Supporting Information for "New York City greenhouse gas emissions estimated with inverse modelling of aircraft measurements"

Joseph R. Pitt[*,1,§], *Israel Lopez-Coto*[2,1], *Kristian D. Hajny*[1,3], *Jay Tomlin*[3], *Robert Kaeser*[3], *Thilina Jayarathne*[3,†], *Brian H. Stirm*[4], *Cody R. Floerchinger*[5], *Christopher P. Loughner*[6], *Conor K. Gately*[5,7,‡], *Lucy R. Hutyra*[7], *Kevin R. Gurney*[8], *Geoffrey S. Roest*[8], *Jianming Liang*[9], *Sharon Gourdji*[2], *Anna Karion*[2], *James R. Whetstone*[2], *Paul B. Shepson*[*,1,3]

[1]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY 11794, USA

[2]National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

[3]Department of Chemistry, Purdue University, West Lafayette, IN 47907, USA

[4]School of Aviation and Transportation Technology, Purdue University, West Lafayette, IN 47906, USA

[5]Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA 02138, USA

[6]Air Resources Laboratory, NOAA, College Park, MD 20740, USA

[7]Department of Earth and Environment, Boston University, Boston, MA 02215, USA

[8]School of Informatics, Computing and Cyber Systems, Northern Arizona University, Flagstaff, AZ 86011, USA

[9]Environmental Systems Research Institute, Redlands, CA 92373, USA

[§]Current address: School of Chemistry, University of Bristol, Bristol, BS8 1TS, UK

[†]Current address: Bristol Myers Squibb, New Brunswick, NJ 08901, USA

[‡]Current address: Metropolitan Area Planning Council, Boston, MA 02111, USA

*Corresponding authors: joseph.pitt@bristol.ac.uk; paul.shepson@stonybrook.edu

- Number of pages: 25
- Number of figures: 11
- Number of tables: 2

S1: Calibration gas

Details of the calibration gases used during the flights are given in Table S1.1. A recent (October 2020) NOAA reanalysis of these cylinders derived mole fractions consistent with the initial analysis, within World Meteorological Organization guidelines (0.1 ppm for CO_2, 2 ppb for CH_4), with the exception of FB03961. For this cylinder, a difference of 0.17 ppm was found for CO_2, which is nonetheless small in comparison to other uncertainties (especially model uncertainties) associated with this study.

Table S1.1: Details of calibration cylinders used during the flights.

Cylinder number	**Date installed**	**Date removed**	**Certified mole fraction**		**Certification date**
			CO_2 (ppm)	**CH_4 (ppb)**	
FB03872	2019-08-04	2020-05-03	396.95	2543.10	2013-10-24
FB03917	2016-11-08	2020-03-16	374.89	1770.42	2014-12-03
FB03949	2017-09-06	2020-03-16	433.80	3336.08	2014-12-03
FB03961	2016-03-23	2019-08-04	399.52	2581.20	2014-12-03

S2: HYSPLIT runs

To perform our dispersion model runs, we modified wrapper scripts from the AER-NOAA branch of STILT (available from http://stilt-model.org) for use with HYSPLIT v5.0.0. These wrappers were modified so that the particle release for each minute was distributed across all aircraft sample locations within the boundary layer during that minute. Example SETUP.CFG and CONTROL files generated by these wrappers are included below. We did not apply a near-field correction of either the form described by Fasoli et al. (2018) or the form described by Sargent et al. (2018) as it is not clear that such corrections are applicable to aircraft data.

We applied a far-field correction to account for uncertainty in single-particle locations when the particle density is low. The footprint was smoothed using a Gaussian weighted spatial kernel, similar to that described by Fasoli et al. (2018), although in our case the bandwith was determined solely by the spread of the particle ensemble in latitude and longitude for a given time period, with an upper limit on the bandwidth equal to twice the resolution of the footprint grid.

We also applied a correction for the impact of mass violation in the driving wind fields (by setting option dmassTF=T in the AER-NOAA wrappers), as described by Lin et al. (2003). Sensitivity tests without this correction showed that it had little impact in the majority of cases, but for certain flights and model ensemble members non-negligible changes in modelled mole fraction were observed, with those using the correction being more similar to the observations. Further investigation into the impacts of mass violation on transport model error is a topic for future work, but based on our sensitivity tests we do not expect that the average posterior emission rates reported here were significantly impacted by the application of this correction.

Each HYSPLIT run was repeated using two turbulence parameterisations: Kantha and Clayson (2000) and Hanna (1982). Input meteorology was taken from four freely available archived products: the European Centre for Medium Range Weather Forecasts Fifth Reanalysis (ERA5), the NOAA Global Forecast System (GFS) model, the NOAA North American Mesoscale Forecast System (NAM) model, and the NOAA High-Resolution Rapid Refresh (HRRR) model. Archived data for GFS, NAM and HRRR, at resolutions of 0.25°, 12 km and 3 km and on 55, 40 and 36 vertical levels respectively, were retrieved from the NOAA archive (https://www.ready.noaa.gov/archives.php), which provides model data in the ARL format required by HYSPLIT. Data for the ERA5 reanalysis product (at 0.25° resolution and on 37 vertical pressure levels) were retrieved from the Copernicus Climate Change Service Climate Data Store (https://cds.climate.copernicus.eu) and converted to ARL format using the HYSPLIT utility era52arl. The ERA5, HRRR and NAM data had a temporal resolution of 1 hour and the GFS data had a temporal resolution of 3 hours. The four different choices of input meteorology and the two different turbulence parameterisations were combined to give us an eight-member transport model ensemble.

Each release consisted of approximately 1000 particles, distributed equally across all aircraft sample locations within the boundary layer for a given minute of the flight. These particles were tracked for 48 hours backwards in time, with the particle location recorded every 1 min (i.e., at each HYSPLIT integration timestep). The effective mixing height used to account for the footprint was set to half the modelled planetary boundary layer height (Gerbig et al., 2003). Every particle was assigned a single-particle footprint value at each recorded particle location, equal to the time the particle spent below the effective mixing height within the previous timestep, divided by the integrated molar air density up to the mixing height (see Lin et al., 2003, for more details). The gridded time-integrated footprints used to calculate the modelled enhancements were obtained by summing the single-particle footprints within each grid cell (of both the d03 and d01 domains) over all 2880 model timesteps and all particles.

Example SETUP.CFG file:

```
&SETUP
CMASS=0,
CONAGE=48,
CPACK=1,
DELT=1,
DXF=1,
DYF=1,
DZF=0.01,
EFILE= ,
FRHMAX=3,
FRHS=1,
FRME=0.1,
FRMR=0,
FRTS=0.1,
FRVS=0.01,
HSCALE=10800,
ICHEM=8,
CAPEMIN=-1,
INITD=0,
ISOT=-99,
KBLS=1,
KBLT=5,
KDEF=0,
KHMAX=9999,
KMIX0=100,
KMIXD=0,
KMSL=0,
KPUFF=0,
KRND=6,
KSPL=1,
KZMIX=0,
MAXDIM=1,
MAXPAR=10000,
MGMIN=1000,
NCYCL=0,
NDUMP=0,
NINIT=1,
NTURB=0,
OUTDT=0,
P10F=1,
QCYCLE=0,
KRAND=2,
SPLITF=1,
TKERD=0,
TKERN=0,
TLFRAC=0.1,
TRATIO=0.75,
TVMIX=1,
VEGHT=0.5,
VSCALE=200,
VSCALES=-1,
WINDERRTF=0,
ZICONTROLTF=0,
IDSP=2,
WVERT=.FALSE.,
NUMPAR=990,
IVMAX=13,
VARSIWANT='time','indx','lati','long','zagl','foot','dmas','mlht','sigw','tlgr','zsfc','pres','de
ns',
 /
```

Example CONTROL file:

```
18 11 09 16 13
45
40.60064202014 -74.04306846587 455.127197885
40.6015926848 -74.04304914372 455.9137415067
40.60245133932 -74.04308266535 454.9592458993
40.60340344065 -74.04317974082 453.9805207748
40.60431101349 -74.04332949445 453.8326739412
40.60518328781 -74.04350759658 454.3034994239
40.6061455101 -74.04371954258 454.7968855452
40.60704556841 -74.04392023225 455.6618639879
40.60783208485 -74.04409449802 456.7384945324
40.60865780833 -74.04427400877 457.5645947041
40.60956152734 -74.04446618455 457.8047095083
40.61040362672 -74.04463921472 458.0249251049
40.6112355552 -74.04480356492 457.8821824064
40.61206989844 -74.04496323873 458.7885658272
40.61294377362 -74.0451293987 459.8765455058
40.61371465654 -74.04527605117 461.198464537
40.61447169466 -74.04541948963 462.3079021688
40.61522699744 -74.04556310903 464.0467691617
40.61598032533 -74.04570265259 465.1654199616
40.61673252545 -74.0458365306 464.6065725193
40.61749738128 -74.04596609883 464.4895949179
40.61824954063 -74.04608392131 464.8334373806
40.61898817499 -74.04619026027 465.4002148002
40.61973887057 -74.04628876109 465.5979593664
40.62049006078 -74.04638364206 465.4478125846
40.62124169715 -74.04646799331 465.335222084
40.62199313403 -74.04653575725 466.0757751414
40.62275580125 -74.04658327992 467.431080902
40.6235037215 -74.04660634237 467.7203670128
40.6242373568 -74.04661007391 466.6348039499
40.62499570109 -74.04660355886 465.7466528489
40.62574235895 -74.04659485148 466.0213384808
40.62649006771 -74.04658584303 466.5654804102
40.62723853635 -74.04657477255 466.5204133131
40.62798660662 -74.04655854866 466.4840465181
40.62874572076 -74.04653208312 465.8281389562
40.62948980951 -74.04649178965 464.7062223462
40.63023133139 -74.04643442729 462.9225338671
40.6309719438 -74.04636082652 461.2886596721
40.63171235094 -74.04627291833 459.5671045543
40.63246621767 -74.04617527564 457.7512087631
40.63319565898 -74.04607152109 455.8651045806
40.63393758086 -74.04595180029 453.8632957782
40.63467896369 -74.04581575191 451.4588603387
40.63541817752 -74.04566012772 449.9242899969
-48
0
25000.0
3
/gpfs/projects/ShepsonGroup/HYSPLIT/run/inputs/met/
ERA5_20181107.ARL
/gpfs/projects/ShepsonGroup/HYSPLIT/run/inputs/met/
ERA5_20181108.ARL
/gpfs/projects/ShepsonGroup/HYSPLIT/run/inputs/met/
ERA5_20181109.ARL
1
t001
1
0.01
18 11 09 16 13
1
0.0 0.0
0.5 0.5
30.0 30.0
./
cdump
1
100
00 00 00 00 00
00 00 00 00 00
00 2 00
1
0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.0
0.0 0.0 0.0
0.0
0.0
```

S3: Domains and prior fluxes

The boundaries of the outer d01 and inner d03 domains are shown in Figure S3.1. The inner d03 domain is centred on New York City, and includes parts of the following US states: New York, New Jersey, Connecticut, Pennsylvania, and Delaware. The Philadelphia urban area is also included within this domain. The outer d01 domain encompasses all or part of 20 US states, as well as part of Canada. Large methane emissions from oil and gas operations in the Appalachian basin are included within this outer domain.

