Response to Reviewer 1

We thank the two reviewers for their efforts and constructive comments (https://www.geosci-model-dev-discuss.net/gmd-2018-123/#discussion). Each reviewer's comments are shown below in *italics*, followed by our point-by-point responses in blue.

Anonymous Referee #1

This paper describes a new modification and application of the STILT model for total column measurements. This will allow X-STILT to be used to interpret satellite (and ground-based) total column abundances, and is a timely contribution, given the rapidly increasing number of satellite greenhouse gas total column measurements. The manuscript is thorough and technical, generally clear and well written, and suitable for this journal. I would recommend its publication after the following comments are addressed.

We thank anonymous referee #1 for the positive feedback and have attempted to address these comments and made several clarifications and changes to the manuscript. Here are major changes made during the review process including

1) the quantification of XCO_2 errors due to vertical mixing errors within X-STILT (**Sect. 2.6.2**);

2) discussions on results using observations from b8 Lite files (**Sect. 4.4**) and X-STILT's potential for broader applications and expected changes (**Sect. 4.3**);

3) details on the wind bias correction within the model (**Appendix C**); and

4) updated simulations and codes built on the STILT-R version 2 (Fasoli et al., 2018) with updated description in **Sect. 2.1.2**.

General Comments

"Is X-STILT restricted to XCO2, or could it be applied to any total column tracer (e.g., XCH4, XCO, XN2O, etc.)? Could it also be applied to ground-based total column measurements? My understanding from reading this paper is that X-STILT could be applied more generally, and that you were showing a rigorous example of its functionality with OCO-2 XCO2. If this is true, its generality should be made more clear – perhaps with a more general title and introduction."

In general, we agree with the reviewer on X-STILT's potentials for wider applicability. We did not study other species than CO_2 or incorporate profiles from other sensors. The codes we currently modified only aim at XCO_2 and incorporate OCO-2 satellite profiles. However, we do foresee that X-STILT could be applied more generally. Still, we are inclined to retain the current title and majority of the manuscript, as our methods (in particular our definition of background) are built on our particular focus over urban areas. Continuous work is ongoing towards a more flexible model framework that can be more easily applied to other column measurements.

We have now clarified the model's generality and expected changes for other tracers within a new subsection (Sect. 4.3) of the main text:

> **"4.3 X-STILT's potential for broader applications**
>
> In theory, X-STILT can be applied to other column measurements and other species. The underlying Lagrangian atmospheric model (STILT) has been applied to simulate other atmospheric species, such as CO, CH_4 and N_2O (Mallia et al., 2015; Kort et al., 2008). One of the key modifications to X-STILT from STILT is the column weighting of STILT footprint values (Sect. 2.1.2). Specifically, X-STILT interpolates the OCO-2 *AK* and *PW* onto each modeled level and then applies weighting of the trajectory-level footprints before generating a horizontal footprint map. The X-STILT code can be easily modified to apply sensor-specific vertical profiles of *AK* and *PW* from other satellites or ground-based column measurements.
>
> Lastly and more importantly, background may need to be derived differently according to different applications, e.g., local urban emissions versus regional fluxes. The overpass specific background (M3) aims at isolating the citywide emissions, so it makes use of the measurements outside the city, but still are quite closed to the city (within the few

degrees latitude). However, if the study focus is to look at emissions over a much broader region (e.g., statewide emissions), background region should be defined farther away from the target region, e.g., taking the advantages of measurements from available upstream overpasses."

Specific Comments

1. P2L30-36.

"I don't think the current suite of CO2 satellites will completely fill in the gaps of the surface in situ networks, especially over specific locations such as cities. Certainly the future looks bright with OCO-3's "city mode", and the geostationary missions on the horizon, but I think you are overstating the impacts over cities as our satellite observing system currently stands."

We agree with the reviewer that the surface in situ networks have their vital and irreplaceable role in the CO_2 measurement system. We might overstate the impacts over cities and be over-optimistic about the current suite of CO_2 satellites. Thus, we reworded the previous text (**P2L34-36**) as:

"Although most carbon-observing satellites have revisit times of multiple days (e.g., 3 days for GOSAT and 16 days for OCO-2), their global coverage, large number of retrievals and multi-year observations may further complement the current surface observing networks. Space-borne CO_2 measurements, in combination with surface CO_2 networks, may help reduce emission uncertainties and benefit urban emissions analysis, especially over regions with no surface observations (Duren and Miller, 2012; Houweling et al., 2004; Rayner and O'Brien, 2001)"

2. Section 2.1.

"I'm having trouble with your definition of the sensitivity of the satellite sensor: it seems incomplete. The column averaging kernel represents the change in the retrieved total column with respect to a perturbation in the abundance at a particular altitude. When the column averaging kernel is 0, the measurement is insensitive to changes at that altitude, and thus relies completely on the information in the a priori profile to construct the column. (Ref: OCO-2 ATBD, P58: https://co2.jpl.nasa.gov/static/docs/OCO-2%20ATBD_140530%20with%20ASD.pdf)

Weighting functions, which you mention on L33, at least to the retrieval community, refer to the Jacobian matrices, and while these are related to the averaging kernel matrices, they are not the same as the column averaging kernels (see P54 on the ATBD document above). I'd ask that this section is clarified further.

We apologize for the many confusions caused by the lack of clarity and a few mistakes made in Sect. 2.1 and Eq. (1). We modified the definition of AK as: "the sensitivity of the change in retrieved XCO_2 due to CO_2 anomaly at each retrieved grid" (**P5L3**). We reframed the entire Sect. 2.1 to clear up these confusions. And, here are some responses and clarifications:

1) In terms of representing the atmospheric column, we mislabeled *n* in Eq. (1) in the previous version, and corrections have been made to **P5L10-13**:

   "$XCO_{2.sim.ak} = \sum_{n=1}^{nlevel}(AK_{norm,n}PW_nCO_{2.sim.n} + (I - AK_{norm,n})PW_nCO_{2.prior,n})$ (1)

   I is the identity vector and *n* stands for the combined vertical levels of X-STILT plus OCO-2. Specifically, we replaced OCO-2 levels with denser model release levels for the lower part of the troposphere (red circles from the surface in Fig. 2), while kept OCO-2 levels for upper part (blue circles in Fig. 2). To reduce computational cost, the air column is only simulated up to the maximum release height (MAXAGL in meters above the ground level, mAGL; Fig. 2)."

2) Background definitions can be quite different among studies depending on their applications (e.g., examined spatiotemporal scales). In this study, because our focus is about the urban emissions from a target city, we defined the background XCO_2 as the portion that is not affected by urban emissions and naturally broke the total modeled XCO_2 down to two components – $XCO_{2.ff}$ and $XCO_{2.bg}$.

- When we solely rely on model to estimate $XCO_{2.bg}$ (the trajectory-endpoint method, M1), the background XCO_2 is the sum of *AK*-weighted modeled biospheric and oceanic perturbations on top of CO_2 boundary conditions using CarbonTracker and a portion from OCO-2 a priori profiles (Eq. (2)).
- When the other background methods (M2H, M2S, M3) are being examined, the background XCO_2 per overpass is simple one number derived from statistical methods or the forward plume. Thus, no simulations on biospheric or oceanic anomalies involving OCO-2 prior or CarbonTracker is used in these cases.

- Clarifications have been made to the relevant text on **P5L23-28**:

  "Eq. (1) can further be rewritten as Eq. (2), since the simulated CO_2 profiles in Eq. (1) is comprised of CO_2 boundary condition plus CO_2 anomalies due to sources/sinks ($FFCO_2$, biospheric and oceanic fluxes):

  $$XCO_{2.sim.ak} = XCO_{2.sim.ff} + XCO_{2.sim.bio} + XCO_{2.sim.ocean} + XCO_{2.sim.bound} + XCO_{2.prior} = XCO_{2.sim.ff} + XCO_{2.bg}. \quad (2)$$

  Given our focus, we defined background value as the XCO_2 portion not "contaminated" by urban emissions. Thus, $XCO_{2.sim.ak}$ is the sum of the XCO_2 enhancement due to $FFCO_2$ ($XCO_{2.sim.ff}$) and estimated background value ($XCO_{2.bg}$). Estimates of XCO_2 anomalies are further explained in Sect. 2.2.2 and four ways to estimate background values ($XCO_{2.bg}$) are proposed in Sect. 2.3."

3) Yes—the "weighting functions" we referred to in Sect. 2.1 is the column averaging kernel and pressure weighing functions, instead of the Jacobian matrices in the retrieval. We have corrected our word choices (**P5L6-7**) to "the satellite's column averaging kernels".

On P5, you mention the interpolation of the measurement onto the model levels. Why wouldn't you do the reverse: interpolate the model onto the retrieval grid? Your method requires that you make several assumptions that seem to complicate your analyses (i.e., these "scaling factors" you mention). Is there a compelling reason not to interpolate the model instead of the measurements? Please explain these "scaling factors" in more detail to walk the reader through Fig. 2."

Here are explanations to the Reviewer's questions on the scaling factors and interpolation of measurements' profiles:

1) *PW* function is primarily estimated based on the air mass or pressure difference (*dp*) between two layers. Fig. 2b shows that OCO-2 levels have relatively **constant pressure difference** (*dp_oco2*) and *PW* values (in black dots, except for the very top and bottom level). On the contrary, model levels are finer (in orange circles in Fig. 2b) with two different vertical spacings in **altitudes** (100 m vs. 500 m) below and above 3 km. So, those scaling factors are to adjust the *PW* function according to difference in *dp_oco2* vs. *dp_stilt*.

   For example, from 0 to 3 km, *dp_stilt* ranges from ~8-10 mb, while *dp_oco2* are mostly ~52 mb (red circles vs. black dots in Fig. 2b). Thus, we further scaled down the interpolated *PW* (red circles in Fig. 2b) by the ratio of *dp_stilt* over *dp_oco2* (e.g., sf = 10 mb/52 mb), because of less air between model levels than initial retrieval grids. Thus, final *PW* after scaling has value of ~0.01 (orange circles in Fig. 2b).

   Comparing air below 3km, fewer model levels are placed from 3-6 km (~650 to 450 mb), which gives larger *dp_stilt*, larger *PW* scaling factors and resultant larger scaled *PW* of ~0.03 –0.04 (orange circles in Fig. 2b).

   Since no model level placed above MAXAGL, *PW* stay the same as the initial OCO-2 *PW* values (blue circles in Fig. 2b). Finally, we made sure that the sum of vertical PW profile ends up being 1 for each sounding/receptor.

2) We agree with the reviewer that we could construct the 'redistribution matrix' and interpolate the model values on retrieval grid, as done in Basu et al. (2013). However, we would like to keep the current method for the following reasons (with reasons explained on **P5L14-16** in the revised version as well):

- It seems that the construction of redistribution matrix may also need to deal with or resolve the mismatch between model levels versus retrieval grids, unless model levels perfectly agree with retrieval grids.
- Most importantly, as our intention is to preserve finer variations in modeled CO_2 benefit from placing denser model levels within the PBL, we foresee the reversed interpolation (from model levels back to retrieval grids) could potentially involve some averaging or smoothing.

3) Clarifications have been made to the main text in **Sect. 2.1 (P5L14-22)** to explain above comments:

"Interpolations are further needed to resolve the mismatch between prescribed OCO-2 retrieval grids and model levels for the lower part of the troposphere. Our intention is to preserve the finer modeled CO_2 variations by performing interpolations of satellite profiles from retrieval grids to model levels. Vertical profiles of AK_{norm}, *PW* and $CO_{2,prior}$ are treated as continuous functions and interpolated linearly to model grids (red circles in Fig. 2). Note that the initial OCO-2 *PW* functions have steady value of ~0.052 (except for the very bottom and top levels; black dots in Fig. 2b), which results from constant pressure spacings (*dp_oco2*) between two adjacent OCO-2 levels. However, X-STILT levels are much denser with smaller pressure spacings (*dp_stilt*) or less airmass between their two adjacent levels. Therefore, the linearly interpolated *PW* (red circles in Fig. 2b) needs an additional scaling via a set of "scaling factors" representing the ratios of pressure spacings in STILT versus OCO-2 retrieval (*dp_stilt/dp_oco2)*, to arrive at the correct *PW* for each finer model grid (orange circles in Fig. 2b)."

3. Section 2.4.

"A recent paper by Nassar et al. (2017) would be relevant to cite in Section 2.4. They use OCO-2 data to quantify power plant emissions, and they choose an overpass-dependent background that would be interesting to compare with your method. Ref: Nassar, R., T. G. Hill, C. A. McLinden, D. Wunch, D. B. A. Jones, and D. Crisp (2017), Quantifying CO2 emissions from individual power plants from space, Geophys. Res. Lett., doi:10.1002/2017GL074702."

We appreciate the reviewer in pointing out this paper. We have added this relevant paper when discussing different ways to determine the background in **Sect. 2.3.3 (P8L27-29)**:

"Nassar et al. (2017) derived overpass-dependent background and its uncertainty based on the averaged OCO-2 observations within four different tested background latitudinal ranges."

as well as discussing the challenges in defining background and associated uncertainties when dealing with column measurements in **Sect. 1 (P4L17-19)**:

"Lastly but more importantly, recent column studies (Nassar et al., 2017; Fischer et al., 2017) studied the impact of potential errors/biases in background values on their emission or fluxes estimates."

In general, both studies derived the background from OCO-2 measurements over the "clean" region and accounted for background uncertainties. However, as Nassar et al. (2017) focuses more on emissions from individual power plants over different regions, we did not conduct direct comparisons against our background values.

4. Section 2.6.

My (admittedly simplistic) understanding of the transport error issue is that it is still important for models to get the vertical transport right when assimilating or inverting total column measurements, because the vertical transport sets the altitude at which advection occurs, and thus the distribution of the gas around the planet. So while the total column measurements themselves are insensitive to the altitude of the molecule within the column, it is not necessarily the case that the models are better able to reproduce the column. Indeed, the abstract of Lauvaux and Davis cited in this paragraph seems to confirm that vertical transport errors are very important for calculating fluxes from column-integrated measurements. Please address this issue further.

We agree with the reviewer that while likely small, we may not completely be able to neglect the impact of vertical transport on column simulations. To quantify this error, we have now conducted another set of transport analysis to quantify XCO_2 errors due to vertical mixing. Sect. 2.6 in previous version has now been divided into two subsections for horizontal (Sect. 2.6.1) and vertical transport errors (Sect. 2.6.2, newly added). Section 2.6.2 has been added as:

> **"2.6.2 Vertical transport errors**
> Vertical turbulent mixing dominants the vertical transport of air parcels and control the dilution of surface emissions within the PBL (e.g., Gerbig et al., 2008). Uncertainties in the vertical mixing or PBL height can affect both the footprint magnitude and the its spatial distribution via different horizontal advections at each altitude. Although column-integrated measurements may be less sensitive to vertical distribution of air particles than in situ measurements, vertical transport errors can have some impacts on column simulations nonetheless, due to wind shear and its interaction with vertical redistribution of air parcels (Lauvaux and Davis, 2014). Comprehensive quantifications of the vertical transport errors in a column sense are performed in Lauvaux and Davis (2014) using ensemble of surface and planetary BL parameterizations involving a regional inverse modeling framework.
>
> Instead, we made use of the stochastic nature of STILT and propagated typical PBL height errors in the model. Changes in STILT-modeled mixed layer height modify the vertical profiles of turbulent statistics that directly control the stochastic motions of the Lagrangian air parcels (Lin et al., 2003). Thus, we obtained different air parcel trajectories with rescaled PBL heights. The resultant vertical transport error in XCO_2 space is calculated as the root-mean-squared errors (RMSEs) between two sets of XCO_2 enhancements among different receptors for each overpass. Due to this calculation, vertical transport errors are only provided at the overpass level (results in Sect. 3.5). Gerbig et al. (2008) reported typical relative PBL errors in the range of ± 20 %. Thus, we rescaled the PBL heights higher and lower by 20 % and evaluated the scaling's impact on XCO_2 enhancements. Because of our focus on the urban emissions and potential small XCO_2 enhancements contributions beyond one day backwards in time, we only rescaled PBL within the first 24 hours of transport before arrival of the air parcels at the column receptors."

We further briefly mentioned the XCO_2 errors due to vertical mixing error as in **Sect. 3.5**:

> "XCO_2 errors solely resulted from vertical mixing errors are in general < 15 % of the modeled signal for each overpass, whereas XCO_2 errors due to horizontal wind errors dominate the overall XCO_2 transport error (Table 1)."

Furthermore, there is little discussion about the atmosphere above MAXAGL (~450 hPa). While this is unlikely to be important for CO2 emissions over regional or smaller scales, it may be important for other tracers (e.g., CH4). Can you comment on how important the tropopause altitude, for example, might impact this work? Over what spatial and temporal scales would X-STILT properly represent the total column?

We strongly agree with the reviewer that atmosphere above MAXAGL can be less important for XCO_2 emissions, but can be important for other tracers like methane, due to chemical productions and losses along the atmospheric transport. In theory, the vertical profile of methane can be sensitive to the tropopause altitude since it rapidly decreases in the stratosphere.

We admit that no chemical reaction is considered in X-STILT at this point and the model levels are only placed over the lower troposphere, given our focus on urban-scale CO_2 surface emissions (that only get mixed within the PBL). As particles released from higher levels can hardly get entrained and make contact with the near-field land surface, the X-STILT column particles may properly represent the total column over the urban scale for less than one day. Due to our specific focus on urban CO_2, we decided not to add related content in the main text, but to address the reviewer's question here.

However, we may expect small impact on our work with reasoning showed below.

- Although variations for atmosphere above MAXAGL are not explicitly modeled or represented by X-STILT particles given the way we release air parcels, those variations are part of the defined background.
- Take methane as an example. Recall that the background air defined by the overpass-specific method (M3) are atmospheric columns located outside (but not far away from) the city plume. The background value is derived from measured XCH_4 over those background atmospheric columns. In fact, those measured XCH_4 are the resultant XCH_4 after being produced or destructed along their way (backward in time) and have contained information about the upper tropospheric and stratospheric variations over upwind regions. Since the background air and the plume-elevated air are not far away from each other, we may assume no big difference in their tropopause altitudes. Thus, both background and plume-elevated XCH_4 contains information about the XCH_4 over atmosphere over MAXAGL. And, the difference between the two gives us the methane enhancements due to urban emissions.
- One note on the background definitions: When using the trajectory-endpoint method (M1), model trajectories are used to model the XCO_2 boundary condition. However, when using the other three background methods (M2H, M2S and M3), model trajectories are only used to estimate the XCO_2 anomalies due to different sources and sinks. The choice of max release height may impact the modeled XCO_2 enhancements. While as shown in the MAXAGL sensitivity test, almost no changes in XCO_2 enhancement due to increase in MAXAGL.

5. OCO-2 v7.

You are using v7 of the OCO-2 data in this paper, but v8 is available and v9 will be available soon. V8 has significant improvements in the treatment of aerosols and throughput, which may be important for work over polluted urban regions. V9 will have improved pointing, especially important over topography. Please comment on whether your results will be robust against these changes to the OCO-2 data.

We thank the reviewer in pointing out different OCO-2 versions and look forward to the upcoming b9 product to help future studies over cities with complex terrain. We have now performed few more simulations and analysis using version 8 Lite product and briefly depicted the changes in observed and simulated urban XCO_2 signals in **Sect. 4.4 (P19L23-33**). As ongoing improvements made in OCO-2 retrievals are modifying the observations, it may not be that fair for us to comment the robustness in our results, especially given model evaluation against observations.

Text has been added:

"Emission evaluations for different regions can be different and affected by different observational constraints. Even changes in different versions of the retrieval (Lite b7 vs. b8) may slightly affect the model-data comparisons and simple inversion results in this work. Modeled XCO_2 enhancements using the newer b8 differ slightly from those using b7 (purple dots in Fig. S8 vs. in Fig. S13) due to changes in the locations of the receptors, column averaging kernels, and data filtering (QF) for measurements around Riyadh. Specifically, observations from b8 may yield more overpasses with sufficient screened soundings than those from b7 (black and red bars in Fig. S1). However, much larger differences in observed enhancements are found and caused by the changes in total observed XCO_2 and estimated background values. Specifically, background uncertainty decreases by up to 0.1 ppm primarily attributed to smaller spread (smaller SD) of the observed XCO_2. Positive shifts in the total observed XCO_2 for b8 from b7 are found over most overpasses (Fig. S11). The M3-derived observed enhancements may be less affected by positive shifts in total observations, given similar positive shift associated with the overpass-specific background near the target urban region (dark green dashed lines in Fig. 6e vs. Fig. S12)."

Technical Comments

6. Section 2.1.1

I found myself wondering which version of OCO-2 data you were using, given the discussion about quality flags and albedo cutoffs, which is version-specific. I realize this is answered later in Section 2.2. I'd suggest either not mentioning the specifics of the quality filtering in section 2.1.1, or mentioning the data version in 2.1.1.

We realize the confusion caused by the introduction on quality flags in Sect. 2.1.1. We have removed this discussion (the data filtering on observations, i.e., QF and aerosols cutoffs) in Sect. 2.1.1.

7. P10L14-7

I'm having trouble understanding these sentences, and I believe this wind correction may be an important step in X-STILT. Please explain in more detail.

We apologize for the confusion and clarify that this attempt to correct wind biases can be an important part in X-STILT, in particular over places with large systematic wind error, less complex terrain and denser wind observations. We briefly mentioned this correction and discussed its limitation in **Sect. 2.6.1 (P11L33- P12L9)** and added an **Appendix C** for details of this bias-correction:

"Appendix C: Correcting for wind biases within X-STILT

While we did not apply the wind bias correction for the overpasses analyzed in this paper due to the biases being generally small (previously explained in Sect. 2.3.3), X-STILT has the capability to account for biases, if necessary. The basic idea is to correct the near-field wind biases in both forward- and backward- time trajectories. Because wind error at each observed pressure level can be quite different, vertically-weighted u- and v- wind biases were calculated by fitting logarithmic mean wind profiles based on available near-fields observed and simulated wind speeds and directions. We then calculated the deviations in latitude and longitude directions (dx, dy, with conversion from distance to degrees) given estimated u- and v- wind biases. These deviations accumulate as air parcels travel further backward or forward in time and are used to correct the location of each particle. After fixing the particle locations, Fig. S6b shows the general distribution of backward trajectory being clockwise rotated, compared to initial trajectory distribution in Fig. S6b. Air parcels in Fig. S6b appear to be "noisier" than those in Fig. S6a, due to inclusion of the random wind error component. Then the new bias-corrected set of column trajectory is used to generate spatial footprint. This correction can also be performed to forward-time trajectory to reduce wind bias impact on best-estimated background value using the M3 method."

Unfortunately, we ended up not performing bias correction to model particles, because the wind observation sites are not perfect around Riyadh to estimate a robust wind error to rotate model particles. We mentioned these limitations and other more comprehensive methods in **Sect. 2.6.1 (P12L1-9)**:

"Unfortunately, only 2 radiosonde stations around Riyadh with 3 vertical pressure levels within the PBL (and sometimes with missing data) may be insufficient to correctly interpolate the near-field vertical wind biases. However, cities with meteorological profiles sampling more levels within the PBL and higher temporal frequency in reporting observed vertical winds will be more suitable sites to retrieve the near-field wind errors. Other methods include rotation and stretching of urban plumes derived from WRF-Chem (Ye et al., 2017), similar to the rotation of X-STILT air parcels, to quantify errors in wind directions and speeds. Deng et al. (2017) sought correction of wind biases in a sophisticated manner via data assimilation. Yet, the near-field correction within X-STILT can be potentially utilized in the future as a quick bias correction to the near-field wind in LPDMs, given denser wind observations and relatively flat terrains. Therefore, we decided to reduce the potential impact of wind bias on model-data comparisons using a latitudinal integration (further in Sect. 3.5)."

8. P13L3-6:

I believe you are saying that M2H is in general lower than M3. But can you say definitively that M3 is correct, and thus M2H has the bias? Also, 0.56 ppm is not small! This can be 25-50% of the enhancement.

We agree that a mean difference of 0.56 ppm (between M2H- and M3-derived background values) is actually quite large, given small column enhancement. And, we cannot definitively say that our method is correct since the "true" background is more of an unknown value. So, we have reworded some relevant sentences (e.g., "bias" to "mean difference").

However, we would argue that M2H may not be suitable for local/urban studies and have rewritten the latter two paragraphs in **Sect. 3.3** (**P13L36-P14L17**) to objectively discuss the pros and cons of M2H and M3 methods (also pasted below):

> "We now focus on the comparison between M3 and M2H with objectively analyzing their advantages and limitations. On average, M2H derived background is lower than our localized "overpass-specific" background by 0.55 ppm (Fig. 6e), which can primarily be attributed to different defined background regions. M3 defined the background region from the same track as the one over Riyadh, which guarantees that the background air contains variations due to long-term atmospheric transport, natural sources/sinks and $FFCO_2$ emissions except for local emissions (e.g., from Riyadh). Whereas the enhanced air contains the enhancements due to local emissions on top of all the information included in the background air. Therefore, the subtraction between M3-defined background and enhanced air correctly represent the XCO_2 portion enhanced by the local emissions. On the contrary, M2H use a fairly broad background region (0° N–60° N, 15° W–60° E in Fig. S7) to estimate gridded anomalies over all places in Europe, Middle East and North Africa. Although may yield more data, this broad spatial region may misrepresent the correct upwind region, because the wind regime can be quite different among different overpass dates or seasons.
>
> We admit M3-defined background range and background value can be affect by potential large wind bias over cities other than Riyadh. However, the impact on background may be small and is implicitly considered in the background uncertainty (previously discussed in the last paragraph of Sect. 2.3.3). As for M2H, all regional OCO-2 measurements are lumped into its background calculation. For example, some measurements on the east-most overpass in Fig. S7 are affected by Riyadh's emissions, whereas atmospheric columns at soundings along the west two overpasses in Fig. S7 may not necessarily be the background air that eventually arrives at region around Riyadh. Thus, the regional median of XCO_2 may not physically indicate the accurate background that is supposed to isolate local-scale fluxes. Therefore, our localized overpass-specific background is designed and more suitable for extracting local-scale XCO_2 anomalies. Given relatively small urban enhancements around our study site, this 0.55 ppm difference may lead to large differences in estimated observed urban signals and emission evaluations (Sect. 4.2)."

Lastly, we may admit that our definition (using one mean value to represent the background) is not perfect and are aware of the potential wind bias impact on background definition. So, we made several attempts in this work:

1) by widening the polluted latitude range by bringing a wind error component (Sect. 2.3.3);
2) by trying to correct the wind bias on forward plume via a near-field correction (Appendix C, better for regions with denser wind observations); and
3) by introducing background uncertainty as the combined error impacts from retrieval error and natural variation (SD) of observed XCO_2 over background latitude range (Sect. 2.3.3). Retrieval errors of measurements over the background range have now been included in the background error as well, based on a comment from the Reviewer #2.

Relevant discussions based on above point 1) to point 3) have been made to **Sect. 2.3.3** on **P9L17-28** (also pasted here):

> "In addition to random errors (that are resolved by the inclusion of the aforementioned wind error component and broadening of the city plume), potential large bias in near-field wind direction may lead to mismatch in modeled and observed background regions and may bring relatively higher XCO_2 values into background XCO_2. However, we do not explicitly account for the potential near-field wind bias's impact on forward-trajectories defined urban plume with following considerations. Firstly, we attempted to propagate a near-field wind bias into the modeled plume by rotating

forward trajectories, whereas the robustness of this near-field bias can be affected by the very few wind measurements near Riyadh (further explained in Sect. 2.6.1). Secondly, the background latitude range defined by M3 with the broadening effect (blue lines in Fig. 5b) in general matches well with that observed from OCO-2 for most overpasses, which implies that the overall wind bias around our study site is not significant. Lastly, even if potential wind bias may result in less accurate background range and bring elevated XCO_2 into the background, the background uncertainty implicitly contains information about the spatial variation in background measurements (green ribbon in Fig. 5b). In addition, the M3-derived background is the mean value of mostly hundreds of background observations (numbers in Fig. 6e), which may not be greatly affected by a few potential urban-enhanced measurements."

