**Review of "Ozone seasonal evolution and photochemical production regime in polluted troposphere in eastern China derived from high resolution FTS observations", Sun et al., ACPD.**

**Summary:**

The authors report on solar absorption FTIR measurements of tropospheric columns of O3, CO, and HCHO at a candidate NDACC IRWG observation station in Hefei, China. High spectral resolution measurements were conducted between 2014 and 2017 and fill a data gap within the NDACC observation network. The data shows higher tropospheric O3, also with higher variability, in spring and summer. The authors compare these O3 measurements to OMI satellite O3 (PROFOZ product), as well as GEOS-Chem (2 x 2.5 deg) and WRF-Chem (20 x 20 km) model O3 outputs. Comparisons are done in both profile and tropospheric partial column form.

Ozone FTS vs. GEOS-Chem model differences (481 coincidences) are attributed to uncertainties in GEOS-Chem input files ("ozone production loss rates and emission inventory"), it is concluded that GEOS-Chem is biased 13% lower (along profile), with r=0.5 for tropospheric column correlation plots.

Ozone FTS vs. WRF-Chem model differences (481 coincidences) are attributed to uncertainties in WRF-Chem input files ("ozone production and loss rates and MEIC inventory"), it is concluded that WRF-Chem is biased 12% lower (along profile), with r = 0.65 for tropospheric column correlation plots.

Comparisons to coincident OMI ozone profiles and partial (tropospheric) OMI columns were done on 53 coincident measurements after filtering for 0.7° spatial coincidence. Coincident FTS profiles were averaged in a 3 hour window around the OMI overpass at 13:30. OMI profiles were smoothed with FTIR averaging kernels. The OMI profiles are biased 2-13% lower than FTIR profiles, with r=0.73 for tropospheric column correlation plots, in which most OMI points sit below the 1:1 line, indicating also a low bias of OMI w.r.t. FTS.

Both sets of model ozone data are described as "smoother" than FTIR data and are "bias corrected" by adding a constant offset to the tropospheric O3 columns throughout the year to shift the model data towards FTIR partial column values. GEOS-Chem partial columns are increased by ~100% while WRF-Chem partial columns are increase by ~33% to increase agreement with FTS. Finally, OMI ozone partial column data were increased by ~20% and only then were monthly mean ozone partial column differences calculated.

24-hour back trajectories were calculated arriving at Hefei at 3000 m.a.s.l. from 2014-2017, presumably for those days with FTS observations (?), and they were grouped into spring/summer (presumably MAM/JJA?) and autumn/winter (presumably SON/DJF?). Summer transport is less vigorous and more varied than winter transport, as expected, bringing more air from highly polluted areas, e.g., east China, and broadly accounting for higher O3 and higher O3 variability in the data in spring/summer.

Finally, the O3 production regime is analyzed by describing correlations to meteorological variables recorded at Hefei, as well as looking at O3 vs. CO, O3 vs NO2 (for ratios of HCHO/NO2 > 2.8, assumed to correspond to NOx-ltd O3 production) and O3 vs HCHO (for ratios of HCHO/NO2 < 1.3, assumed to correspond to VOC-ltd O3 production). The ratios to indicate the O3 production regime were found iteratively until the correlation between O3 and NO2 or O3 and HCHO was > 0.6. 106 days of observations (O3, HCHO, CO from FTS; NO2 from OMI) were identified and of those 60% were NOx-ltd, 11% were VOC-ltd, and the remainder were mixed.

**Major comments:**

The paper is generally well written and presents a thorough error budget and sensitivity analysis of FTIR retrievals (O3, CO, HCHO) from a new candidate station in the NDACC network. The methods used here are well known and figures 2-5 should also move to the appendix, along with the Rodgers & Connor formulation, unless the authors highlight how their averaging kernels and error budget profiles differ from other similar published results. The paper presents a valuable new and growing observational dataset, however, this reviewer recommends major revisions in order to meet the ACP criteria of scientific significance and quality.

The FTS O3 measurements are higher than both models (global and regional) and the OMI measurements. The FTS measures a total column through a particular atmospheric slant column, and is expected to be less sensitive to local O3 events than an in situ sensor. We expect generally good agreement with downward-looking OMI, although coincidences are always a challenge. We also expect differences in the FTS vs. model comparisons because of different representativeness offered by a 20x20 km model vs. as 2.5° x 2° model. This is not discussed in the paper. Also, for the 20x20 km WRF-Chem model, the profile up to 10 km could extend over two horizontal grid boxes for most SZAs > 45°, depending on the location of Hefei within a model grid box. Has this been considered?

