Author responses to Anonymous Referee # 1 RC#2
Referee comments in **boldface**, author comments in normal typeface, locations where revisions are made are prefaced with an underlined heading. All edits are prefaced with their PXXLYY location relative to the original discussions manuscript. For the SI only, all edits are prefixed with their new location.

Thank you for the detailed and excellent comments and suggestions which have strengthened the paper. We address them as responses to (1) your general comments and (2) the detailed comments, below. We direct attention to specific places in the text or supplement where clarifications exist or are being included in future versions of this manuscript. Overall, we find the reviews do not quite recognize the difficulty in obtaining observational data from areas where access is limited for various reasons. We would like to point out that despite lacking a dense observational data set, valuable scientific studies must continue to occur (with the appropriate caveats, as we think we provided in this study) to guide future efforts to fill these obvious data gaps. We feel there is significant value to studies, *particularly* in light of the fact that there is such a dearth of observationally-constrained analyses.

General Comments.

1. **The main conclusion of the paper is that a "Chinese specific inventory based on subnational data and domestic field-studies (20% higher than global inventories) agrees significantly better with observations than the global inventories at all timescales." The target of the paper is timely and interesting and certain to draw attention to the policy community outside of the academic trace gas community.**

   Thank you! The potential for broader impact was an important aspect of conducting this research.

2. **While I believe the comparisons between the inventories to be reasonable, relatively, I don't see how the authors can have faith in the overall accuracy of either of the inventories. The total national Chinese fossil fuel emissions are being estimated from a single 6 meter high tower on the outskirts of Beijing. The height of the tower is not even mentioned in the manuscript (that has to be added in main text and certainly highlighted as a caveat in Discussions/Conclusions). To this reviewer, at a minimum, this paper needs to explore the uncertainty of their result, in particular, with respect to atmospheric transport. At 6 meters CO_2 can even vary in a "well mixed column", let alone in shoulder transition times or as a function of PBL scheme. The authors mention McKain et al 2015 for uncertainty references. That paper itself is far from clear about the averaging up of errors in their "bootstrapping" technique and also appears to provide some evidence of robustness of the results to "configurations" of the system. That paper used 4 towers and multiple species, INFLUX (Indianapolis experiment) used a dozen towers and multiple species, and fossil projects in Paris and Los Angeles have similar constraints. If the authors are going to be bold enough to say that national inventories can be measured with a single tower located near the surface on the**

outskirts of town, they must provide far more evidence on the uncertainties and robustness of their results.

We do include this information in the manuscript, specifically Page 8 Lines 1-2: "The STILT model is configured in backward time mode, with the particle release point set as the Miyun sample inlet height of 158m above sea level (masl), corresponding to 6m above ground level (magl)." The station was strategically located such that its height was greater than the immediate surroundings. We have included this earlier on as well, and include a more detailed discussion of the implications of the receptor height. In particular, we have also now referenced the Karion et al. (2016) study where a 30m apparent inlet height was compensated by the height of the receptor relative to its surroundings.

In addition, we do not claim to estimate *total* national Chinese fossil fuel emissions at all. To do so with a single site would indeed, as you say, be bold. In fact, we emphasize that the regional analysis requires more stations for more robust conclusions to be made. The title of the paper, the abstract, the introduction, and conclusion all note that the evaluation is for northern China. We only mention *all* of China in the context of implications of the study and improvements required for more comprehensive future studies. For instance, in the abstract Page 2 L3-5 we state: "On average, over the study time period, the China-specific inventory has substantially larger (20%) emissions for all China than the global inventories." Furthermore, within the abstract itself we state the value of the study but also the obvious caveat of a single station (P2 L5-9): "Our analysis uses observations to support and justify increased development of China-specific inventories in tracking China's progress towards reducing emissions. Here we are restricted to a single measurement site; effectively optimizing inventories at relevant spatial scales requires multiple high temporal resolution observations. We emphasize the need for a denser observational network in future efforts to measure and verify CO_2 emissions for China both regionally and as a whole." We have re-arranged the text and re-worded to avoid misunderstandings about this statement, as we see how it could have caused confusion.

The error analysis is described in the supplementary information and involves treating all errors as embedded within the model-measurement residuals; as you point out, the single station prevents a more sophisticated error analysis. We discuss this issue further in our responses to Referee #2, RC 1 who raised the same concern (please see General Comments #3 in RC1). In addition, we include a brief comment about this in our revision in the inlet height section below.