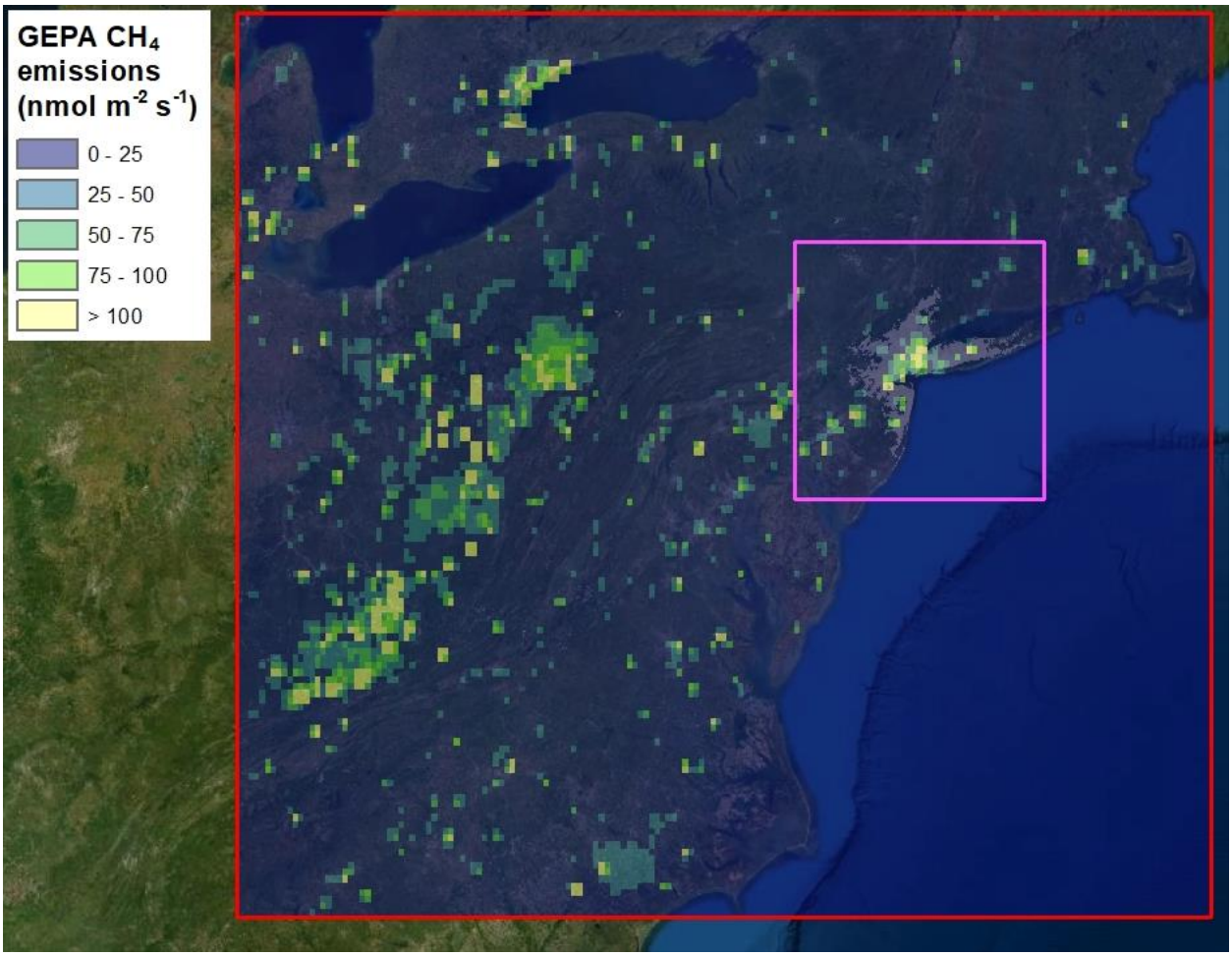

Figure S3.1: Map showing the GEPA prior CH_4 fluxes within the outer d01 (red) and inner d03 (pink) domains (including the Canadian emissions from EDGAR v5.0). The shaded area within d03 represents the New York-Newark urban area. Source: Maasakkers et al. (2016), EDGAR v5.0, ESRI, Maxar, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN and the GIS User Community.

For CO_2, we used three different flux priors: Vulcan v3.0 (Gurney et al., 2020), ACES v2.0 (Gately and Hutyra, 2017) and EDGAR v5.0 (European Commission Joint Research Centre (JRC)/Netherlands Environmental Assessment Agency (PBL), 2019; Crippa et al., 2020). For CH_4, we used four different flux priors: EDGAR v5.0, EDGAR v4.2 (European Commission Joint Research Centre (JRC)/Netherlands Environmental Assessment Agency (PBL), 2011), GEPA (Maasakkers et al., 2016), and a composite prior constructed by taking the average of the GEPA and EDGAR v4.2 fluxes within each grid cell. Of these seven priors, two (Vulcan and ACES) are bottom-up inventories, as described by Gurney et al. (2020), while the remainder rely on spatial proxies to downscale national emission totals for most sources. None of the inventories contained fluxes for the years of our flights, so for each prior we used the most recent year available: 2015 for Vulcan, 2017 for ACES, 2018 for EDGAR v5.0 CO_2, 2015 for EDGAR v5.0 CH_4, 2012 for the GEPA, and 2008 for EDGAR v4.2. We used annual mean fluxes in all cases, such that any sub-annual variability is only informed by the measurements rather than the prior (see section S5 for more details).

We calculated total emission rates for the New York-Newark urban area, as defined by the 2019 version of the US Census Bureau's Topologically Integrated Geographic Encoding and Referencing (TIGER) database (United States Census Bureau, 2019). Totals were calculated as a weighted sum of the emissions from all grid cells, with the weights given by the fraction of each grid cell within the urban boundary. For example, a grid cell lying entirely inside the urban area was assigned a weight of one, a grid cell entirely outside the urban area was assigned a weight of zero, and grid cells straddling the urban area boundary were assigned fractional weights. We also calculated total emission rates for the five boroughs of New York City using the same approach, according to the boundaries defined by the New York City Government (release number 20C; New York City Department of City Planning, 2020)

The GEPA and both EDGAR inventories are available on a 0.1° × 0.1° lat-lon grid, making it trivial to regrid these fluxes to the 0.02° and 0.08° lat-lon grids of d03 and d01 while conserving mean flux. Vulcan and ACES are available at 1 km × 1 km resolution on a Lambert Conformal Conic (LCC) projection; these inventories were regridded using the conservative regridding scheme provided by the Python package xESMF (Zhuang, 2020). However, even though mean flux is conserved by this operation, the emission rate within the New York-Newark urban area is not conserved because the fraction of emissions within the urban area boundary changes.

We tested the impact of regridding the Vulcan and ACES inventories by comparing the urban area emission totals calculated at the resolution of the d01 and d03 grids. We then projected the urban area boundary onto the native LCC coordinate system of Vulcan and ACES and repeated the urban total calculation. The results are shown below in Table S3.1. For Vulcan the total emission rate within the urban area boundary was very similar when calculated on all three grids, however for ACES the total emission rate when calculated on the d01 domain grid was 4% lower than the total flux calculated on the native LCC grid. This shows the benefit of using a high-resolution domain, rather than simply calculating posterior urban emission totals on the coarse grid of the d01 domain.

Table S3.1: Total CO_2 emission rates for the New York-Newark urban area, calculated by aggregating emissions from the Vulcan and ACES inventories on several different grids.

Resolution	**Vulcan 2015 annual average (kmol s^{-1})**	**ACES 2017 annual average (kmol s^{-1})**
Native (1 km on LCC grid)	103.8	117.0
Regridded to 0.02° lat-lon (d03)	104.5	116.2
Regridded to 0.08° lat-lon (d01)	102.8	111.7

S4: Footprint sensitivity

The nine flights were conducted under a variety of wind directions and were designed to maximise sensitivity to the high-emitting urban core, subject to airspace constraints, with some measurements upwind (in most cases) to characterize the influence of upwind sources. Example d03 domain footprints are shown in Figure S4.1 in relation to the flight tracks and New York-Newark urban area boundary.