9. *There are several instances of omitted definite and indefinite articles, and a few typos here and there, but I assume that once this paper has been accepted, the copy editor will find and correct them more thoroughly than I have. However, I will list the ones I caught here. Between ** are the edits I suggest.*
*P2L17: top-down constrain*ts* P3L6-7: shed light** on CO2 emission** monitoring network*s*. P4L18: Riyadh*,* with *a* population P4L20: Saudi Arabia has the largest CO2 emission*s* among... P4L32: "apple*s*-to-apple*s*" P4L33: weighted using *the* satellite's *column averaging kernels*... P5L17: at the same lat/lon as *the* satellite... P5L21: compare ** overall modeled P5L28: *The l*onger the time an air parcel..., *the* higher its footprint value... P6L5: FFCO2 *are* derive*d* from... P6L13: we binned ** the observed... P6L16: estimate *the* increase in observed... P6L22: 1x1 km resolution on ** monthly scale*s*... emission estimates by fuel type** from the... specific ODIAC emission categories on *a* monthly basis... P6L31: line sources and diffuse** sources... P8L32: which are more straightforward** and efficient** than solely *relying* on... P9L7: boundary of *the* city... P10L19: more suitable sites *to* retrieve... P10L25: get around ** the impact on... P11L6: we fit *an* exponential variogram... P12L11: which results in *an* overall smaller footprint... Yet, column foot- print*s* cover**... P12L14: an air column can be one or *a* few orders... P12L18: regardless *of the* adopted meteorological fields. P12L33: Here we emphasi*ze*... P13L29: sensible → sensitive P14L21: according to *the* OCO-2 Lite file... P14L24: scatter*ed* P15L36: latitudinal** integration ←this happens in other locations as well P16L30: exceeding *a* certain averaged... P17L8: emissions of *a* target city P18L2: even large impact*s* on *the* posterior... can be caused by using *a* background derived from simplistic statistic*s* P18L13: hampered ** due to... P18L24: improves biospheric flux** estimation... P20L10: for *a* few levels... P21L17: These small changes *show* that our *latitude band integration*... a second peak or miss** large XCO2 enhancements. P21L21: *widths* P21L27: The word "benefit" seems out of place here. P21L30: Based on three simpl*e* tests...*

We sincerely thank the reviewer in thoroughly reading the manuscript and pointing out these issues listed above. We have corrected them all in the relevant text.

Response to Reviewer 2

We thank the two reviewers for their efforts and constructive comments (https://www.geosci-model-dev-discuss.net/gmd-2018-123/#discussion). Each reviewer's comments are shown below in *italics*, followed by our point-by-point responses in blue.

Anonymous Referee #2

This study is timely as the OCO2 satellite has begun producing data and relevant analyses are being conducted. I think the manuscript can contribute to the OCO2 community and, in general, the GHG community as well. The author did a lot of work including different sensitivity tests, and I think this work deserves publication after addressing issues I raise below.

We thank Reviewer #2 for the positive feedback and constructive comments, which help improve both the scientific contents and the flow of this manuscript. In general, we identified several main concerns raised by Reviewer #2, in terms of 1) the flow of the manuscript, 2) transport error analysis, 3) background estimates, and 4) bias in wind direction and its impact on background estimates.

We have tried to address each comment and make clarifications/modifications to the manuscript accordingly. Also, we recently merged the X-STILT model codes with the newer version of STILT (i.e., STILT-R version 2 by Fasoli et al., 2018) and updated figures and results.

Main Comments

1. *The paper covers a lot of aspects of comparing modeled column simulations and observations. The main manuscript is long and sometimes deviates from the main story to tell; even boring although this paper is technical by nature and the information can be useful. I recommend that the authors remove some sections and technical details to the Supplement and consolidate the main text for a coherent story. Another issue is that the authors do not link the text with figures well; some of the figure captions are enormously long. In many places, the authors finish the sentences with "see Fig. X" without explaining the content of the figure well enough. I strongly recommend that the authors identify more important results (even move some figures to the Supplement, e.g., Figure 4 or 5, 7) and convey those main results with more care and clarity; please explain the figures! For example, Section 3.2 is useful (I am glad that the authors did this), but not essential for the main story given the length of the manuscript. The authors can spend the space (after moving some details) in explaining figures associated with the main results.*

We thank the reviewer for these valuable comments and suggestions that help better re-organize our manuscript. We have moved several figures from the main text to the supplement and modified the legend of almost every figure by removing redundant sentences. More explanations in the main text have now been added when explaining a figure (e.g., "red circles in Fig. X"). We removed some less important results and replaced with more important analysis and discussions suggested by both reviewers. **Table 1** is now added to summarize main results from signal calculations and error quantifications.

Still, we would like to keep some content in the main text, e.g., Section 3.2 and Fig. 3, as they visually show the modifications of X-STILT from STILT and may help readers easily understand the upwind surface influence onto a downwind atmospheric column.

2. *Third, I am not quite satisfied with the transport error analysis. The problem is that the errors (mostly winds) for WRF and GDAS are not clearly defined, so it is hard to understand how good or bad the transport is and how the error can be related to signals (e.g., low winds to high signals or the impact of wrong wind directions – not presented clearly). The authors spend a lot of space to explain transport but it needs some improvement. Referring to the unpublished paper too much is not a good idea.*

We apologize for the lack of clarity in the transport error analysis. The definition of horizontal wind errors of meteorological fields is similar to that in Lin and Gerbig (2005) and was provided in **Appendix B**:

"In terms of the wind error component ($\boldsymbol{u_\varepsilon}$) mentioned in Sect. 2.6, two sets of parameters are used to describe, 1) σ_{uverr}, the standard deviation of horizontal wind errors (RMSE) describing to what extent should we randomly perturb air parcels; and 2) horizontal and vertical length-scales and time-scales (Lx, Lz, and Lt) determining how wind errors are correlated and decayed in space and time. We calculated different sets of wind error statistics over 3 vertical bins, i.e., 0–3 km, 3–6 km and 6–10 km, for randomizing air parcels. To obtain σ_{uverr}, observed winds at mandatory levels (i.e., 925, 850, 700, 500, 400, 300 mb) from surrounding radiosonde sites (Fig. 4) are compared against WRF- or GDAS-interpolated winds. Then, we averaged wind errors at different mandatory levels over aforementioned three vertical bins. In addition, wind errors are considered to be spatiotemporally correlated. To determine error correlation scales, differences in the wind errors are calculated and wind errors at different radiosonde stations or different reported hours (00UTC or 12UTC) are paired up based on their separation length- or time-scales. An exponential variogram is then applied to estimate the horizontal, vertical and temporal correlation scales, which are the separation scales when errors become statistically uncorrelated."

The wind error and transport error statistics over the five overpasses we focused are now summarized in **Table 1**. Wind error statistics (RMSE for lower atmosphere, 0-3km) for several overpasses for Riyadh are labeled as numbers in **Fig. S1**. Additionally, we now add a new set of analysis and subsection about the vertical transport errors, via propagating typical PBL errors into the model, as part of our response to Reviewer #1. Please refer to **Sect. 2.6.2** for the changes.

We agree with the reviewer that a lot of numbers/statistics were listed for the transport error analysis without explicitly discussing the linkage from errors in wind speed to XCO_2 signals in Sect. 3.4 and 3.5 (in the previous paper version). Now, we have added a paragraph in **Sect. 3.5 (P16L22-35)** to discuss this linkage and removed some sentences in simply listing numbers/statistics.

Here are some other main points about the XCO_2 transport errors:

- No large systematic errors in u- and v- component wind is discovered over dozens of overpasses (Fig. S1).
- For each sounding, XCO_2 errors due to the horizontal transport error are calculated from the CO_2 variance differences between the standard trajectory and the perturbed trajectory, for each level. More details regarding the transport error quantification at each model level and for each sounding can be found in **Appendix B**.

- For each overpass, the latitude-integrated XCO_2 error due to horizontal transport is a mixture of several factors. Relevant text has been added to Sect. 3.5:

  "The integrated XCO_2 transport error per track reflects the aggregate effect of several factors which interact, given how we propagate wind errors into XCO_2 space (Sect. 2.6):

  1) The magnitude of the modeled urban XCO_2 enhancements. In general, air parcels that are very far away from potential upstream emitters may hardly "hit" the emission sources or gain their enhancements, even after the wind perturbation. If the estimated signal is large (e.g., 3.04 ppm-deg. on 20151216 in Table 1), its resultant integrated transport error can also be fairly large (1.83 ppm-deg. in Table 1).
  2) The RMSE of u- and v-component winds. In general, larger wind errors will lead to larger changes in model trajectories and larger possibilities for perturbed trajectories in intersecting an emission source.
  3) How air parcels interact with surface emissions, i.e., the geometry/angle between the model footprint (or the wind direction) and satellite swaths. Changes in this angle may fluctuate the width of enhanced latitudinal band along with the final integration latitudinal ranges (i.e., 1.10°–2.25°). If the back-trajectory or backward wind direction is more parallel to the OCO-2 swath (events on 20141227, 20151216 and 20160216 in Fig. S10), the integration range and error covariance among soundings are usually larger, which yields larger integrated XCO_2 errors (e.g., 1.22, 1.83, and 1.05 ppm-degree in Table 1). The averaged latitudinal range for integration is about 1.66° (~189 km) over 5 tracks."

3. *Last, I would like the authors to comment on the utility of OCO2 for urban studies based on this work, because there is some skepticism about OCO2's capability for estimating urban emissions with relatively small areas.*

We addressed this concern in the last paragraph of **Sect. 4.4**. We briefly mentioned the limits of using OCO-2 on urban studies, such as limited temporal coverage or limited screened observations. We expect more data and diurnal variations after the launch of OCO-3 and its orbit on the International Space Station.

"OCO-2 observations have been utilized in several recent studies along with this work with a particular look into relatively small areas, e.g., individual power plants (Nasser et al., 2017) and megacities (Ye et al., 2017). Even though the XCO_2 urban signal over Riyadh may be in general smaller than those over other large cities, both model and observation successfully detect the urban signal. Still, no summertime XCO_2 signal has been derived, due to the lack of screened observations (QF = 0) reported in OCO-2 Lite b7 file over most summertime tracks (black bars in Fig. S1). No diurnal variation, revisit time of 16 days and relatively narrow swath of OCO-2 may still pose challenges to urban emission estimates. We expect the inclusion of more column observations in stationary (target) modes, e.g., by scanning over megacities by OCO-3 (Eldering et al., 2016), which may offer more concrete spatial and diurnal variabilities that benefits urban flux inversions. Many nations are devoting considerable resources in launching carbon-observing satellites that can potentially be coordinated in a larger monitoring system (Tollefson, 2016). Given that X-STILT can potentially work with most satellites (given their sensor-specific vertical profiles), we expect enhanced capability in emission constraints of urban emissions by combining column measurements with X-STILT."

Detailed Comments

4. P1, L19. *Global assimilation data seems to be too coarse for the urban scale CO2 simulation. Why use GDAS?*

The reviewer raised five detailed comments (including *comment **4, 13, 14, 27** and **28***) related to the meteorological field we used. We address them altogether here.

Yes—GDAS is the primary choice in this study. STILT trajectories over all five overpasses are guided by meteorological fields from GDAS. Although the spatial resolution of GDAS (0.5 degree) is coarser than WRF customized in this study, GDAS is the main choice in this work due to the following considerations:

1) The surrounding terrain around Riyadh is relatively flat. For other cities with complex terrain, we may have two options. If we still use global assimilation data, the model may likely "return" larger wind errors and resultant XCO_2 errors around the best estimates. Alternatively, we always have the option to use higher resolution meteorological fields, e.g., customized WRF or HRRR, to better resolve the subgrid scale dynamics and terrain flows with more accurate estimates in ground heights.

2) The regional wind error statistics (compared against observed winds from radiosonde stations) of GDAS is similar to that of WRF for the few cases we examined (Table 1). The reviewer or readers may be concerned about the wind error quantification, as the number of observation sites around the city may not be that large (e.g., comment #27). However, we discussed the pros (i.e., less cloud and vegetation coverage) and cons (i.e., sparser wind observation network) of choosing such city like Riyadh in **Sect. 1 (P4L24-27)**:

"Riyadh, with a population of over 6 million by 2014 (WUP 2014), is chosen as the city of interest because of its low cloud interference, limited vegetation coverage, and isolated location in a barren area, which leads to higher data recovery rates and facilitates the background determination. Saudi Arabia has the largest CO_2 emissions among Middle Eastern countries and ranks eighth globally in 2016 (Boden et al., 2017; BP, 2017; UNFCCC, 2017)."

and in **Sect. 4.4 (P19L15-24):**

"Admittedly, the transport error analysis and near-field correction may work the best with the assistance of denser meteorological observing networks to characterize the error structures of transport errors. Increasing the density of surface networks may modify the wind error statistics including the wind error variances and horizontal

correlated length-scale, and further impact the model transport uncertainties and inversed fluxes. Yet, this shortcoming is not inherent to X-STILT and applies to other means of quantifying the transport errors based on real data as well. The trade-off of choosing a city in the Middle East like Riyadh to minimize cloud and vegetation influences is the relatively sparse observations of surface meteorological network or aircraft. The most recent OCO-2 b8 Lite files include retrieved surface winds for each sounding. Unfortunately, most of those surface wind retrievals are not available over Riyadh, but the retrieved surface winds for other urban areas, if available, may be used for assimilation and assisting X-STILT error analysis."

3) Our ultimate goal is to look at emissions from a couple of cities over the Middle East or even around the world with the assistance of X-STILT in future studies. Customized WRF field for many more cities and overpasses can be relatively computational more expensive and require careful evaluations on their configurations over different regions.

4) Lastly, the scope of this manuscript is to present a modified atmospheric transport model framework with an application over a relatively "simple" city. Our intention is not about evaluating the differences between two meteorological fields and making conclusions about which one is better (comment #28). And the STILT model itself is not fixated to a particular choice of meteorological field.

5. P1, L21.

"68% in posterior scaling factor" should be "68% in posterior signal" because here the bias in background is in the units of signal. Also, it is not clear what 68 % in posterior scaling factor means. Posterior uncertainty in 1-sigma? Or Does it mean the bias in background resulted in 68% higher or lower bias in the posterior scaling factor?

Clarifications: Our intention is to reveal or highlight the impact of different background methods on the posterior scaling factor ($\hat{\lambda}$ in Table 1). The posterior scaling factors (for mean XCO_2 signal) using M2H- and M3- derived observed signals are ~1.78 and ~1.14, as shown in Table 1. We now reworded **P1L21-23** as:

"In addition, a sizeable mean difference of -0.55 ppm in background derived from a previous study employing simple statistics (regional daily median) leads to a higher mean observed urban signal by ~39 % and a larger posterior scaling factor."

6. P1, L22.

It seems to me that the authors are referring to signal calculation, and the impact of uncertainty and bias on the urban signals by "Based on these results". I wonder if the authors can add a couple sentences that are more significant than these. If I put it differently, are these results the most important results we take home from this study?

Yes—the goal of this study is to provide a modified version of STILT for column measurements and associated error quantifications (with a case study over a city in the Middle East). We have changed 'Based on these results' to 'Based on our signal estimates and associated error impacts' on **P1L23**.

7. P3, L31.

Please add references related to "minimal guidance". The authors can simply add few references on uncertainties associated with atmospheric column simulations.

We have reworded the sentence. Although the error impact from receptor setups can be small, most studies simply depicted their model setups without further explaining why they chose those setups or the error impact (due to model configurations) on modeling XCO_2. No study examined this error impact on column simulations, to our best knowledge. Text has been changed to as:

"Previous studies reported negligible to ~20 % of the modeled enhancements are reported as the error impact due to STILT particle number (released from a fixed level), depending on adopted particle numbers, examined species and their components/sources (Zhao et al., 2009; Gerbig et al., 2003; Mallia et al., 2015). When it comes to representing

an atmospheric column using particle ensembles, many studies depicted their setups for receptors/particles without further explaining why they chose those setups or the error impact (due to model configurations) on modeling XCO_2. Although this error impact may be small, we still perform a set of sensitivity tests to provide more guidance on placing column receptors."

8. P3, L33 – 34:

The authors underestimate recent developments in inverse modeling. There are several atmospheric inverse studies that consider transport errors and use full error matrix (not just diagonal), in particular non-CO2 studies (e.g., regional methane studies). The references there are old and does not support the statement. The authors need to be specific. I may agree that there are not many studies to incorporate full error characterizations for column-observation inversion studies, but there are now many studies to consider errors more carefully. The authors should be careful in this statement and need corrections.

We now add a few more references on recent urban CO_2 and regional methane inverse studies, e.g., Jeong et al., 2013, Lauvaux et al., 2016 and Zhao et al., 2009. And, we have changed some of our statements regarding the full error characterizations for column inversion studies. Text in **Sect. 1 (P3L35-P4L2)** has been reworded as:

"Approaches to quantify errors in horizontal wind fields and vertical mixing have been proposed followed by comprehensive error characterizations on atmospheric simulations (Gerbig et al., 2008; Jeong et al., 2013; Lauvaux et al., 2016; Lin and Gerbig, 2005; Zhao et al., 2009). Recent efforts (e.g., Lauvaux and Davis, 2014; Ye et al., 2017) have been made to rigorously examine the column transport errors."

Also, I am surprised that the authors use a very simple inversion – later in the section I find they are not well formulated but rudimentary – I don't see the benefit of including the inversion result in the study. Please note that there are many sophisticated inversion methods that are much more amenable to error characterizations – please do some literature review.

We appreciate the criticism on the simple inversion from the reviewer. However, we note that conducting a comprehensive inversion or making conclusions about inversed urban emissions may be out of the scope of this study. This study focuses on the model descriptions for XCO_2 signal extraction and error quantifications (that helps provide insights into future comprehensive inverse studies), with a case study over Riyadh.

Two reasons for including a simple scaling factor inversion in discussion section:

1) to follow the scaling factor analysis and compare the transport error results in Ye et al. (2017), even though our methods and adopted atmospheric transport models can be different; and
2) to address the importance of background estimates and provide STILT-based error impacts (e.g., posterior covariances).

In addition, we agree with the reviewer that this is a simple inversion, probably because we treated the gridded upwind urban emissions as a whole (i.e., no adjustments for emissions for each gridcell) and integrated latitude-dependent XCO_2 enhancements. More sophisticated inversions on the spatially distributed emissions, given more sampled satellite overpasses or more sampled cities over the Middle East will be considered in future studies. However, we may justify that these simplifications are made for the consideration of reducing error impact, in particular from potential near-fields wind biases.

Relevant text in Sect. 4.2 has been modified and added:

"Estimated background uncertainty is represented by the spatial variation and retrieval errors of background observations and may be reduced given large sampling size. To further demonstrate X-STILT's potential role in inverse modeling and the potential background "bias" via different background methods on inversed results, we conducted a simple scaling factor inversion (Rodgers, 2000), based on 5 pairs of model-data latitudinally-integrated urban signals. Even though our sampling may seem to be small and the gridded urban source emissions are treated as a whole (i.e., no adjustments for emissions for each gridcell), these integrated signals and errors are chosen to reduce the impact of

potential near-field wind bias on model evaluations. Also, we are partially limited by the overpasses over Riyadh (black bars in Fig. S1)."

9. P4, L8-9:

I don't quite understand "Most of these studies aim at extracting relatively large CO2 changes at a fixed level within the PBL or due to large emissions such as of wildfire". Which studies are the authors referring to? The point is tower vs. column or large signal vs. small? Are the authors suggesting that the study site in this work has very little CO2 changes (exchanges?)? The study areas in this study are different from other urban areas in previous studies in terms of CO2 variations or signal-to-noise ratio? Also, related to this, why did the authors choose this study area instead of some US large cities?

We regret the confusions caused by these lines. The main point of this paragraph is to point out some common ways to define background e.g., the trajectory endpoint method, as well as the limitations of those modeled background, especially when trying to estimate background from column observations. Text has been changed to as:

"The aforementioned studies (adopting the trajectory-endpoint method) aim at extracting relatively large CO_2 anomalies (e.g., at a fixed level within the PBL or due to large emissions such as of wildfire) out of the total measured CO_2."

No — we are not suggesting the study site in this work is special or the CO_2 changes for this site is low. We just wanted to bring up the difference in extracting urban enhancements from PBL-based or column observations, where the enhancements are relatively larger and smaller by nature. Because the relatively small column enhancement and SNR when extracting the column signal, even a small error in background as low as 1 ppm can be "harmful" for interpreting XCO_2 variation.

For the reason of choosing Riyadh rather than other large cities in US, we explained in **P4L23-27**: "its low cloud interference, limited vegetation coverage, and isolated location in a barren area, which leads to higher data recovery rates and facilitates the background determination." And, we will expand our study area to examine more cities around the world in future work.

10. P4, L13:

It is not clear why the authors introduce a new background estimation method. I guess this has to do with column simulations, but please state the reason more clearly.

Yes—we introduce a new method because of the relatively small SNR in extracting urban enhancements (of few ppm) out of total XCO_2 concentration as well as limitations of some other methods. We added a very brief limitation of trajectory-endpoint method for column simulation on **P4L10-12:**

"However, for studying XCO_2 that is less variable than near-surface CO_2 (Olsen and Randerson, 2004), potential errors in modeled concentration fields and atmospheric transport may pose more significant adverse impact on derived urban signals."

and now add the limitation of simple statistics on **P4L14-15**:

"These simple statistical methods often neglect the transport and may use the less accurate spatial region to select measurements for deriving background values."

We further discussed these background methods in **Sect. 2.3 and 3.3**.

11. P4, L28:

Please define "prior profile" since many "priors" are used in this paper.

"Prior profile" stands for the "a priori CO_2 profile" from OCO-2 Lite product. Text has been made on **P5L5**.

12. P5, L4:

It seems "ratios of the pressure difference between adjacent model levels over that between adjacent retrieval levels" needs more clarification. Once PW is interpolated to model levels, then the pressure difference between model levels (as the scaling factor) should be enough? Please clarify.

PW function is primarily estimated based on the air mass or pressure difference (*dp*) between two layers. Fig. 2b shows that OCO-2 levels have relatively **constant pressure difference** (*dp_oco2*) and *PW* values (in black dots, except for the very top and bottom level). On the contrary, model levels are finer (in orange circles in Fig. 2b) with two different vertical spacings in **altitudes** (100 m vs. 500 m) below and above 3 km. So, those scaling factors are to adjust the *PW* function according to difference in *dp_oco2* vs. *dp_stilt*.

For example, from 0 to 3 km, *dp_stilt* ranges from ~8-10 mb, while *dp_oco2* are mostly ~52 mb (red circles vs. black dots in Fig. 2b). Thus, we further scaled down the interpolated *PW* (red circles in Fig. 2b) by the ratio of *dp_stilt* over *dp_oco2* (e.g., sf = 10 mb/52 mb), because of less air between model levels than initial retrieval grids. Thus, final *PW* after scaling has value of ~0.01 (orange circles in Fig. 2b).

Comparing air below 3km, fewer model levels are placed from 3-6 km (~650 to 450 mb), which gives larger *dp_stilt*, larger *PW* scaling factors and resultant larger scaled *PW* of ~0.03 –0.04 (orange circles in Fig. 2b).

Since no model level placed above MAXAGL, *PW* stay the same as the initial OCO-2 *PW* values (blue circles in Fig. 2b). Finally, we made sure that the sum of vertical PW profile ends up being 1 for each sounding/receptor.

Relevant text has been clarified (in Sect. 2.1 on **P5L14-22**):

> "Interpolations are further needed to resolve the mismatch between prescribed OCO-2 retrieval grids and model levels for the lower part of the troposphere. Our intention is to preserve the finer modeled CO_2 variations by performing interpolations of satellite profiles from retrieval grids to model levels. Vertical profiles of AK_{norm}, *PW* and $CO_{2,prior}$ are treated as continuous functions and interpolated linearly to model grids (red circles in Fig. 2). Note that the initial OCO-2 *PW* functions have steady value of ~0.052 (except for the very bottom and top levels; black dots in Fig. 2b), which results from constant pressure spacings (*dp_oco2*) between two adjacent OCO-2 levels. However, X-STILT levels are much denser with smaller pressure spacings (*dp_stilt*) or less airmass between their two adjacent levels. Therefore, the linearly interpolated *PW* (red circles in Fig. 2b) needs an additional scaling via a set of "scaling factors" representing the ratios of pressure spacings in STILT versus OCO-2 retrieval (*dp_stilt/dp_oco2)*, to arrive at the correct *PW* for each finer model grid (orange circles in Fig. 2b)."

13. P5, L12:

I wonder what "When WRF fields were available" means. WRF is not used for all days/hours? For the comment on the abstract, I added that GDAS alone is not sufficient for the urban scale. Also, more importantly, the authors must add the minimum description of the WRF model, e.g., vertical and horizonal resolutions unless stated somewhere later in the sections. It is not appropriate to toss everything to another unpublished reference.

STILT trajectories over all five overpasses are guided by meteorological fields from GDAS. These customized WRF fields can be computational expensive and require careful evaluations on its configurations. Thus, model trajectories for the first two overpasses (i.e., 12/27/2014 and 12/29/2014) are driven by nested WRF and GDAS fields.

We have added a brief description on WRF configurations in **Sect. 2.1.1 (P5L35-P6L1**): "Hourly WRF fields contain 51 vertical levels with boundary conditions from 6-hourly 0.5°×0.5° NCEP FNL (Final) Operational Global Analysis data (Ye et al., 2017) are customized and utilized for the first 2 of the total 5 overpasses over Riyadh."

14. P5, L15: *Is GDAS the primary choice?*

Yes. For the reason of choosing GDAS, please refer to our response to comment #4.

15. P5, L19:

Remove "a certain height", but directly use an explicit one, e.g., "the maximum release height" - unnecessary vagueness. I see a few places in this paper that use such a vague expression.

We have replaced 'a certain height' with 'the maximum release height' or 'MAXAGL'.

16. P5, L20:

Please state what constitutes "different setups" so that the reader has a clear sense of the setups that might differ. As written, it is not clear.

Different setups comprise of the maximum release level (MAXAGL), the vertical spacing of release levels (dh), and the particle number per level (dpar), which is now clarified in the main text (**P6L8-9**).

17. P6, L26: *Define "BP".*

BP is the acronym for the British Petroleum Company plc and BP Amoco plc (an oil and gas company). We add it in the main text.

18. P7, L3:

Please say so, if 0.1 degree is the final resolution for signal calculation, which could be coarse for a urban region.

Clarifications: We still kept 1km x 1km anthropogenic emissions from ODIAC and generated 1x1km footprints to calculate the XCO_2 signal. The 1km x 1km should be fine for getting XCO_2 signals from the urban. We further clarified this point on P9L20 ("To calculate modeled XCO2 enhancements, we used the latest (year 2017) version of …")

However, when it comes to emission uncertainty calculations, emissions from ODIAC are aggregated to 0.1° (due to mismatches in the horizontal resolutions of emission grids). Thus, another set of footprints with 0.1° x 0.1° spatial resolution is generated to convolve with the 0.1° x 0.1° spatial emission uncertainty, which propagates the errors in prior emissions to the XCO_2 space (to the 1st order).

19. P7, L10:

Please comment on the 1-degree bio flux relative to the size of the study area and its potential impact (due to coarse resolution) on the inversion.

We agree with the reviewer the 1° x 1° CarbonTracker can be comparable to the size of the urban domain and too coarse to resolve the subgrid scale heterogeneity in biospheric fluxes. However, the potential impact on inversion or the signal calculation is small due to following reasons. And, we did not modify the main text.

1) For the inversion and signal calculations, we actually used the overpass-specific background from M3, instead of the trajectory-endpoint based background that relies on CarbonTracker biospheric fluxes.

2) The biospheric influence has been included over the background latitude range and then get subtracted from the total observed XCO_2. M3 may work fine, unless large gradient of biospheric fluxes exists around the urban area (mentioned as a potential limitation of M3 in **Sect. 4.4, P19L1-2**):

   "When examining summertime tracks or tracks over some other cities, potential local gradients in biospheric fluxes should be considered as those gradients can affect our overpass-specific background."

3) Lastly, the land around Riyadh is relatively barren with minimal biomass coverage. For studying other cities, we can use biospheric fluxes with finer spatial resolution generated from other inventories/models, e.g., MsTMIP.

20. P8, L22:

Please add comments on the potential impact of transport over the city when using Method 3. I note that the authors discussed the potential transport error for Method 1 (i.e., endpoint method). Wind direction could be a serious problem for Method 3. Enough overpasses (both up- and down-wind) are available for Method 3.

