Characterizing Uncertainty in Anthropogenic Point Source Emissions of CO_2

by

Dawn Leigh Woodard

Honors Thesis

Submitted to the Department of Mathematical Sciences

and the Honors College

in partial fulfillment of the requirements for the degree of

Bachelor of Science

May 2014

Approved by:

Dr. Eric S. Marland, Mathematics, Thesis Director

Dr. Gregg H. Marland, Geology, Second Reader

Dr. Vicky Klima, Mathematics, Department Honors Director

Leslie Sargent Jones, Ph.D., Director, The Honors College

Abstract

Large point sources account for as much as 60% of the carbon dioxide emissions for some countries. Further, in the US one third of all CO2 emissions come from only 311 point sources (power plants, industrial sites, etc.). Because CO_2 emissions are seldom measured directly but are generally estimated from related, proxy, and re-purposed data; we also need to understand the uncertainty of these estimates. Simply stated, given a geographic and temporal space on the Earth, what are the CO_2 emissions from that space and what is the uncertainty in this estimate? While the US data on large point sources is largely assumed to have no spatial uncertainty, the actual locations of these sources differ by 0.84km on average from their reported locations. Analysis also reveals quantifiable trends in the uncertainty based on simple characteristics such as proximity to water sources, reported location within political boundaries, local and population density. This paper presents a metric to quantify spatial uncertainty in point sources based on the results of this analysis, and explains why point source data cannot be described with traditional methods. To incorporate resolution and placement within a grid cell, a Monte Carlo simulation is used to calculate expected values for emissions for each point source. The spatial uncertainty is then derived from the simulation output to give a picture of the potential spatial spread of the emissions. This is output as gridded data at the desired resolution and can then be incorporated into other data products reporting estimated emissions from point sources.

Keywords: uncertainty, carbon dioxide, point source, Monte Carlo

1. Introduction

For reasons ranging from the understanding of important human and biogeochemical processes to the monitoring, reporting, and verification of international agreements, there is great interest in describing both the magnitude and distribution of anthropogenic CO_2 emissions on the Earth. Because CO_2 emissions are seldom measured directly but are generally estimated from related, proxy, and re-purposed data; it is also important to understand the uncertainty of these estimates. This concern is additionally driven by the desire to use ground-level data of emissions to calibrate satellites, enabling them to remotely determine locations of sources and the magnitude of gas being emitted. Before data can be used for such an application, it is necessary to have an understanding of its accuracy. The question that needs to be addressed is therefore: given a geographic and temporal space on the Earth, what are the CO_2 emissions from that space and what is the uncertainty in this estimate? Uncertainty in this sense refers to both the confidence in the estimate of the total emissions being produced, but also to the confidence that those emissions are coming from the place they are reported to be. The latter is a quantity that has remained largely unaddressed in the literature and is of high concern due to substantial reporting errors in the available data. The analysis here focuses on this uncertainty in point sources using data of annual sums of carbon dioxide emissions from electric power generation facilities in the United States. Anthropogenic point sources are human-caused, localized, stationary sources of emissions such as from coal, biomass, natural gas, oil and other types of power plants. Due to their localized nature and extremely high levels of emissions compared to other sources, errors in their location make a larger impact on overall emissions totals than does spatial error in reporting of other types of data such as traffic or agricultural emissions estimates. The following sections examine characteristics that affect the overall uncertainty in the location of point sources, and propose a methodology for quantifying that uncertainty.

In section 2 below I give background of the problem and establish the importance of large point sources, describe the sources of data on large point source emissions in the US, and begin to examine the issue of dealing with locational uncertainty when available data are used to describe the actual locations of CO_2 discharges. In section 3 I describe our approach to dealing with locational uncertainty of large point sources and in section 4 I develop a metric to characterize the uncertainty of emissions from a given spatial unit when locational uncertainty is an issue. Section 5 discusses the relationship between spatial uncertainty and spatial resolution, section 6 provides some sample output, and section 7 discusses my analysis.

2. Background and Motivation

2.1. **Climate Change.** The issue of climate change is of increasing global concern, as the planet continues to experience significant impacts from its effects. On average global surface temperatures have increased by 0.85 °C over the last century, and could increase by as much as 4.8 °C by 2100. The mean annual sea level rise has been nearly 2mm since 1901 and is predicted to rise by anywhere from .26m to .98m by 2100 [cite IPCC]. Arctic sea ice cover is expected to decline substantially with these increases in temperature, accompanied by decreases in global glacier volume and spring snow cover in the Northern Hemisphere. Additionally predicted are changes in rainfall patterns and an increase in hot temperature extremes while cold temperature extremes decline [12].

The rising global temperatures that are causing such dramatic changes to the planet are a direct result of increased atmospheric concentrations of what are known as greenhouse gases. The Earth's system maintains a radiative balance between incoming energy in the form of short wave radiation from the sun, and outgoing energy in the form of long wave, or thermal infrared radiation, emitted by the Earth. Greenhouse gases are gaseous constituents of the atmosphere that absorb and re-emit that thermal radiation as it travels away from the Earth effectively slowing its escape from the troposphere. This decreases the net radiative flux, or amount of energy passing through, at the top of the troposphere, shifting the balance between incoming and outgoing energy and resulting in an overall increase in temperature on the planet in order to maintain that balance [5]. The impact a particular gas has on this radiative flux is referred to as its radiative forcing. Thus the higher magnitude of the radiative forcing a particular gas has, the more impact each unit increase in concentration has on the net energy balance on the planet [11]. Taking into account radiative forcing and current concentrations, the four most important greenhouse gases in terms of their contributions to global warming are carbon dioxide, methane, dichlorofluoromethane, and nitrous oxide.

As a direct result of human activity concentrations of carbon dioxide, methane, and nitrous oxide among others in the atmosphere have all increased substantially since the late 1700s with the onset of the Industrial Revolution in the West. The current concentrations of these gases exceed anything recorded in ice cores over the last 800,000 years due to the recent increase in anthropogenic emissions. These gases can be emitted by point and non-point sources. Point sources of pollution are localized and stationary, such as power plants or industrial sites, whereas non-point sources are more dispersed, such as highway traffic [5, 7]. Non-point sources also include land use and land cover change, which has contributed an estimated 33% of the total carbon emissions over the last 150 years. This includes

emissions from deforestation, agricultural management practices that affect the storage of carbon in soils, fire management, land degradation, and others [6].

Efforts through the United Nations, individual governments, and even many corporations have been ongoing in working to mitigate global climate change. The United Nations Framework Convention on Climate Change (UNFCCC) has been working to design a global solution, though it has been hindered by economic concerns of many countries and the enormity of determining a solution across so many countries impacting so many aspects of economics and politics. As of 2010 the UNFCCC reached an agreement to attempt to limit warming of the planet to 2 °C by the end of the century, requiring emissions reductions on the part of both developed and developing nations. Some more localized solutions have also been enacted. In 2005 the European Union enacted a cap and trade system encompassing 31 countries with a goal of reducing emissions of CO_2 by 20% by 2020 and 80-95% by 2050. The program regulates about half of the emissions in Europe, including some 11,000 power plants and manufacturing facilities as well as flights to and from participating nations [2]. Cap and trade systems such as this work by setting an overall cap on emissions levels, which is lowered over time, and issuing or selling emissions allowances to companies within the limits of the cap. These can then be traded and sold between companies as necessary. Also in 2005 24 major corporations met at the G8 Climate Change Roundtable to commit to working to prevent climate change and urge governments to help their efforts [13]. As of 2013 the United States Environmental Protection Agency has proposed new CO_2 emissions standards for future power plants, and a separate, less strict set of standards for power plants already in operation in accordance with President Obama's Climate Action Plan [4]. The Regional Greenhouse Gas Initiative in the northeastern US also uses a cap and trade system to reduce emissions overall within the nine participating states, lowering the cap by 2.5% each year between 2015 and 2020 [3].