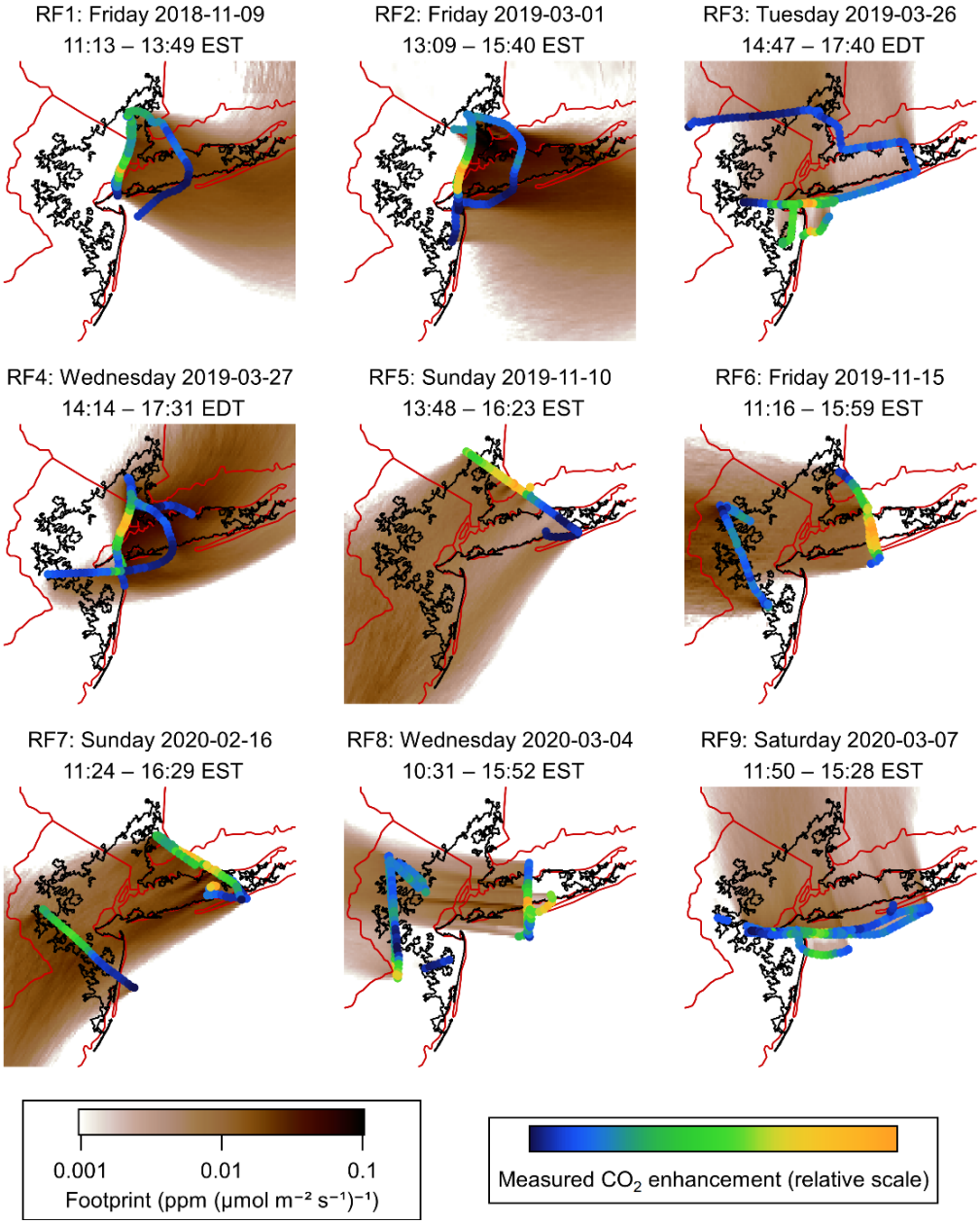

Figure S4.1: Flight tracks and d03 domain footprints for all nine flights. Flight times are given in local time. The footprints shown were calculated using ERA5 meteorology and the Kantha and Clayson (2000) turbulence parameterisation. Flight tracks are coloured by measured CO_2 enhancement on a relative scale bounded by the maximum and minimum values for each flight. The logarithmic footprint colour scale is saturated at the shown limits. The New York-Newark urban area boundary is shown in black, with state boundaries in red. The mole fraction unit ppm is equivalent to the unit µmol mol^{-1}.

S5: Representative inventory totals

The Vulcan and ACES anthropogenic CO_2 inventories are available at hourly resolution, such that diurnal, weekly and seasonal cycles can be resolved. In this study we have used data from the most recent year available for each inventory: 2015 for Vulcan and 2017 for ACES. As we do not have inventory data corresponding to the actual date and time each flight was conducted, we opted to use annual average emissions from these inventories as our prior flux estimates. However, both inventories show a large variability in CO_2 emissions on diurnal and seasonal timescales (Figure S5.1). Therefore, we also compared our average posterior emission rates (and the flight-to-flight variability in posterior emissions) to inventory emissions on dates and at times of day that are more representative of our flights.

For each flight date, we looked for the closest date in the inventory that was on the same day of the week as that flight. For example, flight RF1 was on Friday November 9th 2018. To represent this flight, we took data for Friday November 6th 2015 from Vulcan and Friday November 10th 2017 from ACES. We then took the equivalent days (Fridays in this case) from two weeks either side of these dates so that we had 5 representative inventory days for each flight, resulting in 45 representative days in total.

For every flight, we then calculated a weighted average emission across the hourly totals for each of the representative days. To calculate the weights, we first calculated the mean modelled enhancement (averaged across all minutes of the flight) resulting from emissions within the urban area for each hour of the footprint. The hourly emissions from the inventory were then weighted by the fraction of modelled urban area enhancement corresponding to each hour. For example, if (for a given flight) only emissions between 12pm and 1pm contributed to the modelled urban enhancement, then the inventory emissions from that hour would be used for each of the 5 representative days. However, if there were a 75% contribution from fluxes emitted between 12pm to 1pm and a 25% contribution from fluxes between 1pm to 2pm, then the inventory fluxes would be calculated as a weighted average of these two hours, with weights equal to 0.75 and 0.25 respectively.

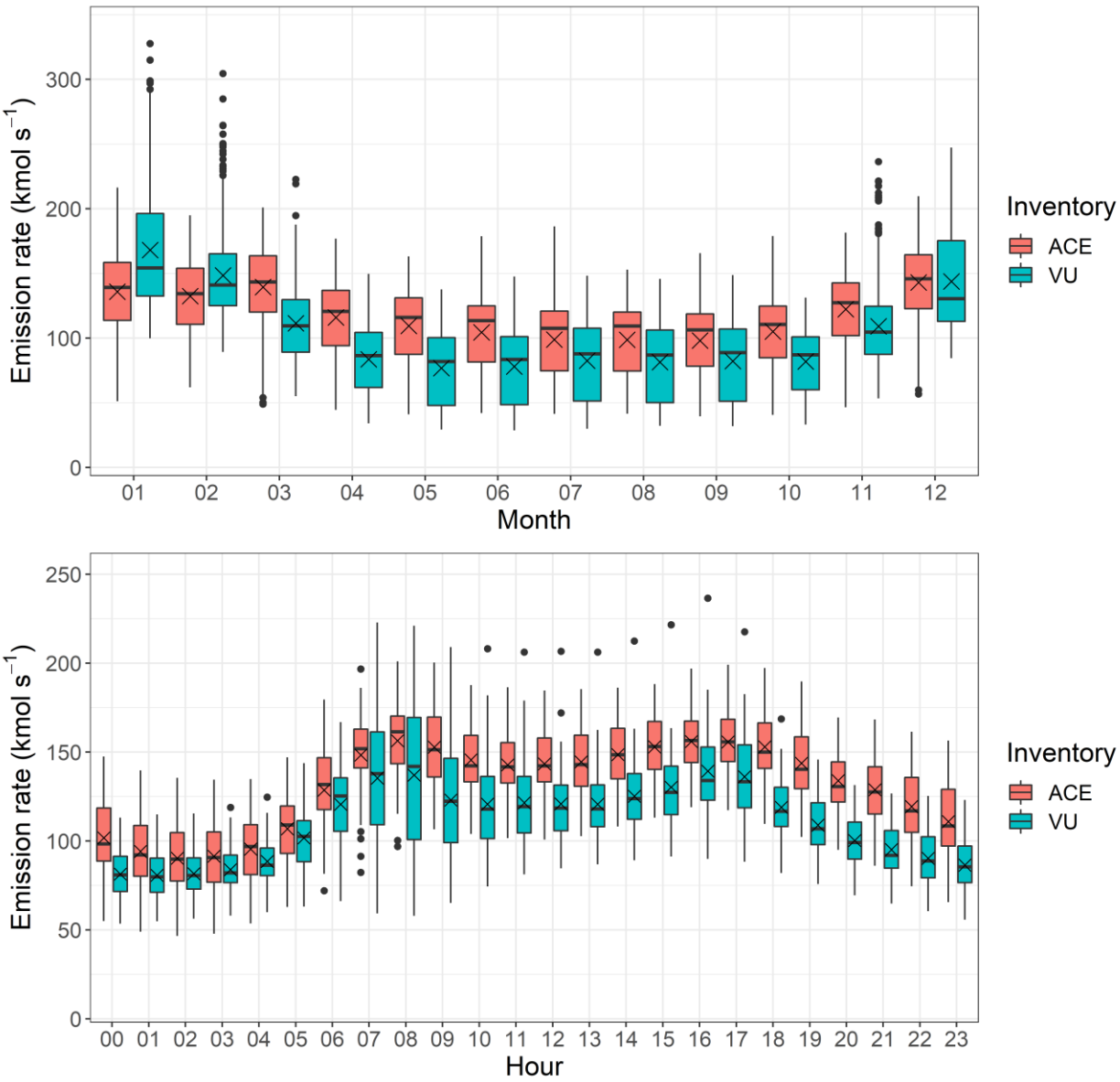

Figure S5.1: Boxplots showing the monthly cycle (upper panel) and March/November diurnal cycle (lower panel) of anthropogenic CO_2 emissions in the 2017 ACES (ACE) and 2015 Vulcan (VU) inventories. The upper panel represents all data within each month, while the lower panel shows emissions for each hour (LT) of the day for all days in March and November (as eight of the nine flights in this study were in one of those two months). The crosses indicate the mean values. The boxes extend between the upper and lower quartiles, with the median values shown as solid horizontal black bars. The whiskers extend to the highest and lowest data points within 1.5 times the interquartile range of the upper and lower quartiles, respectively. All data outside these whiskers are shown as individual points.

S6: Biospheric influence

The measured CO_2 mole fractions are influenced by any biospheric respiration and photosynthesis within the sample footprint, while the inventories used in this study contain only anthropogenic fluxes. We used output from the Vegetation Photosynthesis and Respiration Model (VPRM; Mahadevan et al., 2008) to estimate the impact of biospheric fluxes on our results. VPRM is an empirical model, calibrated with flux measurements for different vegetation classes. It parameterises the Net Ecosystem Exchange (NEE) as a function of remotely sensed Enhanced Vegetation Index and Land Surface Water Index and in situ measurements of temperature and incoming solar radiation. Here we used year-specific hourly fluxes for 2018 and 2019, generated at 0.02° resolution; see Gourdji et al. (2021) for more details of the VPRM version used in this study.

Figure S6.1 shows the daytime-average ratio of total flux (fossil fuel + biosphere) to fossil fuel flux (from both Vulcan and ACES) for the New York-Newark urban area, based on VPRM runs for 2018 and 2019. The ratios were computed at hourly resolution, then for each day the daytime average was taken across all hours between 0800 and 1800 EST. The large impact of biospheric fluxes during summer months is evident. The days of the year on which our flights were conducted are shown as dashed vertical lines. It can be seen that at these times of year the contribution of biospheric fluxes to the total urban area flux is very small, approximately 2% on average. This is smaller than the contribution reported for Boston (Sargent et al., 2018) and Indianapolis (Turnbull et al., 2019) in the winter, but it is not directly comparable to those studies because it represents only fluxes from the census-defined New York-Newark urban area. Better quantification of urban biospheric fluxes from New York City and surrounding areas is a topic for future research, but for the purpose of this study the important conclusion is that, at the times of our flights, these biospheric fluxes are likely to be small compared to anthropogenic fluxes.

