We thank the two reviewers for the detailed and helpful comments which have strengthened the paper. Individual responses to each of the points raised are provided below.

Response to comments from Anonymous Referee #1 (RC1)

1. I see two potential pathways to go ahead with the publication of this work: 1) Re-frame the writing/aim of the work and highlight more and focus more on what you did and what information you can get from the available data (with high certainty) and not what you weren't able to do 2) Extend the work/analysis by including additional methods to evaluate and constrain regional CO2 emissions and additional inventories that potentially became available since the start of this research.
   - We agree that the focus should be more on spotlighting the key results (pathway 1) and our revisions have focused on this. Extension to additional emissions inventories (pathway 2) without additional observations is simply throwing more modeled quantities at sparse observations (this work is observation-limited, not inventory-limited). In this context, a selection of three inventories is sufficient to make our main points which are (i) the power of top-down constraints on emissions assessments in the region, motivating new ground-based observation sites; (ii) there are substantial differences in categories of inventories for China, and these broad categories (China-specific ones that are tuned to actual but difficult to obtain field data, or global subsets based on coarser but more readily available inputs) produce significantly different observational mismatch, highlighting the importance of China-specific field data; (iii) our work suggests that current emissions tracking at the international level based on archaic inventories such as CDIAC can be very different from emissions reality; more ground-based observations are needed (along with remotely sensed observations, which now have a robust timeseries) to test this hypothesis. When sufficiently dense observations are available, a truly comprehensive analysis of all available and relevant inventories can be conducted. Such an analysis is different from the main point of our study.

2. Lots of focus on justifications/limitations early on in the Abstract. I suggest shortening the Abstract, more specifically the part where the authors discuss that they only have one site. Try to re-frame it so that you highlight and emphasize what you have and what you did, and not what you don't have and can't do. Moreover, the limitation of only 1 site is discussed too many times in the text, no need to repeat it over and over again.
   - Agreed. Combining this with (1) above, we have gone through the text and deleted redundant caveats and spotlighted areas of greater certainty so they are not lost in the "caveat weeds". We also recognize there is a certain amount of subjectivity surrounding how much is too much for caveats, as one of the main concerns of previous reviewers was that there were not enough.

3. I would also suggest to merge the caveats section with the conclusions. I understand that including the caveats of this study is quite important to justify the results/work/effort; however, currently the focus on the caveats puts the results in the second plan.
   - Agreed. The stand-alone caveats have now been ingested into the Conclusions. Along with (1) and (2) above, we have pared down the amount of in-line caveats as well.

4. The whole text could be condensed. There are a number of places where the authors repeat the same thing. An example: Introduction lines ~115 where describing the measurements is the same as the beginning of the CO2 observations section. No need to have all the details on both places.
   - Agreed. We have fixed this.

5. The writing needs modification and improvement, hence the paper needs a careful reading/checking. There are a number of sentences (or parts of sentences) that are hard to follow and requires few re-reading in order for the reader to understand the message. I suggest the authors to re-write and clarify the sentences under the Specific Comments.
   - See responses to Specific Comments.

6. Rephrasing/restructuring. Line 22: Comparison of CO2 observations to CO2 predicted from accounting for global background concentration and atmospheric mixing of emissions suggests potential biases in the inventories
   - Fixed. We have made significant changes to the abstract wording to both spotlight our key results and also improve the structure.

7. Rephrasing/restructuring. Line 39: Additionally, we note that averaged over the study time period, the unscaled China-specific inventory has substantially larger annual emissions for China as a whole (20% higher) and the northern China evaluation region (30%) than the unscaled global inventories.
   - Fixed. We have made significant changes to the abstract wording to both spotlight our key results and also improve the structure.

8. Rephrasing/restructuring. Line 42: lend support the rates
   - Fixed. We have made significant changes to the abstract wording to both spotlight our key results and also improve the structure.

9. Rephrasing/restructuring. Line 180: Winter wheat emergence in the spring and corn emergence in later summer shift the seasonal patterns such that regional seasons are more appropriately represented when months of year are grouped as January, February, March (JFM/Winter); April, May, June (AMJ/Spring); July, August, September (JAS/Summer); and October, November, December (OND/Fall), respectively.
   - Fixed.

10. Rephrasing/restructuring. Line 285: This is not intended as an exhaustive sampling of inventory approaches; however, it is sufficient to demonstrate the utility of continuous high-accuracy observations as a top-down constraint for evaluating emissions estimates.
    - Fixed.

11. Line 468: As noted in Sect. 3, the regional growing season does not have a typical pattern in that peak uptake occurs around July/August with the onset of the corn growing season.
    - Fixed.

12. Rephrasing. Line 45: import
    - Fixed.

13. Rephrasing. Line 80: exhaustive
    - We use the word exhaustive in response to previous reviewers who were concerned that we were claiming the three inventories were the only ones that existed for China. We wanted to assure readers that we recognize that these are the only inventories that exist for China. We feel this word should be left in for that reason.

14. Rephrasing. Line 95: judge
    - We use this word in response to previous reviewers who felt we were actually judging the merits of the inventories individually. That has not been our intent. We feel this word should be left in for that reason.

15. Rephrasing. Line 110: while the others do not
    - Fixed.

16. More details needed. Abstract Line 21: “CO2 inventories” – list which ones.
    - We are leaving the named inventories out of the abstract in response to previous reviewers who felt that calling them out specifically at that stage implied criticizing the particular inventories rather than a generalized examination of the approaches they represented. We feel the wording should remain this way until the reader has the deeper context from the text.

17. More details needed. Line 285: ”This is not intended” - this as what, the study? Clarify.
    - Fixed – “Our study is not”.

18. More details needed. Line 320: “Applying the weekly and diurnal Nassar et al. (2013) scaling factors did not generate differences that were statistically significant, suggesting that a more rigorous set of temporal scaling factors need to be developed for China. “ Is this based on work from the authors or Nassar or? Clarify.
    - It was based on our work. We have clarified this.

19. Line 145: "it is not possible to evaluate any error in spatial allocation of emissions. However, we note that the same transport model is applied to all the emission fields. Unresolved transport error undoubtedly contributes to scatter in the model-data comparison but is unlikely to generate consistent biases among the inventories." - could you please explain this better.
    - We have reworded and expanded this section. We have also restructured the paragraph to make the point clearer.

20. Line 170: “Average annual data coverage” – was this calculated based on hourly, daily data? Just add some brief details how was it quantified.
    - Fixed. (Calculated based on hourly data).

21. Line 189: “filtered to include only non-missing observations” – a little bit unclear, does this means that only days are used when we have measurement for each hour between 11 and 16?
    - We have clarified this. The subset is done for each individual hour, not the daily blocks. For example, if we were missing 1100h but had 1200-1600 for a particular day, 1200-1600 would be used but 1100 would not (because we would not be able to compare our modeled quantities for that hour to any observation). But we would be able to compare modeled to obs for 1200-1600h on that same day.

22. Line 190: “background criteria” – if possible, briefly mention what it is in the main text also or additionally refer to section 3.5.
    - Fixed. Referring to section 3.5, where we also included a very brief discussion of what that criteria is (previously mentioned only in the SI).

23. Methods section – when describing why the 11-16 measurements are used, please add a discussion of why the authors didn’t use night time data and how much this affects the results. Although this is briefly mentioned at the end of section 3.3. it would be good to extended it in the Methods section also.
    - We do explain this in that paragraph (~L190) but we have improved the wording to make it more clear as to why we were not using nighttime data.

24. Figure 2: It would be good to add another sentence on what the different percentile regions represent/describe. This could be added around Line 235.
    - We have considerably re-worded our presentation of percentile regions in response to concerns from the second reviewer.

25. Line 259: “which has been noted previously as major uncertainty in Chinese emission inventories” – add reference.
    - Reference added.

26. Feel free to remove the word respectively from everywhere in the text. It is already automatically assumed that the order is respectively.
    - The comment is appreciated, but I would feel more comfortable keeping “respectively” in formal writing as is customary when multiple variables are being described. Even if the order is understood by most people, it eliminates confusion more than it distracts.

27. Line 160 rephrase ‘made’ –> “measured’
    - Fixed.

28. Table 1. – define what the abbreviations are (if used for the first time in the text). And just to clarify, these are the 2005-2009 averages? Add in the caption.
    - Fixed.

29. Line 820: (in press), 2018 – still in press?
    - No longer in press. Fixed.

Response to comments from Anonymous Referee #2 (RC2)

1. The title implies that actual flux estimates for Northern China are the central point of this paper. However, this paper is very technical and focuses much more on the comparison of existing inventories with atmospheric observations and also the regional emission intensity. The title should be updated to better reflect this core content.
   - Agreed. We have changed the title accordingly. It is now "Evaluating China's anthropogenic CO_2 emissions inventories: a northern China case-study using continuous surface observations from 2005-2009."

2. The authors present the L_90 footprint, which usually reflects the theoretical sensitivity of the observations to a unit of flux. Two issues arise with this. First and foremost, the actual area influencing the receptor observations is not reflected by this, as CO2 sources span multiple orders of magnitude. Therefore, a major source e.g. coal-fired power plants just outside the L_90 footprint will have more influence on Miyun CO2 mole fractions than a deserted patch of land without CO2 flux inside the L_90 footprint (e.g. in inner Mongolia). The second problem is that the authors use a plethora of different terms and all seem to refer to nearly(?) the same thing: "L_90 footprint", "influence region", "L_90 region", "L_90 evaluation region", "90th percentile of multiyear mean annual STILT footprint influences", "surface influence maps".
   - We recognize there has been a misunderstanding about the role of the L_0.90 region in our study. We apologize for the confusion caused by the way this concept was explained and introduced, especially with the effect of inconsistent references (we now consistently use L_0.90 region). We have explained the role of the L_0.90 region and the methodology for calculation more thoroughly in the text. In particular, we have clearly stated that at any given time the *entire* STILT footprint is convolved with the flux estimates; the L_0.90 region is simply the region we chose to ascribe the model-observation mismatch. The area of the region informs the conversion of ppm mismatch to mass units and we compare this regionally scaled mass correction to the mass originally estimated by the fluxes in that L_0.90 bounding area. We could have just as easily made our bounding area for evaluation of mismatch be L_0.75 or L_0.99 but we explain in the text why the L_0.90 region is a good balance of capturing enough of the surface sensitivity with not having an unrealistically diffuse spatial area. With this clarification in mind, your concern about neglecting influential sources is alleviated: we *are* taking such sources into account in the main footprint*flux = ppm setup; we just ascribe that ppm to the L_0.90 region in the end to obtain a mass correction over a reasonably influential area. The data set does not really allow us to geographically allocate the mismatch beyond this relatively coarse method.

3. A further limitation is that only one biosphere model is used. The authors seem to ignore this limitation, while nicely highlighting that having 3 anthropogenic prior is very helpful to better understand general results of modelled atmospheric CO2. The fact that biospheric fluxes might be even more uncertain than anthropogenic CO2 fluxes seems to unrecognized. A straightforward analysis to investigate the relevance of natural versus anthropogenic fluxes would be to investigate if the biggest model-observation mismatches systematically occurr during times of high contributions of anthropogenic or natural fluxes to the modeled CO2.

- We have modified the text accordingly. We used only one biosphere model to simplify our assessment of the variations across different anthropogenic emissions inventories (and we have changed the paper title accordingly). Our companion paper (Dayalu et al., 2018) highlights differences across vegetation models when controlling for anthropogenic emissions. We did not intend to imply that only the anthropogenic inventories are needed to understand China's atmospheric CO2 and we have reworded the text to make this clearer. In addition, we make more references to the Dayalu et al. (2018) VPRM-CHINA paper (in agreement with the Turnbull et al. (2011) paper) that specifically highlights that in the heavily agricultural region of the North China Plain, the *peak* growing season sink actually is comparable in magnitude to the anthropogenic source. Outside of this observation, we do state that the Summer-time analysis of anthropogenic vs biogenic cannot really be undertaken absent additional and diverse data sets. Table 1 (Mean Bias and RMSE segment) does highlight the model-observation mismatch by season, as you had suggested. In the winter in northern China, the anthropogenic signal is the dominant signal, swamping the NEE terms. The systematic bias among the three simulations in this season (Table 1) is largely attributable to differences in the anthropogenic emissions alone (all other modeled components being identical among the three modeled CO2 quantities). Based on the results presented in Table 1 (with the confounding exception of the Summer for reasons we just described) we see that during seasons where biological activity is lower or significantly lower than anthropogenic activity, there is a consistent discrepancy between the CO2 modeled by the three different anthropogenic inventories suggesting a systematic difference in the anthropogenic component. In the fall, all three modeled quantities are consistently lower than observations most likely resulting from the known underestimate of ecosystem respiration which is the dominant biological process at this season (Dayalu et al., 2018); but even so China's significant anthropogenic component still dominates at this time. If we assume that the winter represents the "purely anthropogenic" baseline, and we assume a certain percentage impact of temporal activity factors (1.5-8ppm as suggested by Nassar et al. 2013) we could make an estimate as to how much this baseline is expected to shift over the course of seasons – but that would be overextending our analysis as we know very little about temporal activity factors in China.

4. Furthermore, multiple papers that address anthropogenic CO2 emissions in China and specifically the Beijing region are ignored, e.g. the PKU-CO2 inventory (see comment line 35f) or the isotope studies by Niu et al. 2016 (see comment line 364f). A comparison to their results would be an important addition to this study. Lastly, one important result of this study is the trend in regional carbon intensity. Unfortunately, the calculation of GRP and its trend as well as their uncertainty is not clear enough. To really assess the importance of reducing the uncertainty in CO2 emission estimates by using atmospheric observations strongly depends on how well GRP and GRP trends can be calculated and also scaled to GDP and GDP trends. The explanation, data sources and methodologies for the GRP calculation should be expanded.
   - Re: inventories. See responses to 5 and 20 below. Re: GRP calculations: we have expanded the text and descriptions of the calculations.

5. Line 35f: When did this research begin? The PKU-CO2 emissions inventory which is China-specific was published in 2013, but is not considered or even mentioned in this manuscript (Wang et al. 2013; doi:10.5194/acp-13-5189-2013, available through e.g. http://inventory.pku.edu.cn/download/download.html)
   - We include this paper in the references now (and cite it as an additional justification for our not explicitly separating ODIAC and CDIAC for China – a major concern of previous reviewers). That aside, the PKU-CO2 emissions inventory referred to in the Wang et al., 2013 paper was a *global* emissions inventory solely for the year 2007. We selected the Zhao inventory due to the fact that at the time the study was conducted, it was the only readily available China-specific inventory that spanned our observational data set (2005-2009). Furthermore, the paper does specify use of the CARMA power plant inventory for emissions factors; for reasons described in the text, the global emissions factors provided by CARMA for China are known to be problematic. PKU and MEIC have since been leaders in developing China-specific inventories, but unfortunately the study concluded before those were readily available and ingestible into our analysis framework. We agree that future studies in China would benefit greatly from more China-specific inventories being evaluated, and look forward to results from such analysis when longer timeseries of observational data become available.

6. Line 56f: The author's should expand more on the nature of the differences of different inventories. Atmospheric measurements will only be able to consider scope 1 emissions and do always include all sources, while national inventory reporting and provincial reporting might use different methodologies and also different emission category definitions and reporting thresholds. Therefore, it is unclear if the mere fact that there is a discrepancy between provincial and national estimates really means that there is a difference that an atmospheric approach could detect, help to decrease.
   - Agreed -- the observations give an overall constraint on total fluxes and cannot resolve the particulars of estimate methodology. We are looking at whether totals are consistent with observations, but we can't really diagnose the finer scale methodology (we can't attribute the error to any particular source). That being said, if there is a significant difference between the totals reported by provincial vs national inventory estimates then the discrepancy can be suggestive of differences in methodology. With our approach we can assess whether total fluxes over a region are consistent with observations, but with only one species we can't really diagnose which emission source types are too high or too low. One inventory is closest to matching the observations than the other, but we can't say what feature makes this so. The most we can do is highlight the major differences among the inventory methodologies (as we have done in Section 3.3); we don't know which of these changes accounts for the better model-observation match. We have made it clearer in the introduction to ensure we are not suggesting this. "The primary intent of the comparisons presented here is not to judge specific inventories, but to demonstrate that even a single site with a long record of high time resolution observations can identify the potential impact of major differences among inventories that manifest as biases in the model-data comparison."

7. Line 79: see comment line 35f

- See response #5.

8. Line 118: A key element that needs further explaination is the notion of "surface influence map". This seems to be used to describe the footprint, i.e. the sensitivity of the observations to a unit of flux (emission) from a given area. However, in line 645 the "L_90 footprint" is apparently something separate from the "influence region". See general comments.
   - Yes – we have fixed this in the text and included a new set of figures in the SI to further explain the footprint map/notion of surface influences.

9. L130: Figure 10 from Dayalu et al. 2018 does indeed show that natural and anthropogenic fluxes are the same order of magnitude in the growing season. But given the very significant variability (1-sigma is near 100%) it seems unclear why the authors assume that this is only in the peak growing season and not also in other months e.g. May 2006 seems to have high uncertainties in relative importance of natural versus anthropogenic CO2 fluxes.
   - Uncertainty in the modeled biosphere is undoubtedly significant, but contributes equally to the variability across all the modeled-observation quantities. We have made this clearer in the uncertainty discussion as well. See response to #20.

10. L146f: Please elaborate why transport errors (which can be systematic in nature) could not cause a bias when comparing inventories with distinctly different spatial distributions.
    - We have amended the line to include that it may be important in that it could attribute errors in transport to biases in inventories. Although the interaction of transport error with differences in spatial distribution could bias individual observations, averaging over longer timescales (seasons, years) minimizes the bias of individual points. We have made this clearer in the text.

11. L160f: Wang et al. 2010 provide information on instrumental precision and that a calibration strategy was in place to monitor long-term drifts. This seems like an important addition here.
    - Fixed. We also note the citation for Wang et al., 2010 at the end of the section directing the readers to that paper for details on the instrument precision, calibration strategy, etc.

12. Line201f: Given that a short tower is used for observations it seems useful to know what the height of the lower/lowest WRF levels used are. 41 vertical levels are mentioned, but without additional information this seems difficult to interpret.
    - We use the default WRF eta levels generated with the 41-level specification and we would expect about 20 vertical levels in the first 1500m. Our first vertical level would be roughly around 8m (using Arasa et al., 2016 as a guide: https://www.scirp.org/pdf/ACS_2016042911473822.pdf). We also note this is the other reason for restricting our analysis to middle of day -- excluding times when the vertical gradients are sharp.

13. Line 234f and Figure 2: It seems important to expand on how your "L_90 region" was calculated. It is referred "90% of the surface influencing measurements". So this would mean that it is NOT the footprint or the 90th percentile of the surface sensitivity. The surface sensitivity/footprint reflects how a unit of flux will alter the observed mole fraction. However, even regions with very low sensitivity can still have a noticeable influence on the

observed concentrations. Figure S8a clearly shows that some regions within the "90th percentile (Northern part of China) have emission rates that are at least 3 orders of magnitude lower than areas just South of the 90th percentile footprint (Nanjing-Shanghai region). It seems very likely that atmospheric CO2 mole fractions at Miyun would be more affected by these Southern Emissions than from some remote Northern regions that. A true influence/contribution map could be calculated very quickly with the existing data.

- Agreed -- the concept was poorly explained which has led to the misunderstanding. We have now explained our methodology better and included footprint maps and percentile selection illustration in the SI. See detailed comments to your previous point (#2.)

14. Line 238f: Are really 40% influence/contribution coming from outside of L_50 or rather 40% of footprint sensitivity lies outside L_50?
    - Agreed – again, poorly worded. This has been fixed. We also modified to state L_0.75, as this seemed more interesting a comparison.

15. Line 256f: The justification of the interpolation seems to rely on the fact that only a few regions show large differences. However, a more straightforward method would be to convolute the 2005 footprints with 2005 emissions and then with 2009 emissions (using the same 2005 footprints). This way we can directly assess if the flux changes are theoretically noticeable in the atmospheric record used later or if this just adds "random noise" to the observations.
    - While the test proposed would be interesting, evaluating errors in spatial allocation is beyond the scope of this study and the available data set (L143). It would still be relying on a single site to optimize a spatial distribution of emissions and would not be a conclusive test; our approach was the simplest method using the information provided by the raw anthropogenic inventory alone.

16. Line 274/275: citations needed
    - Fixed.

17. Line 287f: see comment line 35f
    - See response 5.

18. Line 332f: Given that only one simple biosphere model is used in this study a discussion of its performance and uncertainty would be very useful here. How well does VPRM-China compare to the local/regional flux towers sites within the L_90 footprint?
    - We direct readers to the companion paper by Dayalu et al (2018) which shows these figures (E.g., Figure 5). Eddy flux sites in the region are sparse, and most of the available data was enough to be used only as calibration. Only two sites had enough data to be used for validation.

19. Line 351 – eq 1? The equation seems to imply that CT2015 was used for all years – maybe add clarification that CT fluxes for the appropriate years was used and not a climatology based on CT2015. $CO2(t) = CO2,obs(t)-CO2,CT2015(t-7d)$ Also, the equation is not numbered/labelled.

- The equation is now labeled. CT2015 is the CarbonTracker version number (not the year of the data). CT2015 was used for all years (ie. that version of CarbonTracker), and we used atmospheric mixing ratios not fluxes. Setting (t-7d) in the subscript implies that was always used – however, as we detail in the supplementary information (and now in the text) background concentrations were selected when the particle reached the domain edges which may or may not be as far back as 7days (7 days being the backward limit).

20. Line 364f and S5: The authors suggest that the anthropogenic fluxes dominate the annual total in the main text and then go even further in the supplement and suggest that natural fluxes are negligible. Quote from S5: ". . . correction at annual scales is therefore applied only to the anthropogenic emissions inventories" This a very strong assumption and seems to require further explanation. During the growing season fluxes seem comparable and VPRM underestimates respiration fluxes in the non-growing season (see line 504). In the absence of other biosphere models in this study to cement this notion it seems necessary to refer to other studies in China to rationalize this. For example, Niu et al. 2016 [https://pubs.acs.org/doi/abs/10.1021/acs.est.5b02591] found that even in Beijing (Haidan district) CO2 from fossil burning only contributes 75% to the annual average CO2 offset. So, it seems unlikely the natural contribution to the CO2 mole fractions at Miyun can be ignored, even at annual average scale.

