Waste Sector: Estimating CH_4 Emissions from Solid Waste Disposal Sites

Krsna Raniga, Aaron Davitt, Christy Lewis, Lekha Sridhar, Lee Gans, and Gavin McCormick
All authors affiliated with WattTime and Climate TRACE

1. Introduction

Methane, or CH_4, is a potent greenhouse gas responsible for 30% of global warming since the industrial revolution, and is 28 times more potent than carbon dioxide over a 100 year timespan (Ayandele et al., 2022). Solid waste disposal sites (SWDS) are the third largest human-generated source responsible for increasing atmospheric methane worldwide, after fossil fuels and enteric fermentation. For example, SWDS accounted for 11% of global anthropogenic methane emissions in 2020 (Ayandele et al., 2022). In the United States alone, SWDS contribute 14.5% of greenhouse gas generation (EPA, 2022). Emissions from SWDS are generated from anaerobic decomposition of organic matter, the target contributor to which is municipal solid waste (MSW), or everyday "trash" or "garbage".

Globally, 37 percent of waste is landfilled, while 33 percent is openly dumped (Kaza et al., 2018). Within landfills, levels of management will range from covered, methane-collecting facilities, to simple designated waste depositing locations for all locally collected municipal waste. In contrast, dumpsites are better described as patches of emergent waste, and are most common in low-income nations, where 90 percent of waste is openly dumped or burned and landfilling is usually not yet available (Kaza et al., 2018).

Data on waste sites, crucial for understanding and mitigating methane emissions, are frequently varied, inconsistent, and/or incomplete. While the U.S. Environmental Protection Agency's (EPA) Greenhouse Gas Reporting Program (GHGRP), the European Pollutant Release and Transfer Register (E-PRTR), and Canada's Greenhouse Gas Reporting Program (Canada GHGRP) provide annual methane emissions data for each facility, these sources offer a fragmented view of the overall activity and emissions profiles at waste sites. Similarly, the EPA's Landfill Methane Outreach Program (LMOP) and Mexico's National Institute of Statistics and Geography (INEGI) offer insights into waste quantities and management practices, but the absence of direct emissions data limits their utility in creating a comprehensive emissions profile.

Global coverage efforts like OpenStreetMap, though broad in scope, suffer from issues of verification, with many cataloged sites likely not being MSW landfills. The Global Plastic Watch

introduces a novel approach by employing remote sensing to identify plastic-rich waste sites, but its focus on plastic might overlook other types of open waste sites or might falsely identify MSW landfills, mirroring some limitations seen with OpenStreetMap data. Lastly, the Waste Atlas, while a global catalog, is outdated, having last been updated in 2013, and thus only serves as a stopgap when other sources are exhausted.

Considering the complexity and variance of the waste data landscape, the Climate TRACE methodology for 2023 implemented Bayesian regression modeling. This was a data-driven, statistical approach, intended to closely align with the available information by generating a consistent framework for estimating emissions on a global scale. By integrating detailed facility-level data and open data with satellite emissions measurements from GHGSat and Carbon Mapper, the Bayesian approach was designed to accommodate varying degrees of data completeness and precision and to focus the effort to align emissions estimates more closely with the realities of the data encountered.

2. Materials and Methods

The waste datasets used by Climate TRACE were selected based on comprehensive data availability for three key parameters: land area, annual capacity (meaning annual incoming waste), and direct CH_4 emissions. If none of these were available, the waste site was omitted from the final composite dataset. The composite dataset was contingent on a combination of data richness, source reliability, and recency, where a hierarchy of preference in data source was implemented if multiple datasets overlapped on certain landfills (see Section 2.1.3).

The Bayesian method used for modeling capacities and emissions provided a probabilistic framework connecting hypotheses about waste sites, process (or causal) models generated from those hypotheses, and statistical models conjectured from these process models. The best model definition was identified by iterating over these steps of the framework and testing and cross-validating models in the context of the available data. This approach differed markedly from the deterministic Intergovernmental Panel on Climate Change (IPCC) first-order decay method used in Climate TRACE's 2022 release, by incorporating prior domain knowledge as probability distributions from the onset and allowing the model to learn from the data to constrain these distributions (IPCC, 2006; IPCC, 2019). Consequently, the chosen model reflected a balance of prior knowledge, data robustness, and computational feasibility.

2.1 Constructing the Climate TRACE Waste Dataset

Constructing the Climate TRACE dataset required pre-processing and integrating a variety of solid waste data sources, which varied widely in their data completeness and regional coverage. The primary goal was to construct a unified dataset, composed of unique waste sites, while

incorporating the most recent and comprehensive information available per location. Three datasets were constructed in the following stages:

1. Pre-modeling inventory: This stage involved creating a deduplicated, filtered concatenation of all individual contributing datasets, deploying a hierarchy of source preference when multiple sources reported on the same site.
2. Training dataset: In this stage, satellite plume data were spatially matched with waste sites, and metadata such as areas or capacities were harmonized between sources for the same sites.
3. Final dataset: This dataset mirrored the pre-modeling inventory in structure but included additional enhancements. Capacities and emissions were filled where needed by the Bayesian model, but self-reported emissions values from the EPA datasets, the Canada GHGRP, and the E-PRTR were included without modifications. This dataset was then cleaned, processed, and filtered one last time.

2.1.1 Core datasets employed

EPA GHGRP: The EPA's Greenhouse Gas Reporting Program (GHGRP) requires over 8,000 large-emitter (>25,000 megatons CO_2 equivalent annual emitting) facilities to report their greenhouse gas emissions each year (EPA, 2022). This is the most broad and comprehensive dataset publicly available on U.S. waste disposal sites. The reported data was acquired through the EPA's Facility Level Information on GreenHouse Gases Tool (FLIGHT), which includes annual accepted waste back to at least 2010 and 2022 emissions (https://ghgdata.epa.gov).

EPA LMOP: Waste disposal site data was collected from the Landfill Methane Outreach Program (LMOP; https://www.epa.gov/lmop). LMOP data sources include facility self-reports, LMOP partner reports, and publicly available data (EPA, 2022). Beyond land area, waste-in-place, annual capacity, and operating years, many sites published landfill gas (LFG) generation and collection values. For those sites which were not included in GHGRP reports, the net published LFG values (LFG generated minus LFG collected) were utilized for the reported year's methane emission value and rescaled to provide an updated estimate for 2022.

Canada GHGRP: Similar to the EPA GHGRP, Canada requires annual reports from facilities that emit at least 10,000 tonnes of carbon dioxide equivalent per year. 142 unique locations reported emissions as of at least 2016, up to 2021. 2021 data was forward-filled to 2022 without rescaling. Only emissions quantities per year and ownership information are published by the program (https://www.canada.ca/en/environment-climate-change/services/climate-change/greenhouse-gas-emissions.html).

