Aggregation of Point Source Emissions to Spatial Emissions with Uncertainty

Michael Pekala[1,3], April Nellis[1,3], Gary Collins[1,3], Krsna Raniga[2,3], Zoheyr Doctor[2,3], Dan Moore[2,3], Elizabeth Reilly[1,3], Marisa Hughes[1,3], and Gavin McCormick[2,3]

1) The Johns Hopkins University Applied Physics Laboratory (JHU/APL) , 2) WattTime, 3) Climate TRACE

1. Introduction

While individual asset emissions estimates and their associated uncertainties are provided for each sector within the Climate TRACE consortium, there is also often a need for spatially aggregated emissions. For example, there are various methods for modeling the transport of gas species throughout the atmosphere (such as [Byrne2023]); these models often require a regular grid of emissions as part of their initial condition. Mapping individual point emissions to such a grid entails aggregating (summing) point-level emissions.

Although Climate TRACE users are free to aggregate data themselves as desired, this document outlines methods for which pre-aggregated data products are available upon request. In this document we model individual asset emissions as random variables and aggregation with uncertainty measures consists of computing the distributions of sums of these random variables (r.v.s). For independent gaussian r.v.s this is straightforward; however, departures from independence or normality can add complexity. The currently available data products are baseline aggregation methods developed by the Climate TRACE coalition, with continued refinement planned for the future.

2. Data and Methods

2.1. Overview

This document focuses on the spatial aggregation of point source emissions of greenhouse gasses—also referred to as "assets"— generated by the Climate TRACE coalition. In particular, we focus on partitions of the surface of the earth, consisting of Global Administrative Areas (GADM) levels 0, 1, and 2 and regular grids (currently, grids whose cells are each 4x5 degrees). We operate at the temporal resolution provided by sectors and do not aggregate temporally. We also assume that each asset is “small” relative to the grid resolution so that each asset is contained entirely within a single grid cell and correlations between grid cells can be neglected. Hereafter, when we refer to an “asset”, we implicitly mean a point source asset. Furthermore, we focus on uncertainty quantification related to emissions; uncertainty quantification associated with other intermediate variables (e.g., emission factors, capacity, capacity factors) are addressed

by sectors. Finally, we explicitly focus on uncertainty quantification and do not consider the (more qualitative) confidence measures also available in the Climate TRACE data product.