- The data from the two sites in Niu et al. study is only for one year (2014), and the contribution of fossil fuels to the Beijing site displays considerable variance (75% +/- 15%); nevertheless we have incorporated these important results in our paper to caveat our statements about annual emissions but we also note that the timing relative to our study (2014 vs 2005-2009) and the variance (60% - 90% contributed by fossil fuels) annually. In line 364 we don't say the biospheric impact if zero; we say the anthropogenic signal dominates (which is true, also according to Niu et al's results). More biospheric models are needed to quantify the regional biospheric impacts and we note this in the conclusions. This is an area of significant uncertainty, as evidenced in Piao et al. (2009) which we also cite at line 478 ("For annual budgeting we follow the assumptions of Piao et al. (2009) and Jiang et al. (2016) that agricultural systems are in annual carbon balance because crop biomass has a short residence time.")

  The quantity relevant to this question is the annual *net* biospheric carbon flux: and annual net carbon balance in this region is highly uncertain with an uncertainty in both magnitude and sign (ie, spanning zero) both for process-based models and inversions (Piao et al. 2009). Piao et al examine this quantity by region of China, noting that the uncertainty is very large (his regional inversions are based on 9 sites across all of Asia). Process-based models and prior models corresponding to our northern China study region (Figure 2, Piao et al) assume either small net emissions or zero (ie. zero in the agriculturally dominated north china plain). To the extent that there is a net biosphere source/sink at the annual scale, it should be included but is currently highly uncertain. Our assumption of dominant anthropogenic influence in northern china is in keeping with the priors (e.g. from Piao et al.) that assume zero and are not significantly corrected by the poorly constrained inversions. We have summarized this discussion in the text.

21. Line 373f: Why is VPRM now classified/labelled as an inventory and not as a biosphere model anymore?

- We have changed the wording to make it clearer (the VPRM model output is biogenic fluxes of CO2).

22. Line 402f: Please clarify the distinction you make between “footprint extent” and “influence region”
    - We have reworded substantially to have consistent phrasing, and have made the linkage clear where we interchange surface influence and footprint. Furthermore, we are now consistent with how we refer to the L_0.90 region without mistakenly overstating its role in our study.

23. Line 415 - 417: Please clarify - on one hand, during winter the receptor is predominantly influenced from low emitting regions northwest (Inner Mongolia), but also subject to CO2 from inefficient district heating? Is that district heating in Mongolia? Or do the more local CO2 sources (e.g. Beijing) dominate the atmospheric CO2 mole fractions at Miyun during this season?
    - Our wording was confusing. As in all seasons the closer the sources are, the more influence they have. We reword to remove dominant, and instead discuss their influence on the site relative to their influence at other times of the year.

24. Line 457f: Suggests that section 4.2.1 implies that better performance of EDGAR and CDIAC is due to an artifact of their lower emissions. However, section 4.2.1 does not explain why EDGAR+VPRM and CDIAC+VPRM being too low has to be due to EDGAR and CDIAC and not a feature of VPRM. One could also easily argue that ZHAO+VPRM matching at hourly scale is an artifact due to too high anthropogenic fluxes. A more detailed discussion why matching hourly data is correct and matching the seasonal data is likely an artifact would be helpful here. One point raised later (line 504) is that VPRM underestimates non-growing season CO2 respiration. Wouldn’t this even further improve the fit of EDGAR/CDIAC+VPRM in Figure 4? And maybe partially explain the underestimation at hourly timescale?
    - If we (justifiably) view wintertime as the anthropogenic baseline where neither GPP nor R are contributing appreciably to the signal, we see there is a consistent offset in the bias relative to observations (mean bias: ZHAO=0.01, EDGAR=-2.2, CDIAC=-3.1). We have modified the text accordingly to better illustrate the discussion that anthropogenic discrepancies are contributing to the model-observation mismatch. In any case, with the limited data and the lack of temporal activity factors we agree with your point of the ZHAO+VPRM providing too high anthropogenic emissions. We have incorporated this into the text.

      We have also clarified the statement at L504: VPRM underestimates respiration across the board (barring winter) not just non-growing season respiration. It’s just that the effects of this underestimated respiration are more pronounced at the time of year where R is high, with lower GPP (e.g, Fall).

      We also wish to clarify that we are certainly not saying “hourly data is best”…we are just identifying the model the model that minimizes errors at all timescales (a multiple constraint).

25. Line 505-510: see comment line 457f.

- See response 24. We have also expanded this part accordingly.

26. Line 550: More detailed needed on the calculation of GRP and a proper assessment of its uncertainty is crucial, see general comments.
    - The GRP for each province was retrieved from the IMF, World Bank, and China Statistical Yearbook. We added the GRP for each province contained in the L_0.90 region. The GRP/GPP values are very uncertain quantities, but in the form they are broadly disseminated it is unfortunately a single value for each province per year. An economic analysis to estimate uncertainty of these values is beyond the scope of our expertise and the study itself. We have, however, made our methodology clearer.

27. Line 551-556: A list of potential reasons for changes in GRP and CO2 emissions is given here, but it is unclear if this is linked to table 3 or just a list of events happening in the discussed time window. For example, the financial crisis of 2008 is mentioned, but no decrease in GRP is visible in Figure 9a.
    - We plot the percent contribution to total GDP for this reason (the raw numbers alone don't necessarily provide the whole picture). In 2008 we do see a plateauing of the regional percent contribution to China's total GDP. We have made this section clearer in terms of strength of conclusions that can be drawn. We also moved the location of the sentence to make the transition from Table 3 and CO2 growth rate discussions to GRP clear.

28. Line 556: Should maybe refer to Figure 9a not 6a?
    - Yes—should be 9a. Fixed.

29. Line 573: Previous section was already 4.4
    - Fixed

30. Line 590: Figure 8 seems not to support an evident increase in CO2 emissions strongly correlated with GRP. Maybe a scatter plot of GRP versus CO2 emissions would help to highlight a possible correlation.
    - We have included a scatterplot in Figure 9.

31. Line 596: Please elaborate why doubling of GRP suggests enlarged production capacity as driver for emission reductions? Would a shift towards more service-oriented businesses or production of higher value goods have the same effect?
    - We have expanded this section and incorporated your point that the same effect would be had with service-oriented shifts.

32. Line 597: The reported decrease of regional carbon intensity by 47% (28%, 65%) is based on which inventory?
    - In the text in the same sentence it states it is calculated by pooling the values across the scaled inventories: "From 2005 to 2009, carbon intensity for the L_0.90 region decreased by 47% (28%,65%), based on a one-sample t-test of pooled emissions intensity changes across scaled inventories."

33. Figure 9: This is a core result of the study and could be discussed in more detail. Are the regional carbon intensity trends of the 3 inventories significantly different after applying the correction when uncertainties in GRP are accounted for?
    - See response #26: accounting for uncertainties in GRP requires access to a level of economic data that is not readily available. We have highlighted the inherent uncertainty in our economic evaluation. We have also expanded the discussion and conclusions from Figure 9 to highlight the truly interesting and main point from this which is that carbon intensity reductions and absolute carbon emissions reductions particularly in emerging countries can be at odds with each other, and therefore distracting from climate goals.

34. Line 608: A longer discussion of the limitations introduced by only using one biosphere model could be added.
    - We have incorporated more of a discussion in the uncertainty section of the conclusions.

35. Line 645f: see comment line 402f
    - Fixed. See response to #2.

**** ADDENDUM TO AUTHOR'S COMMENTS *****
Edited response to Reviewer #2 Question #12:
Line201f: Given that a short tower is used for observations it seems useful to know what the height of the lower/lowest WRF levels used are. 41 vertical levels are mentioned, but without additional information this seems difficult to interpret.

- The first few layers of the 41 wrf-modeled vertical eta levels used by the STILT vertical coordinate system (converted to meters AGL) are provided below, noting that the miyun receptor is within the first layer at 6magl (158masl). There are 9 vertical levels in the first 1500m.

eta	mAGL
0.996500015	27
0.987999976	92
0.976500034	180
0.962000012	293
0.944000006	435
0.921499968	615
0.894500017	836
0.866361856	1070
0.839085698	1302
0.811809599	1540

Evaluating China's anthropogenic CO_2 emissions inventories: a northern China case-study using continuous surface observations from 2005-2009.

Archana Dayalu[1,*], J. William Munger[2,3], Yuxuan Wang[4,5], Steven C. Wofsy[2,3], Yu Zhao[6], Thomas Nehrkorn[1], Chris Nielsen[3], Michael B. McElroy[3], Rachel Chang[7]

[1]Atmospheric and Environmental Research, Lexington, MA, USA
[2]Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA
[3]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
[4]Department of Earth and Atmospheric Sciences, University of Houston, Houston, TX, USA
[5]Department of Earth System Sciences, Tsinghua University, Beijing, China
[6]School of the Environment, Nanjing University, Nanjing, China
[7]Department of Physics and Atmospheric Science, Dalhousie University, Halifax, Canada
*Formerly at Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA

Correspondence to: Archana Dayalu (adayalu@aer.com)

Abstract. China has pledged reduction of carbon dioxide (CO_2) emissions per unit GDP by 60-65% relative to 2005 levels, and to peak carbon emissions overall by 2030. However, the lack of observational data and disagreement among the many available inventories makes it difficult for China to track progress toward these goals and evaluate the efficacy of control measures. To demonstrate the value of atmospheric observations for constraining CO_2 inventories we track the ability of CO_2 concentrations predicted from three different CO_2 inventories to match a unique multi-year continuous record of atmospheric CO_2. Our analysis time window includes the key commitment period for the Paris accords (2005) and the Beijing Olympics (2008). One inventory is China-specific and two are spatial subsets of global inventories. The inventories differ in spatial resolution, basis in national or subnational statistics, and reliance on global or China-specific emission factors. We use a unique set of historical atmospheric observations from 2005–2009 to evaluate the three CO_2 emissions inventories within China's heavily industrialized and populated Northern region accounting for ~33–41 % of national emissions. Each anthropogenic inventory is combined with estimates of biogenic CO_2 within a high-resolution atmospheric transport framework to model the time series of CO_2 observations. To convert the model-observation mismatch from mixing ratio to mass emission rates we distribute it over a region encompassing 90% of the total surface influence in seasonal (annual) averaged back-trajectory footprints (L_0.90 region). The L_0.90 region roughly corresponds to northern China. Except for the peak growing season, where assessment of anthropogenic emissions is entangled with the strong vegetation signal, we find the China-specific inventory based on subnational data and domestic field-

Deleted: Carbon dioxide emissions in Northern China based on atmospheric observations from 2005 to 2009

Formatted: Subscript

Formatted: Subscript

Deleted: In this study, we demonstrate an approach based on a long time series of surface CO_2 observations to evaluate regional CO_2 emissions rates in northern China estimated by three anthropogenic CO_2 inventories—two of which are subsets from global inventories, and one of which is China-specific. Comparison of CO_2 observations to CO_2 predicted from accounting for global background concentration and atmospheric mixing of emissions suggests potential biases in the inventories. The period analyzed focuses on the key commitment period for the Paris accords (2005) and the Beijing Olympics (2008). Model-observation mismatch in concentration units is translated to mass units and is displayed against the original inventories in the measurement influence region, largely corresponding to northern China.

studies agrees significantly better with observations than the global inventories at all timescales. Averaged over the study time period, the unscaled China-specific inventory reports substantially larger annual emissions for northern China (30%) and China as a whole (20%) than the two unscaled global inventories. Our results, exploiting a robust timeseries of continuous observations, lend support to the rates and geographic distribution in the China-specific inventory Though even long-term observations at a single site reveal differences among inventories, exploring inventory discrepancy over all of China requires a denser observational network in future efforts to measure and verify CO_2 emissions for China both regionally and nationally. We find that carbon intensity in the northern China region has decreased by 47% from 2005 to 2009, from approximately 4kgCO_2/USD$_{PPP}$ in 2005 to about 2kgCO_2/USD$_{PPP}$ in 2009 (Figure 9c). However, the corresponding 18% increase in absolute emissions over the same time period affirms a critical point that carbon intensity targets in emerging economies can be at odds with making real climate progress. Our results provide an important quantification of model-observation mismatch, supporting the increased use and development of China-specific inventories in tracking China's progress as a whole towards reducing emissions. We emphasize that this work presents a methodology for extending the analysis to other inventories and is intended to be a comparison of a subset of anthropogenic CO_2 emissions rates from inventories that were readily available at the time this research began. For this study's analysis time period, there was not enough spatially distinct observational data to conduct an optimization of the inventories. The primary intent of the comparisons presented here is not to judge specific inventories, but to demonstrate that even a single site with a long record of high time resolution observations can identify major differences among inventories that manifest as biases in the model-data comparison. This study provides a baseline analysis for evaluating emissions from a small but important region within China, as well a guide for determining optimal locations for future ground-based measurement sites.

Deleted: Owing to limitations from having a single site, addressing the significant uncertainty stemming from transport error and error in spatial allocation of the emissions remains a challenge. Our analysis uses observations to support and justify increased use and 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 evaluating and constraining inventories at relevant spatial scales requires multiple stations of high-temporal resolution observations. At this stage and with observational data limitations, we emphasize that this work is intended to be a comparison of a subset of anthropogenic CO_2 emissions rates from inventories that were readily available at the time this research began. For this study's analysis time period, there was not enough spatially distinct observational data to conduct an optimization of the inventories. Rather, our analysis provides an important quantification of model-observation mismatch. In the northern China evaluation region, emission rates from the China-specific inventory produce the lowest model-observation mismatch at all timescales from daily to annual. Additionally, we note that a

Deleted: has

Deleted: higher

Deleted: and the northern China evaluation region (30%)

Deleted: unscaled

Deleted: . However, exploring this discrepancy for China as a whole …

Moved down [1]: This study provides a baseline analysis for a small but import region within China, as well a guide for determining optimal locations for future ground-based measurement sites.The primary intent of the comparisons presented here is not to

Moved (insertion) [1]

Deleted: ¶

1 Introduction

China's contribution to world CO_2 emissions has been steadily growing, becoming the largest in the world in 2006. China has accounted for 60% of the overall growth in global CO_2 emissions over the past 15 years (EIA, 2017). Under the United Nations Framework Convention on Climate Change (UNFCCC) 2015 Paris Climate Agreement, China has committed to reduce its carbon intensity (CO_2 emissions per unit GDP) by 60-65% relative to the baseline year of 2005, and to peak carbon emissions overall by or before 2030. Demonstration of progress on emissions reduction and evaluation of how well specific policies are working is hindered by large uncertainty in the existing Chinese emission inventories. In 2012 the discrepancy between data reported at national and provincial levels was approximately half of China's 2020 emission reduction goals (EIA, 2017; NDRC, 2015; Guan et al., 2012; Zhao et al., 2012). Moreover, China is under mounting pressure to address severe regional air pollution events that are often associated with CO_2 emissions sources—vehicles, power plants and other fossil fuel-burning operations. China's 11th Five Year Plan (11th FYP) of 2006-2010 included aggressive measures to retire inefficient coal-fired power plants and improve energy efficiency in other industries starting in 2007 (Zhao et al., 2013; Nielsen & Ho, 2013). A number of pollution control measures that were implemented specifically in preparation for the 2008 Beijing Summer Olympics were also largely in effect by the end of 2007 (Nielsen & Ho, 2013; Wang et al., 2010).

A variety of top-down approaches including inverse analysis (Le Quere et al., 2016) and comparison between atmospheric observations and Eulerian forward model predictions (Wang, X. et al., 2013) have been used to evaluate and constrain emission estimates, albeit at coarse spatial resolution. As noted by Wang et al. (2011) grid-based atmospheric models have difficulty in simulating high-concentration pollution plumes at specific receptor sites that are too near the source region. The expanding network of high accuracy CO_2 observations coupled with high spatial resolution transport models is emerging as a viable tool for evaluating high resolution emission inventories (e.g. Sargent et al., 2018). In this paper we adopt a Lagrangian transport model to simulate atmospheric mixing and transport. Continuous observations of CO_2 for the period 2005-2009 at Miyun, an atmospheric observatory about 100km NE of Beijing provide a top-down constraint for evaluating persistent bias among emissions rates obtained from a suite of three independent anthropogenic emission inventories that were readily available as spatially gridded fluxes.

The three inventories that are evaluated span a range of bottom-up inventory approaches. They are not intended to be an exhaustive set, but are examples to demonstrate the capability to identify significant differences in the ability of different inventories to match the long time series of observations. Emerging inventory approaches based on updated (yet non-China-specific) point-source data and satellite-observations of night lights as a proxy for spatial allocation of energy production (Oda et al., 2018) were not readily available when this analysis began. Two of the inventories, the Emissions Database for Global Atmospheric Research (EDGAR; European Commission, 2013) and Carbon Dioxide Information Analysis Center (CDIAC), are spatial subsets from larger global models of CO_2 emissions

Deleted: differences in

(PBL, 2013; Andres et al., 2016). They rely on national-level energy statistics and global default values for sectoral emission factors, and they estimate activity levels using generalized proxies (e.g. population). The third inventory (ZHAO) is specific to China, with greater reliance on energy statistics at provincial and individual facility levels as well as emission factors from domestic field studies (Zhao et al., 2012). The ZHAO inventory was readily accessible at the time of this research and represents increased efforts in recent years to incorporate more China-specific data into emissions inventories. Other China-specific inventories that have been recently developed but were not readily available at the time of this research include the Multi-resolution Emissions Inventory (MEIC, http://www.meicmodel.org/) and an inventory by Shan et al., 2016. The primary intent of the comparisons presented here is not to judge specific inventories, but to demonstrate that even a single site with a long record of high time resolution observations can identify the potential impact of major differences among inventories that manifest as biases in the model-data comparison.

A study by Turnbull et al. (2011) used weekly flask observations to evaluate a hybrid approach to inventory construction where CDIAC and EDGAR estimates were spatially allocated to a provincial emissions-based grid. However, to our knowledge, none of the truly China-specific CO_2 inventories have been evaluated with independent high-temporal resolution atmospheric observations. The official national total for China's 2005 CO_2 emissions from energy related activities, used as the benchmark for the Paris commitment, is approximately 5.4Gton CO_2 (NDRC, 2015). ZHAO, EDGAR, and the CDIAC national total (Boden et al., 2016) report total 2005 energy-related CO_2 emissions that are higher by 31% (7.1Gton), 9%(5.9Gton), and 7%(5.8Gton), respectively. As the official national total is not available in a spatially allocated format, it cannot be tested by observations and we refer to it only as a benchmark in our analysis. We will show that the China-specific inventory (ZHAO) provides excellent agreement with observations, and markedly more so than EDGAR and CDIAC, The result provides guidance for efforts to assess China's emissions at larger scales as well as potential updates for the Paris agreement base year emissions.

In order to independently evaluate and scale existing bottom-up estimates of China's CO_2 emissions, we employ a top-down approach using five years of continuous CO_2 observations. Modeled concentrations of CO_2 are obtained from convolving hourly CO_2 surface flux estimates with surface influence estimates ("footprints") derived from the Stochastic Time-Inverted Lagrangian Transport Model driven with meteorology from the Weather Research and Forecasting Model version 3.6.1 (WRF-STILT; Lin et al., 2003; Nehrkorn et al., 2010). NOAA CarbonTracker (CT2015) provides modeled estimates of advected upwind background concentrations of CO_2 that are enhanced or depleted by processes in the study region. As atmospheric CO_2 concentrations are significantly modulated by photosynthetic and respiratory fluxes, we additionally prescribe hourly biosphere fluxes of CO_2 using data-driven outputs from the Vegetation, Photosynthesis, and Respiration Model (VPRM) adapted for China (Mahadevan et al., 2012; Dayalu et al., 2018). VPRM provides a functional representation of biosphere fluxes based on data from remote sensing platforms and eddy flux towers, with significantly better observationally-validated performance relative to subsets of global vegetation models (Dayalu et al., 2018). The WRF-

Deleted:

Deleted: while the others do not

Deleted: (January 2005 through December 2009)

Deleted: hourly-averaged

Deleted: measured in Miyun, China, at a site 100km northeast of Beijing (Wang et al., 2010)

Deleted: maps

STILT-VPRM framework has been successfully adapted for similar emissions evaluation studies in North America in regions where biogenic fluxes dominate surface processes (e.g., Sargent et al., 2018; Karion et al. 2016; Matross et al., 2008). For the Northern China region, anthropogenic fluxes exceed biogenic fluxes for all but the peak of growing season, when they are roughly comparable (Dayalu et al., 2018), which reduces the magnitude of overall error from incorrect modeling of the biosphere. In contrast to extensive measurement networks that exist in North America, continuous high-temporal resolution measurements of CO_2 necessary for inventory evaluation applications are sparse and very few datasets are available in China (Wang et al. 2010). Despite this limitation, our site provides valuable information and constraints on emissions inventories; the long time series and spatial sampling heterogeneities where the site receives both clean continental air as well as air from one of the heaviest emitting regions of China, present a powerful and unique dataset for the region. Our inventory scaling is confined to the Northern China region, but this region accounts for 33-41% of China's total annual CO_2 emissions from fossil-fuel combustion. Model-observation mismatches can be converted from concentration units (ppm) to mass units (Mton CO_2) across the most relevant area subset from modeled annual average surface sensitivity footprints ($\mu mol^{-1}\ m^2\ s$). Ultimately, we compare the inventories by quantifying model-observation mismatch for seasons (using additive mass units) and annually (using scaling factors). We note that identical transport fields and modeled biogenic fluxes are applied to all the anthropogenic emission fields. Unresolved transport error and error in biogenic fluxes undoubtedly contributes to scatter in the model-data comparison. While random transport errors are unlikely to generate consistent biases among the inventories, systematic transport errors can be attributed to biases among inventories with differing spatial allocations. Although the interaction of systematic transport errors with differences in spatial distribution could bias individual observations, averaging over longer timescales (seasons, years) minimizes the bias of individual points. With the available observational data it is not possible to evaluate the error in spatial allocation of individual emissions inventories. For example, future access to total column measurements and/or aircraft vertical profiles would provide additional constraints on spatial allocations of sources and sinks.

Section 2 of this paper describes the observational CO_2 record used in this analysis. Section 3 details the analysis methods, including WRF-STILT model configuration, a discussion of the main features of the inventories, error evaluation, and inventory scaling methods. We present the results in Sect. 4, beginning with an assessment of seasonality impacts. We then compare inventory performance against observations across multiple timescales from hourly to annual. We conclude Sect. 4 with scaling results, and a brief examination of regional carbon intensity over the study period. Concluding remarks are provided in Sect. 5. Additional methodological details are provided in the accompanying Supplementary Information (SI) and at https://doi.org/10.7910/DVN/OJESO0.

