Investigations into the use of multi-species measurements for source apportionment of the Indianapolis fossil fuel CO2 signal

Brian Nathan, Thomas Lauvaux, Jocelyn Turnbull, Kevin Gurney

ELEMENTA
Science of the Anthropocene

Nathan, B, et al. 2018 Investigations into the use of multi-species measurements for source apportionment of the Indianapolis fossil fuel CO_2 signal. *Elem Sci Anth*, 6: 21. DOI: https://doi.org/10.1525/elementa.131

RESEARCH ARTICLE

Investigations into the use of multi-species measurements for source apportionment of the Indianapolis fossil fuel CO_2 signal

Brian Nathan*, Thomas Lauvaux*, Jocelyn Turnbull† and Kevin Gurney‡

Current bottom up estimates of CO_2 emission fluxes are based on a mixture of direct and indirect flux estimates relying to varying degrees on regulatory or self-reported data. Hence, it is important to use additional, independent information to assess biases and lower the flux uncertainty. We explore the use of a self-organizing map (SOM) as a tool to use multi-species observations to partition fossil fuel CO_2 (CO_2ff) emissions by economic source sector. We use the Indianapolis Flux experiment (INFLUX) multi-species observations to provide constraints on the types of relationships we can expect to see, and show from the observations and existing knowledge of likely sources for these species that relationships do exist but can be complex. An Observing System Simulation Experiment (OSSE) is then created to test, in a pseudodata framework, the abilities and limitations of using an SOM to accurately attribute atmospheric tracers to their source sector. These tests are conducted for a variety of emission scenarios, and make use of the corresponding high-resolution footprints for the pseudo-measurements. We show here that the attribution of sector-specific emissions to measured trace gases cannot be addressed by investigating the atmospheric trace gas measurements alone. We conclude that additional a priori information such as inventories of sector-specific trace gases are required to evaluate sector-level emissions using atmospheric methods, to overcome the challenge of the spatial overlap of nearly every predefined source sector. Our OSSE additionally allows us to demonstrate that increasing the (already high) data density cannot solve the co-localization problem.

Keywords: CO_2; urban; sector; emissions; tracers; mitigation; multi-species

Introduction

Although urban areas account for only ~2% of the global land area, they are responsible for approximately 70% of greenhouse gas (GHG) emissions, with anthropogenic CO_2 being the most important of these (United Nations Human Settlements Programme, 2011). Determination of urban CO_2 emissions, then, is of special importance to policymakers trying to mitigate local fossil fuel consumption. This will be most useful to policymakers if the specific economic sector contributions to urban CO_2 emissions can be accurately quantified at the local level (Hutyra et al., 2014).

Detailed bottom-up greenhouse gas emissions data products exist for few urban areas (*e.g.* Gurney et al. (2012); AIRPARIF (2013)). Considering the variety of ground-based sources and the need for mitigation policies, economic sectors are defined for each city (*e.g.* traffic, residential) and then distributed spatially and temporally using ancillary information such as traffic counts, building use types, temperature, etc. These bottom-up data products often lack uncertainty estimates, because it is difficult to assess the uncertainties in much of the data used. Direct measurements exist for some sectors (*e.g.* stack flow data on power plants), but such data may be limited and affected by large biases (Ackerman and Sundquist 2008; Gurney et al., 2016).

Atmospheric observations may be useful in addressing the current gaps in knowledge. Several studies have evaluated whole-city CO_2 emissions (*e.g.* Turnbull et al. (2011); Font et al. (2015); Staufer et al. (2016); Heimburger et al. (2017)), while only a few studies have attempted to quantify sub-city scale and source-specific emissions. Atmospheric inversions can potentially combine top-down and bottom-up information to address this problem (*e.g.* Lauvaux et al. (2016)). Trace gases and isotopes are already being used to identify and quantify specific sources. The $^{14}CO_2$ isotope, specifically, has the ability to discriminate the anthropogenic, fossil-fuel-related

* Department of Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, Pennsylvania, US
† GNS Science, Gracefield, Wellington, NZ
‡ School of Life Sciences, Arizona State University, Tempe, Arizona, US
Corresponding author: Brian Nathan (bjn5178@psu.edu)

CO_2 (CO_2ff) enhancements from biogenic and other CO_2 sources (Levin et al., 2003; Turnbull et al., 2007, 2011; Vogel et al., 2010; Miller et al., 2012). But $^{14}CO_2$ distinguishes only total locally-added CO_2ff and cannot further partition CO_2ff into source sectors. Some urban studies have augmented the ^{14}C-based CO_2ff determination with source information gleaned from the other isotopes in CO_2 (*e.g.* Djuricin et al. (2010); Newman et al. (2016). $\delta^{13}C$ has proven useful for distinguishing between gasoline-related, coal-related, and natural-gas-related contributions (Clark-Thorne and Yapp, 2003; Widory and Javoy, 2003).

There is also strong evidence that other trace gases are strongly related to CO_2ff and could potentially be used to separate CO_2ff by source sector. We summarize the available information on various species and their relationship to CO_2ff in **Figure 1**. This figure is meant to expose the reader to the full range of atmospheric gases being recorded for this study and the initial state of knowledge of this investigation for the gases' relationship to predefined CO_2ff source sectors.

In the figure, the colors of the boxes qualitatively identify how much CO_2ff source sector knowledge, in any urban environment, is available from existing literature. The sizes of the boxes are also scaled as a qualitative metric for the relative proportion of emissions among the sectors: the absence of boxes means no emissions are expected from that species in that sector, small boxes mean few emissions are expected, and large boxes indicate that most emissions are expected to be from that sector. If the relative proportions among sectors are unknown, all boxes are left as large. This also applies to cases such as CH_4 where the major sources (*e.g.* the landfill and the wastewater treatment plant in Indianapolis (Cambaliza et al., 2015; Lamb et al., 2016)) do not fit into any of the predefined CO_2ff sector categories. The sizes of the boxes are purely qualitative and are not intended to represent any number nor to appear to have been normalized. Additionally, the final 2 columns showcase the raw-data r^2 values from linear regressions against CO_2ff as well as r^2_{pared} for pared-down datasets, as described in detail later. Overall, **Figure 1** introduces some of the problems

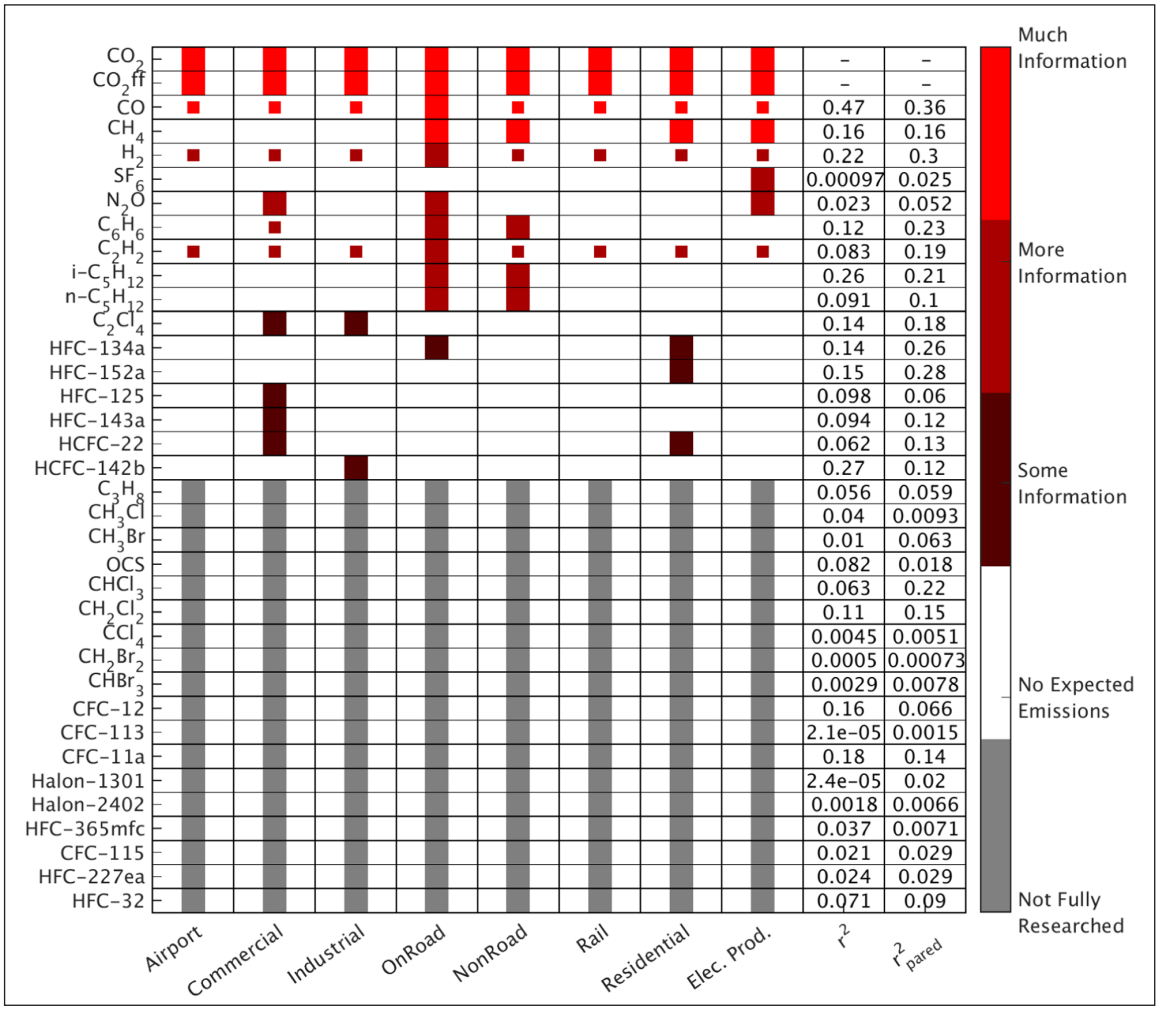

Figure 1: A preliminary literature survey provided some source-sector-related information for the included species, with references stated in the text. One goal of this study is to determine these relationships directly from the atmospheric measurements. The box sizes in the figure qualitatively represent the relative contribution for a species to a sector, if it is known (else they are all large). The r^2 and pared-down r^2_{pared} values from plots against CO_2ff for the INFLUX dataset are included in the final 2 columns, where applicable. DOI: https://doi.org/10.1525/elementa.131.f1

and raises some of the questions which are addressed in this manuscript. These include whether there actually exist unique tracer species for each CO_2ff source sector, whether they would be directly detectable from atmospheric observations if they did exist, and determining how much prior source flux information is needed for accurate detection and attribution.

CO is an example of a species with a strong relationship to CO_2ff. It is co-emitted with CO_2ff in varying ratios depending on combustion conditions, and in U.S. and European cities is predominantly from vehicles (Turnbull et al., 2011, 2015a; Vogel et al., 2010). When the urban plume is well-mixed, the CO:CO_2ff relationship can be diagnosed and used to determine the CO_2ff flux ((Levin and Karstens, 2007; Turnbull et al., 2011). Variability in the CO:CO_2ff relationship in space and time is related to variability in the source mix (Vogel et al., 2010), suggesting that the relative contributions of different sources could be determined from these observations. Turnbull et al. (2015a) showed that the CO:CO_2ff ratio varies diurnally in Indianapolis, driven by diurnal variability in the relative contribution of on-road traffic to total CO_2ff emissions. This result utilized the fact that on-road traffic produces significant *CO* emissions, whereas other CO_2ff in Indianapolis produce little or no *CO* (Turnbull et al., 2015a).

Many other trace gases are emitted from anthropogenic sources, often associated with some subset of CO_2ff sectors, suggesting that they have potential to help partition CO_2ff by source sector. For example, Miller et al. (2012) used 6 years of approximately semi-monthly airborne profiles downwind of the Northeastern US, and found statistically significant correlations for 22 anthropogenic species against CO_2ff. Stronger correlations were found when analyzing the winter and summer measurements separately. Similarly, Turnbull et al. (2011), used aircraft measurements over Sacramento, CA, and found strongly statistically significant correlations for anthropogenically-related hydrocarbons and halocarbons, but a statistically insignificant relationship with biomass burning and ocean tracers. A number of studies have shown strong relationships between *CO* and a host of tracers in urban areas (Warneke et al., 2007; Baker et al., 2008). The current knowledge of the urban budget for these tracers varies, with some fairly well understood, while we have only limited knowledge of others. Acetylene (C_2H_2) for example, is associated with combustion in urban areas and comes primarily from traffic, as demonstrated by strong correlations with *CO* in many studies (*e.g.* Warneke et al., 2007; Baker et al., 2008). It may still have non-negligible contributions in other sectors, particularly heavy industry due to its use in welding (Fortin et al., 2005; Whitby and Altwicker, 1978). In general, trace hydrocarbons are expected to be associated with traffic sources (*e.g.* Colvile et al., 2001; Fortin et al., 2005; Fujita et al., 1995; Warneke et al., 2007; Watson et al., 2001), as they are combustion byproducts, and other combustion-related sectors like industry and electricity production have often optimized the efficiency of their burning methods, which minimizes the emission of such byproducts. Yet light hydrocarbons such as ethane, propane, and butane also enter the atmosphere through evaporation of spills during distribution and use (United States Environmental Protection Agency, 2012), a process which produces no CO_2. Still, the observed strong correlations with *CO* suggest that these hydrocarbons could be good tracers for traffic in an urban area.

Other species may be associated with other CO_2ff sectors, however. Halocarbons are typically associated with distinct processes such as refrigerants, industrial solvents, propellants, and foam-blowing agents (Barletta et al., 2013; Kim et al., 2011; Purohit and Hoglund-Isaksson, 2016, and references therein). This may allow for cases where individual species can be directly associated with a single CO_2ff sector, as in the case of *HFC*-125, which is mainly used for commercial purposes in refrigerant blends (O'Doherty et al., 2009; Velders et al., 2009). Similarly, *HFC*-134*a* enters the atmosphere primarily through leakage from mobile (vehicle) air conditioners (Papasavva et al., 2009), so it can be associated with the traffic emissions sector. It cannot be assumed, however, that *HFC*-134*a* emissions are linear with traffic CO_2ff emissions, since the leaks of *HFC*-134*a* are governed by a different process than CO_2ff from combustion.

SF_6 is an example of a species that had been thought to be a good anthropogenic tracer. It is primarily used in high-voltage gas-insulated switchgears (GIS's) in electrical transmission and distribution systems (as a spark quencher) (Maiss and Brenninkmeijer, 1998). Because of this, it has been thought of as a useful tracer for utilities, and has been used as such in previous atmospheric studies (*e.g.* Bakwin et al., 1998; Geller et al., 1997; Turnbull et al., 2006).