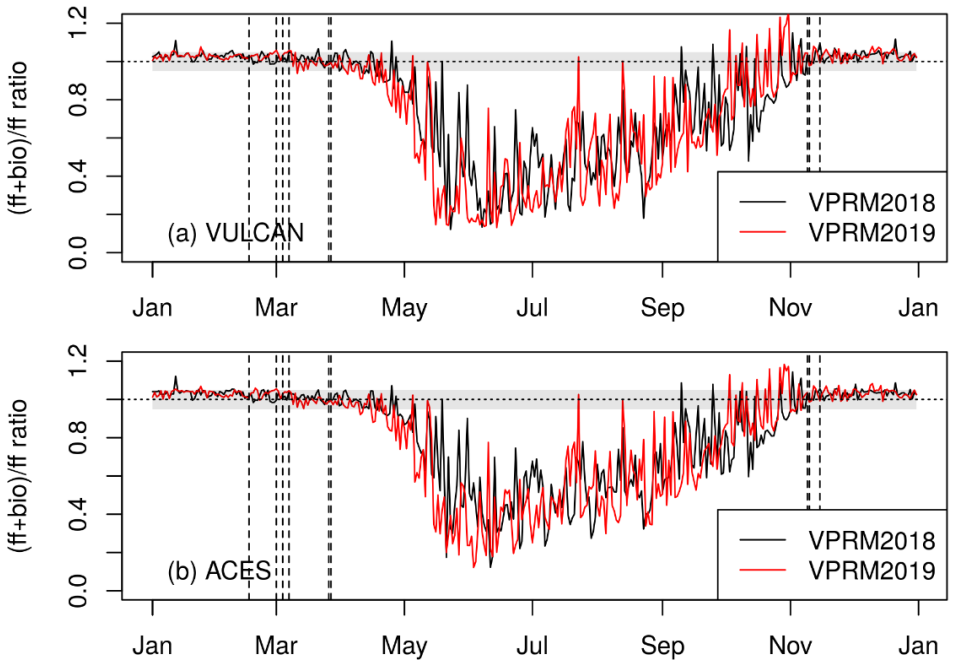

Figure S6.1: The ratio of total CO_2 flux (fossil fuel + biosphere) to fossil fuel CO_2 flux for the New York-Newark urban area, based on VPRM runs for 2018 (black) and 2019 (red). Panel (a) uses 2015 fossil fuel emissions rates from Vulcan, and panel (b) uses 2017 fossil fuel emissions rates from ACES. The dashed vertical lines show the days of the year on which our flights were conducted. The grey shading represents ratios between 0.95 and 1.05.

S7: CH_4:CO_2 ratios

The upper panel of Figure S7.1 shows the ratio of our prior and posterior CH_4 emission rates to our mean posterior CO_2 emission rate for the New York-Newark urban area. Results from two previous studies that measured the CH_4:CO_2 enhancement ratio downwind of the same urban area are also shown. Plant et al. (2019) reported the average of 14 plume measurements taken over 10 separate days in April/May 2018. Floerchinger et al. (2021) reported separate values for a winter flight in March 2018 and a summer flight in September 2019. Figure S7.1 shows the value from the winter flight as this corresponds to the non-growing season conditions of our campaign. The error bar on the Floerchinger et al. (2021) estimate represents only the uncertainty associated with estimating the enhancement ratio for that single flight, hence it is much smaller than the error bars for Plant et al. (2019) and our study, which include flight-to-flight variability.

The lower panel of Figure S7.1 shows CH_4:CO_2 emission ratios for this study broken down by flight. The boxplot for each flight consists of CH_4 emission rates across the ensemble of priors and transport models, all divided by the ensemble-mean CO_2 emission rate for that flight. The mean ratio across all ensemble members and flights is (4.9 ± 0.7) nmol μmol^{-1} (1σ variability across flights). Note that the spread in CH_4:CO_2 ratio for each individual flight is dominated by variability across priors.

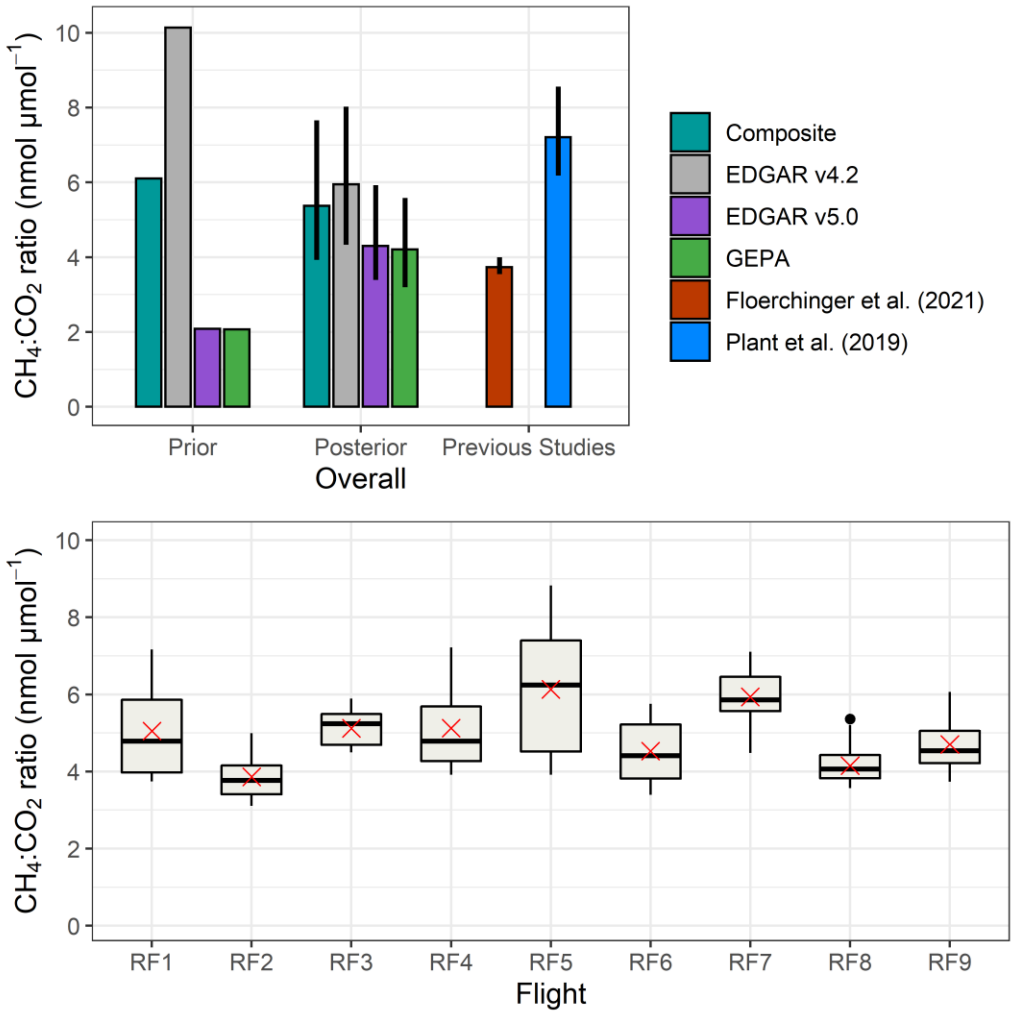

Figure S7.1: Ratio of prior and posterior CH_4 emission rates to the mean posterior CO_2 emission rate for the New York-Newark urban area. Also included are the measured CH_4:CO_2 ratios for this urban area from two previous studies. In the top panel, error bars for this study represent the 97.5th and 2.5th percentiles of the ratios calculated using different flights and models. Error bars for Plant et al. (2019) represent 95% confidence intervals calculated using a bootstrap analysis of the 14 individual plume measurements. Error bars for Floerchinger et al. (2021) represent 95% confidence intervals based on propagation of component uncertainties. The lower panel shows posterior CH_4:CO_2 ratios for this study broken down by flight. See Figure S5.1 caption for boxplot convention.

S8: Posterior−prior spatial differences in central NYC

Figure S8.1 shows a zoomed in view of Manhattan posterior and prior fluxes for the three CO_2 priors. Posterior fluxes for this area are in much closer agreement across the three cases than the prior fluxes. However, further research is required, including higher resolution transport models, to understand whether small scale spatial features observed in these posterior emission maps reflect real spatial patterns in emissions or artefacts due to the coarse resolution of the transport models used.