E-PRTR: The European Pollutant Release and Transfer Register reports pollutant releases, including greenhouse gas emissions, from industrial facilities in the European Union (https://industry.eea.europa.eu/). There were 1,426 landfill locations identified as having reported

emissions since at least 2016 and up to 2021. No 2022 data was available, so 2021 data was forward-filled to 2022 without rescaling. The E-PRTR only publishes landfill locations and annual emissions quantities. Some locations are confidential, so those sites were omitted from the Climate TRACE inventory.

INEGI: Mexico's National Institute of Statistics and Geography (INEGI) published landfill locations, waste quantities, and practices in 2017 under its "National Census of Municipal and Delegational Governments 2017" (https://en.www.inegi.org.mx/). This dataset reports on 2,736 active landfills as of 2017, and reports information including their opening dates, landfill cover types, and existence of gas collection systems.

Waste Atlas: This database was initially released in 2013, designed as an interactive map with data compiled through crowdsourcing and scientific research (http://www.atlas.d-waste.com/). Since its initial release, and a 2014 report highlighting in-depth statistics about the world's 50 largest dumpsites, additional dumpsites and a sanitary landfill catalog have been added (Mavropoulos et al., 2014). The information available per location is a mix of waste-in-place, annual capacity, waste year (last year with updated data), and operating status. Between sanitary landfills and dumpsites, there are 662 unique locations.

OpenStreetMap: The OpenStreetMap (OSM) data, comprising labels, locations, and areas, were compiled by the Stanford METER group. This dataset contained 40,403 locations tagged as "landuse = landfill" in OSM and included any available metadata in the json label for each site. However, it is important to note that not every location in this dataset was confirmed as an MSW landfill. Due to the dataset's size and the visual similarities between MSW and certain non-MSW waste sites, manual verification was impractical. For details on the deduplication and filtering process applied to this data, refer to section 2.1.2 (https://www.openstreetmap.org/).

Global Plastic Watch (GPW): This database consists of plastic-containing waste sites identified by a system of neural networks created to analyze spectral, spatial, and temporal components of Sentinel-2 satellite data to find terrestrial waste aggregations (Kruse et al., 2022; https://globalplasticwatch.org/). Following identification, the footprint of each site was calculated and monitored at monthly intervals. As part of the initial effort, this approach has detected nearly 3,000 waste sites in 26 countries (Figure 1). The GPW method generated contours to estimate the areas of detected waste aggregations, updated at a monthly cadence from the Sentinel-2 observation period between 2017 and early-2021. As no specific timestamp was available for each site when the data was downloaded to use for this work, all areas were treated as the most updated values as of early-2022.

Figure 1 (Top) The layout of the GPW website, highlighting countries with locations identified in yellow. (Bottom) An example of a large waste site in Indonesia, showing available site attributes when selected (top left corner of the image).

2.1.2 Supplemental datasets: satellite emissions & the World Bank

Carbon Mapper: Carbon Mapper provided plume detections from their Airborne Visible InfraRed Imaging Spectrometer (AVIRIS) campaigns from 2016-2023 (https://carbonmapper.org/). 971 plume detections were identified and categorized under "Solid Waste". These plumes were used for the training dataset. Plume uncertainties and wind metadata were also available for each observation, but these were not used in the Climate TRACE modeling process.

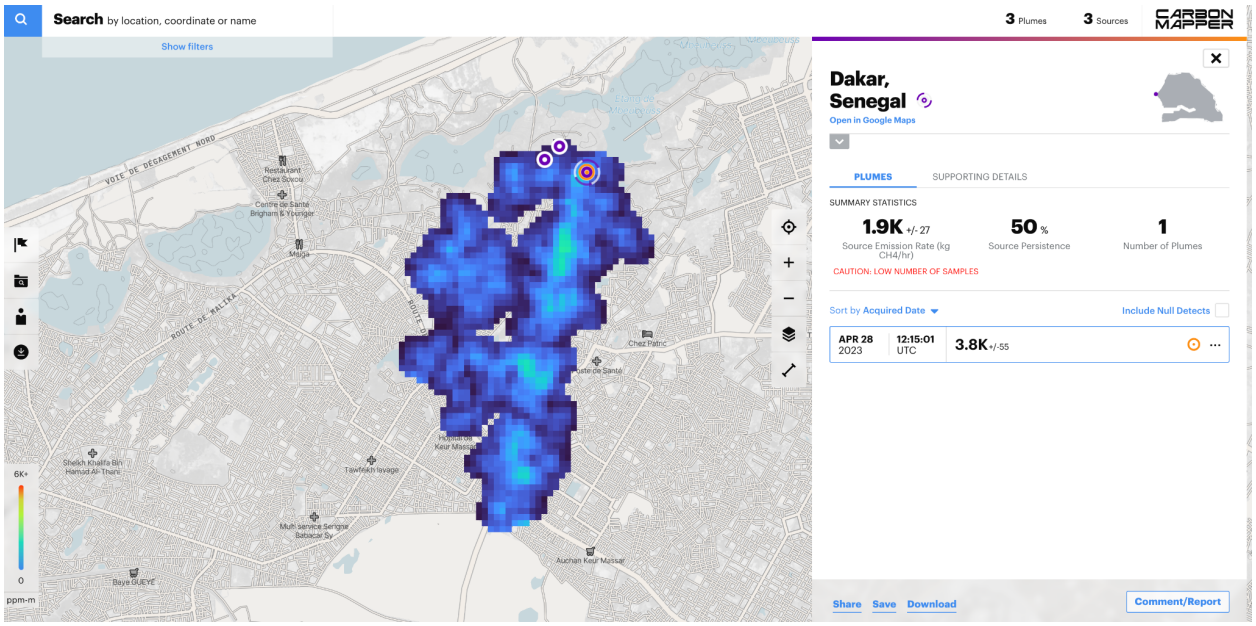

Figure 2 A methane plume example from a waste site (landfill) in Senegal from the Carbon Mapper website.

GHGSat: 3,281 satellite plume detections were available, but these were not categorized by source type or sector. Solid waste plumes were only distinguished by spatial-matching on known waste sites when constructing the training dataset, requiring a small buffer of 500 meters to reduce false-positive risks (https://www.ghgsat.com/).