Policies such as the previous examples drive concerns over what has come to be known as MRV, or measurement, reporting, and verification. With legal and financial interests weighing on the accurate knowledge of emissions estimates, it is important to not only measure and report this data but also be able to verify such reports. Remote sensing techniques are one such method of potential verification, however they require calibration encompassing two main components, namely the magnitude and location of the emissions. We characterize emissions from a given location as the sum of emissions from large point sources such as fossil-fuel-fired power plants and from areal sources such as automobiles and home heating units. Emissions from large point sources can be orders of magnitude greater than for nearby areas and so are of particular concern to characterize. For areal sources we can estimate CO_2 emissions from a given space based on records of things like energy consumption and land-use change, which reflects to the uptake of carbon dioxide by vegetation. For large point sources we are similarly concerned

with data such as fuel consumption in order to estimate the magnitude of the emissions, though in some cases this estimate can be obtained through direct measurement of emissions at the point of release, and we have the additional element of detailing the location in space. In the available data, this question of placement of point sources, given in geographic coordinates, is typically self-reported by the facilities themselves, introducing significant inconsistences depending on whether these facilities provided coordinates corresponding to their street address, in-town office, power generation site, or do not provide any at all.

2.2. **Uncertainty.** Data inconsistencies and other errors affect the accuracy of a reported quantity. As a consequence, reported quantities are provided with range of values (usually expressed as $v \pm x$ with y% confidence), that suggests the probability that the true value lies in the interval around the reported value. This range of probable values where the true value occurs reflects the *uncertainty* of the reported value. The level of uncertainty in any particular value might range by many orders of magnitude. The origins of this uncertainty in the reported value might originate from many sources: a lack of information, disagreement over given data, measurement error, inherent variability, approximations, subjective judgments, or numerous other factors. Understanding and quantifying sources of uncertainty are essential to give a proper reflection of the range in which the true value could potentially fall [9].

In considering emissions estimates of carbon dioxide, there are four important sources of uncertainty: a.) the magnitude of emissions from areal sources, b.) the magnitude of emissions from large point sources, c.) the magnitude uncertainty associated with the emissions estimates, and d.) the spatial uncertainty in the large point sources. The last element is unique because of its binary character and it is the focus of this paper. While important, the spatial uncertainty from areal sources has smoother characteristics and can be ably handled with standard methods and is therefore not specifically addressed here. There are also uncertainties associated with the calculation or measurement of emissions from the stacks of power generation sites or other facilities. Calculations cannot take into account every factor for each individual operation, resulting in approximate values, and the devices to measure emissions are limited by accuracy and precision, leaving uncertainty in the results.

For large point sources we have the binary condition that the source either is or is not in the space under consideration, and small locational errors can result in very large differences in the estimated emissions for two spaces - the space where the facility is reported and the space where it actually exists. The importance of discrepancies in spatial locations increases at finer spatial resolutions. Locational errors with this large point source data arise due mostly to its self-reported nature and because the data are often being re-purposed from other applications. There are instances of lack of information where a

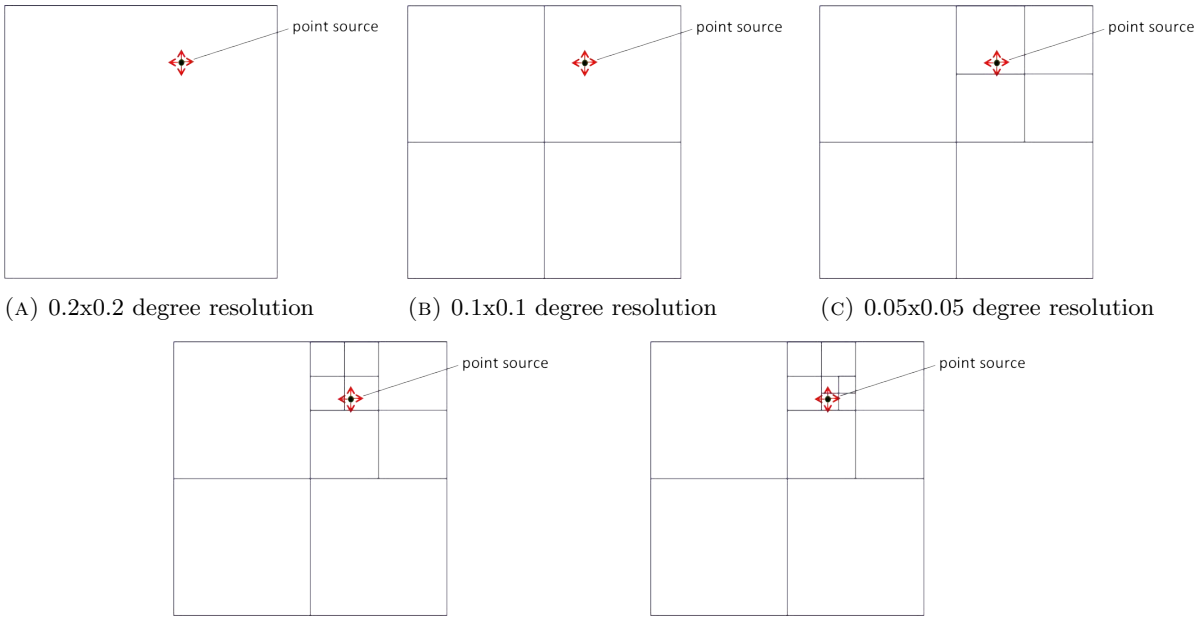
power plant may not report any location at all and the data compilers will place the point source at a default location such as the center of the political unit (county or city) in which it is known to be. This case results in large uncertainty for CO_2 emissions as the point source could theoretically be anywhere within that political unit. In other cases there is simply a lack of precision in the emissions data as power plants and other facilities do not always report the coordinates of point of release of the emissions, but instead an in-town office or street address.

These data issues are only one consideration when looking at uncertainty in emissions data. When reporting point sources as part of gridded data outputs, therefore allocating emissions from a single point source to an entire grid cell, resolution matters. Depending on the dataset errors, finer resolution may not always be better. For a dataset such as the one used in this analysis, which has a mean spatial uncertainty of 0.84km for large point sources (see section 3.1), then at resolutions of smaller than double this distance a given point source reported in the center of one grid cell could actually be in any of the surrounding grid cells with a very high probability (Figure 1). This also brings up the importance of the location of a point source within a grid cell. In the case of resolutions of 11kmx11km, or approximately 0.1x0.1 degrees, a point source reported in the middle of a grid cell would still be expected to actually be in that grid cell with a high degree of confidence. A point source reported closer to the edge of the cell would have a higher chance of its actual location being in a neighboring grid cell.

If the uncertainty is taken to be radially symmetric and normally distributed then the point source would have a chance of being substantially farther off than the average uncertainty measured in the dataset, meaning that even larger resolutions could not guarantee zero spatial uncertainty. It is important to understand this uncertainty and to find means of quantifying it in order to provide information about confidence in this data to the end users.

Methods such as those described in section 3.1 can give information about the average dataset-related uncertainty with the location of a point source, but they cannot take into account other important factors that affect point sources within a gridded reference system such as grid resolution and placement. In order to quantify spatial uncertainty that incorporates both gridding and dataset related uncertainty, we turn to Monte Carlo techniques.