Some gases, such as CH_4 and N_2O, are produced by combustion, but have significant or even dominant sources that are unrelated to CO_2ff. For CH_4, these include landfills, wastewater, and natural gas pipeline leaks (Lamb et al., 2016). For example, Mckain et al. (2015) found that between ~60–100% of methane emissions in Boston could be attributed to natural gas, depending on the season. For N_2O, these dominant sources are mostly agriculture and biomass burning (Ciais et al., 2013; Davidson and Kanter, 2014). Davidson and Kanter (2014) estimate that global anthropogenic emissions are 66% from agriculture, 15% from energy and transport, 11% from biomass burning, and 8% from other sources. H_2, too, may be used as a traffic tracer, but it also has non-combustion sources (Aalto and Lallo, 2009; Barnes, 2003). Indeed, while ~40% of global emissions may be attributed to the burning of fossil fuel and biomass, the oxidation of methane and non-methane hydrocarbons constitute another ~50%, with the remaining ~10% being attributed to volcanic emissions, oceanic emissions, and production by legumes during nitrogen fixation (Barnes, 2003; Novelli et al., 1999).

The analysis presented here attempts to separate out the economic-sector-level emissions within an urban area by utilizing numerous trace gas species and investigating their relationship to CO_2ff values. First, we examine existing flask-based observations of CO_2ff and a suite of 34 other trace gases from the observationally-densest urban mission to date, the INdianapolis FLUX project (INFLUX)

(Turnbull et al., 2012, 2015a; Richardson et al., 2017). We determine the types of relationships that exist between CO_2ff and the other trace gases and use these to constrain some Observing System Simulation Experiments (OSSEs). Next we present the OSSE approach, which establishes a methodology for relating multi-species atmospheric measurements directly to CO_2ff-related economic source sectors. Idealized sector-related tracers are tested in a pseudo-data framework to determine if they can be accurately attributed to their source sector given a wide range of emission scenarios that include both linear and nonlinear relationships with CO_2ff. A self-organizing map (SOM) is utilized to explore the different trace gases and their origin based purely on their co-localization in space. No relationship is assumed between trace gases. The advantages and limitations of this methodology are explored over 495 experimental OSSE analyses with the goal of eventually being able to provide policymakers with a direct, independent verification of urban GHG emissions at the sector level.

Methodology

The INFLUX project was started as a testbed to develop methodologies for measuring urban emissions in 2010. As part of this project, 12 communications towers are instrumented within and just outside of the city of Indianapolis. All 12 measure quasi-continuous CO_2 and CH_4 using Cavity Ring-Down Spectrometers (CRDS) (Crosson, 2008; Rella et al., 2013; Miles et al., 2017), and five measure quasi-continuous CO (http://sites.psu.edu/influx/site-information/). There are additionally continuous LIDAR measurements and aircraft measurement flights using CRDS's for CO_2, CO, and CH_4. Six of the communication towers also collect flask samples for multi-species analysis (and flasks are also collected during the aircraft flights). These flask measurements will be used as a starting point for the OSSE investigation, as explained in detail later.

Multi-species flask measurements

The flask collection and measurement techniques are as described in detail in Turnbull et al. (2012). Whole air samples are collected as one-hour integrated samples using a large, 15L mixing volume and variable flow rate to obtain a rough linear mixture of air from the one-hour integration period, with the final sample stored in 2 flasks for a total of ~4L of air retained. This representative hourly mixed sample is ideally suited for inclusion in models which do not resolve shorter time-scale atmospheric fluctuations. To further ensure the most representative atmospheric samples, the flasks recorded for this analysis were taken in the mid-afternoon, when daily atmospheric mixing is expected to be at its peak (Stull, 1988; Bakwin et al., 1998; Yi et al., 2001; Miles et al., 2017).

Samples are collected only when the air flows from the west, so that Tower One, which is located slightly southwest of Indianapolis, as shown in **Figure 3**, is always upwind. Tower One thus serves as a local background constraint. Urban enhancement values are calculated through subtracting off the corresponding Tower One measurement value for each measured species. We have reason to believe that the measurements recorded at Tower One have not been influenced by background sources that are not being measured (Turnbull et al., 2015a; Lauvaux et al., 2016; Miles et al., 2017). Analyses have shown that Tower One has the lowest CO_2 concentrations on average over time (Lauvaux et al., 2016), including in the dormant season, where it has been shown to be within 0.2 ppm of the lowest CO_2 INFLUX tower measurement 43% of the time, which is the highest percentage of any tower, and potential source signals were only found in the southeast (Miles et al., 2017), which would have no impact on this analysis given the flask sampling strategy. However, it is possible that there are unknown background sources for some of the other measured trace gases which have not yet been fully explored. From December 27, 2010 through June 5, 2015, there were 1,246 tower flask measurements (948 not from Tower One) over 307 unique days.

Each flask sample is analyzed on multiple instruments to retrieve concentrations for 35 different atmospheric trace gas species (Turnbull et al., 2012), which are all included in **Figure 1**. The greenhouse gases, CO_2, CH_4, CO, H_2, N_2O, and SF_6 are analyzed at the National Oceanic and Atmospheric Association's Earth Science Research Laboratory (NOAA/ESRL) using the MAGICC system (Sweeney et al., 2015). The remaining halocarbons and hydrocarbons are analyzed using a Gas Chromatograph Mass Spectrometer (GCMS) (Montzka et al., 1993) at NOAA/ESRL. The $^{14}CO_2$ measurements, which are used to separate out the CO_2ff signal from the CO_2 signal (Meijer et al., 1996; Levin et al., 2003; Turnbull et al., 2006, 2009; Djuricin et al., 2010; Van Der Laan et al., 2010; Turnbull et al., 2015a), are processed to CO_2 gas at the University of Colorado, INSTAAR and graphitized and measured at either University of California Irvine (Turnbull et al., 2007) or GNS Science (Turnbull et al., 2015a; b). Data quality from both analysis laboratories has typical repeatability around 1.8‰ (Turnbull et al., 2007, 2015b).

The CO_2ff value is calculated using the CO_2 and $^{14}CO_2$ measurements following the procedure detailed in Turnbull et al. (2009):

$$CO_2ff = \frac{CO_{2_{obs}}\left(\Delta_{obs} - \Delta_{bg}\right)}{\Delta_{ff} - \Delta_{bg}}, \quad (1)$$

where CO_{2obs} is the observed CO_2 mole fraction, Δ_{obs} is the corresponding observed $^{14}CO_2$ value, Δ_{bg} is the background $^{14}CO_2$ value, and Δ_{ff} is −1000‰. Note that Tower 1 serves as the background for $^{14}CO_2$ measurements, too, which is the best available site when approximating emissions from only Indianapolis, as described in detail in Turnbull et al. (2015a). For this urban study, we assume no significant biases from heterotrophic respiration or other sources (Turnbull et al., 2009).

At sufficiently large sampling distances, tracer-tracer relationships often appear as linearly related to each other (*e.g.* Turnbull et al. (2011); Miller et al. (2012)). As distance between the receptor and the emitter decreases, the true complexity of relationships between trace gases become

increasingly apparent. Sources may be approximately but not exactly co-located (*e.g.* SF_6 is used in electricity junction boxes at powerplants, whereas CO_2ff is emitted directly from smokestacks). Even when co-located, the processes may differ. For example, *HFC*-134*a* will be emitted from mobile air conditioner leaks in vehicles that also produce CO_2ff from combustion. Therefore, we evaluate as a first-order approach the linear relationships between CO_2ff and other trace gases with a series of linear regressions to identify the possible use of tracer-to-tracer relationships.

Hestia: An Indianapolis bottom-up estimate using source sectors

For the INFLUX project, a high-resolution CO_2ff inventory called Hestia was generated for Marion County and the 8 counties surrounding Indianapolis (Gurney et al., 2012). In this study, the building-level-resolution product was aggregated into 1-km^2 pixels filling an 87 × 87-km^2 domain. Hestia's estimates are separated into economic sectors: Airport, Commercial, Industrial, OnRoad, NonRoad, Railroad, Residential, and Electricity Production. The spatial variability for the NonRoad sector was not well-defined at the time of this analysis (every pixel in the domain was assigned some nonzero value), though this has since been improved. In addition to the absence of spatial information, NonRoad emissions represented only 2% of the total city emissions. For these two reasons, we omitted this sector in this study. All emissions estimates are provided at hourly resolution. However, except for electricity-generating processes which are required to report high frequency emissions (United States Environmental Protection Agency, 2006), temporal variability is often based on averaged weekly and diurnal cycles. The OnRoad sector emissions estimates come from traffic flow data when and where available. In this study, considering the limited time coverage due to the flask sampling strategy focusing on the early afternoon, only the spatial information is critical to attribute tracer gases to specific sectors. The Hestia inventory is used in this analysis to set the spatial boundaries for each economic sector, and also for the creation of pseudodata, as is described in detail later.

Using footprints for spatial identification

We used the Weather Research and Forecasting (WRF) (Skamarock and Klemp, 2008) model coupled with the Lagrangian Particle Dispersion Model (LPDM) (Uliasz, 1994) at 1-*km* resolution to produce footprints for the INFLUX tower sites, following the procedure described in detail in Lauvaux et al. (2016). For this study, footprints were calculated only from those towers where flask measurements were recorded during the investigated period, which are Towers 2, 3, 5, and 9, with Tower 1 serving as the background tower, as explained in detail later. The top-left panel of **Figure 3** shows an example 12-hour footprint run from Tower 2 for the flask measurement hour of November 10, 2012 at 19:00 UTC. These footprints were used to tie the atmospheric flask multi-species measurements at the associated towers to the spatially-defined economic source sectors outlined in Hestia as shown in the other panels of **Figure 3**.

The 1-*km* resolution WRF model covering Marion and the surrounding 8 counties provides an 87 × 87 box grid. The LPDM is used as the adjoint of the WRF-FDDA (Four Dimensional Data Assimilation) model to disperse backwards in time 6,300 particles per hour per measurement site. The particle positions are recorded every 2-minute and the trajectories are integrated over 12 hours to ensure all particles have had time to traverse the domain under any meteorological conditions. The surface-source influences are determined via the proportion of particles near the surface (below 50 meters), as detailed in Seibert and Frank (2004). Footprints are created for all days between September 1, 2012 and October 31, 2013, and this is the date range for flask measurements in the portion of the investigation that utilizes footprints.

Observing system simulation experiments (OSSE)

Building the Pseudodatasets

The initial analysis of the recorded multi-species atmospheric data will highlight the complex relationships many of these species have with CO_2ff. This will beg the question of whether it would be possible to disentangle these relationships, and further if this could be performed in a manner that accurately identifies tracer species with their corresponding CO_2ff sectors, if they exist. Since each of the CO_2ff source sectors is distinct, and since unique tracers for each source sector were not able to be identified a priori from the multi-species dataset, a theoretical investigation is undertaken.

In this investigation, a self-organizing map (SOM) is used to test whether sector-based emissions recorded in theoretical multi-species flask measurements can be properly attributed to their appropriate CO_2ff economic source sectors, under a wide range of emission scenarios, using information provided by footprints and the Hestia distributions of those economic sectors. The framework of this analysis is defined as follows.

We define a unique tracer for each source sector of interest:

Airport → *Tracer A*
Commercial → *Tracer C*
Industrial → *Tracer I*
OnRoad → *Tracer OR*
Railroad → *Tracer RR*
Residential → *Tracer Re*
Electricity Production → *Tracer EP*

Starting from unique tracers, we create an "emission matrix". Each row corresponds to one of the emission source sectors, and each column corresponds to a predefined tracer (cf. **Table 1**). Thus, each cell corresponds to the nature (type) of the emissions occurring from that source sector (row) from that tracer (column). We define 4 possible types of emissions in the emission matrices: no emission, emissions linearly scaled with CO_2ff, constantly emitting, and randomly emitting. These four types

Table 1: Example of an emission matrix, from which pseudodatasets are created. "Scaled" means the tracers will be scaled linearly with CO_2ff emissions, as in Equation 3 (where $\alpha = 1$). Similarly, "Random" means the tracer will be randomly emitting in the corresponding sector according to Equation 5. Thirty-three emission matrices are created to build 165 pseudodatasets. DOI: https://doi.org/10.1525/elementa.131.t1

Sector	Tracer A	Tracer C	Tracer I	Tracer OR	Tracer RR	Tracer Re	Tracer EP
Airport	Scaled	0	0	0	0	0	0
Commercial	0	Scaled	0	0	0	0	0
Industrial	0	0	Scaled	0	0	0	0
OnRoad	0	Random	0	Scaled	0	0	0
Railroad	0	0	0	0	Scaled	0	0
Residential	0	0	0	0	0	Scaled	0
Elec. Prod.	0	0	0	0	0	0	Scaled

represent the different scenarios described earlier. When being implemented into our OSSE's, each tracer is represented as follows:

$$No\ Emission = 0 \tag{2}$$

$$Scaled\ With\ CO_2ff_j = \alpha * CO_2ff_j = \alpha * \sum_{i=1}^{n_j}\left(X_{ji} * H_i\right) \tag{3}$$

$$Constant\ Emissions_j = \sum_{i=1}^{n_j}\left(X_{ji} * \beta\right) \tag{4}$$

$$Random\ Emissions_j = R * \sum_{i=1}^{n_j}\left(X_{ji} * \beta\right) \tag{5}$$

where j denotes the source sector of interest, X are the Hestia inventory flux values, H are the footprint (influence function) values, and the sum across $i \rightarrow n_j$ denotes the cumulative total through the n_j pixels where the footprint overlaps Hestia sector j. Here, $\alpha = 1$, $\beta = 1.0e4\ gCkm^{-1}hr^{-1}$, and $R \sim U([0, 1])$ (R is a random number sampled from a uniform distribution between 0 and 1). The constant number for β was determined in a trial-and-error process to approximately match the magnitudes of Hestia emissions at a semi-arbitrary flask measurement time. Note that we construct the emissions to all be around the same order of magnitude so that the SOM is not disproportionately influenced by one species over another. It is presumed that, in the case where this technique is applied to real data, each species' enhancement values would first undergo some normalization process to similarly ensure equitable treatment by the clustering algorithm.

Table 1 shows only one example of an emission matrix. In this example, each sector has its corresponding tracer emitting exactly scaled with CO_2ff following Equation 3, where $\alpha = 1$. Additionally, Tracer C (normally identified with the Commercial sector) will be randomly emitting in the OnRoad sector following Equation 5. Emission matrices like these delineate the framework around which the OSSE is performed.