World Bank: The World Bank's "What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050" report highlights regional and global solid waste generation, management, and impact trends. Regional annual waste generation growth rates were derived from the report and used to project capacities to 2021 and 2022 to construct the pre-modeling and final inventories (see Section 2.1.4). For the capacity modeling stage, the following income groups were assigned to each site by country, per the report's categories: low-income, lower-middle income, upper-middle income, and high-income. For the emissions modeling stage, the following region groups were assigned to each site by country, per the report's categories: East Asia & Pacific, South Asia, Europe & Central Asia, Middle East & North Africa, Sub-Saharan Africa, North America, Latin America & the Caribbean (Kaza et al., 2018).

2.1.3 Filtering & deduplication for the pre-modeling inventory

In creating the Climate TRACE inventory, sources reporting direct emissions were prioritized, such as the EPA datasets, Canada GHGRP, and E-PRTR. For datasets not yet reporting 2022 emissions, any site with reported emissions since 2016 was included as a unique site for 2022, using the most recent emissions value. These values were forward-filled to 2022 without adjustments. When reconciling overlapping data from the LMOP and EPA GHGRP datasets, the GHGRP data was preferred since it directly reported emissions.

A lower bound criterion of 10,000 square meters was set on OSM sites to filter out a bulk of potential false-positive sites for which, in theory, visible confirmation of MSW waste would be difficult. The threshold for GPW sites was set lower, at 1,000 square meters, due to the greater reliability of their remote-sensing techniques.

Further refinement on OSM data required first decoding unicode contents in the json labels and translating non-English labels. Sites were removed based on a list of keywords for non-MSW identifiers, such as “construction” or “tailings”. Additionally, sites with label '{"landuse": "landfill", "_osm_type": "way"}' were omitted entirely, as these contained no information on site type.

Spatial deduplication was necessary both within the same source and between sources. The initial composite dataset, concatenated between all the individual contributing datasets, was spatially matched to itself using the Python package GeoPandas. The largest overlapping site from the same source was retained based on a 1,500-meter radius. For datasets with no reported areas, OSM areas from matching locations were used, a necessary step for the training dataset (see Section 2.1.4).

Finally, OSM, GPW, and Waste Atlas sites overlapping with locations from more informative sources like EPA GHGRP, EPA LMOP, Canada GHGRP, E-PRTR, and INEGI were dropped. For remaining unique Waste Atlas sites, overlaps with OSM and GPW were removed, as Waste Atlas was the next most informative source. Lastly, any remaining OSM sites overlapping with GPW were also excluded. Only unique OSM sites, without any spatial overlaps with other datasets, were included in the final Climate TRACE inventory.

2.1.4 Scaling pre-modeling inventory to 2021 & 2022

Climate TRACE’s goal is to provide emissions for years 2015 to 2022. However, the years 2021 and 2022 had to be estimated for specific waste locations. The following scenarios below highlight each case where annual waste capacity was adjusted up to 2021 and 2022.The annual waste generation growth rate “*r*” as used below was derived from the World Bank, noted in Section 2.1.2.

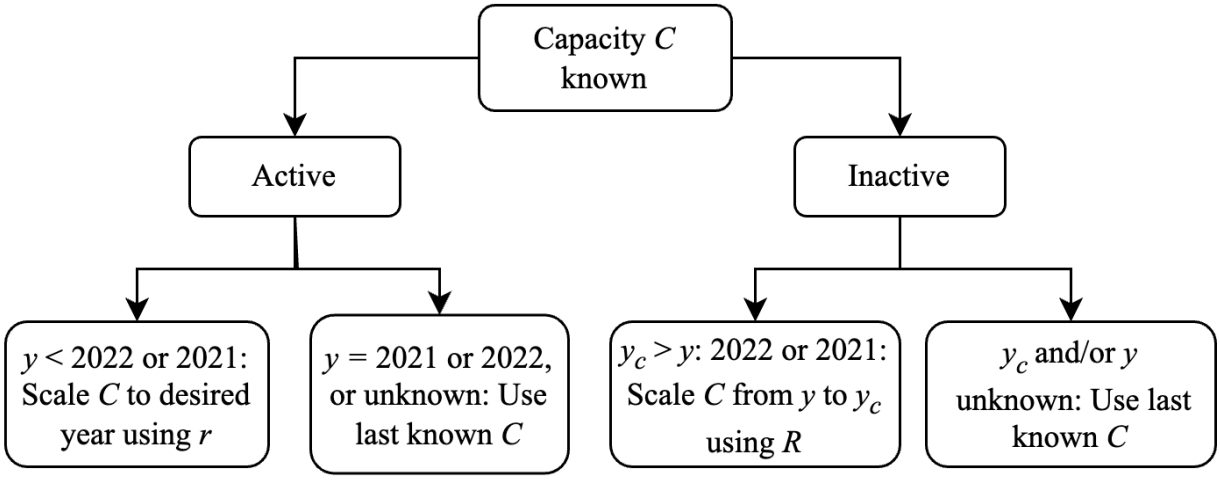

Figure 3 Scenario 1 flowchart to estimate updated capacities for 2021 and 2022, where initial capacities are known.

Scenario 1: Where annual capacity and last reported year were both known.

- If the status was considered active (either explicitly active, labeled under construction, unknown), then either:
    - If the last reported year *"y"* was before 2022, the initial capacity "*C*" was projected yearly up to 2021 and 2022 using growth rate "*r*".
    - If the last reported year *"y"* was 2022 or *"y"* was unknown, then the last reported capacity was retained as the final value.
- Where the status was known as inactive, then either:
    - If the year closed "y_c" was after the known year last reported "*y*", then the initial capacity "*C"* was projected forward up to "y_c" using growth rate "*r*".
    - If "y_c" and "*y*" were unknown, then the last known "*C*" was used.

Scenario 2: Where total waste-in-place and waste-in-place year were both known but capacity was unknown.

- The capacity/waste-in-place ratios were calculated for any sites where both were known.
- Regional average capacity/waste-in-place ratios by World Bank region id were calculated, as this was the most granular spatial resolution possible given data availability.
- Finally, the estimated waste-in-place values were multiplied by these regional ratios to estimate the unknown capacities.

Scenario 3: This scenario addressed waste sites with directly reported emissions but lacking complete data for either 2022 or 2021. In such cases, where 2022 emissions data were unavailable, or only 2022 data was present without 2021 figures, a straightforward method was

applied: no scaling or augmentation of values was performed. Instead, emissions were simply forward or back-filled as needed.

2.1.5 Constructing the training dataset and model development

2.1.5.1 Training dataset

As noted in Section 2.1.3, metadata was shared across overlapping waste sites between different datasets for model training. This integration primarily entailed using area data from GPW, OSM, and/or Waste Atlas for locations that matched with sites from the E-PRTR or any of the GHGRP sources, which lacked area or capacity data. Consequently, a single waste site in our final training dataset could encompass multi-source data including area, capacity, and emissions.