2.3. **Monte Carlo Methods.** Monte Carlo techniques provide a powerful tool that can be used in numerous applications to determine approximate solutions to problems through repeated trials based on random sampling. They are particularly useful for cases where analytical methods are either too time consuming or not yet available. One of the classical applications of these techniques is in the case of Buffon's Needle. This problem describes a scenario where a needle is dropped on a wooden floor and



(A) 0.2x0.2 degree resolution (B) 0.1x0.1 degree resolution (C) 0.05x0.05 degree resolution

(D) 0.025x0.025 degree resolution (E) 0.0125x0.0125 degree resolution

FIGURE 1. A sample point source placed in grid cells of decreasing resolution. The red arrows are scaled to represent 0.84km, or the average distance away from the true location that a point source would be expected to be reported. As the resolution decreases it is shown that the actual location of the point source is more and more likely to be in a neighboring grid cell.

thus could fall completely on one of the boards or it might land so that it is touching one of the cracks between the boards. The problem then asks: how likely it is that the needle falls so that it is touching one of the cracks? An analytical solution to the question is fairly cumbersome, and it is time-consuming to attempt to find an answer by simply dropping physical needles on the floor repeatedly and keeping count when the needle crosses a crack. However, it takes hardly any time at all to allow a computer to do this for us. Modelling the floor as a series of parallel lines d centimeters apart, and the needle as a line with length l and angle θ from parallel, a sufficiently large number of needles can be generated. The needle length is held constant and for each trial a random distance from the edge of the board, y, is generated from within the range $(0, \frac{l}{2})$, and a random θ is selected within the range $(0, \frac{\pi}{2})$. Those two parameters define a needle, and it is then possible to calculate for each trial how often the needle does actually cross a line between the floorboards. This entire computation takes a matter of seconds to run hundreds of thousands of trials, whereas attempting to perform a sufficient number of trials by hand would take hours of time and far more effort [10].

A very similar application of Monte Carlo methods can be applied to point sources in order to determine, instead of how often they fall on a crack, how often they fall in a particular grid cell, such as the one they are reported to be in, and how often they fall into another neighboring cell. This

methodology, detailed in Section 3.2, enables the computation of spatial uncertainty for each point source accounting for dataset properties as well as grid resolution and grid cell placement.

2.4. **CO_2 emissions from large point sources.** Anthropogenic point source emissions comprise a significant portion of total carbon dioxide emissions worldwide (Singer et al., 2014). In the US they represent a full 40-50% of anthropogenic CO_2 emissions, with a third of these emissions coming from only 311 very large point sources (USEPA, 2013a), emphasizing the significant impact of a large point source data on the nation's total carbon dioxide output to the atmosphere. In any effort to characterize the spatial distribution of CO_2 emissions, it is therefore important to accurately report both the magnitude and location of large point sources and to understand any unavoidable uncertainty so that it can be fairly quantified.

There are three data sets in the United States that report point source emissions of carbon dioxide. Carbon Monitoring for Action (CarMA, 2013) provides a global dataset produced and financed by the Confronting Climate Change Initiative. The database of CarMA is comprised of carbon dioxide emissions for over 60,000 power plants and 20,000 power companies worldwide. CarMA relies on data reported to the Environmental Protection Agency (EPA) for all power plants within the United States and by the International Atomic Energy Agency for many power plants in the European Union, Canada, India, and South Africa. Electricity generation and CO_2 emissions for all non-reporting plants are estimated by using statistical models.

The Emissions and Generation Resource Integrated Database (eGRID) (ESEPA, 2013c) is a comprehensive inventory of environmental attributes of electrical power systems in the United States produced by the Environmental Protection Agency (EPA). eGRID integrates many different federal data sources on power plants and power companies from four different federal agencies: the EPA, the Energy Information Administration (EIA), the North American Electric Reliability Corporation (NERC), and the Federal Energy Regulatory Commission (FERC) to produce a detailed emissions and resource profile. The Environmental Protection Agency also supports a data set of self-reported emissions from all large point sources in the US under the Greenhouse Gas Reporting Program (USEPA, 2013a). The Greenhouse Gas Reporting Program data include both power plants and other large facilities and account for almost 7000 large-emitting sites. Because it includes most of the largest point sources in the US and because it provides a model that can be enlarged globally, for most of the analysis presented in this paper eGRID data (USEPA, 2013c) were used.

These datasets are intended for use in reporting carbon emissions totals at various political levels, as well as providing detailed categorical information on each point source. To this end plant locations

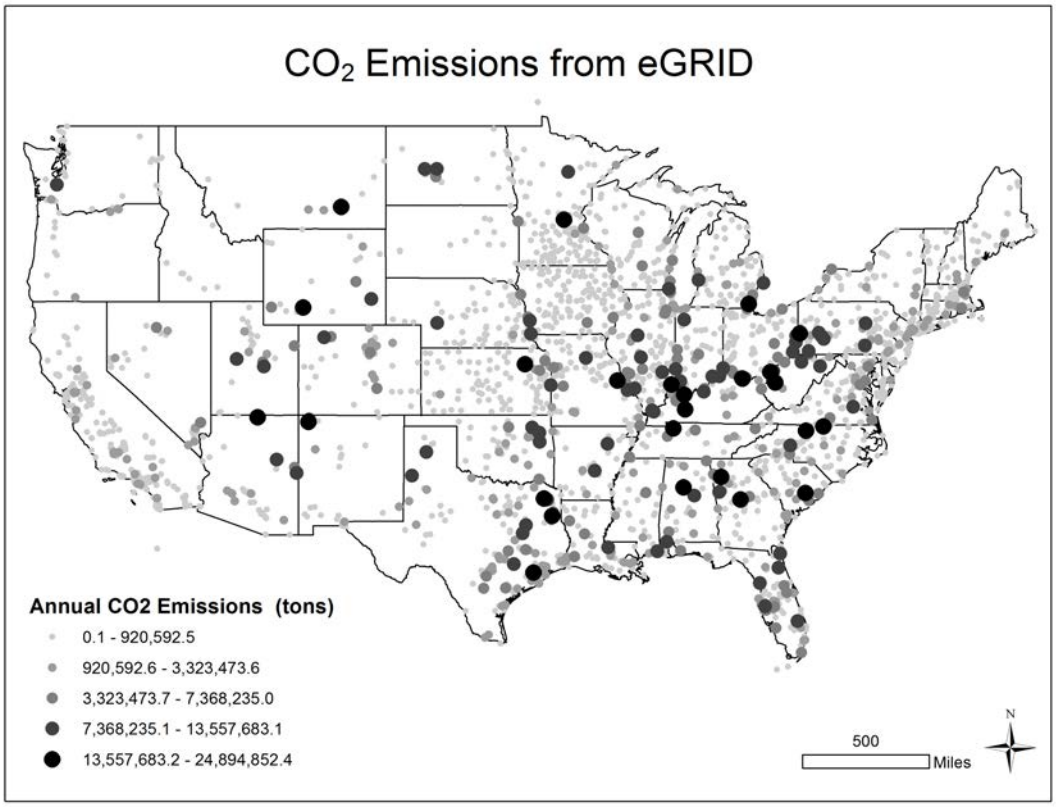
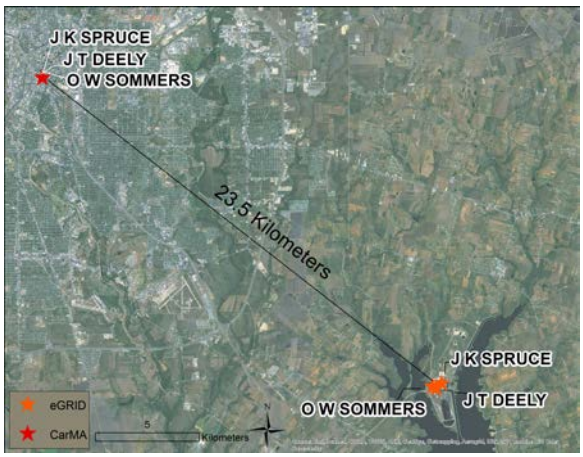


FIGURE 2. Large point sources of CO_2 emissions in the U.S. in 2009 as reported by eGRID [1].

emphasize geopolitical data and not necessarily the exact point of gaseous discharge. Spatial locations of the power plants have been self-reported by the facilities themselves and are allocated by default to the centroid of a county if street address or latitude and longitude coordinates were not given. In using this data for spatial analysis of point source discharge locations significant discrepancies between the reported latitude and longitude of the same point source in eGRID and CarMA, were discovered. Some of this may reflect data revisions that have reached one but not both data sets and some may represent problems with data entry. For some point sources eGRID and CarMA disagree by more than 20 km (Figure 3a).

