Response to Reviewer Comments:

We thank the two Anonymous Reviewers for their thorough comments, particularly Reviewer #1 for catching an error in the presentation of the $\partial\eta/\partial n_s$ curve.

Reviewer #1 Comments:

The approach, fitting a statistical model to a sample of network designs, is novel but intuitive. While the detailed analysis could be more robust in some ways, the conclusions are a significant advance over the current state of knowledge and will have concrete value for the design of future observing systems, and I encourage publication once my concerns are address.

1.) Abstract could lead to false impressions: My primary concern the author's can address readily. As currently written, the abstract could lead to false impressions that this analysis definitively concludes that moderate cost sensors in a denser network is the optimal configuration for any urban area and that weekly CO_2 emissions with uncertainties of less than 5% can be achieved. These sections of the abstract in particular should be re-worked, as the authors actually are finding that with their specific modeling framework, higher density/moderate cost sensors provides an improved basis for flux estimation for the Bay Area. Further, given some of the assumption of diagonal error co-variance matrices, the representativeness of low altitude measurements in an urban region, and the gap presented by neglecting night-time data, the 5% monthly conclusion would appear to be an optimistic/idealized result and needs to be presented as such.

We have updated the text in the abstract to include more qualifiers.

2.) Clarity of presentation: Additionally, error in describing the background condition, or the CO_2 levels before impact of the urban region, have been found to be of high importance in other urban studies, and more discussion on the construction of this and assumptions used would be helpful. Otherwise my recommendations center around the clarity of presentation. In order to be of most use to a wider audience, including to researchers who may wish to perform similar analyses as they design networks in other cities, the methods need to be described more fully and precisely. Some justification should be provided for the choices and assumptions made in the analysis; I point out some examples below, but the authors should make a thorough review. The figures, especially figure 4, should be made more clear.

We have updated the text to include the reviewer's suggestions.

Minor Comments:

1.) How are the representativeness of observations made at just meters above the surface in a dense urban environment addressed? Depending on how these observation sites are

setup, they could be biased in their sampling to see traffic, people, or biosphere in a courtyard. This paper does not need to solve this problem, but it should be discussed as a potential additional source of bias error in 'cheap' network deployments, particularly as more sites are deployed (which can be very challenging to secure sites for deployment) and less ideal deployment locations are used. There is another component of this question, or way to frame it, which is a model designed to work at 1km will not be able to represent the sub-km variability sampled by a network not deployed to make observation representative of 1km areas, and thus potential biases might result.

The reviewer asks about the representativeness of observations made meters above the surface in a dense urban environment. However, as the reviewer notes, this is not a problem our paper aims (or needs) to address. This question is not relevant to a pseudo-data study like ours because our pseudo-data are generated using a model at 1-km resolution, so the pseudo-data are representative of a 1-km area. From a more experimental perspective, it is currently unknown whether sensitivity to local processes necessarily precludes the ability of surface-level sensors to represent domain-wide phenomena, as BEACO$_2$N is the first network with sufficient quantity and density of sensors to empirically investigate this trade-off. Future analyses of the real BEACO$_2$N dataset are positioned to answer these and related questions more quantitatively.

Regarding the potential additional source of bias error in 'cheap' network deployments, we have attempted to address this in the original manuscript through the "systematic bias" sensitivity test that was presented in Section 6.1 and Supplemental Section S6.2 where we added a systematic site-specific bias to each observational site in the network. This site-specific bias could be due to representation error, instrument error, etc. Additionally, the companion manuscript (Shusterman *et al.*, 2016) found the BEACO$_2$N sensors to detect weekly fluctuations in background concentrations to within ±2 ppm.

2.) The abstract should make clear that the statistical models estimate the uncertainty reduction as a function of the number of sites and the model-data mismatch. It should state that the study region is the Bay Area.

Presenting technical details of the statistical models, as the reviewer proposes, would make the abstract overly cumbersome. A reader that is interested in the statistical models will need to consult the main text. However, we have updated the abstract to make it clear that the study area is the Bay Area.

Lines 6-7: "modeled after the BEACO$_2$N network in California's Bay Area"

3.) Line 17-19: Need to specify that with this particular WRF-STILT framework and assumptions on error co-variance the moderate precision array is preferred.

We have updated the text:

Lines 17-18: "Using our inversion framework, we find that…"

4.) Line 19-21: This might technically be accurate, but is a bit misleading as some of the

assumption made here likely will cause issues with absolute flux accuracy that are much larger than 5%. No top-down method has ever been demonstrated to this point to have fidelity greater than 10% (some would argue demonstration of 20% has yet to be actually achieved).

Previous top-down studies have tended to focus on sparser networks and may have been site-limited. Kort *et al.* (2013) claim to constrain fluxes to 10% using a sparser network. Here is the final line of their abstract: "*We estimate that this network can distinguish fluxes on 8 week time scales and 10 km spatial scales to within ~12 g C* m^{-2} d^{-1} *(~10% of average peak fossil* CO_2 *flux in the LA domain).*" Additionally, we are discussing the constraints on the urban region (Area Source), which should be easier to constrain than 10 km grid cells due to it's large size. Given the differences in observational networks, our results are in line with previous estimates reported in the literature.

5.) Line 43: Clarify whether instruments and calibration approaches are mixed within individual networks, between networks, or both.

We have updated the text:

Lines 44-46: "Current monitoring networks use a variety of instruments and approaches to calibration with resulting variations in capital and operating costs, network precision, and potential instrument bias."

6.) Lines 69-70: Specify the temporal resolution of the BAAQMD inventory.

We have updated the text:

Lines 70-71: "The Bay Area Air Quality Management District (BAAQMD) provides detailed **annual** county-level CO_2 emissions information…"

7.) Lines 81-86: Make clear in this paragraph exactly what the FIVE product consists of. Is it a particular representative week of hourly emissions, which can be scaled by the user to fit other weeks? Or is scaled by McDonald et al. and provided for any week desired by the user?

We have updated the text:

Lines 82-85: "The FIVE traffic CO_2 inventory provides a representative week of hourly CO_2 emissions for San Francisco and other nearby Bay Area cities at 10 km, 4 km, 1 km, and 500 m resolution. This representative week can be scaled to different years based on the state fuel sales (see McDonald et al. (2014) for additional details)."

8.) Lines 88-91: As a simple approximation, could agricultural emissions be attributed uniformly to farmland? If this approximation is worse than omitting agricultural emissions entirely, state why.

For the purpose of the OSSE, this is inconsequential because we are using pseudo-observations. We are attempting to generate an inventory that is a reasonable approximation of the true emissions. The agricultural emissions are only 1% of the anthropogenic emissions, including them will have no impact on the results presented here.

Also, see the response to Reviewer #2's minor comment #3.

9.) Line 94-95: Please explain in more detail how you regrid from 1 degree to 1 km – this could be done in different manners.

We have added a detailed explanation to Supplemental Section S3.

10.) Line 102: The assumption of negligible diurnal cycle needs to be more thoroughly justified, especially since Nassar *et al.* (2013) emphasize the importance of diurnal variation.

Nassar *et al.* (2013) show that much of this diurnal variability is in the on-road mobile sector (see their Fig. 1). The other anthropogenic sectors (residential, industrial, electric vehicles, and commercial) show small diurnal variability (diurnal scale factor varies between 0.9 and 1.1). Our bottom up inventory includes diurnal variability from the on-road mobile sector.

11.) Lines 127-132: This description of STILT is confusing. It would be more clear to first explain how STILT is used to calculate influence footprints and only then to describe how the footprints are used to simulate CO_2 concentrations at the site locations.

We have updated the text and our description in Supplemental Section S1. See also our response to Reviewer #2's major comment #1.

12.) Lines 152-153: What does it mean that "the B matrix has an uncertainty of 100% at the native resolution?" One might take this to mean that the prior estimate is assigned a factor-of-two uncertainty. In supplemental section S2, it seems as though "100% uncertainty" means only that a multiplicative factor f_σ is introduced and then set to one.

It means the uncertainty is equal to the standard deviation of the hi-res inventory. We initially considered using other scale factors (f_σ) but ultimately settled on a multiplicative factor of one. However, we left this in the supplement for future studies to potentially modify.

13.) Lines 155-158: The impact on the result of the choice of a diagonal R matrix should be described.

Using a diagonal **R** matrix means that all of our model-data mismatch errors are uncorrelated. As such, using a non-diagonal **R** matrix will mean there is less information in the observations because the observations are not independent. We have used a diagonal **R** matrix because the real observations in the $BEACO_2N$ network are made at 1 Hz. Here we use hourly observations. As such, it seems fair to assume the original 1 second

observations are independent when aggregated to 1 hour. A study with real data could look at the autocorrelation of the model-data mismatch to infer a proper decorrelation length scale, which will almost certainly be much less than 1 hour.

We have updated the text:

Lines 152-153: "Using a diagonal **R** matrix means that we have assumed our mismatch errors are uncorrelated."

14.) Line 173: Why not just use the posterior flux error, which is more intuitive and which is shown in the key figure (figure 4)? Why is it an advantage to use a metric similar in form to the averaging kernel matrix?

The posterior flux error is an absolute metric and, as such, is less generalizable. The error metric chosen here is more similar to previous work (e.g., Kort *et al.*, 2013) and more generalizable to other studies.

15.) Lines 184-185: The single sentence "We vary the number of sites (n_s) and mismatch error (σ_m) and perform an ensemble of 20 inversions for each combination to ensure the results are robust." is not adequate to explain this key step in the analysis. For how many different combinations of n_s and σ_m was the error calculated? Which combinations? How were site locations chosen for non-maximal n_s? What differs between the 20 inversions performed for the same combination of parameters: the choice of site locations, the random errors, the STILT footprint calculation?

We have updated the text:

Lines 195-202: "Fig. 4 shows the error in the estimated CO_2 fluxes using the observations over a wide range of observing system scenarios. We vary the number of sites (n_s = [1, 2, …, 34]), mismatch error (σ_m = [0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20] ppm), and perform an ensemble of 20 inversions for each combination to ensure the results are robust. Each ensemble member uses a unique observational network by randomly drawing n_s sites from the population of 34 possible sites. In total, we perform 8,160 inversions. Fig. 4 shows the mean error in the estimated CO_2 fluxes for the area source, line source, and point source as a function of σ_m and n_s. This figure represents the uncertainty in the estimated emissions at a given hour."