Deleted: being restricted to a single measurement station

Deleted: because

Deleted: it

Deleted: receives

Deleted: at different times

Deleted: and clean air at other time

Deleted: s

Deleted: based on the

Deleted: included in

Deleted: the

Deleted: influence

Deleted: consisten

Deleted: But

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Not Highlight

Deleted: The scaling factors are resolved at the policy-relevant seasonal and annual timescale. With a single receptor our scaling applies to a limited geographical extent (see below) and is limited to a linear scaling (or additive) factor.

Deleted: any

Deleted: However, we note that the same transport model is applied to all the emission fields. Unresolved transport error undoubtedly contributes to scatter in the model-data comparison but is unlikely to generate consistent biases among the inventories.

Deleted: , and a final summary of the caveats and limitations of our study

2 CO_2 observations

This study uses five years (2005-2009) of continuous hourly averaged CO_2 observations (LI-COR Biosciences Li-7000; 2-σ analytical precision of 0.08ppm), measured at a site in Northern China (Miyun; 40°29'N, 116°46.45'E). The Miyun receptor is an atmospheric measurement station in a rural site 100 km northeast of the Beijing urban center (Fig. SI S2). It was established in 2004 by collaborating researchers at the Harvard China Project and operated by researchers at Tsinghua University. The site is strategically located to capture both clean continental background air from the west/northwest and polluted air from the Beijing region to the southwest. Miyun is located south of the foothills of the Yan mountains; the region consists of grasslands, small-scale agriculture intermingled with rural villages and manufacturing complexes, and mixed temperate forest. Land use grades from rural to suburban and dense urban to the south towards Beijing center and sparsely populated and wooded mountains to the north and west. Further descriptions of the site and details of the instrumentation including calibration strategy and assessment of long-term drifts are in provided in Wang et al. (2010). Average annual data coverage (based on hourly data) over the study time period was 83% (range: 78% to 92%).

Formatted: Font: Symbol

Deleted: made

Deleted: of the CO_2 observations

Deleted: in this

3 Methods

We evaluate the performance of the ZHAO, EDGAR, and CDIAC inventories coupled with biogenic fluxes by modelling five years of hourly CO_2 observations using the Stochastic Time-Inverted Lagrangian Transport Model (STILT; Lin et al., 2003) run in backward time mode driven by high resolution meteorology from the Weather Research and Forecasting Model version 3.6.1 (WRF). The WRF-STILT tool models the surfaces that influenced each measurement hour in the study domain (Figure 1). Hourly vegetation CO_2 fluxes are prescribed by the VPRM adapted for China (Mahadevan et al., 2008, Dayalu et al., 2018). We categorize seasons by months based on regional growing season patterns, which are heavily dominated by winter wheat/corn dual-cropping regions in the North China Plain (Dayalu et al. 2018). Winter wheat emergence in the spring and corn emergence in later summer shift the seasonal patterns such that regional seasons are more appropriately represented as January, February, March (JFM/Winter); April, May, June (AMJ/Spring); July, August, September (JAS/Summer); and October, November, December (OND/Fall).

Deleted: when months of year are grouped

Deleted: , respectively

Ultimately, modeled concentrations of CO_2 are obtained from convolving hourly surface flux estimates with footprints derived from the WRF-STILT framework. NOAA CarbonTracker (CT2015) provides estimates of advected upwind background concentrations of CO_2 that are enhanced or depleted by processes in the study region. Our final modeled-measurement data set is the subset consisting of local daytime values (hourly data from 1100h to 1600h). Of this subset, only individual hours for which observational data exists (i.e., non-missing data) is included. The final data set was further filtered to include only CT2015 background values satisfying true background criteria as described in Sect. 3.5

Deleted: surface influence maps

Deleted: filtered to include only non-missing observations and

and in the SI, Sect. S4. As is typical for studies of this nature, our analysis focuses on observations during the 1100 to 1600 local time period. The stronger vertical mixing in the daytime atmosphere (notably absent at night) reduces the influence of extremely local emissions. We select the 1100-1600 window to avoid the presence of shallow inversion layers that are poorly represented in STILT and use the period when vertical mixing through the entire boundary layer is at its maximum (McKain et al., 2015; Sargent et al., 2018). We adjust fluxes based on model-measurement mismatch of this final data subset, focusing on the region that we model as most influential to the signal measured at the receptor. Method details and model components are described individually below.

Deleted: the
Deleted: 6
Deleted: because
Deleted: , shallow inversion layers that STILT represents poorly are absent, and vertical concentration gradients within the boundary layer are at a minimum
Deleted: scale
Deleted: inventories
Deleted: M
Deleted: ¶

3.1 WRF-STILT Model Configuration

The WRF-STILT particle transport framework and optimal configuration have been extensively tested in several studies using mid-latitude receptors (e.g., Sargent et al., 2018; McKain et al., 2014; Kort et

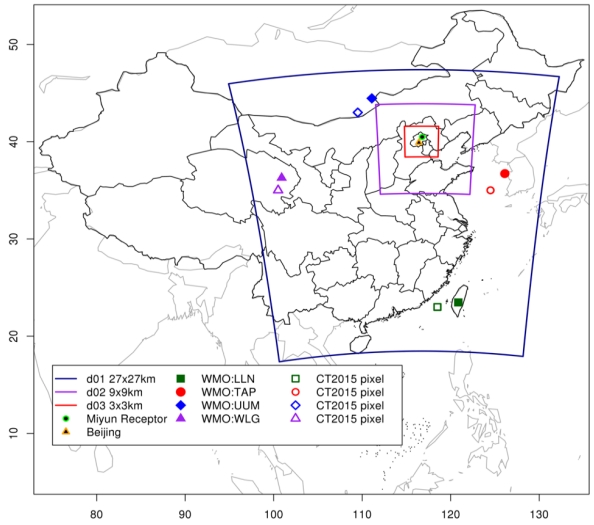


Figure 1. Study domain configuration. Miyun receptor and Beijing center are located within the innermost domain at a resolution of 3x3km. NOAA ESRL/WMO (WMO) flask sampling sites used to evaluate bias in CT2015 modeled backgrounds are the solid shapes; nearest CT2015 comparison pixel is the corresponding unfilled shape.

al., 2013; McKain et al. 2012; Miller et al., 2012). WRF is configured with 41 vertical levels and two-
315 way nesting in three domains, with the outermost domain covering nearly seven administrative regions (Figure 1, Figure 2), defined according to convention in Piao et al. (2009). The domain resolutions from coarsest to finest are 27km (d01), 9km (d02), and 3km (d03). Initial and lateral WRF boundary conditions are provided by NCEP FNL Operational Model Global Tropospheric Analyses at 1°x1° spatial 6-hourly temporal resolution (NCEP, 1999). Nudging of fields is implemented in the outer
320 domain only, and never within the Planetary Boundary Layer (PBL). WRF output is evaluated against

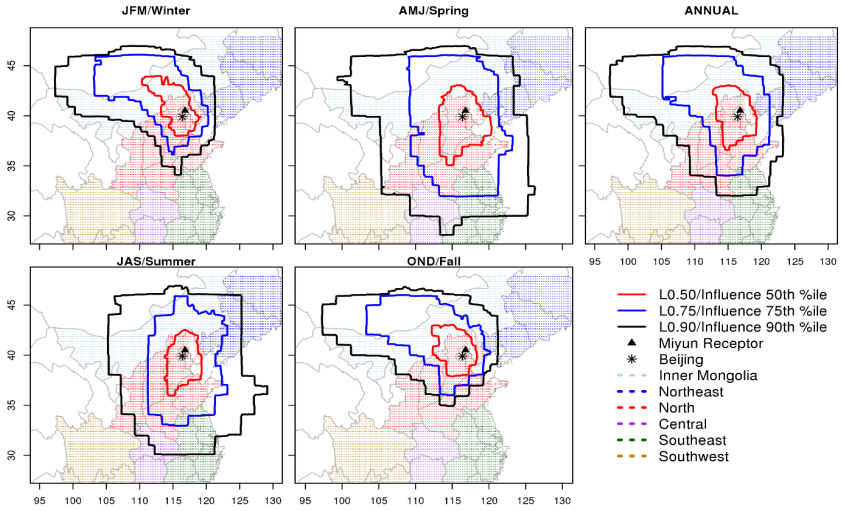


Figure 2. 2005-2009 mean seasonal (a-d) and Annual (e) footprint contours, as percentiles of influence highlighted by administrative region. Red, blue, and black contour lines represent 50th, 75th, and 90th percentile regions, respectively. Stippling represents location of 0.25º x 0.25º footprint and inventory gridcell centers, colored by relevant administrative regions. Northern China (red stippling) is the administrative region with predominant influence on Miyun observations, followed by Inner Mongolia and Northeast China. Southeast and Central China have minimal representation, and only during the spring and summer seasons.

publicly accessible 24-hourly averaged observational datasets from the Chinese Meteorological Administration (CMA); finer temporal resolution meteorological data is not publicly available. WRF run details are presented in Dayalu (2017) and at http://dx.doi.org/10.7910/DVN/OJESO0. A snapshot

of results from comparison with China Meteorological Administration ground-station measurements is presented in SI Sect. S1 and Figures S1-S4.

The STILT model is configured in backward time mode. The particle release point is set as the Miyun measurement sample inlet (the receptor). The inlet height is 158m above sea level (masl), corresponding to 6m above ground level (magl). In our study, the hilltop site was located in an area where the surrounding land was not very productive or intensively cultivated (SI Fig. S2). There is a long history of using short towers in low productivity areas for regional studies (e.g. NOAA Earth Systems Research Laboratory—NOAA ESRL Barrow, Alaska observatory at 11 magl). In addition, the station is located on a small hilltop, so even though the actual inlet height above ground is low, it has a topographic advantage in that it effectively samples air from a greater height relative to the surroundings. Topographic advantage 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. In our study, independent verification from concurrent aircraft measurements (for example) or multi-level inlet locations were not available to quantify the impact of absolute and relative inlet location on transport uncertainty.

Each hourly footprint (CO_2 concentration attributed to each unit of flux as ppm $\mu mol^{-1}\ m^2\ s$) provides an estimate of surface influence on the measurement and is calculated from releasing 500 particles from the measurement site (receptor) until they reach the outer domain boundaries up to seven days back in time. The STILT 0.25º x 0.25º footprint map for each measurement hour up to 7 days back in time enables assessment of regions in the study domain to which the receptor is most sensitive. These entire gridded footprints are convolved with anthropogenic and biogenic CO_2 flux estimates to provide a final modeled concentration (ppm) of CO_2 at the receptor. For clarity, we display the regions of importance to the receptor based on contours calculated from the overall STILT footprints at the 50th (L_0.50 region), 75th (L_0.75 region), and 90th (L_0.90 region) percentile levels (Figure 2). The percentile contours are calculated as follows: the average (seasonal, annual) footprints from 2005 to 2009 are ordered from high to low. We multiply each fraction (0.5,0.75,0.9) with the summed footprints and use cumulative sums of the ordered footprints as a guide to select all points with influence magnitude equal to or greater than this cutoff value. SI Figure 11 illustrates a single footprint map along with the average influence and a plot of cumulative influence to demonstrate the percentile level selection process. We emphasize that we use the entire STILT footprint convolved with fluxes to estimate the receptor CO_2 concentration. We only use the L_0.90 region to provide a reasonable area across which to ascribe the effective inventory adjustment (converted from ppm model-observation mismatch to mass units). As SI Figure 11c shows, the L_0.90 region strikes a balance between capturing sufficient influence while avoiding an unrealistically large adjustment region for a single observation site. Conversely, corrections based on the smaller L_0.75 region would include larger uncertainties from the diffuse influence of emissions outside the L_0.75 region (not accounting for 25% of average surface sensitivity), yet the model-observation mismatch would be ascribed to a region approximately half the area of the L_0.90

Deleted: ¶
Deleted: ,
Deleted: with t
Deleted: of
Deleted: Our study has additional limitations, however, because
Formatted: Subscript
Deleted: We calculate
Formatted: Subscript
Deleted: STILT surface influence
Deleted: displays sample
Deleted: s and illustrates
Deleted: L_0.90—the region estimated as containing 90% of surfaces influencing measurement—is selected as the inventory comparison region…
Formatted: Subscript
Moved (insertion) [5]
Deleted: For example
Deleted: 50
Deleted: 50
Deleted: still
Deleted: 40
Deleted: modeled input
Deleted: significantly smaller

region. Deriving correction factors based on integration over the entire L_0.90 region is a more conservative approach where the model-observation mismatch in mass units is distributed over a larger area.

Further model details are available in SI Sect. S2. Complete WRF-STILT settings and STILT footprint files are available from http://dx.doi.org/10.7910/DVN/OJESO0.

3.3 Anthropogenic CO_2 Emissions Inventories

ZHAO, EDGAR, and CDIAC report estimates of total annual emissions of CO_2 at 0.25º x 0.25º, 0.1º x 0.1º, and 1º x 1º original grid resolutions, respectively. We regridded the EDGAR and CDIAC inventories to the 0.25º x 0.25º resolution, using NCAR Command Language version 6.2.1 Earth System Modeling Framework conserve regridding algorithm to preserve the integral of emissions (Brown et al., 2012). Differences between annual total emissions for EDGAR and CDIAC inventories introduced by regridding are smaller than the interannual trends or differences between the inventories (SI Sect. S3 and Figure S5). We present the main components and defining features of the three anthropogenic CO_2 inventories below.

The ZHAO inventory provides estimates of total annual emissions for 2005 through 2009. In addition, spatial location of emissions is given for years 2005 and 2009 on a 0.25º x 0.25º grid. Using 2005 and 2009 gridded values, we calculate an average percent contribution of each grid cell to the total emissions. The average contributions are used as weights to spatially allocate 2006, 2007, and 2008 total annual emissions. We evaluate and justify this assumption in detail in SI Sect. S3 and Figure S6. The ZHAO inventory represents one of the first statistically rigorous bottom-up CO_2 inventories for China. It relies on provincial- and facility-level data rather than national level data, which has been noted previously as major uncertainty in Chinese emission inventories; total CO_2 emissions estimates based on provincial data are typically higher than those using national statistics (Zhao et al., 2013). Satellite observations of criteria air pollutants (e.g., nitrogen dioxide, which serves as a proxy for fossil fuel combustion) show greater agreement with provincial statistics (Zhao et al., 2012). The increased use of China-specific emission factors and activity levels based on domestic field studies is a shift from other inventories that rely heavily on global averages to estimate processes occurring in China. Despite the increased incorporation of China-specific field data, the largest sources of uncertainty to the ZHAO inventory are industrial emission factors, and activity levels across all sectors. Total uncertainty in the inventory is estimated as -9% to +11%. (Zhao et al., 2012).

The EDGAR emissions database continues to be a major prior in atmospheric studies, and the CO_2 inventory is used to inform key global scientific results considered by the UNFCCC Conference of Parties. The EDGAR global inventory (atemporal EDGAR v4.2 FT2010 gridded emissions) takes total annual estimates of national emissions and downscales emissions to a 0.1º x 0.1º as a function of

Deleted: diffused

Moved up [5]: For example, corrections based on the smaller L_0.50 region would include larger uncertainties from the diffuse influence of emissions outside the L_0.50 region (still 40% of modeled input), yet the model-observation mismatch would be ascribed to a significantly smaller region.

Formatted: Hyphenate

road/shipping networks, population density, energy/manufacturing point sources, and agricultural land. Estimates for China are available for all five years as gridded inventories. Reported uncertainties for global emissions are ±10% (http://themasites.pbl.nl/tridion/en/themasites/edgar/documentation/uncertainties/index-2.html). However, this applies to global averaged uncertainty; we expect uncertainty for China to be much higher.

We include the CDIAC inventory here due to its historical prevalence as a benchmark inventory for global indicators, including evaluations of carbon intensity provided by the World Bank (World Bank, 2017). The CDIAC inventory (v2016; https://dx.doi.org/10.3334/CDIAC/ffe.ndp058.2016) allocates estimates of national emissions to a 1° x 1° grid, primarily distributed according to human population density. A thorough assessment of 2σ uncertainties in the CDIAC spatial allocation of emissions shows considerable spread in regional uncertainties (Andres et al., 2016).

Our study is not intended to be an exhaustive sampling of inventory approaches but serves to demonstrate the utility of continuous high-accuracy observations as a top-down constraint on emissions evaluations. Our inventory list notably does not include emerging spatially resolved global inventories (e.g. Open Data Inventory for Anthropogenic Carbon Dioxide, ODIAC) (Oda et al., 2018) that were not readily available at the time this work was conducted. At 1km x 1km, ODIAC does have a high spatial resolution of nightlight proxy-based emissions; while this is a valuable method for regions in Europe and North America for example, it is less valuable for China where it is analogous to the CDIAC population-based proxy. In China, power plant emissions are typically located far from end-use regions and the night-light proxy can often break down (Wang, R. et al., 2013). Furthermore, ODIAC power plant emissions use the 2012 Carbon Monitoring for Action (CARMA) database, which notably does not incorporate China-specific power plant data; in these instances, CARMA categorizes China's power plants as "non-disclosed plants" and reports using estimates derived from statistical models using averaged emissions factors – comparable to methods in global inventories subset over China (Ummel, 2012). One of our main goals is to quantify model-observation mismatch associated with use of China-specific power plant data, and ODIAC does not address that issue particularly differently from other global emissions inventories subset over China. For completeness, however, evaluation of global inventories like ODIAC and a suite of increasingly available China-specific inventories (e.g., MEIC) would provide value as part of future model-observation comparison efforts.

Based on multi-year means (2005 to 2009) and 95% confidence intervals derived from two-sample t-tests, we find that within the L_0.90 evaluation region EDGAR and CDIAC report emissions that are significantly lower than ZHAO by typically 20% (-24%, -16%) and 36% (-37%, -34%), respectively. Across China's administrative regions, the highest discrepancy between the global and regional inventories is in Northern China (ZHAO is approximately 30% higher than both EDGAR and CDIAC). In addition, Northern China represents one of the administrative regions with the highest CO_2 emissions density (2300 to 3300 Megagrams of CO_2 per square kilometer, compared to the average of 700 $MgCO_2$

Deleted: .
Deleted: the
Deleted: is expected
Formatted: Font color: Auto
Deleted: This
Deleted: as an
Deleted: ;
Deleted: however,
Deleted: it is sufficient
Deleted: for evaluating emissions estimates
Deleted: over China
Deleted: .
Deleted: .
Deleted: kilotonnes
Deleted: 0.
Deleted: $ktCO_2$

km^{-2} averaged across China) and is therefore a particularly rich spatial subset for emissions inventory evaluation. A detailed breakdown of emissions by region of China is provided in the SI Table S1. Spatial differences are displayed in SI Figure S7.

Previous work has found that temporal variations in CO_2 sources can be significant and surface CO_2 can be perturbed from 1.5-8 ppm within source regions based on time of day and/or day of week, resulting from a combination of changes in activity patterns as well as synoptic scale transport effects (Nassar et al., 2013). However, appropriate data for establishing reasonable temporal scaling factors for data-sparse regions such as China are difficult to obtain, and as in the case of Nassar et al. (2013) China's activity factors are based on United States activity factors weighted according to China's EDGARv4.2 emissions patterns. We applied the weekly and diurnal Nassar et al. (2013) scaling factors to our emissions, but these did not generate statistically significant differences from the unscaled versions. These statistically insignificant results suggest that a more rigorous set of temporal scaling factors need to be developed for China. CDIAC does provide monthly gridded inventories with seasonality embedded. However, predictions based on that seasonality deviated even further from the observations than predictions based on constant annual emissions. In the CDIAC global dataset, the seasonality in emissions are based upon generalized global activity factors that are not necessarily appropriate for estimating seasonality of human activity in China. Therefore, in this study we do not explicitly consider diel and seasonal variation in anthropogenic CO_2 fluxes.

Deleted: A
Deleted: ying
Deleted: that were statistically significant
Deleted: ,
Deleted: ting

3.4 Vegetation Flux Inventory

We prescribe biotic contributions to the CO_2 signal by adapting the VPRM model output for the study domain to generate 0.25º x 0.25º gridded estimates of hourly CO_2 net ecosystem exchange (*NEE*) from 2005 to 2009. Details of the VPRM model and output for China are presented in Dayalu et al., 2018. The VPRM is driven by 8-day 500m MODIS surface reflectance values and 10-minute averages of WRF downward shortwave radiation and surface temperature fields. The VPRM parameters are calibrated using eddy flux measurements in the study domain representing each ecosystem type classified according to the International Geosphere-Biosphere Programme (IGBP) scheme. Calibration and evaluation eddy-flux data are obtained from FluxNet and ChinaFlux collaborators. The L_0.90 region is dominated by croplands (Figure S8), in particular the winter wheat and corn dual cropping that characterizes the North China Plain (Dayalu et al., 2018). We use one biosphere model in this study to simplify our assessment of variations across the different emissions inventories. Our selection of the VPRM in particular is based on results from Dayalu et al. (2018), where the VPRM was shown to have significantly lower regional bias than an ensemble of global 3-hourly flux products subset over China.

Deleted: (
Deleted:)
Deleted: Eddy
Deleted:
Deleted: dual-cropping

3.5 Background Concentrations

Appropriate quantification of background CO_2 concentrations (i.e., the CO_2 concentration at the lateral edges of the model domain and/or prior to interaction with domain surface processes) enables realistic

assessment of the study domain's contribution to atmospheric CO_2 at varying timescales. CT2015 estimates of CO_2 concentrations are provided on a 3º x 2º grid at upwind background locations. Background values are selected and corrected for large-scale biases using methodology similar to Karion et al. (2016) where a particle must originate from the outermost domain edge and/or 3000 masl; further details are provided in the SI Sect. S4. The predicted background CO_2 is shown together with observed CO_2 at Miyun for the 1100h-1600h period over the 5-year observational record Figure 3a. For most of the year the measured CO_2 shows large enhancements above background and only in mid-summer is there a small depletion relative to background values.

3.6 Quantifying Regional Changes to Background CO_2 Concentrations: ΔCO_2

We define hourly ΔCO_2 as a regional change (enhancement or depletion) imparted to concentrations of CO_2 advected from the boundary ($CO_{2,CT2015}$) such that for each observation hour $\Delta CO_{2,obs}$:

$$\Delta CO_{2,obs} = CO_{2,obs} - CO_{2,CT2015} \quad (1)$$

For each modeled hour $\Delta CO_{2,mod}$, i and j represent the surface gridcell locations and h represents the hour of the 7-day back trajectory:

$$\Delta CO_{2,mod} = \sum_{0h}^{-168h} \sum_{ij} foot_{ij} \times (ANTH_{ij} + VPRM_{ij}) \quad (2)$$

Note that for the modeled enhancement or depletion, only the VPRM fluxes change hourly; as stated previously, the annual anthropogenic fluxes are atemporal.