Additionally, even if the datasets concur, this is not indicative of accuracy. Using satellite imagery from Google Earth it is possible to visually determine if a point source is actually found at the reported location. In one notable instance both eGRID and CarMA allocate seven different point sources to the same coordinates although there are no emissions stacks evident at that location, and it is clear that all seven are placed incorrectly in both datasets (Figure 3b).



(A) Significant disagreement between eGRID and CarMA on the location of the J. K. Spruce power plant near San Antonio, Texas.

(B) Seven points allocated to the same incorrect location in both eGRID and CarMA in Manassas, Virginia. Also shown is an additional point (unlabeled) placed incorrectly.

FIGURE 3. Examples of locational errors in data sets used to report annual sum data for point source emissions.

An analysis of the top 81 emitters in eGRID reveals that even these hugely significant point sources have considerable uncertainty in representing the location of actual emissions. The geographic coordinates of emissions are accurate in only 15 of these instances, the address is correct only 29 times, and while most are within 16 kilometers of the point of discharge, more than half are still misplaced by more than 1.5 kilometers. However, as demonstrated by the power plant shown in Figure 4, even data accurate to within a few hundred meters can end up placed in an incorrect grid cell. The figure shows that a power plant which ends up in the corner of a grid cell can have the stacks in one grid cell, while the street address and the main offices of the plant are in different grid cells. While this situation may not occur frequently (this plant is in South Africa) misallocation can occur even with very good data. The statistical frequency in which such a situation might arise can be calculated from basic geometry. In a grid cell of size 0.1 degrees by 0.1 degrees (about 11 km by 11 km at mid-latitude in the US), we would expect 36% of power plants to be within 0.01 degrees (about 1.1 km) of a cell boundary. Since more than half of the top emitters are located greater than 1.5 km from the actual discharge, this suggests that around 14 ($= 81 \cdot 0.36 \cdot 0.5$) of the 81 largest US power plants might be placed within 1 km of a boundary and have an actual location farther away than the border of the grid cell. While 14 might seem like a small number, these 14 would be among the largest emitters in the country, accounting for a large fraction of the emissions from the region where they are located. Placing a plant in the wrong grid cell amounts to placing it 11km away from its actual location (based on the center, or reference point, of the grid cell).

This concern over spatial accuracy is particularly aggravated due to the large numbers of point sources along state or other borders. Large water bodies are often used to define state or national political borders and power plants are often along rivers because of their need for cooling water. As a consequence we found in our analysis that 19% of US power plants reported in eGRID are within 10km of a state border, as opposed to 11% of states by area within 10km of state borders.

To quantify this further, keep in mind that emissions data are produced by multiple organizations for different purposes and thus are not reported at a standardized resolution. Much of the data produced ranges from 5x5 degree grids down to less than 0.1x0.1 degrees with plans to refine further. In carbon accounting and analysis multiple datasets are typically used, so it becomes imperative to have an understanding of grid resolution and how changing resolution affects the associated uncertainty.

In an analysis using the initial measured spatial uncertainty for eGRID (1.03 km, see below), the number of the largest 311 point sources that represent one third of US CO_2 emissions from point sources that could expected to be misallocated based on changing grid resolution is shown in Table 1. Using purely geometric arguments it shows that for a 1 by 1 degree grid, 111 of the top 311 point sources are expected to be within 0.1 degree of a border. Based on the spatial uncertainty of these grid cells 15.8%, or 18 of the 111 point sources, would be expected to be reported in the incorrect grid cell. As intuitively would be expected, the increase in grid size decreases the number misallocated, but even at a size as large as 10x10 degrees, there are still likely allocation errors. As data sets move toward finer resolutions, thus raising the likelihood of misallocation, it becomes increasingly critical to understand spatial uncertainty.

Grid Size	Fraction in border	Number of LPS
0.1 by 0.1	100.0%	49
1 by 1	36.0%	18
2 by 2	19.0%	9
3 by 3	12.9%	6
4 by 4	9.8%	5
5 by 5	7.8%	4
6 by 6	6.6%	3
10 by 10	4.0%	2

TABLE 1. The relative size of a 0.1 degree wide border region for different grid sizes and the number of the largest 311 point sources (LPS) potentially mis-allocated as a consequence if the locational uncertainty is 1.03 km.

Similarly placement of a point source within a grid cell is of concern for point sources. Again because of the fixed spatial error found in the data set, the closer to the edge of a grid cell a point source winds up, the higher the likelihood that the inherent error from the data will result in misallocation of that point source to a neighboring cell. A power plant found in South Africa (Figure 4), for example, just

happened to end up at exactly the corner of four grid cells so that the actual stacks are in a separate grid cell from the main offices, which are in a different grid cell than the street address. Thus depending on which is reported by the facility the full value of that plant's emissions could be allocated to any of three different grid cells.

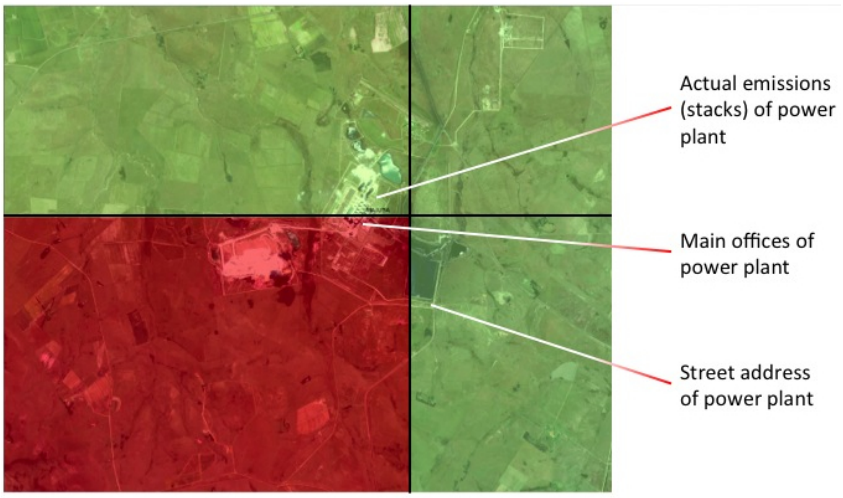


FIGURE 4. South African power plant shown with OMI data, which potentially places the emissions in a different grid cell than the actual emissions stacks.

It therefore becomes relevant to determine the location of point sources with as much accuracy as possible, but even when the location is known to within a few hundred meters, it is still crucial to have a means of reporting the uncertainty associated with each location so that other factors, such as issues of grid placement and resolution can be properly accounted for.