GHGSat and Carbon Mapper emissions data, reported as plumes in kg/hr, were converted to implied annual totals in tonnes. Such extrapolation was intended to counterbalance the more abundant self-reported data, while recognizing the need for temporal granularity capturing both seasonal and diurnal variance to accurately reflect annual emissions rates. Plumes were spatially matched to waste sites from the pre-modeling inventory using a tight 500-meter buffer. Since only 448 matches were identified out of 13,894 locations in the pre-training dataset, satellite emissions data was prioritized over self-reported emissions in these instances.

2.1.5.2 Model development

Two models were developed to predict waste capacity and emissions based on a subset of available data:

1) *Predict capacity-to-area* model: this used the subset of the composite data with known or spatially-matched areas and capacities, and emissions were not required.
2) *Predict capacity-to-emissions* model: This model utilized data with known or spatially-matched capacities and emissions, with no areas required.

Scaled capacities generated by the methods outlined in Section 2.1.4 were used for model-training.

For the capacity-to-emissions model, emissions were averaged by waste site as some landfills had multiple flyover observations from GHGSat and Carbon Mapper. Since single readings were already extended as annual averages, treating multiple such observations as independently informative data points could have introduced unnecessary variability and potential skewness into the dataset. Averaging these observations for each site provided a more stable measure of emissions, attempting to capture the general emission trends of the waste site over time, rather than risking influence from short-term fluctuations whose underlying causal mechanisms were not discernible from the dataset. This limitation stemmed from the insufficient temporal

granularity in available data, coupled with the absence of causal models to explain and predict variability in landfill emissions that could operate within the scope of our dataset.

2.2 Emissions Model

Solid waste capacities and CH_4 emissions quantities were estimated by iterating through the following stages:

1. Causal Models: The first step involved positing plausible causal models, which required a clear understanding of the relationships between variables present in the data and how these relationships could generate emissions. This step was guided by a combination of literature on landfill features and dynamics (such as the IPCC) and plausible reasoning where research proposing deterministic models using these variables did not exist.
2. Statistical Models: Statistical models were defined to capture the predictor-outcome relationships suggested by the causal models in the context of the data. In particular, we used Bayesian regression models, given their capacity to incorporate prior knowledge about predictive relationships of interest and handle uncertainties.
3. Cross-Validation & Model Selection: Between steps 1 and 2, causal models and consequent statistical model definitions were refined continuously. Statistical models were selected by counterbalancing data completeness, model interpretability, degree of accuracy in representing real-world complexity, and generalizability based on rigorous cross-validation.

2.2.1 Causal model

Causal models are best illustrated using Directed Acyclic Graphs (DAGs). A DAG is a way of describing qualitative causal relationships among variables that is, while not not detailed as a full model description, it does contain information that a full statistical model does not. (McElreath, 2018). Figure 4 highlights the DAG casual relationship approach.

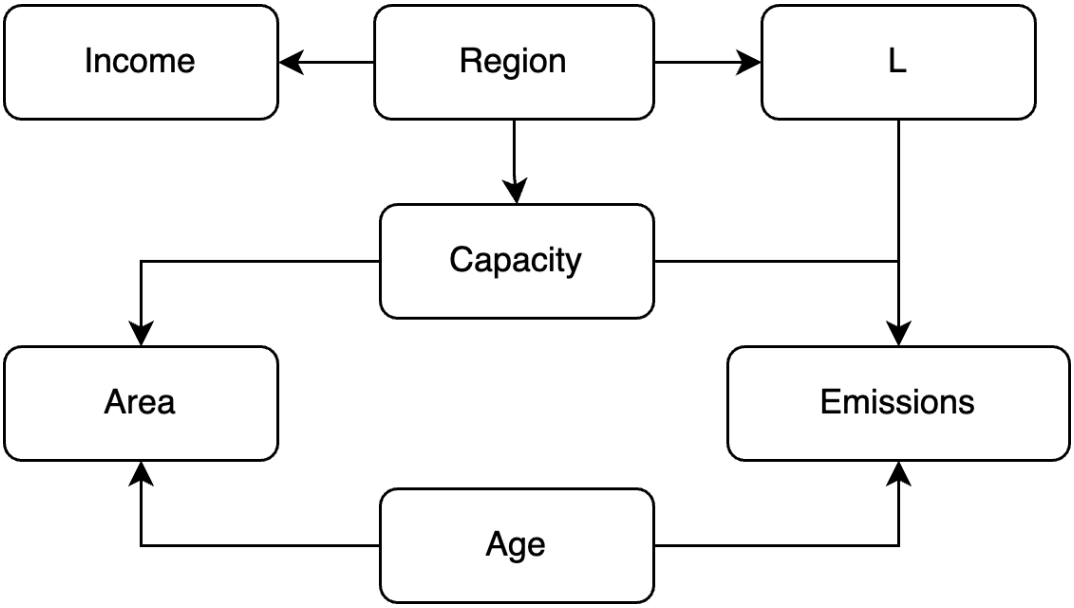

Figure 4 DAG showing high-level causal relationships between variables available in the data.

The causal model and its representative DAG for solid waste were designed by both incorporating hypotheses derived from the available data and integrating the first-order decay equation for solid waste emissions from the 2019-revised guidelines of the IPCC (Eq. 1):

$$CH_4 = L \cdot \sum_{i=1}^{y_f} C_i \cdot e^{-k \cdot t_i}$$

This is adapted notation compared to the IPCC version, where C_i = annual capacity and k = the decay constant. L is as an adapted version of the methane generation potential, defined as follows (Eq. 2):

$$L = A \cdot L_0 = A \cdot MCF \cdot DOC \cdot DOC_f \cdot F \cdot \frac{16}{12}$$

Eq. 2 absorbs any other proportionalities or factors for waste to emissions, in addition to the explicit factors. *MCF* = methane correction factor, which accounts for degree of aerobic decomposition due to waste layering. *DOC* = degradable organic carbon, or the fraction of carbon available for decomposition. DOC_f = the fraction of degradable organic carbon dissimilated, an additional scale factor accounting for the fraction of carbon that is actually released from decomposition. F = fraction of methane in overall landfill gas, typically taken to be 0.5. y_f = the age of the landfill. The final revised equation ultimately used for the statistical model was as follows (Eq. 3):

$$CH_4 = \beta \cdot C_i$$

where (Eq. 4):

$$\beta = L \cdot \sum_{i=1}^{y_f} e^{-k \cdot t_i} \cdot (1+r)^{t_i}$$

Given the data scale and challenges in completeness, without historical data, a constant annual waste generation growth rate *r* based on World Bank data was assumed. If so, then the initial capacity C_i becomes a constant value, and the emissions can be treated as linear with the methane generation potential and the combination of decay and growth rate factors summed over the lifespan of the landfill. In Eq. 3 and 4, β represents that proportionality constant.