For the purposes of this analysis, 33 different emission matrices are created to be representative of all the different possible emission scenarios. The number 33 came about as a consequence of there being 3 perfect-diagonal cases and there being 6 non-diagonal tracers in any row or column. By incrementally turning on the remaining tracers in a row or column, we gain groups of 6 representative matrices. Note that, in all of our scenarios, the idealized tracers were always at least set to be emitting in some capacity within their namesake sector (that is to say, no diagonal was ever set to 0). We pre-established five such groupings-of-six for testing how these differing emission scenarios would affect the SOM's ability to determine an accurate multi-species signature. These 33 matrices are defined as follows.

The first group, matrices 1–6 have each tracer along the diagonal (corresponding to each tracer's namesake sector) emitting scaled with CO_2ff, and Tracer C is also emitting scaled with CO_2ff in an incrementally increasing amount of additional sectors. For example, matrix 1 has Tracer C emitting in the Commercial sector and the OnRoad sector, and matrix 2 has Tracer C emitting in the Commercial sector, the OnRoad sector, and the Airport sector. This continues until matrix 6 has Tracer C emitting in every sector. Matrices 7–12 follow the same pattern, except all tracers that are emitting are constantly emitting rather than being scaled with CO_2ff. Similarly, matrices 13–18 follow the same pattern, but with all emitting tracers randomly emitting. As a variation, matrices 19–24 have constant emissions from the tracers along the diagonal, but the increments of Tracer C have random emissions in the off-diagonal source sectors. Matrices 25–30 start with the same diagonal as matrices 7–12 (constantly emitting), but instead of Tracer C incrementally increasing amount of sectors from which it constantly emits, the Residential sector itself incrementally gains an increasing amount of randomly-emitting tracers. For example, matrix 25 has constant emissions along the diagonal, but has Tracer A also randomly emitting in the Residential sector. Matrix 26 has the same diagonal, but Tracers A and C are both randomly emitting in the Residential sector. This continues until matrix 30, which has all off-diagonal tracers randomly emitting in the Residential sector. Finally, matrices 31–33 are the perfect diagonal sets for scaled emissions, constant emissions, and random emissions, respectively.

Each matrix is used as a template for creating a pseudo-flask measurement during the construction of a pseudodataset. For each case, the structure of a pseudo-flask measurement is similar to **Table 1**, where each row represents that sector's contribution to the pseudo-flask measurement. A full pseudodataset, then, includes one pseudo-flask measurement for every instance that a real flask measurement was recorded in Indianapolis during the period for which their corresponding footprints exist, September 1, 2012 through October 31, 2013. In addition to pseudodata created for the ideal tracers via the corresponding emission matrix, the pseudo-flask measurements which comprise the pseudodataset also include a column for the corresponding CO_2ff value (following Equation 3).

For each emission matrix scenario, five pseudodata cases are created. These all contain the original pseudo-flask measurements constructed from the emission matrices along with some amount of added columns for random ("noise") species: either 0, 1, 2, 5, or 10 additional columns. In these columns, the assigned values for all sectors and all tower measurements follow the exact calculation from Equation 5. The purpose of including varying amounts of noise species is to test how their presence in the dataset being analyzed affects the self-organizing map's ability to properly identify emission signals. By having these five different noise-species pseudodata scenarios for each emission matrix, we end up with 165 pseudodatasets, in total.

Self-organizing maps

This investigation attempts to address potential linear and nonlinear relationships between the emissions of CO_2ff and the other atmospheric species using a self-organizing map. Although only a few of the atmospheric species measured in the INFLUX dataset will be shown to exhibit strong linear correlations with CO_2ff, it still may be possible that non-linear emissions relationships exist. By using an SOM, we can investigate the detectability of these types of signatures, as well. Similar attempts to attribute measured multi-species enhancements to specific sources have been undertaken in the air quality community using techniques such as Positive Matrix Factorization (PMF) (Paatero and Tapper, 1994; United States Environmental Protection Agency, 2008). This technique is suitable in the case where information is available before analysis about the expected multi-species signatures of certain dominant sources (the sources' "factor profiles"). However, where this type of information is not readily available ahead of time (see **Figure 1**), a different multivariate approach may be preferred.

The two most common tools for reducing the dimensionality of a dataset are Principal Component Analysis (PCA) and Self-Organizing Maps (SOM's). The problem presented here also seeks to reduce dimensionality by condensing the multivariate mixing ratio measurements of the 35 trace gas species to distinct CO_2ff source signatures associated with each economic sector. Although both approaches have been used in a wide range of scientific studies, each also have their limitations. The PCA is limited by the fact that it seeks to optimize the problem by assuming that a linear combination of factors are able to represent the observed variability. Here, we attempt to associate tracers to any given sector of the economy without assuming a specific relation with CO_2 emissions or any other trace gases. Therefore, we seek a methodology able to detect co-located emissions which may or may not scale with any other tracers. For this reason, we used a neural network approach (here SOM) for our dimension-reduction problem.

The SOM is a machine learning algorithm which is used to identify similar values in a high-dimensional dataset through the use of a neural network (Kohonen, 1990). The network is established with a predefined number of nodes (neurons) in some predefined topology (usually organized as a series of squares or hexagons; hexagons in this investigation). The network is then inserted into the multivariate dataset and trained in an iterative fashion which deforms the nodes towards any clusters of datapoints within the dataset via competitive learning. An excellent visualization of this process is included in Tamayo et al. (1999). In this investigation, this means that the "classes" discussed later are initialized at a random position in the multispecies dataset, which is composed of enhancement (or pseudo-enhancement) values for all included species. For this analysis, after competitive learning identifies the "winning" neuron, the same procedure updates all neighboring neurons within some neighborhood of radius $d(i \in N_i(d))$ in accordance with the Kohonen rule:

$$W_i(q) = W_i(q-1) + \gamma\left(P(q) - W_i(q-1)\right), \qquad (6)$$

for weight vectors W attached to neuron i at iteration step q for input vector P, where γ is a monotonically decreasing learning coefficient. The final weight vector results can be used to map any dummy input vector to its corresponding representative node (also known as a neuron, and for much of this investigation will be called a "class").

In this analysis, we use the SOM to identify characteristic multi-species signatures of the CO_2ff source sectors. By doing so, we aim to identify which tracers are associated with which sectors, even if those tracers have nonlinear contributions. And to test whether such attribution is able to be achieved accurately with this methodology, the SOM analyses are run in an OSSE capacity, as explained in detail in the preceding subsection.

Defining a pseudodataset as previously described will yield an $a \times b$ two-dimensional matrix. The number of rows, a, is defined as $7 * N_{obs}$, where N_{obs} is the total number of flask (or pseudo-flask) measurements in the time period of interest and 7 represents the number of CO_2ff source sectors which have potential contributions. The number of columns, b, is defined as $7 + N_{noise}$, where N_{noise} is the number of noise species (0, 1, 2, 5, or 10) chosen to be included in that pseudodataset and 7 represents the number of predefined tracers (one for each potential source sector). As explained earlier, each row of the pseudodataset then is filled with pseudodata enhancement values for that sector's contribution in that pseudo-flask measurement from each of the columns' species as

defined by the predetermined emission matrix for this pseudodataset and using Equations 2–5.

First, the procedure for identifying the multi-species signature for each desired source sector is as follows:

1) The self-organizing map is trained on a pseudodataset. (For the uninitiated, this is where the neural network is deformed iteratively, following Equation 6, to fit the dataset).
2) All rows in the pseudodataset corresponding to one specific source sector of interest are averaged together to get a mean multi-species enhancement value signature for that source sector.
3) The mean signature (P in Equation 6) is simulated through the self-organizing map's network to output the corresponding class (i in Equation 6).

As explained earlier, the class number corresponds to a node in the neural network, and its position in the dataset will be the multi-species enhancement value signature for the nearest cluster of datapoints. By determining which class number is associated with the mean multi-species signature for a given CO_2ff source sector, we also gain that source sector's corresponding multi-species signature, which then can be analyzed for accuracy.

Results and Discussion

Justifying the OSSE with INFLUX observations

Linear regressions against CO_2ff

We use linear regressions of CO_2ff against the suite of gases measured in the INFLUX flasks to examine their potential relationships in an urban setting. The linear regressions are executed by fitting a first-order polynomial function to the scattered dataset using the least-squares method. These correlation plots aim to give a sense of potential relationships, and we use the coefficient of determination (r^2) as a simple diagnostic of those relationships. We do not express the ratio of gas: CO_2ff (the slope of the correlation), which would require using a least squares method that allows both variables to be independent (such as ordinary distance regression). In this context, r^2 acts as a first-order check on how well a species correlates with CO_2ff, and we do not explicitly account for measurement uncertainty. Instead, r^2 implicitly includes evaluation of both the noise generated by measurement uncertainty and the scatter due to variability in the sources and emission ratios of each species through the ~5 years of measurements. For most species, this range of variability is much larger than the measurement uncertainty, so that it is negligible by comparison. This is not the case with CO_2ff, which has an uncertainty of approximately 1 *ppm*, so this uncertainty will always contribute to an r^2 of less than unity. Thus, r^2 gives a simple metric to examine the types of relationships that occur in a real urban environment. We consider (a) the full flask dataset for each species, and (b) a pared dataset where the observations with the highest and lowest 5% of values for the non-CO_2ff species are removed. This "pared" dataset removes outlier points that can strongly influence the overall correlation, and we give some examples of reasons these outliers might occur and why it may be useful to exclude them later in this section. The full and pared r^2 values for all species measured in the INFLUX flask network are summarized in **Figure 1** and correlation plots for all species are given in the *Supplementary material*.

Figure 2 demonstrates the three general tracer-to-tracer relationships observed in the INFLUX flask observational dataset: linear, not strictly linear, and no obvious relationship. The regression of CO_{xs} versus CO_2ff_{xs} shows that linear relationships can be found for some trace gases. Because CO and CO_2ff originate from the same combustion processes and there are no other significant CO sources in the urban area (United States Environmental Protection Agency, 2006), the ratio remains fairly similar for the full dataset. However, the combustion ratio itself is known to vary between car engines, furnaces, and the power plant, and therefore may vary across space and time. Further, even in the absence of any atmospheric variability in the ratio, the uncertainty of each CO_2ff of ~1 *ppm* would induce a maximum r^2 of 0.8 in our dataset. Nonetheless, in this dataset, the CO:CO_2ff relationship appears consistent across multiple towers, seasons, and years (Turnbull et al., 2015a).

Figure 2 also shows the example of the *HFC*-134a_{xs} versus CO_2ff_{xs} relationship, that is not strictly linear. Because *HFC*-134*a* corresponds to leaks in mobile air conditioning systems, the emissions of *HFC*-134*a* do not scale with CO_2ff emissions in a direct sense. CO_2ff_{xs} and *HFC*-134a_{xs} do generally scale together, but there are some significant outliers that we speculate may be due to large *HFC*-134*a* leak events (such as air conditioner maintenance). Thus, the "pared" r^2 is also included, where the top and bottom 5% mixing ratio enhancements measured for each non-CO_2ff species are discarded, which improves the regression somewhat, indicating that there may be sufficient relationship between CO_2ff and *HFC*-134*a* to be used in a meaningful way.

Finally, the CO_2ff_{xs} versus $SF_{6,xs}$ plot shows no relationship, linear or otherwise, even when we use the simple paring technique of removing the top and bottom 5% of values. Although the primary use of SF_6 as a spark quencher in electrical facilities would suggest that it should correlate with power generation facilities and some previous studies have shown a relationship at larger spatial scales, none is observed here. In this case, we suspect that the non-co-location of sources at the urban scale and the small signal-to-noise ratio combine to produce this result. For example, CO_2ff emissions from large electrical facilities are emitted from the top of the smoke stack, which may be several hundred meters above ground level, whereas SF_6 leakage from electrical boxes at the same facility will be at or near ground level and possibly hundreds of meters away from the smoke stack. In the case of some other species, it may simply be that there is no relationship with CO_2ff (*e.g.* CH_2Br_2, which is produced from oceanic biological and chemical processes (Fuhlbrügge et al., 2016)).

These three plots in **Figure 2**, then, illustrate three distinct possible linear regression scenarios: an obvious linear relationship, a relationship that may exist but may

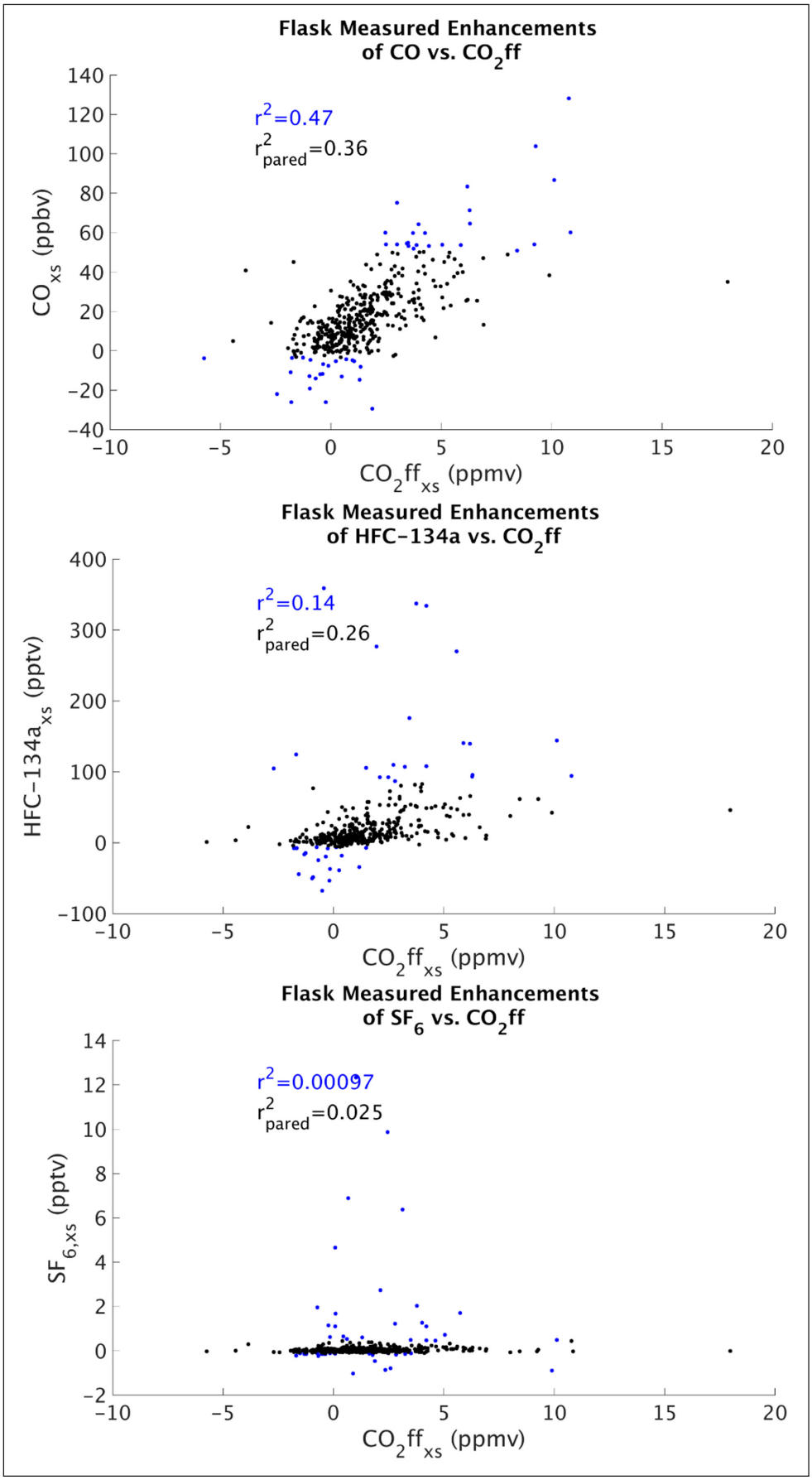

Figure 2: Three linear regression plots of species against CO_2ff_{xs} are shown, demonstrating an apparent linear relationship (CO_{xs}), an apparent relationship that is not strictly linear (*HFC*-134a_{xs}), and an apparent non-relationship ($SF_{6,xs}$). The error bars on CO_2ff are approximately 1 *ppm* and are left off for clarity. DOI: https://doi.org/10.1525/elementa.131.f2