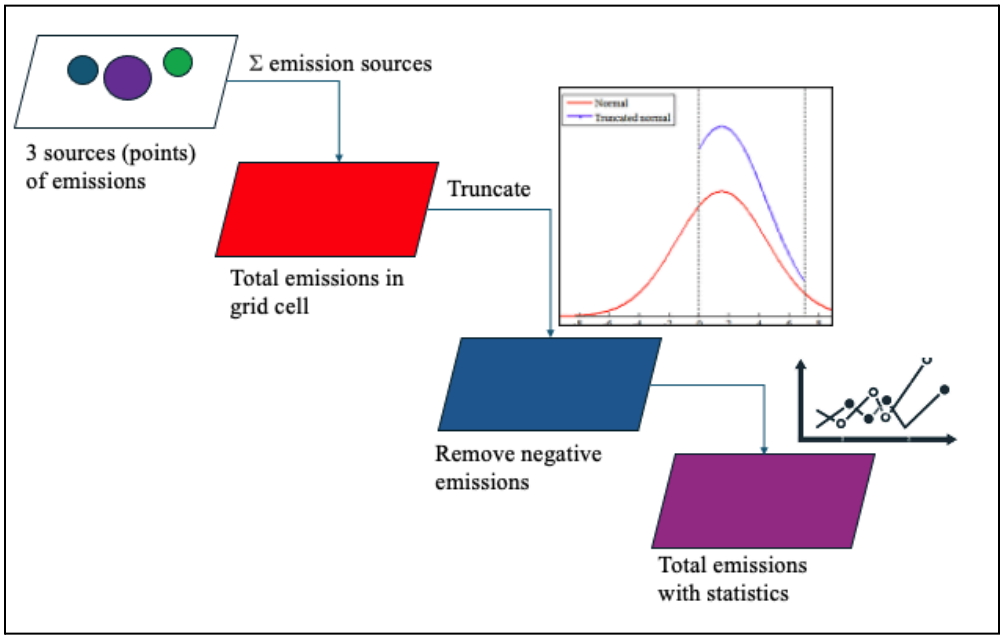

Figure 1: Current data flow and assumptions

2.2. Approach and Results

In order to say things about the sums of r.v.s, we must of course start from some specification of their distribution. The current version of the Climate TRACE data product provides a single scalar uncertainty value for each asset/gas species pair (this uncertainty data is available by request). We assume this uncertainty value represents a standard deviation and interpret the reported emissions value for that same asset/species as the mean. In the absence of any further information (e.g. a covariance matrix) we assume these r.v.s are independent . So, for example, the emissions of two spatially proximate power plants are currently treated as independent even though they may be jointly affected by local customer demand patterns. Cross-sector dependencies might also exist, e.g. the emissions from a power plant may be driven by that of another asset, such as a steel plant that is a power consumer. The topic of dependence could become quite complicated; the approach we take here is to establish a simple baseline from which we can gradually incorporate dependencies that are demonstrably significant (see Future Work).

Consistent with the convention employed by a number of Climate TRACE sectors as of phase 6, we further assume that these asset-level r.v.s follow a normal distribution. Note that these assumptions may restrict the sectors to which we can apply our current approach. Another identified direction for future work is to employ Monte Carlo or other more distribution-agnostic methods to support arbitrary distributions. Of course, this will require sufficient information to be provided by sectors in order to sample from said distributions.

Spatial uncertainty aggregation is implemented in two steps. First, upon ingestion of the sector data, each asset is assigned to spatial cells with respect to its GADM-0, GADM-1, GADM-2, and 4x5 grid location. This spatial information is then leveraged to efficiently sum assets within a given spatial cell.

2.2.1 Baseline

Standard results give that, for independent, normally distributed r.v.s X_i, i=1,…,n with parameters μ_i, σ_i, that $\sum_{i=1}^{n} X_i$ is normally distributed with parameters $\sum_{i=1}^{n} \mu_i$ and $\sum_{i=1}^{n} \sigma_i$ [Ross2019]. This approach is consistent with IPCC Tier 1 uncertainty quantification methodology [IPCC2021]. Under these assumptions, we have a closed form representation for the uncertainty associated with each grid cell. From this basic starting point, we can layer in relevant complexities.

2.2.2 Nonnegativity

One property of asset emissions that we consider here is that these are anthropogenic GHG *producing* sources. Meaning these assets only emit/produce emissions and do not take away, or reduce, emissions. The normal distribution does not enforce nonnegativity; however, provided the mean is sufficiently far above 0 (in units of standard deviations) permitting a negative emissions value with some small probability is not a huge concern. However, if the uncertainty is high enough such that negative values become likely, then we may want to truncate or otherwise adjust the uncertainty representation.