Given that area was the predominantly available proxy for waste activity in the datasets, alongside incomplete information for the particular coefficients in the methane generation potential, the DAG in Figure 4 highlights the high-level variables and relationships based on Eq. 1-4 and additional hypotheses relating them. Note that Fig. 1 does not represent a comprehensive causal diagram of all potential predictors of solid waste emissions, and is instead a constrained representation.

The DAG highlights that the region informs annual capacity for waste sites, but it is a general enough variable that more precise predictors will exist under its umbrella or will be informed by proxy variables. For example, in the DAG, income is a descendant variable of region, considered to contain some but not all information about region-specific effects, although there can be alternative ways to direct this causal flow. In the area-capacity model, however, income was used as a uniquely informative variable for waste generation and landfilling rates based on World Bank research (Kaza et al., 2018). Explicit regional grouping would be considered more informative for emissions, as management practices and ambient climate conditions would not exclusively scale with income but would in principle be better informed geographically.

The capacity in this model is treated as giving rise to the area of the landfill (although this is likely a bi-directional effect) in that at least for dumpsites, annual incoming waste will inform the final size of the waste site. Age must interact with capacity to lead to the landfill's size, consistent with a model that accumulating waste informs the landfill footprint, so age is also a cause of the area.

The "fork" structure between capacity, area, and emissions, in that capacity is treated as a common cause of both area and emissions, indicates that area and emissions are themselves causally independent. Conditioning on capacity, in using it as the predictor, learning area provides no new information for predicting emissions. Consequently, capacity is treated as the priority predictor for emissions directly, in keeping with the IPCC equation. However, note that it is not the only predictor, and the arrows from capacity, age, and methane generation potential *L*

are common causes of emissions implying an interaction between them, mirroring the structure of Eq. 1-4.

Given that capacities are unknown for many waste sites, particularly for OSM and GPW data, area is used as the primary predictor and the proxy variable as it is a "descendant" of capacity and carries some but not all information about capacity. In the statistical models, the first model uses area to predict capacity, and the second model predicts emissions from the "implied capacity". Predicting emissions from capacities that were themselves statistical predictions was motivated by logical continuity, and given that area itself was not a strong predictor of emissions directly, it would not prove to be a worse-performing predictive pipeline than area alone.

2.2.3 Bayesian model

Models were defined, tested, and executed using PyMC3 in Python. PyMC3 is a probabilistic programming package, which uses Markov Chain Monte Carlo (MCMC) sampling to approximate posterior distributions and posterior predictive sampling distributions. In this case, we used the No-U-Turn Sampler (NUTS) algorithm, a self-tuning variant of the Hamiltonian Monte Carlo method (Salvatier et al., 2016).

The first model predicted capacity from area, as motivated in Section 2.2.1. Assuming a constant population or waste generation growth rate contributing to that waste site over the lifespan of the landfill, the area could be modeled as a linear function of accepted waste. Given that land area was the more abundant value over capacity, that function was flipped so that capacity would be predicted from area, in this case using a linear model that initially predicted continuous capacity increase with unit increase in area. Since historical data was unavailable for the majority of landfills, and inconsistent at best if present, the age-capacity interaction for area was not explicitly modeled, with the implication that the model slopes would contain implicit statistical information about such a relationship.

Knowing that waste densities should be contingent on management levels, and that waste management and disposal correlates with incomes of the region or country per the World Bank, waste sites were stratified by country-wise income id groups (low-income, lower-middle income, upper-middle income, high-income) from the World Bank. The ensuing Bayesian model structure was a generalized linear model using a Log-Normal likelihood function for observed capacities. Model 1 was then defined as:

$$C_i = \text{LogNormal}(\log(\mu_i), \sigma_i)$$
$$\text{Capped } \mu_i = \max(\min(\log(\mu_i), \log(1 \times 10^5), 1 \times 10^{-10})$$
$$\mu_i = \beta_{income[i]} \cdot A_i$$
$$\beta_{income[i]} \sim \text{LogNormal}(\log(\beta_\mu), \beta_\sigma)$$
$$\beta_\mu \sim \text{Normal}(0.05, 0.03)$$
$$\beta_\sigma \sim \text{HalfNormal}(0.2)$$
$$\sigma \sim \text{Exponential}(1)$$

The first line is the likelihood function, where C_i is the capacity for the i-th row in the data. The second line is the equation for the mean μ_i of the likelihood function, constructed from the ensuing beta parameters and the area from the i-th row A_i. $\beta_{income[i]}$ represents the distribution from which the income group-wise slopes are drawn, per i-th data point, and β_μ and β_σ are its parameters. Finally, σ is the distribution of the standard deviation parameter for the likelihood function.

The following approach reflects the process for generating Model 1:

1. Hierarchical regression structure: The model allowed for different levels of analysis – in this case, individual landfills within the broader World Bank income groups by country. This captured variability both within and between the income groups, while assessing the relationship between area and capacity for each landfill.
2. Parameters & priors: In the selected hierarchical model, the group-wise parameters defined as β_{income} were drawn from a common Log-Normal distribution, whose parameters β_μ and β_σ were initialized with listed priors. The Log-Normal transformation ensured strictly positive scaling with area and capacity. Priors were selected based on prior predictive simulations and model cross-validation. Note that intercepts were excluded from the model, based on the boundary condition that zero area necessarily entails zero waste.
3. Regression on the mean capacity: The model, at its core, was a regression analysis defining the mean capacity μ of landfills as a deterministic function of the observed areas (predictor variable) and the coefficients β_{income} specific to each income group. The posteriors were generated by sampling from the prior distributions on the parameters and constraining those priors by computing the relative plausibilities of each set of parameter combinations through the structure of the likelihood function. A Log-Normal distribution was chosen for the likelihood, a transformation again designed to ensure strictly positive μ values.
4. Capping: Given some land areas on the order of 1-10 million square meters, and predictions for capacities exceeding reasonable waste generation from local populations, a "saturation effect" was incorporated into the model, capping predictable mean

capacities from the likelihood function at 500,000 tonnes, but setting a lower bound at a vanishingly small value of 10^-10 tonnes.

5. Posterior predictions: Using the posterior distributions on the parameters, as a product of the priors and likelihood incorporating the data, predicted capacities were simulated per landfill based on its area. Selecting an 89% compatibility interval in this case meant identifying the central 89% of the sampling distribution density for that given area, based on the posterior distributions on the parameters. The median of the distribution was used as the point estimate of the landfill's capacity for the waste dataset.