Without a sufficiently dense network of high temporal resolution observations, full-scale inverse modeling approach to inventory scaling is inappropriate. At annual timescales, where anthropogenic sources dominate the CO_2 signal, we compare annual observed and modeled ΔCO_2 to define a mean bias and derive a scale factor to quantify the model-observation mismatch based on the slope of the comparison. Isotopic analysis of atmospheric CO_2 from a site in Beijing in 2014 suggests that annually the fossil fuel burning does dominate the region, contributing 75±15% to the annual signal (Niu et al., 2016). Annually, the biospheric impact in the region is not zero; rather, the anthropogenic signal dominates. The biospheric quantity of relevance annually is the net carbon flux as a balance of GPP and respiration, and is highly uncertain in both sign and magnitude in this region (Piao et al., 2009). In the Piao et al. (2009) study, regional inversions are based on the very limited dataset of nine sites across all of Asia. Our assumption of dominant anthropogenic influence in northern china is in keeping with the

Deleted: and is
Deleted: ed
Moved down [3]: $\Delta CO_{2,obs} = CO_{2,obs} - CO_{2,CT2015}$
Deleted: $\Delta CO_{2,obs} = CO_{2,obs} - CO_{2,CT2015}$
Moved (insertion) [3]
Formatted Table
Moved down [4]: $\Delta CO_{2,mod} = \sum_{0h}^{-168h} \sum_{ij} foot_{ij} \times (ANTH_{ij} + VPRM_{ij})$
Moved (insertion) [4]
Formatted Table
Formatted: Subscript

priors and process-based models from the relevant regions in Piao et al. (2009) that assume zero and are not significantly corrected by relatively poorly constrained inversions. At seasonal timescales, we use the difference between observed and modeled ΔCO_2 normalized by L 0.90 area to obtain a mass flux offset that combines vegetation and anthropogenic inventories. With the available data it is not possible to independently evaluate both the anthropogenic and biogenic CO_2 fluxes. For further details of the scaling technique, please refer to SI Sect. S5.

Deleted: footprint

3.6.1 Uncertainty Analysis

The sources of uncertainty in calculations of ΔCO_2 include uncertainty in CT2015 background concentrations, CO_2 observations, STILT footprints, anthropogenic inventories, and the biogenic CO_2 fluxes from the VPRM. We obtain 95% confidence bounds for ΔCO_2 by following a procedure similar to McKain et al. (2015) and Sargent et al. (2018) that involves bootstrapping daily averages of hourly afternoon values. For monthly and seasonal timescales, we obtain 95% confidence intervals for $\Delta CO_{2,obs}$ by performing a bootstrap on probability distributions of errors in both the CT2015 and observations 1000 times. (See SI Sect. S4 and Figure S9 for details on parameterizing CT2015 uncertainty.) The relevant quantiles are obtained from the resulting distribution, and are reported relative to the mean $\Delta CO_{2,obs}$ of the original data subset. We follow a slightly modified approach for $\Delta CO_{2,mod}$ in that we construct monthly and seasonal residual pools from daily averages of hourly afternoon $CO_{2,mod}$-$CO_{2,obs}$. The residuals—the deviation of the model from the true observed values—represent the total uncertainty in the model and therefore aggregates the effects of uncertainty in the footprints, background, and inventories. Monthly and seasonal 95% confidence intervals of $CO_{2,mod}$-$CO_{2,obs}$ are then obtained from the distribution of bootstrapping the residual pools 1000 times. We then obtain the mean and 95% confidence interval of $\Delta CO_{2,mod}$ by applying the relevant quantiles of the residuals to the mean $\Delta CO_{2,obs}$ of the original data subset. Similar to Sargent et al. (2018) and McKain et al. (2015), distributions of seasonal averages obtained from the above method are used to estimate annual averages and 95% confidence intervals.

Formatted: Subscript

Deleted: vegetation inventory

Sargent et al. (2018) note that applying the same meteorological model over a long time period (15 months) allows for detection of trends in transport uncertainty. In this study, the drawback of a single location is offset somewhat by a much longer time series (60 months). Absent a dense network of observations, a more sophisticated and extensive error analysis cannot be conducted with meaningful results. Turnbull et al. (2011) faced a similar issue, where weekly flask data collected between 2004 and 2010 from two sites in the NOAA ESRL/WMO sampling network were used to evaluate a bottom-up fossil inventory based on CDIAC and EDGAR estimates. Turnbull et al. (2011) noted the difficulty in assessing the transport error given the paucity of regional observations but also demonstrate the power of top-down assessments given improvements in regional transport modeling and density of observations.

4 Results & Discussion

4.1 Impact of Seasonality on Evaluation Region

As shown in Figure 2, we find strong seasonality in the footprint percentile contours, in agreement with previous analysis of Miyun observations by Wang et al. (2010). At annual timescales, the L_0.90 region is comparable to the WRF d02 extent. Northern China, including Inner Mongolia, dominate the L_0.90 region both seasonally and annually. Due to the heavy biosphere influence in the regional growing season, previous work by Wang et al. (2010) used Miyun non-growing season measurements of CO_2 and carbon monoxide (CO) as an anthropogenic tracer to estimate combustion efficiency for China. When compared to bottom-up estimates of national combustion efficiency, observations suggested 25% higher combustion efficiency than bottom-up estimates of national combustion efficiency; however, Wang et al. (2010) note that the regional (Northern China) and seasonal (winter) subsets could contribute to such a discrepancy. The seasonality exhibited in Figure 2 indeed suggests that combustion efficiency estimates derived from non-growing season measurements alone do not represent anthropogenic processes in provinces south of Miyun that are visible in the observations primarily during the growing season. Low emitting regions northwest of Miyun such as Inner Mongolia influence the site more in the fall and winter relative to other seasons. In the spring and summer, higher emitting regions in provinces south of Miyun are more influential. However, non-growing season CO_2 is influenced by often inefficient district heating in the northwest. And, while growing season CO_2 is influenced by intense urban activities from Beijing and other cities to the south, vegetation draws down both background and locally-observed CO_2 significantly (Figure 3a).

Deleted: footprint extent
Deleted: and influence region
Deleted: evaluation
Deleted: evaluation
Deleted: dominate site
Deleted: ;
Deleted: correspond to seasons where the
Deleted: heavily influence the Miyun receptor

4.2 Unscaled Models: Performance at multiple timescales

Table 1. Quantification of model-observation mismatch at hourly timescales averaged over 2005-2009 and pooled by season (W=Winter; Sp=Spring; Su = Summer; F = Fall). We provide Standard Major Axis (SMA) slopes and 95% confidence intervals; R^2 quantities (those > 0.2 are in bold); and mean bias and root mean square error (RMSE) in ppm.

Deleted: for
Deleted: all years
Deleted: .

	SMA Slope (95%CI)				
	All	*W (JFM)*	*Sp (AMJ)*	*Su (JAS)*	*F (OND)*
$\Delta CO_{2,ZHAO+VPRM}$	0.89 (0.88,0.91)	1.0 (1.0,1.1)	0.74 (0.72,0.77)	0.88 (0.84,0.92)	0.92 (0.90,0.95)
$\Delta CO_{2,EDGAR+VPRM}$	0.77 (0.76, 0.78)	0.83 (0.81, 0.86)	0.62 (0.60, 0.65)	0.83 (0.80, 0.87)	0.77 (0.74, 0.79)
$\Delta CO_{2,CDIAC+VPRM}$	0.63 (0.62, 0.64)	0.63 (0.62, 0.65)	0.48 (0.46, 0.50)	0.79 (0.75, 0.82)	0.56 (0.54, 0.58)
	R^2				
	All	*W (JFM)*	*Sp (AMJ)*	*Su (JAS)*	*F (OND)*
$\Delta CO_{2,ZHAO+VPRM}$	**0.49**	**0.56**	**0.26**	**0.22**	**0.56**
$\Delta CO_{2,EDGAR+VPRM}$	**0.47**	**0.55**	**0.21**	0.18	**0.55**

$\Delta CO_{2,CDIAC+VPRM}$	**0.43**	**0.55**	0.17	0.13	**0.54**
	Mean Bias (RMSE), ppm				
	All	*W (JFM)*	*Sp (AMJ)*	*Su (JAS)*	*F (OND)*
$\Delta CO_{2,ZHAO+VPRM}$	0.32 (9.2)	0.014 (7.9)	-0.033 (8.3)	3.1 (11)	-1.1 (9.7)
$\Delta CO_{2,EDGAR+VPRM}$	-2.0 (9.3)	-2.2 (7.7)	-1.9 (8.7)	0.25 (10.8)	-3.4 (10.1)
$\Delta CO_{2,CDIAC+VPRM}$	-3.3 (9.9)	-3.1 (8.1)	-3.3 (9.2)	-1.1 (11.3)	-5.0 (11.1)

We evaluate unscaled model performance relative to observations at hourly, seasonal, and annual timescales. While inventory scaling is performed at the policy relevant scales of seasons and years, examination of the models at shorter timescales provides insight into model bias and error aggregation at longer timescales. Table 1 summarizes hourly model bias across all years and pooled by season.

All modeled hourly quantities include the same biological component from VPRM, background concentrations, and transport model such that the only source of variation among models is the anthropogenic inventory. With a few exceptions that are discussed in the following sections, $CO_{2,EDGAR+VPRM}$, $CO_{2,CDIAC+VPRM}$, $\Delta CO_{2,EDGAR+VPRM}$, and $\Delta CO_{2,CDIAC+VPRM}$ systematically underestimate observations as indicated by larger deviation below the 1:1 line in the comparison of modeled to measured ΔCO_2 (Table 1, Figure 3b-d.)

Deleted: from

4.2.1 Hourly

We examine the distribution of modeled-measured residuals at hourly timescales for each anthropogenic inventory. While standard deviations are consistent across all models of CO_2 flux (1σ=9ppm; Figure 3.e-g) $\Delta CO_{2,ZHAO+VPRM}$ exhibits the least bias relative to observations with a mean residual of 0.32(0.12,0.53) ppm. In contrast, $\Delta CO_{2,EDGAR+VPRM}$ and $\Delta CO_{2,CDIAC+VPRM}$ display significantly greater bias by typically underestimating observations by large amounts: -2.0(-1.8,-2.2) ppm and -3.3(-3.1,-3.5) ppm, respectively. Here, the 95% confidence intervals are derived from a two-sample t-test. The EDGAR and CDIAC underestimation of ΔCO_2 at the hourly scale is consistent across longer timescales of seasons and years as discussed in the following sections, but we note where there are likely aliased effects of the uncertainty in the VPRM biogenic component.

Deleted: aggregates

Deleted: t

4.2.2 Seasonal

The seasonally averaged modeled and measured ΔCO_2 values shown in Figure 4 illustrate the overall biases for the four inventories. Outside of June, July, August, and September, the anthropogenic signal dominates in northern China (Wang et al., 2010). We see from Table 1 that during seasons where biological activity is lower or significantly lower than anthropogenic activity, there is a consistent

discrepancy among the CO_2 modeled by the three different anthropogenic inventories suggesting systematic differences largely attributable to the anthropogenic component (as we do not vary any other component) . In the fall, where respiration is the dominant biological process, all three modeled

Formatted: Subscript

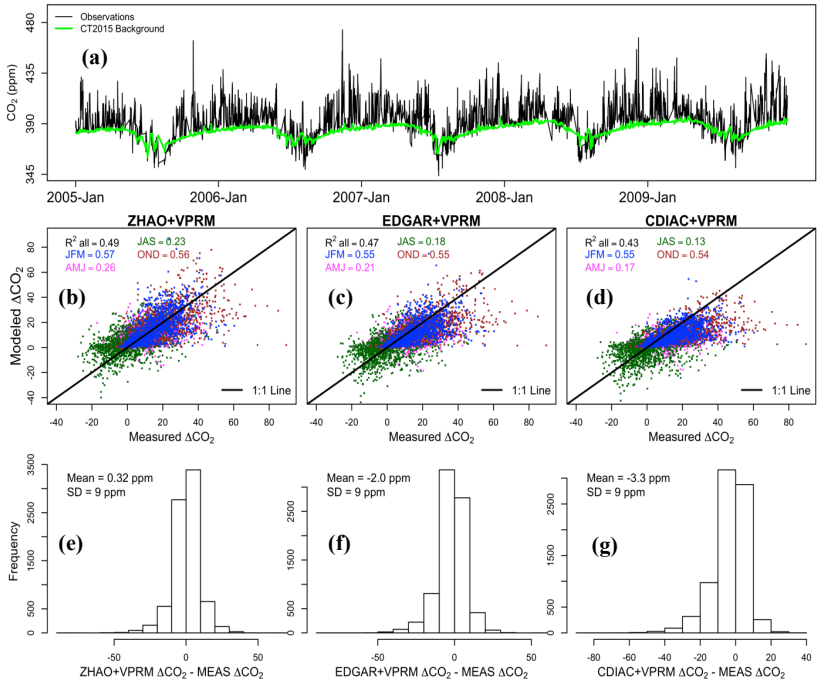


Figure 3. Hourly (1100 to 1600 Local Time) Modeled and Measured CO_2 and ΔCO_2. Measured CO_2 and modeled CT2015 background concentrations are displayed in (a). Modeled versus measured ΔCO_2 for each anthropogenic inventory is shown in (b)-(d), colored by season. Histograms of modeled-measured residuals are shown in (e)-(g). The VPRM vegetation component is included in all modeled ΔCO_2 values.

quantities are consistently lower than observations—a likely a consequence of the known underestimate of ecosystem respiration by the VPRM (Dayalu et al., 2018). Even so, China's significant anthropogenic

component still dominates during these months. During the winter season, where all biospheric activity is at a minimum, the model-observation mismatch is most reflective of biases among anthropogenic inventories rather than aliased impacts from the VPRM. As shown in the winter data in Table 1, ZHAO displays the least bias relative to observations (0.01ppm) followed by EDGAR(-2.2ppm) and CDIAC (-3.1ppm).

With the exception of the peak JAS growing season, $\Delta CO_{2,EDGAR+VPRM}$ and $\Delta CO_{2,CDIAC+VPRM}$ typically underestimate $\Delta CO_{2,OBS}$, even within the 95% uncertainty bounds. The VPRM has a limited calibration network that contributes to an underestimate of regional CO_2 drawdown during the growing season (Dayalu et al., 2018). Therefore, while $\Delta CO_{2,ZHAO+VPRM}$ agrees within 95% confidence bounds with $\Delta CO_{2,OBS}$ during the non-growing seasons, $\Delta CO_{2,ZHAO+VPRM}$ generally overestimates CO_2 concentrations in the growing season (Figure 4a). $\Delta CO_{2,EDGAR+VPRM}$ (Figure 4b) and $\Delta CO_{2,CDIAC+VPRM}$ (Figure 4c) display lower CO_2 concentrations and generally result in better agreement with observations during the peak growing season than at other times of the year; however, our wintertime and overall analysis at hourly timescales (Figure 4, Table 1) suggests this is an artifact of lower anthropogenic emissions

Deleted: sparse

Deleted: , lea

Deleted: ding

Deleted:

Deleted: based on

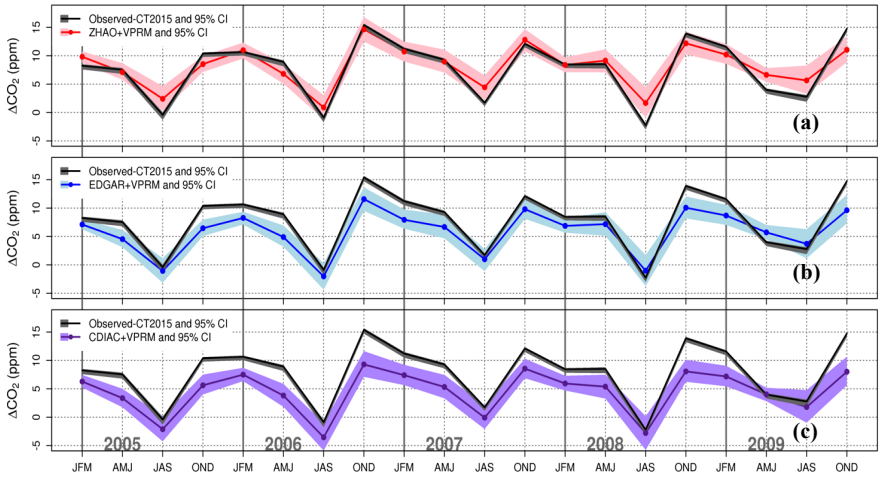


Figure 4. Modeled and Measured Seasonal ΔCO_2. CT2015 background is subtracted from observations to provide observed ΔCO_2 (black line). 95% confidence bounds are derived from bootstrapping hourly afternoon concentrations for each season.

estimates relative to ZHAO that counteracts the VPRM underestimating drawdown. Even during the growing season, $\Delta CO_{2,CDIAC+VPRM}$ agrees with observations typically at its upper confidence limits. However, during times of the year where the impacts of underestimated respiration become more significant (e.g., Fall) it is possible that the seemingly better agreement of ZHAO+VPRM is linked to a counteracting effect of overestimated anthropogenic emissions.

As ZHAO+VPRM demonstrates the least bias relative to observations at hourly and seasonal scales, we model the relative contributions to the monthly signal during the May through September peak regional growing season as defined by Wang et al. (2010). Figure 5 displays the results from partitioning the mean monthly $\Delta CO_{2,ZHAO+VPRM}$ signal as a multi-year average into anthropogenic and vegetation contributions. While the WRF-STILT-VPRM framework has been successfully adapted for similar CO_2 inventory evaluation studies in North American regions where biogenic fluxes dominate surface processes (Karion et al., 2016; Matross et al., 2006), Figure 5 shows the relative magnitude of biogenic fluxes and anthropogenic emissions in the Northern China region is comparable during peak summer, making it difficult to independently constrain them with observational data. As noted in Sect. 3, the regional peak uptake during the growing season occurs with the onset of the corn growing season around July and August. The atypical lower uptake during June represents the winter wheat/corn transition period. These results are consistent with the biological component estimated by Turnbull et al. (2011). Furthermore, knowledge of the relative contribution of vegetation and anthropogenic processes

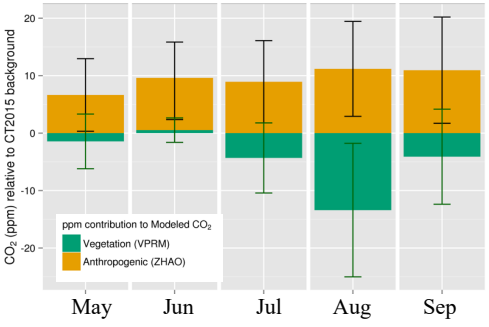


Figure 5. Modeled mean monthly contribution (ppm) to Miyun CO_2 concentrations from vegetation (VPRM) and anthropogenic (ZHAO) sources. Enhancement and depletion are relative to advected CT2015 background concentrations during the regional growing season (MJJAS), averaged over 2005 to 2009. Vertical lines represent 1-σ of monthly averages (Green: Vegetation; Black: Anthropogenic). Negative values represent depletion from CT2015 background; positive values represent enhancement of CT2015 background.

Formatted: Normal

Deleted: <object>

Deleted: of

Deleted: regional

Deleted: does not have a typical pattern in that peak uptake

Deleted: around July/August

to the CO_2 signal during the peak growing season is necessary to interpret satellite retrievals of CO_2 over the region (Dayalu et al., 2018).

4.2.3 Annual

Aggregation of uncertainty and anthropogenic inventory biases at shorter timescales becomes most apparent at the annual timescales. For annual budgeting we follow the assumptions of Piao et al. (2009) and Jiang et al. (2016) that agricultural systems are in annual carbon balance because crop biomass has a short residence time. In the absence of data on regional transfer of agricultural products and proportion of grains used in situ for livestock vs. human consumption in China this is the most conservative assumption to make. Given the dense population in most of Beijing province we expect there may be net import of agricultural products from outside the L_0.90 region, which would show up as additional respiration not captured by VPRM, but that term will be small relative to the anthropogenic CO_2 (Figure 5) (Dayalu et al., 2018). Therefore, while the VPRM is implicitly included in the modeled annual CO_2 and ΔCO_2, vegetation carbon stocks (including harvested products and crop residues) portions of the L_0.90 region with widespread agriculture largely turn over such that only the anthropogenic inventories dominate the modeled CO_2 signal. We evaluate annual CO_2 including

Deleted: ¶
¶

Deleted: -

Deleted: influence

Deleted: influence

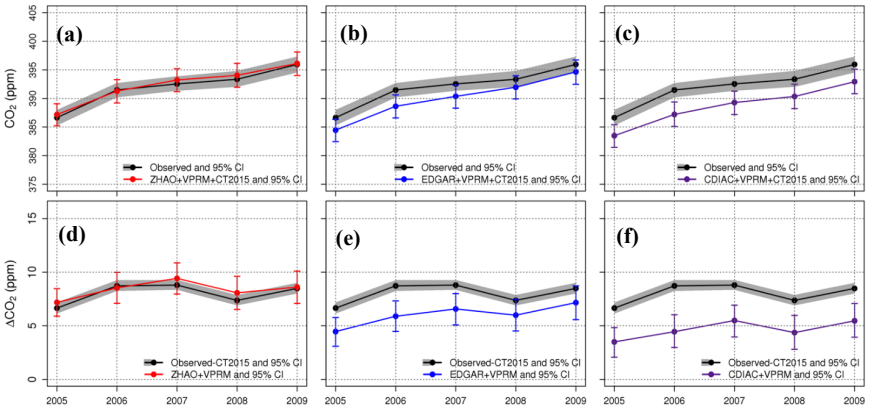


Figure 6. Mean annual CO_2 and ΔCO_2 over entire study time period. (a-c) CO_2 annual concentration; (d-f) ΔCO_2 (regional enhancement, after removal of advected CT2015 background) with bootstrapped 95% confidence intervals.

CT2015 background (Figure 6a-c) and as regional enhancement relative to background (Figure 6d-f). We show that for all years, $CO_{2,ZHAO+VPRM}$ and $\Delta CO_{2,ZHAO+VPRM}$ agree tightly within 95% uncertainty to observations (Figure 6a, Figure 6d). EDGAR+VPRM and CDIAC+VPRM are consistently biased significantly lower than observations.