not be linear, or obviously no relationship. Figures S1–S9 in the *Supplemental material* show the raw-data scatter plots for all species against CO_2ff for the interested reader. The calculated r^2 values from linear regressions of the raw data against CO_2ff are included as a column in **Figure 1** and indicate that, in this dataset, very few species have an appreciable correlation–only 4 of the 34 species even have an r^2 value above 0.2. These are much lower than those found in previous studies such as Turnbull et al. (2011) and Miller et al. (2012). We believe that this is a consequence of the relatively short distance between emission and detection in the INFLUX urban sample network. This leaves very little time for signals from multiple source sectors with different (or zero) emission ratios to become well-mixed before detection. Conversely, the low r^2 values imply that different sources are observed on different days, and consequently suggests that indeed it might be possible to separate these sources using observations.

Miller et al. (2012) also found that, by separating out the seasonality of their signal, the winter measurements yield stronger correlation coefficients. For the Indianapolis dataset, separating out the winter signal (defined as measurements in December through February) yielded only marginal improvements in the correlation coefficients, with now 6 species having an r^2 value above 0.2, where Benzene (C_6H_6) is the only species besides *CO* (0.58) with a value above 0.3 at 0.31. The relative changes in r^2 values from the INFLUX winter signal compared to the full dataset are small enough as to not warrant further investigation for the purposes of this manuscript.

Towards direct spatial attribution to CO_2ff sectors

Identifiable relationships between gases and CO_2ff source sectors, even if nonlinearly related to CO_2ff, are useful to inform about the sector origins of any enhancements. We present here the spatial attribution of flask measurements by comparing the exact footprints derived from our backward model simulations for each flask sampling time against the known spatial extent of the source sectors as defined in Hestia.

Figure 3 shows that the footprint from a single tower overlaps with multiple economic sectors within the Indianapolis domain. Here, the footprint is plotted in the top left panel, and its overlap with each sector fills in the remaining panels (overlapping pixels being colored yellow). Given enough tower footprints, we expect an appreciable amount of variation in the number and type of sectors being overlapped. By relating the corresponding flask measurements to this wide variety of footprints, we want to determine which measured species are associated with different source sectors considering their observed enhancements.

In a case lacking sufficient information from the footprint/sector overlaps, there is no mathematical way to properly distinguish which sectors have strong (or any) relationships with any given species' mole fraction enhancements, without including additional a priori information about sectoral emissions. Here, five of the seven economic sectors of interest are overlapped by every footprint except for the OnRoad sector (intersected 88 out of 89 footprints). In this study, we will refer to the "domain filling problem" to describe this lack of sector-specific observations. **Figure 4** illustrates the attribution problem assuming that any sector located within the tower footprint can be associated with a corresponding atmospheric enhancement. This problem is presented in detail in subsection *Introducing the Domain-Filling Problem.*

The self organizing map analysis

Observing system simulation experiments

As discovered during the *Linear regressions against* CO_2ff subsection, the assumption that a tracer's emissions will be linearly correlated with CO_2ff emissions from a given sector is unlikely (only 4 of 34 species has an r^2 above 0.2). Trace gas emissions originate from processes which have been grouped into economic sectors rather than process-based sectors. The greater the number of processes associated with a sector, the greater the chance that the CO_2ff emissions will not have a singular tracer for which emissions are linearly proportional. There may even be variability in the proportionality of emissions from within the same process (*e.g.* different engines for different vehicles may emit the same chemicals in different proportions, but all are counted as traffic emissions in the OnRoad sector). Thus, we implement Observing System Simulation Experiments (OSSEs), with the framework described in the *Methodology* section, to explore under what circumstances accurate tracer attribution may be gleaned. This can help determine what types of tracers would be able to be detected–including nonlinear emissions with respect to CO_2ff–and how much prior information is necessary for accurate source sector attribution. Nonlinear relationships can be caused by irregular emissions within a CO_2ff sector or by variations in the relative contributions of CO_2ff sectors with consistent gas: CO_2ff emission ratios. The attribution problem described here assumes that the spatial distributions of sectors from Hestia are correct. We also assume that the background mole fractions, which introduce additional errors in atmospheric enhancements as explained in Lauvaux et al. (2016), are known quantities.

Sector signatures and measures of success

Using the framework described in the *Methodology* section above, we want to identify multi-species signatures for each CO_2ff source sector and evaluate the SOM's capabilities for sector attribution of any pseudodatasets. Because every row of enhancement values in a pseudodataset will be assigned to a class (neuron) of the SOM, we evaluate a CO_2ff source sector's signature by finding the class assigned to the mean of all pseudodata rows associated with that sector's contributions and comparing it against the emission matrix which was used to define that pseudodataset. We describe hereafter how we determine quantitatively whether the assigned class matches the original pseudodata. **Figure 5** is used for demonstrative purposes. The SOM's classification result for the Industrial sector is chosen for this example, because it exhibits

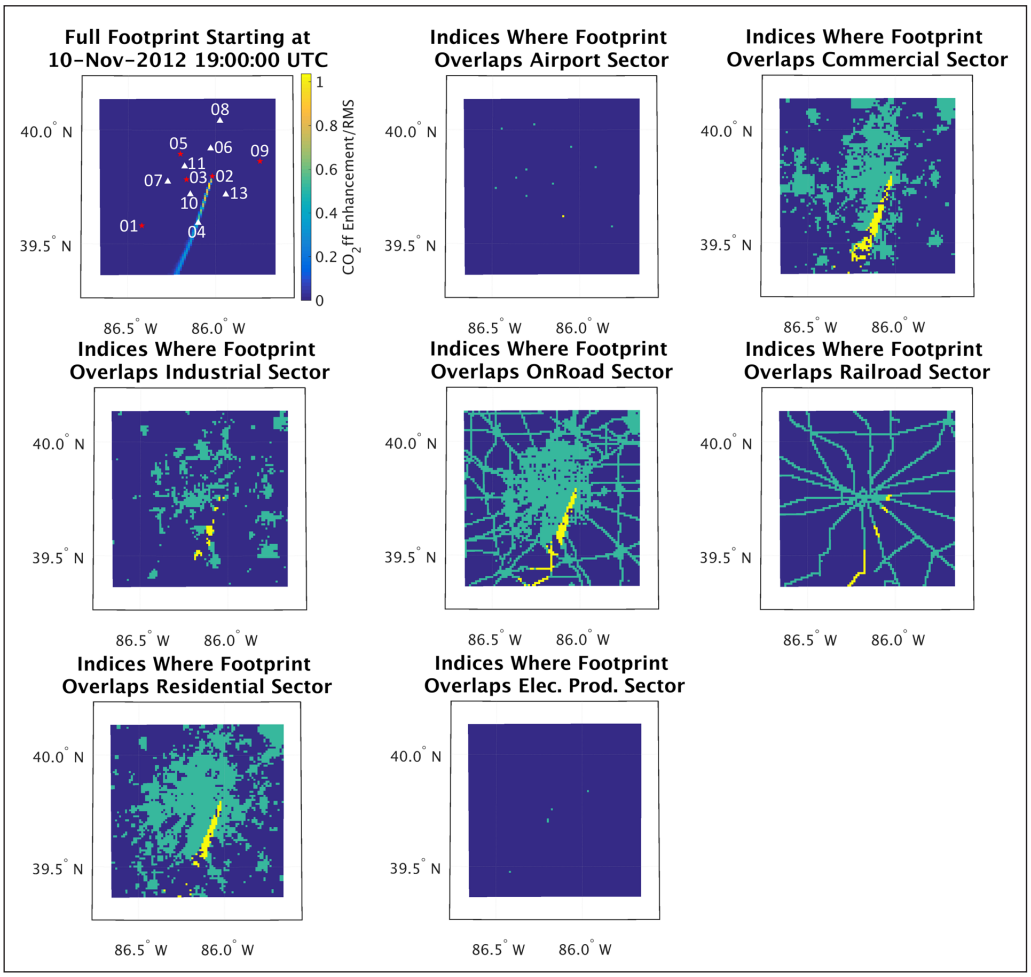

Figure 3: The top-left panel shows an example footprint from the one-hour flask measurement at Tower 2 on November 10, 2012 at 19:00 UTC. White triangles denote INFLUX tower sites and red stars indicate those towers where flask measurements were recorded during the period of this investigation. The remaining panels show the spatial extent of the Hestia flux maps for each source sector in green, with the pixels overlapping the footprint in yellow. These maps show only whether any emissions could occur and do not indicate the relative magnitude. DOI: https://doi.org/10.1525/elementa.131.f3

many of the potential difficulties which will be discussed in detail, especially around the idea of properly defining if an assigned class has been "successful" in its tracer assignments; it is not meant to be necessarily viewed as a representative case.

As **Figure 5** demonstrates, the question of a successful classification is wrought with nuance. The classification result in the bar chart on the right is the result of a 1000-class SOM classification for a pseudodataset based on the emission matrix on the left. For this source sector (here the Industrial sector), the emission matrix defines Tracer C and Tracer I emitting linearly with CO_2ff emissions. The scale factor (α in Equation 3) is set to 1, so Tracer C and Tracer I should have the same enhancement values as CO_2ff. No other tracer emits in this source sector, and 10 noise species are added to evaluate the impact of irrelevant flask data that may have been mistakenly included if this were a real-data multi-species analysis.

The bar chart in **Figure 5** shows that four of the random noise species are found to have higher average enhancement values than any of the known tracers and CO_2ff. Recall that the noise species are defined to fluctuate around the same order of magnitude as CO_2ff (and that any real-data analysis will be expected to similarly put each species' measurements in the dataset through some normalization process before analysis to avoid biasing the clustering), so this represents an analysis scenario where 10 species in a dataset are unknowingly being emitted unpredictably and indiscriminately within the domain of interest. Although the we can test for apparent relationships before analysis, these noise species account for cases of misjudgment. Tracer C and Tracer I are correctly classified by the SOM to have enhancement magnitudes equal to CO_2ff, however these magnitudes are only the fifth through the seventh largest among measured species for this source sector, respectively. Thus, the question of whether or not the SOM's classification is successful is inherently dependent on how the analyzer (the subjective interpreter) defines that success. Three metrics are created here for gauging the success of any given classification, called "Success Rate", "Top Rank", and "Harsh Top Rank", which are defined as follows:

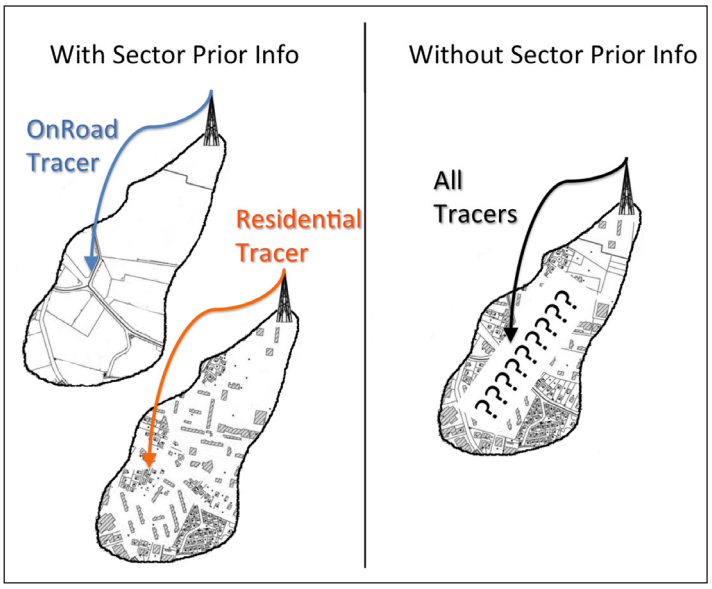

Figure 4: Cartoon demonstrating the domain-filling problem. The left panel shows the ideal case where each source sector's spatial distribution and corresponding tracer are known a priori. In the right panel, representing the real-world case, the lack of a priori bottom-up flux inventory information for most of the measured atmospheric species used in this analysis is a problem. Emissions could have come from anywhere within the atmospheric footprint. There is no way to precisely and accurately identify a source within this area using just this individual footprint, although it may be possible to overcome this limitation given enough footprints with a wide enough variety of source sectors being overlapped. DOI: https://doi.org/10.1525/elementa.131.f4