16.) Line 193: The first two parameters are motivated by the assumption of Gaussian errors; what motivates the choice of the other five parameters?

The x-axis is a linear scale and the y-axis is a log-scale and we can qualitatively see structure in the figure. As such, we assumed that linear and log relationships could yield a significant relationship.

17.) Lines 208-218: This critical part of the procedure is not clear. The derivative of η with respect to n_s expresses the error reduction to be obtained by adding additional sites. In order

to say whether a particular network configuration is noise-limited or site-limited, this reduction should be compared to the reduction to be obtained by reducing the mismatch error, expressed by the derivative of η with respect to σ_m. But the latter is never calculated. Furthermore, it's not clear what is meant by "the $\partial\eta/\partial n_s$ curve," how such a curve can be plotted on axes neither of which corresponds to that derivative (as in Figure 4), or what it means for a particular system to be above or below the curve.

We appreciate the reviewer for catching this mistake. We no longer use the derivatives to estimate these regimes. We now estimate them as the ridgeline from the statistical models.

18.) Lines 232-237: This is an important point and should be explained more clearly. Exactly what fluxes do you estimate in the averaged case, and what errors are you comparing? Precisely what does "10× better" mean? In the figure it looks as though the errors do not decrease as quickly as predicted by the CLT but seem to level off after about 96 hours; you might explain why this is to be expected.

The errors in emissions aggregated over 1 week are about a factor of 10 less than the error in hourly emissions. We have updated the text:

Lines 243-249: "Our work is primarily focused on estimating hourly fluxes, however we can further reduce the uncertainty in our estimates by considering temporally averaged fluxes (e.g., what are the weekly or monthly emissions?). Fig. 5 shows the error in our estimate of the area source emissions aggregated over various time-scales. We find the error in our estimate greatly decreases over the first 72 hours. The central limit theorem provides a lower bound on the error reduction we might expect and the error reductions follow this limit reasonably well over the first 72 hours. This implies that our weekly-averaged emission estimate would be 10× better than our hourly emission estimate."

19.) Line 273-274: How is this statement about the large systematic error consistent with the 5% uncertainty conclusion highlighted in the abstract?

The 5% uncertainty is referring to the case shown in the main text (case without the imposed systematic bias).

20.) Section 6.1: Since the text of the supplemental section S5 contains little additional information, consider integrating it into the main text, possibly combining figures S6-S8. Also, specify whether all the observing systems tested in Section 4, or only a subset, were included in the test of sensitivity to domain size.

We feel that this content is better suited to the supplement.

21.) Section 7: The conclusions should include at least some description of which system designs were found to be site-limited and which noise-limited, since that information is of immediate use to other researchers designing or evaluating their own networks.

We have refrained from putting that in the Conclusions section because it is, somewhat, dependent on the specification of transport error. We allude to this in the conclusion when we qualify one of our findings with the following statement: "*if the dominant source of error is instrument precision.*"

22.) Figure 3: The color scheme in the top row is not intuitive to perceive, especially at low resolution as in panel a. Is white used for no estimate as well as for zero flux?

It is unclear what "*low resolution*" the reviewer is referring to, all of the fluxes are plotted at the same resolution (1 km^2). As for the color scheme, we tested many different color schemes and settled on this one because it facilitated comparison between the different panels (other color schemes were much worse).

White indicates a flux of zero. There is no "*no estimate*", all panels have fluxes at all grid cells.

23.) Figure 4: This figure is crucially important, and the design is generally good. However, the shading needs to be reworked so that the gradient is more visible. Also, as mentioned above, it's not clear what defines the red line that separates noise- from site-limited regimes.

We have updated the coloring and the method for defining noise- and site-limited regimes.

24.) Supplement line 56: Why were the decay parameters chosen as they were?

The decay parameters were judiciously chosen through discussion with the co-authors who have experience creating the bottom-up inventory (Brian McDonald and Robert Harley). Future work could include these parameters in the inversion by defining them as hyperparameters. However, we would no longer have a conjugate prior and would need to move to a sampling approach to obtain the posterior. Given the large number of inversions performed in this study (32,640 inversions), this would be computationally infeasible.

25.) Figure S2: Four judiciously chosen panels would probably be sufficient and could be shown at a larger size.

This is a vector graphics image in the Supplement. Readers should be able to zoom in on the panels without losing quality.

26.) Figure S3: Panels c-e are not as informative and could be omitted.

We disagree with the reviewer on this point. The usefulness of these panels varied quite a bit depending on the audience. We have found these panels to be useful for explaining the methodology to scientists who do not typically construct state vectors themselves. However, the panels are not crucial to the manuscript, which is why we put them in the Supplement.

27.) Figures S6-S8: As in Figure 4, the gradient is not visible enough. Also, the left column corresponds to main text Figure 4, not Figure 3.

We thank the reviewer for pointing out the incorrect labeling.

28.) Appendix A: In my opinion, this table is not necessary.

We feel that his table could be useful to some readers interested in the robustness of the different statistical models, which is why we have included it in the Supplement.

Reviewer #2 Comments:

It is certainly a novel piece of work and is beneficial to other urban measurement network designs and associated studies. However, some parts of the manuscript need improvements or additional details to better understand the results and their interpretations. Also some clarifications are necessary to improve the manuscript (see the comments below). Hence I would recommend this manuscript for publication after addressing my concerns and comments listed below.

1.) Footprint calculations: My major concern is about the footprint calculations presented in the manuscript. As far as I understand, what it is shown in Fig. 2 is the averaged footprints for all sites in the network in which the footprints are calculated separately for each sites. In that case, I am surprised with such a low value for the averaged footprints on the western side of the model domain even if there are many sites (especially the line source/high-way is on that side, Fig. 1). Although a part of this can be explained with the prevailed wind direction, I don't find enough reasons to justify the shown structure. i.e., it is difficult to believe that those sites don't give much information on surface fluxes for this period. Please clarify and also give additional details (e.g. set up of STILT receptor locations, how strong is the advection, details of vertical mixing etc.).

The reviewer thinks the observations should have more sensitivity to the "*western side of the model domain*". This may stem from how other work has presented the footprints. Much of the previous work showing footprints has plotted them on a log-scale (e.g., Lin *et al.*, 2003, 2004, 2007; Kort *et al.*, 2008) or as percentiles (e.g., Miller *et al.*, 2012, 2013; McKain *et al.*, 2015). We have added a Supplemental Figure (Fig. S3) that shows the footprint on a linear-scale and log-scale. The largest footprint values are in locations where we have an observation site, however there are diffuse signals that cover much of the domain and have a non-trivial contribution to the total signal. Further, the spatial footprints found here are broadly consistent with previous work by Bastien *et al.* (2015; their Fig. 2a-e) who used an adjoint model to determine the sources influencing air pollution in California's East Bay.

As for the request for additional details, it seems that the reviewer may have missed or overlooked some of the model description. Most of what the reviewer requested is already included in the manuscript or supplement. Specifically, the STILT receptor locations

(latitude, longitude, and height above ground level) are listed in Table 1 and the details of the WRF simulations (including PBL and LSM schemes) are included in Supplemental Section 1. The advection and vertical mixing are, largely, determined by the WRF model. The cited literature (Lin et al., 2003; Nehrkorn et al., 2010) describe the advection, vertical mixing, and coupling of STILT to WRF in exhaustive detail.

2.) Inverse framework is not well explained: Another criticism is that the inverse framework, although it is a critical component of this study, is not well explained (Sect.4). For example, it is not very clear to me how the state vector is defined for this experiment. What is the spatial and temporal resolutions of the posterior fluxes? This is important to follow the inversion results. This section needs major improvement w.r.t giving additional details.

We have updated the text in Section 4. It now includes a more detailed explanation of the inverse framework including a paragraph describing the state vector.

3.) Reported error estimate: The reported error estimate of the posterior fluxes (5%) is for the best case OSSE and the inversion experiment (rather I would say that it is for "the most idealized case") in which the total model-data mismatch error is assumed to be 0.005 ppm. Since this mismatch error is totally unrealistic in the current scenario, it is not fair to include this "best case" result in the abstract unless the model-data mismatch error (+ other assumptions) is explicitly specified here. Since it is misleading, I would recommend authors to either remove this sentence or provide an error estimate for more reasonable scenario.

It seems that this comment stems from a misunderstanding in the reported errors; our abstract does not report errors using the 0.005 ppm case. The reported error estimate of the posterior fluxes (5%) is for "Network A" (dense network with moderate-precision instruments) from Section 6.2, not the most optimistic case with 0.005 ppm. Fig. 3 shows the most optimistic case because there is only one combination of sites for that case, whereas networks with less sites have multiple configurations that could be shown.

We have updated the abstract to clarify this:

Line 20: "The dense network considered here (modeled after the BEACO$_2$N network)…"

3.) Fig. 3 and associated statements: I can't see a remarkable performance of inversion in retrieving posterior fluxes as one would expect here, given that the inversion uses a loose prior (100% uncertainty), used all 34 sites, and "unrealistically" low mismatch error (=0.005 ppm which includes model error, representation error, and instrument error). The spatial structure in the CO_2 fluxes is captured only for a few parts of the domain. Unfortunately, this says to me that the most of other sites are not much useful in this case, which is hard to believe. This again points back to my concern regarding the footprint calculation. Need to clarify.

It is unclear which "*associated statements*" the reviewer is referring to. Fig. 3 is merely presented as an example of the posterior fluxes. One of the few statements we make about Fig. 3 is in reference to the diurnal cycle in Fig. 3: "*We find substantial improvements in the diurnal cycle (see panel d).*"

As for the comment about the unremarkable performance of the inversion, we are surprised by this comment. We find the improvement, relative to the prior, quite remarkable. Panel d in Fig. 3 highlights this. The diurnal cycle in the prior is completely incorrect (wrong magnitude and out of phase) but the posterior is able to recover the true diurnal cycle. The spatial structure is recovered reasonably well in the region where we have measurements sites (California's East Bay). Not surprisingly, we see little improvement in distant regions because we do not have sites located there (so the observations are unable to constrain those regions). The results follow pretty much exactly with what one might intuitively expect.