4.3 Evaluation of inventories at seasonal and annual timescales

We quantify model-observation mismatch by estimating the additive flux corrections at seasonal timescales and multiplicative corrections at annual timescales. We emphasize that these "corrections", or scalings, are not optimizations; rather, they simply reflect the extent to which the individual anthropogenic+VPRM flux models deviate from the observations. Complete seasonal and annual scaling results are provided in the SI Sect. S5, and Tables S2-S3.

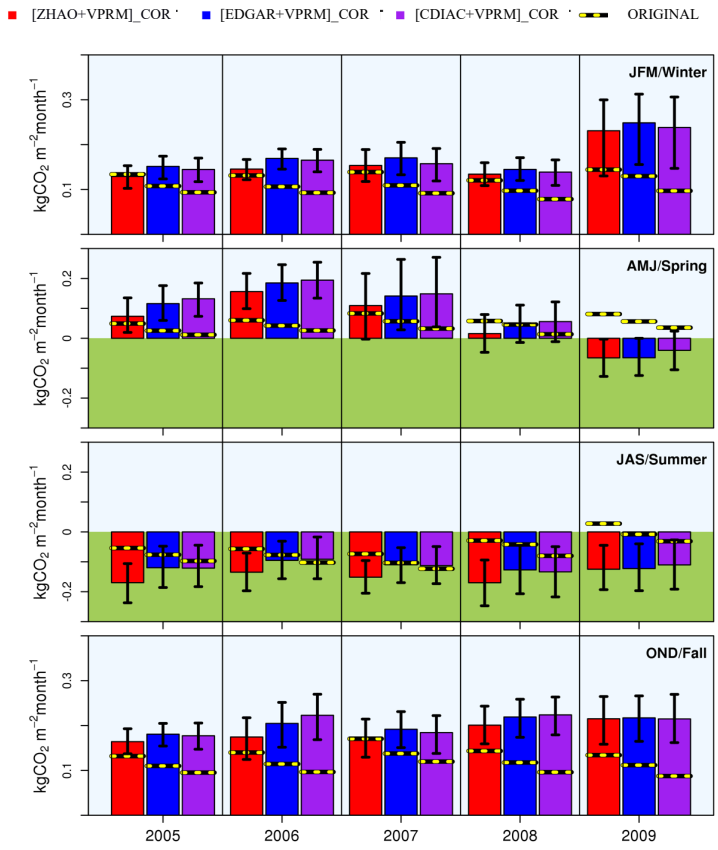


Figure 7. Scaled Seasonal Fluxes in the L_0.90 region (kg CO_2 m^{-2} $month^{-1}$). Anthropogenic and vegetation inventories are scaled together ([ANTH+VPRM_COR]). Black and yellow dashed line is the seasonal flux estimated by the original ANTH+VPRM model. All models have the same vegetation component (VPRM) and differ only in the anthropogenic inventory source. Shaded green represents negative flux (uptake by biosphere). Scaling based on additive corrections; difference among scaled inventories is due to differing spatial allocations by anthropogenic inventories. Bootstrapped 95% confidence intervals are represented by the black vertical lines.

The observational record informing the scaling integrates the biological and anthropogenic signals. At the seasonal scale, where biological processes are significant contributors to the signal, we scale the sum of the anthropogenic and biological fluxes (Figure 7). Scaled non-growing season flux estimates

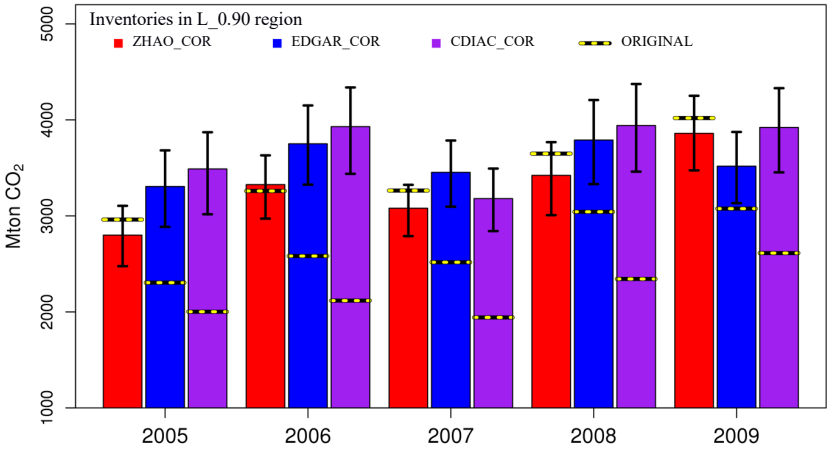


Figure 8. Annually scaled emissions in L_0.90 region. Scaling is based on multiplicative scaling factors. Difference among scaled inventory means is due to differing spatial allocations in original anthropogenic inventories. Bootstrapped 95% confidence intervals are represented by the black vertical lines. **Note the y-axis origin begins at 1000 Mton CO_2 for visual clarity.*

Deleted: for 90th percentile of influence

are higher than unscaled values, partially accounting for the VPRM generally underestimating ecosystem respiration by an additive offset throughout the year (Dayalu et al., 2018). The multi-year seasonal results in Table 1 suggest that this offset can aggregate to a 1-2ppm difference; the result would be a shift in baseline rather than overall pattern for each of the three simulations. As the vegetation and all other components are controlled across models, the inter-model variance reflects the relative performance of the anthropogenic estimates. We find that in the non-growing months the original ZHAO+VPRM inventory typically remains within the 95% confidence bounds of the scaled inventory. However, both EDGAR+VPRM and CDIAC+VPRM are consistently significantly lower than their scaled counterparts. At least in the winter, where biogenic processes are at a minimum, this suggests that both EDGAR and CDIAC underestimate anthropogenic emissions, and that ZHAO estimates are closer to actual emissions. Improved representation of temporal anthropogenic activity factors and biosphere processes are needed to extend the conclusions of anthropogenic inventory performance to all

Deleted: is

Deleted: T

Deleted: implies

seasons. In the absence of such data, it is not possible to conclusively state whether model-data mismatch is rooted in anthropogenic emissions biases or biogenic biases. During the growing seasons, however, the afternoon vegetation signal is significant, and the picture is more complex. In the spring, the CO_2 signal at Miyun is significantly affected by the North China Plain winter wheat growing season. The effect of scaling in the spring from 2005 to 2007 is to increase CO_2 emissions with a net positive seasonal flux; however, in 2008 and 2009 we find the net seasonal flux becomes negative such that uptake dominates emissions. The prior models in all cases predict positive flux. During the summer months, ZHAO+VPRM predicts more emissions and/or less uptake relative to EDGAR+VPRM and CDIAC+VPRM. Scaling of summertime fluxes serves to significantly increase ZHAO+VPRM uptake estimates; the EDGAR+VPRM and CDIAC+VPRM prior estimates are within the 95% confidence bounds of the scaling for reasons discussed previously.

Table 2. Annual scaling factors (95% CI) and corresponding corrected emissions for L_0.90 inventory evaluation region.

		Scaling Factor (95% CI)	Corrected Emissions, $MtCO_2$ (95% CI)	Original emissions, $MtCO_2$
2005	**ZHAO**	0.95 (0.84, 1.0)	2800 (2476, 3105)	3015
	EDGAR	1.4 (1.3, 1.6)	3306 (2886, 3683)	2322
	CDIAC	1.7 (1.5, 1.9)	3489 (3017, 3871)	1930
2006	**ZHAO**	1.0 (0.91, 1.1)	3326 (2972, 3631)	3273
	EDGAR	1.5 (1.3, 1.6)	3751 (3325, 4150)	2586
	CDIAC	1.9 (1.6, 2.0)	3930 (3438, 4338)	2160
2007	**ZHAO**	0.94 (0.85, 1.0)	3080 (2789, 3324)	3588
	EDGAR	1.4 (1.2, 1.5)	3454 (3096, 3785)	2799
	CDIAC	1.6 (1.5, 1.8)	3180 (2842, 3493)	2260
2008	**ZHAO**	0.94 (0.82, 1.0)	3422 (3008, 3768)	3685
	EDGAR	1.2 (1.1, 1.4)	3790 (3332, 4207)	3095
	CDIAC	1.7 (1.5, 1.9)	3941 (3461, 4374)	2395
2009	**ZHAO**	0.96 (0.86, 1.1)	3860 (3474, 4251)	3974
	EDGAR	1.1 (1.0, 1.3)	3518 (3133, 3874)	3298
	CDIAC	1.5 (1.3, 1.7)	3921 (3454, 4330)	2543

We report annual scaled anthropogenic inventories in the L_0.90 region in Fig. 8 and Table 2 as $MtCO_2yr^{-1}$. As discussed previously, the annual scalings are applied only to the anthropogenic inventory, as the signal at the annual timescale is effectively dominated by anthropogenic emissions; net ecosystem fluxes are expected to be relatively minor in the L_0.90 region in comparison. For all years, the emissions estimated by the original ZHAO inventory lie within the 95% confidence bounds of the scaled ZHAO inventory. However, for EDGAR and CDIAC, the original inventories consistently underestimate observations. Averaged over the five-year study period, EDGAR and CDIAC lead to modeled estimates of CO_2 mixing ratios that are typically lower than observations by 30% and 70%

Deleted:
Deleted: significant
Deleted:
Deleted: at
Deleted: extent

respectively (Fig. 6). Averaged across the five years, this translates to EDGAR and CDIAC being scaled relative to their unscaled values in the L_0.90 region by 1.3 and 1.7, respectively (Fig. 8; Table 2). In the case of EDGAR, we note a general increase in observational agreement from 2005 to 2009.

4.4 Potential Contributions to Regional Carbon Emissions Patterns from 2005 to 2009

We examine the statistical significance of the inter-annual observed concentration and enhancement differences using a two-sample t-test (Table 3). The observed concentrations including advected global background (Figure 6, top row) display an overall increasing trend of 1.87 (1.8, 1.9) ppm CO_2 yr^{-1} between 2005 and 2009, in agreement with flask samples obtained from nearby WMO sites between 2007 and 2010 (Liu et al., 2014). The inter-annual increases are statistically significant (Table 3). However, when we remove the modeled background to more closely examine regional patterns that would otherwise be drowned out by the global signal, we find that the regional ΔCO_2 trend (Figure 6, bottom row; Table 3) does not parallel the increasing global CO_2 trend (Figure 6 top row; Table 3). Regionally, the observed enhancements increase from 2005 to 2006 and plateau in 2007 before decreasing in 2008. Regional ΔCO_2 increases again in 2009. Earlier work by Wang et al. (2010) extended the Miyun observations of CO_2 growth rate to all of China and estimates a lower CO_2 growth rate than previously suggested. However, Figure S6 suggests local reductions in regions influencing Miyun, possibly in preparation for the Beijing Olympics, are partially offset by increases elsewhere. A larger network of sites would be needed to quantify this further in order to evaluate the CO_2 growth rate for other regions in China and for China as a whole.

In Figure 9a we estimate Gross Regional Product (GRP) for eight of China's 34 provincial-level administrative units, specifically those encompassed significantly by the L_0.90 region: Beijing, Tianjin, Henan, Shanxi, Shandong, Hebei, Inner Mongolia, and Liaoning. Using data from the International Monetary Fund (IMF; https://www.imf.org/en/Data) and World Bank (World Bank, 2017, we retrieved the GDP for each of the above provinces and summed them to estimate the GRP. GDP calculations are inherently uncertain and were available as single values for each province per year. A more extensive economic analysis to estimate uncertainty of these values is beyond the scope of this study. Key economic events occurred during the study time period and are likely contributors to the observed interannual variation in regional CO_2 emissions (Figure 6d-e) and a doubling of GRP from 2005 to 2009 (Figure 9a). In particular, the time period from 2005-2009 saw industrial energy efficiency improvements beginning in 2007 under the 11th FYP; preparations for and staging of the 2008 Beijing Summer Olympics; the global financial crisis in late 2008; and a large Chinese fiscal stimulus in 2009. We further note that the global financial crisis of 2008 correlates with a plateauing of the percentage contribution of northern China GRP to national GDP (Figure 9a).

Deleted: ¶

Deleted: Patterns in

Deleted: two

Deleted: (

Deleted:)

Deleted: , bottom row

Deleted: Enhancements

Deleted: ¶
¶
In Figure 9a we estimate Gross Regional Product (GRP) for eight of China's 34 provincial-level administrative units, specifically those encompassed significantly by the L_0.90 influence contour: Beijing, Tianjin, Henan, Shanxi, Shandong, Hebei, Inner Mongolia, and Liaoning. We suggest that industrial energy efficiency improvements beginning in 2007 under the 11th FYP, preparations and staging of the 2008 Beijing Summer Olympics, and the global financial crisis in late 2008 followed by a large Chinese fiscal stimulus in 2009 are likely contributors to the observed interannual variation in regional CO_2 emissions (Figure 6d-e) while also compatible with a doubling of GRP from 2005 to 2009 (Figure 6a). In addition, e

Deleted: extends

Deleted: uncertain, and

Table 3. Inter-annual observed CO_2 and ΔCO_2 differences. Differences are of observations between consecutive years. 95% confidence intervals are derived from a two-sample t-test. Italicized entries denote instances where the inter-annual difference is not statistically significant (confidence interval includes zero).

Time Interval (y_2-y_1)	$CO_{2,OBS}$ (ppm) Mean Difference (95% CI)	$\Delta CO_{2,OBS}$ (ppm) Mean Difference (95% CI)
2006-2005	4.86 (4.5, 5.2)	2.08 (1.9, 2.3)
2007-2006	1.08 (0.69, 1.5)	*0.0693 (-0.15, 0.29)*
2008-2007	0.772 (0.37, 1.2)	-1.43 (-1.6, -1.2)
2009-2008	2.60 (2.2, 3.0)	1.12 (0.88, 1.4)
2009-2005	9.31 (8.9, 9.7)	1.84 (1.6, 2.0)

As policy targets are often measured as relative changes over multiple years, an important component of emissions inventories is their ability to accurately capture multi-year changes. Observations indicate enhancements above background CO_2 increased by 28% (22%, 34%) between 2005 and 2009. ZHAO+VPRM estimates a 20% increase over the same time period while EDGAR+VPRM and CDIAC+VPRM estimate 61% and 56% increases respectively.

4.5 Implications for Assessing National Carbon Emission Targets

China has pledged a 60-65% reduction in carbon intensity by 2030 and has additionally set a benchmark of 40-45% reduction in carbon intensity by 2020, where both targets are relative to the baseline year 2005 (NDRC, 2015; Guan et al., 2014). However, Guan et al. (2014) found that provincial trends in carbon intensity can vary significantly from national trends. Using the GRP values shown in Figure 9a, we calculate a Northern China regional carbon intensity incorporating the eight provinces encompassed significantly by the L_0.90 region (Figure 9c). We also estimate an L_0.90 regional carbon intensity based on the official national energy-related CO_2 emissions in NDRC (2015); we scale the national total by 39% (35%,42%) which is the mean (range) contribution of the L_0.90 region to the national emissions in 2005, averaged across the three unscaled gridded emissions inventories. We emphasize that carbon intensity values are inherently uncertain due to complexities in GRP and Gross Domestic Product (GDP) calculations such as double-counting due to inter-provincial trade or spatial mismatch between emissions and economic data. Nevertheless, the analysis provides valuable insight into trends rather than precise values.

Over the study time period, the GRP of the L_0.90 region more than doubled (Figure 9a), exhibiting a moderate, positive correlation with the increasing trend in emissions (Figure 9b). Coinciding with the 2008 Beijing Summer Olympics, the region's contribution to China's GDP grew from approximately 13.5% in 2007 to nearly 16% in 2008, representing a 20% increase, before plateauing into 2009 (Figure

Deleted: 4

Deleted: (Figure 9b). The

Deleted: are those that are

Deleted: influence contour

Deleted: : Beijing, Henan, Shanxi, Tianjin, Shandong, Hebei, Inner Mongolia, and Liaoning

Deleted: , evidently

Deleted: ed to a significant increase in emissions

9a). As noted in Guan et al. (2014), reductions in carbon emissions intensity can come about via two main pathways: the first, within industries, through increased energy efficiency combined with expanded production capacity; the second, across the economy, through structural shifts from energy-intensive industrial sectors to service sectors. The doubling of GRP with the apparent reduction in regional carbon intensity suggests a combination of enlarged production capacity (including production of higher valued goods) and a shift toward service-oriented economy. In the former instance, a larger production capacity tends to reduce the overall energy (and, therefore, carbon) consumption of a single production unit. In the latter instance, the energy consumption by the service sector is considerably lower than that required by industrial and manufacturing processes. In the northern China region, however, industry continues to dominate the economy suggesting that carbon intensity reductions are more due to enlarged production capacity. From 2005 to 2009, carbon intensity for the L_0.90 region decreased by 47% (28%,65%), based on a one-sample t-test of pooled emissions intensity changes across scaled inventories. Analysis presented by organizations such as the World Bank (World Bank, 2017) suggests China's carbon intensity at the national level decreased by 20% in 2009 relative to 2005. However, we note that the carbon emissions data source for the World Bank carbon intensity calculations is CDIAC. We have shown that at least for the L_0.90 region, CDIAC emissions lead to significant underestimates of observations. Our work here suggests that carbon accounting organizations such as the World Bank would benefit from basing their national estimates for China on a variety of inventories, incorporating increasingly available China-specific approaches (including but not limited to MEIC and PKU), EDGAR, and newer global inventories such as ODIAC. However, we emphasize a crucial point with respect to the value of carbon intensity targets, in agreement with Guan et al. (2014): carbon intensity targets are especially misleading in developing countries where absolute emissions continue to significantly grow in concert with economic development goals. We see that despite the decreasing carbon intensity of the region, pooled emissions estimates from the three scaled inventories suggest an 18% increase in absolute emissions from 2005-2009 (Table 2, Figure 9b). In terms of the climate impact, it is the absolute carbon emissions rather than the carbon intensity that ultimately matters.

Deleted:

Deleted: as a driver for regional carbon intensity reductions

Deleted: /Northern China

Deleted: *<object>*

Deleted: (yet to be tested with observations in China)

930

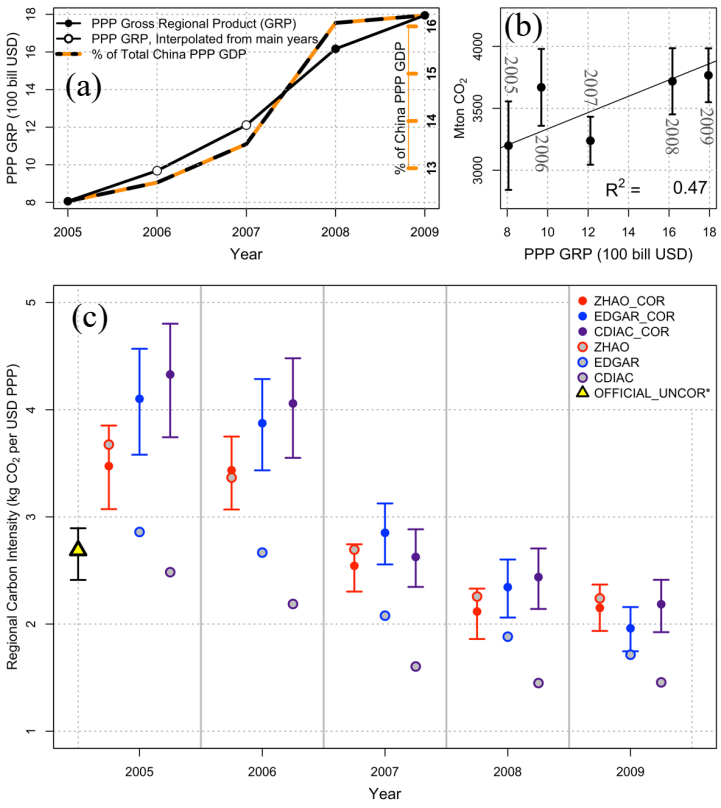


Figure. 9. Estimates of Regional Carbon Intensity (kg CO_2 USD_{PPP}^{-1}). (a) PPP GRP by year and as a % of China's national GDP. No PPP GRP values were available for 2006 and 2007; PPP GRP for these years was derived from linearly interpolated ratio of Nominal GRP/PPP GRP for 2005, 2008, and 2009. (b) Correlating corrected regional emissions from Table 2 with PPP GRP; values are pooled annual means among ZHAO, EDGAR, and CDIAC with 1-σ error bars. (c) Regional Carbon Intensity using scaled (solid) and unscaled (grey) CO_2 estimates. Error bars are bootstrapped 95% confidence intervals. GRP, GDP data from IMF and World Bank. Provinces used in GRP calculation are those significantly encompassed by L_0.90 region Beijing, Henan, Shanxi, Tianjin, Shandong, Hebei, Inner Mongolia, and Liaoning. **Estimated by scaling the official national emissions total by the average contribution (39%) of L_0.90 region to total emissions in 2005. Uncertainty bars represent the % contribution range estimated by ZHAO, EDGAR, and CDIAC in 2005 (35%, 42%).*

Formatted: Subscript
Deleted: instead calculated
Deleted: by
Deleted: interpolating
Formatted: Font: Symbol
Deleted: b
Deleted: i
Formatted: Subscript
Deleted: Uncertainty
Deleted: , World Bank, China Statistical Yearbook
Deleted: contour:
Formatted: Font: Italic

5 Conclusions

Continuous hourly CO_2 observations, significantly influenced by the heavily CO_2-emitting Northern China region, are used in a top-down evaluation and scaling of three bottom-up CO_2 flux inventories. We focus on the policy-relevant time interval from 2005 to 2009, noting that 2005 is China's baseline year for carbon commitments. The three inventories are distinct in their anthropogenic component, with a common biogenic flux component provided by the VPRM, a simple satellite data-driven biosphere model calibrated with ground-level ecosystem observations. The ZHAO anthropogenic emissions inventory incorporates a regional approach to China's CO_2 emissions estimation, using activity data at the provincial and facility-levels as well as domestic emission factors. The EDGAR and CDIAC emissions inventories incorporate a greater reliance on global averages and China's national statistics and international default emission factors, and depend more heavily on proxies (e.g., population) to allocate the emissions geographically. The three anthropogenic inventories represent a range of methods used to estimate emissions for China.