However, there is not currently an established methodology to deal with the spatial uncertainty of large point sources. In previous analyses spatial uncertainty in the United States has largely been assumed to be zero, and globally it has remained unaddressed (see, for example, [8]), despite having significant influence on carbon accounting and policy decisions.

The following sections develop a method for comparing emissions data sets and evaluating their associated uncertainty. Within a single data set the spatial uncertainty can be quantified and ways to reduce that uncertainty are discussed. The uncertainty in the total emissions value, hereafter referred to as magnitude uncertainty, is not part of this analysis. With two different types of uncertainty for each large data point, however, I do address effective means of reporting total uncertainty for various emissions. While atmospheric flux models may propagate these uncertainty values through separately, with their associated location or emissions value, two separate numbers are difficult to present on a map and fail to give a clearly understandable picture of confidence in the data. Thus a combined uncertainty

measure has been developed to allow the reporting of a single value that describes the uncertainty in the data at each location based on uncertainty in both the emissions total and reported location.

3. Methods

3.1. **Calculating Spatial Uncertainty.** In order to determine the spatial uncertainty in eGRID reported locations, a sample of 500 random points was selected from the dataset. Using Google Earth satellite imagery the reported location of each point was found and the surrounding area was searched to visually identify the actual location of the power plant stacks. Where none were immediately apparent, common locations were targeted as first search areas and then verified with addresses and company information. These included landfill sites, outskirts of small towns, bodies of water, and rail lines. The actual location, once found and verified, was recorded and the difference between the actual and original values was determined. The sample mean of the separation distance was 0.84 km, and this is then used as the spatial uncertainty for the eGRID data in further simulations.

3.2. **Monte Carlo Simulation.** The calculated spatial uncertainty provides a basis for investigating the confidence in the reported emissions values, but as data are normally aggregated into gridded formats it is necessary to incorporate the dependence on grid resolution and the location of a point source within a grid cell into an uncertainty metric. In order to take these factors into consideration a Monte Carlo simulation was used, inputting the reported location and calculated spatial uncertainty of a power plant and calculating the proportion of the time the emissions would fall in the original or surrounding cells. The simulation can compute expected emissions values for each grid cell from which a final spatial uncertainty value can be computed and it can concurrently attempt to refine and reduce that uncertainty.

3.2.1. *Computational Algorithm.* The Monte Carlo simulation takes an input emissions value and the spatial uncertainty for a single power plant, as well as geographic coordinates for that plant, and calculates a bivariate normal distribution using the reported coordinates of the power plant as the mean and the previously collected sample data to derive the standard deviation. This distribution is then used to generate 10,000 sample points placed on a 9x9 grid of (for example 0.1x0.1 degree) cells surrounding the cell in which the power plant was reported. The sample points are summed up for each cell and divided by the number of sample points to give a total emissions value for each cell. Points which fall outside of the county in which the original point source was reported are excluded from the total because our random sample suggested that the point sources were almost always placed within the correct political jurisdiction even when the latitude and longitude were uncertain. In summary form our approach was to:

- Generate a temporary grid around the reported location of the source
- Generate sample points on the grid around the reported location
- Exclude points falling outside of the reported county
- Compute the expected values by grid cell
- Calculate the uncertainty and combine this with uncertainty values from other sources

The result is the expected value of the emissions in each grid cell. This acts to distribute the original total carbon dioxide emissions over multiple surrounding cells based on the proportion of the sample points that fell in each cell, which, as discussed previously, is dependent on both placement of the point source within the cell as well as the grid resolution. If a power plant is located in the center of the cell at large enough resolution, all the emissions will be allocated to the same grid cell by the simulation because the probability that the point is actually in the reported cell is extremely high. However, for points near an edge of a grid space, the simulation redistributes the emissions to neighboring cells to reflect the higher probability that the point might fall in an adjacent space based on its spatial uncertainty (Figure 5). The simulation output provides the expected values for emissions from a given point source based solely on spatial uncertainty.

The expected values can then be used to calculate a final spatial uncertainty measure of the reported data for each grid cell, an uncertainty measure for the expected values, and a magnitude uncertainty, taken to be the expected values multiplied by a measure of the magnitude uncertainty for the large point source. The spatial uncertainty calculations are discussed in detail in the following section. A single grid space can, of course, have non-zero probability of containing multiple large point sources.

- Inputs: calculated spatial uncertainty, point source magnitude and reported location
- Expected value output: expected value distribution of the true location
- Uncertainty output: measures of spatial uncertainty in reported and expected values

As a simplified example, consider two test points, one in the center of a 0.1x0.1 degree grid cell, and the other in the corner. The first produces expected values of 100 in the original location and 0 elsewhere. Since the spatial uncertainty is less than the distance to any side of the cell, this is reasonable. The second point gives expected values that spread the emissions into three neighboring grid cells (Figure 5). The placement of the source puts this source a little over 1 km from one edge and a little over 1.5 km from the other. Since only spatial uncertainty is incorporated into the simulation, the total emissions of all cells should remain the same.

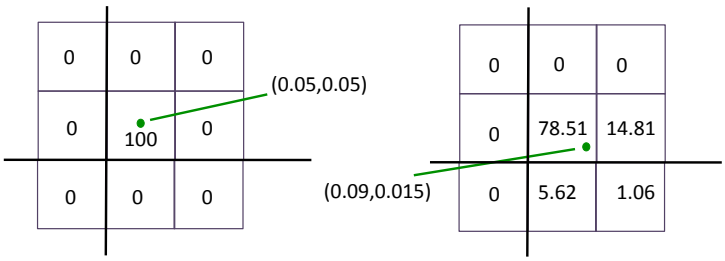
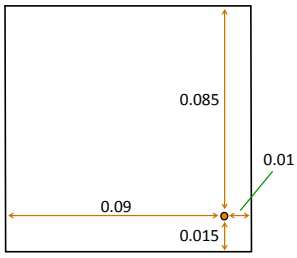


FIGURE 5. The expected values that would be output by the Monte Carlo simulation for each test point in tons of CO_2.

4. Statistical Metrics

Suppose we look at the emissions from a single point source with emissions of 100 tons of CO_2. We locate the source near the corner of a grid cell as shown in Figure 6a. The spacing is not critical to the present discussion, but the grid is assumed to be a 0.1 by 0.1 degree grid. The placement of the source puts the source a little over 1 km from one edge and a little over 1.5 km from the other.



(A) A single source located near the corner of a grid cell.

0	0	0
0	100	0
0	0	0

(B) The single point source from 6a shown with an example emissions value of 100 tons of CO_2 and neighboring grid cells, all with no emissions.

0.00	0.00	0.00
0.00	78.51	14.81
0.00	5.62	1.06

(C) The expected values of the emissions from the indicated point sourse based on the Monte Carlo simulation methodology described above.

0.00	0.00	0.00
0.00	100	100
0.00	100	0.00

(D) The 95th percentile confidence interval widths based on the Monte Carlo simulation results.

FIGURE 6. Example calculation of expected values and confidence interval for a sample point on a 0.1x0.1 degree grid.

If we look at neighboring grid cells, the reported emissions from this point source would look like Figure 6b, where the location of the point source is included in the figure for reference.

As outlined above, a Monte Carlo simulation determines the expected values of the emissions.

The result of the simulation is a grid of expected values for the emissions. Intuitively this can be thought of as a combination of the probability that the emissions occur in a particular grid cell combined with the quantity of emissions. In our test case, we get the grid shown in Figure 6c.