Identifying a suitable statistical model for predicting emissions was trickier than for capacity. Based on the causal model, capacity, age, and methane generation potential were all theoretically necessary predictors. First, as with the capacity model, age could not be accommodated as an explicit predictive variable, so it was omitted as an input to the model. Next, the effective methane generation potential in Eq. 2 represents a large number of degrees of freedom for which data was mostly unavailable. These degrees of freedom include the coefficients in the IPCC equation alongside additional unmeasured influences such as diurnal or seasonal effects on emissions. Instead of imputing regional-level IPCC coefficients, the decision was made to use a simple linear model partially-pooled by region, which achieved a similar degree of grouping variability but allowed learning from the data directly instead of fixing initial proportionalities (Johnson et al., 2022).

A linear relationship was proposed between capacity and emissions as in Eq. 3, with slope parameters grouped by world region to predict regional-methane generation potential effects. This model was considered in lieu of a more complex model that risked overfitting without a sufficient theoretical basis motivating them (see Section 2.2.4). We recognize that regional factors which are explicitly unmeasured and undefined, as implied by Fig. 1, are a common cause of capacity directly and emissions directly. Thus, although capacity and methane generation potential may covary, their interaction was not explicitly modeled. Model 2 was structured as:

$$
\begin{aligned}
E_i &= \text{LogNormal}(\log(\mu_i), \sigma_i) \\
\mu_i &= \beta_{region[i]} \cdot C_i \\
\beta_{region[i]} &\sim \text{LogNormal}(\log(\beta_\mu), \beta_\sigma) \\
\beta_\mu &\sim \text{Normal}(0.005, 0.001) \\
\beta_\sigma &\sim \text{HalfNormal}(0.3) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

As in model 1, the first line is the likelihood function, where E_i is the emissions for the i-th row in the data. The second line defines μ_i for the likelihood function, constructed from the beta

parameters and the capacity of the i-th row C_i. $\beta_{region[i]}$ represents the distribution for regional slopes, per i-th data point, with parameters β_μ and β_σ. Finally, σ is the standard deviation distribution for the likelihood.

Evidently, the model was similar to the area-to-capacity model, in that they were both hierarchical linear regression models with Log-Normal group parameter distributions and likelihood functions. However, note these important differences:

1. Regions were used as groups for hierarchical modeling instead of income.
2. The parameter priors were, of course, unique to this model, tested using prior predictive checks and model cross-validation.
3. Emissions were not explicitly capped, as the emissions predictions did not extend to unreasonable values (such as annual emissions > 1 MT).

3. Result Highlights

Figure 5 shows hotspots of waste sites globally. The locations have been filtered for emissions within the 90th percentile of Climate TRACE estimates. The locations show a high concentration of waste sites in North America and western Asia, due to datasets identified that provide better coverage relative to other regions.

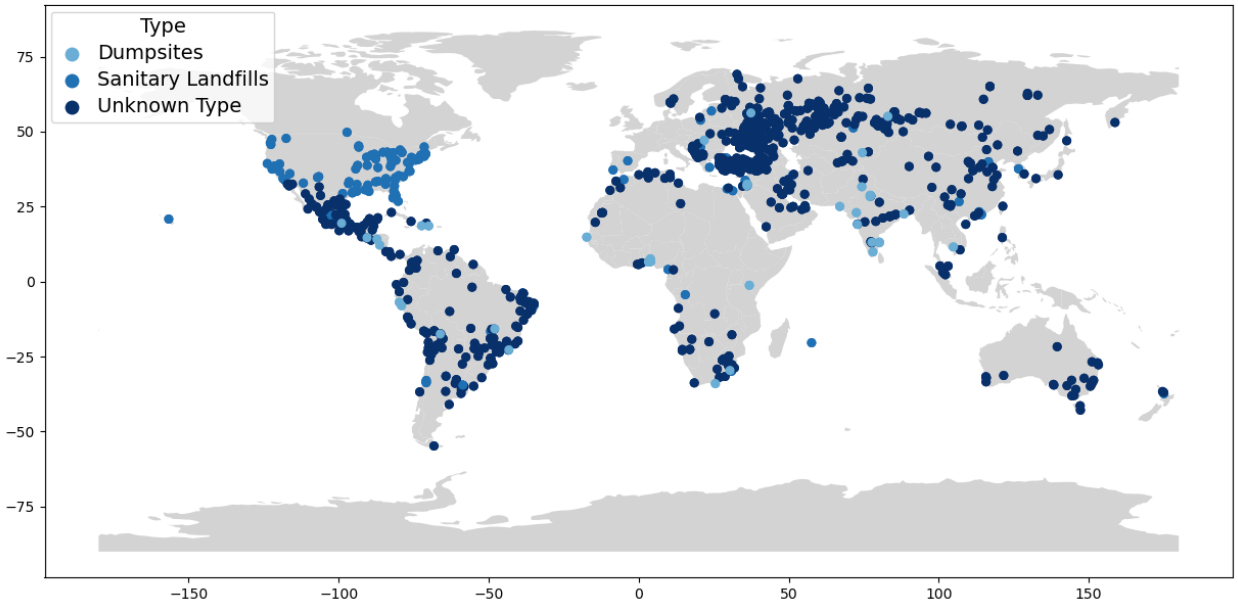

Figure 5 Identified waste sites worldwide whose emission estimates are in the 90th percentile, from dumpsites, sanitary landfills, to unclassified sites.

Figure 6 compares emissions estimates between Climate TRACE's 2022 and 2023 approaches, for LMOP sites which did not publish both gas generated and gas collected. These had site-specific metadata available, which were used in a modified first-order decay equation in the

2022 method, and serve as a comparison mechanism for the deterministic versus statistical methods. Overall, this year's approach tends to overestimate emissions, though with some scatter as attested to by the R^2 = 0.22, indicating that only 22% of the variation in the data can be explained by a linear relationship. The plot highlights greater scatter in the 2023 axis, suggesting that the combination of self-reported and satellite emissions data predicted a more varied range of emissions, as would be expected from a statistical technique.