<u>Our manuscript revisions based on the comments above are as follows:</u>

P8L2: The topography in the vicinity of the station justified the low apparent inlet height, as the station height was greater relative to the immediate surroundings. The topographic advantage of a ground station relative to immediate surroundings was exploited in a similar manner in Karion et al. (2016) in the context of an Alaskan CO_2 study; however Karion et al. (2016) were able to use a suite of additional data to confirm the validity of their assumption including comparisons to concurrent aircraft measurements and multiple inlets at 31.7magl, 17.1magl, and 4.9magl. We caveat our study with the recognition that concurrent aircraft measurements (for example) and multi-level inlet locations were not available and therefore prevents a more thorough analysis on impact of absolute and relative inlet location on transport uncertainty.

P2L3-9: Our analysis uses observations to support and justify increased development of China-specific inventories in tracking China's progress as a whole towards reducing emissions. Here we are restricted to a single measurement site; effectively optimizing inventories at relevant spatial scales requires multiple stations of high-temporal resolution observations. We note that averaged over the study time period, the unoptimized China-specific inventory has substantially larger (20%) annual emissions for all China than the unoptimized global inventories. Exploring this discrepancy for China as a whole requires a denser observational network in future efforts to measure and verify CO_2 emissions for China both regionally and as a whole, and this study provides a baseline analysis as well a guide for determining optimal locations for future ground-based measurement sites.

Detailed Comments:

1. **I see little value in including Supp Figures: S6, S12, S13, S14, S15**

   Thank you for pointing this out, but we included them for informational purposes in the event it could be useful. There had been some interest in seeing these figures, so they are left there in the event that others might find them useful references.

2. **Optimization method: You can succinctly say that you are "scaling" the whole pattern of fossil up or down based on the mean difference between modeled CO_2 and observed CO_2, instead of pushing readers to the supplemental information. Please just summarize this in Sect 3.6. Unless I'm misunderstanding the technique.**

   To clarify, at the seasonal level, the technique is more than a simple scaling factor between predicted and observed concentrations. It involves translating the mole fraction difference to a mass of carbon dioxide normalized by optimization region area. We explain the technique in the supplementary information to avoid unnecessarily lengthy explanations in the main text. We think the details are important enough to warrant inclusion in the SI at the very least, should readers be interested. We agree that a succinct explanation is sufficient for the main text.

   <u>We will reword the text to be clearer based on your recommendation:</u>

   P13L9-11: "Instead, we use the ratio between observed and modeled ΔCO_2 to scale the anthropogenic inventories annually. At seasonal timescales, we use the difference between observed and modeled ΔCO_2 normalized by footprint area to obtain a mass flux correction to the combined vegetation and anthropogenic inventories. For further details of the optimization technique, please refer to SI Sect. S8."

3. **Background Inflow (Fig S17): It concerns me that your comparisons of CT2016 bgd from the south (LLN), "urban impact" appear to show a couple ppm bias (model > obs). The difference between fossil inventories, as seen in the site CO_2, have almost the same pattern (or lack of) as the bgd inflow. These would seem difficult to distinguish.**

   We have revised the SI to be clearer about the information provided by LLN:

   PS13: "As shown in Fig. S17, LLN shows CO_2 depletion relative to CT2015 suggesting that for this analysis it is not representative as a background site. (CT is not responsive to all sites). The LLN observed CO_2 drawdown compared to modeled CT2015 suggests that LLN sees a lot of surface influence on account of its location in the middle of an island in vegetated surroundings. Moreover, LLN is not an important sector for the influence region of this study; we include it primarily for reference for future studies considering regions of China that would be more sensitive to the sector associated with LLN."

4. **Supp Sect 8: "At annual scales, the dominant contributor to the CO2 signal are anthropogenic emissions; optimization at annual scales is therefore applied only to the anthropogenic emissions inventories." I'm confused here. At annual scales the dominant contributors to CO2 are *fossil* as stated, but other significant contributors are biological/ocean carbon sink as well as any interannual variability. Please revise this statement.**

   We have revised the statement to emphasize the regional profile:

   Now PS15: "At annual scales, the dominant contributor to the regional CO_2 signal are anthropogenic emissions; optimization at annual scales is therefore applied only to the anthropogenic emissions inventories. The other significant contributors include longer term biological and ocean carbon sinks and interannual variability within these components, but for this study region, these components are embedded in the background concentrations. In particular, Dayalu et al. (2018) note that 13% of the northern China ecosystems and 20% of northeastern China's ecosystems are mixed forests. However, Dayalu et al. (2018) also demonstrate that the ecosystems with greatest influence on this single site are croplands with high intra-annual carbon turnover rates."