Minor Comments:

1.) L24: Radiative forcing is variable over the years. Please give the value w.r.t year. 1.82 W m^{-2} looks more like the 2011 year values.

We have updated the text.

Line 25: "with a radiative forcing of 1.82 W m^{-2} in 2011 relative to preindustrial times (IPCC 2013)."

2.) L60: The issue is not only with the spatial resolution, but also with the large uncertainty ranges (reported or expected). This issue needs to be addressed clearly in the manuscript to draw the importance of the high resolution inversion modeling, which is to reduce the uncertainty of the emission fluxes. Also mention about the temporal resolution. This is also important especially when cities have peak traffic, industrial, or commercial hours. Need to be mentioned/addressed in the manuscript.

We have updated the text.

Lines 28-29: "…yet current bottom-up inventories still have large uncertainties."

3.) L92-95: From Fig.1 (bottom panel), I see that the natural sources accounts for about 17% (peak to peak, according to CT2013B) of the total fluxes and are varying as expected. This is considerable in comparison with the Bay area traffic sources which accounts for ~50% of the total fluxes. Hence I would expect that using the natural fluxes at coarse resolution (1×1) can generate additional uncertainty and may not be appropriate in this high resolution modeling scenario. Please comment on this.

While a good point, it's not really relevant here because we are performing an OSSE. As such, we have two main goals in constructing the bottom-up inventory: (1) create a bottom-up inventory that is a reasonable approximation of the true emissions and (2) create a bottom-up inventory that is fundamentally different from the prior inventory. For the former goal, the CarbonTracker natural fluxes should provide a reasonable approximation to the true diurnal cycle, albeit with coarse spatial resolution, while the anthropogenic inventory provides high spatio-temporal information about the urban region. Therefore, our bottom-up

information should be a decent approximation of the true emissions. As for the latter goal, we are interested in learning what an observational network could tell us about the emissions. So we are using fundamentally different bottom-up inventories to generate the pseudo-observations and serve as the prior for the inversion.

4.) Fig.1: What is "other Anthro" (red line) based on?

We have updated the caption:

Fig. 1 Caption: "Other anthropogenic sources in the BAAQMD inventory (red)."

5.) Section 4: This section needs further improvements to better explain the inversion technique used in this study. Please modify. Also indicate the dimension of "m" and "n".

The dimensions of m was presented in Supplemental Section S2: "m = 2,133,120, m_t = 240, m_x = 88, and m_y = 101." We now also included this in the main text Section 4:

Lines 144-147: "The resulting state vector has 2,133,120 elements ($m = m_t \bullet m_x \bullet m_y$ with m_t = 240, m_x = 88, and m_y = 101) and the posterior fluxes will have hourly temporal resolution and 1 km^2 spatial resolution. The dimension of n will depend on the number of sites in the observational network."

6.) Mathematical formulas (e.g. Sect. 4): Please use standard formatting as followed by the most of the authors/textbooks. For e.g. prior fluxes, x_b in which "b" is subscript.

We have updated the notation.

References:

Bastien, L. A., McDonald, B. C., Brown, N. J., & Harley, R. A. (2015). High-resolution mapping of sources contributing to urban air pollution using adjoint sensitivity analysis: benzene and diesel black carbon. Environ Sci Technol, 49(12), 7276-7284. doi: 10.1021/acs.est.5b00686

Kort, E. A., Eluszkiewicz, J., Stephens, B. B., Miller, J. B., Gerbig, C., Nehrkorn, T., Wofsy, S. C. (2008). Emissions of CH4 and N2O over the United States and Canada based on a receptor-oriented modeling framework and COBRA-NA atmospheric observations. Geophysical Research Letters, 35(18). doi: 10.1029/2008gl034031

Kort, E. A., Angevine, W. M., Duren, R., and Miller, C. E.: Surface observations for monitoring urban fossil fuel CO2 emissions: Minimum site location requirements for the Los Angeles megacity, Journal of Geophysical Research: Atmospheres, 118, 1577–1584, doi:10.1002/jgrd.50135, 2013.

Lin, J. C., Gerbig, C., Wofsy, S. C., Andrews, A. E., Daube, B. C., Davis, K. J., and Grainger, C. A.: A near- field tool for simulating the upstream influence of atmospheric observations: The Stochastic Time-Inverted Lagrangian Transport (STILT) model, Journal of Geophysical Research-Atmospheres, 108, doi:10.1029/2002jd003161, 2003.

Lin, J. C., Gerbig, C., Wofsy, S. C., Andrews, A. E., Daube, B. C., Grainger, C. A., Hollinger, D. Y. (2004). Measuring fluxes of trace gases at regional scales by Lagrangian observations: Application to the CO2 Budget and Rectification Airborne (COBRA) study. Journal of Geophysical Research-Atmospheres, 109(D15). doi: 10.1029/2004jd004754

Lin, J. C., Gerbig, C., Wofsy, S. C., Chow, V. Y., Gottlieb, E., Daube, B. C., & Matross, D. M. (2007). “Designing Lagrangian experiments to measure regional-scale trace gas fluxes”. Journal of Geophysical Research, 112(D13). doi: 10.1029/2006jd008077

McDonald, B. C., McBride, Z. C., Martin, E. W., and Harley, R. A.: High-resolution mapping of motor vehicle carbon dioxide emissions, Journal of Geophysical Research-Atmospheres, 119, 5283–5298, doi:10.1002/2013jd021219, 2014.

McKain, K., Down, A., Raciti, S. M., Budney, J., Hutyra, L. R., Floerchinger, C., Wofsy, S. C. (2015). Methane emissions from natural gas infrastructure and use in the urban region of Boston, Massachusetts. Proc Natl Acad Sci U S A, 112(7), 1941-1946. doi: 10.1073/pnas.1416261112

Miller, S. M., Kort, E. A., Hirsch, A. I., Dlugokencky, E. J., Andrews, A. E., Xu, X., Wofsy, S. C. (2012). Regional sources of nitrous oxide over the United States: Seasonal variation and spatial distribution. Journal of Geophysical Research, 117(D6). doi: 10.1029/2011jd016951

Miller, S. M., Wofsy, S. C., Michalak, A. M., Kort, E. A., Andrews, A. E., Biraud, S. C., Sweeney, C. (2013). Anthropogenic emissions of methane in the United States. Proc Natl Acad Sci U S A, 110(50), 20018-20022. doi: 10.1073/pnas.1314392110

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 CO2 emissions from global fossil fuel emission data sets, Journal of Geophysical Research: Atmospheres, 118, 917–933, doi:10.1029/2012jd018196, 2013.

Nehrkorn, T., Eluszkiewicz, J., Wofsy, S. C., Lin, J. C., Gerbig, C., Longo, M., and Freitas, S.: Coupled weather research and forecasting stochastic time-inverted lagrangian transport (WRF–STILT) model, Meteorology and Atmospheric Physics, 107, 51–64, doi:10.1007/s00703-010-0068-x, 2010.

Shusterman, A. A., Teige, V., Turner, A. J., Newman, C., Kim, J., and Cohen, R. C.: The BErkeley Atmospheric CO2 Observation Network: initial evaluation, Atmospheric Chemistry and Physics Discussions, pp. 1–23, doi:10.5194/acp-2016-530, 2016.

A. J. Turner's manuscript prepared
with version 4.2 of the LaTeX class copernicus.cls.
Date: 29 August 2016

Network design for quantifying urban CO_2 emissions: Assessing trade-offs between precision and network density

Alexander J. Turner[1,2], Alexis A. Shusterman[3], Brian C. McDonald[4,*], Virginia Teige[3], Robert A. Harley[4], and Ronald C. Cohen[3,5]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA.
[2]Environmental Energy and Technologies Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
[3]Department of Chemistry, University of California at Berkeley, Berkeley, CA, USA.
[4]Department of Civil and Engineering, University of California at Berkeley, Berkeley, CA, USA.
[5]Department of Earth and Planetary Sciences, University of California at Berkeley, Berkeley, CA, USA.
[*]now at: Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado, USA.

Correspondence to: Ronald C. Cohen
(rccohen@berkeley.edu)

Abstract. The majority of anthropogenic CO_2 emissions are attributable to urban areas. While the emissions from urban electricity generation often occur in locations remote from consumption, many of the other emissions occur within the city limits. Evaluating the effectiveness of strategies for controlling these emissions depends on our ability to observe urban CO_2 emissions and attribute them to specific activities. Cost effective strategies for doing so have yet to be described. Here we characterize the ability of a prototype measurement network, modeled after the BEACO$_2$N network in California's Bay Area, in combination with an inverse model based on WRF-STILT to improve our understanding of urban emissions. The pseudo-measurement network includes 34 sites at roughly 2 km spacing covering an area of roughly 400 km^2. The model uses an hourly 1×1 km^2 emission inventory and 1×1 km^2 meteorological calculations. We perform an ensemble of Bayesian atmospheric inversions to sample the combined effects of uncertainties of the pseudo-measurements and the model. We vary the estimates of the combined uncertainty of the pseudo-observations and model over a range of 20 ppm to 0.005 ppm and vary the number of sites from 1 to 34. We use these inversions to develop statistical models that estimate the efficacy of the combined model-observing system at reducing uncertainty in CO_2 emissions. We examine uncertainty in estimated CO_2 fluxes at the urban scale, as well as for sources embedded within the city such as a line source (e.g., a highway) or a point source (e.g., emissions from the stacks of small industrial facilities). ~~We~~ Using our inversion framework, we find that a dense network with moderate precision is the preferred setup for estimating area, line, and point sources from a combined uncertainty and cost perspective. The

dense network considered here (modeled after the BEACO$_2$ N network) could estimate weekly CO_2 emissions from an urban region with less than 5% error, given our characterization of the combined observation and model uncertainty.

1 Introduction

Carbon dioxide (CO_2) is an atmospheric trace gas and the single largest anthropogenic radiative forcer, with a radiative forcing of 1.82 W m^{-2} ~~since~~ in 2011 relative to preindustrial times (IPCC, 2013). CO_2 has increased from 280 ppm in preindustrial times to greater than 400 ppm in the present, largely due to changes in fossil fuel emissions. Over 70% of these fossil fuel CO_2 emissions in the United States (US) are attributable to urban areas (EIA, 2015; Hutyra et al., 2014), yet current bottom-up inventories still have large uncertainties. As such, quantifying and monitoring the emissions from urban areas is crucial to strategies for reducing future increases in CO_2.