We agree with the reviewer that It is possible that higher XCO_2 values may be included in the background ranges, due to mismatches between modeled and observed plumes. We have added a paragraph in **Sect. 2.3.3 (P9L17-28)** to discuss this impact on background value (also pasted here):

"In addition to random errors (that are resolved by the inclusion of the aforementioned wind error component and broadening of the city plume), potential large bias in near-field wind direction may lead to mismatch in modeled and observed background regions and may bring relatively higher XCO_2 values into background XCO_2. However, we do not explicitly account for the potential near-field wind bias's impact on forward-trajectories defined urban plume with following considerations. Firstly, we attempted to propagate a near-field wind bias into the modeled plume by rotating forward trajectories, whereas the robustness of this near-field bias can be affected by the very few wind measurements near Riyadh (further explained in Sect. 2.6.1). Secondly, the background latitude range defined by M3 with the broadening effect (blue lines in Fig. 5b) in general matches well with that observed from OCO-2 for most overpasses, which implies that the overall wind bias around our study site is not significant. Lastly, even if potential wind bias may result in less accurate background range and bring elevated XCO_2 into the background, the background uncertainty implicitly contains information about the spatial variation in background measurements (green ribbon in Fig. 5b). In addition, the M3-derived background is the mean value of mostly hundreds of background observations (numbers in Fig. 6e), which may not be greatly affected by a few potential urban-enhanced measurements."

21. P9, L20:

I agree with the authors that STILT configurations can affect the results. But I don't understand the use of bootstrapping here. The original sample here is from the 401 levels (too many in my opinion). However, what we are interested in is the results from different set-ups, e.g., 20, 40, levels, which can be different from the original samples of the 401 levels. In practice, 401 levels are unrealistic, e.g., for annual analysis.

Clarifications: Yes— what we are interested in is the results from different setups, e.g., whether 20 or 40 levels can be enough. The number of levels (nlevel) is further decomposed into the vertical spacing dh and the MAXAGL. By increasing dh from 50 m to 100 m, the number of levels reduces by half with fixed MAXAGL of 6 km.

Note that the original sample is release from 0 to 10 km with a spacing of 25 m (n = 1, 2, 3, …, 401). For example, if we wanted to test the one case with dh = 50m, we randomly resampled trajectories released from every other level from 0 to 6 km (n = 1, 3, 5, …, 241) for 100 times. In other words, we got 100 sets of resampled trajectories with the same combination (MAXAGL = 6km, dpar = 100 and dh = 50m). From those 100 new sets of trajectories and resultant 100 XCO_2 enhancements, we calculated mean and SD of those enhancements. SD is used to reveal the random uncertainty (error bar in Fig. 5c), while the mean for one case is compared with other means, to reveal any systematic bias (e.g., the decreasing trend of red dots in Fig. 5c). Therefore, we actually do not care about difference between the resampled trajectories against the original sample.

22. P11, L15:

It is surprising that MAXAGL < 2.5 km did not fully capture CO2 enhancements. I would expect that there is not much surface influence above 2 km. Is it because the study region is associated with really high PBLH? As the authors stated in L30-32, the lower portion of the column should matter most. Then why would MAXAGL of ∼2.5 km not capture the full enhancement of CO2? Please add sentences that dis- cuss the reason for this. Actually, looking at Figure 8(a), I realize that there are only two cases below 2.5 km. So, 2.5 km itself looks fine. My guess is that even 2 km should be fine. I think the authors give the reader somewhat wrong information here, considering the fact that using a higher altitude for MAXAGL increases the computational cost significantly. My understanding from this is: 1) use 100 – 200 m vertical resolutions be- tween 0 – 2 km and 2) above 2 km,

use 500 m. If the authors can show even MAXAGL of 2 km is comparable to 2.5 or 3 km, this will reduce the computational cost significantly. I don't understand why the authors use 100 m for up to 3 km given the result shown in Figure 8(a), which in my opinion is too much without good reasoning. I think that some other studies will easily show denser vertical resolutions between 0 – 2 km is good enough.

We thank the reviewer in pointing out this detail and now add the one case with MAXAGL of 2 km in Sect. 3.1 and Fig. 6a. The one simulation using MAXAGL of 2km looks much better than the one with 1.5km MAXAGL, but can be slightly lower than simulation using even larger MAXAGL. We have further run another set of uneven vertical spacing test to see whether a cutoff level of 2 km is enough.

Results are added and explained in **Sect. 3.1 (P13L3-10**, also pasted as below):

> "We further performed two cases with uneven vertical spacing below and above a "cutoff level". Both tested three different lower spacings (of 50, 100 or 150 m) with a fixed upper spacing of 500 m. Two cases differ only in their cutoff levels (2 or 3 km). The comparison of the uneven *dh* against the constant *dh* experiment shows that their results in XCO_2 enhancements are fairly similar, suggesting that the lower spacing below the cutoff level matters mostly to model results, because most anthropogenic XCO_2 enhancements are confined within the PBL. Also, results for uneven *dh* case with the cutoff level of 3 km (blue triangles in Fig. 6c) are more closed to the "truth" implied by the constant *dh* case (red dots in Fig. 6c). To be safe, column receptors are placed from 0–3 km with a spacing of 100 m and from 3–6 km with a spacing of 500 m."

Yes—the PBLHs or mixing height are generally high over the upwind region of our city. Information about model-interpolated mixing depths can be found in Appendix D.

23. P11, L34:

Please clarify what the fractional uncertainty means here. How did the total particle number become >12500 with 100 particle every 100 m within 3 km?

Clarifications: The fractional uncertainty is calculated as the ratio of random uncertainty (in ppm, error bar in Fig. 6a-c) over the averaged simulated enhancement (in ppm, red dots in Fig. 6a-c) of results by resampling trajectories for 100 times. We have now clarified the fractional error in **Sect. 2.5 (P10L31-34)** as well:

> "100 urban enhancements are calculated from 100 new sets of trajectories for each test. Basic statistics—i.e., mean values and standard deviations (or fractional uncertainty, i.e., SD/mean) among these 100 enhancements—are used to infer systematic and random uncertainties in each test, respectively (with results showed in Sect. 3.1)."

For testing the sensitivity of XCO_2 due to changes in one receptor parameter, the other two parameters are fixed.
Specifically, dpar = 100 and dh = 100m are used for testing different MAXAGLs from 1 to 10 km (Fig. 6a);
dh = 100m and MAXAGL = 6km are used for testing different dpar (Fig. 6b);
dpar = 100 and MAXAGL = 6km are used for testing different dh (Fig. 6c).
Thus, the one simulation (dpar = 100, dh = 50 m, MAXAGL = 6 km) has 12,000 particles.

We now clarify the use of fixed MAXAGL of 6km for dpar test in **Sect. 3.1 (P12L37)** and in Fig. 6a-c:

> "In addition, we conducted two experiments using constant and uneven vertical spacings with the fixed MAXAGL of 6 km and *dpar* of 100."

24. P12, L32: *"incorporates both" to "both incorporates"*

Text changed.

25. P12, L37:

I wonder what "we added a wind error component to broaden the urban plume (Sect. 2.4.3 and Sect. 2.6) that helps reduce the inclusion of enhanced values in the background region" means. I can understand this could help reduce strong local sources under the assumption that broadening plumes with additional errors reproduces the reality more accurately. But broadened background does not necessarily solve the bias in the wind direction that is directly related to the enhancement in the background region.

We agree with the reviewer and are aware of the impact of this wind error component onto our defined background values. Please refer to our response for comment #20.

26. P13, L3-6:

How did the authors judge which one is the more accurate background that is assumed to be close to the (unknown) true background? The impact of the background bias (0.56 ppm here) on the emission estimation depends on the magnitude of the observation; it can have only a small impact when the local observations are large.

We agree with the reviewer that it can be different to judge which method is the "truth", since the background value is an unknown and our examined sample size could be small.

However, we would still argue that M2H may not be suitable for local/urban studies, like this study. We have now added two paragraphs to try to discussion the limitations and advantages of M2H and M3 in **Sect. 3.3 (P13L37 – P14L18**, pasted as below):

"We now focus on the comparison between M3 and M2H with objectively analyzing their advantages and limitations. On average, M2H derived background is lower than our localized "overpass-specific" background by 0.55 ppm (Fig. 6e), which can primarily be attributed to different defined background regions. M3 defined the background region from the same track as the one over Riyadh, which guarantees that the background air contains variations due to long-term atmospheric transport, natural sources/sinks and $FFCO_2$ emissions except for local emissions (e.g., from Riyadh). Whereas the enhanced air contains the enhancements due to local emissions on top of all the information included in the background air. Therefore, the subtraction between M3-defined background and enhanced air correctly represent the XCO_2 portion enhanced by the local emissions. On the contrary, M2H use a fairly broad background region (0° N–60° N, 15° W–60° E in Fig. S4) to estimate gridded anomalies over all places in Europe, Middle East and North Africa. Although may yield more data, this broad spatial region may misrepresent the correct upwind region, because the wind regime can be quite different among different overpass dates or seasons.

We admit M3-defined background range and background value can be affect by potential large wind bias over cities other than Riyadh. However, the impact on background may be small and is implicitly considered in the background uncertainty (previously discussed in the last paragraph in Sect. 2.3.3). As for M2H, all regional OCO-2 measurements are lumped into its background calculation. For example, some measurements on the east-most overpass in Fig. S4 are affected by Riyadh's emissions, whereas atmospheric columns at soundings along the west two overpasses in Fig. S4 may not necessarily be the background air that eventually arrives at region around Riyadh. Thus, the regional median of XCO_2 may not physically indicate the accurate background that is supposed to isolate local-scale fluxes. Therefore, our localized overpass-specific background is designed and more suitable for extracting local-scale XCO_2 anomalies. Given relatively small urban enhancements around our study site, this 0.55 ppm difference may lead to large differences in estimated observed urban signals and emission evaluations (Sect. 4.2)."

27. P13, L34:

It depends on which wind observations are used. The number of sites for wind obs. in this study is too small to make a statement as shown here.

We have changed the statement on **P14L34-36** to "Based on available radiosonde sites over the Middle East with relatively flat terrain (white crosses in Fig, 4)".

We are aware of the limitation of sparse wind observation network for this city and discussed the trade-off of choosing this city in **Sect. 4.4**:

> "Yet, this shortcoming is not inherent to X-STILT and applies to other means of quantifying the transport errors based on real data as well. The trade-off of choosing a city in the Middle East like Riyadh to minimize cloud and vegetation influences is the relatively sparse observations of surface meteorological network or aircraft."

Also, please refer to the relevant response to comment #*4*.

28. P13, Section 3.4.1

Comparisons against OCO-2 XCO2 at selected soundings: What is the small conclusion here? After all the analysis, the authors state "we suspect that mismatch in the model-data enhancement widths is primarily due to errors in wind speeds". I expected that the authors state, e.g., "model X is better or worse than model Y in terms of wind' simulations compared to observations, and we also see better or worse in model X or Y for 'signal' comparison between model simulations and observations". Any advantage of WRF due to higher resolutions?

As stated in Sect. 2.1.1 (P6L2-3):

> "We note that the primary focus is to assess the resulting errors given the choice of a particular wind field (i.e., GDAS 0.5°), rather than to carry out analyses of differences between WRF and GDAS."

Even though the shape of resultant XCO_2 contribution maps appear to be different between two models (e.g., Fig. 7b and 7f), the two latitude-integrated XCO_2 contributions (**Fig. 7d and 7h**) appear to be quite identical. The GDAS and WRF regional wind RMSEs are also listed in **Table 1**.

29. P17, L30:

How large was the random error (S_lambda) relative to the background-subtracted enhancements? The 5 x 5 error matrix (if this is the model-data mismatch error covariance, i.e., the irreducible error component in the linear model) suggests that only 5 obs were used? If it is true, that seems to be too small, even for a simple linear regression. The scaling factor suggests the prior emissions are consistent with the observation. Is this the conclusion and what the authors expect from the comparison between modeled XCO2 and obs? The description for this simple inversion doesn't sound good at all.

The random error (square root of the observational error variance; in ppm) are about 63 % to 85 % of background-subtracted enhancements for different overpasses for Riyadh. These random error per overpass are assumed to be independent (due to mostly long separation time) and reduced when aggregating over 5 overpasses. In this revised version, we further added 1) retrieval errors in the background error and 2) error in vertical mixing in the X-STILT transport error (based on a comment from reviewer 1). Thus, the random error is slightly higher than that previously reported.

Yes—we only use 5 pairs of latitude integrated observed and modeled XCO_2 signals. And various errors at each sounding have been properly aggregated to the overpass level to reduce impact from wind bias. We carefully examined every possible overpass based on number of soundings and screened soundings, wind errors and distance to the city (Fig. S1). Then, those overpasses are under manual check to see whether there's promising enhancements. Although we may be limited by our stringent criteria, we are inclined not to perform simulations or model evaluations over some other tracks with insufficient soundings.

We appreciate the criticism and agree with the reviewer that this is a simple inversion and have now commented the limitation of this simple inversion on its lack of consideration of the spatiotemporal structures in **Sect. 4.2**. We will perform more sophisticated column inversion urban studies and analysis on urban emissions in future studies.

The posterior scaling factor is about 1.14 given observed signals using M3 background, which does not suggest the prior emissions are consistent with the observations. No -- our intention is not to make conclusions about the urban emissions for Riyadh. We will perform more comprehensive inverse analysis over more cities in future studies. For reasons of conducting this simple inversion, please refer to our response for comment #8.

30. P18, L4-6:

I wonder if the background estimation for column CO2 from OCO2 can be improved. Somewhat disappointing. I hope to see some discussions (a few sentences) on the utility of OCO2 for urban studies including the retrieval error (this urban region has relatively low enhancements, difficult for OCO2 to tell something), not only for this study area, but for future other regions, more generally.

We appreciate the reviewer for this constructive comment on background estimates. We have now updated the background uncertainty by including the retrieval errors of observations over the background latitude range.

The reviewer is making a good point, and it will be great that we examined the background estimation over other urban regions. However, this may be beyond the scope of this model description paper (with application applied to a city). We will examine more urban regions given background from OCO-2 in future studies.

31. Figure 7.

The trajectories seem to be stratified, with each streak (looks like thick streak) somewhat disconnected from each other, which looks strange. Any explanation? Is it because of different levels?

Yes -- Figure 7 (now Fig S4) contains all air parcels released from different vertical levels. Air parcels at higher levels are driven by higher wind speed and different wind directions aloft than winds within the PBL. Those air parcels released from levels within the PBL are more concentrated near the receptor while parcels released from higher levels are displayed more to the west.

32. Figure 8-e*: Please use the same labels for the legend, e.g., M3.*

Have changed the label in panel e.

A Lagrangian Approach Towards Extracting Signals of Urban CO_2 Emissions from Satellite Observations of Atmospheric Column CO_2 (XCO_2): X-Stochastic Time-Inverted Lagrangian Transport model ("X-STILT v1")

Dien Wu[1], John C. Lin[1], Benjamin Fasoli[1], Tomohiro Oda[2], Xinxin Ye[3], Thomas Lauvaux[3], Emily G. Yang[4], Eric A. Kort[4]

[1]Department of Atmospheric Sciences, University of Utah, Salt Lake City, USA
[2]Goddard Earth Sciences Technology and Research, Universities Space Research Association, Columbia, Maryland/Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA
[3]Department of Meteorology and Atmospheric Science, Pennsylvania State University, USA
[4]Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, USA

Correspondence to: Dien Wu (Dien.Wu@utah.edu)

Abstract. Urban regions are responsible for emitting significant amounts of fossil fuel carbon dioxide ($FFCO_2$), for which emissions at finer, city scales are more uncertain than those aggregated at the global scale. Carbon-observing satellites may provide independent *top-down* emission evaluations and compensate for the sparseness of surface CO_2 observing networks, especially in urban areas. Although some previous studies have attempted to derive urban CO_2 signals from satellite column-averaged CO_2 data (XCO_2) using simple statistical measures, less work has been carried out to link upwind emission sources to downwind atmospheric columns using atmospheric models. In addition to Eulerian atmospheric models that have been customized for emission estimates over specific cities, the Lagrangian modeling approach—in particular, the Lagrangian Particle Dispersion Model (LPDM) approach—has the potential to efficiently determine the sensitivity of downwind concentration changes to upwind sources. However, when applying LPDMs to interpret satellite XCO_2, several issues have yet to be addressed including quantifying uncertainties in urban XCO_2 signals due to receptor configurations and errors in atmospheric transport and background XCO_2.

In this study, we present a modified version of the Stochastic Time-Inverted Lagrangian Transport (STILT) model, "X-STILT", for extracting urban XCO_2 signals from NASA's Orbiting Carbon Observatory 2 (OCO-2) XCO_2 data. X-STILT incorporates satellite profiles and provides comprehensive uncertainty estimates towards urban XCO_2 enhancements on a per sounding basis. Several methods to initialize receptors/particle setups and determine background XCO_2 are presented and discussed via sensitivity analyses and comparisons. To illustrate X-STILT's utilities and applications, we examined five OCO-2 overpasses over Riyadh, Saudi Arabia, during a two-year time period and performed a simple scaling factor-based inverse analysis. As a result, the model is able to reproduce most observed XCO_2 enhancements. Error estimates show that the 68 % confidence limit of XCO_2 uncertainties due to transport (horizontal wind plus vertical mixing) and emission uncertainties contribute to ~33 % and ~20 % of the mean latitudinally-integrated urban signals, respectively, over the five overpasses, using meteorological fields from the Global Data Assimilation System (GDAS). In addition, a sizeable mean difference of -0.55 ppm in background derived from a previous study employing simple statistics (regional daily median) leads to a higher mean observed urban signal by ~39 % and a larger posterior scaling factor. Based on our signal estimates and associated error impacts, we foresee X-STILT serving as a tool for interpreting column measurements, estimating urban enhancement signals, and carrying out inverse modeling to improve quantification of urban emissions.

1 Introduction

Carbon dioxide (CO_2) is a major greenhouse gas in the atmosphere in terms of radiative forcing, with its concentration increasing significantly over the past century (Dlugokencky and Tans, 2015). The largest contemporary net source of CO_2 to the atmosphere over the decadal time scales is anthropogenic emissions, namely from fossil fuel burning and net land-use change (Ciais et al., 2013). Urban areas play significant roles in the global carbon cycle and are responsible for over 70 % of the global energy-related CO_2 emissions (Rosenzweig et al., 2010). Global fossil fuel CO_2 ($FFCO_2$) emission uncertainty (8.4%, 2σ, Andres et al., 2014) may be smaller than other less-constrained emissions such as of wildfire (Brasseur and Jacob, 2017). Still, uncertainties associated with national $FFCO_2$ emissions derived from *bottom-up* inventories typically range from 5–20 % per year (Andres et al., 2014). These estimated emission uncertainties result primarily from differences in emission inventories, such as the emission factors and energy consumptions data used. Moreover, heightened interests in regional- and urban-scale emissions require modelers to investigate $FFCO_2$ emissions at finer spatiotemporal resolutions (Lauvaux et al., 2016; Mitchell et al., 2018) as well as uncertainties in gridded emissions (Andres et al., 2016; Gately and Hutyra, 2017; Hogue et al., 2016; Oda et al., 2018). Dramatic increases in emission uncertainties are associated with finer scales, with these uncertainties being mostly biases due to different methods disaggregating national-level emissions (Marland, 2008; Oda and Maksyutov, 2011). For instance, emission uncertainties of 20 % at regional scales increased to 50–250 % at city scales even for the northeastern United States (Gately and Hutyra, 2017), an area that is considered relatively "data-rich".

Given the large differences/discrepancies in emission inventories at urban scales, the use of atmospheric *top-down* constraints could be helpful for quantifying urban emissions and possibly providing a monitoring support (Pacala et al., 2010). Observed concentrations used in the top-down approach can often be obtained from ground-based instruments (Kim et al., 2013; Mallia et al., 2015; Wunch et al., 2011) and aircraft observations (Gerbig et al., 2003; Lin et al., 2006). Each type of measurement offers valuable information and has both advantages and disadvantages. Most ground-based measurements provide reliable, continuous CO_2 concentrations from fixed locations/heights. Unfortunately, current ground-based observing sites are too sparse to constrain urban emissions around the globe. Most National Oceanic and Atmospheric Administration (NOAA) sites are designed to measure background concentrations and few others aim at measuring concentration changes from few vertical levels within the planetary boundary layer (PBL). Other than a few notable examples (Feng et al., 2016; Lauvaux et al., 2016; Mitchell et al., 2018; Verhulst et al., 2017; Wong et al., 2015; Wunch et al., 2009), near-surface CO_2 measurements may not be available over many other cities around the world. Alternatively, airborne measurements from field campaigns provide better vertical and regional coverages (Cambaliza et al., 2014). Yet, continuous airborne operations over months to years are often impractical due to limited resources, which limits researchers' capability to track temporal variability of anthropogenic carbon emissions (Sweeney et al., 2015).

The carbon cycle community has entered a new era with advanced carbon-observing satellites—i.e., Greenhouse gases Observing SATellite (GOSAT; Yokota et al., 2009), TanSat (Liu et al., 2013) and Orbiting Carbon Observatory (OCO-2) satellite (Crisp et al., 2012)—routinely in orbit to measure variations of atmospheric column-averaged CO_2 mole fraction (XCO_2). Although most carbon-observing satellites have revisit times of multiple days (e.g., 3 days for GOSAT and 16 days for OCO-2), their global coverage, large number of retrievals and multi-year observations may further complement the current surface observing networks. Space-borne CO_2 measurements, in combination with surface CO_2 networks, may help reduce emission uncertainties and benefit urban emissions analysis, especially over regions with no surface observations (Duren and Miller, 2012; Houweling et al., 2004; Rayner and O'Brien, 2001).

Previous studies have demonstrated the potential for detecting and deriving urban CO_2 emission signals from satellite CO_2 observations, in the form of XCO_2 enhancements above the background, without making use of much atmospheric transport

information (Hakkarainen et al., 2016; Kort et al., 2012; Schneising et al., 2013; Silva and Arellano, 2017; Silva et al., 2013). However, the linkage between their derived urban CO_2 emission signals and upstream sources is tenuous, as downwind XCO_2 can be enhanced by not only near-field upwind urban activities (e.g., traffic, houses, and power plants/industries), but regional-scale advection of upwind sources/sinks as well. Simulations using transport models are able to isolate the portion of satellite observations influenced by urban regions from the portion affected by natural fluxes or long-range transport (e.g., Ye et al., 2017). Therefore, accurate knowledge of atmospheric transport is essential in top-down assessment. As importantly, transport modeling is a necessary step within inverse modeling, which can help improve fossil fuel emission estimates and shed light on CO_2 emission monitoring networks (Kort et al., 2013; Lauvaux et al., 2009). Uncertainties in transport modeling have been identified as a significant error source that affects inferred surface fluxes (Peylin et al., 2011; Stephens et al., 2007; Ye et al., 2017). Yet, by analyzing an increased number of satellite overpasses, uncertainties from atmospheric inversions due to non-systematic transport errors in emission estimates can be reduced (Ye et al., 2017).

Two main approaches can be considered for atmospheric transport modeling. Eulerian models, in which fixed grid cells are adopted and CO_2 concentrations within the grid cells are calculated by forward numerical integrations, have been widely utilized and customized to understand urban emissions and quantify model uncertainties over specific metropolitan regions worldwide (Deng et al., 2017; Lauvaux et al., 2013; Palmer, 2008; Ye et al., 2017). The Lagrangian approach, especially the time-reversed approach in which atmospheric transport is represented by air parcels moving backward in time from the measurement location ("receptor"), is efficient in locating upwind sources and facilitating the construction and calculation of the "footprint" (e.g., Lin et al., 2003) or "source-receptor matrix" (Seibert and Frank, 2004)—i.e., the sensitivity of downwind CO_2 variations to upwind fluxes.

In particular, the receptor-oriented Stochastic Time-Inverted Lagrangian Transport (STILT) model, one of the Lagrangian Particle Dispersion Models (LPDM), has the ability to more realistically resolve the sub-grid scale transport and near-field influences (Lin et al., 2003). STILT has been used to interpret CO_2 observations within the PBL (Gerbig et al., 2006; Kim et al., 2013; Lin et al., 2017) and, in recent years, to analyze column observations, i.e., XCO_2 (Fischer et al., 2017; Heymann et al., 2017; Macatangay et al., 2008; Reuter et al., 2014). Among STILT-based column studies, most aim at either natural CO_2 sources and sinks like wildfire emissions and biospheric fluxes, or anthropogenic emissions at regional or state scales. Very few studies focus on city-scale $FFCO_2$ using column data and LPDMs. Moreover, when applying LPDMs to interpret column CO_2 data, three key issues have yet to be carefully examined and will be addressed in this paper:

I. Uncertainty of modeled XCO_2 enhancements due to model configurations. Very few studies have examined model uncertainties resulting from model configurations—i.e., receptors and particles in LPDMs. Previous studies reported negligible to ~20 % of the modeled enhancements are reported as the error impact due to STILT particle number (released from a fixed level), depending on adopted particle numbers, examined species and their components/sources (Zhao et al., 2009; Gerbig et al., 2003; Mallia et al., 2015). When it comes to representing an atmospheric column using particle ensembles, many studies depicted their setups for receptors/particles without further explaining why they chose those setups or the error impact (due to model configurations) on modeling XCO_2. Although this error impact may be small, we still perform a set of sensitivity tests to provide more guidance on placing column receptors.

II. Horizontal and vertical transport error impact on XCO_2 simulations. Flux inversions, e.g., Bayesian Inversion (Rodgers, 2000) involving LPDMs have been widely adopted to constrain emissions. Approaches to quantify errors in horizontal wind fields and vertical mixing have been proposed followed by comprehensive error characterizations on atmospheric simulations (Gerbig et al., 2008; Jeong et al., 2013; Lauvaux et al., 2016; Lin and Gerbig, 2005; Zhao et al., 2009). Recent efforts (e.g., Lauvaux and Davis, 2014; Ye et al., 2017) have been made to rigorously examine the column transport

errors. The uncertainties in horizontal wind fields and vertical mixing within X-STILT will be propagated into column CO_2 space in this study.

III. Determining background XCO_2 and characterizing its uncertainties. Here we define background value as the CO_2 "uncontaminated" by fossil fuel emissions from the city of interest. As urban emission signals are defined as the enhancements of XCO_2 over the background, errors in the background value introduce first-order errors into the derived urban XCO_2 signal from total XCO_2, with such errors propagating directly into fluxes calculated from atmospheric inversions (e.g., Göckede et al., 2010). Consequently, background determination is another critical task.

One commonly used method in determining model boundary conditions of various species in LPDMs is the "trajectory-endpoint" method that establishes the background based on CO_2 extracted at endpoints of back trajectories from modeled regional/global concentration fields (Lin et al., 2017; Macatangay et al., 2008; Mallia et al., 2015). The aforementioned studies (adopting the trajectory-endpoint method) aim at extracting relatively large CO_2 anomalies (e.g., at a fixed level within the PBL or due to large emissions such as of wildfire) out of the total measured CO_2. However, for studying XCO_2 that is less variable than near-surface CO_2 (Olsen and Randerson, 2004), potential errors in modeled concentration fields and atmospheric transport may pose more significant adverse impact on derived urban signals. Other ways of defining background include geographic definitions (Kort et al., 2012; Schneising et al., 2013) and simple statistical estimates (Hakkarainen et al., 2016; Silva and Arellano, 2017). These simple statistical methods often neglect the atmospheric transport and may use a less accurate upwind region to select measurements for deriving background values. Lastly but more importantly, recent column studies (Nassar et al., 2017; Fischer et al., 2017) studied the impact of potential errors/biases in background values on their emission or fluxes estimates. In this work, we introduce a new background determination that combines OCO-2 observations and the STILT-based atmospheric transport and account for errors in our background estimates.