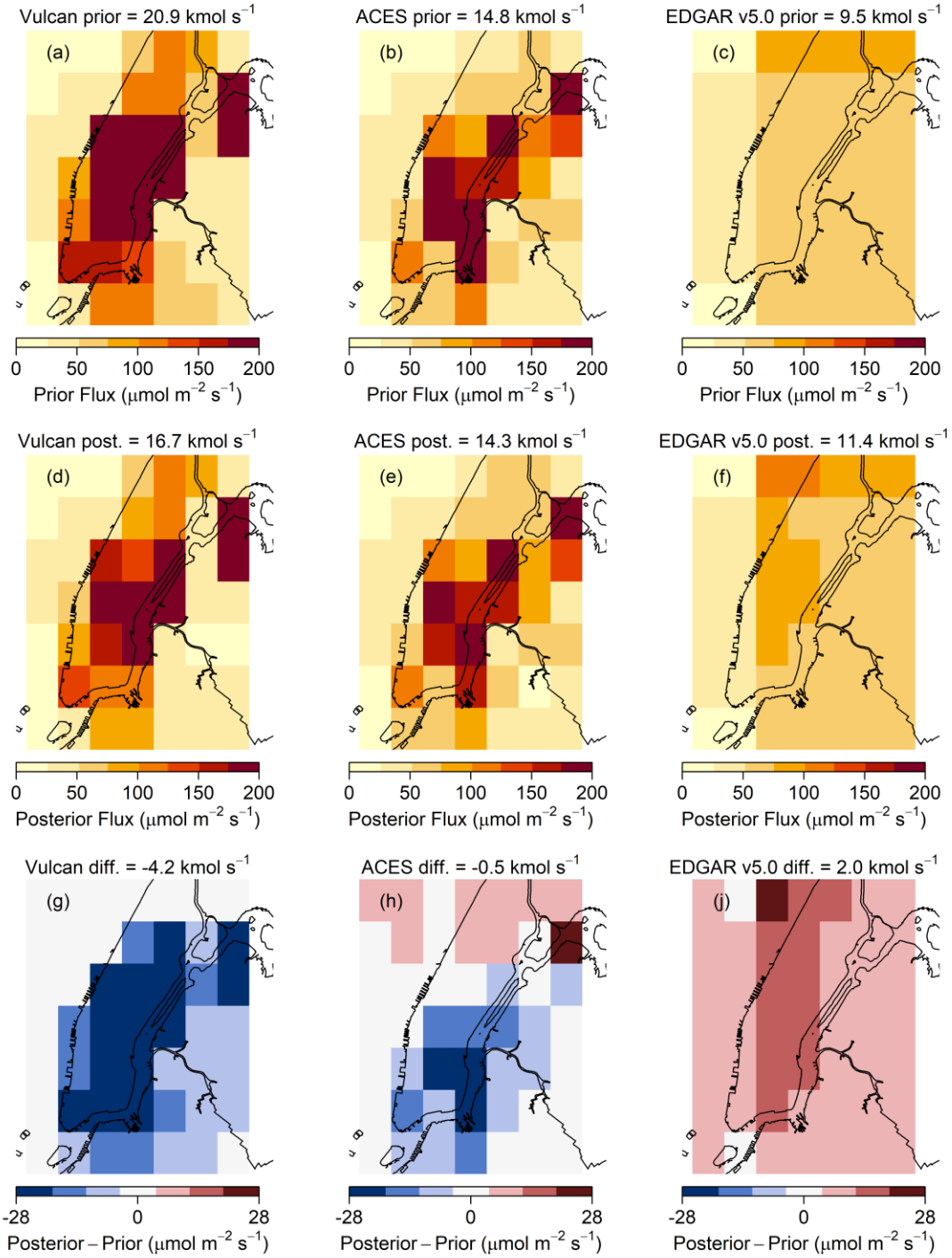

Figure S8.1: Maps showing the prior (a–c) and posterior (d–f) CO_2 fluxes from Manhattan and surrounding areas, and the posterior−prior differences (g–j), for all priors (averaged over all models and flights). Total emission values are taken over all grid cells shown. Colour scale is saturated at the shown limits.

S9: Nested inversion approach

To calculate the model-measurement mismatch ($\mathbf{Hx} - \boldsymbol{y}$), a background mole fraction must first be subtracted from the measurements. This background value represents the CO_2 or CH_4 mole fraction of the measured air prior to entering the domain. First, we describe the background determination for our initial d01 inversion.

In our base case, we calculated the background mole fraction based on the measured and modelled mole fractions during minutes of the flight where both values were low. In an idealised scenario, the locations of the lowest measured and modelled mole fractions would be the same. However, spatial misallocation of fluxes in the prior, error in the transport model and variability in the mole fraction of the air prior to entering the d01 domain can all cause the lowest measured and modelled mole fractions to occur in separate locations from each other. Therefore, we choose to define the background points in such a way as to ensure they represent low values of both the measured and modelled mole fractions.

To determine which data to select, we first added the measured mole fraction to the modelled mole fraction for each minute of the flight. Then we took all points between the 1st and 5th percentiles of this combined dataset to be background points. The measured enhancements for the d01 domain ($\boldsymbol{y_{d01}}$) were then calculated according to:

$$\boldsymbol{y_{d01}} = \boldsymbol{y_{tot}} - \left(\overline{\boldsymbol{y}}_{\boldsymbol{bg}} - \overline{(\mathbf{H}\boldsymbol{x_b})}_{\boldsymbol{bg}}\right) \quad \text{(S1)}$$

Here $\boldsymbol{y_{tot}}$ represents the average measured mole fraction for each minute of the flight, $\overline{\boldsymbol{y}}_{\boldsymbol{bg}}$ is the measured mole fraction averaged over the background points and $\overline{(\mathbf{H}\boldsymbol{x_b})}_{\boldsymbol{bg}}$ is the modelled mole fraction averaged over the background points, which represents the contribution from sources inside our domain to the background measurements.

Once posterior fluxes had been derived for the d01 domain, using the background from Eq. S1 in Eq. 2 (see main manuscript), these posterior fluxes were used to construct a background for the subsequent inversion on the d03 domain. For this d03 inversion the measured enhancements ($\boldsymbol{y_{d03}}$) were calculated according to:

$$\boldsymbol{y_{d03}} = \boldsymbol{y_{d01}} - (\mathbf{H}\boldsymbol{x_{d01}})_{\boldsymbol{oc}} \quad \text{(S2)}$$

Here $\boldsymbol{y_{d01}}$ represents the measured mole fraction enhancements calculated in Eq. S1 and $(\mathbf{H}\boldsymbol{x_{d01}})_{\boldsymbol{oc}}$ represents the contribution of fluxes outside the d03 domain, calculated as modelled mole fractions based on the posterior fluxes from the d01 inversion, with all fluxes within the bounds of the d03 domain set to zero. Posterior fluxes for the high-resolution d03 domain were then obtained using Eq. 2 with the measured enhancements $\boldsymbol{y_{d03}}$ and the modelled enhancements calculated using the d03 domain footprints.

We tested the impact of the nested inversion approach by repeating the d03 inversion without accounting for spatial variability in the mole fraction of air flowing into the d03 domain. In this non-nested sensitivity test we used a single background value for the d03 inversion, as opposed to the nested approach where the background included the spatially-varying contribution from Eq. S2. The measured enhancements for the non-nested sensitivity test were calculated using Eq. S1, with $\overline{(\mathbf{H}\boldsymbol{x_b})}_{\boldsymbol{bg}}$ calculated using the d03 footprints and d03 prior flux map. The locations of the background points were recalculated for this sensitivity test based on the d03 modelled mole fractions, but often they coincided with the background locations used in the d01 inversion.

The differences in posterior emission rate using the non-nested inversion, relative to the posterior emission rate for the corresponding nested inversion, are shown for each flight in Figure S9.1. Using the non-nested inversion resulted in mean posterior CO_2 and CH_4 emission rates for the New York-Newark urban area that were 2.8% higher and 6.6% higher, respectively, than for the nested inversion case. Mean posterior results for individual flights differed from the nested inversion by up to 8.3% for CO_2 and 13.3% for CH_4, with individual ensemble members showing much larger differences in some cases.

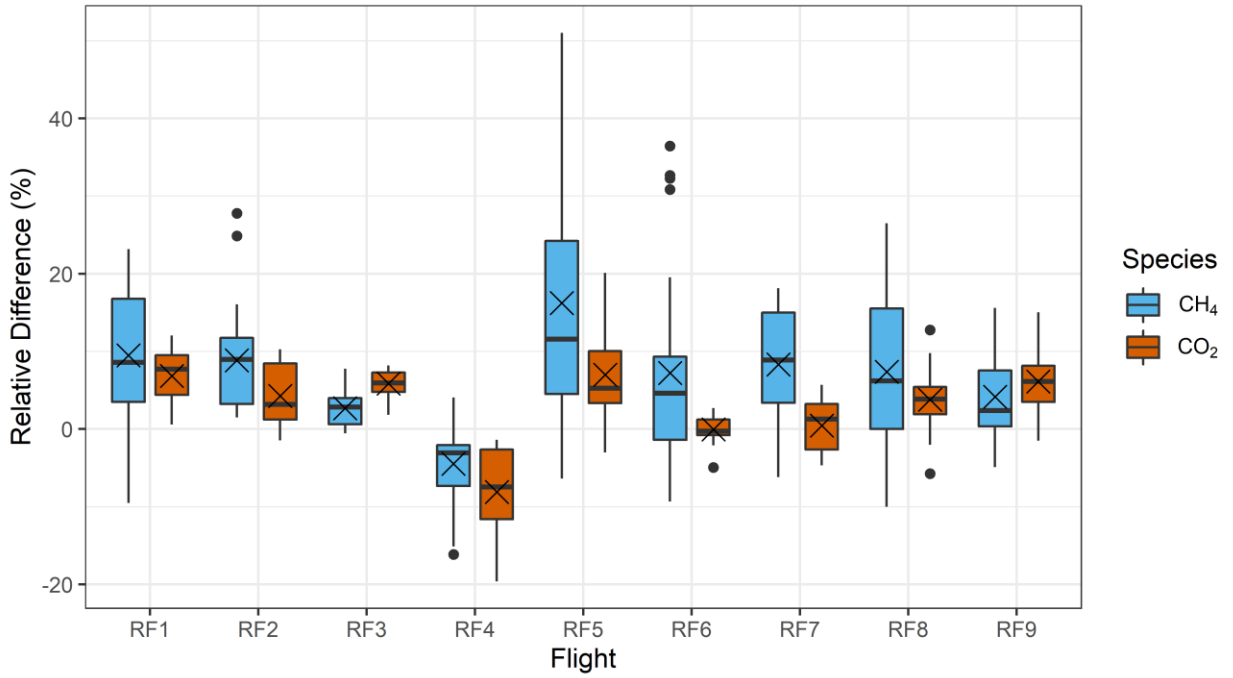

Figure S9.1: Differences between the urban area posterior emission rate for the non-nested inversion and the corresponding nested inversion value for the same ensemble member, expressed relative to the nested inversion value. A positive relative difference indicates the posterior emission rate from the non-nested inversion was larger than the corresponding nested inversion value. See Figure S5.1 caption for boxplot convention.

To understand the source of these differences, consider the following two situations:

1. The spatially-varying background (used in the nested inversion, calculated using d01 posterior fluxes) is larger in magnitude within the urban plume than it is at the background points of the non-nested inversion. In this case, using a single background value, calculated at the background points of the non-nested inversion, will underestimate the background at the in-plume points. Therefore the non-nested inversion will result in a larger posterior emission rate for the urban area.
2. The spatially-varying background is larger in magnitude at the background points of the non-nested inversion that it is within the urban plume. In this case the opposite is true and the in-plume background is overestimated by the non-nested inversion, resulting in a smaller posterior emission rate for the urban area.

As can be seen in Figure S9.1, for most flights the majority of ensemble members show results consistent with situation 1), with the notable exception of RF4 (which is consistent with situation 2). The spatially-varying background in RF4 was enhanced at the background points of the non-nested inversion

(relative to points in the urban plume) due to the influence of emissions from Providence, RI, and Boston, MA.

To further demonstrate the impact of the nested inversion approach, example cases showing the correction of distant sources by the d01 inversion are shown in Figure S9.2. It is clear that sources far upwind of the area of interest were corrected in the initial d01 inversion. For CO_2, the main upwind sources being corrected in these examples were the urban areas of Philadelphia, Baltimore and Washington, DC, although isolated point sources across states were also corrected. For CH_4, the main upwind sources being corrected were the oil and gas operations in the Appalachian basin. It is important to note that these corrections outside the d03 domain do not reflect an accurate quantification of emissions from the upwind sources, as they are only weakly constrained by the measurements, but they do help to account for spatial variability in the composition of air flowing into the d03 domain.