5. **Supp Sect 8: "The heavily cropped L_0.90 influence region implies rapid turnaround of vegetation carbon stocks at the annual scale, justifying this assumption (18)." First of all, saying that the L_0.90 is "all crops" seems like quite a stretch in order to justify quick turnover of C. Do you have information to support this? Even if it was 100% crops, the nature of crop-C is a redistribution laterally and seasonally, which is likely very similar from year to year, could easily induce a "low bias" in the croplands (where the tower is), and a high bias in the cities and livestock areas, where the C is utilized, as well as seasonal effects. Again, these biases should preserve the relative differences between the inventories but affect the overall accuracy of them taken together.**

   The main administrative regions in the L_0.90 contour are northern China, northeastern China, and Inner Mongolia. In Table 4 of Dayalu et al. (2018), it is shown that the ecosystems of these regions consist of primarily of croplands and/or grasslands (over 50%). Northern china in particular consists of 51% croplands, 23% grasslands, and 13% mixed forests. Land use types over the span of the 5-year study period was found to be steady. We recognize that it is not "all crops" and we do not state that it is "all crops". But as we do state, it is indeed "heavily cropped" and is the dominant portion of the biogenic signal. This is also supported by Fig. S11a, where the annual influence contours are overlayed on the IGBP land use map (MODIS data), showing a dominant grassland/cropland influence. Mixed forests (green) and urban (red) are certainly more minor. Given this dominance of croplands/grasslands to the modeled Miyun signal, we make the same reasonable assumption that Piao et al. (2009) make (as we have already cited in the main text and SI) in that annual C balance in agricultural regions is zero. The Miyun CO_2 signal is certainly affected by other biological/oceanic/interannual variability; but as also explained in comment #4 above, these are not significant parts of the *regional* signal (i.e., minus background). These are longer term features embedded in the background concentrations. See response to comment 4, above.

<u>We have expanded upon this to make the concept clearer:</u>
Now PS15: "The annual influence contours are overlayed on the IGBP land use map in Fig. S11a, and shows the dominant grassland/cropland influence on the modeled Miyun signal at annual scales. As stated previously, The Miyun CO_2 signal is certainly affected by other biological/oceanic/interannual variability; but these are not demonstrated to be significant parts of the regional (ΔCO_2) signal. These are longer term features embedded in the background concentrations."

6. **FigS18: The bio signal contribution looks very odd. Max drawdown in PBL in middle of night? Or is there a strong diurnal component to fossil from local sources? This is not what one normally considers a diurnal time series of biological CO2 to look like in the summer. The "observations" show the classic local drawdown impact on CO2 in spring/summer w/ flatter diurnal cycle in periods of low GPP, but the models don't seem to reproduce this at all. I'm very concerned that uncertainty in this biological component, whose mean diurnal cycle is much larger than the spread of fossil estimates, could swamp the contrast between the fossil inventories. Furthermore, this figure doesn't seem to match Fig S19, which seems to show the standard respiration buildup (presumably at night).**

   We include this figure to highlight the importance of using the "well-mixedness" time window for the data subset used in this study. That apparent "drawdown" during those early hours (which are not used in the analysis) is actually an artifact of the regional upwind net uptake signal during the growing season. At nighttime, the receptor can be modeled to be above the nocturnal boundary layer; as the WRF-STILT model is unable to capture nighttime mixing and transport, the result in the modeled time series is a residual layer with an upwind uptake signal. In addition, the modeled receptor does not see anthropogenic emissions from the local surface during these hours as it is above the nocturnal boundary layer which further contributes to the apparent (not real) drop in modeled CO_2 concentrations. As noted in Page 8 of the SI, "When the receptor release occurs outside of peak daylight hours, stratification of the PBL becomes significant. Therefore, as is common practice in virtually all emissions optimization studies, we model the 1100 to 1600 (local time) subset. These daylight hours represent a typical window for which STILT reliably models transport…"

As described in the caption of Fig. S19, the data points included are from the afternoon "well-mixed" hours data subset, and, as also noted in the main text (P7 L7 and P9 L1-L2) is the subset used for all the analyses in this study. There is actually no standard respiration build-up in this figure as it is the daytime hours only; the figure reflects the larger seasonal cycle (January 2005 through December 2009, 11am to 4pm LT values only).

We have edited the main caption of Fig S19 to make it clearer. The usage of "timeseries" was inaccurate, as it is a disjointed subset of mid-day points:

Now PS32: "Observed and modeled CO_2 and ΔCO_2 (1100-1600 LT hourly values only) over entire study time period"