Without discussing representativeness, the authors attribute FTS vs. model differences to model "input files", e.g., "ozone production loss rates and emission inventory" which is superficial. As a consequence, we learn little, if anything, about specific model processes and emission inventories that may be responsible. Also, why is the data from this candidate station considered as "truth" in the comparison to OMI and the models? The total errors are estimated as 10% but they are dominated by smoothing error and based on very tight Sa values for O3 (10%), so (as the authors note), they are an underestimate. If the authors plotted OMI vs. FTS trop O3 column data with both data sets' error bars they would still not overlap, but presumably OMI data has been validated – is it generally found to be low compared to other data?

The addition of a simple offset to model O3 values before looking at fractional monthly mean differences w.r.t. FTS is problematic because it is evident in figures 9 and 11 that such a simple manipulation does not bring the data points onto the 1:1 line. Instead, we have the highest O3 values below the FTS measurements and the lowest values above. This is even more dramatic in GEOS-Chem data, presumably because of lower model resolution, which homogenizes high O3 values over a large grid cell, while raising the background O3 values. Since the highest values occur in spring/summer and the lowest in autumn/winter, the bias is seasonally dependent and therefore not just due to spatial representativeness. Is it due to incorrect emissions or chemistry? What are the main chemistry and emissions differences between the two models being compared to FTS? WRF-Chem is running with the MEIC inventory, presumably optimized for China, as well as biogenic emissions from MEGAN – why does it only do a little bit better than GEOS-Chem?

About smoothing the OMI profile by the FTIR averaging kernels, this method is meant to be applied to high vertical resolution correlative data, which OMI is not. It has about ~1 DOF in the troposphere itself. This may explain why there is still a lot of "shape" left in the fractional difference between FTIR and smoothed OMI profiles. What do OMI kernels look like and where is its peak of sensitivity – is it the same as for FTS?

The trajectory cluster analysis is difficult to follow without familiarity with China's geography. That can easily be fixed by adding the major city and region names referred to in the discussion to Figure 13. Without this information, it is hard to quickly judge if 1-day trajectories are long enough for transport to occur to Hefei. It is also not clear how the trajectories are clustered and the mean cluster trajectories (in color) are hard to see. Another way to represent this data would be to count trajectory elements crossing, e.g., 0.5° x 0.5° grid boxes. Also, why 3000 m? That seems much higher than the typical boundary layer height in winter, and probably also in summer. This choice will influence strongly both the speed and footprint of the pollution regions influencing Hefei. Have the authors tried 1500 m?

Finally, regarding O3 production regimes, ratios of HCHO/NO2 were varied until the correlation was > 0.6 in plots of O3 vs. HCHO and O3 vs. NO2. The outcome is that the correlation for the NOx-ltd plot of O3 vs. NO2 is 0.66 (moderate) while the correlation for the VOC-ltd plot of O3 vs. HCHO is 0.92, with far fewer points remaining in the fit. This seems rather arbitrary and needs justification. Also of the 106 days available for this analysis, which are from spring/summer and which are from autumn/winter? Are all VOC-ltd days in winter?

When I look at the full O3 data in Figure 12, I wonder why there isn't a stronger signature of JJA O3 enhancements in Hefei? (Is it related to filtering out days affected by haze, App B?) Many high values seem to be in May, although the x-axis is hard to read and should really be changed to, Jan 1, June 1, etc., throughout the paper where dates are shown. Or possibly at boundaries between MAM, JJA, SON, DFJ, if these are the groupings for the seasons in the paper. Have the FTS partial columns been compared to in situ O3 monitors in Hefei to see if they also show enhancements in May/June 2015 and 2016? What about the low values in Jan 2015 and 2017 vs. the higher ozone in Jan 2016?

Finally, the Pearson coefficient of 0.35 – 0.6 was taken to mean "moderately correlated" in this work. Typically moderate correlation is associated with values of 0.5 – 0.8, since the lower bound would mean that the model fit to the data explains only 25% of the variations in the data. At 0.35 that drops to only 12%.

**Further detailed technical comments:**

Fig. 1a: most names in this figure are illegible. Use a cleaner map to reduce clutter.

Fig. 1b: no red hexagons are visible, but I assume the red arc is the azimuth and the un-described yellow circles are the SZA.