In general, we attempt to address the aforementioned issues by extending STILT with column features and comprehensive error analyses, referred to as the column-STILT, "X-STILT". We illustrate the model's applications in extracting urban XCO_2 signals from OCO-2 retrievals (Fig. 1) and evaluate model performances via a case study focusing on Riyadh, Saudi Arabia. Riyadh, with a population of over 6 million by 2014 (WUP 2014), is chosen as the city of interest because of its low cloud interference, limited vegetation coverage, and isolated location in a barren area, which leads to higher data recovery rates and facilitates the background determination. Saudi Arabia has the largest CO_2 emissions among Middle Eastern countries and ranks eighth globally in 2016 (Boden et al., 2017; BP, 2017; UNFCCC, 2017). We examine several satellite overpasses and focus on a small spatial domain adjacent to Riyadh for each overpass.

2 Data and methodology

Before demonstrating model details, Fig. 1 highlights several X-STILT characteristics, e.g., column transport errors quantifications, background XCO_2 approximations, and the identification of upwind emitters using backward-time runs from column-receptors. Our goal is to evaluate the model by comparing both the latitude-dependent model-data XCO_2 urban enhancements (Sect. 3.4) and the overall latitude-integrated urban signals within a small latitudinal range (Sect. 3.5). We selected and examined five OCO-2 overpasses during the time period of Sept 2014–Dec 2016, based on four stringent criteria (Appendix A).

2.1 STILT-based approach for XCO_2 simulation ("X-STILT")

The OCO-2's column averaging kernel is the product of normalized averaging kernel (AK_{norm}) and pressure weighting (PW) function and represents the sensitivity of the change in retrieved XCO_2 due to CO_2 anomaly at each retrieved grid. Column AK_{norm} peaks near the surface and exhibits values near unity throughout most of the troposphere (Boesch et al., 2011). Lower AK_{norm} values are mainly found aloft, which requires more information in the a priori CO_2 profiles ($CO_{2,prior}$; Fig. 2a). For direct comparisons against OCO-2 retrieved XCO_2, CO_2 anomalies at model grids should be properly weighted using the satellite's column averaging kernels (Basu et al., 2013; Lin et al., 2004). Thus, the final AK-weighted simulated XCO_2 ($XCO_{2.sim.ak}$) are weighted between model-derived CO_2 profiles and OCO-2 a priori profiles (O'Dell et al., 2012):

$$XCO_{2.sim.ak} = \sum_{n=1}^{nlevel}(AK_{norm,n}PW_nCO_{2.sim.n} + (I - AK_{norm,n})PW_nCO_{2.prior,n}) \quad (1)$$

I is the identity vector and n stands for the combined vertical levels of STILT plus OCO-2. Specifically, we replaced OCO-2 levels with denser model release levels for the lower part of the troposphere (red circles from the surface in Fig. 2), while kept OCO-2 levels for upper part (blue circles in Fig. 2). To reduce computational cost, the air column is only simulated up to the maximum release height (MAXAGL in meters above ground level, mAGL; Fig. 2).

Interpolations are further needed to resolve the mismatch between prescribed OCO-2 retrieval grids and model levels for the lower part of the troposphere. Our intention is to preserve the finer modeled CO_2 variations by performing interpolations of satellite profiles from retrieval grids to model levels. Vertical profiles of AK_{norm}, PW and $CO_{2,prior}$ are treated as continuous functions and interpolated linearly to model grids (red circles in Fig. 2). Note that the initial OCO-2 PW functions have steady value of ~0.052 (except for the very bottom and top levels; black dots in Fig. 2b), which results from constant pressure spacings (*dp_oco2*) between two adjacent OCO-2 levels. However, X-STILT levels are much denser with smaller pressure spacings (*dp_stilt*) or less airmass between their two adjacent levels. Therefore, the linearly interpolated PW (red circles in Fig. 2b) needs an additional scaling via a set of "scaling factors" representing the ratios of pressure spacings in STILT versus OCO-2 retrieval (*dp_stilt/dp_oco2)*, to arrive at the correct PW for each finer model grid (orange circles in Fig. 2b).

Eq. (1) can further be rewritten as Eq. (2), since the simulated CO_2 profiles in Eq. (1) is comprised of CO_2 boundary condition plus CO_2 anomalies due to sources/sinks ($FFCO_2$, biospheric and oceanic fluxes):

$$XCO_{2.sim.ak} = XCO_{2.sim.ff} + XCO_{2.sim.bio} + XCO_{2.sim.ocean} + XCO_{2.sim.bound} + XCO_{2.prior} = XCO_{2.sim.ff} + XCO_{2.bg}. \quad (2)$$

Given our focus, we defined background value as the XCO_2 portion not "contaminated" by urban emissions. Thus, $XCO_{2.sim.ak}$ is the sum of the XCO_2 enhancement due to $FFCO_2$ ($XCO_{2.sim.ff}$) and estimated background value ($XCO_{2.bg}$). Estimates of XCO_2 anomalies are further explained in Sect. 2.2.2 and four ways to estimate background values ($XCO_{2.bg}$) are proposed in Sect. 2.3.

2.1.1 X-STILT setup ("column receptors")

The linkage between the observed XCO_2 concentration by a given OCO-2 sounding and upwind carbon sources and sinks is determined by atmospheric transport. We adopt the STILT model to describe this connection. Fictitious particles, representing air parcels, are released from a "receptor" (location of interest) and are dispersed backward in time. The Lagrangian air parcels within STILT are transported along with the mean wind ($\bar{\boldsymbol{u}}$), turbulent wind component ($\boldsymbol{u}'$), and other meteorological variables, which are derived from Eulerian meteorological fields. In this study, we used meteorological fields simulated by the Weather Research and Forecasting (WRF; Skamarock and Klemp, 2008) and the 0.5° × 0.5° Global Data Assimilation System (GDAS; Rolph et al., 2017; Stein et al., 2015). Hourly WRF fields contain 51 vertical levels with boundary conditions from 6-hourly 0.5°×0.5° NCEP

FNL (Final) Operational Global Analysis data (Ye et al., 2017) are customized and utilized for the first 2 of the total 5 overpasses over Riyadh. We note that the primary focus is to assess the resulting errors given the choice of a particular wind field (i.e., GDAS 0.5°), rather than to carry out analyses of differences between WRF and GDAS.

To represent the air arriving at the atmospheric column of each OCO-2 sounding, we release air parcels from multiple vertical levels, "column receptors" (Fig. 3e), using the same lat/lon as the satellite sounding at the same time and allow those parcels to disperse backward for 72 hours (see Appendix D2 for model impact from backward durations). About 10–20 satellite soundings are selected for simulations over every 0.5° latitude with data filtering using criteria explained in Sect. 2.2. Sensitivity tests are conducted regarding different configurations—the maximum release level (MAXAGL), the vertical spacing of release levels (*dh*), and the particle number per level (*dpar*), when placing column receptors (Sect. 2.5).

2.1.2 Modeling XCO_2 anomalies

Air parcels traveling back in time provide valuable information about how upwind sources and sinks impact the air arriving at a receptor. However, since particles within the ensemble are subject to stochastic motion, the surface fluxes observed by any single particle caries limited information. The influence of upstream surface fluxes on a receptor is given by summing the sensitivities of all particles in the ensemble over a surface grid $f(x, y, t)$, which is referred to as the "footprint" (Lin et al., 2003; Fasoli et al., 2018) or the "source-receptor matrix" (Seibert and Frank, 2004). Formally, the sensitivity of the receptor located at $\boldsymbol{x}_r$ at time t_r to surface fluxes originating from x_i, y_j is given by summing $\Delta t_{p,i,j,z\leq h}$, the time spent by particle p over grid position i, j within the surface layer of height h for each discrete time step m:

$$f(\boldsymbol{x}_r, t_r | x_i, y_j, t_m) = \frac{m_{air}}{h\bar{\rho}(x_i,y_j,t_m)} \frac{1}{N_{tot}} \sum_{p=1}^{N_{tot}} \Delta t_{p,i,j,z\leq h} \quad (3)$$

where N_{tot} is the total number of particles in the ensemble, m_{air} is the molar mass of dry air, $\bar{\rho}$ is the average air density below h. The dilution of surface fluxes to half of the PBL height $h = 0.5 z_{pbl}$ is often used. In general, f increases if particles travel at heights $z \leq h$ and if h is low, concentrating surface fluxes within a shallower atmospheric column.

To reduce grid noise caused by aggregation of a finite number of dispersed particles, a kernel density estimator is used to variably smooth f as a function of elapsed time and particle location uncertainty. Refer to Fasoli et al. (2018) for additional details pertaining to the formulation of f.

We introduce the weighted column footprint f_w that describes the sensitivity of changes in column concentration due to potential upstream sources/sinks and incorporates satellite profiles. The formulation of f_w is similar to (3) but scales the sensitivity with $AK_{norm}(n, r)$ and $PW(n, r)$:

$$f_w(\boldsymbol{x}_{n,r}, t_{n,r} | x_i, y_j, t_m) = \frac{m_{air}}{h\bar{\rho}(x_i,y_j,t_m)} \frac{1}{N_{tot}} \sum_{p=1}^{N_{tot}} \Delta t_{p,i,j,z\leq h}\, AK_{norm}(n, r)\, PW(n, r) \quad (4)$$

where $\boldsymbol{x}_{n,r}, t_{n,r}$ denotes a column receptor. Multiplying f_w by gridded flux estimates yields a change in CO_2 at the downwind column receptor. Thus, surface fluxes $F(x_i, y_j, t_m)$ cause a change in column integrated mole fraction ΔXCO_2 as

$$\Delta XCO_2(\boldsymbol{x}_{n,r}, t_{n,r} | x_i, y_j, t_m) = F(x_i, y_j, t_m)\, f_w(\boldsymbol{x}_r, t_r | x_i, y_j, t_m). \quad (5)$$

For our OCO-2 case study, modeled XCO_2 enhancements due to $FFCO_2$ emissions are derived from the convolution of spatially-varying f_w and ODIAC emissions (Sect. 2.4.1). Also, we account for modeled uncertainties that includes errors in prior $FFCO_2$ emissions (Sect. 2.4.1), receptor configurations (Sect. 2.5), and atmospheric transport (Sect. 2.6).

2.2 OCO-2 retrieved XCO_2 and data pre-processing

The OCO-2 algorithm of retrieving XCO_2 from radiances employs an optimal estimation approach (Rodgers, 2000) involving a forward model, an inverse model, and prior information regarding the vertical CO_2 profiles (O'Dell et al., 2012). We used the bias-corrected XCO_2 values from OCO-2 Lite files (version 7R; OCO-2 Science Team/Michael Gunson, Annmarie Eldering, 2015). The impacts of different versions of the OCO-2 datasets on our results are briefly discussed in Sect. 4.4. Satellite measurements over Riyadh were carried out in Land Nadir and Glint modes. Soundings with quality flags equal zero (QF = 0) are selected, which implies selected observations have passed the cloud and aerosol screening (with removal of albedo > 0.4) and their retrievals have converged (Mandrake et al., 2013; Patra et al., 2017). For smoothing noisy observations, we binned the screened XCO_2 data according to the lat/lon of model receptors (served as the midpoints of each bin) and calculated the mean and standard deviation of screened measurements within each bin. Next, background values were defined (Sect. 2.4) and subtracted from the bin-averaged observed XCO_2 to estimate the increase in observed XCO_2 (step 3 in Fig. 1). The impacts of different bin-widths on bin-averaged observed signals are shown in Appendix E1. Total observed errors contain the spatial and natural variation of observed XCO_2 in each bin, background uncertainties (Sect. 2.3.3), and retrieval errors provided by Lite files. Retrieval error variances per sounding are then averaged within each observed bin to obtain bin-averaged retrieval error variances.

2.3 Estimates of background XCO_2

Definitions of "background" vary among studies with different applications. Here, we define background values as atmospheric XCO_2 that is not "contaminated" by the urban emissions around our study site. Determination of background XCO_2 is crucial, as it can significantly affect the magnitude of inferred observed anthropogenic signals. If the background is underestimated, then the detected signal may be overestimated, and vice versa. In this study, we seek to develop best-estimated background values given five tracks, where 3 methods are proposed and investigated as follows,

M1. A "trajectory-endpoint" method by assigning CO_2 values extracted from global models (i.e., CT-NRT) to trajectory endpoints plus simulated biospheric, oceanic, prior components (Sect. 2.3.1);

M2. Statistical methods estimated solely from XCO_2 observations based on two previous studies (Sect. 2.3.2);

M3. An "overpass-specific" background requires model-defined urban plume and measurements outside the plume (Sect. 2.3.3).

We devote considerable efforts to compare the aforementioned three ways (Sect. 3.3) and investigate the background impact on model-data comparisons and emission estimates (Sect. 4.2). We choose M3-based background for the following analysis as it is designed specifically for examining a particular city and specific overpasses downwind of the city.

2.3.1 Trajectory-endpoint method (M1)

Modeled background XCO_2 comprises modeled boundary condition confined by four-dimensional CO_2 fields from CT-NRT and contributions from biospheric fluxes, oceanic fluxes, and OCO-2 prior profiles (M1 in Fig. 1 and Eq. (2)). Specific for modeling CO_2 boundary condition, CO_2 values for upper levels above MAXAGL are estimated based on CT CO_2 at those OCO-2 pressure levels (purple circles in Fig. 2c). And, averaged CT CO_2 values at trajectory endpoints are used for boundary conditions at model release levels (orange circles in Fig. 2c). Then, modeled boundary conditions at vertical levels are weighted accordingly via OCO-2's column averaging kernel (red and blue circles in Fig. 2c). Model trajectories are properly subsetted according to the boundary of the footprint domain (i.e., 20° × 20°) used for simulating XCO_2 anomalies.

However, potential uncertainties in transport may strongly influence the distribution of Lagrangian parcels as backward duration time increases and may lead to potential spatial mismatch of the background region. Furthermore, potential biases and relatively coarse resolution of 2° × 3° of the global CarbonTracker may add inaccuracies to CO_2 values at trajectory-endpoints.

2.3.2 Statistic method (M2)

Hakkarainen et al. (2016) (referred to as M2H) extracted local XCO_2 anomalies from the daily median of screened measured XCO_2 within a relatively broad region (0° N–60° N, 15° W–60° E over the Middle East; Fig. S7). Their detected anomalies vary from 1–2 ppm over 0.5° × 0.5° gridcells near Riyadh. Silva and Arellano (2017) (referred to as M2S) used measurements within a 4° × 4° combustion region centered around the "urban and dense settlements" inferred from the anthropogenic biomes dataset ("anthromes", Ellis and Ramankutty, 2008). Then, they derived background as the mean minus one standard deviation of available observations within their studied urban extents.

Both statistical methods are highly efficient in estimating background values but can be limited to certain applications. For instance, M2H may be not suitable for determining background values when zooming into specific cities. Measurements within their broad spatial domain are lumped together, regardless of their locations (whether over rural or urban areas) and atmospheric transport. Silva and Arellano (2017) have pointed out that their defined 4° × 4° combustion region is suitable for studying the "bulk" characteristics and may be too coarse for studying urban emissions. Also, the Gaussian statistics assumed in M2S may be less applicable when multiple observed peaks are entangled together caused by a cluster of cities. Therefore, without incorporating much atmospheric transport information, accuracies in the transport from an urban center to the downwind satellite overpass cannot be guaranteed. It may be difficult for either statistical method to locate the exact XCO_2 peak elevated by target city or background region. These difficulties motivate us to introduce a new approach in the next subsection.

2.3.3 Overpass-specific background (M3)

A few space-based studies defined the background values as the averaged observed XCO_2 values over a "clean" upwind region. For instance, Kort et al. (2012) and Schneising et al. (2013) defined the "clean" region based on geographic information (e.g., rural area to the north of LA basin). Although OCO-2 has relatively narrow swaths, transport models can be used to differentiate the enhanced versus background portions along an overpass. For example, Janardanan et al. (2016) calculated background XCO_2 as the averaged GOSAT observations among gridcells with modeled anthropogenic signals < 0.1 ppm. This 0.1 ppm threshold is determined from the average simulated fossil fuel abundance over desert areas worldwide using the FLEXible PARTicle dispersion model (FLEXPART; Stohl et al., 2005), a model similar to STILT in that both are time-reversed LPDMs. Nassar et al. (2017) derived overpass-dependent background and its uncertainty based on the averaged OCO-2 observations within four different tested background latitudinal ranges.

We present an alternative method using a forward-time run from an urban box to reveal the urban influence on satellite soundings, which are more straightforward and efficient than solely relying on backward-time runs. Fictitious particles are released from a box around the city center (pink dots in Fig. 1) as a feature implemented with STILT (T. Nehrkorn, personal communication) to track air parcels over a city and the transport of the urban plume. Specifically, the model continuously releases air parcels over a 30-minute window from a 0.4° × 0.4° box around the city center, with multiple 30-minute releases of 1000-particle ensembles over the 10-hours ahead of the satellite overpass hour (00–10UTC). Then, an urban plume can be derived from the parcels' distribution during the ~3 minutes OCO-2 passing window (purple dots in Fig. 5a). Note that air parcels are tracked forward in time for 12 hours, allowing for equal contributions from parcels released initially from different

time intervals (every 30 mins) onto defined urban plume. We are aware of potential model errors and their adverse impacts on defined urban plume. Therefore, a wind error component (details in Sect. 2.6) is further added in the forward run to broaden the polluted range (solid black line in Fig. 5a). Next, two-dimensional kernel density estimation (Venables and Ripley, 2002) is applied to determine the boundary of the city plume based on the air parcels' distributions. We normalized the two-dimensional kernel density by its maximum value and "sketched" the boundary of the city plume based on a threshold of 0.05, which is sufficient to include most air parcels. No dramatic change in the shape/size of resultant urban plume was found as testing other thresholds < 0.05. The urban-influenced latitude range is defined as the intersection of the urban plume and OCO-2 overpass (Fig. 5a). Overall, the urban-influenced latitudinal band represented by 5 % of the maximum kernel density covers from 23.5° N–26° N, given multiple overpasses for Riyadh (Fig. S2). The background latitudinal range unaffected by Riyadh's urban plume for estimating background then extend ~100 km from the north-most and/or south-most of derived urban plume (Fig. 5b). We abandon observations with latitudes > 26° N and < 23° N, because those retrievals are too scattered (black triangles in Fig. 5b) and indicate a second peak during few other overpasses. If the near-field wind vectors point more towards the north, screened measurements over the southern background latitudinal range is utilized, vice versa. Eventually, the background value is calculated as the mean value of the screened observations over background region (dashed green line in Fig. 5b). Two error sources are incorporated into the background error— i.e., the measured (standard deviation, SD) and the retrieval errors of background observations.

In addition to random errors (that are resolved by the inclusion of the aforementioned wind error component and broadening of the city plume), potential large bias in near-field wind direction may lead to mismatch in modeled and observed background regions and may bring relatively higher XCO_2 values into background XCO_2. However, we do not explicitly account for the potential near-field wind bias's impact on forward-trajectories defined urban plume with following considerations. Firstly, we attempted to propagate a near-field wind bias into the modeled plume by rotating forward trajectories, whereas the robustness of this near-field bias can be affected by the very few wind measurements near Riyadh (further explained in Sect. 2.6.1). Secondly, the background latitude range defined by M3 with the broadening effect (blue lines in Fig. 5b) in general matches well with that observed from OCO-2 for most overpasses, which implies that the overall wind bias around our study site is not significant. Lastly, even if potential wind bias may result in less accurate background range and bring elevated XCO_2 into the background, the background uncertainty implicitly contains information about the spatial variation in background measurements (green ribbon in Fig. 5b). In addition, the M3-derived background is the mean value of mostly hundreds of background observations (numbers in Fig. 6e), which may not be greatly affected by a few potential urban-enhanced measurements.

2.4 Sources of information for CO_2 fluxes

2.4.1 Fossil fuel emission (ODIAC) and prior emission uncertainties

To calculate modeled XCO_2 enhancements, we used the latest (year 2017) version of the Open-Data Inventory for Anthropogenic Carbon dioxide (ODIAC2017 dataset, Oda et al., 2018; Oda and Maksyutov, 2011, 2015) with monthly fossil fuel CO_2 emissions at 1×1 km resolution (Fig. 4). ODIAC starts with annual national emission estimates, separated by fuel type, from the Carbon Dioxide Information Analysis Center (CDIAC, Andres et al., 2011), which are then re-categorized into specific ODIAC emission categories on a monthly basis, i.e., point source, non-point source, cement production, international aviation and marine bunker (Oda et al., 2018). Because CDIAC only covers years up to 2013, ODIAC extrapolates emissions in 2013 for emissions in 2014 and 2015 based on BP (i.e., the British Petroleum Company) global fuel statistical data (BP, 2017). Also, ODIAC estimates point sources emissions according to a global power plants database—the Carbon Monitoring and Action (CARMA) Database (Wheeler and Ummel, 2008), and collects and distributes non-point sources using an advanced nighttime lights dataset from the Defense

Meteorological Satellite Program Operational Line Scanner (DMSP/OLS). The use of the nightlight dataset allows ODIAC to characterize the spatial patterns of the anthropogenic sources such as point sources, line sources, and diffuse sources.

To estimate emission uncertainties, we followed a method similar to those reported in Oda et al. (2015) and Fischer et al. (2017). Three emission inventories derived from different methods are inter-compared: ODIAC, the Fossil Fuel Data Assimilation System (FFDASv2; Asefi-Najafabad et al., 2014; Rayner et al., 2010) and the Emission Database for Global Atmospheric Research (EDGARv4.2; http://edgar.jrc.ec.europa.eu; Janssens-Maenhout et al., 2017). To resolve different spatial grid spacing among three inventories, we aggregated ODIAC emissions from 1 km to 0.1° gridcells. The fractional uncertainty for gridded emissions is characterized by the emission spread (1-σ, among three inventories) and mean values (μ) of estimated emissions for each gridcell within a given region (10° N–40° N, 25° E–60° E; Fig. S3). In general, fractional uncertainties in gridded emissions mostly range from 60–130 % (Fig. S3) around Riyadh. Ultimately, these fractional emission uncertainties and ODIAC emissions are convolved with X-STILT's weighted column footprints to provide the XCO_2 uncertainties due to prior emission uncertainties.

2.4.2 Natural fluxes (CarbonTracker)

The trajectory-endpoint method (M1 in Sect. 2.3.1) requires the oceanic and terrestrial biospheric fluxes that come from the 3-hourly 1° × 1° product—CarbonTracker-NearRealTime (CT-NRTv2016 and v2017, http://carbontracker.noaa.gov). CT-NRT, an extension of the standard CarbonTracker (Peters et al., 2007), is designed for the OCO-2 program and uses different prior flux models and "real-time" ERA-Interim reanalysis in its transport model than regular CT, which allows for more timely model results. To calculate oceanic and biospheric XCO_2 changes, we multiplied these natural fluxes with column weighted footprint according to Eq. (5). Although wildfire emissions can greatly modify atmospheric XCO_2 (e.g., Heymann et al., 2017), we expected relatively small XCO_2 contributions from wildfire and hence excluded wildfire-elevated XCO_2 estimations, considering the studied times (wintertime overpasses) and study region (the Middle East).

2.5 Sensitivity analyses for X-STILT column receptors

The goal of carrying out sensitivity tests is to understand any systematic/random errors towards STILT simulations brought by receptor configurations. Under the premise of limited computational resources, proper column receptors are set up with allowable random errors. The total number of particles (NUMPAR) released from column receptors are decomposed into three parameters, i.e., the maximum release level (MAXAGL), the vertical spacing of release levels (*dh*), and the particle number per level (*dpar*).

Instead of regenerating model trajectories (Jeong et al., 2013; Mallia et al., 2015), we adopted the bootstrap method to resample model ensembles. The bootstrap approach helps construct hypothesis tests and infer error statistics (Efron and Tibshirani, 1986). The initial sample size before the bootstrap should be sufficiently large to ensure the performance of the bootstrap method and its related statistics. Thus, a "base run" of trajectories starting from the surface to 10 km with a vertical spacing of 25 m and 200 particles per level are generated and stored. For testing each parameter (MAXAGL, *dh*, or *dpar*), we fixed the other two parameters and randomly selected/resampled model particles from the base run for 100 times (allowing for repetitions). 100 urban enhancements are calculated from 100 new sets of trajectories for each test. Basic statistics—i.e., mean values and standard deviations (or fractional uncertainty, i.e., SD/mean) among these 100 enhancements—are used to infer systematic and random uncertainties in each test, respectively (with results showed in Sect. 3.1).

2.6 X-STILT column transport errors

Uncertainty in atmospheric transport modeling has been identified to significantly affect emission constraints (Cui et al., 2017; Lauvaux et al., 2016; Stephens et al., 2007). Here we quantify uncertainties in modeled XCO_2 due to transport errors caused by uncertainties in both horizontal wind fields (Sect. 2.6.1) and vertical mixing (Sect. 2.6.2).

2.6.1 Horizontal transport errors

Previous studies (Lin and Gerbig, 2005; Mallia et al., 2017) aimed at estimating transport error at one particular level, whereas for XCO_2 we account for transport error in a column sense (i.e., column transport error). Macatangay et al. (2008) briefly explained the column transport error as the weighting of transport error variances with respect to pressures. Similarly, we treated each model level separately and calculated one CO_2 transport error per level, denoted as σ_ε^2 $(CO_{2.sim.ak.n})$, following Lin and Gerbig (2005). In short, an additional wind error component ($\boldsymbol{u_\varepsilon}$) is added to the mean wind ($\bar{\boldsymbol{u}}$) and turbulent wind component ($\boldsymbol{u'}$) that are embedded in normal STILT runs (Lin et al., 2003), to randomly perturb air parcels for each level. RMS errors of u- and v-component modeled wind, error correlation timescales and length scales describe the $\boldsymbol{u_\varepsilon}$ in space and time. Details about the wind error calculations are explained in Appendix B.

For each model level (n), we obtained two sets of parcel distributions—i.e., one without and one with the wind error component ($\boldsymbol{u_\varepsilon}$). Then, the difference in the spread of these two distributions, or mathematically the difference in the variances of derived CO_2 distributions among air parcels (Lin and Gerbig, 2005), serve as the XCO_2 uncertainty (in ppm) due to transport error. We tested both the normal or log-normal statistics for describing this XCO_2 transport uncertainty. Since transport error using log-normal statistics did not show very distinct improvement from that using normal statistics, we ended up adopting normal statistics for the consideration of benefiting inverse modeling. Because the parcel distribution with $\boldsymbol{u_\varepsilon}$ (orange dots in Fig. S4) is more dispersed than the parcel distribution without $\boldsymbol{u_\varepsilon}$ (blue dots in Fig. S4), the increase in CO_2 variance with $\boldsymbol{u_\varepsilon}$ from that without $\boldsymbol{u_\varepsilon}$ describes the transport error for each level. However, negative values of transport error can occasionally occur, due to statistical sampling from insufficient model parcel trajectories. To resolve this technical issue, we modified Lin and Gerbig (2005) by using a regression-based approach. A weighted linear regression slope is used to describe the increase in CO_2 variances and then estimate transport error. More descriptions about this regression-based method are demonstrated in Appendix B. Overall, transport errors at levels within the PBL are expected to be larger than those at higher levels that approach zero.

Lastly, vertical profiles of transport errors are weighted against OCO-2's weighting functions. Following the definition of modeled AK-weighted XCO_2 in Eq. (1), the weighted column transport error follows Eq. (6),

$$\sigma_\varepsilon^2(XCO_{2.sim.ak}) = \sum_{n=1}^{nlevel} w_n^2\, \sigma_\varepsilon^2\, (CO_{2.sim.ak.n}) + 2\sum_{1 \le n<m \le nlevel} w_n\, w_m\, cov_\varepsilon(CO_{2.sim.ak.n}, CO_{2.sim.ak.m}), \quad (6)$$

where w_n denotes the product of AK_{norm} and PW at level n; and cov_ε represents the correlation of transport errors between every two levels n and m ($1 \le n < m \le nlevel$). To calculate a typical vertical error correlation length scale, we fit an exponential variogram according to transport errors and their separation distances between levels. Results of transport error at each sounding and its latitudinal integration for each track are shown in Sect. 3.4 and Sect. 3.5.