7. **Section S2: This section would benefit from 1) Summary stats for the wind speed differences 2) An analysis across more than one WRF pixel. For example, you should be able to pull the matching WRF pixel for a dozen or so CMA stations. It's hard to tell whether this cell was just cherry-picked as it isn't the tower site. 3) There is no wind direction information, which in my mind would be much more useful than most of the met data analysis in this section. Windroses by season for example. Mostly we're interested in differences in wind speed, PBL height, and direction across the WRF domain.**

i) Agreed, we have now included this in the section and it encompasses stations in d02 and d03 for 2006. We would like to emphasize that this validation is against 24-hourly averaged data (the only publicly available CMA observations at the time) which results in an underestimated assessment of WRF windspeed bias. We have noted this accordingly in the section.
ii)Yes. This analysis was indeed done for all CMA stations in the study domain (including d01 and d02). There are hundreds of stations, so including all the graphics them all was not feasible. However, we have now uploaded the files to the data portal containing the supporting data sets.
iii) We agree. It would be valuable to include such a wind direction analysis if this data were readily accessible. Unfortunately public access to meteorological observations is limited for this region, and the analysis presented here was conducted on what was available including being restricted to 24-hour averages which in itself limits the power of the model evaluation.

We have made the following revisions:

PS6: Comparing against publicly available 2006 CMA data from 35 stations across the d02 and d03 domains (Fig S2), the median modeled wind speed was 15% higher than observations, with a median absolute deviation of 16%. We emphasize that a more robust evaluation of WRF windspeed (or other meteorological) biases relative to observations would require access to higher temporal resolution meteorological observations. Currently, we are restricted by data availability to 24-hour averages which blur smaller timescale processes and therefore likely underestimates the WRF surface wind speed bias relative to observations. We do not include d01 comparisons in this analysis, as the distance between nearest station and WRF gridcell center can be on the order of tens of kilometers, decreasing the information and value of the comparison. The graphics associated with the d02 and d03 comparisons are available from https://dx.doi.org/10.7910/DVN/OJESO0 as "006_WRFvCMAplots_2006_d0X.pdf".

8. **L25, Pg28: " .. so this conclusion applies strictly to the other three seasons." What does this mean? Please clarify this statement.**

   We agree that this wording is confusing. We have re-arranged and re-worded accordingly:

   P28 L23: "We find the ZHAO+VPRM inventory agrees very closely with observations, much better than the nationally referenced inventories at all timescales and seasons, with the exception of the peak growing season. During the peak growing season, the regional enhancement to background CO_2 concentrations is modeled as approximately zero, due to an agriculturally dominated vegetation signal that is equal in magnitude and opposite in sign to the anthropogenic signal (Dayalu et al., 2018). While this agrees with previous work by Turnbull et al. (2011), in both that study and the present study the sparse data prevents a more conclusive statement about anthropogenic inventory performance during the regional growing season.

9. **Conclusions: There is no discussion of limitations or caveats anywhere that I could see. The obvious ones include what appear to be a (locally?) poorly modeled biological flux term, biased low respiration (in general) from VPRM, and a fossil "signal" whose pattern may be difficult to distinguish from errors in background inflow. Furthermore, there is no characterization of transport error. This shouldn't always be required but the degree to which the authors are trying to extract information from a single 6m tower seems like it should require caveats as well as some quantitative characterization of uncertainty.**

   We do specifically caveat our study in multiple locations in the text with the fact that we do not have access to more than a single site and are therefore restricting our analysis to a specific region (northern China). In addition to the caveats included as part of the revision detailed in the points above, and in RC#1, the text from the original discussions manuscript detailing caveats is listed below. We feel this is sufficient; unless the editor and referees would like to direct us about what else we should be saying, and if so where.

   i. Abstract. P2 L6-9: "Here we are restricted to a single measurement site; effectively optimizing inventories at relevant spatial scales requires multiple high temporal resolution observations. We emphasize the need for a denser observational network in future efforts to measure and verify CO2 emissions for China both regionally and as a whole."
   ii. Introduction. P5 L14-21: "Despite being restricted to a single measurement station, our site provides valuable information and constraints on emissions inventories because it receives air at different times from one of the heaviest emitting regions of China, and

clean air at other times. … we conduct a basic benchmark optimization of the inventories for the 2005-2009 measurement time period…"

iii. Conclusion.

P29L4: "Since the ZHAO methodology gives comparatively accurate and higher results for the influence region dominated by Northern China, we hypothesize that the proxies used in the global inventories have biases that likely result also in overestimation in other regions of China. However, observational data from strategically located stations in and around the eastern half of China are required to explore this hypothesis."

P29 L16-21: "The single station available for the 2005-2009 period was strategically located to provide information on one of the highest CO2 emitting regions of China. Within that limitation, the observations provide strong evidence supporting the use of China-specific methods, such as those employed in ZHAO, for China's CO2 emissions inventory derivation. A denser network of CO2 measurement stations in China is required as a basis for effective monitoring, reporting, and verification of regional and national inventories."

In terms of quantitative characterization of uncertainty, this was brought up in your general comment #2 above and we have responded there and also to RC#1.