Numerous studies have performed top-down estimations of CO_2 emissions using observations from urban surface monitoring networks of various sizes (e.g., Gratani and Varone, 2005; McKain et al., 2012; Newman et al., 2013; Lauvaux et al., 2013; Breon et al., 2015; Turnbull et al., 2015). However, it's not immediately clear how many sites are necessary to monitor the emissions from an urban area. Kort et al. (2013) found that a surface monitoring network would need at least 8 sites operating for 8 weeks to accurately estimate CO_2 emissions in Los Angeles. Yet most current urban monitoring networks have fewer than 8 sites but operate for much longer than 8 weeks. For example, Gratani and Varone (2005) used a single site in Rome, Newman et al. (2013) used a single site in Los Angeles, Lauvaux et al. (2013) used two sites in Davos, Switzerland, McKain et al. (2012) used a network of 5 sites in Salt Lake City, and Breon et al. (2015) used 5 sites in Paris. Recent work from Turnbull et al. (2015) employed a denser network of 12 sites in Indianapolis.

This issue is further complicated by bias and noise in both the measurements and the modeling framework. The combined model and measurement error is known as the model-data mismatch error (hereafter referred to as the "mismatch error"). Current monitoring networks use a mix of instruments and approaches to calibration with resulting variations of capital and operating costs, network precision, and potential instrument bias. Monitoring networks located in regions with complex orography are challenging for atmospheric transport calculations, making it more difficult to determine the dispersion from sources.

The tradeoff between measurement network density and mismatch error has yet to be characterized. Understanding these tradeoffs is crucial to reducing the uncertainty in emissions from urban regions and to developing cost-effective urban monitoring networks. Here we present a high-resolution inventory of CO_2 fluxes and a numerical model that relates atmospheric observations to high resolution surface fluxes. We then use this inventory and model in a series of observing system simulation experiments (OSSEs) to investigate the tradeoff between reductions in the mismatch error

and increases in the measurement network density. We develop statistical models to characterize this relationship for different types of sources in the San Francisco Bay Area, identify limiting regimes, and recommend future observing strategies.

2 Constructing a high resolution regional CO_2 inventory

McDonald et al. (2014) demonstrated that 1×1 km^2 spatial resolution is necessary to resolve the gradients in urban CO_2 fluxes from highways. However, most of the existing CO_2 anthropogenic inventories are not available at this resolution. For example, EDGAR (European Commission, 2011) and VULCAN (Gurney et al., 2009) are only available at $0.1^\circ \times 0.1^\circ$ and 10×10 km^2, respectively. A notable exception is the Odiac fossil fuel CO_2 inventory (Oda and Maksyutov, 2011) which is based on satellite-observed nightlight data and available globally at 1×1 km^2 resolution. High resolution fossil fuel CO_2 emissions are available for select cities and sectors such as Paris through the AirParif inventory (Breon et al., 2015, http://www.airparif.asso.fr/en/index/index) and Indianapolis, Los Angeles, Salt Lake City, and Phoenix through the HESTIA project (Gurney et al., 2012, http://hestia.project.asu.edu/); three recent studies (Gately et al., 2013; McDonald et al., 2014; Gately et al., 2015) developed high resolution CO_2 emissions from vehicular traffic.

The Bay Area Air Quality Management District (BAAQMD) provides detailed annual county-level CO_2 emissions information for San Francisco and California's Bay Area (Mangat et al., 2010). The BAAQMD found that the transportation sector accounted for 36% of the Bay Area anthropogenic emissions, industrial and commercial for 36%, electricity for 16%, residential fuel usage for 7%, off-road equipment for 3.0%, and agriculture for 1%. The BAAQMD also reports CO_2 emissions for 4,375 point sources in the Bay Area. We geocode these point sources based on the addresses provided by the BAAQMD. These point sources capture the emissions from the industrial, commercial, and electricity sectors. We map residential fuel usage to population using block level population data from the 2010 US Census and apply a temporal temperature scaling based on Deschłnes and Greenstone (2011); the resulting temporal scaling effect is small due to the temperate climate in the East Bay region of the SF Bay Area.

Here we use the traffic CO_2 emissions from the fuel-based inventory for vehicle emissions (FIVE) developed by McDonald et al. (2014). The FIVE traffic CO_2 inventory provides a representative week of hourly CO_2 emissions for San Francisco and other nearby Bay Area cities at 10 km, 4 km, 1 km, and 500 m resolution. This representative week can be scaled to different years based on the state fuel sales (see McDonald et al. (2014) for additional details). The FIVE inventory is constructed by partitioning CO_2 emissions using state-level fuel data to individual roads with road-specific traffic count data and temporal patterns from weigh-in-motion data. In this manner, CO_2 emissions from the FIVE inventory will be consistent with state and national CO_2 budgets and can easily be scaled to different years.

Combining the industrial, commercial, electricity, residential, and traffic emissions account for 95.8% of the anthropogenic CO_2 emissions in the Bay Area. We do not have high resolution proxy data for the off-road equipment or agriculture sectors in the Bay Area and have chosen to assume their contributions are smaller than the uncertainty in the total budget; therefore we neglect these sectors in the construction of our inventory.

CarbonTracker CT2013B (http://www.esrl.noaa.gov/gmd/ccgg/carbontracker/; Peters et al., 2007) provides 3 hourly fossil fuel, ocean, biogenic, and fire CO_2 fluxes at $1^\circ \times 1^\circ$ resolution. These fluxes are optimized to agree with atmospheric CO_2 observations. We regrid these fluxes to 1×1 km^2 spatial resolution (see Supplemental Section S3) and use the fire, ocean, and biogenic sectors to account for our natural fluxes.

Fig. 1 shows snapshots of the CO_2 fluxes from our inventory at 4 different times of day and the a-temporal fluxes from EDGAR v4.2 FT2010 (European Commission, 2011). From Fig. 1 we can see the inventory clearly resolves the large CO_2 gradients from highways, confirming that 1×1 km^2 spatial resolution is sufficient to resolve urban CO_2 fluxes from highways. The bottom panel of Fig. 1 shows a time series of Bay Area CO_2 fluxes broken down by source. The diurnal cycle in our inventory is largely driven by the traffic emissions with modest uptake from the biosphere during the middle of the day. Other anthropogenic sources were assumed to have a negligible diurnal cycle (Nassar et al., 2013). In what follows, we use EDGAR as the prior and the high spatio-temporal resolution inventory as the "truth".

[Fig. 1 about here.]

3 The Berkeley Atmospheric CO_2 Observation Network (BEACO$_2$N)

The Berkeley Atmospheric CO_2 Observation Network ~~("BEACO$_2$N", see http://beacon.berkeley.edu)~~ ("BEACO$_2$N", see http://beacon.berkeley.edu and Shusterman et al., 2016) was founded in 2012 as a web of approximately 25 carbon dioxide sensing "nodes" stationed atop schools and museums in the Oakland, CA metropolitan area (see Table 1). With sensors installed on an approximately 2 km square grid, BEACO$_2$N is the only surface-level (3 to 130 m a.g.l.) greenhouse gas monitoring system with roughly the same spatial resolution as the emissions inventories described above. Each node requires only a standard, 120V power source and is sited on pre-existing structures based on voluntary, no-cost partnerships. The BEACO$_2$N configuration therefore represents a reasonable expectation and is one model for future monitoring networks aimed at constraining CO_2 fluxes at neighborhood scales within an urban dome.

[Table 1 about here.]

BEACO$_2$N's unprecedented spatial density is achieved by exploiting lower cost instrumentation than has traditionally been utilized for ambient CO_2 detection. The non-dispersive infrared (NDIR)

absorption sensor used in each BEACO$_2$N node (http://www.vaisala.com/en/products/carbondioxide/Pages/GMP343.aspx) has been seen to possess adequate sensitivity to resolve diurnal as well as seasonal phenomena relevant to urban environments (Rigby et al., 2008) and costs one to two orders of magnitude less than the commercial cavity ring-down instruments commonly used in other networks. However, the low-cost NDIR sensor is more susceptible to factors such as temporal drift and environmental instability that can negatively impact data quality. This trade-off between mismatch error and network density is explored below.

4 Observing system simulation experiments

CO_2 concentrations were simulated at 34 sites in the BEACO$_2$N network with the Stochastic Time-Inverted Lagrangian Transport (STILT) model (Lin et al., 2003), coupled to the Weather Research and Forecasting (WRF) meso-scale meteorological model run at 1×1 km^2 grid resolution (WRF-STILT; Nehrkorn et al., 2010). WRF-STILT computes footprints (Δ CO_2 per surface flux, or ppm per μmol·m^{-2}·s^{-1}; See Supplemental Section S1 and Lin et al. (2003) for additional details) for each observation that relate the hourly 1 km^2 CO_2 fluxes ($\mathbf{x}$; an $m\times1$ vector) to the observations ($\mathbf{y}$; an $n\times1$ vector):

$$\mathbf{y} = \mathbf{H}\mathbf{x} \tag{1}$$

Each row of the $n\times m$ Jacobian matrix ($\mathbf{H} = \partial\mathbf{y}/\partial\mathbf{x}$) is a reshaped footprint. Fig. 2 shows the location of the sites and the average network footprint for Sept 15 to 22.

[Fig. 2 about here.]

Our aim is to estimate hourly CO_2 fluxes at 1 km^2 over a one week period. As such, the model domain is 88 km $\times$ 101 km and we solve for 240 hours of fluxes (1 week plus 3 additional days of back trajectories). The resulting state vector has 2,133,120 elements ($m = m_t \cdot m_x \cdot m_y$ with $m_t = 240$, $m_x = 88$, $m_y = 101$) and the posterior fluxes will have hourly temporal resolution and 1 km^2 spatial resolution. The dimension of n will depend on the number of sites in the observational network.