In addition to above random error component, we are aware of potential systematic wind errors in certain areas, e.g., positive wind speed bias reported over Los Angeles (Ye et al., 2017), and their impacts on both the forward- and backward- time simulations. As an attempt to resolve these obstacles, X-STILT can incorporate a near-field wind bias correction (to both backward- and forward-time simulations). By rotating model trajectories, this bias correction aims at "correcting" air parcel distributions and resultant footprints, given knowledge that the near-field wind bias can be properly interpolated. Details about this wind bias

correction are described in Appendix C. Unfortunately, only 2 radiosonde stations around Riyadh with 3 vertical pressure levels within the PBL (and sometimes with missing data) may be insufficient to correctly interpolate the near-field vertical wind biases. However, cities with meteorological profiles sampling more levels within the PBL and higher temporal frequency in reporting observed vertical winds will be more suitable sites to retrieve the near-field wind errors. Other methods include rotation and stretching of urban plumes derived from WRF-Chem (Ye et al., 2017), similar to the rotation of X-STILT air parcels, to quantify errors in wind directions and speeds. Deng et al. (2017) sought correction of wind biases in a sophisticated manner via data assimilation. Yet, the near-field correction within X-STILT can be potentially utilized in the future as a quick bias correction to the near-field wind in LPDMs, given denser wind observations and relatively flat terrains. Therefore, we decided to reduce the potential impact of wind bias on model-data comparisons using a latitudinal integration (further in Sect. 3.5).

2.6.2 Vertical transport errors

Vertical turbulent mixing dominants the vertical transport of air parcels and control the dilution of surface emissions within the PBL (e.g., Gerbig et al., 2008). Uncertainties in the vertical mixing or PBL height can affect both the footprint magnitude and the its spatial distribution via different horizontal advections at each altitude. Although column-integrated measurements may be less sensitive to vertical distribution of air particles than in situ measurements, vertical transport errors can have some impacts on column simulations nonetheless, due to wind shear and its interaction with vertical redistribution of air parcels (Lauvaux and Davis, 2014). Comprehensive quantifications of the vertical transport errors in a column sense are performed in Lauvaux and Davis (2014) using ensemble of surface and planetary BL parameterizations involving a regional inverse modeling framework.

Instead, we made use of the stochastic nature of STILT and propagated typical PBL height errors in the model. Changes in STILT-modeled mixed layer height modify the vertical profiles of turbulent statistics that directly control the stochastic motions of the Lagrangian air parcels (Lin et al., 2003). Thus, we obtained different air parcel trajectories with rescaled PBL heights. The resultant vertical transport error in XCO_2 space is calculated as the root-mean-squared errors (RMSEs) between two sets of XCO_2 enhancements among different receptors for each overpass. Due to this calculation, vertical transport errors are only provided at the overpass level (results in Sect. 3.5). Gerbig et al. (2008) reported typical relative PBL errors in the range of ± 20 %. Thus, we rescaled the PBL heights higher and lower by 20 % and evaluated the scaling's impact on XCO_2 enhancements. Because of our focus on the urban emissions and potential small XCO_2 enhancements contributions beyond one day backwards in time, we only rescaled PBL within the first 24 hours of transport before arrival of the air parcels at the column receptors.

3 Results

3.1 X-STILT sensitivity tests with column receptors

Fig. 6 shows test results given a sounding on 12/29/2014 around Riyadh. In general, urban enhancements increase as MAXAGL increases from 1–2 km and then stabilizes (Fig. 6a). When MAXAGL is small (< 2 km), the model fails to fully capture the CO_2 enhancements within the mixing height and causes underestimations on the elevated XCO_2. Random errors reflect the stochastic nature of the model particles leading to small fluctuations in parcel distributions and resultant signals. In this experiment, *dpar* and *dh* are fixed to 100 particles and 100 m. For testing particle number per level *(dpar),* MAXAGL is set to 6 km (well above the top of the PBL; see Appendix D for the choice of 6 km). No obvious bias is associated with mean XCO_2 enhancements. The random error reduces as particle numbers increases (error bars in Fig. 6b). We ended up placing 100 particles for each level, as random errors do not change dramatically from 100 to 200 particles.

In addition, we conducted two experiments using constant and uneven vertical spacings with the fixed MAXAGL of 6 km and

dpar of 100. Vertical spacing in the constant *dh* experiment ranges from 50 m to 1 km. Mean enhancements generally decrease as vertical spacing increases (red dots in Fig. 6c), likely because fewer release levels are insufficient to represent air parcels in a column and their interactions with surface emissions, especially under strong wind shear. We further performed two cases with uneven vertical spacing below and above a "cutoff level". Both tested three different lower spacings (of 50, 100 or 150 m) with a fixed upper spacing of 500 m. Two cases differ only in their cutoff levels (2 or 3 km). The comparison of the uneven *dh* against the constant *dh* experiment shows that their results in XCO_2 enhancements are fairly similar, suggesting that the lower spacing below the cutoff level matters mostly to model results, because most anthropogenic XCO_2 enhancements are confined within the PBL. Also, results for uneven *dh* case with the cutoff level of 3 km (blue triangles in Fig. 6c) are more closed to the "truth" implied by the constant *dh* case (red dots in Fig. 6c). To be safe, column receptors are placed from 0–3 km with a spacing of 100 m and from 3–6 km with a spacing of 500 m. See Appendix D for the derivation of the cutoff level.

To summarize, the fractional uncertainties in modeled XCO_2 enhancements reduce rapidly as total particle number increases (blue triangles in Fig. 6d). Our choice of column receptors and particle numbers has no noticeable bias and a fractional uncertainty of ~4 % per simulation (dashed green line). Overall changes in X-STILT column receptors have a fairly small impact on modeled anthropogenic signals, which is consistent with the finding (for biospheric signals) in Reuter et al. (2014).

3.2 X-STILT column footprints and upwind emission contributions

Upstream source regions and their contributions to downwind air column can be identified as the "footprint" using backward-time simulations. Here we illustrate the differences in parcel distributions and footprint patterns derived from 500 m, 3 km, and multiple levels, for one sounding at 24.4961° N on 12/29/2014 (when southwestern winds dominated). Air parcels released at 500 m are associated with large footprints in the adjacent area of Riyadh (Fig. 3b). While parcels released from a higher level of 3 km travel much faster to their upwind regions, where most parcels barely get entrained back into the PBL (Fig. 3c) and make minimal contact with the surface implied from zero footprint values in Fig. 3d. When air parcels are released from multiple levels, the column footprints (Fig. 3f) cover a broader spatial domain with relatively smaller values than footprints derived from 500 m (Fig. 3b). The intention here is to illustrate the difference in upwind influences from a PBL-based tower-like measurement versus a column-integrated measurement (e.g., satellite). As expected, surface influence arriving at an air column can be one or a few orders of magnitude smaller than that arriving at a given location. Consequently, CO_2 changes within the PBL are expected to be larger than column changes. If zooming into the near-field land surface, westerly winds dominated during the 12/29/2014 event. XCO_2 contribution maps indicate large contributions due to urban emissions of Riyadh (Fig. 7b, 7f) and small contributions from regions to the south of Riyadh (Fig. 7a, 7e), regardless of the adopted meteorological fields.

3.3 Comparisons between methods to calculate background XCO_2

As Silva and Arellano (2017) have pointed out their 4° × 4° urban extent may be too coarse for studying urban emissions we only borrowed their statistical assumption (μ-σ) and used a 2° × 2° domain for computing M2S' background. M2S and M3 calculate background values from local observations. Therefore, M2S may agree better with M3 in their derived background regarding both the temporal variations and their magnitudes (black diamonds and green squares in Fig. 6e). M1 modeled background differs significantly from the other three and exhibits positive biases spanning roughly from 0.5–1.5 ppm (orange dots in Fig. 6e). Reasons to this large bias may be the accumulated transport errors as backward duration increases together with potential errors in the global concentration fields with its coarse resolution (2° × 3°).

We now focus on the comparison between M3 and M2H with objectively analyzing their advantages and limitations. On

average, M2H derived background is lower than our localized "overpass-specific" background by 0.55 ppm (Fig. 6e), which can primarily be attributed to different defined background regions. M3 defined the background region from the same track as the one over Riyadh, which guarantees that the background air contains variations due to long-term atmospheric transport, natural sources/sinks and $FFCO_2$ emissions except for local emissions (e.g., from Riyadh). Whereas the enhanced air contains the enhancements due to local emissions on top of all the information included in the background air. Therefore, the subtraction between M3-defined background and enhanced air correctly represent the XCO_2 portion enhanced by the local emissions. On the contrary, M2H use a fairly broad background region (0° N–60° N, 15° W–60° E in Fig. S7) to estimate gridded anomalies over all places in Europe, Middle East and North Africa. Although may yield more data, this broad spatial region may misrepresent the correct upwind region, because the wind regime can be quite different among different overpass dates or seasons.

We admit M3-defined background range and background value can be affect by potential large wind bias over cities other than Riyadh. However, the impact on background may be small and is implicitly considered in the background uncertainty (previously discussed in the last paragraph of Sect. 2.3.3). As for M2H, all regional OCO-2 measurements are lumped into its background calculation. For example, some measurements on the east-most overpass in Fig. S7 are affected by Riyadh's emissions, whereas atmospheric columns at soundings along the west two overpasses in Fig. S7 may not necessarily be the background air that eventually arrives at region around Riyadh. Thus, the regional median of XCO_2 may not physically indicate the accurate background that is supposed to isolate local-scale fluxes. Therefore, our localized overpass-specific background is designed and more suitable for extracting local-scale XCO_2 anomalies. Given relatively small urban enhancements around our study site, this 0.55 ppm difference may lead to large differences in estimated observed urban signals and emission evaluations (Sect. 4.2).

3.4 Latitude-dependent urban enhancements and associated uncertainties

We compare both the magnitude and shape of modeled and observed anthropogenic enhancements along the track. Models using GDAS and WRF report fairly similar XCO_2 peaks as bin-averaged observations for the 12/29/2014 overpass (Fig. 8). Although XCO_2 contributions using GDAS and WRF can differ in their spatial distributions for some receptors (Fig. 7b vs. 7f), the overall XCO_2 contributions integrated from all receptors along the overpass share fairly similar spatial distributions and magnitudes (Fig. 7d vs. 7h). Regarding the shape of latitude-dependent XCO_2 enhancements, large enhancements inferred from bin-averaged observations (solid black line in Fig. 8) cover a wider range compared to narrower modeled enhancements (dashed blue or purple lines in Fig. 8). Also, modeled versus observed enhancements exhibit a 0.1° latitudinal shift for event on 12/29/2014 (Fig. 8) and vary from 0.1°–0.4° for other events (Fig. S8). Column simulations with strong near-field influences can be sensitive to potential errors in the near-field wind speeds and directions along with errors in the gridded emissions. And the limited wind observations within the near-field land surface around Riyadh make it even harder to estimate representative wind statistics that can be directly linked with model-data mismatch in XCO_2 shapes and magnitudes. All these challenges lead us to perform a latitude integration on the urban XCO_2 enhancements over a certain latitudinal band to reduce near-field sensitivity on model-data comparisons and emission evaluations (further discussed in Sect. 3.5).

Based on available radiosonde sites over the Middle East with relatively flat terrain (white crosses in Fig, 4), regional RMS errors associated with the GDAS u- and v-component winds are mainly < 2 m s^{-1} (Fig. S1) and generally smaller than those from previous studies over regions with relatively more complex terrains (Henderson et al., 2015; Lin et al., 2017). Even though positive/negative biases may exist per overpass, the averaged wind bias over a dozen tracks is fairly small, with absolute values close to zero. That is, no obvious systematic error over times is found in GDAS wind field around Riyadh. Similarly, Ye et al. (2017) reported no bias in the transport for Riyadh using WRF-Chem. Because of the spatial inhomogeneity in urban emissions,

wider parcel distributions after randomization may have higher possibilities in making contact with more emission sources than those before. Take the 12/29/2014 track as an example. Small transport errors can often be found over less polluted latitudinal range (< 24.3° N and > 24.9° N in Fig. 8). Transport errors then start to increase as few randomized parcels tend to "hit" some emission sources, even though simulated enhancements are still small (24° N–24.5° N and 24.7° N–24.8° N in Fig. 8). Although air parcels at higher altitudes are also under perturbations, the change in parcel distribution may hardly impact the column transport errors due to minimal contact of those parcels with surface emissions. As a result, the transport error per sounding for this overpass ranges from 0.07–2.87 ppm (Fig. 8). For the other tracks with more intense urban enhancements, maximum transport error per sounding can reach > 5 ppm, e.g., 2016011510 in Fig. S8. In addition, XCO_2 errors due to vertical mixing error are not provided at the sounding level given our method described in Sect. 2.6.2 but are reported on a per overpass basis later in Sect. 3.5.

Spatial fractional uncertainties in gridded emissions over the Middle East (Fig. S3) can be comparable to few prior studies. For instance, several commonly-used inventories differ by > 100 % over half of examined 0.1° gridcells (Gately and Hutyra, 2017) in the northeastern U.S. Resultant XCO_2 uncertainties due to prior emission errors range from 0.1–1.48 ppm per sounding for the overpass on 12/29/2014 (Fig. 8) and 0.04 – 2.82 ppm for all 5 overpasses (orange ribbon in Fig. S8).

Retrieval errors are reported for each sounding by the OCO-2 Lite file and exhibit a Gaussian-like distribution with the most frequent values of 0.45–0.5 ppm. Background uncertainty (e.g., green ribbon in Fig. 5b) varies from 0.77–1.00 ppm among tracks. Overall observed uncertainty per sounding varies from 0.8–1.27 ppm. Worden et al. (2017) accounted for the natural variability in observed XCO_2, the measurement noise errors with error covariance within spatial domain of 100 km × 10.5 km. They found the overall precision of a measurement (WL < 10) over the land is ~0.75 ppm. Our larger observed uncertainties per sounding may be attributed to no filter of observations using WL, different examined regions and time periods and inclusion of background uncertainty (for the purpose of inverse analysis).

On a per sounding basis, XCO_2 resulted from horizontal wind errors are comparable to or higher than XCO_2 emission errors. Both errors are higher than observed uncertainties. Yet, uncertainty reductions are expected as sounding-level uncertainties are aggregated along the track (Sect. 3.5).

3.5 Latitudinally-integrated urban signals and uncertainties

Because shapes and locations for XCO_2 peaks between models and observations did not line up perfectly (Sect. 3.4; Fig. 8), direct model-data comparison may lead to significant deviations for each sounding. Thus, we compare urban signals with their associated errors integrated over a latitude band for each overpass.

Firstly, we integrated bin-averaged observed or modeled anthropogenic enhancements (i.e., differences between total XCO_2 and overpass-specific background) along their latitudes. While multiple degrees of freedom are sacrificed by this integration, this calculation gains a larger benefit of potentially reducing the impact of near-field wind bias on emission evaluations, as long as the latitude band for aggregation is representative. Secondly, a representative latitudinal range for integration (e.g., ~24° N–25.2° N in Fig. 8) is required. Note that negative observed urban enhancements may occur when the bin-averaged total observed XCO_2 is slightly lower than background value. The occurrence of these negative values is partially caused by the natural variations in measured XCO_2 and have been included as the background uncertainty. To minimize the inclusion of those negative values, we start with the enhanced latitudinal range (e.g., 24.2° N–24.9° N in Fig. 5b) and further account for latitudinal mismatch in model-data XCO_2 peaks. To further include urban enhancements over the "tails" outside the distinct XCO_2 peaks, we then extend the previous latitudinal range by 20 % on both sides. We tested percentages other than 20 % and found no dramatic changes in estimated signals due to small enhancements outside the plume (Appendix E).

Overall, the latitude-integrated modeled XCO_2 signals range from 0.64–3.04 ppm-degree with a mean signal of 1.57 ppm-degree, whereas the observed signals detected by OCO-2 vary from 1.09–2.92 ppm-degree with a mean value of 1.65 ppm-degree (Table 1, Fig. 9a). The magnitudes of observed signals can be slightly affected by how observations are selected and binned up (Appendix E1).

To arrive at integrated errors per overpass, error variance-covariance matrices can be built. For example, diagonal elements comprise transport error variance per sounding/receptor with off-diagonal elements filled with error covariance between each two soundings/receptors (Fig. S9a). A correlation length scale of transport errors (~25 km) among receptor locations is estimated by fitting exponential variograms (Fig. S9b), given the transport errors (further driven by plume structures) and our choice of grid spacing between receptors. And, similar calculations are performed to integrate sounding-level errors to overpass-level errors due to various error sources. Moreover, assuming errors are independent given the multiple days to months between overpasses, overpass-level XCO_2 errors (Table 1) are further aggregated to arrive at an overall error for all 5 overpasses.

XCO_2 errors solely resulted from vertical mixing errors are in general < 15 % of the modeled signal for each overpass, whereas XCO_2 errors due to horizontal wind errors dominate the overall XCO_2 transport error (Table 1). The random uncertainties due to the choices of column receptors/parcels are negligible, < 1 % of the latitude-integrated modeled XCO_2 signal per track. The 68 % confidence limits of XCO_2 uncertainties due to errors in prior emission and transport (i.e., horizontal wind fields and vertical mixing) are 0.32 ppm-degree and 0.52 ppm-deg., which is ~20 % and 33 % of the mean modeled urban signal over 5 tracks, respectively (Table 1). The integrated XCO_2 transport error per track reflects the aggregate effect of several factors which interact, given how we propagate wind errors into XCO_2 space (Sect. 2.6):

1) The magnitude of the modeled urban XCO_2 enhancements. In general, air parcels that are very far away from potential upstream emitters may hardly "hit" the emission sources or gain their enhancements, even after the wind perturbation. If the estimated signal is large (e.g., 3.04 ppm-deg. on 20151216 in Table 1), its resultant integrated transport error can also be fairly large (1.83 ppm-deg. in Table 1).
2) The RMSE of u- and v-component winds. In general, larger wind errors will lead to larger changes in model trajectories and larger possibilities for perturbed trajectories in intersecting an emission source.
3) How air parcels interact with surface emissions, i.e., the geometry/angle between the model footprint (or the wind direction) and satellite swaths. Changes in this angle may fluctuate the width of enhanced latitudinal band along with the final integration latitudinal ranges (i.e., 1.10°–2.25°). If the back-trajectory or backward wind direction is more parallel to the OCO-2 swath (events on 20141227, 20151216 and 20160216 in Fig. S10), the integration range and error covariance among soundings are usually larger, which yields larger integrated XCO_2 errors (e.g., 1.22, 1.83, and 1.05 ppm-degree in Table 1). The averaged latitudinal range for integration is about 1.66° (~189 km) over 5 tracks.

Retrieval errors between OCO-2 soundings are found to be correlated in both space and time, with correlation coefficients (for land nadir) of 0.45 and 0.31 as a function of satellite footprint and time, respectively (Worden et al., 2017). Uncertainties of bin-averaged observed XCO_2 share similar source as the background uncertainties, both of which rely on spatial variation in noisy observations (in each bin or over background region). Different types of observed uncertainties are assumed to be uncorrelated. Because observations along with their uncertainties have been binned up and the satellite footprints for bin-averaged observed uncertainties are hard to track, we only account for the temporal correlation of retrieval errors between every two soundings. As a result, total observed uncertainty per track vary from 0.33–0.50 ppm-degree and the 68 % confidence limit of observed error is 0.19 ppm-deg., i.e., ~ 11 % of the mean observed signals (1.65 ppm-deg.) over total 5 overpasses.

4 Discussions

4.1 Model capabilities and performances

In this study, we demonstrate the coupling of forward- and backward-time Lagrangian particle dispersion model simulations within X-STILT and model applications in locating the urban plume, determining background XCO_2, identifying upwind sources, and estimating enhanced XCO_2 caused by sources/sinks (Fig. 1). Specifically, backward-time simulations over an atmospheric column connect upwind emission sources with downwind atmospheric columns and generate spatial maps of this connection with additional information from satellite retrieval profiles. Although forward-time simulations from an urban box are an alternative and optional portion of X-STILT, these simulations help gain information regarding the location and size of the time-varying urban plume (Fig. 5a) and locate downwind polluted range on a satellite overpass.

Model sensitivity tests suggest two implications on simulating urban XCO_2 enhancements using LPDMs: 1) Receptor levels need to reach levels exceeding a typical mean PBL height to fully capture influences from surface emissions. 2) The model may capture a larger urban signal as number of levels increases. But, to minimize computational costs, one may try sparser and denser levels above and within a representative mean PBL height (the cutoff level) over upwind regions. Users can adopt their own setup of receptors in X-STILT according to combined results from sensitivity tests (Fig. 6d).

Additionally, X-STILT offers alternative solutions in dealing with errors in the meteorological fields, including regional random wind error perturbations and potential near-site wind bias corrections on model trajectories (Appendix C). For several satellite overpasses over Riyadh, models using WRF and GDAS are capable of capturing XCO_2 enhancements due to urban emissions, even though there remains small mismatch in the locations of model-data XCO_2 peaks. Model-to-model discrepancy between GDAS and WRF in latitudinally-integrated urban signals is not large, benefiting from relatively flat terrain and similar interpolated terrain heights around Riyadh. No noticeable difference in overall RMSE in u- and v- component winds derived from radiosonde comparisons with WRF versus GDAS is reported in this case. Thus, global meteorological fields such as 0.5° GDAS can be used for studying "flat cities" like Riyadh.

When dealing with enhancements in column concentration with small signal-to-noise ratio, careful examination to modeled background XCO_2 should be taken care of. Although one can possibly "eyeball" the city plume from observed XCO_2 (especially when a signal XCO_2 peak is visually distinctive), forward-time simulations with additional accounts for transport errors implemented in X-STILT may provide a more objective and efficient way (in that valuable human time is unnecessary) in figuring out the potential downwind sections along track that are affected by the city plume and extrapolating background region and its value. These advantages of overpass-specific background will become more important as more satellite tracks are incorporated within the analyses and future flux inversions.

4.2 Implications on error analysis and future inversion using LPDMs

Column transport uncertainties have not been rigorously examined for studies employing LPDMs like STILT and column measurements like OCO-2. In this work, we conducted comprehensive analysis towards observed errors incorporating natural and spatial XCO_2 variabilities, background and retrieval uncertainties; simulated errors including errors resulted from model configurations, horizontal and vertical atmospheric transport and prior emissions. On average, column transport errors (with 68 % confidence limit) contribute to 33 % of the mean modeled urban signal over 5 overpasses, whereas the horizontal transport error on a per track basis are still substantial. We also accounted for horizontal transport error correlations among X-STILT release levels and among multiple soundings. For instance, the horizontal transport error covariance between soundings is responsible for

about 67 % of the latitude integrated errors, which emphasizes the importance of error covariance on model evaluations (e.g., Lin and Gerbig, 2005).

Estimated background uncertainty is represented by the spatial variation and retrieval errors of background observations and may be reduced given large sampling size. To further demonstrate X-STILT's potential role in inverse modeling and the potential background "bias" via different background methods on inversed results, we conducted a simple scaling factor inversion (Rodgers, 2000), based on 5 pairs of model-data latitudinally-integrated urban signals. Even though our sampling may seem to be small and the gridded urban source emissions are treated as a whole (i.e., no adjustments for emissions for each gridcell), these integrated signals and errors are chosen to reduce the impact of potential near-field wind bias on model evaluations. Also, we are partially limited by the overpasses over Riyadh (black bars in Fig. S1). The prior emissions from ODIAC are assumed to be "unbiased", which yields a prior scaling factor of unity (λ_a = 1). The prior error (S_a) represents the overall uncertainties of the sum and spatial spread of ODIAC emissions around Riyadh (further calculated from the inter-comparisons against FFDAS and EDGAR). Observational error covariance matrix (S_λ) contains error variances related to observation and horizontal and vertical transport errors (Table 1). Errors between every two overpasses are assumed to be independent.

Our conservative results based on GDAS suggest that the posterior scaling factor ($\hat{\lambda}$; of mean XCO_2 signal) and its posterior uncertainty (of the scaling factor) is 1.14 ± 0.31 using background from M3. However, potential errors in background XCO_2 defined by other methods may affect resultant observed signals and posterior scaling factors. Since the M2H- and M1-derived background values are generally lower and higher than M3 background, M2H- and M1-derived background values result in a higher and lower mean observed signal (2.30 and 0.88 ppm-degree in Table 1) than that based on M3 (1.65 ppm-degree). Furthermore, $\hat{\lambda}$ based on M2H is about 2.30, larger than that using M3 by 40 %. The $\hat{\lambda}$ derived from M3 background (1.14) can be more comparable to the WRF-Chem-based emission estimate in Ye et al. (2017). These results again emphasize the significant role of background definitions played in estimated observed signals and emission estimates. In particular, simple statistical approaches without considering the atmospheric transport may lead to erroneous conclusions (previously discussed in Sect. 3.3).

4.3 X-STILT's potential for broader applications

In theory, X-STILT can be applied to other column measurements and other species. The underlying Lagrangian atmospheric model (STILT) has been applied to simulate other atmospheric species, such as CO, CH_4 and N_2O (Mallia et al., 2015; Kort et al., 2008). One of the key modifications to X-STILT from STILT is the column weighting of STILT footprint values (Sect. 2.1.2). Specifically, X-STILT interpolates the OCO-2 *AK* and *PW* onto each modeled level and then applies weighting of the trajectory-level footprints before generating a horizontal footprint map. The X-STILT code can be easily modified to apply sensor-specific vertical profiles of *AK* and *PW* from other satellites or ground-based column measurements.

Lastly and more importantly, background may need to be derived differently according to different applications, e.g., local urban emissions versus regional fluxes. The overpass specific background (M3) aims at isolating the citywide emissions, so it makes use of the measurements outside the city, but still are quite closed to the city (within the few degrees latitude). However, if the study focus is to look at emissions over a much broader region (e.g., statewide emissions), background region should be defined farther away from the target region, e.g., taking the advantages of measurements from available upstream overpasses.

4.4 Limitations and future plans

Robust constraints on urban emission can be hampered due to their alternating-sign nature and signals potentially comparable to anthropogenic emissions (Shiga et al., 2014; Ye et al., 2017), which are also inferred from tracks we modeled over Cairo with non-

negligible biomass (results not shown in this paper). When examining summertime tracks or tracks over some other cities, potential local gradients in biospheric fluxes should be considered as those gradients can affect our overpass-specific background. Although biospheric fluxes or their resultant changes in XCO_2 concentrations are beyond the scope of this work, many studies have been working to address this challenge. Ye et al. (2017) incorporated biospheric fluxes from the North American Carbon Program (NACP) Multi-scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP; Fisher et al., 2016; Huntzinger et al., 2013) and performed downscaling on biospheric fluxes using MODIS-derived Green Vegetation Fraction (GVF), to provide high-resolution biospheric flux fields and estimated background XCO_2 by modeling. Besides, radiocarbon and terrestrial solar-induced chlorophyll fluorescence (SIF) data are helpful to isolate fossil fuel CO_2 and biospheric CO_2 (Fischer et al., 2017; Levin et al., 2003; Sun et al., 2017). In particular, recent studies have identified SIF as a better indicator/proxy of gross or net primary production than some other greenness indices over several different vegetation types (Shiga et al., 2018; Sun et al., 2017; Zuromski et al., 2018), which improves biospheric flux estimation in ecosystem models and benefits the interpretation of OCO-2/OCO-3 retrievals (Luus et al., 2017).

X-STILT extends its way to account for transport errors, background uncertainties and particle statistics in a column sense within LPDMs. Admittedly, the transport error analysis and near-field correction may work the best with the assistance of denser meteorological observing networks to characterize the error structures of transport errors. Increasing the density of surface networks may modify the wind error statistics including the wind error variances and horizontal correlated length-scale, and further impact the model transport uncertainties and inversed fluxes. Yet, this shortcoming is not inherent to X-STILT and applies to other means of quantifying the transport errors based on real data as well. The trade-off of choosing a city in the Middle East like Riyadh to minimize cloud and vegetation influences is the relatively sparse observations of surface meteorological network or aircraft. The most recent OCO-2 b8 Lite files include retrieved surface winds for each sounding. Unfortunately, most of those surface wind retrievals are not available over Riyadh, but the retrieved surface winds for other urban areas, if available, may be used for assimilation and assisting X-STILT error analysis.