The Northern China administrative region, excluding Inner Mongolia, dominates the L_0.90 region which is the region over which we distribute the model-observation mismatch (Figure 2). We find strong seasonality in the L_0.90 region, : the northwest features more strongly in the non-growing season and there is a more symmetric influence in the growing season. Within the L_0.90 region, EDGAR and CDIAC are—on average across the five study years—lower than ZHAO by 20% and 36%, respectively. Across administrative regions, the highest discrepancy between the global and regional inventories is in Northern China, where the ZHAO inventory estimates emissions that are on average 30% higher than both EDGAR and CDIAC (SI, Table S1).

We find the ZHAO+VPRM inventory generally agrees very closely with observations, often significantly better than the nationally referenced inventories at all timescales (hourly through annually), 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. At annual timescales, the anthropogenic signal dominates, and we find that emission rates from EDGAR and CDIAC lead to underestimated emissions in the Northern China region by an average of 30% and 70%, respectively, averaged across all study years. We note that the discrepancy between the EDGAR-based timeseries and the observations generally decreases over the five-year study period. In contrast, emission rates from the ZHAO inventory gives *a priori* results very close to observations throughout and is not significantly affected by the scaling: the error bars for the scaled estimates consistently include the original estimate. Note that the EDGAR and CDIAC inventories can differ from -10% to -20% relative to ZHAO in their national emissions totals (Table S1). The inventories evaluated here exhibit distinct differences in their

Moved down [2]: Despite the limitations of having data from a single site, this analysis demonstrates how a long time series of continuous observations can identify apparent overall biases in some inventories. Our results, while specific to northern China regional emissions in particular, also provide some insight into current methods of carbon emissions accounting for China as a whole. We do, however, wish to summarize multiple caveats and limitations of our study that have been presented throughout the text. First, we emphasize that this work is intended to be a comparison of emission rates from a subset of anthropogenic CO_2 inventories over northern China that were readily available at the time this research began and is not intended to be an advocate or criticism of any single published inventory. Rather, we use a long observational record to examine model-data mismatch in an important carbon emitting region where local data is difficult to access and global datasets are forced to rely on the best available public data which are not necessarily accurate assumptions of China-specific activity. Second, while we recognize the height limitations –and therefore the footprint—of the Miyun receptor its topographic advantage along with the low-productivity vicinity, make it similar to other short-tower sites suitable for regional analysis. In addition, addressing the significant uncertainty stemming from transport error and error in spatial allocation of the emissions remains a challenge. Independent verification from concurrent aircraft measurements (for example) or multi-level inlet locations were not available to quantify the impact of absolute and relative inlet location on transport uncertainty. In this study, the drawback of a single location is offset somewhat by the long 60-month timeseries. Absent a dense network of observations, a more sophisticated and extensive error analysis than what was provided cannot be conducted with meaningful results. Finally, we emphasize our implied "corrections", or scalings, of modeled CO_2 relative to observations are not optimizations; rather, they simply reflect the extent to which the individual anthropogenic+VPRM CO_2 flux models deviate from the observations. Effectively evaluating and constraining inventory emissions rates at relevant spatial scales requires multiple stations of high-temporal resolution observations.¶

Deleted: 4.5 Summary of study caveats and limitations¶

Deleted:

Deleted: to

Deleted: attribute

Deleted: footprint extent and influence

Deleted: with

Deleted: t

Deleted: dominating

Deleted: a

Deleted: uniform

Deleted: The Northern China administrative region, excluding Inner Mongolia, dominates the L_0.90 influence region (Figure 2).

Deleted: inventory evaluation

Deleted: dominates

ability to match observations. However, observational data from a network of sites strategically located in and around the eastern half of China would be required to (1) examine whether differences in spatial allocation approaches contribute to differences among the inventories and (2) conduct actual optimizations of the inventories.

We find that carbon intensity in the region has decreased by 47%(28%, 65%) from 2005 to 2009, from approximately 4kgCO_2/USD_{PPP} in 2005 to about 2kgCO_2/USD_{PPP} in 2009 (Figure 9c). However, we see that despite the decreasing carbon intensity of the region, there is an 18% increase in absolute emissions over time, affirming the point made by Guan et al. (2014) that meeting carbon intensity targets in emerging economies can be at odds with making real climate progress (Table 2, Figure 9b).

Despite the limitations of having data from a single site, this analysis demonstrates how a long time series of continuous observations can identify apparent overall biases in some inventories. Our results, while specific to northern China regional emissions in particular, also provide some insight into current methods of carbon emissions accounting for China as a whole. We emphasize that this work is intended to be a comparison of emission rates from a subset of anthropogenic CO_2 inventories over northern China that were readily available at the time this research began and is not intended to be an advocate or criticism of any single published inventory. Rather, we use a long 60-month continuous observational record to examine model-data mismatch in an important carbon emitting region where local data is difficult to access and global datasets are forced to rely on the best available public data, which are not necessarily accurate assumptions of China-specific activity. Second, while we recognize the height limitations –and therefore the footprint—of the Miyun receptor its topographic advantage along with the low-productivity vicinity, make it similar to other short-tower sites suitable for regional analysis. In addition, a detailed assessment of uncertainty stemming from errors in transport, biogenic inventories, and inventory spatial allocation remains a challenge. Independent verification from concurrent aircraft measurements (for example) or multi-level inlet locations were not available to quantify the impact of absolute and relative inlet location on transport uncertainty. Finally, we emphasize our implied seasonal and annual "corrections", or scalings, of modeled CO_2 relative to observations are not optimizations; rather, they simply reflect the extent to which the individual anthropogenic+VPRM CO_2 flux models deviate from the observations. At least in the winter, where biogenic processes are at a minimum, the low bias of ZHAO-modeled CO_2 concentrations suggests the ZHAO inventory is closer to actual emissions. However, improved representation of temporal anthropogenic activity factors and biosphere processes are needed to extend the conclusions of anthropogenic inventory performance to all seasons. Effectively evaluating and constraining inventory emissions rates at relevant spatial scales requires multiple stations of high-temporal resolution observations, as well as improvements and greater diversity in observationally-constrained biogenic flux models. In its current configuration, the single biogenic flux model precludes a comprehensive multi-seasonal and annual disentangling of contributions to CO_2; particularly in our annual scale analysis, we are ascribing more uncertainty to the anthropogenic inventories over the biogenic contributions. Absent data from a dense network of ecosystem flux and atmospheric measurements, there will constantly be a tradeoff between drawing

Formatted: Subscript

Formatted: Subscript

Moved (insertion) [2]

Deleted: We do, however, wish to summarize multiple caveats and limitations of our study that have been presented throughout the text. First, w

Deleted: In addition

Deleted: addressing the significant uncertainty

Deleted: error

Deleted: and

Deleted: error in

Deleted: of the emissions

Deleted: In this study, the drawback of a single location is offset somewhat by the long 60-month timeseries. Absent a dense network of observations, a more sophisticated and extensive error analysis than what was provided cannot be conducted with meaningful results. …

Formatted: Subscript

Formatted: Subscript

Deleted:

conclusions using low-temporal resolution flask measurements from a few sites and continuous data from a single location.

In situ CO_2 observations interpreted within a high-resolution model framework such as described in this study provide a powerful constraint to test and correct spatially explicit inventories. The observation station available for the 2005-2009 period was strategically located to provide information on one of the highest CO_2 emitting regions of China. Within the limitations described above, the observations provide strong evidence supporting the use of China-specific methods, such as those employed in ZHAO, for China's CO_2 emissions inventory derivation. In future, access to a spatially dense network of measurements will allow for a sophisticated error analysis that can more readily assess uncertainty in key model components such as transport, flux fields, and background concentrations. Along with the results presented here, previous studies (e.g., Turnbull et al., 2011) provide key information that is necessary to guide and motivate more extensive future measurement and emissions evaluation efforts. Such future efforts will benefit substantially from incorporating newly available information from column-average CO_2 concentrations acquired by orbiting instruments or ground-based spectrometers to increase observational coverage. A number of existing (OCO-2, OCO-3) and planned satellite missions will significantly reduce the observational gap in China, though surface observations provide additional constraints and a link to absolute calibration scales. A denser network of CO_2 measurement stations in China is required as a component for effective monitoring, reporting, and verification of regional and national inventories. The results of this research present a necessary baseline for a key CO_2-emitting region of China. Our results have broad implications toward designing future analyses as more observations of China's CO_2 continue to become available, particularly in the era of increased CO_2 satellite coverage. However, as the quality of satellite retrievals can be compromised by factors such as aerosol loading, surface observations continue to be crucial for the region both in their own right and as a key component of cross-platform evaluations.

Deleted: single

Deleted: that

Deleted: Absent data from a dense network of high temporal resolution measurements, there will constantly be a tradeoff between drawing conclusions using low-temporal resolution flask measurements from a few sites and continuous data from a single location. …

Deleted: particular

Deleted: However, despite the dearth of observational data, past

Deleted: and studies such as this one

Deleted: studies

Deleted: F

Formatted: Subscript

Deleted: have

Deleted: in general

Code and Data Availability

Code and data are available through the Harvard Dataverse at https://doi.org/10.7910/DVN/OJESO0. The code and data supplement includes observational and modeled CO_2 time series, WRF and STILT parameter files, and STILT footprint files.

Author Contributions

A.D., J.W.M, and S.C.W designed the research. A.D. performed the research with guidance from all co-authors. Y.W. and J.W.M monitored, maintained, and provided access to the Miyun hourly observational data set. Y.Z. provided the China-specific anthropogenic inventory. WRF-STILT simulations were performed by A.D. with assistance from T.N. A.D. constructed the vegetation CO_2 inventory. A.D. and J.W.M wrote the paper with contributions from all other co-authors.

Competing Interests

The authors declare no competing interests.

Acknowledgments

We acknowledge the Harvard-China Project and the Harvard Global Institute for funding this study. We thank Zhiming Kuang for providing computational resources. We also thank Jenna Samra, Maryann Sargent, and Victoria Liublinska for helpful discussion.

References

Andres, R.J., Boden, T.A., and Marland, G.. Annual Fossil-Fuel CO_2 Emissions: Mass of Emissions Gridded by One Degree Latitude by One Degree Longitude v2016. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A. doi 10.3334/CDIAC/ffe.ndp058.2016, 2016a.

Andres, R. J., Boden, T. A., and Higdon, D. M.: Gridded uncertainty in fossil fuel carbon dioxide emission maps, a CDIAC example, Atmos. Chem. Phys., 16, 14979-14995, doi:10.5194/acp-16-14979-2016, 2016b.

Boden, T.A., Marland, G., and Andres, R. J.: Global, Regional, and National Fossil-Fuel CO_2 Emissions. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A. doi 10.3334/CDIAC/00001_V2016, 2016.

Brown D, Brownrigg R, Haley M, & Huang W (2012) *The NCAR Command Language (NCL) v6.0. 0*. UCAR/NCAR Computational and Information Systems Laboratory, Boulder, CO. Available at http://dx.doi. org/10.5065/D6WD3XH5.

Dayalu, A (2017) "Exploring the Wide Net of Human Energy Systems: From Carbon Dioxide Emissions in China to Hydraulic Fracturing Chemicals Usage in the United States". PhD thesis (Harvard University, Cambridge, MA).

Dayalu, A., Munger, J. W., Wofsy, S. C., Wang, Y., Nehrkorn, T., Zhao, Y., McElroy, M. B., Nielsen, C. P., and Luus, K.: Assessing biotic contributions to CO2 fluxes in northern China using the Vegetation, Photosynthesis and Respiration Model (VPRM-CHINA) and observations from 2005 to 2009, Biogeosciences, 15, 6713-6729, https://doi.org/10.5194/bg-15-6713-2018, 2018.

European Commission, Joint Research Centre (JRC)/Netherlands Environmental Assessment Agency (PBL): Emission Database for Global Atmospheric Research (EDGAR), release EDGARv4.2 FT2010, http://edgar.jrc.ec.europa.eu, Accessed 2013.

Guan, D., Liu, Z., Geng, Y., Lindner, S., and Hubacek, K.: The gigatonne gap in China's carbon dioxide inventories, Nat. Clim. Chg, 2,672–675, doi:10.1038/nclimate1560, 2012.

Guan, D., Klasen, S., Hubacek, K., Feng, K., Liu, Z., He, K., Geng, Y., and Zhang Q.: Determinants of stagnating carbon intensity in China, Nat. Clim. Chg, 4, 1017-1023, doi:10.1038/nclimate2388, 2014.

Hegarty, J., Draxler, R., Stein, A., Brioude, J., Mountain, M., Eluszkiewicz, J., Nehrkorn, T., Ngan, F., and Andrews, A.: Evaluation of Lagrangian Particle Dispersion Models with Measurements from Controlled Tracer Releases, J. Appl. Meteorol. Climatol., 52, 2623-2637, doi: 10.1175/JAMC-D-13-0125.1, 2013.

Karion, A., Sweeney, C., Miller, J. B., Andrews, A. E., Commane, R., Dinardo, S., Henderson, J. M., Lindaas, J., Lin, J. C., Luus, K. A., Newberger, T., Tans, P., Wofsy, S. C., Wolter, S., and Miller, C. E.: Investigating Alaskan methane and carbon dioxide fluxes using measurements from the CARVE tower, Atmos. Chem. Phys., 16, 5383-5398, doi:10.5194/acp-16-5383-2016, 2016.

Kort, E. A., Angevine, W.M., Duren, R., and Miller, C.E.: Surface observations for monitoring urban fossil fuel CO_2 emissions: Minimum site location requirements for the Los Angeles megacity, J. Geophys. Res. Atmos., 118, 1577–1584, doi:10.1002/jgrd.50135, 2013.

Le Quere, C., et al. (2016), Global Carbon Budget 2016, *Earth System Science Data*, *8*(2), 605-649, doi:10.5194/essd-8-605-2016.

Lin, J. C., Gerbig, C., Wofsy, S. C., Andrews, A. E., Daube, B. C., Davis, K. J., and Grainger, C. A.: A near-field tool for simulating the upstream influence of atmospheric observations: The Stochastic Time-Inverted Lagrangian Transport (STILT) model, Journal of Geophysical Research-Atmospheres, 108, 4493, doi:10.1029/2002JD003161, 2003.

Liu, Z., Guan, D., Wei, W., Davis, S.J., Ciais, P., Bai, J., Peng, S., Zhang, Q., Hubacek, K., Marland, G., Andres, R.J., Crawford-Brown, D., Lin, J., Zhao, H., Hong, C., Boden, T.A., Feng, K., Peters, G.P., Xi, F., Liu, J., Li, Y., Zhao, Y., Zeng, N., and He, K. :Reduced carbon emission estimates from fossil fuel combustion and cement production in China. Nature 524, 335–338, 2015.

Mahadevan, P., Wofsy, S.C., Matross, D.M., Xiao, X., Dunn, A.L., Lin, J.C., Gerbig, C., Munger, J.W., Chow, V.Y. and Gottlieb, E.W.: A satellite-based biosphere parameterization for net ecosystem CO2 exchange: Vegetation Photosynthesis and Respiration Model (VPRM), Global Biogeochem. Cycles, 22, GB2005, doi:10.1029/2006GB002735, 2008.

Matross, D. M., Andrews, A., Pathmathevan, M., Gerbig, C., Lin, J. C., Wofsy, S. C., Daube, B. C., Gottlieb, E. W., Chow, V. Y., Lee, J. T., Zhao, C. L., Bakwin, P. S., Munger, J. W., and Hollinger, D. Y.: Estimating regional carbon exchange in New England and Quebec by combining atmospheric, ground-based and satellite data, Tellus Series B-Chemical and Physical Meteorology, 58, 344-358, 2006.

McKain, K., Down, A., Raciti, S. M., Budney, J., Hutyra, L. R., Floerchinger, C., Herndon, S. C., Nehrkorn, T., Zahniser, M. S., and Jackson, R. B.: Methane emissions from natural gas infrastructure and use in the urban region of Boston, Massachusetts Proc. Natl. Acad. Sci. U.S.A., 112 (7) 1941– 1946, 2015.

McKain, K., Wofsy, S.C., Nehrkorn, T., Eluszkiewicz, Ehleringer, J.R., and Stephens, B.B.: Assessment of ground-based atmospheric observations for verification of greenhouse gas emissions from an urban region, Proc. Nat. Acad. Sci., 109(22), 8423–8428, 2012.

Miller, S.M., Kort, E.A., Hirsch, A.I., Dlugokencky, E.J., Andrews, A.E., Xu, X., Tian, H., Nehrkorn, T. Eluszkiewicz, J., Michalak, A.M., and Wofsy, S.C.: Regional sources of nitrous oxide over the

United States: Seasonal variation and spatial distribution, J. Geophys. Res., 117, D06310, doi:10.1029/2011JD016951, 2012.

Nassar, R., Napier-Linton, L., Gurney, K.R., Andres, R.J., Oda, T., Vogel, F.R., and Deng, F.: Improving the temporal and spatial distribution of CO_2 emissions from global fossil fuel emission data sets, J. Geophys. Res. Atmos., 118, 917–933, doi:10.1029/2012JD018196, 2013.

Nielsen, C. and Ho, M.: Clearer Skies Over China: Reconciling Air Quality, Climate, and Economic Goals, MIT Press, ISBN-13: 9780262019880, DOI:10.7551/mitpress/9780262019880.001.0001, 2013.

Niu, Z., Zhou, W., Wu, S., Cheng, P., Lu, X., Xiong, X., Du, H., Fu, Y., Wang, G.: Atmospheric Fossil Fuel CO2 Traced by Δ14C in Beijing and Xiamen, China: Temporal Variations, Inland/Coastal Differences and Influencing Factors, *Env. Sci. & Tech. 50* (11), 5474-5480, DOI: 10.1021/acs.est.5b02591, 2016.

NCEP National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce: NCEP FNL Operational Model Global Tropospheric Analyses, continuing from July 1999, http://dx.doi.org/10.5065/D6M043C6, Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, Colo. (Updated daily.) Accessed† 05 Feb 2014, 2000.

NDRC National Development Reform Commission: Enhanced Actions on Climate Change: China's Intended Nationally Determined Contributions, 2015.

Nehrkorn, T., Eluszkiewicz, J., Wofsy, S.C., Lin, J., Gerbig, C., Longo, M., and Freitas, S.: Coupled weather research and forecasting–stochastic time-inverted lagrangian transport (WRF–STILT) model, Meteorol. Atmos. Phys., 107: 51. doi:10.1007/s00703-010-0068-x, 2010.

Oda, T., Maksyutov, S., Andres, R.J.: The Open-source Data Inventory for Anthropogenic CO2, version 2016 (ODIAC2016): a global monthly fossil fuel CO2 gridded emissions data product for tracer transport simulations and surface flux inversions, Earth Syst. Sci. Data, 10, 87-107, https://doi.org/10.5194/essd-10-87-2018, 2018.

Piao, S., Fang, J., Ciais, P., Peylin, P., Huang, Y., Sitch, S., and Wang, T.: The carbon balance of terrestrial ecosystems in China, Nature, 458(7241), 1009–1013, 2009.

Sargent, M., Barrera, Y., Nehrkorn, T., Hutyra, L., Gately, C., Jones, T., McKain, K., Sweeney, C., Hegarty, J., Hardiman, B., Wang, J., Wofsy, S.: Anthropogenic and biogenic CO2 fluxes in the Boston urban region, Proc. Natl. Academy Sci, https://doi.org/10.1073/pnas.1803715115, 2018.

Shan, Y., Liu, J., Liu, Z., Xu, X., Shao, S., Wang, P., Guan, D.: New provincial CO2 emission inventories in China based on apparent energy consumption data and updated emission factors, Applied Energy, v184, 742-750, https://doi.org/10.1016/j.apenergy.2016.03.073, 2016.

Formatted: Font: Not Bold

Formatted: Font: Not Bold

Formatted: English (US)

Deleted: (*in press*)

Deleted:

Formatted: Font color: Custom Color(RGB(34,34,34)), English (US)

Turnbull, J.C., Tans, P.P., Lehman, S.J., Baker, D., Chung, Y., Gregg, J.S., Miller, J.B., Southon, J.R., Zhao, L.: Atmospheric observations of carbon monoxide and fossil fuel CO_2 emissions from East Asia. Journal of Geophysical Research Atmospheres 116. DOI:10.1029/2011JD016691, 2011.

Ummel, K. "CARMA Revisited: An Updated Database of Carbon Dioxide Emissions from Power Plants Worldwide." CGD Working Paper 304. Washington, D.C.: Center for Global Development. http://www.cgdev.org/content/publications/detail/1426429. 2012.

Wang, R., Tao, S., Ciais, P., Shen, H. Z., Huang, Y., Chen, H., Shen, G. F., Wang, B., Li, W., Zhang, Y. Y., Lu, Y., Zhu, D., Chen, Y. C., Liu, X. P., Wang, W. T., Wang, X. L., Liu, W. X., Li, B. G., and Piao, S. L.: High-resolution mapping of combustion processes and implications for CO_2 emissions, Atmos. Chem. Phys., 13, 5189–5203, https://doi.org/10.5194/acp-13-5189-2013, 2013.

Wang, X., Y. X. Wang, J. M. Hao, Y. Kondo, M. Irwin, J. W. Munger, and Y. J. Zhao, Top-down estimate of China's black carbon emissions using surface observations: Sensitivity to observation representativeness and transport model error, *Journal of Geophysical Research-Atmospheres*, *118*(11), 5781-5795, doi:10.1002/jgrd.50397. 2013.

Wang, Y., X. Wang, Y. Kondo, M. Kajino, J. W. Munger, and J. Hao, Black carbon and its correlation with trace gases at a rural site in Beijing: Top-down constraints from ambient measurements on bottom-up emissions, *Journal of Geophysical Research-Atmospheres*, *116*, doi:10.1029/2011jd016575. 2011.

Wang, Y., Munger, J., Xu, S., McElroy, M., Hao, J., Nielsen, C., and Ma, H.: CO_2 and its correlation with CO at a rural site near Beijing: Implications for combustion efficiency in China, Atmos. Chem. and Phys., 10: 8881-8897, 2010.

World Bank (2017). CO2 emissions (kg per PPP $ of GDP). Available at https://data.worldbank.org/indicator/EN.ATM.CO2E.PP.GD?locations=CN. Accessed May 12, 2017.

Zhao, Y., Nielsen, C.P., and McElroy, M.: China's CO_2 emissions estimated from the bottom up: Recent trends, spatial distributions, and quantification of uncertainties, Atmos. Envir., 59, 214-223, 2012.

