergence.
7. Line 159-160: what are the supporting statistics for using a depth of 15 splits?
8. Line 164-167: why not keep the inputs the same between the CTM-based and ML-based approaches? In other words, why not choosing GEOS-FP meteorological fields as used in the CTM, as the inputs for training ML model? How would the differences between meteorological fields contribute to the differences of CTM-based and ML-based estimates?
9. Line 203: the simulation is not for the full year and thus it is not annual mean actually. It is recommended to conduct full year simulations over each year for temporal representativeness and to be consistent with the ground-based observations.
10. Figure 3: are the comparisons against all ground-based sites or over the same independent testing sites used in ML-based approach? Also, how are the simulation-observation pairs sampled? As the simulation is at very coarse resolution of 2 degree, while monitoring sites are single points. Are ground-based observations averaged over each grid box?
11. Line 282: what are the predictors used in Yang et al. (2023b) and Zhang et al. (2022)? What is the supporting evidence to restrict predictors to the specific 5 predictors used in this study? Are there any other physical predictors that need to be included in this study?
12. Figures 6 and 7: only one month in each season in a single year of 2015 is not representative for seasonality analyses over 4 years. It is recommended to conduct complete 4-year simulations.