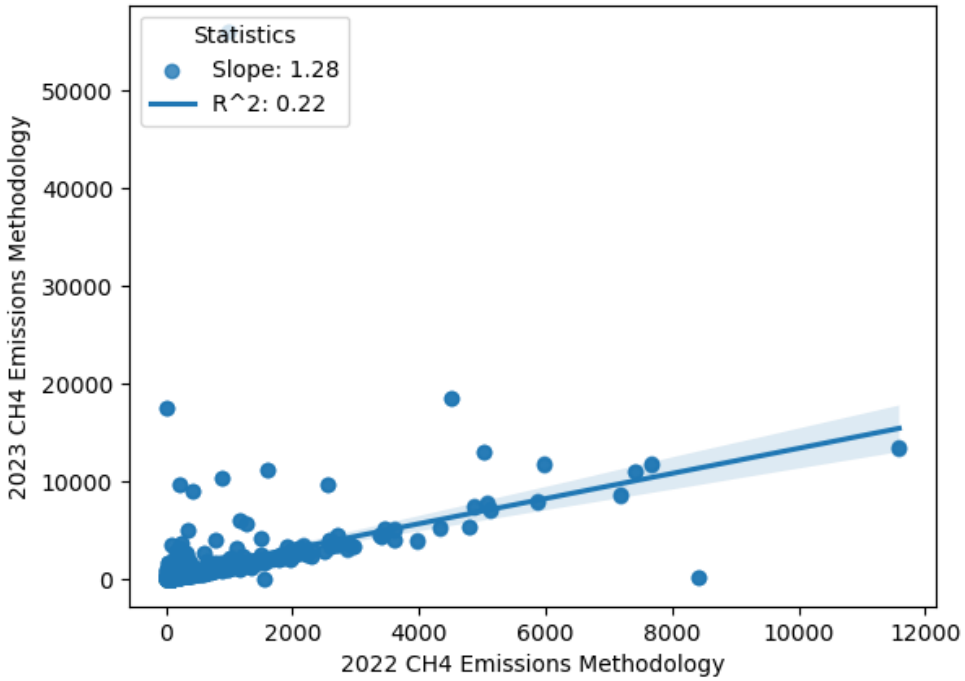

Figure 6 Scatter-plot comparison for LMOP sites between Climate TRACE's 2022 waste emissions estimates vs 2023 emissions estimates.

4. Discussion & Conclusion

The approach utilized by Climate TRACE for the 2023 data release is a critical next step towards establishing a pipeline from constructing a cutting-edge, comprehensive global scale inventory of solid waste sites to predicting interpretable, bounded emissions estimates for them. Existing site-level emissions inventories tend to either be executed for specific countries or only probe values such as waste-in-place and nearby populations, without estimating emissions. Additionally, past projects at generating such inventories can become outdated, without continuous updates. By contrast, the TRACE approach constructs a data-driven, statistical modeling pipeline, informed by a combination of domain knowledge and deterministic models while immediately accommodating the realities and complexities of the data available.

The first extension to dramatically improve the Climate TRACE emissions-accounting from solid waste disposal sites is to construct a higher-confidence, global dataset of at least all

medium to large landfills and dumpsites. This is critical because the 2023 Climate TRACE inventory relied on a combination of data sources with varying degrees of confidence regarding the type of waste site identified, with lower confidence in OpenStreetMap data. An improved inventory would require a higher volume of waste sites identified, targeted validation by location, and final cataloging of verified waste sites.

Improvements in waste activity measurements and detection would best stem from remote-sensing data, which is independent of crowd-sourced or self-reported waste-site activity data. For the TRACE emissions estimates, Global Plastic Watch served as the remote-sensing source to provide previously uncatalogued solid-waste aggregations. A crucial next step is to use remote-sensing technology capable of fully measuring waste activity and identifying salient emissions causing or mitigating mechanisms within landfill boundaries. The best case expansion of solid waste emissions monitoring is the extension of remote sensing data collection of methane plumes, from groups like Carbon Mapper and GHGSat, with extensive temporal granularity and increased global coverage.

Lastly, the Bayesian statistical modeling techniques implemented in Climate TRACE's 2023 data release would need to improve, as they are designed to do, by increasing their complexity while maintaining their interpretability and generalizability. As better data become available, the model would need to synthesize a potential abundance of predictor variables, both previously known and unknown. Additionally, more detailed deterministic models need to be considered, particularly those based on case-studies done that started with the IPCC approach. For example, with greater temporal granularity, seasonal variance and time-evolution would need to be explicitly accounted for. The Bayesian modeling paradigm welcomes any and all such developments, and as such, the models proposed by Climate TRACE can only improve and strike a stronger balance between the causal implications of more informative deterministic models and more robust consequent statistical models.

A final consideration is that waste sites, particularly open dumpsites in emerging and developing nations, can be volatile and unpredictable locations. Many people live near or even within dumpsites - some picking waste for scraps to sell for subsistence, others merely living near them through no will of their own - all impacted by the continuous release of CH_4 (amid other gasses) and uncollected liquid leachate. These large volumes of gas released from unmanaged waste aggregations can be highly flammable, commonly resulting in massive, toxic fires (Gupta, 2022). Lastly, an additional complication is that some of these waste sites play host to violent gang activity, serving as direct threats to locals while potentially stymying efforts at introducing management and gas collection practices (D'Aubuisson, 2022). Altogether, it is imperative to recognize that the hazards of solid waste disposal sites are substantial in both the short and long-term, and that any attention that the scope of Climate TRACE's solid waste emissions

estimates may bring to improving waste-management practices to reduce global greenhouse gas emissions can produce equally substantial public health and safety benefits.

5. Supplementary metadata information

The solid waste emissions dataset reports CH_4, and 20 and 100 year GWP emissions from individual solid waste disposal sites. CO_2 and N_2O are not explicitly modeled or estimated, but are republished from facilities where provided.

Site-level emissions estimates were reported for the years 2021 and 2022. Country-level emissions estimates span 2015-2022, and reflect a combination of the sum of site-level emissions by country and EDGAR data where coverage was lacking. All data is freely available on the Climate TRACE website (https://climatetrace.org/). A detailed description of what is available is described in Table S1 and S2.

