Review of "Evolution and drivers of urban–suburban ozone contrasts in China during 2013–2020"

General comment

This manuscript combines surface observations and GEOS-Chem simulations to examine summertime urban-suburban O_3 contrasts in China, with particular emphasis on the effect of urban/suburban classification, the representativeness of monitoring sites, and the relative roles of meteorology, anthropogenic emissions, and biogenic VOC emissions. The topic is relevant, and the manuscript contains several potentially useful results. In particular, the comparison between the original monitoring-site classification and the population-density-based classification is informative. The regional analysis of eastern and western China and the JJJ and YRD case studies also provide a coherent framework for discussing spatial heterogeneity in ozone response.

I nevertheless recommend **major revision**. Currently, several of the central conclusions depend on methodological choices that are not yet tested or documented sufficiently. Here are more detailed comments that can hopefully improve the manuscript.

Major comments

1. **The population-density threshold is central to the paper, but its robustness is not demonstrated.**

   The manuscript shows convincingly that the inferred urban-suburban contrast can change substantially when the classification changes. That makes the choice of the new threshold particularly important. The value of 342 persons km^{-2} is obtained by averaging the minimum reported suburban densities from four cities and then selecting the closest mean density among national city-size categories. However, this threshold is still arbitrary. Because the main conclusions change with classification, the paper should show that they are not an artefact of choosing exactly 342 persons km^{-2}.

   I suggest, instead of having a binary classification, a more "continuous" approach; for instance, the authors could plot the trends in Figure 2 vs population density (using a bigger number of population density categories).

2. **Model evaluation needs to be substantially expanded.**

   The model evaluation is currently too limited, considering that most of the central conclusions after Sect. 3.2 rely on simulated regional means and sensitivity experiments. The evaluation is based mainly on NMB, temporal correlation, and linear trends of summertime O_3.

   In particular, I suggest that the authors include model-observation scatter plots for urban and suburban sites, preferably also separated between eastern and western China. These should report, in addition to r, standard performance metrics. Such plots would make it possible to

assess whether the model bias is systematic and whether model performance differs across the observed concentration range.

I would also strongly encourage an evaluation of the mean/median diurnal cycle of O_3, separately for urban and suburban environments and, for eastern and western China. Reproducing the diurnal evolution is particularly relevant here because the interpretation of urban-suburban differences involves photochemical production, NO titration, boundary-layer evolution, and precursor sensitivity. Agreement in seasonal mean MDA8 O_3 does not necessarily demonstrate that these processes are represented correctly.

In addition, because the ozone-regime analysis relies directly on simulated HCHO and NO_2, the authors should provide some evaluation of these species, or at minimum a quantitative discussion of their known model uncertainty and the implications for the derived HCHO/NO_2 ratios and regime classifications.

Finally, the model evaluation section should, in my opinion, be the first section of the results.

3. **The attribution to anthropogenic emissions, BVOCs, and meteorology is a sequential decomposition and is therefore path dependent in a nonlinear O_3 system.**

   The attribution appears to be constructed as Base − FixA for the anthropogenic contribution, FixA − FixAB for the biogenic contribution, and FixAB relative to 2013 for the meteorological contribution. This produces an additive decomposition, but it is not unique because O_3 chemistry is nonlinear and the contributions include interaction terms that are implicitly assigned according to the order in which the perturbations are applied. For example, the "anthropogenic" effect is evaluated under year-specific meteorology and BVOCs, whereas the "biogenic" effect is evaluated with anthropogenic emissions fixed at 2013 levels.

   The paper should, at minimum, discuss this order dependence. Ideally, the authors should test whether the signs and ranking of the major contributions are robust to an alternative perturbation order. If additional simulations are not feasible, the attribution should be described as a sequential sensitivity decomposition rather than as a unique quantitative partitioning, and the resulting uncertainty should be acknowledged when interpreting changes in the dominant driver.

Minor comments

1. Some information about the statistical analysis is currently missing. Please specify the regression method, what the reported slope uncertainties represent, the test used for statistical significance, how the reported correlation coefficient r is calculated, and how error bars are calculated.

2. **L69-72:** Please provide a figure showing the extent of the domain used for the simulations.

3. **L105:** Why was a sixth-order polynomial function chosen for the fit?

4. **L143:** Please define MDA8.

5. **L147-149:** "Grid cells for which the aggregated monitoring-site classification was inconsistent with the population density-based model-grid classification were excluded from subsequent model-observation comparisons". I cannot understand this sentence at all; please reformulate.

6. **Figure 2:** It is not clear what the exact meaning of Model-Red and Model-Reg-Ext is.

7. **Figure 5:** It would be useful to include the red circles in all the panels.