Zhao, Y., Zhang, J., and Nielsen, C. P.: The effects of recent control policies on trends in emissions of anthropogenic atmospheric pollutants and CO_2 in China, Atmos. Chem. Phys., 13, 487-508, doi:10.5194/acp-13-487-2013, 2013.

Field Code Changed

Formatted: Font: Times New Roman, 12 pt

Formatted: Indent: Left: 0", Hanging: 0.31"

Formatted: Indent: Left: 0", First line: 0"

Table of Contents

S1 WRF Model: Post-processing and Evaluation 2

S2 STILT Model Set-up and Run Details 3

S3 Anthropogenic CO_2 inventories 3

S4 CT2015: Background Concentration Selection and Evaluation of Model Bias 4

S5 Scaling Results and Methodology 5

Fig. S1. CMA Station Map (2006, 2008) with WRF domain boundaries 7

Fig. S2. Miyun Receptor and surroundings, April 2007 7

Fig. S3. Evaluation of WRF output against observational data 8

Fig. S4. Observed vs WRF Modeled (Forecast) meteorology for sample WRF gridcell 9

Fig. S5. Q-Q plots of Observed and WRF Modeled (Forecast) meteorology for sample WRF gridcell 9

Fig. S6. ZHAO, EDGAR, and CDIAC estimates of total annual CO_2 emissions for Mainland China, 2005 to 2009 10

Fig. S7 Spatial Allocation of ZHAO inventories (2006-2008). 11

Fig. S8 Mean annual anthropogenic emissions (Mt CO2 yr-1, 2005-2009) zoomed to approximate d02 extent. 11

Fig. S9. IGBP land use categories in domain overlaid with STILT influence contours. 12

Fig. S10. Evaluation of CT2015 model bias. 12

Fig. S11. Example surface influence maps and basis for percentile contours. 13

Table S1. Comparison of unadjusted annual anthropogenic CO_2 emissions ($TgCO_2$) by region 14

Table S2. Seasonal Flux Corrections and 95% CI (kg CO_2 m^{-2} $month^{-1}$) for L_0.90 region. 15

Deleted: S1 WRF Model: Post-processing and Evaluation→ 2¶
S2 STILT Model Set-up and Run Details→ 3¶
S3 Anthropogenic CO_2 inventories→ 3¶
S4 CT2015: Background Concentration Selection and Evaluation of Model Bias→ 4¶
S5 Scaling Results and Methodology→ 5¶
Fig. S1. CMA Station Map (2006, 2008) with WRF domain boundaries→ 7¶
Fig. S2. Miyun Receptor and surroundings, April 2007→ 7¶
Fig. S3. Evaluation of WRF output against observational data→ 8¶
Fig. S4. Observed vs WRF Modeled (Forecast) meteorology for sample WRF gridcell→ 9¶
Fig. S5. Q-Q plots of Observed and WRF Modeled (Forecast) meteorology for sample WRF gridcell→ 9¶
Fig. S6. ZHAO, EDGAR, and CDIAC estimates of total annual CO_2 emissions for Mainland China, 2005 to 2009.→ 10¶
Fig. S7 Spatial Allocation of ZHAO inventories (2006-2008).→ 11¶
Fig. S8 Mean annual anthropogenic emissions (Mt CO2 yr-1, 2005-2009) zoomed to approximate d02 extent.→ 11¶
Fig. S9. IGBP land use categories in domain overlaid with STILT influence contours.→ 12¶
Fig. S10. Evaluation of CT2015 model bias.→ 12¶
Fig. S11. Sample footprint maps.→ 13¶
Table S1. Comparison of unadjusted annual anthropogenic CO_2 emissions ($TgCO_2$) by region→ 14¶
Table S2. Seasonal Flux Corrections and 95% CI (kg CO_2 m^{-2} $month^{-1}$) for L_0.90 region.→ 15¶
S1 WRF Model: Post-processing and Evaluation→ 2¶
S2 STILT Model Set-up and Run Details→ 3¶
S3 Anthropogenic CO_2 inventories→ 3¶
S4 CT2015: Background Concentration Selection and Evaluation of Model Bias→ 4¶
S5 Scaling Results and Methodology→ 5¶
Fig. S1. CMA Station Map (2006, 2008) with WRF domain boundaries→ 7¶
Fig. S2. Miyun Receptor and surroundings→ 7¶
Fig. S3. Evaluation of WRF output against observational data→ 8¶
Fig. S4. Observed vs WRF Modeled (Forecast) meteorology for sample WRF gridcell→ 9¶
Fig. S5. Q-Q plots of Observed and WRF Modeled (Forecast) meteorology for sample WRF gridcell→ 9¶
Fig. S6. ZHAO, EDGAR, and CDIAC estimates of total annual CO_2 emissions for Mainland China, 2005 to 2009.→ 10¶
Fig. S7 Spatial Allocation of ZHAO inventories (2006-2008).→ 11¶
Fig. S8 Mean annual anthropogenic emissions (Mt CO2 yr-1, 2005-2009) zoomed to approximate d02 extent.→ 11¶
Fig. S9. IGBP land use categories in domain overlaid with STILT influence contours.→ 12¶
Fig. S10. Evaluation of CT2015 model bias.→ 12¶
Table S1. Comparison of unadjusted annual anthropogenic CO_2 emissions ($TgCO_2$) by region→ 13¶
Table S2. Seasonal Flux Corrections and 95% CI (kg CO_2 m^{-2} $month^{-1}$) for L_0.90 region.→ 14¶

** *Note: Complete details of model set-up are available as part of our Replication Data Set at* https://dx.doi.org/10.7910/DVN/OJESO0 **

S1 WRF Model: Post-processing and Evaluation

We evaluate WRF output against publicly available, 24h-averaged Chinese Meteorological Administration (CMA) observational data. CMA observational data is not used in the NCEP FNL reanalysis WRF initialization fields. CMA provides daily averages of surface pressure, wind speed, temperature, and relative humidity. Access to higher temporal resolution observational data is limited. We convert hourly (d01) and half-hourly (d02, d03) WRF output to daily averages before evaluation. We use a combination of NCAR Command Language v6.1.2 (NCL; http://dx.doi.org/10.5065/D6WD3XH5) and R v2.9.0 (https://www.r-project.org/) to process the observed and simulated output. The standard post-processing toolbox, consisting of the WRF Unified Post Processor and METv4.1 Point-Stat Tool (http://www.dtcenter.org/code/) is not used here because of the low temporal resolution of observational data and file format mismatches. However, we base our evaluation method and procedures on the METv4.1 Point-Stat Tool. Both the METv4.1 and our version of the Point-Stat tool match WRF forecast fields to observation point locations for comparison. For surface observations, no interpolation is performed. Forecasts are instead matched to nearest CMA surface station observation point. Fig. S1 displays a map of the CMA surface network in 2006 and 2008, with approximate WRF domains overlaid with CMA station 54511 (C54511; 39.8N, 116.47E) highlighted in d03. We display sample evaluation results from C54511 in Fig. S3 through Fig. S5, using observed and simulated fields from 2006. In the evaluation, WRF forecast fields are matched to the nearest observation point.

Comparing against publicly available 2006 CMA data from 35 stations across the d02 and d03 domains (Fig S1), 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".

S2 STILT Model Set-up and Run Details

The version of WRF-STILT[1] used in this study corresponds to STILT release r701 of the AER-NOAA branch at the STILT svn repository[2], and Release-3-5 of the WRF-STILT interface[3]. Spin-up periods are removed from the WRF meteorological data and the WRF netcdf output files are converted to .arl format (Air Research Laboratory; https://ready.arl.noaa.gov/HYSPLIT_data2arl.php#INFO) prior to being ingested into STILT.

In this study, we transport an ensemble of 500 particles 7-days back in time to model footprints for each measurement hour at the receptor. The receptor (Miyun; 40°29′N, 116°46.45′E, 152 m above sea level (asl)) has the measurement inlet (STILT particle "release" point) located 6m above ground level (agl) (Fig. S2). We employ dynamic regridding, which accounts for resolution changes among the nested WRF domains. Mixing height is derived from WRF PBL heights; we set the surface layer as 50% of the mixed layer height. Footprints are integrated hourly. We set up the STILT runs as "pleasantly parallel" by running each month of a year simultaneously; hours within a month are run serially.

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/assessment studies, we model the 1100 to 1600 (local time) subset. These daylight hours represent a typical window for which STILT reliably models transport (e.g., 4). We examine the unadjusted model performance at all times, averaged seasonally and diurnally, in Sec S7.

S3 Anthropogenic CO_2 inventories

In order to facilitate comparison among the three anthropogenic inventories used in this study, we interpolate the two global inventories (EDGAR, 0.1ºx0.1º; CDIAC, 1ºx1º) to the same 0.25ºx0.25º grid as the regional inventory (ZHAO). We use the NCL Earth System Modeling Framework (ESMF) Conserve regridding method which minimizes deviation of the variable's integral between source and destination grids. We evaluate the impact of regridding in Fig. S6 by comparing annual totals ($MtCO_2$) before and after regridding. The ZHAO inventory remains on its native grid. We show that regridding does not appreciably affect the total emissions reported for mainland China by EDGAR and CDIAC, providing confidence in our representation of the two original inventories.

The ZHAO inventory provides estimates of total annual emissions for 2005 through 2009. In addition, the 2005 and 2009 ZHAO emissions are spatially allocated to a 0.25º x0.25º grid. We average the 2005 and 2009 percent contributions of each grid cell to the total emissions to provide weights for spatially allocating 2006 through 2008 total annual emissions. Fig. S7 evaluates the validity of this assumption by identifying regions where the 2009 gridcell contribution to the total emissions is outside +/- 2% of its 2005 contribution (Fig. S7a) and +/-50% of its 2005 contribution

[1] https://www.bgc-jena.mpg.de/bgc-systems/projects/stilt/pmwiki/pmwiki.php?n=WRFSTILT.WRF-STILT
[2] https://projects.bgc-jena.mpg.de/STILT/svn/branches
[3] available from http://files.aer.com/external/CarbonSoftware

(Fig. S7b). We find the assumption to be valid; the mean change per gridcell from 2009 relative to 2005 is -0.011% with a 2-σ of 15%.

Total uncorrected emissions for each anthropogenic inventory are calculated on the 0.25ºx0.25º grids and provided in Table S1. We provide emissions summed for each administrative region in the study domain, each STILT influence contour, and all China. Differences among the inventories zoomed to the L_0.90 region, are displayed in Fig. S8. Miyun and Beijing are encompassed by the L_0.25 contour. We display the average gridcell emissions of ZHAO (Fig. S8a) and the differences of EDGAR and CDIAC relative to ZHAO (Fig. S8b and Fig. S8c, respectively). In heavily emitting regions, ZHAO is typically higher than EDGAR and CDIAC. In regions where ZHAO is consistently lower than CDIAC, the differences are lower than the instances where ZHAO is higher. Note that, in the case of CDIAC, the uniformity of the differences includes artefacts from downscaling the gridded CDIAC inventory from 1°x1° to 0.25°x0.25°.

Deleted: influence

S4 CT2015: Background Concentration Selection and Evaluation of Model Bias

We derive estimates of background CO_2 concentrations from NOAA CarbonTracker (CT2015; https://www.esrl.noaa.gov/gmd/ccgg/carbontracker/CT2015/). CT2015 enables us to estimate concentrations of CO_2 prior to interaction with the surfaces in the study domain. Background value selection is summarized as follows. For each hour, the end x-y-z-time coordinates for each of 500 particles is found and linked to its corresponding CT2015 CO_2 concentrations using a spatiotemporal nearest neighbor approach. Only instances where a particle originated at the edge of the outermost domain and/or an altitude greater than or equal to 3000masl is included in the average background concentration calculation for that hour. If less than 75% of particles for an hour have valid background concentrations, that hour is not used in subsequent analyses. This selection criteria for background CO_2 mole fractions enables realistic modeling of true background conditions that have not interacted with the domain within each hourly measurement's maximum seven-day regional influence period. For the five-year study period, this method of boundary selection retains approximately 85% of hourly modelled values per year and across years.

The CT2015 model for the study domain is heavily trained by observations made approximately weekly via flask sampling at four World Meteorological Organization (WMO) sites in the region (https://www.esrl.noaa.gov/gmd/dv/site/). Mt. Waliguan to the west of the receptor (WLG) represents free tropospheric background air; Ulaan Uul (UUM) in Mongolia represents clean continental air; Tae-ahn Peninsula (TAP) in South Korea represents urban-influenced air from the east; Lulin (LLN) in Taiwan represents urban-influenced air from the southeast. TAP and LLN become more prominent in their representation upwind/background air sites during the spring and summer months when the East Asian Monsoon begins to influence regional air trajectory patterns. WLG and UUM are prominent in their representation of upwind/background air at all times of the year but particularly weight background air during the winter and fall seasons.

We quantify bias in the background model by evaluating observations against the nearest CT2015 model pixel and level. Observations are filtered using highest quality flask sample points only. Fig. S10(top panel) displays the time series of 3-hourly modeled CT2015 values and observed WMO measurements. Deviation of residuals from a normal distribution are displayed in Fig. S10

(bottom panel). The typical 1-σ model bias is 2ppm, but not all of the distributions are normal. For UUM, and therefore, CT2015 parameterization of clean continental background, the model-measurement residuals largely follow a normal distribution centered around 0. The clean continental background generally exhibits well-mixed behavior and is not defined by large excursions in the CO_2 signal. At the high-altitude WLG site representative of the free troposphere, the residuals follow a normal distribution centered around 0 but deviate from normal during instances where significant excursions in the CO_2 signal are present. This is also the case at LLN (distribution centered near 2.5ppm). TAP residuals deviate significantly from normal. In general CT2015 does not capture CO_2 events that are significantly different from global means; CT2015 underestimates uptake processes and overestimates lower or higher than global mean.

As not all deviations from observations can be represented as normal distributions, we place the model-measurement residuals at the four WMO sites in an error pool and select as part of the overall bootstrapping procedure for the modeling framework.

As shown in Fig. S10, 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.

S5 Scaling Results and Methodology

We translate the resulting mole fraction (ppm) mismatch between observed and modeled ΔCO_2 to inventory corrections at annual and seasonal timescales. We scale in the L_0.90 region (Fig. S9) which represents regions that substantially influence the receptor without disproportionally weighting pixels that contribute very little to the observed signal (Fig. S11). As discussed in the main text, we are still using surface influences from the entire STILT footprint to derive the CO_2 concentration at the receptor, but we ascribe the resulting model-observation mismatch as dominated by the L_0.90 region. Table S2 provides seasonal fluxes for each year before and after scaling. Annual scaling results are in Table 2 of the main text.

At annual scales, the dominant contributor to the CO_2 signal are anthropogenic emissions; correction 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, 13% of the northern China ecosystems and 20% of northeastern China's ecosystems are mixed forests. However, the ecosystems with greatest influence on this single site are croplands with high intra-annual carbon turnover rates. The heavily cropped L_0.90 region implies rapid turnaround of vegetation carbon stocks at the annual scale, justifying this assumption. At these timescales, we derive the $\Delta CO_{2,obs}/\Delta CO_{2,mod}$ ratio which represents the factor by which the annual anthropogenic inventory must be scaled in order to match observations. We use a model of the mean method to derive the annual scaling factors,

Deleted: contour

Formatted: Subscript

Deleted: influence

$$SF = \frac{\overline{\Delta CO_2}_{,obs_{hh}}}{\overline{\Delta CO_2}_{,mod_{hh}}}$$

where *hh* represents each local afternoon hour (1100 to 1600) in the year. SF>1 implies the model underestimates CO_2 concentrations while SF<1 implies the model overestimates CO_2 concentrations. We obtain 95% confidence bounds by bootstrapping uncertainties in the numerator and denominator separately, and obtaining the 0.025 and 0.975 quantiles from the ratio of the means of the two distributions. The annual influence contours are overlayed on the IGBP land use map in Fig. S9, 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.

At the seasonal timescales, however, evaluation of CO_2 processes is complicated by the biogenic flux contribution during the growing season and, to a lesser extent, the effects of ecosystem respiration in the dormant season. At these timescales, we derive additive corrections from converting observation-model mole fraction mismatch to the total CO_2 to be added or subtracted from the inventories. We correct the anthropogenic and vegetation inventories together as it is not possible to distinguish the contributions from our existing observational data set. For each modeled hour we derive a residual-based flux correction, $\Delta\Phi_{hh}$, in $\mu molCO_2 m^{-2} s^{-1}$:

$$\Delta\Phi_{hh} = \frac{\Delta CO_{2,obs_{hh}} - \Delta CO_{2,mod_{hh}}}{\sum_0^{-168h} foot_{hh}}$$

where *hh* represents each local afternoon hour (1100 to 1600) in the season and *h* represents the STILT footprint back-trajectory hour up to 7 days back in time. Given that anthropogenic emissions are positive terms and the biogenic component is a net balance of two opposing terms (uptake and release) of CO_2 during the growing seasons, use of inventory scaling factors for growing season scaling is inappropriate. That is, even a small mole fraction difference between modeled and observed in the growing season can result in meaningless scaling factors when there is a difference in sign involved. While scaling factors are appropriate during dormant seasons, for consistency we apply the same method of additive corrections across all seasons and report the adjusted inventory as fluxes (kg CO_2 m^{-2} $season^{-1}$). The methods are comparable; inventory corrections obtained by both methods during the winter and fall exhibit converging 95% confidence intervals.

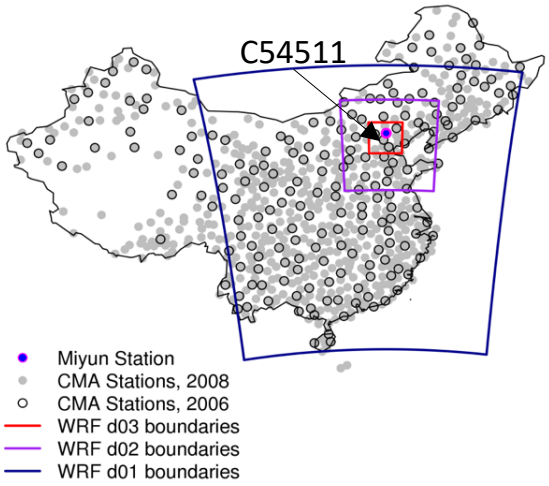


Fig. S1. CMA Station Map (2006, 2008) with WRF domain boundaries. Sample WRF evaluation results are provided for Station 54511 (indicated by arrow on map), near Miyun receptor.

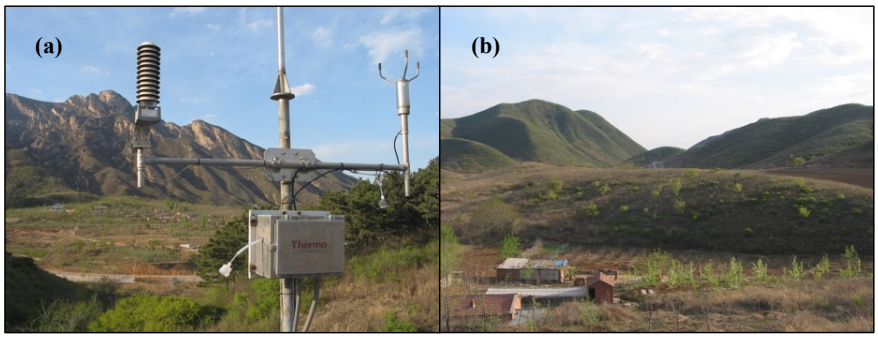


Fig. S2. Miyun Receptor and surroundings, April 2007. (a) Miyun inlet at 6magl/158masl, looking ENE, shows a small rural village in the valley below site, a small patch of short pines, that are generally in downwind direction. Even in spring there is still considerable bare ground. (b) view from Miyun sampling site, looking SW. Foreground shows a farmhouse and various outbuildings that were no longer in active use. Small-scale agricultural fields that were being converted to fruit-tree orchards. Unmanaged lands were grassy/shrub vegetation on hillsides.

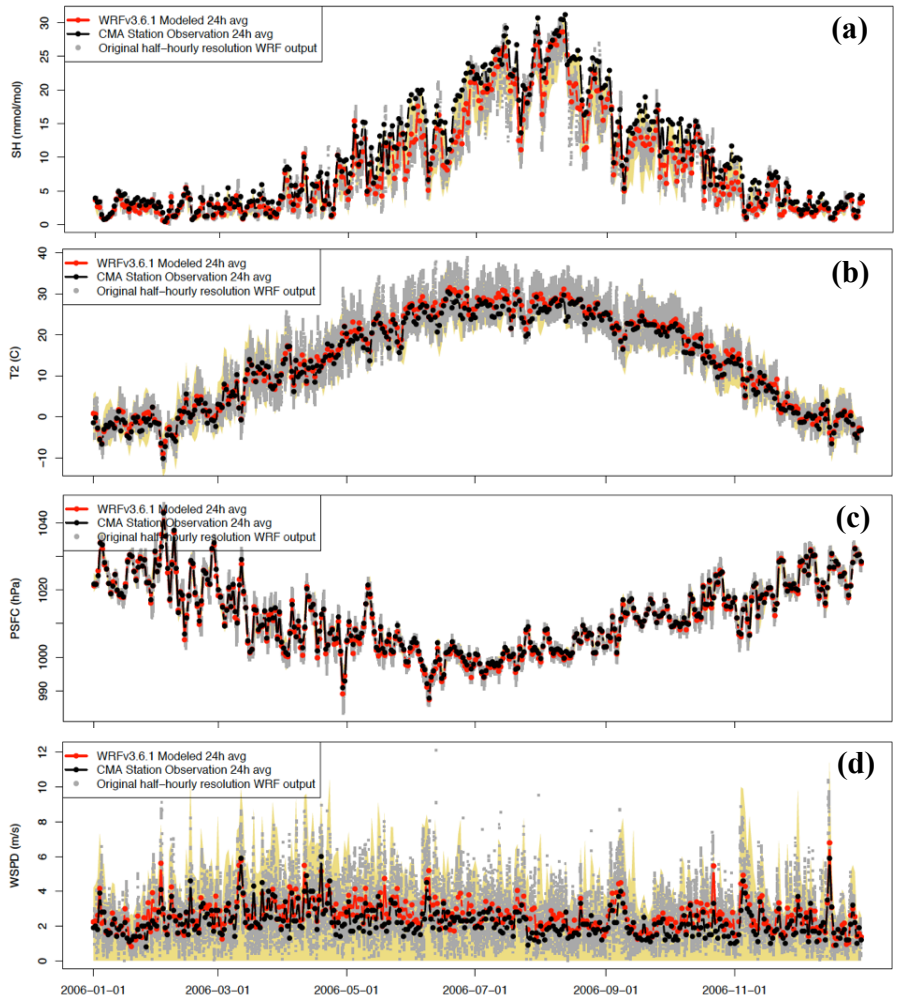


Fig. S3. Evaluation of WRF output against observational data. 2006 Meteorology timeseries for sample WRF gridcell (39.825N, 116.51E) evaluated against nearest CMA Station C54511 (39.800N 116.47E). WRF Meteorology averaged from half-hourly to daily for (a) Specific Humidity; (b) Surface Temperature; (c) Surface Pressure; (d) Surface Wind Speed. Original half-hourly output displayed in grey. Shaded yellow region represents observed daily range; daily minimum for windspeed is not available, but assumed to be 0m/s.