Other flights also exhibited non-negligible corrections in sources outside the d03 domain, with the specific adjustments changing from flight to flight based on synoptic conditions. In some cases corrections were even larger than in the urban area of interest. Just using the prior to estimate the influence of upwind sources, or not considering them at all (as in our sensitivity test above), leads to a misrepresentation of their contribution, inducing error in our urban emission estimates. From a practical perspective, we note that measurements upwind of the urban area provide an important constraint on these upwind sources.

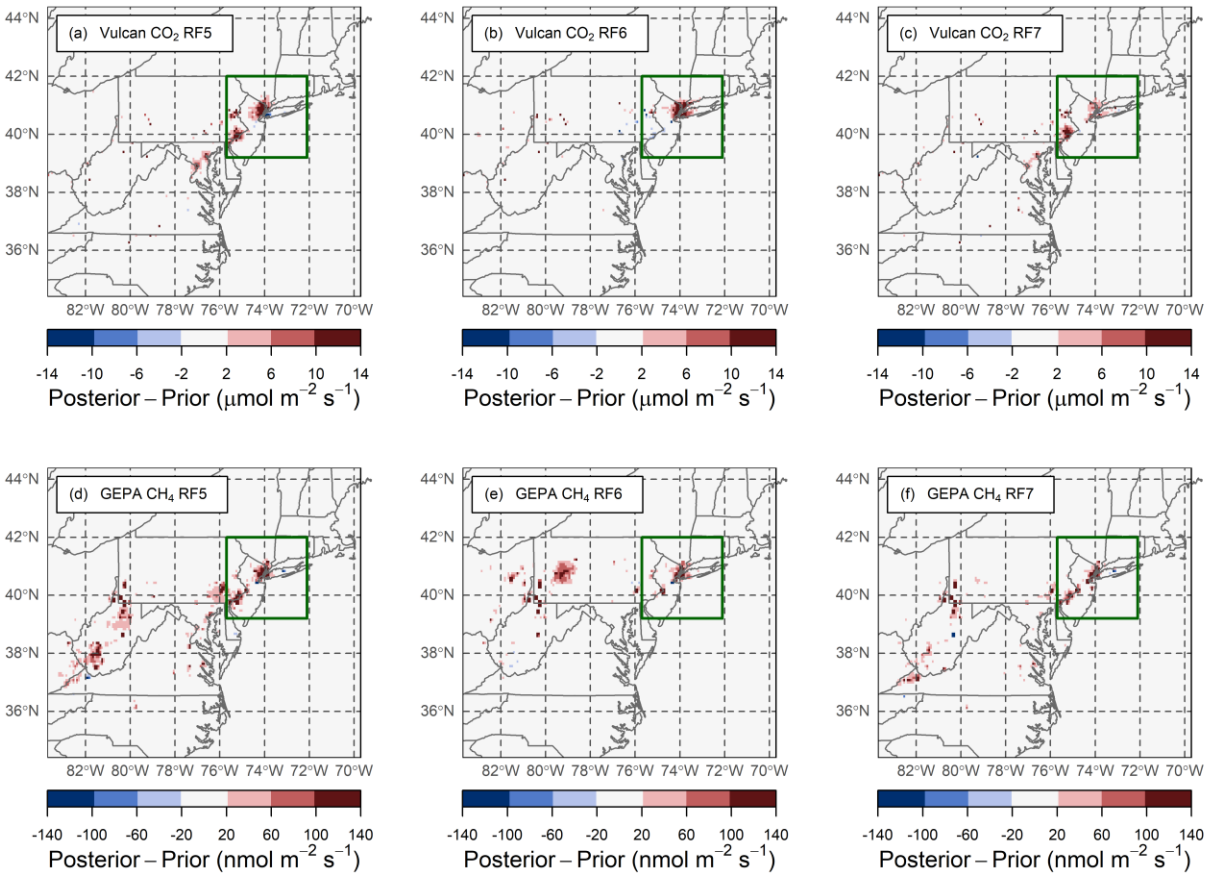

Figure S9.2: Difference between posterior and prior fluxes for certain flights (RF5–RF7) and priors (Vulcan and GEPA) from the d01 inversion (averaged over all models). The boundary of the d03 domain is shown in green. Colour scale is saturated at the shown limits.

S10: Background sensitivity tests

In addition to our base case ensemble of inversions, we performed a sensitivity analysis to better understand the impact of different background choices (within the context of the nested inversion approach). The background options used in our base case and in our sensitivity analysis were defined as follows:

Base case:
As described above in section S9, we added the measured mole fraction and modelled mole fraction for each minute of the flight, and took all points between the 1st and 5th percentile of this combined dataset to be the background points. The background mole fraction, as given by $\boldsymbol{y_{tot} - y_{d01}}$ (see Eq. S1), was then calculated as the difference between the mean measured mole fraction over these background points and the mean modelled mole fraction over these background points. We used a separate modelled mole fraction for each ensemble member (i.e., every combination of transport model and prior has its own background).

B1:
We used the same approach as the base case, but we took the mean modelled mole fraction over the eight transport model ensemble members (instead of using a separate background for each transport model). So in this case each prior still had its own background, but the background was the same for all the transport model ensemble members.

B2:
We used the same approach as the base case, but defined the background points differently. In this case we took all points between the 1st and 5th percentile of the measured dataset to be the background points.

B3:
We used the same approach as the base case, but defined the background points differently. In this case we took all points between the 1st and 5th percentile of the modelled dataset to be the background points.

B4:
We used the same approach as the base case, but defined the background points differently. In this case we took all points below the 5th percentile of the combined (measured mole fraction + modelled mole fraction) dataset to be the background points (as opposed to all points between the 1st and 5th percentile in the base case).

B5:
We used the same approach as the base case, but defined the background points differently. In this case, we calculated the measured and modelled backgrounds independently. The mean measured mole fraction over the set of points that fell between the 1st and 5th percentiles of the measured dataset was used as the measured background. The mean modelled mole fraction over the set of points that fell between the 1st and 5th percentile of the modelled dataset was used as the modelled background. The background mole fraction was then calculated as the difference between these two average values.

The urban area posterior emission rates calculated using each background definition are shown in Figure S10.1. The 1σ variability in mean urban area posterior emission rate across the six background options was 4.3% for CO_2 and 7.7% for CH_4, relative to the respective base case posterior emission rates. Background choices B1, B4 and B5 had little impact on the mean posterior emission rate relative to the base case (less than 0.7% difference for CO_2 and 1.7% difference for CH_4). On the other hand, using B2 resulted in a mean posterior emission rate that was 3.8% larger for CO_2 and 11% larger for CH_4. This makes qualitative sense, because choosing background points based on the locations of the lowest measured mole fractions (B2), rather than locations where both the measured and modelled mole fractions are low (base case), inevitably results in larger modelled mole fractions at the background points. Based on Eq. S1, larger modelled mole fractions at the background points ($\overline{(\mathbf{H}\boldsymbol{x_b})}_{\boldsymbol{bg}}$) result in larger measured enhancements ($\boldsymbol{y_{d01}}$) and therefore larger posterior fluxes. By the same logic, using B3 resulted in lower posterior emission rates (by 9.1% for CO_2 and 14% for CH_4) because it inherently leads to smaller measured enhancements.

These impacts arise due to transport and prior flux errors, which can result in significant modelled enhancements at points with low measured mole fractions and vice versa. Any variability in the long-range background (i.e., the mole fraction of air as it flows into the d01 domain) could also result in differences between fluxes calculated using B2, B3 and the base case scenario. These background sensitivity tests demonstrate the importance of carefully considering the background definition. Options such as B3 that are particularly sensitive to transport and prior flux errors should be avoided.

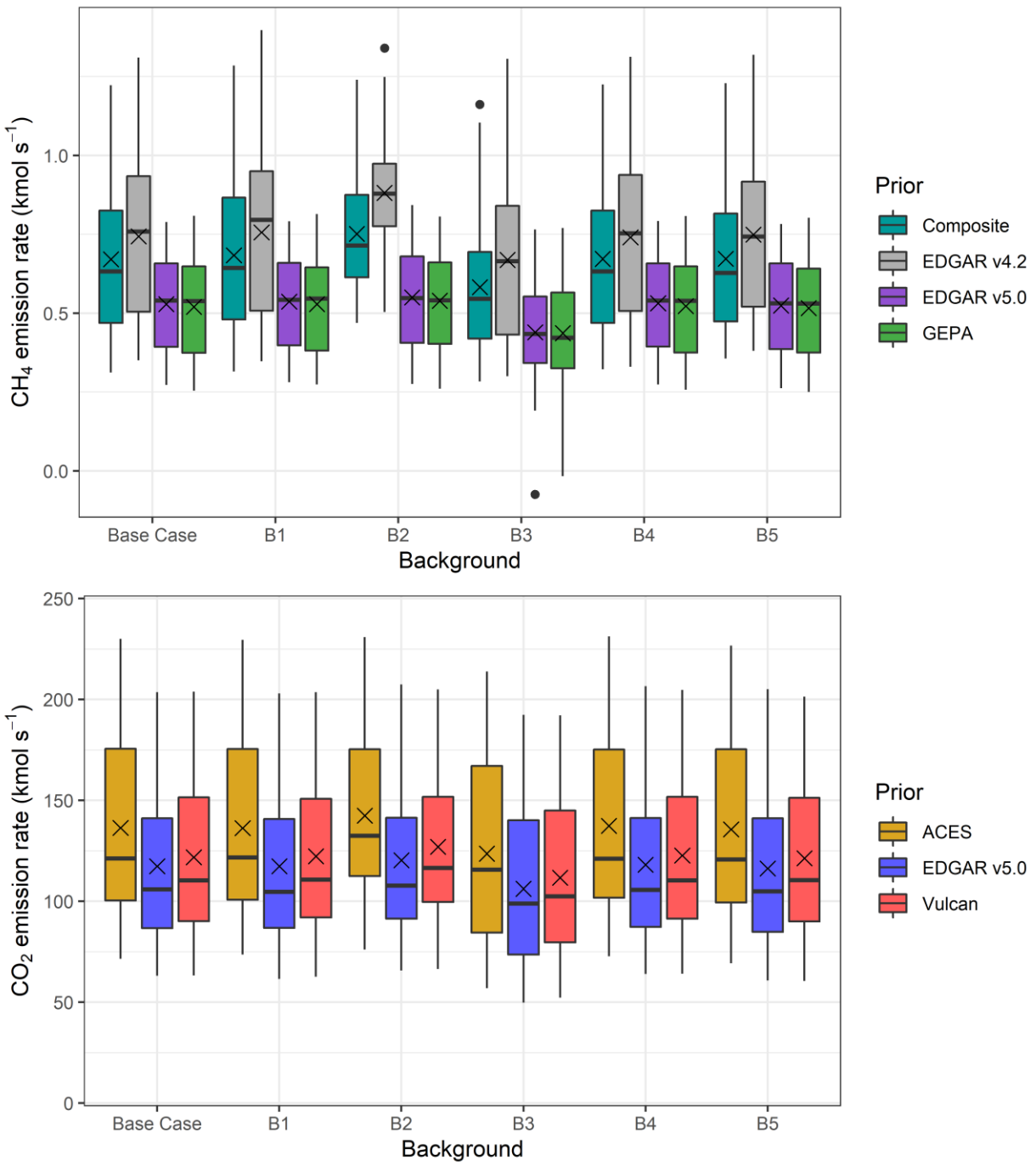

Figure S10.1: Boxplots showing emission rates for the New York-Newark urban area for CH_4 (top panel) and CO_2 (bottom panel), using six different background definitions. The values represented by the boxplots are the posterior emission rates across all flights and transport models for each background. The mean posterior emission rates are shown as crosses. B3 is particularly sensitive to transport model error; for certain ensemble members this results in a negative estimate of CH_4 emission rate. See Figure S5.1 caption for boxplot convention.