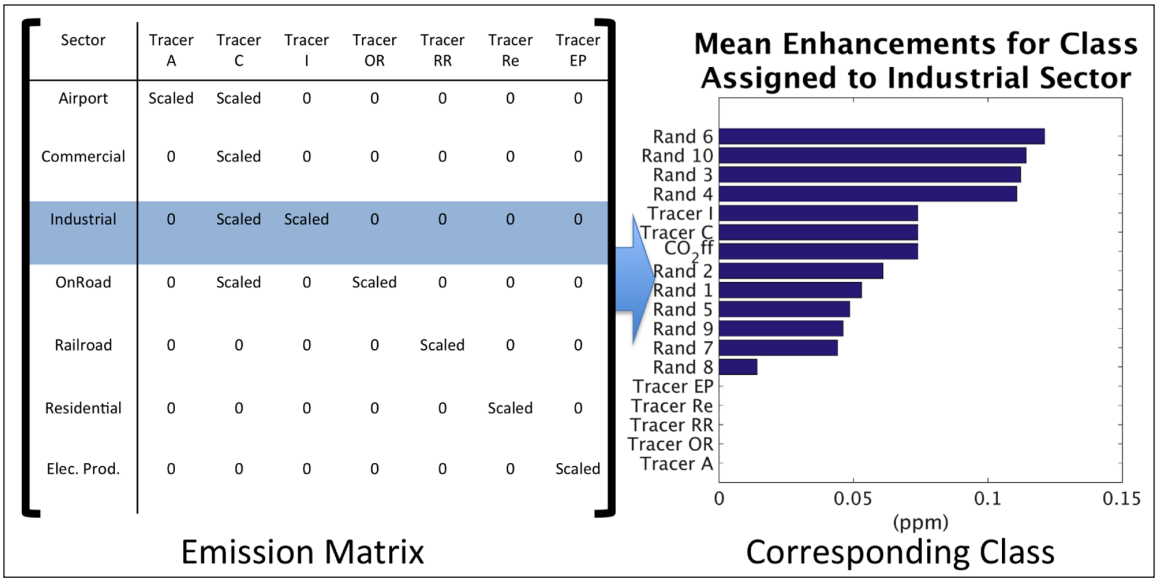

Figure 5: An example of a final class definition for the Industrial sector, derived by a self-organizing map of 1000 nodes (classes), using an emission matrix where Tracer C and Tracer I are the only tracers expected to be emitting, as highlighted by the light blue box over the Industrial sector's row. Here "Scaled" means that the tracers should have enhancements with magnitudes linearly scaled with CO_2ff (in this case, the scale factor = 1, so they should have identical values). In this example, four of the noise species in the dataset have higher mean enhancements than the known emitting tracers, which may be problematic in a real-world case where it was not known ahead of time which tracers are expected to be emitting. DOI: https://doi.org/10.1525/elementa.131.f5

The "Success Rate" looks only at the tracers of the emission matrix (not the noise species or CO_2ff) and is defined as the fraction of those tracers which are correctly identified as emitting or not emitting based on their emission matrix designation. In mathematical terms:

$$Success\ Rate = \frac{N_{cor}}{N_{tracer}}, \tag{7}$$

where N_{cor} is the number of tracers correctly identified as on or off and N_{tracer} is the total number of potential tracers (N_{tracer} = 7, for all cases in this investigation). In the

example of **Figure 5**, the Success Rate would be $\frac{2+5}{7} = 1$. The Success Rate metric does not concern itself with the magnitude of the enhancement.

The Top Rank metric evaluates specifically among the subset of known tracers from the emission matrix which are supposed to be emitting, meaning all of those non-noise, non-CO_2ff species from the emission matrix which are not set as zero when building the pseudodataset. This thus also assumes that any random noise species will be able to be identified and ignored prior to analysis of any real dataset. All of the known possible tracers are sorted based on their enhancement values. The Top Rank asks the question of if the tracers that were supposed to be emitting in a given sector (based on the corresponding emission matrix) are the same tracers found to have the highest mean enhancement values among all potential tracers in that sector:

$$Top\ Rank = \frac{N_{cor\_top}}{N_{emit}}, \tag{8}$$

where N_{emit} is the number of tracers which are supposed to be emitting (number of emitters), and N_{cor_top} are the number of emitters which correctly occupy the top N_{emit} spots of the N_{tracer} values that have been sorted by enhancement value.

In the example of **Figure 5**, we can determine the Top Rank as follows. We know that there are always 7 potential non-noise, non-CO_2ff tracers in the emission matrix corresponding to the 7 CO_2ff source sectors. Of these 7, the emission matrix on the left of **Figure 5** shows that 2 tracers were assigned to be emitting when the pseudodataset was built: Tracer C and Tracer I. This means that $N_{emit} = 2$. If we then sort the enhancement values for all tracers of the emission matrix (that is, ignoring the noise species), then how many of the top 2 spots (largest enhancement values) are occupied by either Tracer C or Tracer I? By looking at the sorted enhancement values plotted in the right of **Figure 5**, we see that the answer is that both of them occupy the top 2 spots, so $N_{cor\_top} = 2$. Thus, the Top Rank would be $\frac{2}{2} = 1$.

While the Top Rank metric has the advantage of accounting for the magnitude of the enhancements, it also introduces the assumption that the tracers which are emitting will have the largest mean enhancements. One of the issues with this assumption is that the accuracy of this metric becomes highly dependent on the assumption that any noise species have been accurately identified and removed from the dataset prior to analysis.

To address this secondary assumption, the Harsh Top Rank metric is defined as follows:

$$Harsh\ Top\ Rank = \frac{N^*_{cor\_top}}{N_{emit}}, \tag{9}$$

which is calculated after sorting every species in the dataset (N_{tot}) by enhancement value–including the random noise species and CO_2ff. Similar to Top Rank, $N^*_{cor\_top}$ is the number of expected emitters (from the corresponding emission matrix) correctly occupying the top N_{emit} positions of the sorted N_{tot} dataset.

In the example of **Figure 5**, the Harsh Top Rank is thus determined as follows. The total number of species in this pseudodataset is 18: 7 tracers from the emission matrix, 10 noise species, and CO_2ff. Of these, we know that two tracers are expected to be emitting based on the emission matrix in the left of **Figure 5**: Tracer C and Tracer I. This means that $N_{emit} = 2$, exactly as with Top Rank. After having sorted the species by enhancement value, as shown on the right of **Figure 5**, we ask how many of the top 2 spots among all 18 species are occupied by Tracer C and Tracer I. The top 2 spots among all species in the dataset are clearly occupied by Rand 6 and Rand 10, however, so $N_{cor\_top} = 0$. The Harsh Top Rank value for this classification, then, would be $\frac{0}{2} = 0$. This is considered to be a more realistic metric, as noise species can negatively influence the success value.

Control experiments: Proof of concept

A total of 495 self-organizing map classifications are performed, consisting of 165 pseudodatasets from 33 emission matrices with 5 different amounts of noise species, all run with 3 different amounts of classes (nodes): 100, 200, and 1000. The runs with different amounts of classes are performed since the optimal number will depend on the geometry of the dataset and is not known prior to classification. The ideal number of classes has enough to capture all of the similar groupings of datapoints (clusters) without having multiple classes assigned to the same grouping.

Since this analysis is being done to address whether proper CO_2ff source sector attribution is possible under ideal circumstances, these 495 trials represent the most ideal case for the INFLUX flask dataset. In these original 165 pseudodatasets, it is already known ahead of time which tracers are emitting from which sector, and the exact spatial bounds of the economic sectors are known (from Hestia). We also assume that the number of noise species not tied to a source sector is known. This initial evaluation will provide our initial control experiments from which to evaluate the loss of information on the sector attribution.

Figure 6 shows the mean results of the OSSE SOM classifications across all success metrics for each of the source sectors. The Airport and Electricity Production sectors have low and sporadic successes across all metrics, presumably because these sectors have small spatial signatures–only a few scattered pixels in the domain–which are not able to be well captured in the model. In all of the other sectors, though, the self-organizing map performs well across all success metrics. For these sectors, the Success Rate and Top Rank metrics are around the 90–100% range. The Harsh Top Rank is closer to the 70% range, though, showing the dramatic negative impact that is brought by the inclusion of extraneous species measurements in a dataset. While 90–100% may arguably be enough to justify using this method for identifying the tracers emitting in a source sector, 70% is low enough that this becomes difficult to justify, and starts to beg the question of where the acceptability threshold really lies. Restated as a real-world example: it may be easier to justify using the attribution

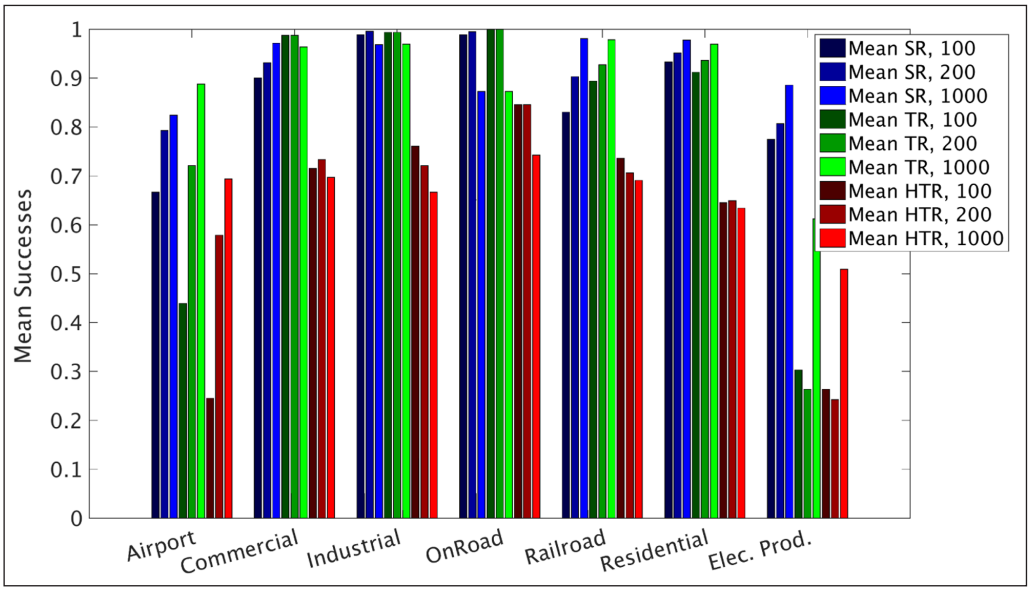

Figure 6: The mean successes in the control case across all pseudodatasets are shown, for each source sector, specified by the type of success metric (SR = Success Rate, TR = Top Rank, and HTR = Harsh Top Rank) and by the number of classes in the SOM. The Success Rate and Top Rank have very good success (between 90–100%) for all source sectors except the Airport and Electricity Production sectors. Presumably this is because the Airport and Electricity Production sectors are only comprised of a few pixels in the domain, so the model has difficulty distinguishing their signals. For the sectors which the model was able to characterize well, the Harsh Top Rank only had success values around 70%. DOI: https://doi.org/10.1525/elementa.131.f6

results in a future analysis such as an inversion if only 0–1 attributions out of 10 are potentially wrong, compared to the case where 3 in 10 may be wrong. We will not attempt to determine what this acceptability range is, only stating when results are clearly unacceptable (*e.g.* worse than a coin-flip).

In keeping with the idea that this control scenario is aiming for the most generous conditions to determine whether proper source sector identification is at all possible, **Figure 7** shows the relative successes for each pseudodata noise scenario across all emission matrices, but only for those source sectors which provided good results in **Figure 6**. Additionally, **Figure 7** is focused on the 200-class SOM case, which seemed to perform the best of the three cases (100-class, 200-class, and 1000-class), and it shows only the results from the Top Rank success metric, for which the classification had the highest success values. These choices of 200-classes and the Top Rank metric will maintain consistent through the rest of the analysis as adjustments are made.

For **Figure 7**, each shaded region moving across the x-axis represents one of the incremental group-of-six emission matrices explained in the Methodology section (cf. subheading "Building the Pseudodatasets"). A brief description of how the tracers are defined in each scenario is included underneath the scatter plot. Emission matrices 1–6 (shaded gray) represent the emission matrices where the "Expected Tracers" (i.e. related to the sectors) down the diagonal of the emission matrices were all set by applying a multiplicative factor (set to 1 here) to the original CO_2ff emissions (thus referred to as "Scaled"). The off-diagonal "Unexpected Tracers" are also emitting "Scaled" and are incrementing "Across Sectors", moving down a column of the emission matrix. (For example, emission matrix 1 is set "Scaled" along the diagonal, but also has Tracer C emitting "Scaled" in the OnRoad sector. Emission matrix 2 is identical but also includes Tracer C emitting "Scaled" emissions from the Airport sector.) Because each emission matrix has 5 pseudodatasets associated with it, corresponding to the inclusion of 5 amounts of noise tracers, the success values for all 5 pseudodatasets are plotted in **Figure 7** for each emission matrix. When results remain independent of the number of noise tracers, dots overlap. According to this Top Rank metric, the success values for tracer assignment are consistently high across our different cases.

Figure 7 shows that the metrics tend to perform better in scenarios with fewer noise species in the dataset. The real-world implication of this is that species whose sources are not well-constrained should be removed from a dataset before beginning analysis. For the INFLUX flask data, this would at least mean removing all species with gray boxes in **Figure 1** unless one is able to create or obtain emission maps for them. This point is accentuated if compared against the Harsh Top Rank metric, shown in Figure S12 in the *Supplemental material*. Additionally, in scenarios where the tracers were scaled with CO_2ff emissions, proper identification was lowest among all emission matrix scenario groups. This is believed to be a result of the real CO_2ff emission data, after being divided into sector contributions, having values close enough to 0 on the *ppm* scale that the SOM's nodes for one sector may be influenced by contributions in another, despite being defined as independent in the emission matrix from which the dataset was constructed. These influences lead to misclassifications, and, being near-zero, these misclassifications are proportionally greater in magnitude, which

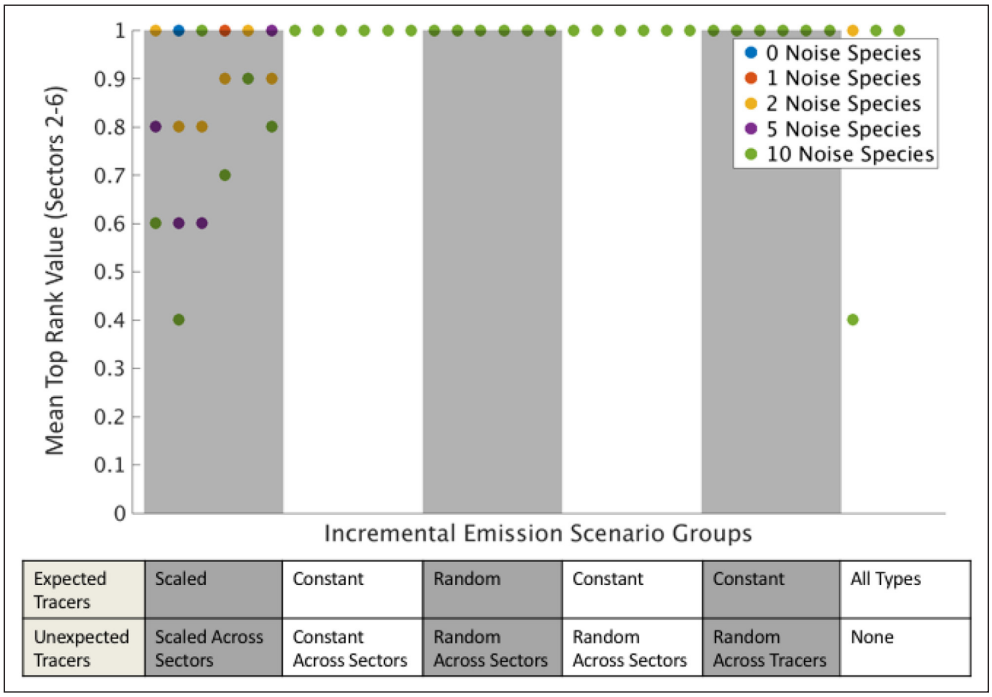

Figure 7: The mean successes using the "Top Rank" metric for each noise scenario across all the different emission matrices. The table below the scatter plot identifies which incremental emission scenario the points correspond to. The 200-class case is chosen for each of the success metrics, as it generally had greater agreement than the 100- and 1000-class cases. The Top Rank is the most generous metric of success, so will be the benchmark metric for the most ideal conditions. DOI: https://doi.org/10.1525/elementa.131.f7

can lead to mis-ordering and therefore failures according to the Top Rank formula.