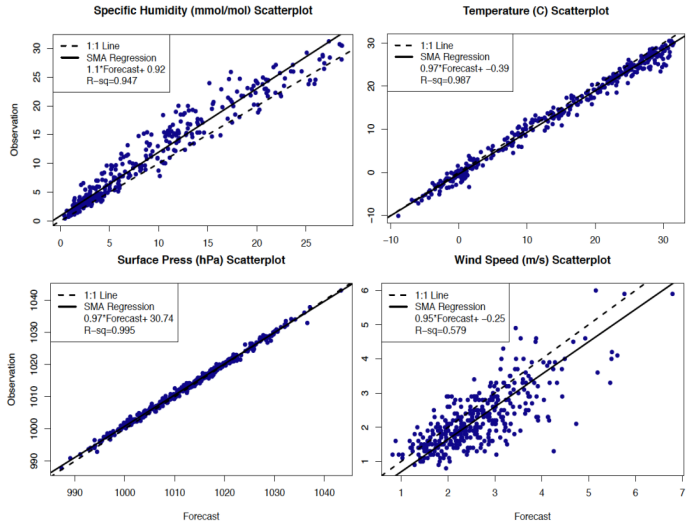


Fig. S4. Observed vs WRF Modeled (Forecast) meteorology for sample WRF gridcell. Gridcell (39.825N, 116.51E) evaluated against nearest CMA Station C54511 (39.800N 116.47E). Time-base of fields is daily average.

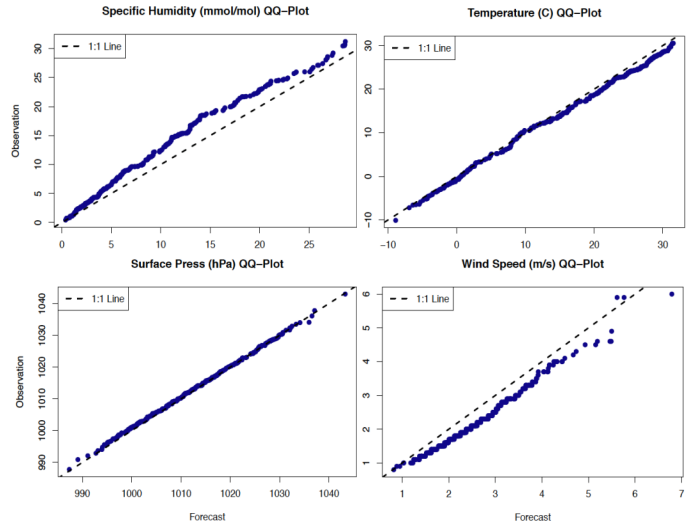


Fig. S5. Q-Q plots of Observed and WRF Modeled (Forecast) meteorology for sample WRF gridcell. Gridcell (39.825N, 116.51E) evaluated against nearest CMA Station C54511 (39.800N 116.47E). Time-base of fields is daily average.

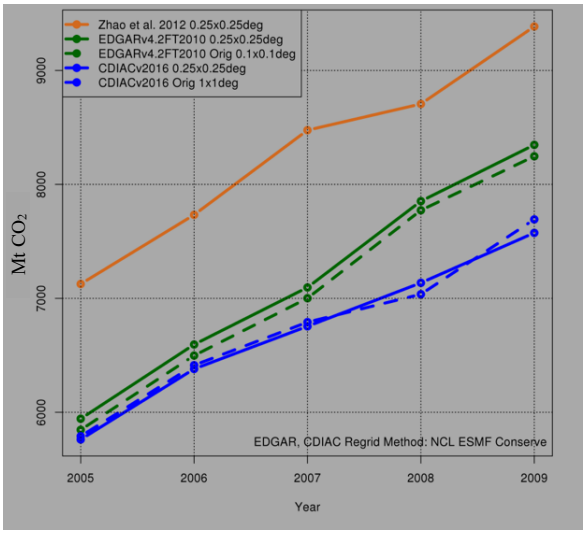


Fig. S6. ZHAO, EDGAR, and CDIAC estimates of total annual CO_2 emissions for Mainland China, 2005 to 2009. EDGAR and CDIAC are regridded to 0.25ºx0.25º grid using the NCL Earth System Modeling Framework Conserve regridding function.

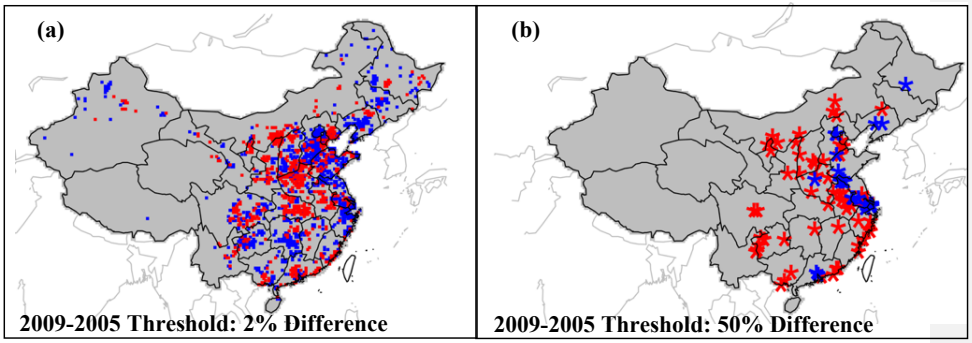


Fig. S7 Spatial Allocation of ZHAO inventories (2006-2008). Mean percent difference of gridcell contribution to total emissions is -0.011% ±15% (2-σ). We highlight instances where 2009 gridcell contribution to total annual emissions differs from its 2005 value by (a) more than 2% and (b) more than 50%. Blue represents a relative DECREASE in 2009 relative to 2005; red represents a relative INCREASE; grey represents values WITHIN the specified threshold.

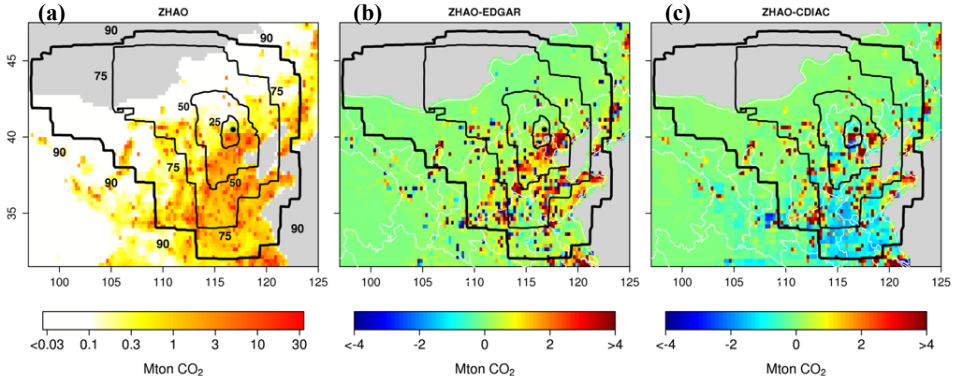
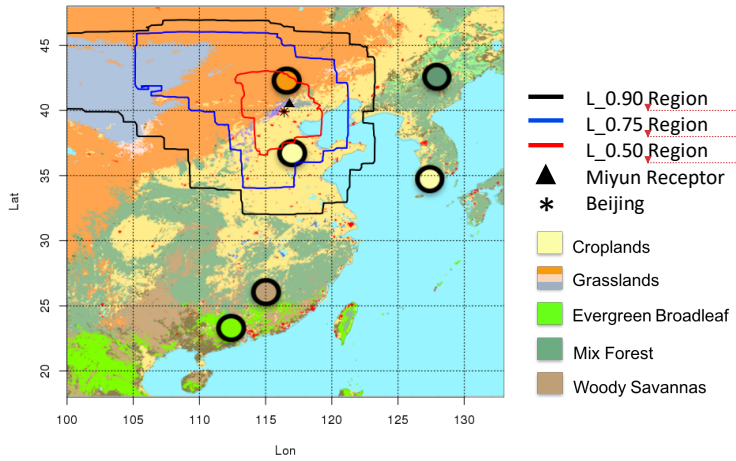


Fig. S8 Mean annual anthropogenic emissions (Mt CO_2 yr-1, 2005-2009) zoomed to approximate d02 extent. Black contour lines represent the 25th, 50th, and 75th, and 90th percentiles of multi-year mean annual STILT footprint influences. (3a) displays emissions estimated by ZHAO; black and green circle represents Miyun receptor. (3b) displays EDGAR inventory difference relative to ZHAO; (3c) displays CDIAC inventory difference relative to ZHAO. ZHAO is consistently higher than EDGAR and CDIAC in the Beijing area. Both EDGAR and CDIAC are regridded from their original grids to the ZHAO grid via ESMF Conserve regridding technique.

Formatted: Subscript



Deleted: Influence

Deleted: Influence

Deleted: Influence

Fig. S9. IGBP land use categories in domain overlaid with STILT influence contours. Note western edge of domain is slightly truncated.

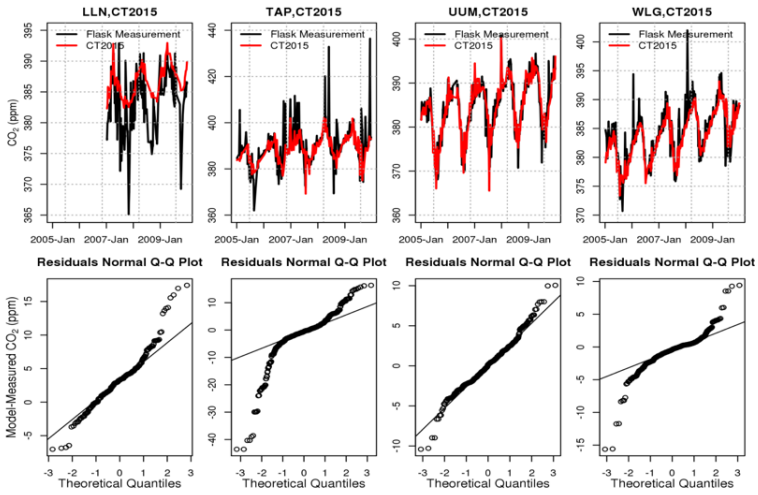


Fig. S10. Evaluation of CT2015 model bias. ~Weekly flask samples from WMO sites (LLN, TAP, UUM, WLG) used to train CT2015 compared with nearest CT2015 pixel. Top row: timeseries. Bottom row: QQ plots of model-measurement residuals.

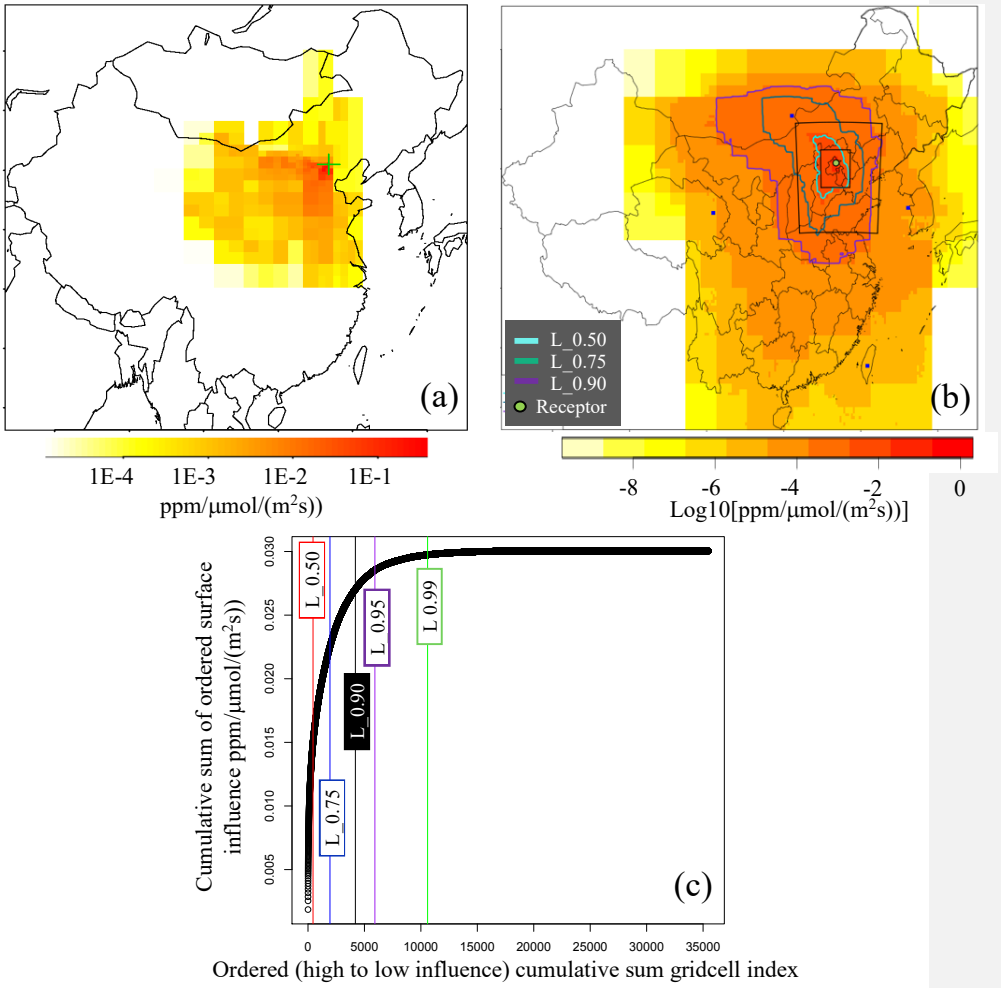


Fig. S11. Example surface influence maps and basis for percentile contours. (a) Sample hourly STILT footprint. Measurement hour on January 23, 2005 at 0700UTC (1500 Local). Surface influences are provided in ppm $\mu mol^{-1}m^{-2}s^{-1}$. Receptor release point is indicated by the green cross. (b) Example of annual average footprint for 2005 as log_{10}(ppm $\mu mol^{-1}m^{-2}s^{-1}$). Influence of gridcells on receptor drops by 5 orders of magnitude from L_0.90 contour to d01 edges. Note scale differences in sensitivity axes for (a) and (b). Black rectangles are d02 and d03 domain boundaries. (c) Cumulative sum of sorted (high to low) mean annual footprint from 2005-2009. Percentiles selected as points at or below fractional cutoff (0.5, 0.75, 0.9, 0.95, 0.99) of summed ordered footprint. Effect of excluding points outside each contour region is evident by steepness of curve beyond the respective vertical cutoff.

Formatted: Default Paragraph Font
Formatted: Default Paragraph Font
Formatted: Font: Not Bold
Formatted: Subscript
Formatted: Font: Symbol
Formatted: Not Superscript/ Subscript
Formatted: Font: Bold

Table S1. Comparison of unadjusted annual anthropogenic CO_2 emissions ($TgCO_2$) by region. EDGAR and CDIAC are reported as percent differences relative to ZHAO. *: Based on sums AFTER spatial allocation of ZHAO inventories but are <0.1% different from original inventory totals.

		STILT L_0.25	**STILT L_0.50**	**STILT L_0.75**	**STILT L_0.90**	**IM**	**NE**	**N**	**C**	**SE**	**S**	**SW**	All China
2005	ZHAO	135.1	697.0	1796	3015	252.1	682.8	2244	502.4	1486	519.6	759.5	7126
	EDGAR	*-31%*	*-35%*	*-28%*	*-23%*	*+1.9%*	*+1.2%*	*-32%*	*+1.2%*	*-12%*	*-19%*	*-25%*	*-17%*
	CDIAC	*-49%*	*-44%*	*-42%*	*-36%*	*-48%*	*-32%*	*-32%*	*+13%*	*-23%*	*-1.7%*	*+17%*	*-19%*
2006	ZHAO	124.8	734.4	1922	3273	311.7	690.6	2440	558.9	1590	567.6	822.9	7726*
	EDGAR	*-17%*	*-32%*	*-26%*	*-21%*	*-8.2%*	*+13%*	*-31%*	*+2.1%*	*-7.9%*	*-19%*	*-23%*	*-15%*
	CDIAC	*-39%*	*-41%*	*-40%*	*-34%*	*-54%*	*-26%*	*-31%*	*+13%*	*-21%*	*-0.74%*	*+20%*	*-17%*
2007	ZHAO	136.8	805.0	2107	3588	341.6	757.0	2675	612.6	1743	622.1	902.0	8469*
	EDGAR	*-18%*	*-33%*	*-27%*	*-22%*	*-9.8%*	*+12%*	*-32%*	*+0.76%*	*-9.3%*	*-21%*	*-25%*	*-16%*
	CDIAC	*-41%*	*-43%*	*-42%*	*-37%*	*-55%*	*-29%*	*-33%*	*+9.2%*	*-23%*	*-4.2%*	*+15%*	*-20%*
2008	ZHAO	140.5	826.8	2164	3685	350.9	777.5	2747	629.2	1790	639.0	926.4	8699*
	EDGAR	*-12%*	*-27%*	*-21%*	*-16%*	*-3.8%*	*+18%*	*-26%*	*+7.5%*	*-1.9%*	*-14%*	*-20%*	*-9.7%*
	CDIAC	*-39%*	*-41%*	*-40%*	*-35%*	*-54%*	*-27%*	*-31%*	*+12%*	*-21%*	*-1.4%*	*+19%*	*-18%*
2009	ZHAO	125.1	864.7	2301	3974	424.6	777.2	2967	694.8	1903	693.4	997.2	9370
	EDGAR	*+5.4%*	*-26%*	*-20%*	*-17%*	*-16%*	*+25%*	*-27%*	*+3.5%*	*-1.8%*	*-15%*	*-21%*	*-11%*
	CDIAC	*-26%*	*-40%*	*-40%*	*-36%*	*-60%*	*-22%*	*-32%*	*+8.0%*	*-21%*	*-3.5%*	*+17%*	*-19%*

Table S2. Seasonal Flux Corrections and 95% CI (kg CO_2 m^{-2} $month^{-1}$) for L_0.90 region. Original fluxes are in regular font; corrected fluxes and 95% CI are in bold

		JFM/Winter	AMJ/Spring	JAS/Summer	OND/Fall
2005	**ZHAO**	0.133	0.0492	-0.0540	0.132
		0.129 (0.103, 0.105)	**0.0735 (0.0195, 0.135)**	**-0.170 (-0.237,-0.106)**	**0.164 (0.137, 0.193)**
	EDGAR	0.108	0.0256	-0.076	0.110
		0.151 (0.124, 0.174)	**0.116 (0.0597, 0.176)**	**-0.120 (-0.186, -0.0478)**	**0.181 (0.154, 0.204)**
	CDIAC	0.0937	0.0117	-0.0972	0.0951
		0.144 (0.117, 0.170)	**0.132 (0.0734, 0.185)**	**-0.121 (-0.183, -0.0445)**	**0.177 (0.147, 0.206)**
2006	**ZHAO**	0.131	0.0601	-0.0568	0.140
		0.146 (0.122, 0.167)	**0.156 (0.0990,0.217)**	**-0.135 (-0.197,-0.0708)**	**0.174 (0.124, 0.217)**
	EDGAR	0.106	0.0421	-0.0771	0.114
		0.169 (0.145, 0.190)	**0.185 (0.126, 0.246)**	**-0.0951 (-0.157, -0.0310)**	**0.204 (0.152, 0.251)**
	CDIAC	0.0929	0.0260	-0.102	0.0965
		0.165 (0.139, 0.189)	**0.194 (0.134, 0.254)**	**-0.0912 (-0.157, -0.0171)**	**0.223 (0.168, 0.270)**
2007	**ZHAO**	0.139	0.0831	-0.0735	-0.171
		0.154 (0.118, 0.189)	**0.109 (-0.00290, 0.217)**	**-0.151 (-0.205, -0.0958)**	**0.174 (0.129, 0.214)**
	EDGAR	0.109	0.0569	-0.103	0.138
		0.171 (0.133,0.205)	**0.141 (0.0282, 0.264)**	**-0.110 (-0.170, -0.0528)**	**0.192 (0.151, 0.231)**
	CDIAC	0.0917	0.0323	-0.123	0.119
		0.157 (0.119, 0.191)	**0.149 (0.0381, 0.271)**	**-0.113 (-0.173, -0.490)**	**0.184 (0.138, 0.222)**
2008	**ZHAO**	0.120	0.0577	-0.0290	0.143
		0.134 (0.109, 0.160)	**0.0157 (-0.0470,0.0794)**	**-0.170 (-0.247, -0.0940)**	**0.201 (0.159, 0.243)**
	EDGAR	0.0973	0.0459	-0.419	0.118
		0.145 (0.120, 0.171)	**0.0492 (-0.0140,0.111)**	**-0.127 (-0.207,-0.0447)**	**0.219 (0.174, 0.259)**
	CDIAC	0.0785	0.0135	-0.0800	0.0960
		0.139 (0.109, 0.166)	**0.0559 (-0.0114, 0.122)**	**-0.134 (-0.217, -0.0494)**	**0.224 (0.179, 0.264)**
2009	**ZHAO**	0.144	0.0809	0.0277	0.134
		0.231 (0.130, 0.300)	**-0.0655 (-0.127, -0.00290)**	**-0.125 (-0.193, -0.0449)**	**0.215 (0.158, 0.265)**
	EDGAR	0.130	0.0563	-0.00797	0.112
		0.249 (0.156, 0.313)	**-0.0653 (-0.124, 0.00)**	**-0.122 (-0.197, -0.0399)**	**0.217 (0.165, 0.266)**
	CDIAC	0.0970	0.0355	-0.0312	0.0874
		0.238 (0.147, 0.306)	**-0.0404 (-0.105, 0.0239)**	**-0.110 (-0.192, -0.0267)**	**0.215 (0.162, 0.270)**