Here we use our high resolution CO_2 inventory (~~$\mathbf{x}^a$~~ $\mathbf{x}^\star$; an $m\times1$ vector) to generate synthetic observations (~~$\mathbf{y}^a$~~ $\mathbf{y}^\star$; an $n\times1$ vector):

$$\mathbf{y}^{a\star} = \mathbf{H}\mathbf{x}^{a\star} + \varepsilon \tag{2}$$

where ε is an $n\times1$ vector of normally distributed noise with mean $\boldsymbol{\epsilon}_b$ and diagonal covariance matrix $\mathbf{R}$: $\varepsilon \sim \mathcal{N}(\boldsymbol{\epsilon}_b, \mathbf{R})$. Using a diagonal $\mathbf{R}$ matrix means that we have assumed our mismatch errors are uncorrelated. Our base case inversion assumes the mean bias is zero: $\boldsymbol{\epsilon}_b = \mathbf{0}$. We evaluate the sensitivity to this assumption in Section 6 and Supplemental Section ~~S5~~S6.2. These synthetic observations can then be used in a Bayesian inference framework to estimate the optimal CO_2 fluxes (c.f.

Rodgers, 2000). Assuming the prior and likelihood distributions are Gaussian gives us a closed-form solution for the posterior CO_2 fluxes:

$$\hat{\mathbf{x}} = \mathbf{x}^b_{\_p} + (\mathbf{HB})^T \left(\mathbf{HBH}^T + \mathbf{R}\right)^{-1} \left(\mathbf{y}^{a\star}_{\_} - \mathbf{Hx}^b_{\_p}\right) \tag{3}$$

where ~~$\mathbf{x}^b$~~ $\mathbf{x}_p$ is an $m \times 1$ vector of prior CO_2 fluxes, comprised of a coarse (10×10 km^2) a-temporal EDGAR v4.2 FT2010 anthropogenic CO_2 inventory and natural fluxes from CarbonTracker CT2013B, regridded to 1×1 km^2. $\mathbf{B}$ is the $m \times m$ prior error covariance matrix. The prior error covariance matrix can be expressed as a Kroenecker product (cf. Meirink et al., 2008; Singh et al., 2011; Yadav and Michalak, 2013) of temporal and spatial covariance matrices: $\mathbf{B} = \mathbf{D} \otimes \mathbf{E}$ where $\mathbf{D}$ is the temporal covariance matrix and $\mathbf{E}$ is the spatial covariance matrix. The $\mathbf{B}$ matrix has an uncertainty of 100% at the native resolution and the spatial and temporal covariance matrices are fully populated (see Supplemental Section S2 for more details).

We do not explicitly represent the individual error terms contributing to the $\mathbf{R}$ matrix (instrument error, model error, and representation error). Instead, we have assumed that the $\mathbf{R}$ matrix is diagonal and can be characterized by a single parameter: the total mismatch error (σ_m; $\mathbf{R} = \sigma_m^2 \mathbf{I}$), which represents the combined effects of the different error components.

Fig. 3 shows an example of the estimated CO_2 fluxes. We can see that the posterior fluxes capture more of the spatial variability in the CO_2 fluxes than the prior fluxes in the region where the network is deployed. We find substantial improvements in the diurnal cycle (see panel d). Previous work has used the posterior covariance matrix ($\mathbf{Q} = \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1}\right)^{-1}$), averaging kernel matrix ($\mathbf{A} = \mathbf{I} - \mathbf{QB}^{-1}$), and the degrees of freedom for signal ($\mathrm{DOFs} = \mathrm{tr}(\mathbf{A})$) as metrics to evaluate the information content of different observing systems ~~(e.g., Kort et al., 2013; Wu et al., 2015)~~(e.g., Kort et al., 2013; Wu et al., 2016). However, it is computationally infeasible to construct these $m \times m$ matrices for our application as $m > 10^6$ and storing them would require $\sim$36 Tb of memory (assuming double precision, dense matrices).

[Fig. 3 about here.]

Instead, we evaluate the efficacy of the posterior fluxes by taking the norm of the difference between the posterior fluxes and the true fluxes: ~~$||\hat{\mathbf{x}} - \mathbf{x}^a||_2$~~ $||\hat{\mathbf{x}} - \mathbf{x}^\star||_2$. We express this as a relative improvement by comparing the norm of the difference between the prior fluxes and the true fluxes:

$$\eta = 1 - \frac{||\hat{\mathbf{x}} - \mathbf{x}^a||_2}{||\mathbf{x}^b - \mathbf{x}^a||_2} \frac{||\hat{\mathbf{x}} - \mathbf{x}^\star||_2}{||\mathbf{x}_p - \mathbf{x}^\star||_2} \tag{4}$$

This error metric, η, was chosen as it has a similar form to the averaging kernel matrix but it also allows us to directly compare the posterior fluxes to the true fluxes. This relative error metric can be related to the flux error (see Supplemental Section ~~S4~~S5). As such, we can use the error metric to evaluate the ability of the observing system to resolve three types of emission sources: (1) area, (2) line, and (3) point sources, by examining a subset of grid cells in the domain (see Section S3 for

more details). The area source (AS) examined here is the East Bay urban dome (147 $\pm$ 55 tC hr^{-1}; uncertainty is the 1-σ range of hourly fluxes from the high resolution inventory), the line source (LS) is Interstate 880 and the Bay Bridge (45 $\pm$ 20 tC hr^{-1}), and the point sources (PS) are 4 large CO_2 sources in the East Bay (9 $\pm$ 4 tC hr^{-1}). For comparison, Salt Lake City emits $\sim$300 $\pm$ 50 tC hr^{-1} (McKain et al., 2012). The top panel of Fig. 2 shows these three source types.

Fig. 4 shows the error in the estimated CO_2 fluxes using the observations over a wide range of observing system scenarios. We vary the number of sites (~~n_s)and~~ $n_s = [1,2,\ldots,34]$), mismatch error (~~σ_m)~~$\sigma_m = [0.005,0.01,0.02,0.05,0.1,0.2,0.5,1,2,5,10,20]$ ppm), and perform an ensemble of 20 inversions for each combination to ensure the results are robust. Each ensemble member uses a unique observational network by randomly drawing n_s sites from the population of 34 possible sites. In total, we perform 8,160 inversions. Fig. 4 shows the mean error in the estimated CO_2 fluxes for the area source, line source, and point source as a function of σ_m and n_s. This figure represents the uncertainty in the estimated emissions at a given hour.

[Fig. 4 about here.]

5 Simplified statistical models of error reduction

We develop statistical models to predict the error reduction and quantify the importance of the different factors governing the error reduction. We tested all combinations of models with the following 7 parameters (127 possible combinations): $\sqrt{\sigma_m}$, $\sqrt{n_s}$, $\ln(\sigma_m)$, $\ln(n_s)$, σ_m, n_s, and a constant. These statistical models were evaluated using Akaike information criterion (AIC) and Bayesian information criterion (BIC). The following statistical models were found to be best:

$$\hat{\eta}_{\mathrm{AS}} = \beta_6\sqrt{\sigma_m} + \beta_5\sqrt{n_s} + \beta_4\ln(\sigma_m) + \beta_3\ln(n_s) + \beta_2\sigma_m + \beta_0 \quad (5)$$

$$\hat{\eta}_{\mathrm{LS}} = \beta_6\sqrt{\sigma_m} + \beta_5\sqrt{n_s} + \beta_4\ln(\sigma_m) + \beta_3\ln(n_s) + \beta_2\sigma_m + \beta_1 n_s \quad (6)$$

$$\hat{\eta}_{\mathrm{PS}} = \beta_6\sqrt{\sigma_m} + \beta_5\sqrt{n_s} + \beta_4\ln(\sigma_m) + \beta_2\sigma_m + \beta_0 \quad (7)$$

All the regression coefficients (β_i) in the statistical models yielded statistically significant ($p < 0.001$) parameters based on F-tests (see the Supplemental Section ~~S6~~ S7 for the regression coefficients and model selection criterion).

We find the $\sqrt{\sigma_m}$, $\sqrt{n_s}$, $\ln(\sigma_m)$, and σ_m parameters in all three statistical models (Eq. 5–7). This dependence on $\sqrt{n_s}$ and $\sqrt{\sigma_m}$ logically follows from the assumption of Gaussian errors in the derivation of the posterior CO_2 fluxes (Eq. 3) and the basic properties of variance. These two parameters tend to be dominant and generally explain more than 50% of the variance. As such, we suspect that these two parameters are the most important and that other terms are capturing higher-order effects.

These statistical models can also be used to define the regimes where increasing the number of sites in the observing system is more important and those where reducing the mismatch error is more

important~~by taking the derivative of $\hat{\eta}$ with respect to n_s :~~

$$\frac{\partial \hat{\eta}_{\mathrm{AS}}}{\partial n_s} = \frac{\beta_5}{2\sqrt{n_s}} + \frac{\beta_3}{n_s}$$

$$\frac{\partial \hat{\eta}_{\mathrm{LS}}}{\partial n_s} = \frac{\beta_5}{2\sqrt{n_s}} + \frac{\beta_3}{n_s} + \beta_1$$

$$\frac{\partial \hat{\eta}_{\mathrm{PS}}}{\partial n_s} = \frac{\beta_5}{2\sqrt{n_s}}$$

. We estimate these regimes using the ridge line from the statistical models (Eq. 5–7). From Fig. 4 we can see two distinct regimes: *noise-limited* and *site-limited*. Observing systems that lie above the ~~$\partial\hat{\eta}/\partial n_s$ curve~~ ridge line are in the the noise-limited regime where the error reduction is largely governed by the mismatch error in the observing system. Conversely, observing systems below the ~~$\partial\hat{\eta}/\partial n_s$ curve~~ ridge line are in the the site-limited regime where the error reduction is largely governed by the number of sites in the observing system.

The mismatch error is controlled by the instrument, representation, and model error. In the noise-limited regime reducing these errors will provide the greatest benefit. Whereas, in the site-limited regime the greatest benefit will come from increasing the number of sites in the observing system and there will only be marginal benefit from reducing the instrument, representation, and model error.

6 Discussion

Three conclusions we can draw from Fig. 4 for California's East Bay are:

1. Achieving $\sigma_m = 1$ ppm adds value. There is relatively little additional benefit to reducing mismatch error to 0.1 ppm, particularly for estimating line or point source emissions.

2. At $\sigma_m = 1$ ppm there is a benefit to increasing the number of sites, but this benefit increases slower than $\sqrt{n_s}$.