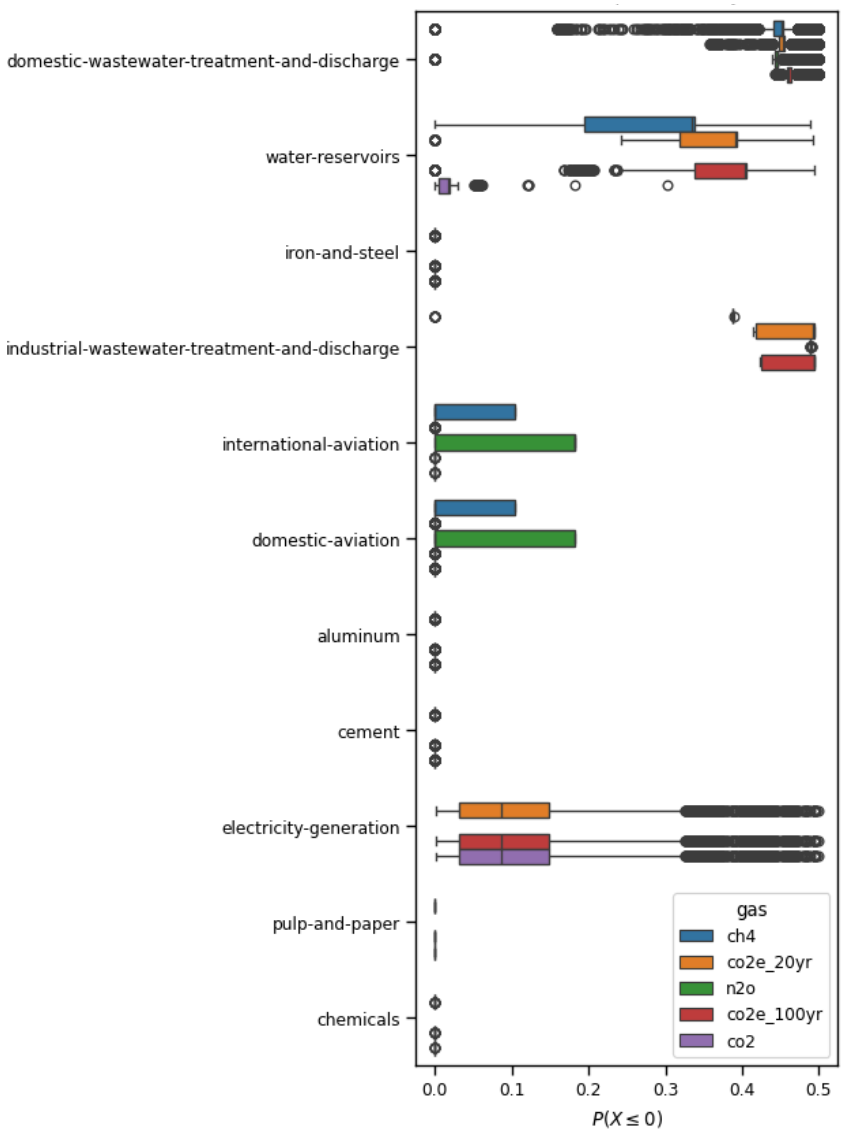

Figure 2: Distributions of P(X<=0) for individual assets (data from 2023). Sector/gas pairs where this probability is large indicate significant departures from nonnegativity.

To assess whether this is relevant for Climate TRACE we explore the magnitude of asset uncertainty in the context of this nonnegativity property. For each asset, we compute $P(X \leq 0)$ and Figure 2 shows the distribution of these values by gas species and sector. For some sector/species pairs, negative emissions represent a significant proportion of the distribution. Indeed, if we look at individual entries within the database (e.g. Figure 3) we observe that this is caused by asset-level uncertainty values where the standard deviation is large relative to the mean.

original_inventory_sector	asset_id	gas	emissions_quantity	emissions_quantity_std	coefficient of variation
domestic-wastewater-treatment-and-discharge	3881524	co2e_20yr	22.915533	1660.730676	72.471833
electricity-generation	4407210	co2e_100yr	15.250000	7300.000000	478.688525
industrial-wastewater-treatment-and-discharge	4417474	co2e_20yr	231.230509	13815.492986	59.747708
water-reservoirs	4414702	co2e_100yr	3920.053928	240000.000000	61.223648

Figure 3: Example assets for which the coefficient of variation (the ratio of the standard deviation to the mean) is large, suggesting that negative values are likely for this distribution.

Ideally, additional analysis could be done at the sector level to reduce the uncertainties, but if this is not possible then another approach is to change the uncertainty representation. One option is to employ a truncated normal distribution, where we truncate from the left at x=0 and from the right at some distance that is substantially "far" from the mean (say, 5 standard deviations above the mean) since we do not want to substantially modify the right tail. We perform this truncation after summing (i.e. within the context of each spatial cell). The result of truncation will be a new distribution derived from the original "parent" normal distribution where the mean and variance will be perturbed [Burkhardt]. In the case of the mean, this perturbation will correspond to a shift to the right (increased emissions) in cases where the left tail is substantially truncated. A substantial perturbation to the mean may be undesirable if it is the case that an asset's emissions estimate is reliable but the associated uncertainty estimate is not.