For reporting purposes, we will want to retain both the original reported values and the expected values. In addition, we want to describe the uncertainty in the reported value in a style suggestive of a 95th percentile confidence level. There are some sensitive issues related to reporting the uncertainty, and we propose a model for discussion.

Why can we not use a 95th percentile confidence interval? To help get a good handle on this basic issue, we refer to our sample case again. What does the 95th percentile interval tell us? It tells us the interval in which we are 95 percent confident that the true value lies within. For our case, we are 78.51 percent confident that the source actually lies in the center grid cell. This calculation is simplified since we used 100 as our total emissions, but we can make the same calculation in other cases in the following manner.

Since the emissions all must occur in the same location (it is after all a single point source), the values in the grid cells can be converted to the percentage of the emissions in the grid cell by dividing by the total emissions shown in each cell. Then we assume that the percentage of emissions correspond to the likelihood of the source being in that grid cell. So,

$$\% \text{ confidence} = \text{grid cell expected emissions/sum of grid cells}$$

The result is that we are not 95 percent confident that the source is located in the central grid cell where the source is reported. If the source is actually located in another grid cell, what would the emission be in the central cell? It would be 0 since there is no source there, and this would occur 21.49 percent of the time. So in order to create an interval in which we are 95 percent confident that the actual emissions lie in the interval, we must include 0. So the uncertainty (the plus/minus value) must be 100 so that the reported value and associated uncertainty would be 100 ± 100.

The same is true in the cells reported as 14.81 and 5.62, except in these cases we are not 95 percent confident that the emissions do *not* lie in the cell. To allow an interval in which we are 95 percent confident that the true value lies in the interval, we must expand the interval to include the value 100. This produces the grid in Figure 6d of the 95th percentile confidence widths.

Unfortunately the results shown in Figure 6d are not particularly enlightening. Our intuition tells us that we are less confident in the central cell being 100 than the outlying cells being 0. We should ask our statistic to give us more useful information than the grid of 95th percentile intervals.

So we move on to analyze expected values in an alternate way that provides a better reflection of the differences in uncertainty in the grid cells. We consider two alternatives based on: the uncertainty in the expected value, and the uncertainty in the reported value.

4.1. **Uncertainty in the Expected Value.** We might look to the standard deviation in the expected value for a measure of the uncertainty. If we look at the standard deviation formula, we get a straightforward way of calculating this measure.

$$SD = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

Here we calculate the standard deviation at each grid cell. For a grid cell, the expected value from the Monte Carlo simulation is the mean and the individual runs of the simulation are the x values. The individual runs are all either 100 or 0 in this case and we let p be the proportion of the time that the value is 100 and $(1 - p)$ be the proportion of the time that the value is 0. Since our total is 100, the expected value divided by 100 is this proportion. So,

$$SD = \sqrt{\frac{\overline{x}}{100}(100 - \overline{x})^2 + \frac{100 - \overline{x}}{100}(\overline{x})^2}$$

Factoring out common terms inside the square root and simplifying, the get

$$SD = \sqrt{\overline{x}(100 - \overline{x})}$$

This then provides a measure of the uncertainty in the expected value calculation. If we assume that twice the standard deviation provides roughly a 95 percent confidence interval, we have a measure of the uncertainty. It is of course asymmetrical since the interval should not go above 100 or fall below 0, but we leave the full interval for now to give an indication of the strength of uncertainty. For our test case above, this gives an uncertainty measure for the nearby grid cells shown in Figure 7.

This measure is a nice measure of the uncertainty in the expected value calculations. However depending on the application it may be more relevant to give the reported values from the data set along with an uncertainty measure, so we still require a metric to describe the confidence in those values. The standard deviation for the expected values has a maximum when the expected value is at 50 (i.e. when there is only a 50% chance that the point is actually in the grid space where it was reported) and diminishes to both sides. For a measure of the uncertainty of reported values, the uncertainty for a cell containing the reported value should continue to increase as the expected value decreases below 50.

0.00 ±0.00	0.00 ±0.00	0.00 ±0.00
0.00 ±0.00	78.51 ±82.16	14.81 ±71.04
0.00 ±0.00	5.62 ±46.06	1.06 ±20.48

FIGURE 7. The confidence interval widths based on the standard deviation of the expected values from the Monte Carlo simulation results.

Would this ever happen that the expected value decreases below 50? There are two situation where it definitely occurs. The first could be argued should not really happen because it suggests that the grid spacing is much too small for the magnitude of the spatial error. This happens if the expected value is spread out over many cells because the expected value is large. The second case is simply if the reported value is near a corner with three other grid cells. No matter how small the spatial error might be, if you are very close to the corner, the expected value in each of the four grid cells will approach 25. That is, each of the four grid cells has roughly a 25 percent chance of actually containing the point source.

At the 25 percent level, the uncertainty for the cell in which the source is reported should be larger than if the expected value is 50. This does not happen if we use the standard deviation of the expected values.

The reason is that this is the standard deviation of the expected values and not a measure of the reported value. To understand the uncertainty in the reported values, we perform a very similar calculation, but use the reported value as the mean rather than the expected value.

4.2. **An alternate measure.** Recall that the basic calculation for a standard deviation of a value in a population is

$$SD = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

which we might rewrite as

$$SD = \sqrt{\sum \left(\frac{1}{n} (x - \overline{x})^2 \right)}$$

This second relation reminds us that this is basically an average where each element is given equal weight in the sum. Now if there were multiple entries with the same value, we might combine them and form a weighted sum according to their frequency or probability.

$$SD = \sqrt{\sum (p(x - \overline{x})^2)}$$

where $p = f/n$ is the frequency that the particular value $(x - \overline{x})^2$ occurs.

We now apply this idea to the expected values of the simulation by looking at the differences between the simulation value at each grid cell and the reported value in that grid cell. Each outcome in the simulation produces either a 100 or a 0 in each cell and therefore the difference is either 100 or 0, but with a frequency related to the expected value from the simulation. In our case, we only have to add up two different outcomes for our sum, the times when a 100 appears and the times when a 0 appears. Therefore we defined the uncertainty to be

$$U = \sqrt{p_1(100)^2 + (1 - p)(0)^2}$$

where p is the frequency that the simulation result differs from the reported value by 100, and hence $1 - p$ is the frequency that the simulation result differs from the reported value by 0. Since the source is either in the cell or not, these are the only two options. Of course, this simplifies to

$$U = \sqrt{p_1(100)^2}$$

where p remains the frequency that the simulation value and the reported value differ by the 100 (the total emissions as reported). For our sample case, this gives the grid of uncertainty values shown in Figure 8.

Using these values, we can create a grid describing the reported value at each location and an associated uncertainty value. Standard practice might suggest that since the 95th percentile confidence interval is roughly twice the standard deviation, we should double these values in reflecting the level of uncertainty. However, the values computed from this method become large if doubled and outweigh the actual emissions values for frequency values below 0.75. The grid in Figure 9 shows the reported values in the sample case along with the uncertainty values.

Since the calculation depends entirely on the presence or lack thereof of the emissions occurring in a grid cell, a table can be created to provide an idea of what these numbers would be in different situations.

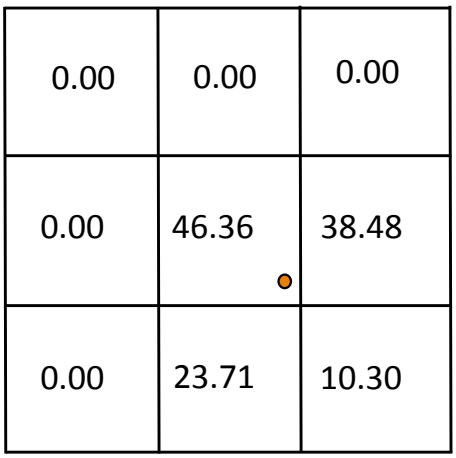


FIGURE 8. The standard deviations around the reported values based on the Monte Carlo simulation results.