Emission evaluations for different regions can be different and affected by different observational constraints. Even changes in different versions of the retrieval (Lite b7 vs. b8) may slightly affect the model-data comparisons and simple inversion results in this work. Modeled XCO_2 enhancements using the newer b8 differ slightly from those using b7 (purple dots in Fig. S8 vs. in Fig. S13) due to changes in the locations of the receptors, column averaging kernels, and data filtering (QF) for measurements around Riyadh. Specifically, observations from b8 may yield more overpasses with sufficient screened soundings than those from b7 (black and red bars in Fig. S1). However, much larger differences in observed enhancements are found and caused by the changes in total observed XCO_2 and estimated background values. Specifically, background uncertainty decreases by up to 0.1 ppm primarily attributed to smaller spread (smaller SD) of the observed XCO_2. Positive shifts in the total observed XCO_2 for b8 from b7 are found over most overpasses (Fig. S11). The M3-derived observed enhancements may be less affected by positive shifts in total observations, given similar positive shift associated with the overpass-specific background near the target urban region (dark green dashed lines in Fig. 6e vs. Fig. S12).

OCO-2 observations have been utilized in several recent studies along with this work with a particular look into relatively small areas, e.g., individual power plants (Nasser et al., 2017) and megacities (Ye et al., 2017). Even though the XCO_2 urban signal over Riyadh may be in general smaller than those over other large cities, both model and observation successfully detect the urban signal. Still, no summertime XCO_2 signal has been derived, due to the lack of screened observations (QF = 0) reported in OCO-2 Lite b7 file over most summertime tracks (black bars in Fig. S1). No diurnal variation, revisit time of 16 days and relatively narrow swath of OCO-2 may still pose challenges to urban emission estimates. We expect the inclusion of more column observations in

stationary (target) modes, e.g., by scanning over megacities by OCO-3 (Eldering et al., 2016), which may offer more concrete spatial and diurnal variabilities that benefits urban flux inversions. Many nations are devoting considerable resources in launching carbon-observing satellites that can potentially be coordinated in a larger monitoring system (Tollefson, 2016). Given that X-STILT can potentially work with most satellites (given their sensor-specific vertical profiles), we expect enhanced capability in emission constraints of urban emissions by combining column measurements with X-STILT.

Code availability. X-STILT is built on STILT (Lin et al., 2003) and STILT-R version 2 (Fasoli et al., 2018), which can be downloaded from GitHub repository (https://github.com/wde0924/X-STILT). The version of the X-STILT code coinciding with the work described in this manuscript is on Zenodo (http://doi.org/10.5281/zenodo.1432528). Model developments are still ongoing.

Appendices

Appendix A: Four conservative criteria to select overpasses over Riyadh

We accounted for four factors, including 1) the prevailing wind directions and downwind regions; 2) the portion of soundings with QF = 0; 3) the distance between satellite track and the city center, and 4) regional wind errors in modeled meteorological fields. In the end, we selected 5 overpasses via manual check.

I. First of all, we defined a spatial domain (2° latitudes by 3° longitudes) centered around Riyadh (i.e., 24.71° N, 46.74° E) and counted the total sounding numbers that fall into this domain for each overpass. This spatial domain can be determined by examining prevailing wind directions and locating downwind regions based on wind rose plot from radiosonde stations at the city center and the airport of Riyadh (with 4-character international ID of OERK and OERY) during each overpass date. Alternatively, forward-time model runs starting from a box around the center of Riyadh allows us to determine polluted latitudinal ranges on satellite overpass (Fig. 5). Detailed demonstrations about the forward-time runs are in Sect. 2.3.3. Total 43 overpasses with at least one measurement fall into this designed spatial domain for Riyadh (gray bars in Fig. S1). **II.** Next, we ensured the amount of screened observed data using warn levels/quality flags (QF). Because high warn level is associated with high total aerosol optical depth inferred from soundings (Lite b7) near Riyadh, we only used quality flag to control data quality in this study. After selecting overpasses with > 100 soundings with QF = 0, 11 overpasses remain. Most spring- and summer- time tracks (during Mar–Aug) fail to satisfy this criterion (black bars in Fig. S1). Further, we ensured enough amounts of screened observations are falling within a prescribed urban domain (1° x 1° box) around the city center (red bars in Fig. S1). Only 8 overpasses have > 50 screened soundings (red dashed line in Fig. S1). **III.** Overpasses with distinct enhancements in retrieved XCO_2 due to urban emissions are preferred. Near-field domain affected by PBL processes may extend over 100–1000 km based on the globally averaged ventilation time for PBL (Lin et al., 2003). We made a conservative assumption on the impacted near-field domain being a circle with a radius of 50 km around the city center. Thus, we calculated the smallest distance between soundings and city center (orange dots in Fig. S1) and most pass this filter given our examined spatial domain. **IV.** As a final step, since model results can potentially be affected by meteorological fields, regional u- and v- wind RMS errors below 3 km (derived from comparison against radiosonde stations, white crosses in Fig. 4) are calculated (numbers in brown in Fig. S1). Details on the wind error calculation are in Appendix B.

Appendix B: Wind error calculation and regression-based transport error method in X-STILT

In terms of the wind error component ($\boldsymbol{u_\varepsilon}$) mentioned in Sect. 2.6, two sets of parameters are used to describe, 1) σ_{uverr}, the standard deviation of horizontal wind errors (RMSE) describing to what extent should we randomly perturb air parcels; and 2) horizontal and vertical length-scales and time-scales (Lx, Lz, and Lt) determining how wind errors are correlated and decayed in space and time. We calculated different sets of wind error statistics over 3 vertical bins, i.e., 0–3 km, 3–6 km and 6–10 km, for randomizing air parcels. To obtain σ_{uverr}, observed winds at mandatory levels (i.e., 925, 850, 700, 500, 400, 300 mb) from surrounding radiosonde sites (Fig. 4) are compared against WRF- or GDAS-interpolated winds. Then, we averaged wind errors at different mandatory levels over aforementioned three vertical bins. In addition, wind errors are considered to be spatiotemporally correlated. To determine error correlation scales, differences in the wind errors are calculated and wind errors at different radiosonde stations or different reported hours (00UTC or 12UTC) are paired up based on their separation length- or time-scales. An exponential variogram is then applied to estimate the horizontal, vertical and temporal correlation scales, which are the separation scales when errors become statistically uncorrelated.

Solution of negative transport errors: The CO_2 variance derived from model trajectories after the randomizations ($\sigma^2_{\varepsilon+u'}$) can occasionally be smaller than that before the randomization ($\sigma^2_{u'}$) for a few levels, due to insufficient parcel numbers (green dots in Fig. S5). Instead of abandoning these data, we developed a regression-based method to deal with the reduction in CO_2 variances. Specifically, we applied linear regression lines to the two sets of CO_2 variances before and after the randomizations, with weights of $1/\sigma^2_{\varepsilon+u'}$. That means larger variances are weighted lesser. When we used several other ways (without the weights) to apply linear regression, extremely large regression slopes and negative y-intercept occur, which potentially leads to unreasonable large transport errors (in ppm) at lower levels within the PBL and negative transport errors aloft. Then, we scaled and recalculated $\sigma^2_{\varepsilon+u'}$ based on weighted regression slope S_{WLR} and $\sigma^2_{u'}$. The regression line indicates the overall increase in CO_2 variance that serve as transport error in ppm:

$$\sigma^2_{\varepsilon}\,(CO_{2.sim.ak.n}) = (S_{WLR} - 1)\,\sigma^2_{u'}\,(CO_{2.sim.ak.n}), \quad \text{(A1)}$$

where the weighted linear regression is fitted for variances with versus without wind error component (dashed blue line in Fig. S5). Extremely large anthropogenic enhancement (e.g., >1000 ppm) for a given parcel may exist for a few cases. Thus, outliers (i.e., the upper 1st percentile of both parcel distributions before and after the randomizations) are removed for each level, before calculating variances in both CO_2 distributions.

Appendix C: Correcting for wind biases within X-STILT

While we did not apply the wind bias correction for the overpasses analyzed in this paper due to the biases being generally small (previously explained in Sect. 2.3.3), X-STILT has the capability to account for biases, if necessary. The basic idea is to correct the near-field wind biases in both forward- and backward- time trajectories. Because wind error at each observed pressure level can be quite different, vertically-weighted u- and v- wind biases were calculated by fitting logarithmic mean wind profiles based on available near-fields observed and simulated wind speeds and directions. We then calculated the deviations in latitude and longitude directions (dx, dy, with conversion from distance to degrees) given estimated u- and v- wind biases. These deviations accumulate as air parcels travel further backward or forward in time and are used to correct the location of each particle. After fixing the particle locations, Fig. S6b shows the general distribution of backward trajectory being clockwise rotated, compared to initial trajectory distribution in Fig. S6b. Air parcels in Fig. S6b appear to be "noisier" than those in Fig. S6a, due to inclusion of the random wind error component. Then the new bias-corrected set of column trajectory is used to generate spatial footprint.

This correction can also be performed to forward-time trajectory to reduce wind bias impact on best-estimated background value using the M3 method.

Appendix D: The determinations of MAXAGL and cutoff level

MAXAGL and a cutoff level (below which more model levels are placed) are the most important factors in determining modeled urban signals and can be determined based on few model trajectories starting from few satellite soundings for each overpass. Modeled mixing height *h* reported for an individual air parcel at a timestamp, $h(p, t_m)$, can be very high over the upwind desert region near Riyadh. We determine MAXAGL to be the maximum mixing height for each individual air parcel. To determine a cutoff level, we calculate the averaged *h* over all parcels as a function of backward time, as follows

$$\bar{h}(t_m) = \frac{1}{N_{tot}} \sum_{p=1}^{N_{tot}} h(p, t_m), \quad \text{(A2)}$$

where t_m represents the backward timestamp, ranging from 0 to 72 hours back. The mean modeled mixing heights among air parcels at each timestamp $h(t_m)$ exhibit a diurnal cycle, where expected high values present during the daytime. Also, $\bar{h}(t_m)$ typically display relatively high values where parcels are more concentrated within a day backward, and low values as parcels disperse outwards few days back. We ended up using the maximum value of mean mixing heights over parcels and over time as a representative cutoff level.

The maximum $h(p, t_m)$ and maximum $\bar{h}(t_m)$ are 5816 m and 2420 m, for the specific sounding we showed (Sect. 3.1, Fig. 6). Considering potential uncertainties in modeled PBL or mixing heights, these two numbers are rounded to 6 km and 3 km for a representative MAXAGL and cutoff level. In addition, we generalize the rules for placing column receptors to other seasons, based on aforementioned calculations. Maximum $h(p, t_m)$ and maximum $\bar{h}(t_m)$ over the upwind region vary slightly among different soundings during different seasons. Typically, maximum $h(p, t_m)$ are mostly under 6 km for wintertime soundings (Dec, Jan, and Feb), but can reach ~7 km and 10 km for soundings in spring/fall and summer. Maximum $\bar{h}(t_m)$ are < 3 km for wintertime tracks and ~4 km and 6 km for tracks in spring/fall and summer. Therefore, column receptors are placed from the surface to 3 km with 100 m spacing and 3–6 km with 500 m spacing for wintertime overpasses with 100 parcels per level (Fig. 3e). For other seasons such as the summertime, additional receptors are placed from 6–10 km with a spacing of 1 km, to ensure the model captures entire contributions from surface emissions. Although we expect relatively similar MAXAGLs and cutoff levels for most soundings over the Middle East, due to overlaps in upwind regions, these values should be recalculated when other cities are examined (Eq. A2).

Appendix E: Factors that may influence observed or modeled enhancements/signals

In Section 3.5, we integrated XCO_2 enhancements along latitudes to estimate modeled and observed signals within a certain latitudinal band for each overpass. This latitudinal band starts with enhanced latitudinal range, then gets corrected based on model-data latitudinal shift in XCO_2 peaks, and finally extends by 20% of its length. Also, we tested the impact of different percentages other than 20 % on latitudinally-integrated signals. Because of relatively small XCO_2 enhancements over background range, the impact due to different percentages (i.e., 10 %, 15 %, 20 %, 25 %) are relatively small—i.e., with changes of 0.03 ppm and 0.06 ppm in averaged modeled and observed signals, respectively. These small changes show that our latitude integration band is representative as it does not include a second peak or miss large XCO_2 enhancements.

E1 Influences on observed signals (bin-widths, warn levels)

These modeled and observed signals reported in Sect. 3.5.1 are calculated based on the uneven sampling choice for model receptor lat/lon described in Sect. 2.1.1; i.e., with smaller bin widths of 0.025° and larger bin widths of 0.05° over which urban influences

are stronger and weaker. In addition, we tested the impact on observed signals resulted from different bin widths with constant values starting from 0.01° to 0.5°. Both the latitudinal variation and the overall observed signal for an overpass generally decrease as bin widths increase, because bin-averaged observed XCO_2 enhancements get smoothed out, especially over latitudes with strong urban influences. Some information is lost in latitude-integrated observed signals based on our sampling choices when comparing against the signals calculated using constant bin widths such as 0.02°. Yet, binning observations based on the lat/lon of model receptors ensures a fair comparison with the model and our uneven sampling choices may better resolve XCO_2 enhancements within much finer grid spacing (particularly under urban influences) in the premise of limited computational resources. In addition, warn levels (WLs) may impact the filtering of observed data, bin-averaged observed XCO_2, defined background and conclusion regarding the model-data comparisons. Based on 3 simple tests by selecting measurements with QF = 0 and additional WL filters (WL < 10, 12, and 15), observed signals slightly increase, as more conservative WL filtering is applied. Changes in linear regression slopes and correlation between best-estimated modeled and observed signals due to sample choices and WL filtering are small.

E2 Influences on modeled signals (hourly vs. monthly emissions, nhrs, averaging kernel)

An additional set of hourly scaling factors (Nassar et al., 2013) can be applied to ODIAC to downscale the monthly mean emissions down to hourly values. In this study, we use monthly mean $FFCO_2$ emissions from ODIAC and apply TIMES to only 1 of the total 5 overpasses. Simulations including TIMES are slightly larger than those without the hourly scaling factors. Also, numbers of hours may impact the modeled enhancements at each sounding/receptor. We also conducted another simulation for 12/27/2014 event using model trajectories with only 24 hours back (different from 72 hours used in main text). The decrease in anthropogenic enhancements is < 0.05 ppm per sounding, which is small due to very small surface influence from far-away emission sources. Lastly, we report overall discrepancy in the modeled anthropogenic enhancements with or without weighting by OCO-2 prior profiles to be small. The difference is about 1–2 % of the weighted modeled anthropogenic enhancements, which is much smaller than impact caused by uncertainties in transport, emissions, or different setups. Note that XCO_2 portion from OCO-2's prior profile is zero and averaging kernel is simply unity everywhere for non-AK weighted simulations.

Acknowledgements. This work is based upon work supported by the National Aeronautics and Space Administration funding under Grant No. NNX15AI41G and the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1256260. We gratefully acknowledge Thomas Nehrkorn for providing modifications code to facilitate the forward-time box runs and thank Derek Mallia, Feng Deng, Arlyn Andrews, Andy Jacobson for their valuable advice on STILT-based modeling. The OCO-2 data were produced by the OCO-2 project at the Jet Propulsion Laboratory, California Institute of Technology, and obtained from the OCO-2 data archive maintained at the NASA Goddard Earth Science Data and Information Services Center. The authors acknowledge the NOAA Air Resources Laboratory (ARL) for the provision of the HYSPLIT transport and dispersion model and/or READY website (http://www.ready.noaa.gov) used in this publication. CarbonTracker CT-NRT.v2017 results provided by NOAA ESRL, Boulder, Colorado, USA from the website at http://carbontracker.noaa.gov. The support and resources from the Center for High Performance Computing (CHPC) at the University of Utah are gratefully acknowledged.

References

Andres, R. J., Gregg, J. S., Losey, L., Marland, G. and Boden, T. A.: Monthly, global emissions of carbon dioxide from fossil fuel consumption, Tellus, Ser. B Chem. Phys. Meteorol., 63(3), 309–327, doi:10.1111/j.1600-0889.2011.00530.x, 2011.

Andres, R. J., Boden, T. A. and Higdon, D.: A new evaluation of the uncertainty associated with CDIAC estimates of fossil fuel carbon dioxide emission, Tellus B Chem. Phys. Meteorol., 66(1), 23616, doi:10.3402/tellusb.v66.23616, 2014.

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(23), 14979–14995, doi:10.5194/acp-16-14979-2016, 2016.

Asefi-Najafabad, S., Rayner, P. J., Gurney, K. R., Mcrobert, A., Song, Y., Coltin, K., Huang, J., Elvidge, C. and Baugh, K.: A multiyear, global gridded fossil fuel emission data product: Evaluation and analysis of results, J. Geophys. Res. Atmos., 119(17), 10213–10231, doi:10.1002/2013JD021296, 2014.

Brasseur, G. P., and Jacob, D. J. Modeling of Atmospheric Chemistry. Cambridge University Press, 2017.

Basu, S., Guerlet, S., Butz, A., Houweling, S., Hasekamp, O., Aben, I., Krummel, P., Steele, P., Langenfelds, R., Torn, M., Biraud, S., Stephens, B., Andrews, A. and Worthy, D.: Global CO_2 fluxes estimated from GOSAT retrievals of total column CO_2, Atmos. Chem. Phys., 13(17), 8695–8717, doi:10.5194/acp-13-8695-2013, 2013.

Boden, T. A., Marland, G., and Andres, R. J.: Global, Regional, and National Fossil-Fuel CO2 Emissions, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., USA, doi:10.3334/CDIAC/00001_V2017, 2017.

Boesch, H., Baker, D., Connor, B., Crisp, D. and Miller, C.: Global characterization of CO_2 column retrievals from shortwave-infrared satellite observations of the Orbiting Carbon Observatory-2 mission, Remote Sens., 3(2), 270–304, doi:10.3390/rs3020270, 2011.

BP: Statistical Review of World Energy, available at http://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html, 2017.

Cambaliza, M. O. L., Shepson, P. B., Caulton, D. R., Stirm, B., Samarov, D., Gurney, K. R., Turnbull, J., Davis, K. J., Possolo, A., Karion, A., Sweeney, C., Moser, B., Hendricks, A., Lauvaux, T., Mays, K., Whetstone, J., Huang, J., Razlivanov, I., Miles, N. L., and Richardson, S. J.: Assessment of uncertainties of an aircraft-based mass balance approach for quantifying urban greenhouse gas emissions, Atmos. Chem. Phys., 14(17), 9029-9050, doi:10.5194/acp-14-9029-2014, 2014.

Ciais, P., Sabine, C., Govindasamy, B., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Quéré, C., Myneni, R., Piao, S., and Thornton, P.: Chapter 6: Carbon and Other Biogeochemical Cycles, in: Climate Change 2013 The Physical Science Basis, Cambridge University Press, Cambridge, 2013.

Crisp, D., Fisher, B. M., O'Dell, C., Frankenberg, C., Basilio, R., Bösch, H., Brown, L. R., Castano, R., Connor, B., Deutscher, N. M., Eldering, A., Griffith, D., Gunson, M., Kuze, A., Mandrake, L., McDuffie, J., Messerschmidt, J., Miller, C. E., Morino, I., Natraj, V., Notholt, J., O'Brien, D. M., Oyafuso, F., Polonsky, I., Robinson, J., Salawitch, R., Sherlock, V., Smyth, M., Suto, H., Taylor, T. E., Thompson, D. R., Wennberg, P. O., Wunch, D. and Yung, Y. L.: The ACOS CO_2 retrieval algorithm – Part II: Global XCO_2 data characterization, Atmos. Meas. Tech., 5(4), 687–707, doi:10.5194/amt-5-687-2012, 2012.

Cui, Y. Y., Brioude, J., Angevine, W. M., Peischl, J., McKeen, S. A., Kim, S. W., Neuman, J. A., Henze, D. K., Bousserez, N., Fischer, M. L., Jeong, S., Michelsen, H. A., Bambha, R. P., Liu, Z., Santoni, G. W., Daube, B. C., Kort, E. A., Frost, G. J., Ryerson, T. B., Wofsy, S. C. and Trainer, M.: Top-down estimate of methane emissions in California using a mesoscale inverse modeling technique: The San Joaquin Valley, J. Geophys. Res. Atmos., 122(6), 3686–3699, doi:10.1002/2016JD026398, 2017.

Dayalu, A., Munger, W., Wofsy, S. C., Wang, Y., Nehrkorn, T., Zhao, Y., McElroy, M. B., Nielsen, C. and Luus, K.: VPRM-CHINA: Using the Vegetation, Photosynthesis, and Respiration Model to partition contributions to CO_2 measurements in Northern China during the 2005–2009 growing seasons, Biogeosciences Discuss., in review, 2017.

Deng, A., Lauvaux, T., Brian, Gaudet, Kenneth, James, Davis, Kevin, Robert, Gurney, Risa, Patarasuk, Robert, Michael, Hardesty, And, Alan and Brewer: Toward Reduced Transport Errors in a High Resolution CO_2 Inversion System, Environ. Sci. Technol., 5, 5–20, doi:10.1021/es3011282, 2017.

Dlugokencky, E. and Tans, P.: Trends in atmospheric carbon dioxide, National Oceanic & Atmospheric Administration, Earth System Research Laboratory (NOAA/ESRL), available at: http://www.esrl.noaa.gov/gmd/ccgg/trends, 2015.

Duren, R. M. and Miller, C. E.: Measuring the carbon emissions of megacities, Nat. Clim. Chang., 2(8), 560–562, doi:10.1038/nclimate1629, 2012.

Efron, B. and Tibshirani, R.: Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science, 54-75, 1986.

Eldering, A., Bennett, M. and Basilio, R.: The OCO-3 Mission: overview of science objectives and status. In EGU General Assembly Conference Abstracts, 18, 5189, 2016.

Ellis, E.C. and Ramankutty, N.: Putting people in the map: anthropogenic biomes of the world. Frontiers in Ecology and the Environment, 6(8), 439-447, 2008.

Fasoli, B., Lin, J. C., Bowling, D. R., Mitchell, L., and Mendoza, D.: Simulating atmospheric tracer concentrations for spatially distributed receptors: updates to the Stochastic Time-Inverted Lagrangian Transport model's R interface (STILT-R version 2), Geosci. Model Dev., 11, 2813-2824, https://doi.org/10.5194/gmd-11-2813-2018, 2018.

Feng, S., Lauvaux, T., Newman, S., Rao, P., Ahmadov, R., Deng, A., Díaz-Isaac, L. I., Duren, R. M., Fischer, M. L., Gerbig, C., Gurney, K. R., Huang, J., Jeong, S., Li, Z., Miller, C. E., O'Keeffe, D., Patarasuk, R., Sander, S. P., Song, Y., Wong, K. W. and Yung, Y. L.: Los Angeles megacity: A high-resolution land-atmosphere modelling system for urban CO_2 emissions, Atmos. Chem. Phys., 16(14), 9019–9045, doi:10.5194/acp-16-9019-2016, 2016.

Fischer, M. L., Parazoo, N., Brophy, K., Cui, X., Jeong, S., Liu, J., Keeling, R., Taylor, T. E., Gurney, K., Oda, T. and Graven, H.: Simulating estimation of California fossil fuel and biosphere carbon dioxide exchanges combining in situ tower and satellite column observations, J. Geophys. Res. Atmos., 122(6), 3653–3671, doi:10.1002/2016JD025617, 2017.

Fisher, J. B., Sikka, M., Huntzinger, D. N., Schwalm, C. and Liu, J.: Technical note: 3-hourly temporal downscaling of monthly global terrestrial biosphere model net ecosystem exchange, Biogeosciences, 13(14), 4271–4277, doi:10.5194/bg-13-4271-2016, 2016.

Gately, C. K. and Hutyra, L. R.: Large Uncertainties in Urban-Scale Carbon Emissions, J. Geophys. Res. Atmos., 242–260, doi:10.1002/2017JD027359, 2017.

Gerbig, C., Lin, J. C., Wofsy, S. C., Daube, B. C., Andrews, A. E., Stephens, B. B., Bakwin, P. S. and Grainger, C. A.: Toward constraining regional-scale fluxes of CO_2 with atmospheric observations over a continent: 2. Analysis of COBRA data using a receptor-oriented framework, J. Geophys. Res. Atmos., 108(D24), 4757, doi:10.1029/2003JD003770, 2003.

Gerbig, C., Lin, J. C., Munger, J. W. and Wofsy, S. C.: What can tracer observations in the continental boundary layer tell us about surface-atmosphere fluxes?, Atmos. Chem. Phys., 6(2), 539–554, doi:10.5194/acp-6-539-2006, 2006.

Gerbig, C., Körner, S. and Lin, J. C.: Vertical mixing in atmospheric tracer transport models: error characterization and propagation, Atmos. Chem. Phys., 8, 591–602, doi:10.5194/acp-8-591-2008, 2008.

Göckede, M., Turner, D. P., Michalak, A. M., Vickers, D. and Law, B. E.: Sensitivity of a subregional scale atmospheric inverse CO_2 modeling framework to boundary conditions, J. Geophys. Res. Atmos., 115(D24), 2010.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Clais, P., Fan, S., Fung, I. Y., Gloor, M., Helmann, M., Higuchi, K., John, J., Maki, T., Maksyutov, S., Masarie, K., Peylin, P., Prather, M., Park, B. C., Randerson, J., Sarmiento, J., Tuguchi, S., Takahashi, T. and Yuen, C.-W.: Towards robust regional estimates of CO_2 sources and sinks using atmospheric transport models, Nature, 415(6872), 626–630, 2002.

Hakkarainen, J., Ialongo, I. and Tamminen, J.: Direct space-based observations of anthropogenic CO_2 emission areas from OCO-2, Geophys. Res. Lett., 43(21), 11,400-11,406, doi:10.1002/2016GL070885, 2016.

Henderson, J. M., Eluszkiewicz, J., Mountain, M. E., Nehrkorn, T., Chang, R. Y. W., Karion, A., Miller, J. B., Sweeney, C., Steiner, N., Wofsy, S. C. and others: Atmospheric transport simulations in support of the Carbon in Arctic Reservoirs Vulnerability Experiment (CARVE), Atmos Chem Phys, 15(8), 4093–4116, 2015.

Heymann, J., Reuter, M., Buchwitz, M., Schneising, O., Bovensmann, H., Burrows, J. P., Massart, S., Kaiser, J. W. and Crisp, D.: CO_2 emission of Indonesian fires in 2015 estimated from satellite-derived atmospheric CO_2 concentrations, Geophys. Res. Lett., 44(3), 1537–1544, doi:10.1002/2016GL072042, 2017.

Hogue, S., Marland, E., Andres, R. J., Marland, G. and Woodard, D.: Uncertainty in gridded CO_2 emissions estimates, Earth's Futur., 4(5), 225–239, doi:10.1002/2015EF000343, 2016.

Houweling, S., Breon, F. M., Aben, I., Rodenbeck, C., Gloor, M., Heimann, M. and Ciais, P.: Inverse modeling of CO_2 sources and sinks using satellite data: a synthetic inter-comparison of measurement techniques and their performance as a function of space and time, Atmos. Chem. Phys., 4, 523–538, doi: 10.5194/acp-4-523-2004, 2004.

Huntzinger, D. N., Schwalm, C. R., Michalak, A. M., Schaefer, K., King, A. W., Wei, Y., Jacobson, A. R., Liu, S., Cook, R. B., Post, W. M., Berthier, G., Hayes, D., Huang, M., Viovy, N., Lu, C., Tian, H., Ricciuto, D. M., Mao, J. and Shi, X.: The north american carbon program multi-scale synthesis and terrestrial model intercomparison project – Part 2: Environmental driver data, Geosci. Model Dev., 6(6), 2121–2133, doi:10.5194/gmd-7-2875-2014, 2013.

Janardanan, R., Maksyutov, S., Oda, T., Saito, M., Kaiser, J. W., Ganshin, A., Stohl, A., Matsunaga, T., Yoshida, Y. and Yokota, T.: Comparing GOSAT observations of localized CO_2 enhancements by large emitters with inventory-based estimates, Geophys. Res. Lett., 43(7), 3486–3493, doi:10.1002/2016GL067843, 2016.

Janssens-Maenhout, G., Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Dentener, F., Bergamaschi, P., Pagliari, V., Olivier, J. G. J., Peters, J. A. H. W., van Aardenne, J. A., Monni, S., Doering, U. and Petrescu, A. M. R.: EDGAR v4.3.2 Global Atlas of the three major Greenhouse Gas Emissions for the period 1970–2012, Earth Syst. Sci. Data Discuss., 10.5194/essd-2017-79, in review, 2017.