Table S1 General dataset information for *Estimating CH_4 Emissions from Solid Waste Disposal Sites*

General Description	Definition
Sector definition	*Individual landfill and solid waste disposal site emissions*
UNFCCC sector equivalent	*4.A Solid Waste Disposal*
Temporal Coverage	*2021-2022 at facility-level, 2015-2022 at country-level*
Temporal Resolution	*Annual*
Data format(s)	*CSV*
Coordinate Reference System	*EPSG:4326, decimal degrees*
Number of site and countries available for download	*10,314 solid waste disposal sites, from 155 countries.*
Total emissions for 2022	*2.056 billion tonnes of 100 year CO2-equivalent*
Ownership	*We used public data and research to identify ownership information*
What emission factors were used?	*Model-based emissions factors, motivated by IPCC's first-order decay method from updated 2019 guidelines*
What is the difference between a "NULL / none / nan" versus "0" data field?	*"0" values are for true non-existent emissions. If we know that the sector has emissions for that specific gas, but the gas was not modeled, this is represented by "NULL/none/nan"*
total_CO2e_100yrGWP and total_CO2e_20yrGWP conversions	Climate TRACE uses IPCC AR6 CO_2e GWPs. CO_2e conversion guidelines are here: https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_FullReport_small.pdf

Table S2 Facility level metadata description confidence and uncertainty for *Estimating CH_4 Emissions from Solid Waste Disposal Sites*

Data attribute	Confidence Definition	Uncertainty Definition
type	• *Low*: Sites labeled dumpsites, from the Global Plastic Watch, as these types are unconfirmed • *High*: Sanitary landfills, if from EPA, Canada, E-PRTR, or Waste Atlas (where type reported) sources	Not used; N/A
capacity_description	• *Low:* OpenStreetMap areas, or where capacities were actively rescaled • *Medium*: Waste Atlas sites, out of date data • *Very High*: Values from EPA sources (self-reported) or Global Plastic Watch (satellite measured)	+/- 10% from reported values
capacity_factor_description	Not used; N/A	Not used; N/A
capacity_factor_units	Not used; N/A	Not used; N/A
activity	• *Very Low:* Modeled values for OpenStreetMap • *Low:* Modeled values for sources other than OpenStreetMap • *Medium*: Self-reported values scaled to 2022 from any source • *High*: Self-reported, unscaled values from Waste Atlas • *Very High*: Self-reported, unscaled values not from Waste Atlas	• +/- 10% from reported values • 89% confidence interval upper and lower bounds where modeled
CO2_emissions_factor	Not used; N/A	Not used; N/A
CH4_emissions_factor	• *Very Low*: Modeled values for OpenStreetMap • *Low:* Modeled values for sources other than OpenStreetMap • *Medium*: LMOP data where gas generated and gas collected both reported • *High*: All direct self-reported emissions data	• +/- 10% from reported values • 89% confidence interval upper and lower bounds where modeled
N2O_emissions_factor	Not used; N/A	Not used; N/A
other_gas_emissions_factor	Not used; N/A	Not used; N/A
CO2_emissions	• *High*: Self-reported emissions values where available	• +/- 10% from reported values
CH4_emissions	Constructed as a weighted combination of activity and emissions factor confidences, quantifying those by scoring from 1-5 as: 1 = very low, 2 = low, 3 = medium, 4 = high, 5 = very high	• +/- 10% from reported values • 89% confidence interval upper and lower bounds where modeled

N2O_emissions	*High*: Self-reported emissions values where available	• +/- 10% from reported values
other_gas_emissions	Not used; N/A	Not used; N/A
total_CO2e_100yrGWP	Same as CH4_emissions	• +/- 10% from reported values • 89% confidence interval upper and lower bounds where modeled
total_CO2e_20yrGWP	Same as CH4_emissions	• +/- 10% from reported values • 89% confidence interval upper and lower bounds where modeled

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: Raniga, K., Davitt, A., Lewis, C., Sridhar, L., Gans, L., McCormick, G. (2023). *Solid Waste Sector: Estimating CH4 Emissions from Solid Waste Disposal Sites.* 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) Ayandele, E., Huffman, K., Jungclaus, M., Tseng, E., Duren, R., Cusworth, D. and Fisher, B., 2022. *Key Strategies for Mitigating Methane Emissions from Municipal Solid Waste*, RMI, https://rmi.org/insight/mitigating-methane-emissions-from-municipal-solid-waste/ (Accessed 01 September 2022).
2) D'Aubuisson, J.J.M. (2022) 'How the MS13 Became Lords of the Trash Dump in Honduras', *Insight Crime*, 19 Jan [online]. Available at: https://insightcrime.org/news/honduras-how-ms13-became-lords-trash-dump/ (Accessed 01 September 2022).
3) Gupta, A. (2022) 'Delhi landfill burning', *The Times of India*, 25 May [online]. Available at: https://timesofindia.indiatimes.com/blogs/voices/delhi-landfill-burning/ (Accessed 01 September 2022).
4) IPCC (Intergovernmental Panel on Climate Change), 2006. Chapter 3. *2006 IPCC Guidelines for National Greenhouse Gas Inventories, Volume 5, Waste*. Published: IGES, Japan.
5) IPCC (Intergovernmental Panel on Climate Change), 2019. Chapter 5. *2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories*. Published: IPCC, Switzerland.
6) Johnson, Alicia A., Miles Q. Ott, and Mine Dogucu. *Bayes' Rules! An Introduction to Applied Bayesian Modeling*. 1st ed., Chapman and Hall/CRC, 2022.
7) Kaza, S., Yao, L., Bhada-Tata, P. and Van Woerden, F., 2018. *What a waste 2.0: a global snapshot of solid waste management to 2050*. World Bank Publications.
8) Kruse, C., Boyda, E., Chen, S., Karra, K., Bou-Nahra, T., Hammer, D., Mathis, J., Maddalene, T., Jambeck, J. and Laurier, F., 2022. Satellite Monitoring of Terrestrial Plastic Waste. *arXiv preprint arXiv:2204.01485*.
9) Mavropoulos, A., Mavropoulou, N., Anthouli, A. and Tsakona, M., 2014. *Waste Atlas 2014 Report - The World's 50 Biggest Dumpsites.* 10.13140/RG.2.2.33122.45763.
10) McElreath, R., 2018. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.
11) United Nations Climate Change *Parties & Observers*. Available at: https://unfccc.int/parties-observers (Accessed 01 September 2022).
12) United States Environmental Protection Agency (2022) *LMOP Landfill and Project Database*. Available at: https://www.epa.gov/lmop/lmop-landfill-and-project-database (Accessed 01 July 2023).

13) United States Environmental Protection Agency (2022) *Facility Level Information on GreenHouse gases Tool (FLIGHT)*. Available at: https://ghgdata.epa.gov (Accessed Accessed 01 July 2023).
14) EPA (United States Environmental Protection Agency) (2022) *Basic Information about Landfill Gas* (2022). Available at: https://www.epa.gov/lmop/basic-information-about-landfill-gas (Accessed 01 July 2023).
15) EPA (United States Environmental Protection Agency) (2022) *Using GHG Inventory and GHGRP Data* (2022). Available at: https://cfpub.epa.gov/ghgdata/inventoryexplorer/data_explorer_flight.html (Accessed 01 July 2023).
16) Salvatier J, Wiecki TV, Fonnesbeck C. 2016. "Probabilistic programming in Python using PyMC". *PeerJ Computer Science* 2:e55 https://doi.org/10.7717/peerj-cs.55.