3. At $\sigma_m = 5$ ppm there is little benefit from increasing the number of sites; reducing the noise would add more value.

Our work is primarily focused on estimating hourly fluxes, however we can further reduce the uncertainty in our estimates by considering temporally averaged fluxes ~~.~~ (e.g., what are the weekly or monthly emissions?). Fig. 5 shows the error in our estimate of the area source emissions aggregated over various time-scales. We find the error in our estimate greatly decreases over the first 72 hours~~and agrees well with~~ . The central limit theorem provides a lower bound on the error reduction ~~predicted by the central limit theorem~~we might expect and the error reductions follow this limit reasonably well over the first 72 hours. This implies that our weekly-averaged emission estimate would be 10× better than our hourly emission estimate.

[Fig. 5 about here.]

6.1 Additional factors affecting observing system design

We considered three additional factors that could adversely impact an observing system: (1) inversion domain size, (2) site-specific systematic biases, and (3) using only daytime observations.

Our results are found to be largely insensitive to the inversion domain size (see Fig. S6). This is discerned through a set of sensitivity OSSEs with a reduced domain size. We find that inversions on the reduced domain were only marginally worse at reducing the error ($\sim 1\%$) than inversions on the full domain (see Supplemental Section ~~S5~~S6.1). This is due to the strong local signal in the footprint of the measurements (see bottom panel of Fig. 2). As such, the non-local emission sources do not adversely impact our ability to estimate urban emissions.

Biases can adversely impact the observing system (see Fig. S7). To test the impacts of biases in the modeling-measurement framework, we repeated the OSSEs outlined in Section 4 but included a systematic bias. The bias was unique to each site and was drawn from a normal distribution ($\boldsymbol{\epsilon}_b \sim \mathcal{N}\left(\mathbf{0}, \sigma_b^2 \mathbf{I}\right)$; $\sigma_b = 1$ ppm). There are three major findings from the OSSEs with systematic biases:

1. Systematic biases become particularly problematic when the spread of the potential biases (defined here as σ_b) is larger than the mismatch error ($\sigma_b > \sigma_m$). This is because we have defined the observational error covariance matrix as: $\mathbf{R} = \sigma_m^2 \mathbf{I}$. However, if $\sigma_b > \sigma_m$ with a dense observing system then the site-specific biases will artificially inflate the observational error covariance matrix: $\mathbf{R} \approx \left(\sigma_m^2 + \sigma_b^2\right) \mathbf{I}$ and the errors will be incorrectly characterized in the observing system. As long as $\sigma_b < \sigma_m$ then $\mathbf{R} = \sigma_m^2 \mathbf{I}$ and the characterization of the errors will be appropriate.

2. Observing systems with more sites are generally less affected by site-specific systematic biases. This is because observing systems with a small number of sites rely heavily on those few sites. An observing system with many sites is less reliant on a single site and the site-specific systematic biases act more like additional noise in the observing system.

3. Systematic biases have a greater impact when estimating an area source than line and point sources. This is because an airmass sensitive to a line or point source will have a greater enhancement relative to the background compared to a diffuse area source, thus there is a larger signal-to-noise ratio for these sources and a systematic bias is less important.

During the day, model calculations of the PBL height are more reliable leading to a temptation to omit the nighttime data from the analysis. However, emissions at night can be as much as 30% of the total and ignoring them makes estimates of urban emissions strongly dependent on prior assumptions. Our observing system would be unable to correct the misrepresented nighttime emissions of our a-temporal prior without using nighttime observations. As a result, even our most optimistic

observing system would have a systematic $\sim$50 tC hr^{-1} error ($\sim$30%) in the estimated area source emissions due to the misrepresented nighttime emissions.

6.2 Potential cost tradeoffs

We consider two potential observing systems:

1. "Network A" ($n_s = 25$, $\sigma_m = 1$ ppm): A dense network with moderate-precision instruments. This network is similar to the BEACO$_2$N network described in Section 3. We assume a cost of \$5,000 per instrument giving a total cost of \$125,000. This network is shown as a purple star in the left column of Fig. 4.

2. "Network B" ($n_s = 3$, $\sigma_m = 0.1$ ppm): A sparse network with of high-precision instruments. This network uses cavity-ring down instruments. We assume a cost of \$50,000 per instrument giving a total cost of \$150,000. This network is shown as a green star in the left column of Fig. 4.

We note that the assumed mismatch error for these two potential observing systems is defined as the instrument error and assumes there is no contribution from model or transport errors.

The cost for these two networks is comparable. From Fig. 4, we find that the sparse "Network B" is site-limited in all cases whereas the dense "Network A" is ~~near the noise/site-limited boundary~~in the noise-limited regime. Further, we find that the dense "Network A" has less error in the estimate of all source types in San Francisco's East Bay. Networks sitting ~~exactly~~ on the ridge line are at the optimal balance between precision and number of sites.

6.3 The relationship between network density and transport error

In this work we have treated transport error and the number of measurement sites as independent. However, in practice, there would be a relationship between the transport error and measurement network density. This can be understood with a thought experiment using two different observing systems to estimate emissions: a sparse network with a single site and an infinitely dense network (sites at each grid cell in our domain). Estimating emissions with the sparse network would require us to simulate the atmospheric transport with high fidelity if we are to reliably say anything about emissions upwind of our site. This is especially true for point sources. Any errors in the simulated atmospheric transport would adversely impact the estimated emissions, whereas the infinitely dense network could potentially neglect atmospheric transport and use data from only the local grid cell to estimate emissions. This is because the differential signal at each site would be largely governed by the local emissions. Explicitly quantifying this relationship between transport error and measurement network density should be the focus of future work.

7 Conclusions

Understanding the factors that govern our ability to estimate urban greenhouse gas emissions are crucial to improving an observing system and reducing the uncertainty in emission estimates. Here we have quantitatively mapped the errors in CO_2 emission estimates from different observing systems for three different types of sources in California's Bay Area: area sources, line sources, and point sources. Our results show that different observing systems may fall into noise or site-limited regimes where reducing the uncertainty in the estimated emissions is governed by a single factor; these regimes differ for the source types. Identifying the regime an observing system is in will help inform future improvements to the observing system. A number of prior urban CO_2 experiments have defined as a goal, the understanding of emissions to less than 10% ~~(e.g., Kort et al., 2013; Wu et al., 2015)~~(e.g., Kort et al., 2013; Wu et
We find that a BEACO$_2$N-like network could achieve this accuracy and precision with 1 week of observations, if the dominant source of error is instrument precision. This conclusion may motivate a re-examining of the conventional instrument quality-oriented design of CO_2 observing systems, according to the stated goal of a given network.

Acknowledgements. This work was supported by a Department of Energy (DOE) Computational Science Graduate Fellowship (CSGF) to AJT, a National Science Foundation (NSF) Grant 1035050 to RCC, and a Bay Area Air Quality Management District (BAAQMD) Grant 2013.145 to RCC. AAS was supported by a National Science Foundation Graduate Research Fellowship. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank M. Sulprizio (Harvard University) for gridding the US Census population data and the UC Berkeley Academic Computing center for access to computing resources.

References

Breon, F. M., Broquet, G., Puygrenier, V., Chevallier, F., Xueref-Remy, I., Ramonet, M., Dieudonn, E., Lopez, M., Schmidt, M., Perrussel, O., and Ciais, P.: An attempt at estimating Paris area CO_2 emissions from atmospheric concentration measurements, Atmospheric Chemistry and Physics, 15, 1707–1724, doi: 10.5194/acp-15-1707-2015, 2015.

Deschłnes, O. and Greenstone, M.: Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the US, American Economic Journal: Applied Economics, 3, 152–185, doi: 10.1257/app.3.4.152, 2011.

EIA, U.: Emissions of Greenhouse Gases in the U.S., Tech. rep., U.S. Energy Information Administration, 2015.

European Commission: Emission Database for Global Atmospheric Research (EDGAR), release version 4.2, Tech. rep., Joint Research Centre (JRC)/Netherlands Environmental Assessment Agency (PBL), 2011.

Gately, C. K., Hutyra, L. R., Wing, I. S., and Brondfield, M. N.: A bottom up approach to on-road CO_2 emissions estimates: improved spatial accuracy and applications for regional planning, Environ Sci Technol, 47, 2423–30, doi:10.1021/es304238v, 2013.

Gately, C. K., Hutyra, L. R., and Sue Wing, I.: Cities, traffic, and CO_2: A multidecadal assessment of trends, drivers, and scaling relationships, Proc Natl Acad Sci U S A, 112, 4999–5004, doi:10.1073/pnas. 1421723112, 2015.

Gratani, L. and Varone, L.: Daily and seasonal variation of CO_2 in the city of Rome in relationship with the traffic volume, Atmospheric Environment, 39, 2619–2624, doi:10.1016/j.atmosenv.2005.01.013, 2005.

Gurney, K. R., Mendoza, D. L., Zhou, Y., Fischer, M. L., Miller, C. C., Geethakumar, S., and de la Rue du Can, S.: High resolution fossil fuel combustion CO_2 emission fluxes for the United States, Environ Sci Technol, 43, 5535–41, 2009.

Gurney, K. R., Razlivanov, I., Song, Y., Zhou, Y., Benes, B., and Abdul-Massih, M.: Quantification of fossil fuel CO_2 emissions on the building/street scale for a large U.S. city, Environ Sci Technol, 46, 12 194–202, doi:10.1021/es3011282, 2012.

Hutyra, L. R., Duren, R., Gurney, K. R., Grimm, N., Kort, E. A., Larson, E., and Shrestha, G.: Urbanization and the carbon cycle: Current capabilities and research outlook from the natural sciences perspective, Earth's Future, 2, 473–495, doi:10.1002/2014ef000255, 2014.

IPCC: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Tech. rep., 2013.

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, Journal of Geophysical Research: Atmospheres, 118, 1577–1584, doi:10.1002/jgrd.50135, 2013.

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, 2654–2668, doi:10.1175/jamc-d-13-038.1, 2013.

Lin, J. C., Gerbig, C., Wofsy, S. C., Andrews, A. E., Daube, B. C., Davis, K. J., and Grainger, C. A.: A near-field tool for simulating the upstream influence of atmospheric observations: The Stochastic Time-Inverted

Lagrangian Transport (STILT) model, Journal of Geophysical Research-Atmospheres, 108, ACH 2–1–ACH 2–17, doi:10.1029/2002jd003161, 2003.