S11: Error covariance matrices

The model-measurement mismatch error covariance is defined as:

$$R_{ij} = \sigma_{R,i}\sigma_{R,j}e^{-d_{R,ij}/L_R}e^{-t_{R,ij}/\tau_R} \quad \text{(S3)}$$

Here $\sigma_{R,i}$ represents the uncertainty on the model-measurement mismatch for observation number i, $d_{R,ij}$ is the distance between the locations of observation i and observation j, $t_{R,ij}$ is the time between observation i and observation j, $L_R = 1\ \mathrm{km}$ is the spatial correlation length for model-measurement mismatch errors and $\tau_R = 1\ \mathrm{h}$ is the correlation time for model-measurement mismatch errors.

Prior flux error covariance is defined as:

$$P_{b,ij} = \sigma_i\sigma_j e^{-d_{ij}/L} \quad \text{(S4)}$$

Here σ_i represents the uncertainty on the emissions in grid cell i, d_{ij} is the distance between cells i and j and L is the spatial correlation length for prior flux errors. In our base case, we take σ_i to be equal to the prior emissions in grid cell i (i.e., 100% uncertainty) and $L = 10\ \mathrm{km}$ based on Lopez-Coto et al. (2017) and the variogram analysis below. A minimum value of 1 μmol m^{-2} s^{-1} for CO_2 and 1 nmol m^{-2} s^{-1} for CH_4 is imposed to ensure that grid cells with very low prior fluxes can still be allocated some posterior flux in the inversion (e.g., due to emissions not represented in anthropogenic inventories, such as plant/soil respiration and uptake).

Experimental variograms of the spread within the prior flux ensemble were computed for CO_2 and CH_4 (see Figure S11.1). The spread was calculated as the standard deviation, at 10 km and 2 km resolution, of ACES and Vulcan for CO_2, and EDGAR v5.0 and GEPA for CH_4. We tested for anisotropy by calculating directional variograms at 0, 45, 90 and 135 degrees. We then fitted an exponential model to the experimental variograms showing that the exponential model is appropriate and correlation lengths are on the order of 10 km.

We conducted sensitivity tests on the prior flux error covariance parameters σ_i and L, using prior flux uncertainties of 50% and 200% (based on the uncertainty ranges reported by Andres et al., 2016, and Gately and Hutyra, 2017) and prior-flux-error correlation lengths of 4 km and 12 km based on previously used values for urban emissions (Lauvaux et al., 2016). The results are shown in Figure S11.2. The 1σ variability in mean urban area posterior emission rate across the nine sensitivity tests was 3.7% for CO_2 and 2.8% for CH_4 (relative to the base case: 100% uncertainty, 10 km correlation length), indicating that our conclusions would not change significantly under other reasonable assumptions regarding the prior flux error covariance. As the prior constraints are relaxed (larger prior uncertainties or larger correlation lengths) the mean posterior emission rate for each prior generally moves away from the prior value, and the variability in posterior emission rate across all flights and models increases, as expected. The largest difference in mean posterior emission rate, relative to the base case, was produced by using a correlation length of 4 km and a prior uncertainty of 50%. These stronger prior constraints resulted in a mean posterior emission rate that was lower than the base case by 7.1% for CO_2 and 5.5% for CH_4.

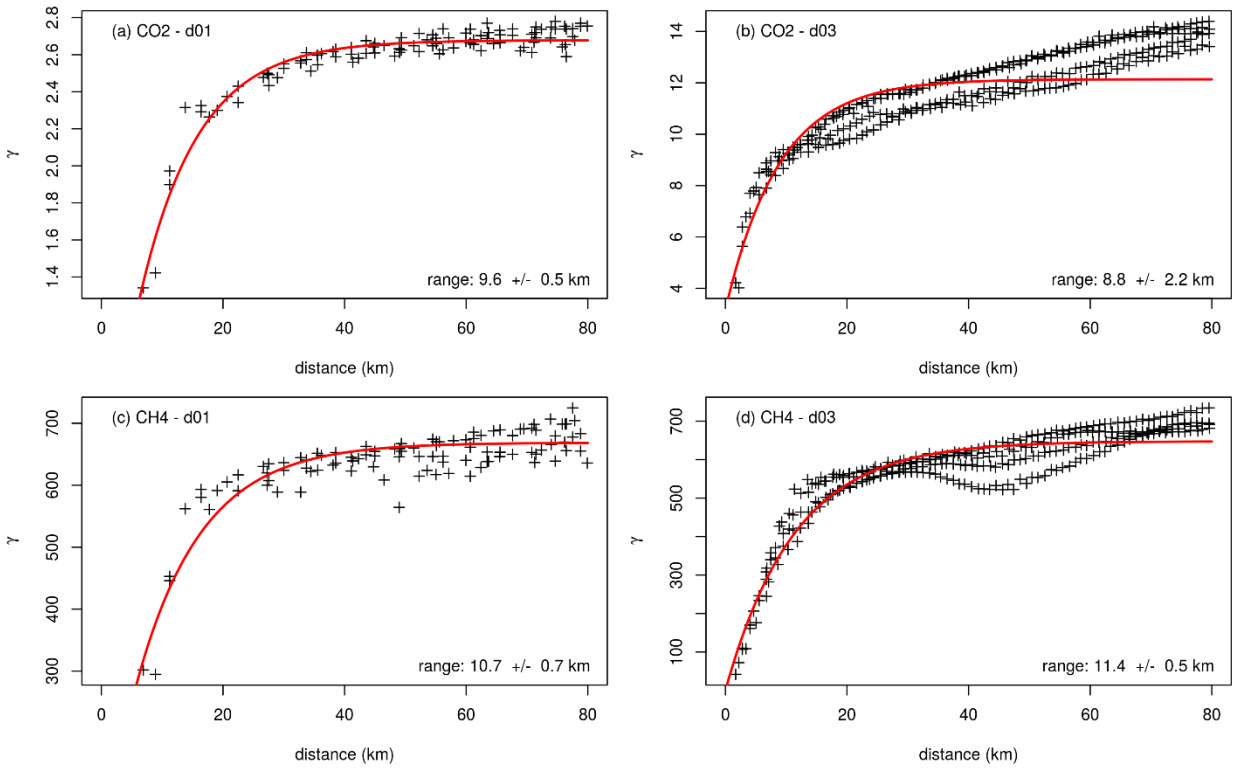

Figure S11.1: Variograms showing the semivariance of the spread of the ACES and Vulcan inventories for CO_2 (a, b) and the EDGAR v5.0 and GEPA inventories for CH_4 (c, d), for the d01 (a, c) and d03 (b, d) domains.

In those cases where the correlation length was equal to the base case value of 10 km, reducing the prior flux grid cell uncertainty to 50% resulted in mean posterior emission rates that were 3.5% and 2.5% lower than the base case for CO_2 and CH_4 respectively. Using a grid cell uncertainty equal to 200% of emissions increased the mean posterior emission rates by 4.1% for CO_2 and 3.1% for CH_4. Daily variability was 29% for both CO_2 and CH_4 in the 50% case, in comparison to 32% for CO_2 and 33% for CH_4 in the 200% case.

In cases where the prior grid cell uncertainty was equal to the base case value of 100%, using a prior flux covariance length of 4 km reduced the mean posterior emission rates by 2.9% for CO_2 and 2.7% for CH_4 relative to the base case, with a daily variability of 28% for CO_2 and 27% for CH_4. Using a 12 km length scale increased the mean posterior emission rates by 0.10% for CO_2 and 0.14% for CH_4 and resulted in a daily variability of 32% for CO_2 and 31% for CH_4.