Jeong, S., Hsu, Y. K., Andrews, A. E., Bianco, L., Vaca, P., Wilczak, J. M. and Fischer, M. L.: A multitower measurement network estimate of California's methane emissions, J. Geophys. Res. Atmos., 118(19), 11339–11351, doi:10.1002/jgrd.50854, 2013.

Kim, S. Y., Millet, D. B., Hu, L., Mohr, M. J., Griffis, T. J., Wen, D., Lin, J. C., Miller, S. M. and Longo, M.: Constraints on carbon monoxide emissions based on tall tower measurements in the US Upper Midwest, Environ. Sci. Technol., 47(15), 8316–8324, 2013.

Kort, E. A., J. Eluszkiewicz, B. B. Stephens, J. B. Miller, C. Gerbig, T. Nehrkorn, B. C. Daube, J. O. Kaplan, S. Houweling, and S. C. Wofsy (2008), Emissions of CH_4 and N_2O over the United States and Canada based on a receptor-oriented modeling framework and COBRA-NA atmospheric observations, Geophys. Res. Lett., 35, L18808, doi:10.1029/2008GL034031.

Kort, E. A., Frankenberg, C., Miller, C. E. and Oda, T.: Space-based observations of megacity carbon dioxide, Geophys. Res. Lett., 39(17), 1–5, doi:10.1029/2012GL052738, 2012.

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(3), 1–8, doi:10.1002/jgrd.50135, 2013.

Lauvaux, T., Pannekoucke, O., Sarrat, C., Chevallier, F., Ciais, P., Noilhan, J. and Rayner, P. J.: Structure of the transport uncertainty in mesoscale inversions of CO_2 sources and sinks using ensemble model simulations, Biogeosciences, 6(6), 1089–1102, doi:10.5194/bg-6-1089-2009, 2009.

Lauvaux, T. and Davis, K. J.: Planetary boundary layer errors in mesoscale inversions of column-integrated CO_2 measurements. Journal of Geophysical Research: Atmospheres, 119(2), 490-508, 2014.

Lauvaux, T., Miles, N. L., Richardson, S. J., Deng, A., Stauffer, D. R., Davis, K. J., Jacobson, G., Rella, C., Calonder, G. P. and DeCola, P. L.: Urban emissions of CO_2 from Davos, Switzerland: The first real-time monitoring system using an atmospheric inversion technique. Journal of Applied Meteorology and Climatology, 52(12), 2654-2668, 2013.

Lauvaux, T., Miles, N. L., Deng, A., Richardson, S. J., Cambaliza, M. O., Davis, K. J., Gaudet, B., Gurney, K. R., Huang, J., O'Keefe, D., Song, Y., Karion, A., Oda, T., Patarasuk, R., Razlivanov, I., Sarmiento, D., Shepson, P., Sweeney, C., Turnbull, J. and Wu, K.: High-resolution atmospheric inversion of urban CO_2 emissions during the dormant season of the Indianapolis Flux Experiment (INFLUX), J. Geophys. Res. Atmos., 121(10), 5213–5236, doi:10.1002/2015JD024473, 2016.

Law, R. M., Rayner, P. J., Denning, A. S., Erickson, D., Fung, I. Y., Heimann, M., Piper, S. C., Ramonet, M., Taguchi, S., Taylor, J. A., Trudinger, C. M. and Watterson, I. G.: Variations in modeled atmospheric transport of carbon dioxide and the consequences for CO_2 inversions, Global Biogeochem. Cycles, 10(4), 783–796, doi:10.1029/96GB01892, 1996.

Levin, I., Kromer, B., Schmidt, M. and Sartorius, H.: A novel approach for independent budgeting of fossil fuel CO_2 over Europe by $^{14}CO_2$ observations, Geophys. Res. Lett., 30(23), 2194, doi:10.1029/2003GL018477, 2003.

Lin, J. C. and Gerbig, C.: Accounting for the effect of transport errors on tracer inversions, Geophys. Res. Lett., 32(1), 2005.

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, J. Geophys. Res. Atmos., 108(D16), 2003.

Lin, J. C., Gerbig, C., Daube, B. C., Wofsy, S. C., Andrews, A. E., Vay, S. A. and Anderson, B. E.: An empirical analysis of the spatial variability of atmospheric CO_2: Implications for inverse analyses and space-borne sensors, Geophys. Res. Lett., 31(23), 1–5, doi:10.1029/2004GL020957, 2004.

Lin, J. C., Gerbig, C., Wofsy, S. C., Daube, B. C., Matross, D. M., Chow, V. Y., Gottlieb, E., Andrews, A. E., Pathmathevan, M. and Munger, J. W.: What have we learned from intensive atmospheric sampling field programmes of CO_2?, Tellus, Ser. B Chem. Phys. Meteorol., 58(5), 331–343, doi:10.1111/j.1600-0889.2006.00202.x, 2006.

Lin, J. C., Mallia, D. V., Wu, D. and Stephens, B. B.: How can mountaintop CO_2 observations be used to constrain regional carbon fluxes?, Atmos. Chem. Phys., 17(9), 5561–5581, doi:10.5194/acp-17-5561-2017, 2017.

Liu, Y., Yang, D. and Cai, Z.: A retrieval algorithm for TanSat XCO_2 observation: Retrieval experiments using GOSAT data, Chin. Sci. Bull. 58(13), 1520–1523, doi:10.1007/s11434-013-5680-y, 2013.

Luus, K. A., Commane, R., Parazoo, N. C., Benmergui, J., Euskirchen, E. S., Frankenberg, C., Joiner, J., Lindaas, J., Miller, C. E., Oechel, W. C., Zona, D., Wofsy, S. and Lin, J. C.: Tundra photosynthesis captured by satellite-observed solar-induced chlorophyll fluorescence, Geophys. Res. Lett., 44(3), 1564–1573, doi:10.1002/2016GL070842, 2017.

Macatangay, R., Warneke, T., Gerbig, C., Ahmadov, R., Heimann, M. and Notholt, J.: A framework for comparing remotely sensed and in-situ CO_2 concentrations, Atmos. Chem. Phys., 8, 2555–2568, 2008.

Mallia, D. V, Lin, J. C., Urbanski, S., Ehleringer, J. and Nehrkorn, T.: Impacts of upwind wildfire emissions on CO, CO_2, and $PM_{2.5}$ concentrations in Salt Lake City, Utah, J. Geophys. Res. Atmos., 120(1), 147–166, 2015.

Mallia, D. V., Kochanski, A., Wu, D., Pennell, C., Oswald, W. and Lin, J. C.: Wind-blown dust modeling using a backward-Lagrangian particle dispersion model, J. Appl. Meteorol. Climatol., 56(10), 2845–2867, doi:10.1175/JAMC-D-16-0351.1, 2017.

Mandrake, L., Frankenberg, C., O'Dell, C. W., Osterman, G., Wennberg, P. and Wunch, D.: Semi-autonomous sounding selection for OCO-2, Atmos. Meas. Tech., 6(10), 2851–2864, 2013.

Marland, G.: Uncertainties in accounting for CO_2 from fossil fuels, J. Ind. Ecol., 12(2), 136–139, doi:10.1111/j.1530-9290.2008.00014.x, 2008.

Mitchell, L., Lin, J. C., Bowling, D. R., Pataki, D. E., Strong, C., Schauer, A. J., Bares, R., Bush, S., Stephens, B. B., Mendoza, D., Mallia, D. V, Holland, L., Gurney, K. R. and Ehleringer, J. R.: Long-term urban carbon dioxide observations reveal spatial and temporal dynamics related to urban characteristics and growth, Proc. Natl. Acad. Sci., 115(12), 2912–2917, doi:10.1073/pnas.1702393115, 2018.

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(2), 917–933, doi:10.1029/2012JD018196, 2013.

Nassar, R., T. G. Hill, C. A. McLinden, D. Wunch, D. B. A. Jones, and D. Crisp (2017), Quantifying CO_2 emissions from individual power plants from space, Geophys. Res. Lett., doi:10.1002/2017GL074702.

OCO-2 Science Team/Michael Gunson, Annmarie Eldering, OCO-2 Level 2 bias-corrected solar-induced fluorescence and other select fields from the IMAP-DOAS algorithm aggregated as daily files, Retrospective processing V7r, Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), https://disc.gsfc.nasa.gov/datacollection/OCO2_L2_Lite_SIF_7r.html, 2015.

O'Dell, C. W., Connor, B., Bösch, H., O'Brien, D., Frankenberg, C., Castano, R., Christi, M., Eldering, D., Fisher, B., Gunson, M., McDuffie, J., Miller, C. E., Natraj, V., Oyafuso, F., Polonsky, I., Smyth, M., Taylor, T., Toon, G. C., Wennberg, P. O. and

Wunch, D.: The ACOS CO_2 retrieval algorithm-Part 1: Description and validation against synthetic observations, Atmos. Meas. Tech., 5(1), 99–121, doi:10.5194/amt-5-99-2012, 2012.

Oda, T. and Maksyutov, S.: A very high-resolution (1km × 1 km) global fossil fuel CO_2 emission inventory derived using a point source database and satellite observations of nighttime lights, Atmos. Chem. Phys., 11(2), 543–556, doi:10.5194/acp-11-543-2011, 2011.

Oda, T., Ott, L., Topylko, P., Halushchak, M., Bun, R., Lesiv, M., Danylo, O. and Horabik-Pyzel, J.: Uncertainty associated with fossil fuel carbon dioxide (CO_2) gridded emission datasets, 2015.

Oda, T. and Maksyutov, S.: ODIAC Fossil Fuel CO_2 Emissions Dataset (Version name: ODIAC2017), Center for Global Environmental Research, National Institute for Environmental Studies, doi:10.17595/20170411.001, 2015.

Oda, T., Maksyutov, S. and Andres, R. J.: The Open-source Data Inventory for Anthropogenic CO_2, version 2016 (ODIAC2016): a global monthly fossil fuel CO_2 gridded emissions data product for tracer transport simulations and surface flux inversions, , 2, 87–107, 2018.

Olsen, S. C. and Randerson, J. T.: Differences between surface and column atmospheric CO_2 and implications for carbon cycle research, J. Geophys. Res., 109(D2), D02301, doi:10.1029/2003JD003968, 2004.

Pacala, S. W., Breidenich, C., Brewer, P. G., Fung, I., Gunson, M. R., Heddle, G., Marland, G., Paustian, K., Prather, M., Randerson, J. T., Tans, P., and Wofsy, S. C.: Verifying Greenhouse Gas Emissions: Methods to Support International Climate Agreements, Tech. rep., Committee on Methods for Estimating Greenhouse Gas Emissions, Washington, DC, 2010.

Palmer, P. I.: Quantifying sources and sinks of trace gases using space-borne measurements: current and future science, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., 366(1885), 4509–4528, doi:10.1098/rsta.2008.0176, 2008.

Patra, P. K., Crisp, D., Kaiser, J. W., Wunch, D., Saeki, T., Ichii, K., Sekiya, T., Wennberg, P. O., Feist, D. G., Pollard, D. F., Griffith, D. W. T., Velazco, V. A., De Maziere, M., Sha, M. K., Roehl, C., Chatterjee, A. and Ishijima, K.: The Orbiting Carbon Observatory (OCO-2) tracks 2–3 peta-gram increase in carbon release to the atmosphere during the 2014–2016 El Niño, Sci. Rep., 7(1), 13567, doi:10.1038/s41598-017-13459-0, 2017.

Peters, W., Jacobson, A. R., Sweeney, C., Andrews, A. E., Conway, T. J., Masarie, K., Miller, J. B., Bruhwiler, L. M. P., Petron, G., Hirsch, A. I., Worthy, D. E. J., van der Werf, G. R., Randerson, J. T., Wennberg, P. O., Krol, M. C. and Tans, P. P.: An atmospheric perspective on North American carbon dioxide exchange: CarbonTracker, Proc. Natl. Acad. Sci., 104(48), 18925–18930, doi:10.1073/pnas.0708986104, 2007.

Peylin, P., Houweling, S., Krol, M. C., Karstens, U., Rödenbeck, C., Geels, C., Vermeulen, A., Badawy, B., Aulagnier, C., Pregger, T., Delage, F., Pieterse, G., Ciais, P. and Heimann, M.: Importance of fossil fuel emission uncertainties over Europe for CO_2 modeling: Model intercomparison, Atmos. Chem. Phys., 11(13), 6607–6622, doi:10.5194/acp-11-6607-2011, 2011.

Rayner, P. J. and O'Brien, D. M.: The utility of remotely sensed CO_2 concentration data in surface source inversions, Geophys. Res. Lett., 28(1), 175–178, doi:10.1029/2000GL011912, 2001.

Rayner, P. J., Raupach, M. R., Paget, M., Peylin, P. and Koffi, E.: A new global gridded data set of CO_2 emissions from fossil fuel combustion: Methodology and evaluation, J. Geophys. Res., 115(D19), D19306, doi:10.1029/2009JD013439, 2010.

Reuter, M., Buchwitz, M., Hilker, M., Heymann, J., Schneising, O., Pillai, D., Bovensmann, H., Burrows, J. P., Bösch, H., Parker, R., Butz, A., Hasekamp, O., O'Dell, C. W., Yoshida, Y., Gerbig, C., Nehrkorn, T., Deutscher, N. M., Warneke, T., Notholt, J.,

Hase, F., Kivi, R., Sussmann, R., Machida, T., Matsueda, H. and Sawa, Y.: Satellite-inferred European carbon sink larger than expected, Atmos. Chem. Phys., 14(24), 13739–13753, doi:10.5194/acp-14-13739-2014, 2014.

Rodgers, C.D.: Inverse methods for atmospheric sounding: theory and practice. World scientific, 2000.

Rolph, G., Stein, A. and Stunder, B.: Real-time Environmental Applications and Display sYstem: READY, Environ. Model. Softw., 95, 210–228, doi:10.1016/j.envsoft.2017.06.025, 2017.

Rosenzweig, C., Solecki, W., Hammer, S. A. and Mehrotra, S.: Cities lead the way in climate-change action, Nature, 467(7318), 909–911, doi:10.1038/467909a, 2010.

Schneising, O., Heymann, J., Buchwitz, M., Reuter, M., Bovensmann, H. and Burrows, J. P.: Anthropogenic carbon dioxide source areas observed from space: Assessment of regional enhancements and trends, Atmos. Chem. Phys., 13(5), 2445–2454, doi:10.5194/acp-13-2445-2013, 2013.

Seibert, P. and Frank, A.: Source-receptor matrix calculation with a Lagrangian particle dispersion model in backward mode, Atmos. Chem. Phys., 4(1), 51–63, doi:10.5194/acp-4-51-2004, 2004.

Shiga, Y. P., Michalak, A. M., Gourdji, S. M., Mueller, K. L. and Yadav, V.: Detecting fossil fuel emissions patterns from subcontinental regions using North American in situ CO_2 measurements, Geophys. Res. Lett., 41(12), 4381–4388, doi:10.1002/2014GL059684, 2014.

Shiga, Y. P., Tadić, J. M., Qiu, X., Yadav, V., Andrews, A. E., Berry, J. A. and Michalak, A. M.: Atmospheric CO_2 Observations Reveal Strong Correlation Between Regional Net Biospheric Carbon Uptake and Solar-Induced Chlorophyll Fluorescence, Geophys. Res. Lett., 1122–1132, doi:10.1002/2017GL076630, 2018.

Silva, S. and Arellano, A.: Characterizing Regional-Scale Combustion Using Satellite Retrievals of CO, NO_2 and CO_2, Remote Sens., 9(7), 744, doi:10.3390/rs9070744, 2017.

Silva, S. J., Arellano, A. F. and Worden, H. M.: Toward anthropogenic combustion emission constraints from space-based analysis of urban CO_2/CO sensitivity, Geophys. Res. Lett., 40(18), 4971–4976, doi:10.1002/grl.50954, 2013.

Skamarock, W. C. and Klemp, J. B.: A time-split nonhydrostatic atmospheric model for weather research and forecasting applications, J. Comput. Phys., 227(7), 3465–3485, 2008.

Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J. B., Cohen, M. D. and Ngan, F.: Noaa's hysplit atmospheric transport and dispersion modeling system, Bull. Am. Meteorol. Soc., 96(12), 2059–2077, doi:10.1175/BAMS-D-14-00110.1, 2015.

Stephens, B.B., Gurney, K.R., Tans, P.P., Sweeney, C., Peters, W., Bruhwiler, L., Ciais, P., Ramonet, M., Bousquet, P., Nakazawa, T. and Aoki, S., Machida, T., Inoue, G., Vinnichenko, N., Lloyd, J., Jordan, A., Heimann, M., Shibistova, O., Langenfelds, R.L., Steele, L.P., Francey, R.J., Denning, A.S.: Weak Northern and Strong Tropical Land Carbon Uptake from Vertical Profiles of Atmospheric CO_2, Science, 316(5832), 1732–1736, doi:10.1126/science.1137004, 2007.

Stohl, A., Forster, C., Frank, A., Seibert, P. and Wotawa, G.: Technical note: The Lagrangian particle dispersion model FLEXPART version 6.2, Atmos. Chem. Phys., 5(9), 2461–2474, doi:10.5194/acp-5-2461-2005, 2005.

Sun, Y., Frankenberg, C., Wood, J.D., Schimel, D.S., Jung, M., Guanter, L., Drewry, D.T., Verma, M., Porcar-Castell, A., Griffis, T.J. and Gu, L., Magney, T. S., Kohler, P., Evans, B. and Yuen, K.: OCO-2 advances photosynthesis observation from space via solar-induced chlorophyll fluorescence. Science, 358(6360), eaam5747, 2017.

Sweeney, C., Karion, A., Wolter, S, Newberger, T, Guenther, D, Higgs, J.A., Andrews, A.E., Lang, P.M., Neff, D., Dlugokencky, E., Miller, J.B.: Seasonal climatology of CO_2 across North America from aircraft measurements in the NOAA/ESRL Global Greenhouse Gas Reference Network. Journal of Geophysical Research: Atmospheres, 120(10), 5155-5190, 2015.

Tollefson, J.: Carbon-sensing satellite system faces high hurdles, Nature, 533, 446–447, 2016.

UNFCCC, 2017. National Inventory Submissions 2017. United Nations Framework Convention on Climate Change. Available at: http://unfccc.int/national_reports/annex_i_ghg_inventories/national_inventories_submissions/items/9492.php; accessed June 2017.

Venables, W. N. and Ripley, B.D.: Random and mixed effects. In Modern applied statistics with S, Springer, New York, NY, 271-300, 2002.

Verhulst, K. R., Karion, A., Kim, J., Salameh, P. K., Keeling, R. F., Newman, S., Miller, J., Sloop, C., Pongetti, T., Rao, P., Wong, C., Hopkins, F. M., Yadav, V., Weiss, R. F., Duren, R. and Miller, C. E.: Carbon Dioxide and Methane Measurements from the Los Angeles Megacity Carbon Project: 1. Calibration, Urban Enhancements, and Uncertainty Estimates, Atmos. Chem. Phys., 17, 8313-8341, doi:10.5194/acp-17-8313-2017, 2017.

Wheeler, D. and Ummel, K.: Calculating CARMA: Global Estimation of CO_2 Emissions from the Power Sector, Work. Pap., (145), 37, 2008.

Wong, K. W., Fu, D., Pongetti, T. J., Newman, S., Kort, E. A., Duren, R., Hsu, Y. K., Miller, C. E., Yung, Y. L. and Sander, S. P.: Mapping CH_4 : CO_2 ratios in Los Angeles with CLARS-FTS from Mount Wilson, California, Atmos. Chem. Phys., 15(1), 241–252, doi:10.5194/acp-15-241-2015, 2015.

Worden, J., Doran, G., Kulawik, S., Eldering, A., Crisp, D., Frankenberg, C., O'Dell, C. and Bowman, K.: Evaluation And Attribution Of OCO-2 XCO_2 Uncertainties, Atmos. Meas. Tech., 10, 2759-2771, doi:10.5194/amt-10-2759-2017, 2017.

Wunch, D., Wennberg, P. O., Toon, G. C., Keppel-Aleks, G. and Yavin, Y. G.: Emissions of greenhouse gases from a North American megacity, Geophys. Res. Lett., 36(15), 1–5, doi:10.1029/2009GL039825, 2009.

Wunch, D., Wennberg, P. O., Toon, G. C., Connor, B. J., Fisher, B., Osterman, G. B., Frankenberg, C., Mandrake, L., O'Dell, C., Ahonen, P., Biraud, S. C., Castano, R., Cressie, N., Crisp, D., Deutscher, N. M., Eldering, A., Fisher, M. L., Griffith, D. W. T., Gunson, M., Heikkinen, P., Keppel-Aleks, G., Kyrö, E., Lindenmaier, R., MacAtangay, R., Mendonca, J., Messerschmidt, J., Miller, C. E., Morino, I., Notholt, J., Oyafuso, F. A., Rettinger, M., Robinson, J., Roehl, C. M., Salawitch, R. J., Sherlock, V., Strong, K., Sussmann, R., Tanaka, T., Thompson, D. R., Uchino, O., Warneke, T. and Wofsy, S. C.: A method for evaluating bias in global measurements of CO_2 total columns from space, Atmos. Chem. Phys., 11(23), 12317–12337, doi:10.5194/acp-11-12317-2011, 2011.

World Urbanization Prospects: The 2014 Revision (WUP 2014), United Nations, Department of Economic and Social Affairs, Population Division (2014)., CD-ROM Edition.

Ye, X., Lauvaux, T., Kort, E. A., Oda, T., Feng, S., Lin, J. C., Yang, E., and Wu, D.: Constraining fossil fuel CO_2 emissions from urban area using OCO-2 observations of total column CO_2, Atmos. Chem. Phys. Discuss., in review, doi: 10.5194/acp-2017-1022, 2017.

Yokota, T., Yoshida, Y., Eguchi, N., Ota, Y., Tanaka, T., Watanabe, H. and Maksyutov, S.: Global Concentrations of CO_2 and CH_4 Retrieved from GOSAT: First Preliminary Results, Sola, 5, 160–163, doi:10.2151/sola.2009-041, 2009.

Zhao, C., Andrews, A. E., Bianco, L., Eluszkiewicz, J., Hirsch, A., MacDonald, C., Nehrkorn, T. and Fischer, M. L.: Atmospheric inverse estimates of methane emissions from Central California, J. Geophys. Res. Atmos., 114(16), 1–13, doi:10.1029/2008JD011671, 2009.

Zuromski, L.M., Bowling, D.R., Köhler, P., Frankenberg, C., Goulden, M.L., Blanken, P.D. and Lin, J.C.: Solar-Induced Fluorescence Detects Interannual Variation in Gross Primary Production of Coniferous Forests in the Western United States. Geophysical Research Letters. 45(14), 7184-7193, doi:10.1029/2018GL077906, 2018

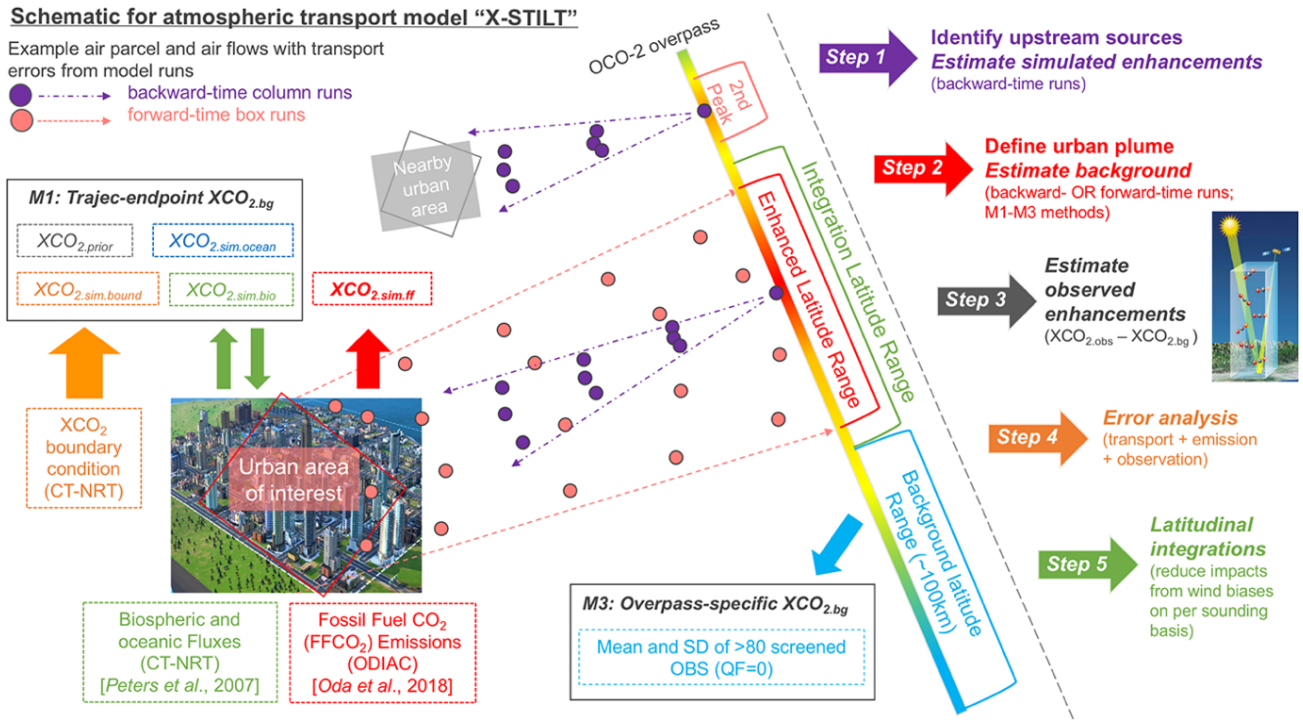

Figure 1. A schematic of X-STILT in 5 steps (with arrows on the right). Pink and purple dots and arrows represent the air parcels and overall air flows based on forward-time box runs and backward-time column runs with wind error component accounted for. Rainbow band is an example of one OCO-2 overpass with warmer color indicating higher observed XCO_2. M1 include modeled-derived biospheric, oceanic XCO_2 changes, CO_2 boundary conditions, and prior CO_2 portion from OCO-2. M3 requires enhanced latitude range based on either backward-time XCO_2 enhancements or forward-time urban plume.

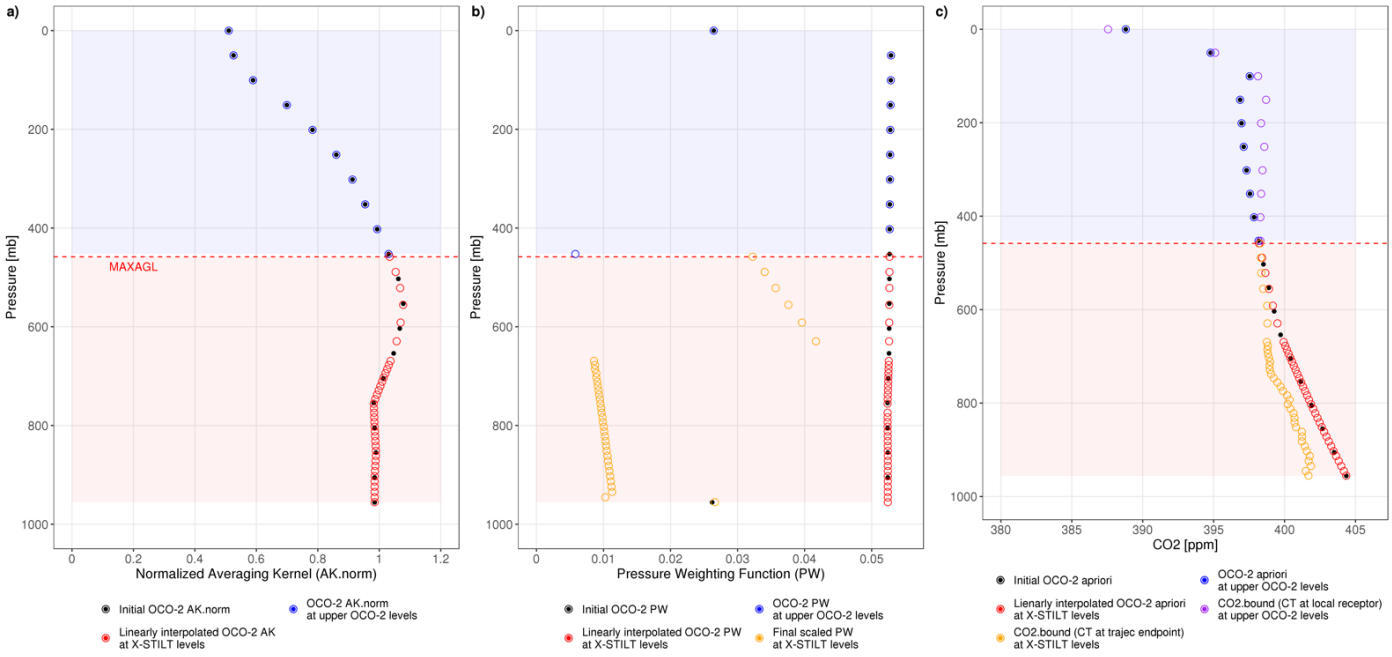

Figure 2. Demonstrations of interpolations on **a)** normalized averaging kernel profile, **b)** pressure weighting function and **c)** CO_2 boundary conditions (derived from CT-NRT) and OCO-2 a priori profile, given one sounding (lat/lon same as column receptors). Red and blue shadings denote the X-STILT release levels from the surface up to MAXAGL and upper OCO-2 levels.