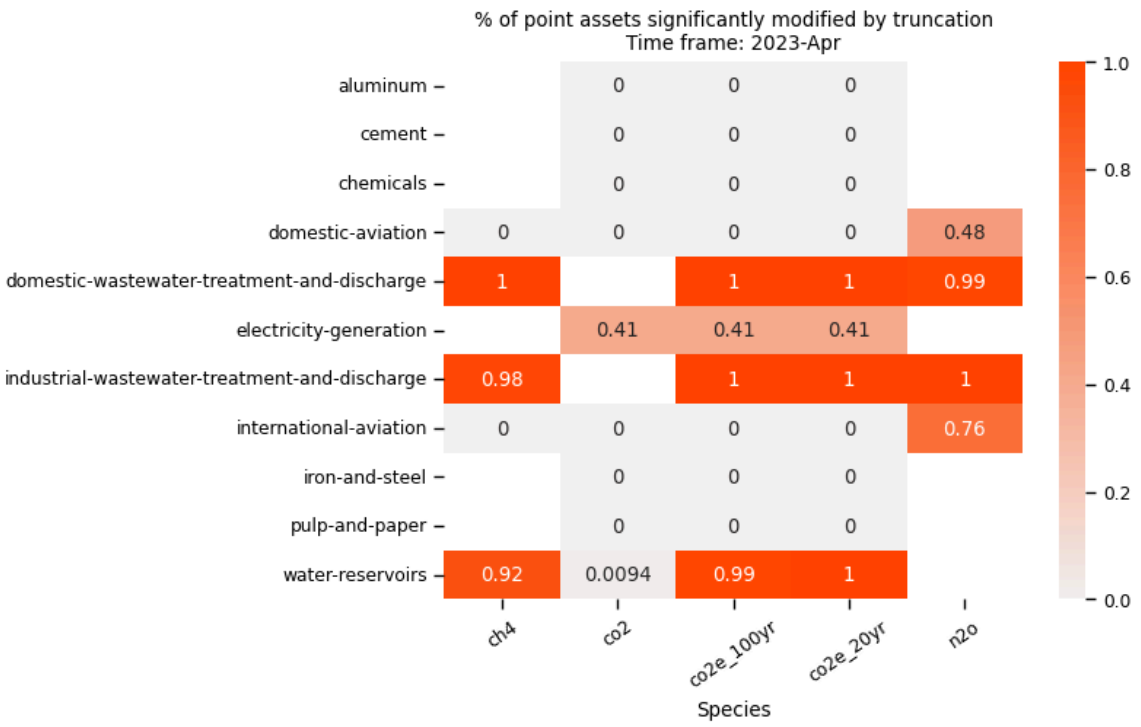

Figure 4: Heatmap indicates the proportion of point assets which, after moving to a truncated normal distribution, would have a mean emissions value that is "significantly" different from the original mean, with "significantly" defined as a relative change of at least 20%.

In Figure 4 we visualize the impact of truncation to this same 2022 data. The figure depicts the proportion of asset-level emissions that would have been substantially modified by truncation, where here "substantial" corresponds to more than a 20% relative change (δ) in the mean

$$\delta = \frac{\mu_t - \mu_0}{\mu_0}, \text{ (Eq. 1)}$$

where μ_0 corresponds to the original mean from the parent distribution, and μ_t is that of the truncated normal. These large perturbations map to situations where $P(X \leq 0)$ is large (Figure 2) as expected. In the data product, we provide the mean and standard deviation of the truncated distribution, along with 0.05 and 0.95 quantiles. Figure 5 shows an excerpt of this data product.

	0	1	2	3	4
original_inventory_sector	aluminum	aluminum	aluminum	aluminum	aluminum
gas	co2	co2e_100yr	co2e_20yr	co2	co2e_100yr
spatial_support	ARE	ARE	ARE	ARE	ARE
temporal_support	2023-Apr	2023-Apr	2023-Apr	2023-Aug	2023-Aug
method	ind. truncated normal	ind. truncated normal	ind. truncated normal	ind. truncated normal	ind. truncated normal
emissions_quantity_agg	698832.688023	698832.688023	698832.688023	734050.398864	734050.398864
emissions_quantity_std_agg	82053.955092	82053.955092	82053.955092	85956.638719	85956.638719
q_0.025	538009.891254	538009.891254	538009.891254	565578.482743	565578.482743
q_0.975	859655.484792	859655.484792	859655.484792	902522.314984	902522.314984
n_assets	3	3	3	3	3
version	v0.0	v0.0	v0.0	v0.0	v0.0
timestamp	October 30, 2024	October 30, 2024	October 30, 2024	October 30, 2024	October 30, 2024

Figure 5: Example of aggregated emissions data (5 entries only). The table (transposed above, for readability) includes column specifying the sector/species pair, identifies the spatial (here, GADM 0) and temporal (here, monthly data from 2023) support, the aggregation method used (here, truncated gaussians), basic statistics (mean, standard deviation, 0.025, and 0.975 quantiles), the number of assets contributing to the sum, the software version, and the time when the data set was created.