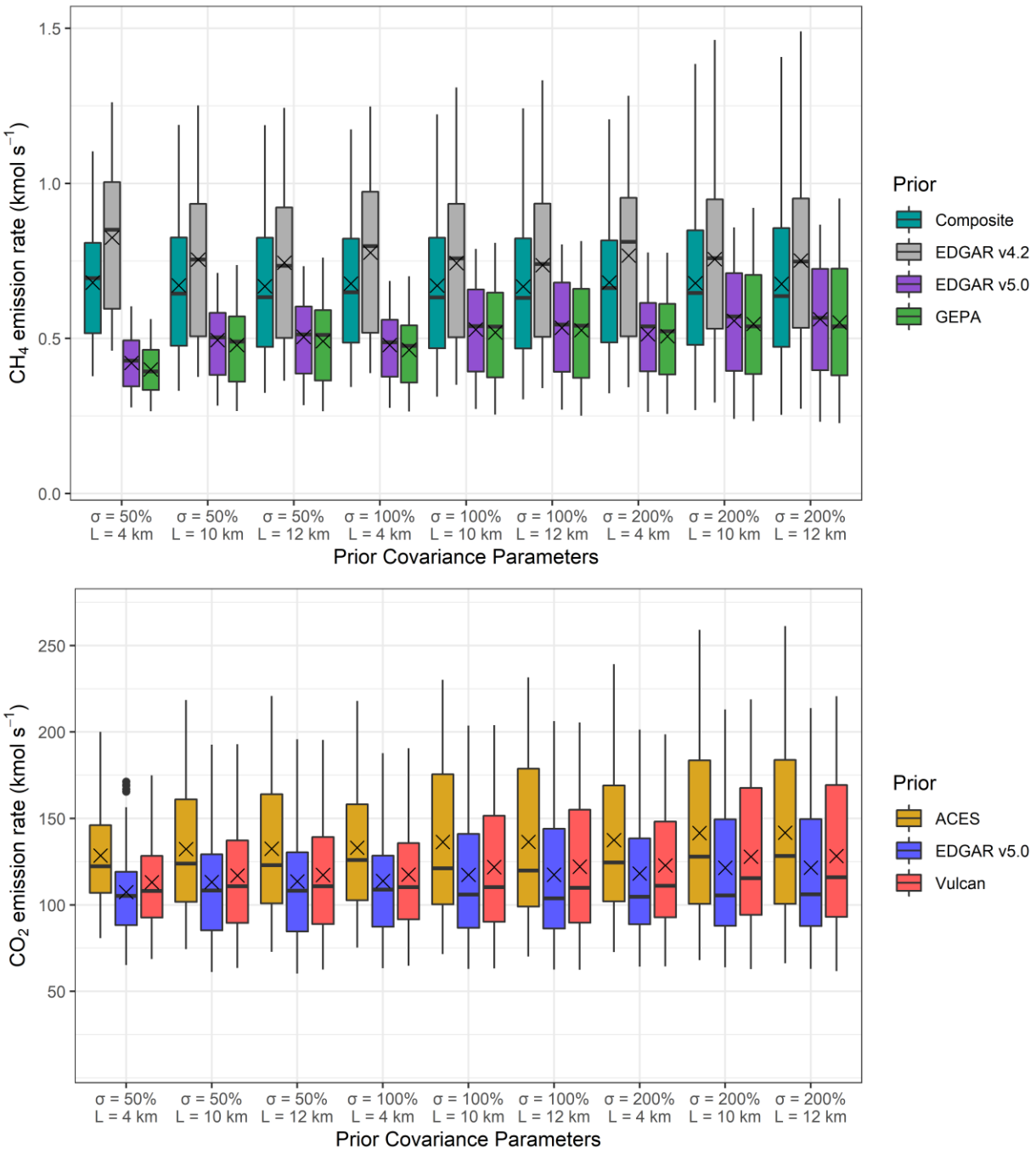

Figure S11.2: Boxplots showing total emission rates for the New York-Newark urban area for CH_4 (top panel) and CO_2 (bottom panel), using nine different combinations of the prior flux uncertainty and the spatial correlation length for prior flux errors. The values represented by the boxplots are the posterior emission rates across all flights and transport models for each prior flux covariance configuration. The mean posterior emission rates are shown as crosses. See Figure S5.1 caption for boxplot convention.

References:

Andres, RJ, Boden, TA, Higdon, DM. 2016. Gridded uncertainty in fossil fuel carbon dioxide emission maps, a CDIAC example. *Atmospheric Chemistry and Physics* **16**(23): 14979–14995. DOI: http://dx.doi.org/10.5194/acp-16-14979-2016.

Crippa, M, Solazzo, E, Huang, G, Guizzardi, D, Koffi, E, Muntean, M, Schieberle, C, Friedrich, R, Janssens-Maenhout, G. 2020. High resolution temporal profiles in the Emissions Database for Global Atmospheric Research. *Scientific Data* **7**(1): 121. DOI: http://dx.doi.org/10.1038/s41597-020-0462-2.

European Commission Joint Research Centre (JRC)/Netherlands Environmental Assessment Agency (PBL). 2011. Emission Database for Global Atmospheric Research (EDGAR), release version 4.2 [dataset]. Available at https://edgar.jrc.ec.europa.eu/overview.php?v=42. Accessed October 2, 2020.

European Commission Joint Research Centre (JRC)/Netherlands Environmental Assessment Agency (PBL). 2019. Emission Database for Global Atmospheric Research (EDGAR), release version 5.0 [dataset]. Available at https://edgar.jrc.ec.europa.eu/overview.php?v=50_GHG. Accessed October 2, 2020.

Fasoli, B, Lin, JC, Bowling, DR, Mitchell, L, Mendoza, D. 2018. Simulating atmospheric tracer concentrations for spatially distributed receptors: updates to the Stochastic Time-Inverted Lagrangian Transport model's R interface (STILT-R version 2). *Geoscientific Model Development* **11**(7): 2813–2824. DOI: http://dx.doi.org/10.5194/gmd-11-2813-2018.

Floerchinger, C, Shepson, PB, Hajny, K, Daube, BC, Stirm, BH, Sweeney, C, Wofsy, SC. 2021. Relative flux measurements of biogenic and natural gas-derived methane for seven U.S. cities. *Elementa: Science of the Anthropocene* **9**(1): 1–15. DOI: http://dx.doi.org/10.1525/elementa.2021.000119.

Gately, CK, Hutyra, LR. 2017. Large uncertainties in urban-scale carbon emissions. *Journal of Geophysical Research: Atmospheres* **122**(20): 11242–11260. DOI: http://dx.doi.org/10.1002/2017JD027359.

Gerbig, C, Lin, JC, Wofsy, SC, Daube, BC, Andrews, AE, Stephens, BB, Bakwin, PS, Grainger, CA. 2003. Toward constraining regional-scale fluxes of CO2 with atmospheric observations over a continent: 2. Analysis of COBRA data using a receptor-oriented framework. *Journal of Geophysical Research: Atmospheres* **108**(D24): 4757. DOI: http://dx.doi.org/10.1029/2003JD003770.

Gourdji, SM, Karion, A, Lopez-Coto, I, Ghosh, S, Mueller, K, Zhou, Y, Williams, C, Baker, I, Haynes, K, Whetstone, J. 2021. A modified Vegetation Photosynthesis and Respiration Model (VPRM) for the eastern USA and Canada, evaluated with comparison to atmospheric observations and other biospheric models. *Journal of Geophysical Research: Biogeosciences*, in press. DOI: http://dx.doi.org/10.1029/2021JG006290.

Gurney, KR, Liang, J, Patarasuk, R, Song, Y, Huang, J, Roest, G. 2020. The Vulcan version 3.0 high-resolution fossil fuel CO2 emissions for the United States. *Journal of Geophysical Research: Atmospheres* **125**(19). DOI: http://dx.doi.org/10.1029/2020JD032974.

Hanna, SR. 1982. Applications in air pollution modelling, in Nieuwstadt, FT, Van Dop, H eds. *Atmospheric turbulence and air pollution modelling*. Dordrecht, The Netherlands: Reidel. 275–310.

Kantha, LH, Clayson, CA. 2000. *Small scale processes in geophysical fluid flows*. San Diego: Academic Press.

Lauvaux, T, Miles, NL, Deng, A, Richardson, SJ, Cambaliza, MO, Davis, KJ, Gaudet, B, Gurney, KR, Huang, J, O'Keefe, D, Song, Y, Karion, A, Oda, T, Patarasuk, R, Razlivanov, I, Sarmiento, D, Shepson, P,

Sweeney, C, Turnbull, J, Wu, K. 2016. High-resolution atmospheric inversion of urban CO2 emissions during the dormant season of the Indianapolis Flux Experiment (INFLUX). *Journal of Geophysical Research: Atmospheres* **121**(10): 5213–5236. DOI: http://dx.doi.org/10.1002/2015JD024473.

Lin, JC, Gerbig, C, Wofsy, SC, Andrews, AE, Daube, BC, Davis, KJ, Grainger, CA. 2003. A near-field tool for simulating the upstream influence of atmospheric observations: The Stochastic Time-Inverted Lagrangian Transport (STILT) model. *Journal of Geophysical Research* **108**(D16): 4493. DOI: http://dx.doi.org/10.1029/2002JD003161.

Lopez-Coto, I, Ghosh, S, Prasad, K, Whetstone, J. 2017. Tower-based greenhouse gas measurement network design—The National Institute of Standards and Technology North East Corridor Testbed. *Advances in Atmospheric Sciences* **34**(9): 1095–1105. DOI: http://dx.doi.org/10.1007/s00376-017-6094-6.

Maasakkers, JD, Jacob, DJ, Sulprizio, MP, Turner, AJ, Weitz, M, Wirth, T, Hight, C, DeFigueiredo, M, Desai, M, Schmeltz, R, Hockstad, L, Bloom, AA, Bowman, KW, Jeong, S, Fischer, ML. 2016. Gridded national inventory of U.S. methane emissions. *Environmental Science & Technology* **50**(23): 13123–13133. DOI: http://dx.doi.org/10.1021/acs.est.6b02878.

Mahadevan, P, Wofsy, SC, Matross, DM, Xiao, X, Dunn, AL, Lin, JC, Gerbig, C, Munger, JW, Chow, VY, Gottlieb, EW. 2008. A satellite-based biosphere parameterization for net ecosystem CO 2 exchange: Vegetation Photosynthesis and Respiration Model (VPRM). *Global Biogeochemical Cycles* **22**(2): GB2005. DOI: http://dx.doi.org/10.1029/2006GB002735.

New York City Department of City Planning. 2020. Borough boundaries (water areas included) [dataset]. Available at https://www1.nyc.gov/assets/planning/download/zip/data-maps/open-data/nybbwi_20c.zip. Accessed November 13, 2020.

Plant, G, Kort, EA, Floerchinger, C, Gvakharia, A, Vimont, I, Sweeney, C. 2019. Large fugitive methane emissions from urban centers along the U.S. east coast. *Geophysical Research Letters* **46**(14): 8500–8507. DOI: http://dx.doi.org/10.1029/2019GL082635.

Sargent, M, Barrera, Y, Nehrkorn, T, Hutyra, LR, Gately, CK, Mckain, K, Sweeney, C, Hegarty, J, Wofsy, SC, Sargent, M, Barrera, Y, Nehrkorn, T, Hutyra, LR, Gately, CK, Jones, T. 2018. Anthropogenic and biogenic CO2 fluxes in the Boston urban region. *Proceedings of the National Academy of Sciences* **115**(40): E9507–E9507. DOI: http://dx.doi.org/10.1073/pnas.1815348115.

Turnbull, JC, Karion, A, Davis, KJ, Lauvaux, T, Miles, NL, Richardson, SJ, Sweeney, C, McKain, K, Lehman, SJ, Gurney, KR, Patarasuk, R, Liang, J, Shepson, PB, Heimburger, A, Harvey, R, Whetstone, J. 2019. Synthesis of urban CO2 emission estimates from multiple methods from the Indianapolis Flux Project (INFLUX). *Environmental Science & Technology* **53**(1): 287–295. DOI: http://dx.doi.org/10.1021/acs.est.8b05552.

United States Census Bureau. 2019. TIGER/Line® Shapefiles [dataset]. Available at https://www2.census.gov/geo/tiger/TIGER2019/UAC/tl_2019_us_uac10.zip. Accessed September 23, 2021.

Zhuang, J. 2020. xESMF: Universal Regridder for Geospatial Data. DOI: http://dx.doi.org/10.5281/zenodo.1134365.