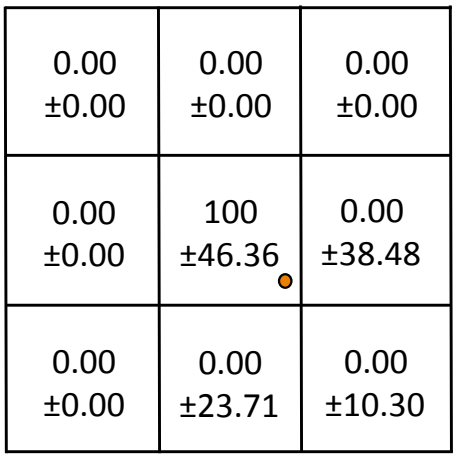


FIGURE 9. The reported values in the sample case, along with values of uncertainty expressed as a simple plus/minus notation. These of course should be viewed as asymmetric intervals since the values of 100 and 0 cannot be exceeded. Another representation would be to simply include a second grid labeled as the uncertainty measure.

In table 2, the values are calculated based on a single point source with emissions of 100 tons of CO_2. The uncertainty values can be scaled to other emissions numbers by multiplying integral units of 100.

4.2.1. *A combined metric for spatial and magnitude uncertainty.* In some applications, such as maps for public use, it is relevant to have a measure of uncertainty that takes into account both spatial and magnitude uncertainty for a given point source. Magnitude is not the focus of this analysis and we use the value of 10.62% derived for coal-fired power plants in eGrid point-source data by Quick (2014) and Quick personal communications in 2013. This particular value serves to demonstrate the process and is not critical to the present discussion.

Prob.	U	Prob.	U	Prob.	U
1.00	0.00	0.85	38.73	0.67	57.45
0.999	3.16	0.83	41.23	0.65	59.16
0.99	10.00	0.81	43.59	0.63	60.83
0.97	17.32	0.79	45.83	0.61	62.45
0.95	22.36	0.77	47.96	0.59	64.03
0.93	26.46	0.75	50.00	0.57	65.57
0.91	30.00	0.73	51.96	0.55	67.82
0.89	33.17	0.71	53.85	0.53	68.56
0.87	36.06	0.69	55.68	0.51	70.00

TABLE 2. Uncertainty values for different expected probabilities (the expected probability that the source is in the reported grid cell). Because these values are based on an emissions quantity of 100 tons of CO_2, these values are equivalent to a percentage uncertainty of the reported value.

In order to combine these uncertainties, one key assumption was made, namely 100% correlation between spatial and magnitude uncertainties. Because of this, the two two uncertainty types then can be added linearly to create a combined uncertainty metric. This takes into account all the associated uncertainty in the data and presents a comprehensive value for the uncertainty in a gridded data product of emissions. Similar to the spatial uncertainty metric described in the previous section, this is different in concept from a 95% confidence interval and cannot be thought of as a plus or minus value on a grid cell. Instead it should be envisioned as a quantitative representation of the total uncertainty on a grid cell based on the emissions in or near that cell. If divided by the total emissions it would represent the maximum fraction of the emissions total that could be found in that cell, although, again, as the emissions are binary in nature it is not reasonable to think of a fraction of their total in any given cell.

5. Spatial Resolution

With the calculation of spatial uncertainty defined, we can now produce a gridded map containing the point sources along with a companion map showing the accumulated uncertainty of the point sources. These maps, or data sets if you like, can then be incorporated into larger efforts to define and characterize global carbon emissions, sequestration, and stocks.

One of the remaining tasks from the standpoint of the point sources is to determine the appropriate grid on which to report the values. Part of this tasks lies with the other pieces of the puzzle. If we can report the values on the same level of resolution used in other data products, then the integration will be less cumbersome. On the other hand, if we report the data on too fine a grid, the uncertainty quantities will be so large that the values will be undermined. Therefore, we investigate the resolution of the point source data to determine the minimal grid size on which the data are meaningful.

We recognize that there is always some probability that a point source will lie near a grid border and that the spatial uncertainty will therefore be correspondingly large. In light of this, we look toward defining an averaged uncertainty measure to determine the overall level of spatial uncertainty.

Here we run an additional simulation, placing a large number of random points within a grid of a specified size. We calculate the average uncertainty measure of all such points and use this number to quantify the level of uncertainty on that grid size. By repeating this simulation on multiple grid sizes, we propose a threshold of 5% for which the grid size is appropriate, i.e. when the mean spatial uncertainty due to the dataset is equal to 5% of the grid dimension.

Since it is the relative size of the grid to the spatial uncertainty that matters, we report the grid size as a function of the mean locational uncertainty. For example, the data we collected from eGRID suggested an average uncertainty of 0.84 km. If we report the data on a grid of 0.1 by 0.1 degree grid, then the uncertainty is a little less than one tenth the size of the grid. We use this ratio to evaluate a useful resolution at which to report the data.

6. Sample Simulation Outputs

The Monte Carlo simulation elaborated above was applied to eGRID data for each state in the continental United States to produce expected and uncertainty values on a 0.1 x 0.1 degree grid. The expected values are shown in Figure 10 for the Southeastern United States based on reported values in eGRID for 2009. It is clear that emissions from large point sources are distributed over grid spaces adjacent to the reported locations to reflect locational uncertainty, but that relatively few of the total number of grid spaces are affected by this locational uncertainty and that this locational uncertainty will be absorbed as the spatial resolution is increased.

For further illustration, expected and uncertainty values are shown for Iowa specifically in Figures 11a and 11b. Iowa is bordered on the right by the Mississippi River and the concentration of large power plants along this line is evident. These large emissions values are accompanied by large spatial uncertainties with significant uncertainties potentially falling across the state line. However, because the computed points are constrained to the boundary of the reported county, cells shown which cross over into Wisconsin and Illinois still only contain the sum of the contribution of emissions to that cell from power plants in Iowa. Thus when summing expected values over regions or the entire nation there is no duplicate accounting.