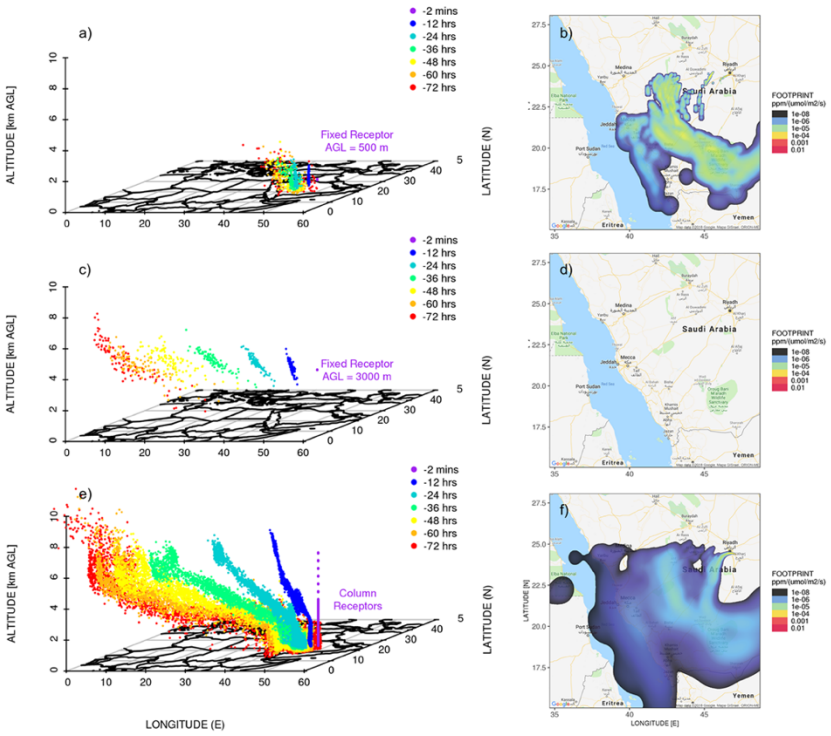

Figure 3. Left panels (**a, c, e**): 3D scatter plot of STILT ensembles that are initially released from a fixed receptor of 500 m, 3 km and column receptors for Riyadh on 10UTC 12/29/2014. Colors differentiate hours backwards (-2 mins, -12, -24, -36, -48, -60, and -72 hours) for each trajectory. Column receptors (e) are placed every 100 m within 3 km and every 500 m from 3–6 km. Right panels (**b, d, f**): Modeled fixed footprints vs. column footprints are plotted in blue to red gradient. Column footprints are weighted by pressure weighting functions. Only footprints values >1E-8 ppm/(μmol m^{-2} s^{-1}) are displayed.

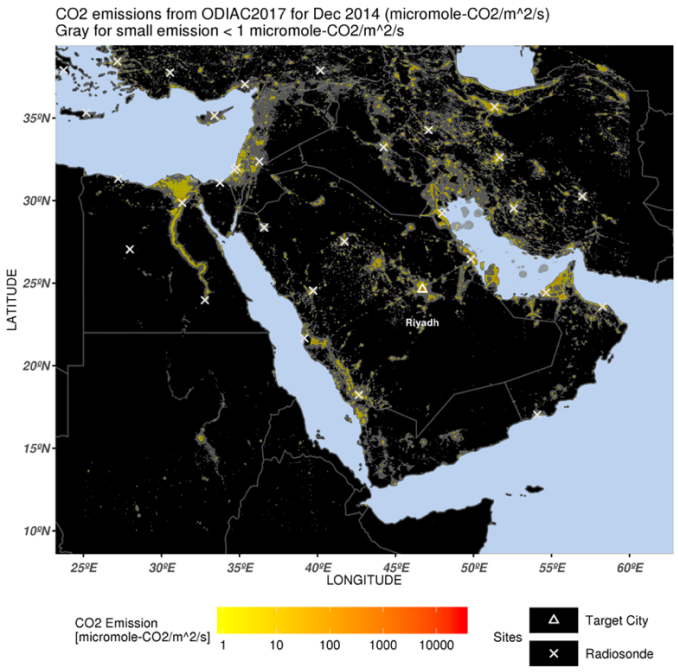

Figure 4. Monthly ODIAC emissions (yellow to orange) in log-scale at 1 km×1 km grid spacing for Dec 2014. White crosses and triangle denote the radiosonde networks used to evaluate provide wind error statistics and our study site of Riyadh. Small emissions (< 1μmole m^{-2} s^{-1}) are shaded in gray.

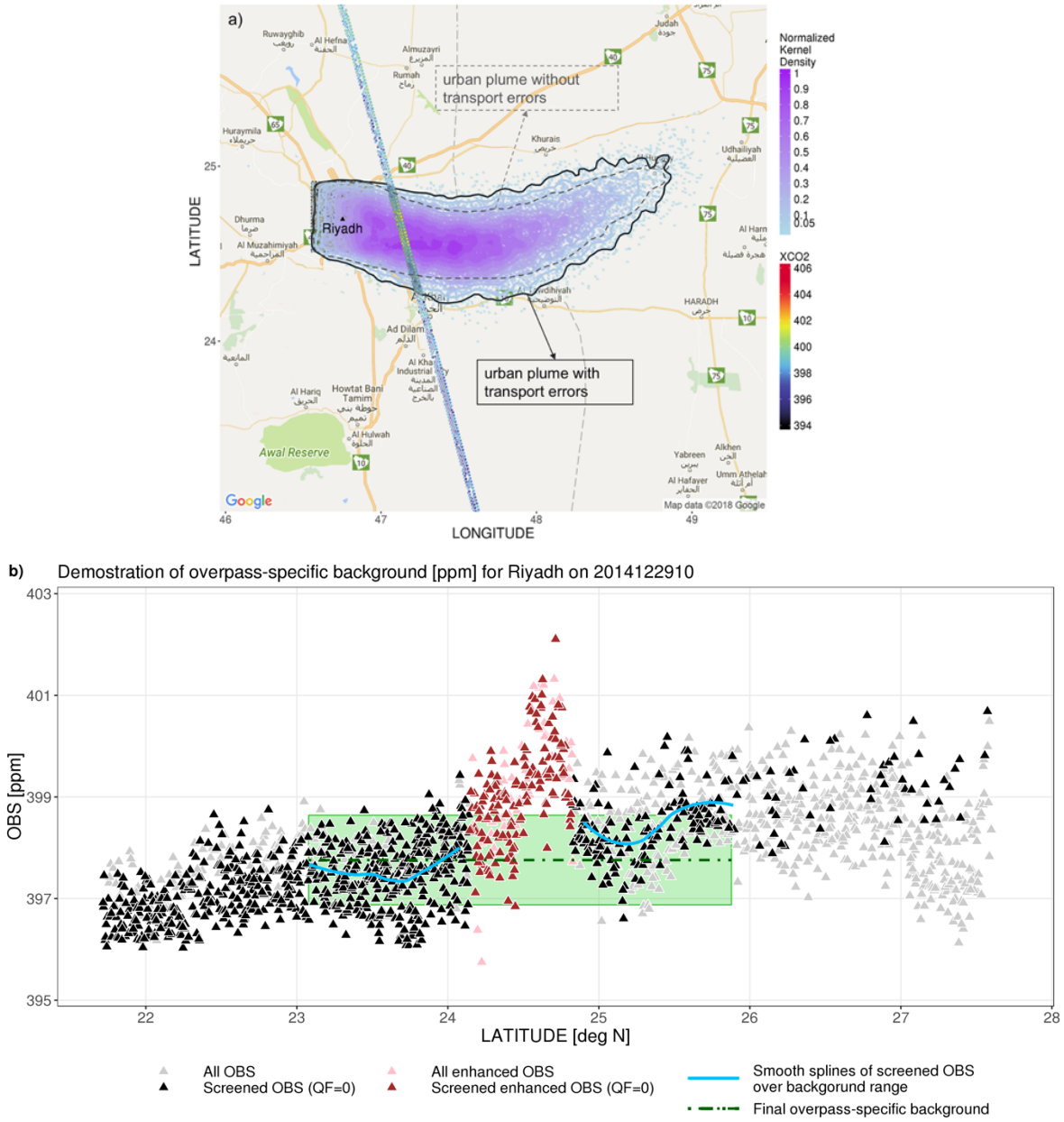

Figure 5. Demonstrations of overpass-specific background with an example of 12/29/2014 overpass for Riyadh. **a)** Forward particle distributions with random transport error included (blue and purple dots) and their derived normalized kernel density (solid purple contours) during OCO-2 overpass time (~3 mins) with observed XCO_2 (blue to red dots). Urban plumes defined based on 5 % of the max 2D kernel density estimated from parcels' distributions without (grey dashed line) and with (black solid line) transport errors. **b)** Latitude-series of observed XCO_2 with demonstration of background estimates. Smooth splines (solid blue lines) are drawn to visually reveal the variation of observed XCO_2 over background latitudinal band. Background uncertainty (green ribbon) includes both spatial uncertainty and retrieval uncertainty of observations over the background latitude range.

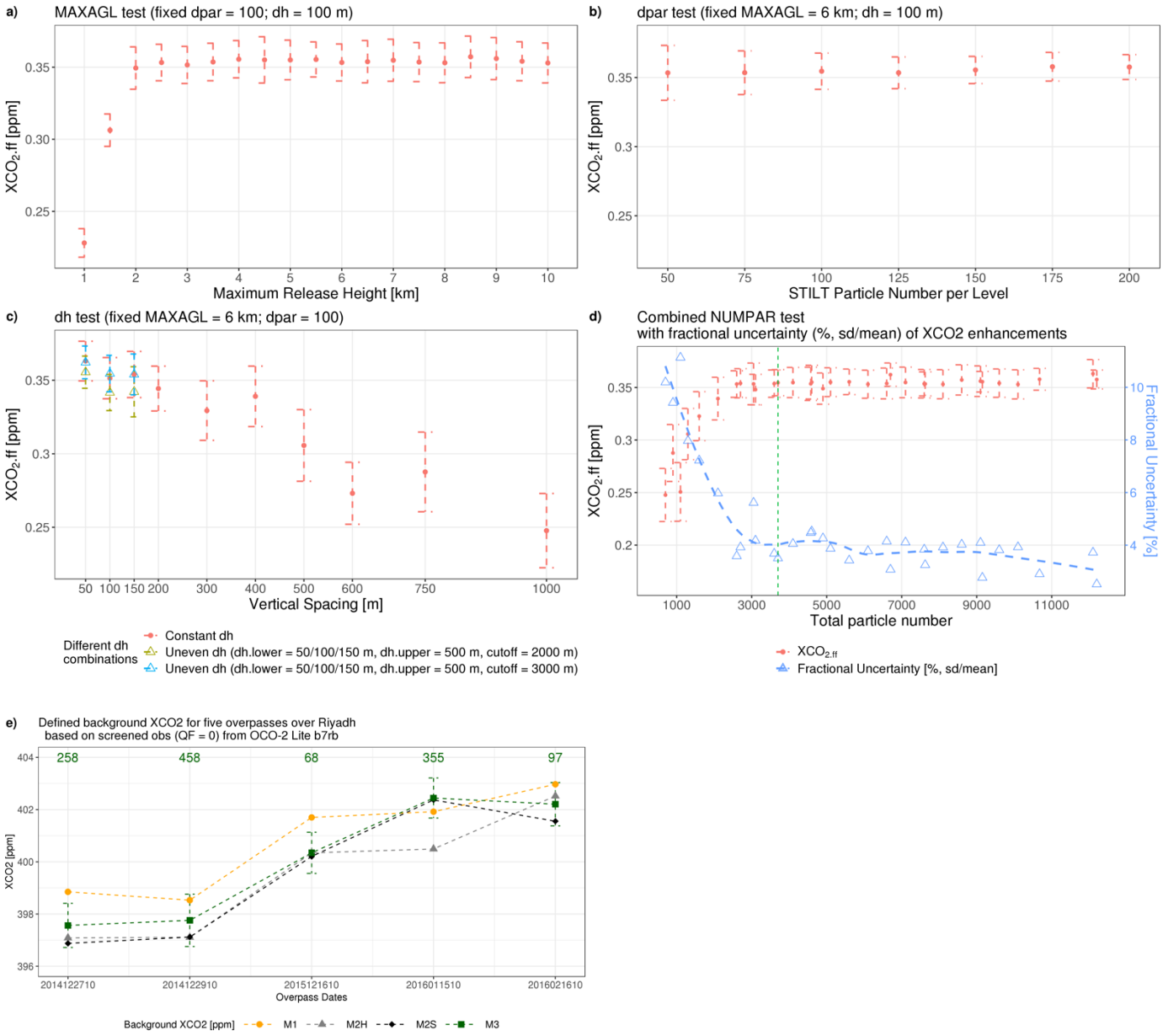

Figure 6. Results of sensitivity tests (**a-d**) shown for the one sounding with the largest retrieved XCO_2 for Riyadh and background comparisons for 5 tracks (**e**). Random error for each simulation is indicated as dashed red error bars (**a-d**); and potential biases are shown as the trend of the mean XCO_2 enhancements (red dots; **a-d**) derived from 100 times of bootstrap. **c)** For vertical spacing test, besides tests with constant *dh* (red dots and error bars), two other cases with uneven *dh* above and below a cutoff level are carried out. Case 1) tested 3 different lower *dh* (50, 100, and 150 m) with a fixed upper *dh* = 500 m and a cutoff level of 2000 km (yellow triangles and dashed error bars); Case 2 used the same upper and lower spacings as Case 1) but with a different cutoff level of 3000 km (blue triangles and dashed error bars). **d)** A summary plot of mean and SD of XCO_2 enhancements (red dots and dashed red error bars) and fractional uncertainties (%, blue triangles and dashed line) as functions of total particle number (NUMPAR). Green dashed vertical line denotes the configuration used in this study. **e)** Background comparisons using different methods (M1, M2H, M2S, and M3) for 5 tracks. The amounts of screened observations used for M3 background are labeled in dark green. M3 background errors (including spatial variation and retrieval errors over the background region) are indicated as dashed green error bars.

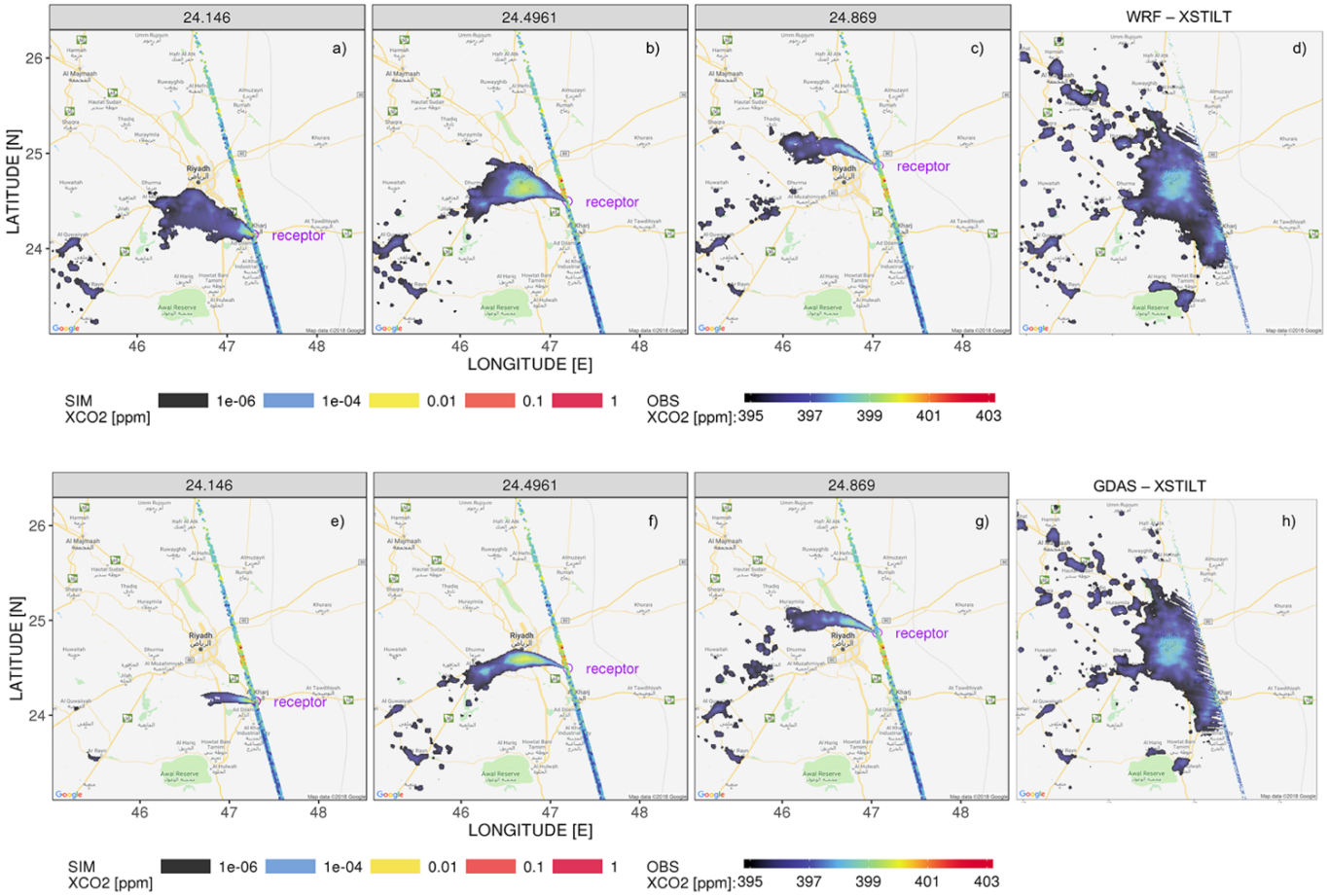

Figure 7. Spatial maps of 1 km × 1 km modeled XCO_2 contributions (ppm; log-scale) from 3 selected soundings along with screened observations (QF = 0) on 1000UTC 12/29/2014 over Riyadh, with meteorological fields driven by WRF (a-d) and GDAS (e-h). Panel d and h denote the latitude-integrated $XCO_{2,ff}$ contributions (with weights of receptor spacings, e.g., 0.02°) using WRF and GDAS, derived from spatial XCO_2 enhancements for over 60 column receptors along each overpass. The sum of the latitude-weighted spatial XCO_2 enhancements over all gridcells (in panel d or h) equals to the latitude-integrated $XCO_{2,ff}$ signal (ppm-degree) reported in Sect. 3.5. Only large enhancements > 10^{-6} ppm are plotted.

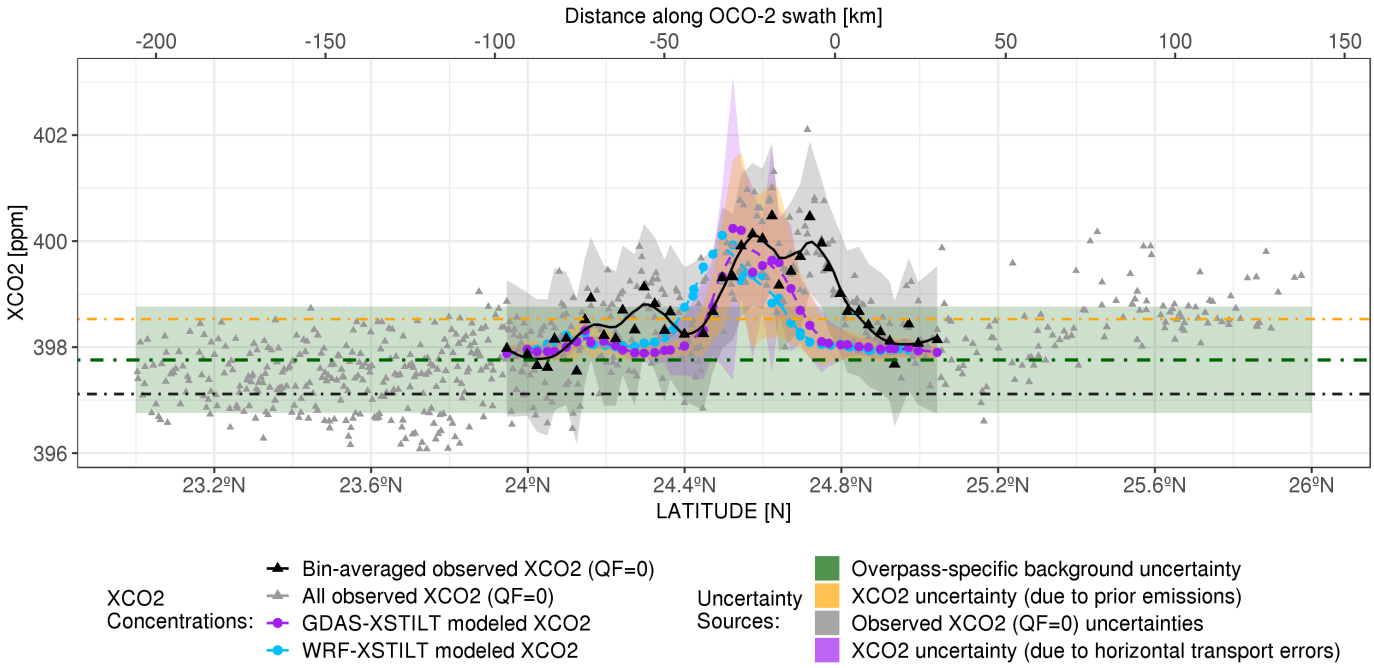

Figure 8. Latitude-series of sounding-level signal comparisons and error estimates for Riyadh. Screened observations with QF=0 and bin-averaged observed XCO_2 are shown in grey and black triangles. GDAS- and WRF- derived XCO_2 are displayed in purple and light blue dots, with smooth splines applied to visually reveal the main variations (purple and blue dashed lines). XCO_2 errors due to errors in emissions, transport and observation are drawn as yellow, purple, and light grey ribbons. Overpass-specific (M3) background XCO_2 is drawn as dark green dotted dashed line with its background uncertainty in light green ribbon. Background values using M1, M2H and M2S are drawn as orange, gray and black dotted dashed lines, respectively. The latitudinal range for integrating XCO_2 enhancements and associated various uncertainties is ~24° N–25° N in this case. The top x-axis is the distance (in km) along the OCO-2 swath from a "minimum distance sounding" that has the smallest distance from the city center.

Overpass Dates	*Lat-integrated Observed XCO_2 signal [ppm-deg.]*				*Lat-int. Sim. XCO_2 signal [ppm-deg.]*	*GDAS u, v-wind RMSE [m/s]*	*WRF u, v-wind RMSE [m/s]*	*Lat-integrated Modeled XCO_2 Errors [ppm-deg.]*				*Lat-integrated Observed XCO_2 Errors [ppm-deg.]*			
	M1	*M2H*	*M2S*	*M3*				*Emiss*	*U, V-*	*PBL*	*Tot*	*Bg.*	*Bin*	*Retrieval*	*Tot*
20141227	0.25	2.41	2.74	1.62	1.76	1.94 (2.06)	2.15 (1.85)	0.73	1.22	0.23	1.44	0.24	0.14	0.32	0.42
20141229	0.47	1.75	1.74	1.09	0.64	1.81 (2.23)	1.75 (2.03)	0.34	0.41	0.06	0.54	0.18	0.11	0.28	0.35
20151216	1.01	2.92	3.20	2.92	3.04	1.74 (2.03)		1.04	1.83	0.36	2.14	0.26	0.21	0.39	0.51
20160115	2.04	3.63	1.54	1.47	1.06	1.81 (1.77)		0.56	0.60	0.17	0.84	0.15	0.10	0.27	0.33
20160216	0.65	0.77	2.11	1.17	1.37	1.78 (1.90)		0.70	1.05	0.16	1.27	0.24	0.15	0.33	0.43
Mean signal *or* ***SDOM*** [ppm-deg.]	0.88	2.30	2.27	1.65	1.57			***0.32*** *(20 % of sim)*	***0.52*** *(hor. & ver.; 33% of sim)*		***0.61*** *(39 % of sim)*				***0.19*** *(11 % of obs)*
$\hat{\lambda}$ [unitless]	0.75 (M1 obs vs. sim); 1.78 (M2H obs vs. sim); 1.52 (M2S obs vs. sim); 1.14 (M3 obs vs. sim)														

Table 1. Results of signal and errors calculations for the examined five overpasses, including latitude-integrated observed versus modeled XCO_2 enhancements and errors (ppm-degree), regional wind RMSE (m/s), standard deviation of mean (SDOM) for various errors and posterior scaling factors (unitless) of the mean modeled XCO_2 signal. The GDAS (purple) and WRF (light blue) regional wind RMSEs from 0-3 km or 3-6 km are shown within or outside the bracket.

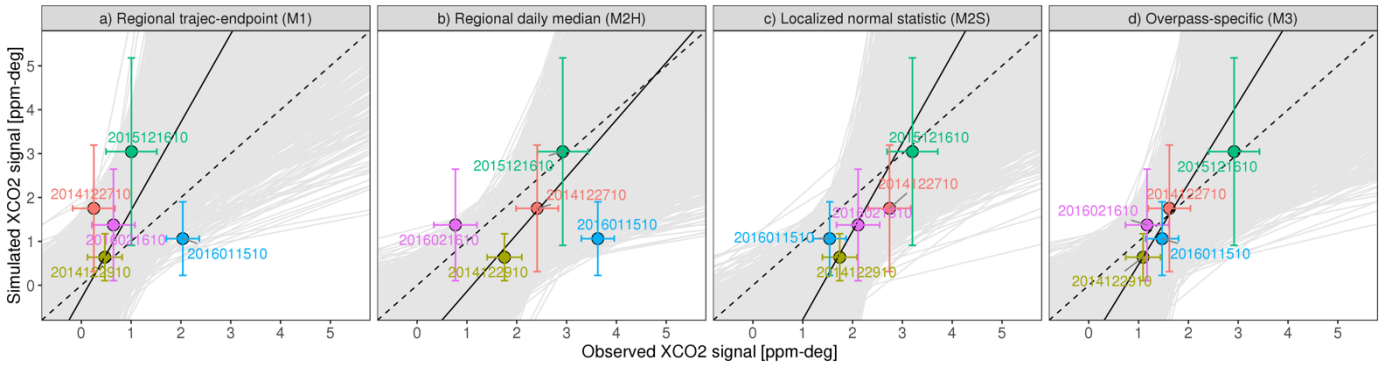

Figure 9. Correlation between observed and simulated anthropogenic XCO_2 signals for 5 overpasses. Colors differentiate different satellite overpass dates. Model-data comparisons using GDAS-derived XCO_2 signals and observed signals based on different background methods. Error bars along x-axis and y-axis represent the overall observed uncertainty (represented as 1-σ, including XCO_2 spatial variability, background uncertainty and retrieval errors) associated with observed signals and the overall modeled uncertainty (σ, including emission uncertainty and transport uncertainty) around modeled signals. Dashed line represents the 1:1 line. Monte Carlo experiments are performed to fit linear regression lines based on sampled model-data signals and associated errors. Regression lines with positive slopes are shaded in light grey. Median values of slopes and y-intercepts from those multiple regression lines (with positive slopes) are used to draw a linear regression (black solid line).