Mangat, T. S., Claire, S. J., Dinh, T. M., Fanai, A. K., Nguyen, M. H., and Schultz, S. A.: Source inventory of Bay Area greenhouse gas emissions, Tech. rep., Bay Area Air Quality Management District, 2010.

McDonald, B. C., McBride, Z. C., Martin, E. W., and Harley, R. A.: High-resolution mapping of motor vehicle carbon dioxide emissions, Journal of Geophysical Research-Atmospheres, 119, 5283–5298, doi:10.1002/2013jd021219, 2014.

McKain, K., Wofsy, S. C., Nehrkorn, T., Eluszkiewicz, J., Ehleringer, J. R., and Stephens, B. B.: Assessment of ground-based atmospheric observations for verification of greenhouse gas emissions from an urban region, Proc Natl Acad Sci U S A, 109, 8423–8, doi:10.1073/pnas.1116645109, 2012.

Meirink, J. F., Bergamaschi, P., and Krol, M. C.: Four-dimensional variational data assimilation for inverse modelling of atmospheric methane emissions: method and comparison with synthesis inversion, Atmospheric Chemistry and Physics, 8, 6341–6353, doi:10.5194/acp-8-6341-2008, 2008.

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, Journal of Geophysical Research: Atmospheres, 118, 917–933, doi:10.1029/2012jd018196, 2013.

Nehrkorn, T., Eluszkiewicz, J., Wofsy, S. C., Lin, J. C., Gerbig, C., Longo, M., and Freitas, S.: Coupled weather research and forecastingstochastic time-inverted lagrangian transport (WRF–STILT) model, Meteorology and Atmospheric Physics, 107, 51–64, doi:10.1007/s00703-010-0068-x, 2010.

Newman, S., Jeong, S., Fischer, M. L., Xu, X., Haman, C. L., Lefer, B., Alvarez, S., Rappenglueck, B., Kort, E. A., Andrews, A. E., Peischl, J., Gurney, K. R., Miller, C. E., and Yung, Y. L.: Diurnal tracking of anthropogenic CO_2 emissions in the Los Angeles basin megacity during spring 2010, Atmospheric Chemistry and Physics, 13, 4359–4372, doi:10.5194/acp-13-4359-2013, 2013.

Oda, T. and Maksyutov, S.: A very high-resolution (1 km$\times$1 km) global fossil fuel CO_2 emission inventory derived using a point source database and satellite observations of nighttime lights, Atmospheric Chemistry and Physics, 11, 543–556, doi:10.5194/acp-11-543-2011, 2011.

Peters, W., Jacobson, A. R., Sweeney, C., Andrews, A. E., Conway, T. J., Masarie, K., Miller, J. B., Bruhwiler, L. M., Petron, G., Hirsch, A. I., Worthy, D. E., 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 U S A, 104, 18 925–30, doi:10.1073/pnas.0708986104, 2007.

Rigby, M., Toumi, R., Fisher, R., Lowry, D., and Nisbet, E. G.: First continuous measurements of CO_2 mixing ratio in central London using a compact diffusion probe, Atmospheric Environment, 42, 8943–8953, doi:10.1016/j.atmosenv.2008.06.040, 2008.

Rodgers, C. D.: Inverse Methods for Atmospheric Sounding, World Scientific, Singapore, 2000.

Shusterman, A. A., Teige, V., Turner, A. J., Newman, C., Kim, J., and Cohen, R. C.: The BErkeley Atmospheric CO_2 Observation Network: initial evaluation, Atmospheric Chemistry and Physics Discussions, pp. 1–23, doi:10.5194/acp-2016-530, 2016.

Singh, K., Jardak, M., Sandu, A., Bowman, K., Lee, M., and Jones, D.: Construction of non-diagonal background error covariance matrices for global chemical data assimilation, Geoscientific Model Development, 4, 299–316, doi:10.5194/gmd-4-299-2011, 2011.

Turnbull, J. C., Sweeney, C., Karion, A., Newberger, T., Lehman, S. J., Tans, P. P., Davis, K. J., Lauvaux, T., Miles, N. L., Richardson, S. J., Cambaliza, M. O., Shepson, P. B., Gurney, K., Patarasuk, R., and Razlivanov, I.: Toward quantification and source sector identification of fossil fuel CO_2 emissions from an urban area: Results from the INFLUX experiment, Journal of Geophysical Research-Atmospheres, 120, 292–312, doi: 10.1002/2014JD022555, 2015.

Wu, L., Broquet, G., Ciais, P., Bellassen, V., Vogel, F., Chevallier, F., Xueref-Remy, I., and Wang, Y.: Atmospheric inversion for cost effective quantification of city CO_2 emissions, Atmospheric Chemistry and Physics Discussions, 15, 30 693–30 756, doi:10.5194/acpd-15-30693-2015, 2015.

Wu, L., Broquet, G., Ciais, P., Bellassen, V., Vogel, F., Chevallier, F., Xueref-Remy, I., and Wang, Y.: What would dense atmospheric observation networks bring to the quantification of city CO_2 emissions?, Atmospheric Chemistry and Physics, 16, 7743–7771, doi:10.5194/acp-16-7743-2016, 2016.

Yadav, V. and Michalak, A. M.: Improving computational efficiency in large linear inverse problems: an example from carbon dioxide flux estimation, Geoscientific Model Development, 6, 583–590, doi: 10.5194/gmd-6-583-2013, 2013.

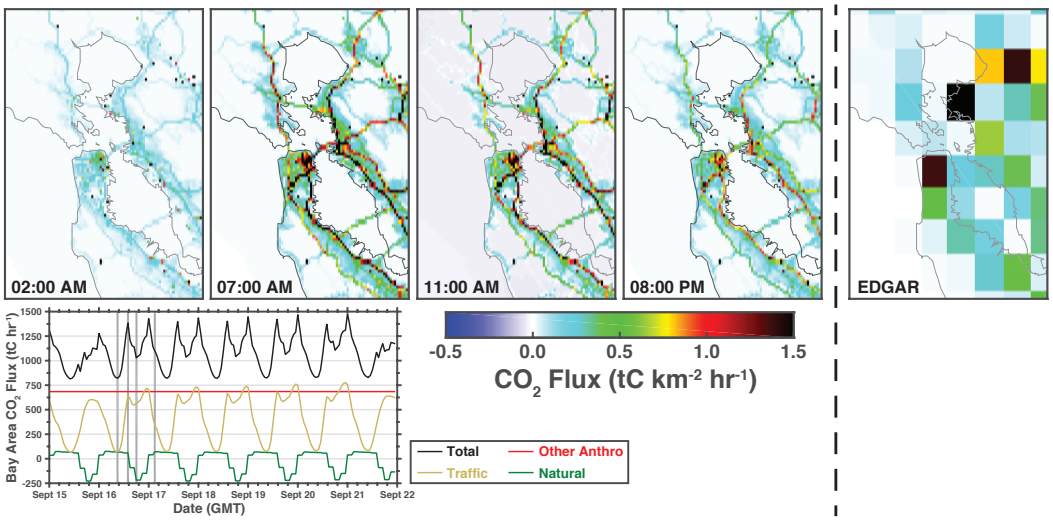

Fig. 1. September 2013 CO_2 fluxes from bottom-up inventories. Top row shows the fluxes in the Bay Area (122.0357° – 122.7683°W, 37.3771° – 38.2218°N) at four representative hours (hour in local time). Right panel shows the a-temporal EDGAR v4.2 FT2010 CO_2 flux in the Bay Area. Bottom panel shows the total Bay Area CO_2 flux (black), traffic (orange), other anthropogenic (red), and natural (green) sources. Vertical gray shading indicates the time slices plotted in the top and middle panels.

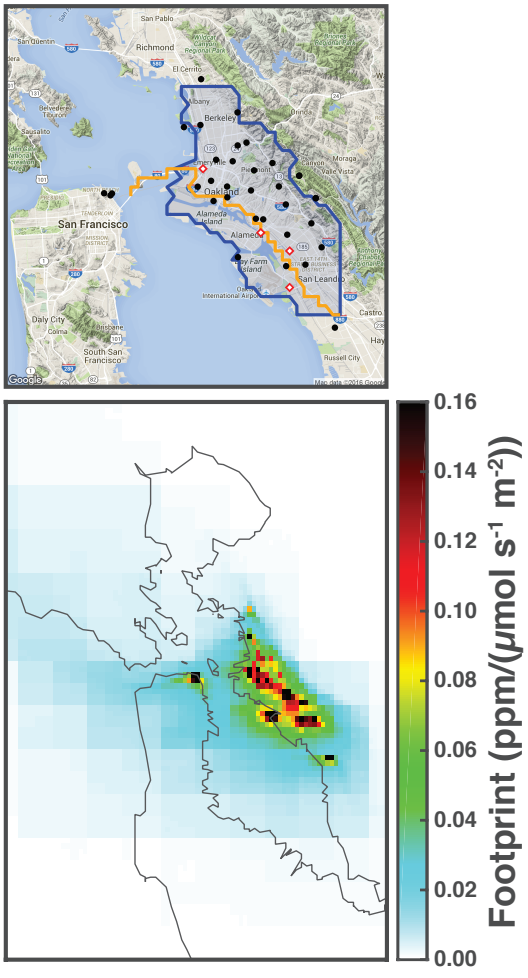

Fig. 2. Top panel shows the location of the sites (black circles), the area source (blue region), the line source (orange line), and point sources (red diamonds). Bottom panel shows the September 15 to 22 average footprint for the 34 sites in the network, see Table 1 for a list of the sites. The bottom panel is the full domain used for the inversion. Supplemental Fig. S3 shows the footprint on a log-scale.