Introducing the domain-filling problem

Considering that, in the control case, the identification of most sectors with the self-organizing map is successful, we evaluate whether these successes maintain their performance after introducing the domain-filling problem identified earlier and illustrated in **Figure 4**. For the control experiment, the pseudodatasets were constructed as a series of pseudo-flask measurements whereby the exact contributions from each source sector were known (defined by the corresponding emission matrix) and were explicitly stated in their own rows. The domain-filling case uses the same pseudo-flask measurements as a starting point, mimicking the real-world case, and projects them back into the domain using only the tower footprints. The new domain-filled sector values, then, are derived with the following formula:

$$\Delta_{yj} = \Delta_y * \frac{n_j}{n_{tot}}, \tag{10}$$

where Δ_y is the pseudo-flask enhancement for species y, n_j is the number of pixels where the footprint overlaps Hestia sector j, and n_{tot} is the total number of pixels contained in the footprint. An approach like formula 10 is necessary to properly account for there not being any a priori information about any species' spatial emissions distribution within any given footprint, as is true for nearly every species in the real flask data case. The formula here in equation 10 is chosen instead of maintaining the form of equation 3, because direct reliance on the influence function would heavily favor identifying sources as being nearby the tower.

Using this methodology, all 165 pseudodatasets are reconstructed, and the 495 self-organizing map trials are rerun. **Figure 8** shows the new results for each noise scenario across all the emission matrices. In keeping with the control analysis, the 200-class case and the Top Rank success metric are used.

The results of **Figure 8** show success values around 40%, incrementally increasing to near 70%. The apparent increase in success (up to 70%) is an artifact based on the formulation of the Top Rank equation (Equation 8). In the domain-filling problem, by definition, every tracer will be present in every spatially-defined sector within a footprint. According to equation 10, the relative sectoral enhancement will be proportional to the relative spatial fraction of the footprint occupied by that sector. To understand this artifact, note first that the denominator of the Top Rank equation is the number of tracers in the emission matrix expected to be emitting. Thus, as one steps across any of the groups of 6 matrices which are incrementally increasing the number of tracers which are emitting in the sectors, the denominator grows. If one also notes that all tracers are always present within a footprint in the domain-filling case, then one can see how N_{cor_top} grows as N_{emit} grows, and the Top Rank metric artificially looks more successful. This feature is ignored in the results.

Considering the other scenarios, the highest success values are around 20–40%, depending on the emission scenario group. This shows that the Top Rank, considered the most generous success metric, indicates that the inclusion of the domain-filling problem alone makes proper

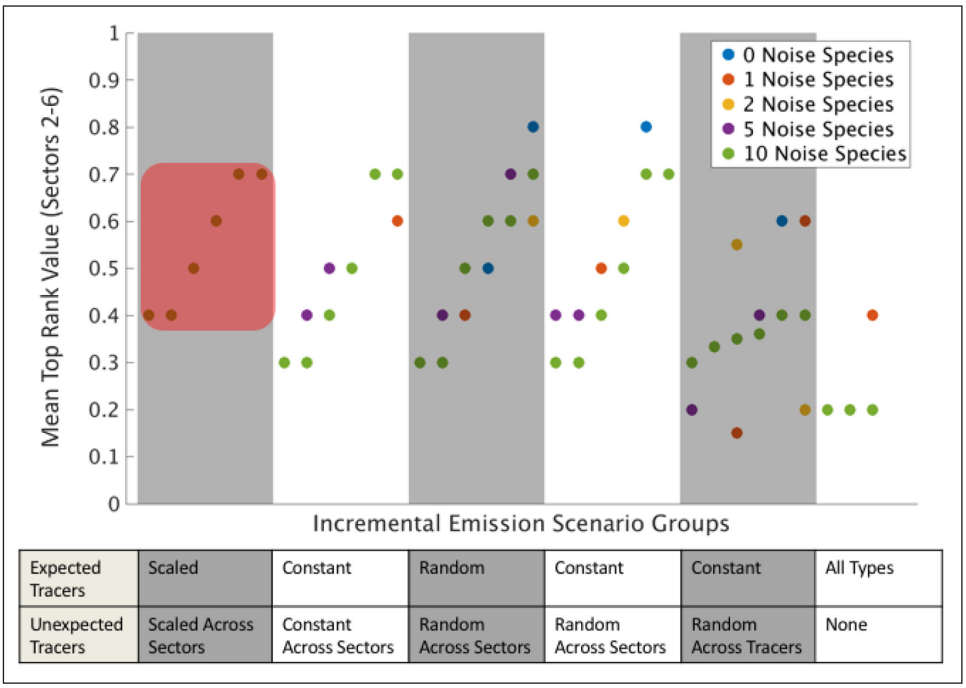

Figure 8: The mean successes for each noise scenario across all the different emission matrices for the pseudodatasets with the domain-filling issue. As with Figure 7, the 200-class case is chosen for the Top Rank success metric. Highlighted with a red box, the Top Rank values start at near 40% and apparently improve, but this is an artifact. Since the improvements are artifacts, the numbers across the emissions scenarios should be considered as in the 20–40% range. DOI: https://doi.org/10.1525/elementa.131.f8

sector identification with an SOM nonviable. Indeed, the other success metrics do fare worse. The Success Rates are in the 20–30% range after the control case had them around 90–100%, as can be seen in the *Supplemental material* Figure S13. The Harsh Top Rank metric benefits in part from the same artifact as the Top Rank metric, but even with this some of the scenarios with higher amounts of noise species hover around 0%, as can be seen in the *Supplemental material* Figure S14. With none of the success metrics having a trustworthy value above 50%, none of them could be considered useful for source sector identification.

The "all towers" case: Attempting to improve results through an expanded dataset

The two pseudodata scenarios (with and without the domain-filling problem) are rerun with an expanded dataset. Rather than only creating a pseudodata measurement from each tower with a real simultaneous flask measurement (on average ~2 towers per measurement time), a pseudodata measurement is simultaneously created for all functional towers in the domain every time any real flask measurement was recorded. In this way, we evaluate the impact of data availability on the domain-filling problem, assuming that there were not enough data points previously to allow the self-organizing map to compensate for the information lost through domain-filling. These tests are referred to as the "all towers" cases.

First, the control results are rerun under the "all towers" conditions. **Figure 9** shows the comparison between the two, with the original control results in blue and the "all towers" results in red. In contrast to **Figures 7** and **8**, only the results of the first emission matrix in any incremental group are included, as well as the results of each perfect diagonal emission matrix. This change is made so that the domain-filled comparison later will no longer include the aforementioned Top Rank artifact which falsely implies improvement of the success values as they increment within the emission matrix scenario groups. The new "all towers" control results are very close to the original control results.

Figure 10 shows that expanding the dataset to include additional measurement locations from all available towers in the INFLUX domain at any measurement time does not overcome the domain-filling problem. The relative successes among the emission matrices are almost identical to the original analysis. Additionally, there is an apparent dip for the perfect diagonal cases, which appears to be a result of the same artifact from **Figure 8**. If anything, though, this implies that the real Top Rank success values in the domain-filled scenarios are closer to 20% than 40%. Regardless, neither of these results for the domain-filling cases would be viable for justifying this methodology for source sector attribution.

Conclusions

In this study, we examined the use of multi-species flask measurements to attribute atmospheric signals emanating from specific CO_2ff economic source sectors at the scale of an urban domain. Initial simple correlation plots of actual observations from the INFLUX network showed that many species are related to CO_2ff; the relatively weak correlations suggest that multiple sources contribute to the overall signal and in principle these could be separated

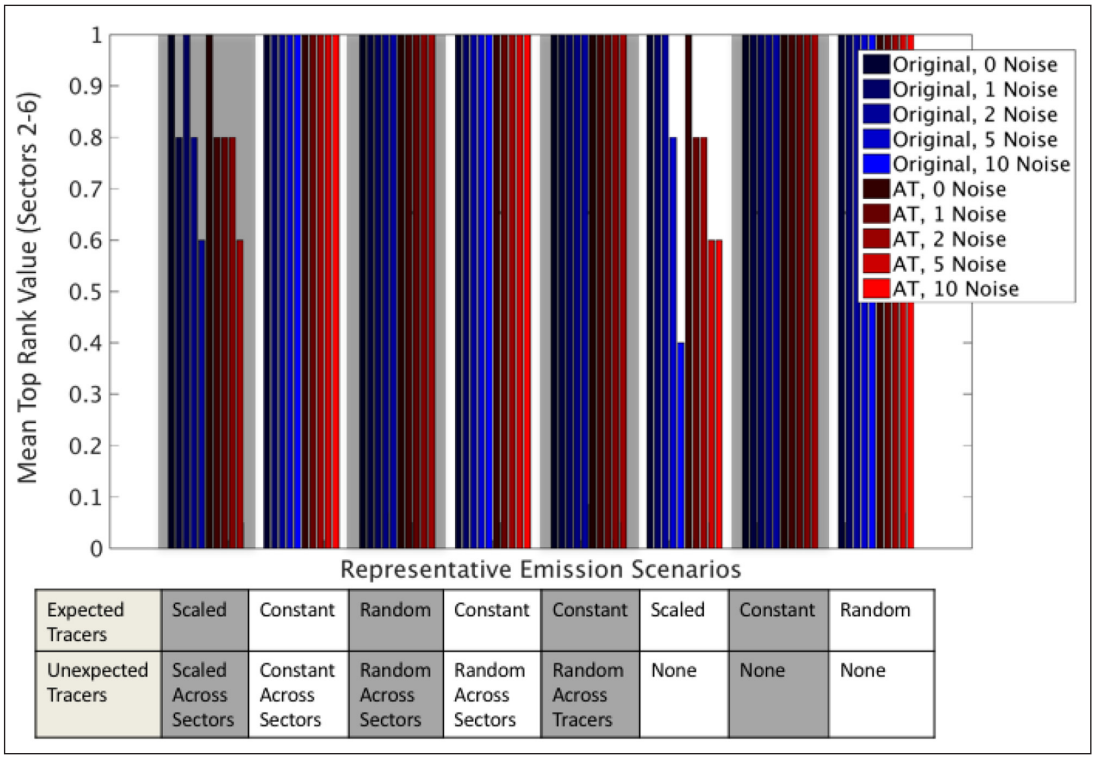

Figure 9: The mean successes using the Top Rank metric are shown for each noise scenario for both the original control case (blue) and the "all towers" case (red). Compared to Figures 7 and 8, only the first emission matrix scenario from any incremental group is included, as well as each perfect diagonal emission matrix scenario. As would be expected, the Top Rank successes in the "all towers" case are at least as good as those in the control case. DOI: https://doi.org/10.1525/elementa.131.f9

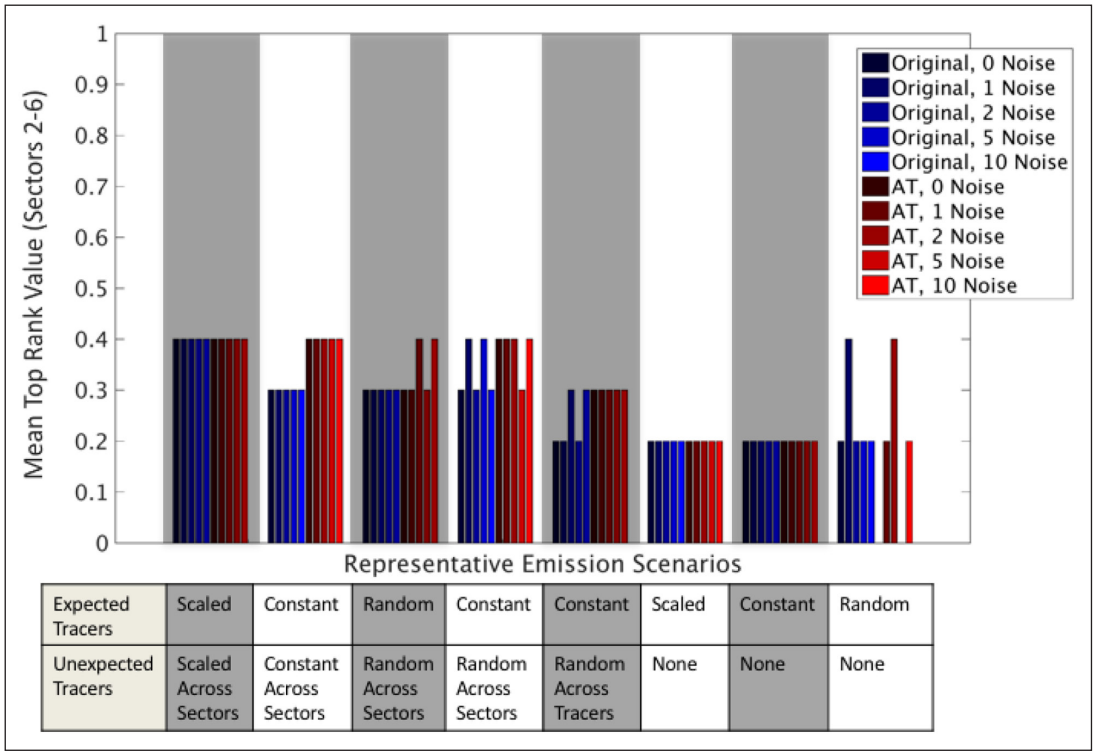

Figure 10: The mean successes with the domain-filling problem according to the Top Rank metric are shown for each noise scenario, with the original in shades of blue and the "all towers" case in shades of red. The same emission matrix scenarios as in Figure 9 are included to avoid having the Top Rank false-improvement artifact of Figure 8. Unfortunately, the expansion of the dataset for the "all towers" case did not greatly improve the success results. DOI: https://doi.org/10.1525/elementa.131.f10

to evaluate the contribution of these different sources to the total emissions.