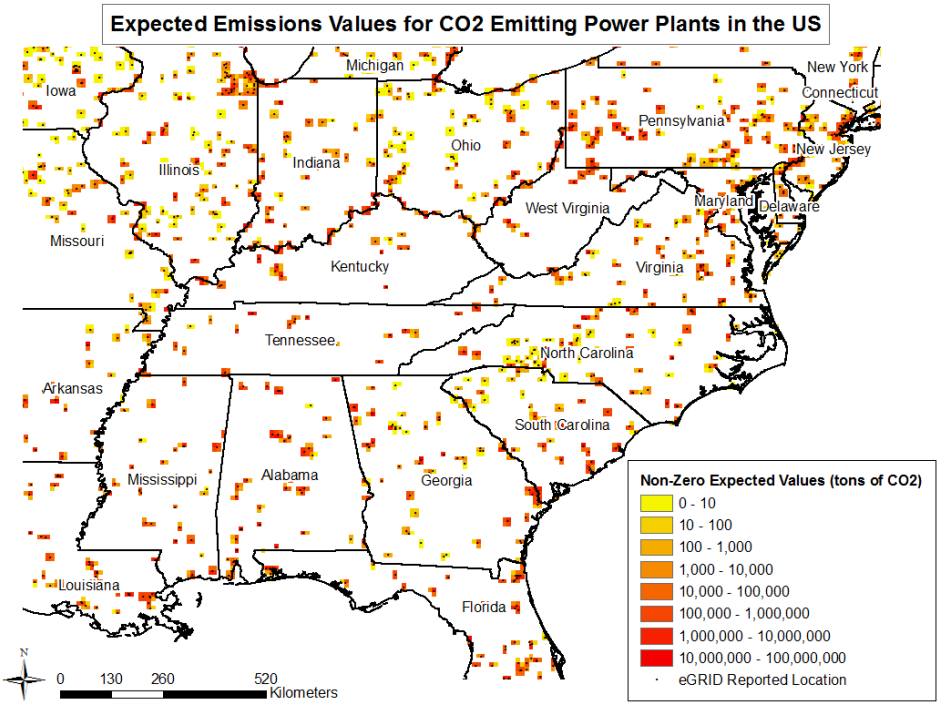
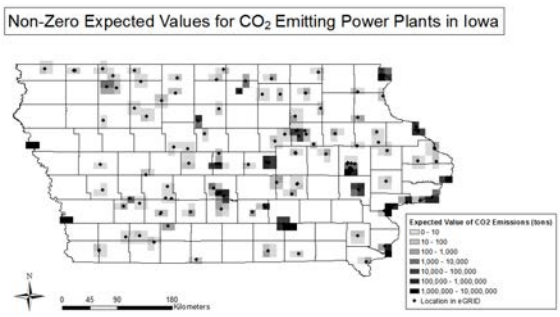
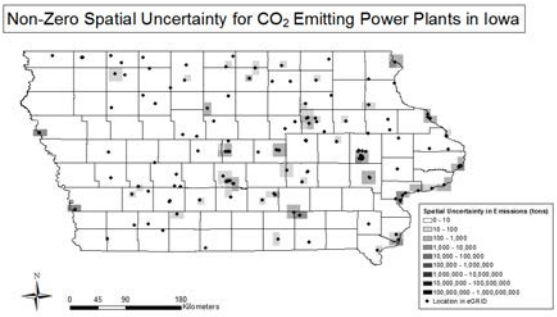


FIGURE 10. Expected values produced by the Monte Carlo simulation for the eastern United States from 2009 eGRID data of electric power generation in the USarcos on a 0.1x0.1 degree grid.



(A) Expected emissions values output by the Monte Carlo simulation for Iowa, computed from 2009 reported values from eGRID. Units are tons of carbon dioxide per year.



(B) Spatial uncertainty of reported emissions values in Iowa, given in tons of carbon dioxide per year. The uncertainty is computed from the expected values output by the Monte Carlo simulation and reported on 0.1x0.1 degree grid size. Only non-zero data are displayed.

FIGURE 11. Simulation output in Iowa as an example, showing both river borders, and variation in expected values and uncertainty between power plants depending on grid placement.

7. DISCUSSION

Large point source totals are highly influential in overall totals of CO_2 emissions totals and therefore it is critical to understand the issues associated with them and to have a means of quantifying and reporting

their uncertainty in both magnitude and location. However, the binary nature of these emissions sources precludes traditional methods of dealing with their uncertainty. Small spatial errors and uncertainty may have order of magnitude effects on emissions totals in a grid cell. The approach presented here allows for the computation of spatial uncertainty values associated with large point source emissions. Placing a confidence interval on spatial uncertainty is already problematic, but the enormous impact arising from large point sources makes traditional confidence intervals irrelevant. Instead, we have presented a measure of uncertainty that provides a very clear qualitative understanding of the uncertainties involved in the point sources according to their grid placement and spatial resolution and that is also quantitatively useful in comparing uncertainties across the data set.

The data used here as an example are from the United States, which is widely believed to have one of the most reliable data sets on large point sources. We still believe this to be true; however, we now begin to show the potential consequences of these relatively small uncertainties. Continuing studies will explore what happens when we look at data sets that have more spatial uncertainty. One of the main difficulties, which we will address relates to the disproportionate number of point sources that lie on political boundaries (e.g. sources of water). Misplacing a few point sources along these boundaries can potentially reallocate significant emissions to another party - unless special care is taken in keeping the data disjoint.

This doesn't seem particularly troublesome at first, since we would hope that the locations of the large point sources will be corrected in the near future. We expect that this will happen in the US fairly soon as the owners of the data sets realize that the variety of uses of the data sets extend beyond their original scope, however it may take some time in other countries. This also becomes a major issue as other types of data are also repurposed. If the goal is to use remote sensing data as a verification or validation tool, the idea of spatial error in large point sources becomes increasingly important.

Again, we re-emphasize the importance to treating large point sources very carefully. For these reasons, we recommend keeping large point source data from different political domains separate (layered) unless the conglomerate data is truly necessary. This layered approach might also be useful for other transitions as well, such as between land and ocean, where the dynamics of carbon transport are very different.

In addition to providing a measure of uncertainty for spatial causes, we also find that combining these uncertainties with magnitude uncertainty still provides a useful measure and we can calibrate the magnitudes to provide a semblance of consistency. This enables the data to be implemented for varying purposes. The utility of combined or separate uncertainties is dependent on the purpose and intended audience.

The general approach developed here extends beyond point source data and intuitively can be used to combine multiple data types and obtain uncertainty values. Further analysis is needed to continue to refine the uncertainty outputs based on characteristics of the point sources. Preliminary analysis suggests that uncertainties can be refined based on proximity to water sources, political structures, and population centers for example (depending on the country). The fuel source of the emissions also likely will help to clarify levels of spatial uncertainty.

References

[1] Clean Energy: eGRID. http://www.epa.gov/cleanenergy/energy-resources/egrid/, 2013.

[2] The EU Emissions Trading System (EU ETS). `http://ec.europa.eu/clima/policies/ets/index_en.htm`, 2014.

[3] Regional Greenhouse Gas Initiative. `http://www.rggi.org/`, 2014.

[4] United States Environmental Protection Agency. Carbon Pollution Standards. `http://www2.epa.gov/carbon-pollution-standards/what-epa-doing`, 2014.

[5] U. Cubasch, D. Wuebbles, D. Chen, M.C. Facchini, D. Frame, N. Mahowald, and J.-G. Winther. *Introduction*. University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[6] R.A. Houghton, G.R. van der Werf, R.S. DeFries, M.C. Hansen, J.I. House, C. Le Quere, J. Pongratz, and N. Ramankutty. Chapter G2 Carbon emissions from land use and land-cover change. *Biogeosciences Discussions*, 9:835–878, 2012.

[7] G. Myhre, D. Shindell, F.-M. Breon, W. Collins, J. Fuglestvedt, J. Huang, D. Koch, J.-F, Lamarque, D. Lee, B. Mendoza, T. Nakajima, A. Robock, G. Stephens, T. Takemura, and H. Zhang. *Anthropogenic and Natural Radiative Forcing*. University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[8] T. Oda and s. Maksyutov. A very high-resolution (1 km x 1 km) global fossil fuel co2 emission inventory derived using a point source database and satellite observation of night lights. *Atmospheric Chemistry and Physics*, 11:543–556, 2011.

[9] S.H. Scheider and K. Kuntz-Duriseti. *Uncertainty and Climate Change Policy*, chapter 2. Island Press, Washington D.C., USA, 2002.

[10] R.W. Shonkwiler and F. Mendivil. *Explorations in Monte Carlo Methods*. Springer, 2009.

[11] T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley. *Annex III: Glossary*. University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[12] T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley, editors. *Summary for Policymakers*. University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[13] World Economic Forum in collaboration with Her Majesty's Government, United Kingdom. *Statement of G8 Climate Change Roundtable*, June 2005.