Fig. 2: what does "with measured ILS" mean in this caption? Is the ILS characterized with linefit and then fixed in the retrievals, or are some ILS parameters still being retrieved? Why is there a loss of sensitivity to HCHO right at the surface? Is this a priori related?

Fig. 3: the HCHO trop column AK seems unhealthy for growing so far past 1 quickly above ~3 km, even if there is little HCHO there. What is the reason for this shape?

Fig. 4: What is the explanation for the peak in the CO error at around ~3 km?

Fig. 5: Legend seems reversed for total random error and z shift for CO.

Fig. 9 and 11: it's hard to judge seasons with the date labels as presented. Also, why do these figures not have the identical number of O3 data points if they are derived from the same data filtering applied to FTS data that is described in App B?

Fig. 14: is based on Fig 12, not 13 as the caption says. Again, what are the model process and inventory differences leading to this? Panel a) says smoothed model, but is OMI not also smoothed in this figure?

Fig. 15: The wind sensor appears to be installed in a poor location as the wind speed never exceeds 0.3 m/s or ~1 km/h! If that is the case, then the wind direction data is also spurious. That's too bad, because I wanted to see a plot of Hefei O3 vs. wind direction to see if O3 is higher when winds blow from the city.

Fig 16: In spite of problems above, the highest O3 values occur for the lowest of the low wind speeds, pointing to the accumulation of local pollution. There is a "moderate" negative correlation between O3 and RH – why? We could learn more if these data were colored according to spring/summer and autumn/winter.

Fig. 19: hard to judge seasons with x-axis labels. Panel b is based on data in panel a that does not seem to sample seasons evenly. This should be discussed.

Table 1: retrieved interfering gases → as columns, I presume, except for H2O, as noted? Also, WM → MW. I'm not sure what footnote b means, please clarify.

**Manuscript:**

P1L74: sun spectra → solar absorption spectra

P1L3: replace wiki reference with something from the many, many refereed papers on Chinese modernization and growing air pollution problems.

P4L89: what are China's AQ standards in ppb for long- and short-term exposure?

P4L95: greatly contribute to ozone pollution controls → contribute to the evaluation of O3 pollution controls

P4L117: … after it is itself validated as an NDACC site and it moves from candidate to regular status.

P5L129: then increases → then SZA increases

P5L129-133: what region influences the measurements depends on the azimuth of observation, yes, but also on the direction and wind speed pushing air masses above Hefei, especially for the lowest parts of the atmosphere. This could be significant when local pollution events are occurring as some events can be completely swept away from the FTS obs path.

P6L173: cited references missing from references section

P6L178: please explain deweighting more clearly. What are instrument SNR levels without deweighting?

P7L187: how are the Sa diagonal element magnitudes chosen? WACCM?

P7L191: is the ILS retrieved in all retrievals or is it done with LINEFIT and then held constant?

P11L315-317: tagged O3 runs are mentioned, which would be nice and would allow the attribution of pollution to various source regions, but these 3 lines are very unclear (i.e., also about restart files)

P14L393-4: basically reproduced … but with slight shifts in timing → July is wrong in both models; why are they low in August? When is the local Hefei smog season?

P14L403: Logan (1985) “observed” → I presume this is a model study?

P16L448: basically consistent throughout all seasons → it really doesn’t look like that to me; would be easier to think about if time series started with MAM as opposed to JFM.

P16L457-63: this really is a shallow explanation of what may be causing the differences, from which we learn nothing concrete. Also, how does larger air pollution increase uncertainty in either emission inventories or the photochemical regime? Isn’t the latter, especially, something that is diagnosed from the emission rates and relative abundances of NOx and VOCs?

P17L495/6: which emissions are being discussed here: biogenic? anthropogenic? What are the expected magnitudes and timing of each?

P18L528: straightly applied → straight forwardly applied

P19L554: “validate OMI” → that’s a strong statement given the unproven nature of these particular FTIR measurements, and given there’s no reference to other OMI validation efforts and what they have typically revealed

P19L560: WRF-Chem agreement is “better” → it has a lower “bias” but greater summer differences. It’s not clear if that is better given it is a high res model using optimized emissions for China

P22 L651: would not screening out hazy days eliminate a lot of JJA O3 pollution days? Haze isn’t a problem for FTIR as much as non-constant intensity (e.g., clouds floating by during a ~20 minute observation time).