We used an OSSE to explore how an SOM could be used to attribute these urban observations to source sectors. All analyses were performed in a highly unconstrained fashion–without any other a priori information (such as emission ratios or bottom-up inventories) about any of the non-CO_2ff species. A high-resolution meteorological model was applied to each flask measurement to find the footprints within which emissions could have occurred. The overlaps of these footprints with predefined CO_2ff emission source sectors from the Hestia inventory were scrutinized to identify any measured atmospheric species that could constrain atmospheric CO_2ff measurements to their appropriate source sectors.

The direct-attribution attempts with real data did not yield useful results, because of the large effects of several real-world sources of uncertainty. One such source of uncertainty in this analysis concerns which measured species are specifically known a priori to be related to which (if any) CO_2ff source sectors, as defined in the Hestia dataproduct. Further, a consistent and prominent source of uncertainty comes from the lack of a priori knowlege for the expected spatial distribution of emissions for any non-CO_2ff measured species within any projected footprint. This last source of uncertainty, labeled the "domain-filling" problem, makes it unfeasable to assign sectors for the real-world case, where virtually every tower footprint overlaps 5 of the 7 source sectors every time. This problem forced the investigation into the realm of pseudodata.

The OSSE pseudodata investigation showcased the self-organizing map's ability to successfully identify emissions from tracers that are or are not emitting in source sectors with adequate spatial coverage within a flask measurements' footprints. This is, on its surface, very encouraging for solving the problem of source sector attribution of greenhouse gas measurements. However, we know that in the real-world case, for any flask collected at a given time, the exact locations of the trace gas sources are nearly always undefined within the flask footprint. We showed here that the self-organizing map approach is unable to overcome this "domain-filling" problem to correctly attribute trace gases to specific sectors, even in a case with a greatly expanded dataset. The inability to overcome this limitation suggests that the real-world source apportionment problem may be irreconcilable without additional a priori information related to the processes emitting all observed gases of interest. It is additionally shown in the OSSE analysis that the inclusion of extraneous species in the dataset significantly hinders the ability to associate tracer emissions to their source sector. Only atmospheric species that are specifically relevant to the source attribution problem should be included in further analyses. Future investigations will aim at pre-identifying individual species-of-interest and at gathering first-guess spatial inventory estimates to address both the domain-filling problem and the issue of including unnecessary species. Our results suggest that an obvious path forward is to partition CO_2ff by source sector at the whole city scale, and/or to develop higher resolution a priori information to inform future studies.

Data Accessibility Statement

The data used in this study are available at http://sites.psu.edu/influx/data/description-univ-of-coloradonoaa-flasks/flask-data/.

Supplemental Files

The supplemental files for this article can be found as follows:

- **Figure S1.** Scatter plot of raw flask data against CO_2ff. DOI: https://doi.org/10.1525/elementa.131.s1
- **Figure S2.** Scatter plot of raw flask data against CO_2ff. DOI: https://doi.org/10.1525/elementa.131.s2
- **Figure S3.** Scatter plot of raw flask data against CO_2ff. DOI: https://doi.org/10.1525/elementa.131.s3
- **Figure S4.** Scatter plot of raw flask data against CO_2ff. DOI: https://doi.org/10.1525/elementa.131.s4
- **Figure S5.** Scatter plot of raw flask data against CO_2ff. DOI: https://doi.org/10.1525/elementa.131.s5
- **Figure S6.** Scatter plot of raw flask data against CO_2ff. DOI: https://doi.org/10.1525/elementa.131.s6
- **Figure S7.** Scatter plot of raw flask data against CO_2ff. DOI: https://doi.org/10.1525/elementa.131.s7
- **Figure S8.** Scatter plot of raw flask data against CO_2ff. DOI: https://doi.org/10.1525/elementa.131.s8
- **Figure S9.** Scatter plot of raw flask data against CO_2ff. DOI: https://doi.org/10.1525/elementa.131.s9
- **Figure S10.** Success metric comparison with domain filling. DOI: https://doi.org/10.1525/elementa.131.s10
- **Figure S11.** Success Rate metric comparison across emission scenario groups. DOI: https://doi.org/10.1525/elementa.131.s11
- **Figure S12.** Harsh Top Rank metric comparison across emission scenario groups. DOI: https://doi.org/10.1525/elementa.131.s12
- **Figure S13.** Success Rate metric comparison with domain filling. DOI: https://doi.org/10.1525/elementa.131.s13
- **Figure S14.** Harsh Top Rank metric comparison with domain filling. DOI: https://doi.org/10.1525/elementa.131.s14
- **Figure S15.** Success Rate "all towers" metric comparison. DOI: https://doi.org/10.1525/elementa.131.s15
- **Figure S16.** Harsh Top Rank "all towers" metric comparison. DOI: https://doi.org/10.1525/elementa.131.s16
- **Figure S17.** Success Rate "all towers" metric comparison with domain filling. DOI: https://doi.org/10.1525/elementa.131.s17
- **Figure S18.** Harsh Top Rank "all towers" metric comparison with domain filling. DOI: https://doi.org/10.1525/elementa.131.s18

Acknowledgements

The authors would like to acknowledge the work performed by Drs. Natasha Miles and Scott Richardson in maintaining the operations at the Indianapolis Flux Towers. They further would like to acknowledge Drs.

Anna Karion and Colm Sweeney for their assistance in the acquisition and maintenance of the NOAA flasks, with additional gratitude to Dr. Colm Sweeney for his time and feedback in the editorial process of this manuscript. The authors are also grateful for the feedback of the other scientists involved in the greater INFLUX project during the course of this research.

Funding information

This work has been funded by the National Institute for Standards and Technology (project 70NANB10H245) and the National Oceanic and Atmospheric Administration (grant NA13OAR4310076).

Competing interests

The authors have no competing interests to declare.

Author contributions

- Contributed to conception and design: Brian Nathan, Thomas Lauvaux, and Jocelyn Turnbull
- Contributed to acquisition of data: Jocelyn Turnbull and Kevin Gurney
- Contributed to analysis and interpretation of data: All authors
- Drafted and/or revised the article: All authors
- Approved the submitted version for publication: All authors

References

Aalto, T and **Lallo, M** 2009 Atmospheric hydrogen variations and traffic emissions at an urban site in Finland. *Atmospheric Chemistry and Physics* **9**: 7387–7396. ISSN 1680-7324. http://www.atmos-chem-phys.net/9/7387/. DOI: https://doi.org/10.5194/acp-9-7387-2009

Ackerman, KV and **Sundquist, ET** 2008 Comparison of two U.S. power-plant carbon dioxide emissions data sets. *Environmental Science and Technology* **42**(15): 5688–5693. ISSN 0013936X. DOI: https://doi.org/10.1021/es800221q

AIRPARIF 2013 Bilan Des Émissions de Pollutants Atmosphériques et de Gaz à Effet de Serre à Paris Pour L'année 2010 et Historique 2000/2005. http://www.airparif.asso.fr/_pdf/publications/Emissions_2010_CG75.pdf.

Baker, AK, Beyersdorf, AJ, Doezema, LA, Katzenstein, A, Meinardi, S, et al. 2008 Measurements of nonmethane hydrocarbons in 28 United States cities. *Atmospheric Environment* **42**(1): 170–182. ISSN 13522310. DOI: https://doi.org/10.1016/j.atmosenv.2007.09.007

Bakwin, PS, Tans, PP, Hurst, DF and **Zhao, C** 1998 Measurements of carbon dioxide on very tall towers: results of the NOAA/CMDL program. *Tellus* **50**: 401–415. http://www.blackwell-synergy.com/links/doi/10.1034/j.1600-0889.1998.t01-4-00001.x. DOI: https://doi.org/10.1034/j.1600-0889.1998.t01-4-00001.x

Barletta, B, Carreras-Sospedra, M, Cohan, A, Nissenson, P, Dabdub, D, et al. 2013 Emission estimates of HCFCs and HFCs in California from the 2010 CalNex study. *Journal of Geophysical Research Atmospheres* **118**(4): 2019–2030. ISSN 21698996. DOI: https://doi.org/10.1002/jgrd.50209

Barnes, DH 2003 Hydrogen in the atmosphere: Observations above a forest canopy in a polluted environment. *Journal of Geophysical Research* **108**(D6): 1–10. ISSN 0148-0227. DOI: https://doi.org/10.1029/2001JD001199

Cambaliza, MOL, Shepson, PB, Bogner, J, Caulton, DR, Stirm, B, et al. 2015 Quantification and source apportionment of the methane emission flux from the city of Indianapolis. *Elem Sci Anth* **3**: 000037. ISSN 2325-1026. http://elementascience.org:80/article/info:doi/10.12952/journal.elementa.000037. DOI: https://doi.org/10.12952/journal.elementa.000037

Ciais, P, Sabine, C, Govindasamy, B, Bopp, L, Brovkin, V, et al. 2013 Carbon and Other Biogeochemical Cycles. In: Stocker, T, Qin, D, Plattner, GK, Tignor, M, Allen, S, et al. (eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge, UK and New York, NY, USA: Cambridge University Press: Chap. 6.

Clark-Thorne, ST and **Yapp, CJ** 2003 Stable carbon isotope constraints on mixing and mass balance of CO_2 in an urban atmosphere: Dallas metropolitan area, Texas, USA. *Applied Geochemistry* **18**(1): 75–95. ISSN 08832927. DOI: https://doi.org/10.1016/S0883-2927(02)00054-9

Colvile, RN, Hutchinson, EJ, Mindell, JS and **Warren, RF** 2001 The transport sector as a source of air pollution. *Atmospheric Environment* **35**(9): 1537–1565. ISSN 13522310. DOI: https://doi.org/10.1016/S1352-2310(00)00551-3

Crosson, ER 2008 A cavity ring-down analyzer for measuring atmospheric levels of methane, carbon dioxide, and water vapor. *Applied Physics B: Lasers and Optics* **92**(3 Special issue): 403–408. ISSN 09462171. DOI: https://doi.org/10.1007/s00340-008-3135-y

Davidson, EA and **Kanter, D** 2014 Inventories and scenarios of nitrous oxide emissions. *Environmental Research Letters* **9**(10): 105012. ISSN 1748-9326. http://iopscience.iop.Org/1748-9326/9/10/105012/article/. DOI: https://doi.org/10.1088/1748-9326/9/10/105012

Djuricin, S, Pataki, DE and **Xu, X** 2010 A comparison of tracer methods for quantifying CO_2 sources in an urban region. *Journal of Geophysical Research: Atmospheres* **115**(11): 1–13. ISSN 01480227. DOI: https://doi.org/10.1029/2009JD012236

Font, A, Grimmond, CSB, Kotthaus, S, Morguí, J, Stockdale, C, et al. 2015 Daytime CO_2 urban surface fluxes from airborne measurements, eddy-covariance observations and emissions inventory in Greater London. *Environmental Pollution* **196**: 98–106. ISSN 0269-7491. DOI: https://doi.org/10.1016/j.envpol.2014.10.001

Fortin, TJ, Howard, BJ, Parrish, DD, Goldan, PD, Kuster, WC, et al. 2005 Temporal changes in U.S. benzene emissions inferred from atmospheric measurements. *Environmental Science and Technology* **39**(6): 1403–1408. ISSN 0013936X. DOI: https://doi.org/10.1021/es049316n

Fuhlbrügge, S, Quack, B, Atlas, E, Fiehn, A, Hepach, H, et al. 2016. Meteorological constraints on oceanic halocarbons above the Peruvian upwelling. *Atmospheric Chemistry and Physics* **16**: 12205–12217. DOI: https://doi.org/10.5194/acp-16-12205-2016

Fujita, EM, Watson, JG, Chow, JC and **Magliano, KL** 1995 Receptor Model and Emissions Inventory Source Apportionments of Nonmethane Organic Gases in California's San Joaquin Valley and San Francisco Bay Area. *Atmospheric Environment* **29**(21): 3019–3035. ISSN 13522310.

Geller, LS, Elkins, JW, Clarke, AD, Hurst, DF, Butler, JH, et al. 1997 Tropospheric SF6: Observed latitudinal distribution and trends, derived emissions and interhemispheric exchange time. *Geophysical Research Letters* **24**(6): 675–678. DOI: https://doi.org/10.1029/97GL00523

Gurney, KR, Huang, J and **Coltin, K** 2016 Bias present in US federal agency power plant CO_2 emissions data and implications for the US clean power plan. *Environmental Research Letters* **11**(6): 064005. ISSN 1748-9326. http://stacks.iop.org/1748-9326/11/i=6/a=064005?key=crossref.dc145a36bd1d325c11a4e4fe3b10aa00. DOI: https://doi.org/10.1088/1748-9326/11/6/064005

Gurney, KR, Razlivanov, I, Song, Y, Zhou, Y, Benes, B, et al. 2012 Quantification of Fossil Fuel CO_2 Emissions on the Building/Street Scale for a Large U.S. City. *Environmental Science & Technology* **46**: 12194–12202. DOI: https://doi.org/10.1021/es3011282

Heimburger, AMF, Harvey, RM, Shepson, PB, Stirm, BH, Gore, C, Turnbull, J, et al. 2017 Assessing the optimized precision of the aircraft mass balance method for measurement of urban greenhouse gas emission rates through averaging. *Elem Sci Anth* **5**: 26. DOI: http://doi.org/10.1525/elementa.134

Hutyra, LR, Duren, R, Gurney, KR, Grimm, N, Kort, EA, et al. 2014 Urbanization and the carbon cycle: Current capabilities and research outlook from the natural sciences perspective. *Earth's Future* **2**: 473–495. DOI: https://doi.org/10.1002/2014EF000255

Kim, KH, Shon, ZH, Nguyen, HT and **Jeon, EC** 2011 A review of major chlorofluorocarbons and their halocarbon alternatives in the air. *Atmospheric Environment* **45**(7): 1369–1382. ISSN 13522310. DOI: https://doi.org/10.1016/j.atmosenv.2010.12.029

Kohonen, T 1990 The self-organizing map. *Proceedings of the IEEE* **78**(9): 1464–1480. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=58325. DOI: https://doi.org/10.1109/5.58325

Lamb, BK, Cambaliza, MOL, Davis, KJ, Edburg, SL, Ferrara, TW, et al. 2016 Direct and Indirect Measurements and Modeling of Methane Emissions in Indianapolis, Indiana. *Environmental Science & Technology*. ISSN 0013-936X. http://pubs.acs.org/doi/abs/10.1021/acs.est.6b01198. DOI: https://doi.org/10.1021/acs.est.6b01198