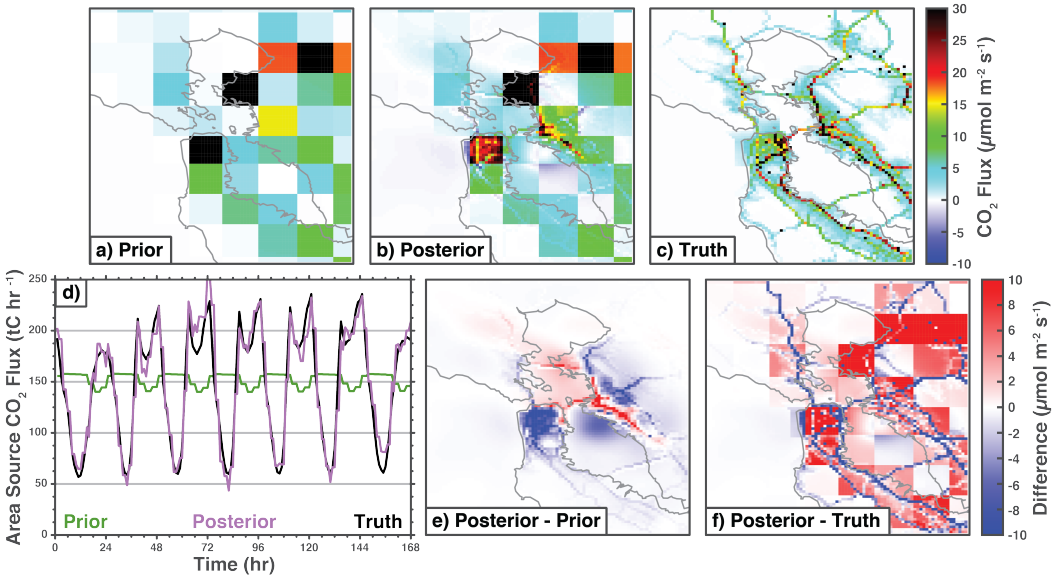

Fig. 3. Example of estimated CO_2 fluxes. Top row shows the average emissions from **(a)** the prior, **(b)** the posterior, and **(c)** the true emissions. Panel **(d)** shows a time series of the emissions from the area source with the prior (green), posterior (pink), and true emissions (black). Panel **(e)** shows the difference between the posterior and the prior. Panel **(f)** shows the difference between posterior and the truth. Posterior output is from the best case scenario ($n_S = 34$ and $\sigma_m = 0.005$ ppm).

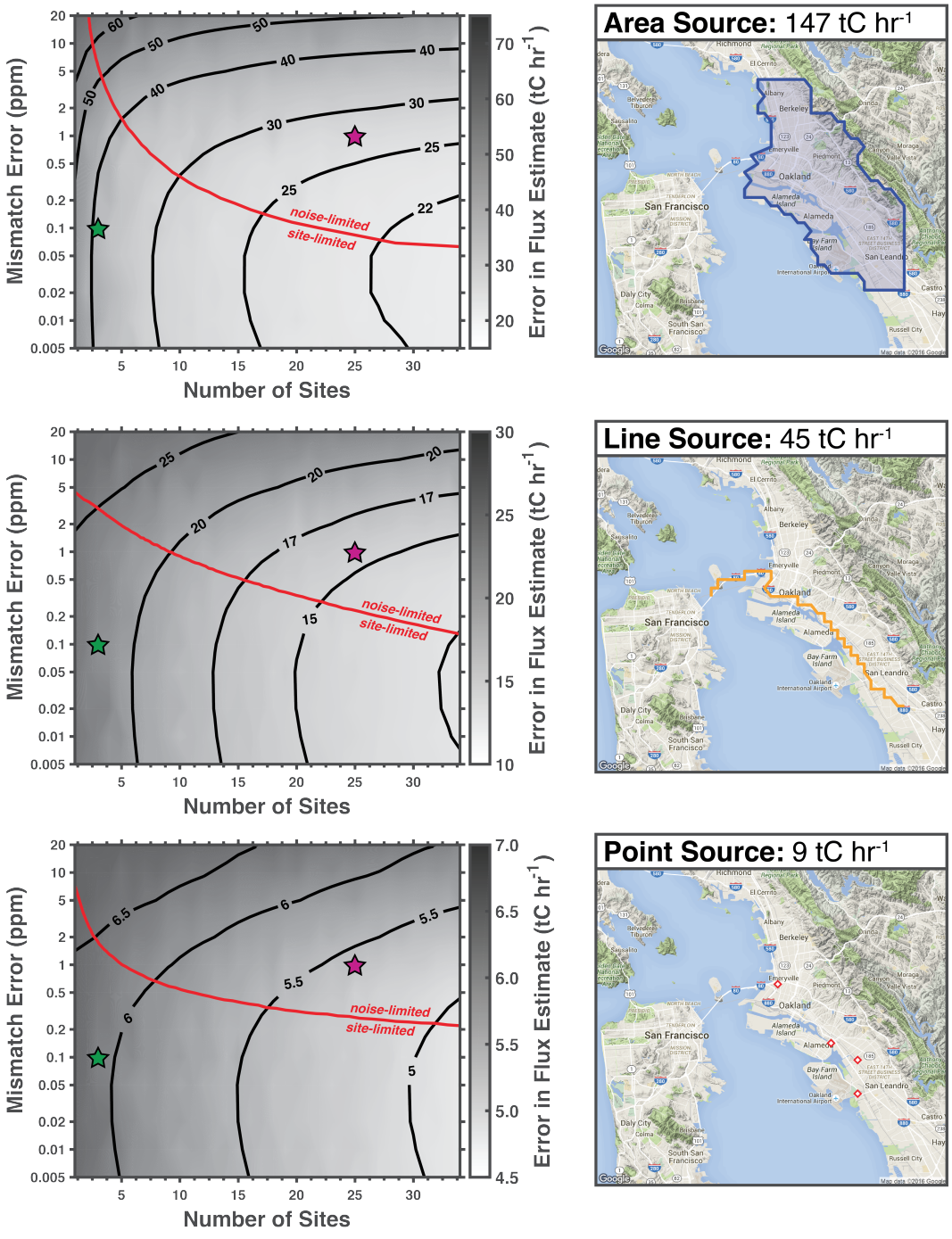

Fig. 4. Left column shows the error in the posterior CO_2 fluxes. Right column shows the fluxes being estimated. Top row is the area source, middle row is the line source, and bottom row is the point source. Inversions were performed using $n_s = [1,2,\ldots,34]$ sites and $\sigma_m = [0.005,0.01,0.02,0.05,0.1,0.2,0.5,1,2,10,20]$ ppm mismatch error. Results shown are the mean of a monte carlo analysis using 20 different combinations of sites for each n_s, σ_m pair. Contours are from the statistical models $\hat{\eta}$ (see Eq. 5–7) converted to flux errors and the red lines are the ~~partial derivative of the statistical models with respect to the number of sites, $\partial\hat{\eta}/\partial n_s$ (Eq. 8–8),~~ ridge lines that define the cutoff between the noise-limited and site-limited regimes. Purple star shows an observing system with 25 sites and 1 ppm noise. Green star shows an observing system with 3 sites and 0.1 ppm noise. Note the log-scale on the y-axis.

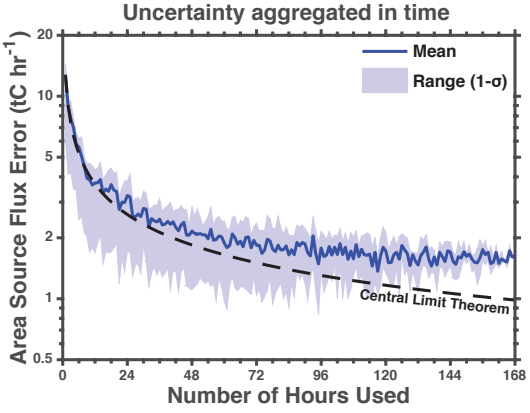

Fig. 5. Uncertainty aggregated in time for the best case inversion (see Fig. 3). The CO_2 flux estimate in this study has an hourly temporal resolution. The uncertainty in the emissions estimate declines as the estimate is averaged to longer temporal scales. Solid blue line is the mean uncertainty, shading is the 1-σ range, and the dashed black line is the uncertainty predicted by the central limit theorem. Note the log scale on the y-axis.

Table 1. 34 sites in the network[a] used in this study.

Site Code	Site name	Latitude (°N)	Longitude (°W)	Height (m a.g.l.)
AHS	Arroyo High School	37.680	122.139	3
BEL	Burckhalter Elementary School	37.775	122.167	5
BFE	Bayfarm Elementary School	37.744	122.251	3
BOD	Bishop O'Dowd High School	37.753	122.155	3
CES	Claremont Elementary School	37.846	122.252	3
CHA	Chabot Space & Science Center (low)	37.819	122.181	3
CHB	Chabot Space & Science Center (high)	37.819	122.181	9
COI	Coit Tower	37.8030	122.406	5
CPS	College Preparatory School	37.849	122.242	24
EBM	W. Oakland EBMUD Monitoring Station	37.814	122.282	3
ELC	El Cerrito High School	37.907	122.294	8
EXB	Exploratorium (Bay)	37.803	122.397	6
EXE	Exploratorium (Embarcadero)	37.801	122.399	3
FTK	Fred T. Korematsu Discovery Academy	37.738	122.174	3
GLE	Greenleaf Elementary School	37.765	122.194	3
HRS	Head Royce School	37.809	122.204	7
ICS	International Community School	37.779	122.231	3
KAI	Kaiser Center	37.809	122.264	127
LAU	Laurel Elementary School	37.792	122.197	12
LBL	Lawrence Berkeley National Lab, Bldg. 70	37.876	122.252	3
LCC	Lighthouse Community Charter School	37.736	122.196	3
MAR	Berkeley Marina	37.863	122.314	3
MON	Montclair Elementary School	37.830	122.212	3
NOC	N. Oakland Community Charter School	37.833	122.277	3
OMC	Oakland Museum of California	37.799	122.264	3
PAP	PLACE at Prescott Elementary	37.809	122.298	3
PDS	Park Day School	37.832	122.257	3
PHS	Piedmont Middle & High School	37.824	122.233	3
POR	Port of Oakland Headquarters	37.796	122.280	3
OHS	Oakland High School	37.805	122.236	3
ROS	Rosa Parks Elementary School	37.865	122.295	3
SHA	Skyline High School (low)	37.798	122.162	3
SHB	Skyline High School (high)	37.798	122.162	13
STL	St. Elizabeth High School	37.779	122.222	3

[a] This study uses both operational and proposed sites. See Shusterman et al. (2016) and "http://beacon.berkeley.edu/" for more information on the network.