3. Conclusions and Future Work

We describe a simple baseline for generating spatially aggregated uncertainty values under a particular parametric assumption for asset uncertainties and also explore one possibility for enforcing a nonnegativity property. This approach is an initial baseline and there are abundant opportunities for extensions that the team is considering for future phases, including:

- Selectively relaxing the assumption that asset-level emissions can be treated as independent random variables. A key here is to identify sources of dependence that are significant to the aggregated result (not all dependence may matter). There are various ways that dependence could enter - one we are actively exploring is how dependence arising from disaggregation might be incorporated [Pekala2024].
- As mentioned above, asset-level emissions may not follow a normal distribution - they only emit emissions and do not reduce emissions - in which case Monte Carlo methods can provide an alternative method for estimating net emissions [IPCC2021]; this requires that asset-level uncertainties be sufficiently well specified (by sector teams) to permit sampling.
- In situations where high-confidence mean emissions for an asset can be obtained but further distributional assumptions cannot be made, another option we are exploring is to leverage concentration inequalities to characterize uncertainty.

For access to spatially aggregated data, please reach out to coalition@climatetrace.org.

4. Supplemental Information

Permissions and Use: All Climate TRACE data is freely available under the Creative Commons Attribution 4.0 International Public License, unless otherwise noted below.

Data citation format: Pekala, M., Nellis, A., Collins, G., Raniga, K., Doctor, Z., Moore, D., Reilly, E., Hughes, M., and McCormick, G. (2024). *Aggregation of Point Source Emissions to Spatial Emissions with Uncertainty.* The Johns Hopkins University Applied Physics Laboratory (JHU/APL) and WattTime, USA, Climate TRACE Emissions Inventory. https://climatetrace.org [Accessed date]

Geographic boundaries and names (iso3_country data attribute): The depiction and use of boundaries, geographic names and related data shown on maps and included in lists, tables, documents, and databases on Climate TRACE are generated from the Global Administrative Areas (GADM) project (Version 4.1 released on 16 July 2022) along with their corresponding ISO3 codes, and with the following adaptations:

- HKG (China, Hong Kong Special Administrative Region) and MAC (China, Macao Special Administrative Region) are reported at GADM level 0 (country/national);
- Kosovo has been assigned the ISO3 code 'XKX';
- XCA (Caspian Sea) has been removed from GADM level 0 and the area assigned to countries based on the extent of their territorial waters;
- XAD (Akrotiri and Dhekelia), XCL (Clipperton Island), XPI (Paracel Islands) and XSP (Spratly Islands) are not included in the Climate TRACE dataset;
- ZNC name changed to 'Turkish Republic of Northern Cyprus' at GADM level 0;
- The borders between India, Pakistan and China have been assigned to these countries based on GADM codes Z01 to Z09.

The above usage is not warranted to be error free and does not imply the expression of any opinion whatsoever on the part of Climate TRACE Coalition and its partners concerning the legal status of any country, area or territory or of its authorities, or concerning the delimitation of its borders.

Disclaimer: The emissions provided for this sector are our current best estimates of emissions, and we are committed to continually increasing the accuracy of the models on all levels. Please review our terms of use and the sector-specific methodology documentation before using the data. If you identify an error or would like to participate in our data validation process, please contact us.

5.References

1. Burkardt, John. "The truncated normal distribution." *Department of Scientific Computing Website, Florida State University* 1.35 (2014): 58.

2. Byrne, B., et al. (2023). National CO2 budgets (2015–2020) inferred from atmospheric CO2 observations in support of the global stocktake. Earth System Science Data, 15 (2), 963–1004. DOI: **10.5194/essd-15-963-2023**

3. Intergovernmental Panel on Climate Change (IPCC). "Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories". June 2021.

4. Janson, Svante. "Large deviations for sums of partly dependent random variables." *Random Structures & Algorithms* 24.3 (2004): 234-248.

5. Lampert, Christoph H., Liva Ralaivola, and Alexander Zimin. "Dependency-dependent bounds for sums of dependent random variables." *arXiv preprint arXiv:1811.01404* (2018).

6. Pekala et al. "Aggregation of Point Source Emissions Data with Uncertainty," American Geophysical Union (AGU) poster presentation, December 2024.

7. Ross, Sheldon. *First Course in Probability, A*. Pearson Higher Ed, 2019.

8. Wainwright, Martin J. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.