Lauvaux, T, Miles, NL, Deng, A, Richardson, SJ, Cambaliza, MO, et al. 2016 High resolution atmospheric inversion of urban CO_2 emissions during the dormant season of the Indianapolis Flux Experiment (INFLUX). *Journal of Geophysical Research: Atmospheres,* 1–24. ISSN 2169-8996. http://onlinelibrary.wiley.com/doi/10.1002/2015JD024473/abstract. DOI: https://doi.org/10.1002/2015JD024473

Levin, I and **Karstens, U** 2007 Inferring high-resolution fossil fuel CO_2 records at continental sites from combined $^{14}CO_2$ and CO observations. *Tellus, Series B: Chemical and Physical Meteorology* **59**(2): 245–250. ISSN 02806509. DOI: https://doi.org/10.1111/j.1600-0889.2006.00244.x

Levin, I, Kromer, B, Schmidt, M and **Sartorius, H** 2003 A novel approach for independent budgeting of fossil fuel CO_2 over Europe by $^{14}CO_2$ observations. *Geophysical Research Letters* **30**(23): 2194. ISSN 0094-8276. http://doi.wiley.com/10.1029/2003GL018477, http://archiv.ub.uni-heidelberg.de/volltextserver/volltexte/2006/6727/pdf/LevinGRL2003.pdf. DOI: https://doi.org/10.1029/2003GL018477

Maiss, M and **Brenninkmeijer, CAM** 1998 Atmospheric SF6: Trends, Sources, and Prospects. *Environmental Science & Technology* **32**(20): 3077–3086. ISSN 0013-936X. DOI: https://doi.org/10.1021/es9802807

Mckain, K, Down, A, Raciti, SM, Budney, J, Hutyra, LR, et al. 2015 Methane emissions from natural gas infrastructure and use in the urban region of Boston, Massachusetts. *Proceedings of the National Academy of Sciences* **112**(7): 1941–1946. DOI: https://doi.org/10.1073/pnas.1416261112

Meijer, H, Smid, H, Perez, E and **Keizer, M** 1996 Isotopic characterisation of anthropogenic CO_2 emissions using isotopic and radiocarbon analysis. *Physics and Chemistry of the Earth* **21**(5–6): 483–487. ISSN 00791946. http://www.sciencedirect.com/science/article/pii/S0079194697811469, http://linkinghub.elsevier.com/retrieve/pii/S0079194697811469. DOI: https://doi.org/10.1016/S0079-1946(97)81146-9

Miles, NL, Richardson, SJ, Lauvaux, T, Davis, KJ, Balashov, NV, Deng, A, et al. 2017 Quantification of urban atmospheric boundary layer greenhouse gas dry mole fraction enhancements in the dormant season: Results from the Indianapolis Flux Experiment (INFLUX). *Elem Sci Anth* **5**: 27. DOI: http://doi.org/10.1525/elementa.127

Miller, JB, Lehman, SJ, Montzka, SA, Sweeney, C, Miller, BR, et al. 2012 Linking emissions of fossil fuel CO_2 and other anthropogenic trace gases using atmospheric $^{14}CO_2$. *Journal of Geophysical Research: Atmospheres* **117**(8). ISSN 01480227. DOI: https://doi.org/10.1029/2011JD017048

Montzka, SA, Myers, RC, Butler, JH, Elkins, JW and **Cummings, S** 1993 Global Tropospheric Distribution and Calibration Scale of HCFC-22. *Geophysical*

Research Letters **20**(8): 703–706. ISSN 00948276. DOI: https://doi.org/10.1073/pnas.0703993104

Newman, S, Xu, X, Gurney, KR, Hsu, YK, Li, KF, et al. 2016 Toward consistency between trends in bottom-up CO_2 emissions and top-down atmospheric measurements in the Los Angeles megacity. *Atmospheric Chemistry and Physics* **16**(6): 3843–3863. ISSN 16807324. DOI: https://doi.org/10.5194/acp-16-3843-2016

Novelli, PC, Lang, PM, Masarie, KA, Hurst, DF, Myers, R, et al. 1999 Molecular hydrogen in the troposphere: Global distribution and budget. *Journal of Geophysical Research* **104**(D23): 30427–30444. DOI: https://doi.org/10.1029/1999JD900788

O'Doherty, S, Cunnold, DM, Miller, BR, Mu, J, Mcculloch, A, et al. 2009 Global and regional emissions of HFC-125 (CHF2CF3) from in situ and air archive atmospheric observations at AGAGE and SOGE observatories. *Journal of Geophysical Research* **114**(D2): 3304. DOI: https://doi.org/10.1029/2009JD012184

Paatero, P and **Tapper, U** 1994 Positive Matrix Factorization – A Nonnegative Factor Model with Optimal Utilization of Error Estimates of Data Values. *Environmetrics* **5**(2): 111–126. ISSN 11804009. http://doi.wiley.com/10.1002/env.3170050203\n<GotoISI>://WOS:A1994NZ66000002. DOI: https://doi.org/10.1002/env.3170050203

Papasavva, S, Luecken, DJ, Waterland, RL, Taddonio, KN and **Andersen, SO** 2009 Estimated 2017 refrigerant emissions of 2,3,3,3-tetrafluoropropene (HFC-1234yf) in the United States resulting from automobile air conditioning. *Environmental Science and Technology* **43**(24): 9252–9259. ISSN 0013936X. DOI: https://doi.org/10.1021/es902124u

Purohit, P and **Hoglund-Isaksson, L** 2016 Global emissions of fluorinated greenhouse gases 2005–2050 with abatement potentials and costs. *Atmospheric Chemistry and Physics Discussions* (August). ISSN 1680-7375. DOI: https://doi.org/10.5194/acp-2016-727

Rella, CW, Chen, H, Andrews, AE, Filges, A, Gerbig, C, et al. 2013 High accuracy measurements of dry mole fractions of carbon dioxide and methane in humid air. *Atmospheric Measurement Techniques* **6**(3): 837–860. ISSN 18671381. DOI: https://doi.org/10.5194/amt-6-837-2013

Richardson, SJ, Miles, NL, Davis, KJ, Lauvaux, T, Martins, DK, Turnbull, JC, et al. 2017 Tower measurement network of in-situ CO_2, CH_4, and CO in support of the Indianapolis FLUX (INFLUX) Experiment. *Elem Sci Anth* **5**: 59. DOI: http://doi.org/10.1525/elementa.140

Seibert, P and **Frank, A** 2004 Source-receptor matrix calculation with a Lagrangian particle dispersion model in backward mode. *Atmospheric Chemistry and Physics* **4**(1): 51–63. ISSN 1680-7324. DOI: https://doi.org/10.5194/acp-4-51-2004

Skamarock, WC and **Klemp, JB** 2008 A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational Physics* **227**(7): 3465–3485. ISSN 00219991. DOI: https://doi.org/10.1016/j.jcp.2007.01.037

Staufer, J, Broquet, G, Bréon, FM, Puygrenier, V, Chevallier, F, et al. 2016 The first 1-year-long estimate of the Paris region fossil fuel CO_2 emissions based on atmospheric inversion. *Atmospheric Chemistry and Physics* **16**(22): 14703–14726. ISSN 16807324. DOI: https://doi.org/10.5194/acp-16-14703-2016

Stull, RB 1988 *Boundary Layer Meteorology*.

Sweeney, C, Karion, A, Wolter, S, Newberger, T, Guenther, D, et al. 2015 Seasonal climatology of CO_2 across North America from aircraft measurements in the NOAA/ESRL Global Greenhouse Gas Reference Network. *Journal of Geophysical Research: Atmospheres*, 5155–5190. DOI: https://doi.org/10.1002/2014JD022591

Tamayo, P, Slonim, D and **Mesirov, J** 1999 Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the …* **96**: 2907–2912. http://www.pnas.Org/content/96/6/2907.short. DOI: https://doi.org/10.1073/pnas.96.6.2907

Turnbull, J, Guenther, D, Karion, A, Sweeney, C, Anderson, E, et al. 2012 An integrated flask sample collection system for greenhouse gas measurements. *Atmospheric Measurement Techniques* **5**(9): 2321–2327. ISSN 18671381. DOI: https://doi.org/10.5194/amt-5-2321-2012

Turnbull, J, Rayner, P, Miller, J, Naegler, T, Ciais, P, et al. 2009 On the use of $^{14}CO_2$ as a tracer for fossil fuel CO_2: Quantifying uncertainties using an atmospheric transport model. *Journal of Geophysical Research: Atmospheres* **114**(22): 113. ISSN 01480227. DOI: https://doi.org/10.1029/2009JD012308

Turnbull, JC, Karion, A, Fischer, ML, Faloona, I, Guilderson, T, et al. 2011 Assessment of fossil fuel carbon dioxide and other anthropogenic trace gas emissions from airborne measurements over Sacramento, California in spring 2009. *Atmospheric Chemistry and Physics* **11**(2): 705–721. ISSN 16807316. DOI: https://doi.org/10.5194/acp-11-705-2011

Turnbull, JC, Lehman, SJ, Miller, JB, Sparks, RJ, Southon, JR, et al. 2007 A new high precision $^{14}CO_2$ time series for North American continental air. *Journal of Geophysical Research-Atmospheres* **112**(D1): 1310. ISSN 01480227. http://www.agu.org/pubs/crossref/2007/2006JD008184.shtml. DOI: https://doi.org/10.1029/2006JD008184

Turnbull, JC, Miller, JB, Lehman, SJ, Tans, PP, Sparks, RJ, et al. 2006 Comparison of $^{14}CO_2$, CO, and SF6 as tracers for recently added fossil fuel CO_2 in the atmosphere and implications for biological CO_2 exchange. *Geophysical Research Letters* **33**(1): 2–6. ISSN 00948276. DOI: https://doi.org/10.1029/2005GL024213

Turnbull, JC, Sweeney, C, Karion, A, Newberger, T, Lehman, SJ, et al. 2015a 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: https://doi.org/10.1002/2014JD022555

Turnbull, JC, Zondervan, A, Kaiser, J, Norris, M, Dahl, J, et al. 2015b High-Precision Atmospheric $^{14}CO_2$ Measurement at the Rafter Radio-Carbon Laboratory. *Radiocarbon* **57**(3): 377–388. ISSN 00338222. http://arxiv.org/abs/1306.2418.

Uliasz, M 1994 Lagrangian particle modeling in mesoscale applications.

United Nations Human Settlements Programme 2011 Global Report on Human Settlements 2011: Cities and Climate Change. http://books.google.es/books?id=mDTsgiMgnHwC. DOI: https://doi.org/10.1787/9789264091375-en

United States Environmental Protection Agency 2006 Documentation for the final 2002 point source national emissions inventory. Research Triangle Park, North Carolina: Emission Inventory and Analysis Group, Air Quality and Analysis Division, EPA.

United States Environmental Protection Agency 2008 EPA Positive Matrix Factorization (PMF) 3.0 Fundamentals & User Guide. http://www.epa.gov/heasd/products/pmf/pmf.htm.

United States Environmental Protection Agency 2012 Oil and Natural Gas Sector: New Source Performance Standards and National Emission Standards for Hazardous Air Pollutants Reviews. http://www.epa.gov/airquality/oilandgas/pdfs/20120417finalrule.pdf.

Van Der Laan, S, Karstens, U, Neubert, REM, Van Der Laan-Luijkx, IT and **Meijer, HAJ** 2010 Observation-based estimates of fossil fuel-derived CO_2 emissions in the Netherlands using $\Delta^{14}C$, CO and 222Radon. *Tellus, Series B: Chemical and Physical Meteorology* **62**(5): 389–402. ISSN 02806509. DOI: https://doi.org/10.1111/j.1600-0889.2010.00493.x

Velders, GJM, Fahey, DW, Daniel, JS, Mcfarland, M and **Andersen, SO** 2009 The large contribution of projected HFC emissions to future climate forcing. *Proceedings of the National Academy of Sciences* **106**(27): 10949–10954. DOI: https://doi.org/10.1073/pnas.0902817106

Vogel, FR, Hammer, S, Steinhof, A, Kromer, B and **Levin, I** 2010 Implication of weekly and diurnal ^{14}C calibration on hourly estimates of CO-based fossil fuel CO_2 at a moderately polluted site in southwestern Germany. *Tellus, Series B: Chemical and Physical Meteorology* **62**(5): 512–520. ISSN 02806509. DOI: https://doi.org/10.1111/j.1600-0889.2010.00477.x

Warneke, C, McKeen, SA, de Gouw, JA, Goldan, PD, Kuster, WC, et al. 2007 Determination of urban volatile organic compound emission ratios and comparison with an emissions database. *Journal of Geophysical Research Atmospheres* **112**(10). ISSN 01480227. DOI: https://doi.org/10.1029/2006JD007930

Watson, JG, Chow, JC and **Fujita, EM** 2001 Review of volatile organic compound source apportionment by chemical mass balance. *Atmospheric Environment* **35**(9): 1567–1584. ISSN 13522310. DOI: https://doi.org/10.1016/S1352-2310(00)00461-1

Whitby, R and **Altwicker, E** 1978 Acetylene in the atmosphere:Sources,representativeambientconcentrations and ratios to other hydrocarbons. *Atmospheric Environment (1967)* **12**: 1289–1296. ISSN 00046981. DOI: https://doi.org/10.1016/0004-6981(78)90067-7

Widory, D and **Javoy, M** 2003 The carbon isotope composition of atmospheric CO_2 in Paris. *Earth and Planetary Science Letters* **215**(1–2): 289–298. ISSN 0012821X. DOI: https://doi.org/10.1016/S0012-821X(03)00397-2

Yi, C, Davis, KJ, Berger, BW and **Bakwin, PS** 2001 Long-Term Observations of the Dynamics of the Continental Planetary Boundary Layer. *Journal of the Atmospheric Sciences* **58**(10): 1288–1299. ISSN 0022-4928. DOI: https://doi.org/10.1175/1520-0469(2001)058<1288:LTOOTD>2.0.CO;2

How to cite this article: Nathan, B, Lauvaux, T, Turnbull, J and Gurney, K 2018 Investigations into the use of multi-species measurements for source apportionment of the Indianapolis fossil fuel CO_2 signal. *Elem Sci Anth,* 6: 21. DOI: https://doi.org/10.1525/elementa.131

Domain Editor-in-Chief: Detlev Helmig, University of Colorado Boulder, US

Associate Editor: Lori Bruhwiler, National Oceanic & Atmospheric Administration, US

Knowledge Domain: Atmospheric Science domain

Part of an *Elementa* Special Feature: Quantification of Urban Greenhouse Gas Emissions: The Indianapolis Flux Experiment

Submitted: 18 October 2016 **Accepted:** 11 December 2017 **Published:** 01 March 2018

ELEMENTA Science of the Anthropocene

Elem Sci Anth is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS