Bayesian Methods for Enhanced Greenhouse Gas Emissions Inventories

Michael Pekala[1,3] **Zoheyr Doctor**[2,3] **Elizabeth Reilly**[1,3] **Gavin McCormick**[2,3]
Gary Collins[1,3] **Daniel Moore**[2,3] **Amy Piscopo**[2,3]
Krsna Raniga[2,3] **Michael Robinette**[1,3] **Ting So**[2,3]
[1]Johns Hopkins Applied Physics Laboratory [2]WattTime [3]Climate TRACE
{*firstname.lastname*}@jhuapl.edu
{zoheyr.doctor,gavin,dan.moore,amy.piscopo,krsna,ting.so}@wattime.org

Abstract

Developing effective mitigation strategies for greenhouse gas reduction hinges on accurate emissions and metadata tracking to identify the most impactful reduction opportunities. Given that emissions cannot be perfectly and ubiquitously observed, constructing inventories entails fusing data from multiple sources that are of varying levels of fidelity, quality, and completeness. This proposal suggests that Bayesian models, powered by modern probabilistic programming frameworks, can integrate multiple data sources data into posterior emissions estimates while also accounting for incompleteness and leveraging data from less granular spatiotemporal scales. A preliminary analysis combining country-level steel production data and facility-level activity data shows promise for estimating emissions reduction potential when there is a population of facilities that have not been directly observed.

1 Emissions inventories: potential impact and open challenges

There is an urgent need to rapidly and effectively reduce anthropogenic greenhouse gas (GHG) emissions to blunt the impact of global temperature increases [1]. Accelerating progress in climate change mitigation requires the ability to rapidly identify and prioritize high-impact, under-resourced strategies for reducing atmospheric greenhouse gases. Central to this effort are emissions inventories which quantify sources and sinks of greenhouse gases across regions, sectors, and time scales. Modern inventories have advanced considerably, with some spanning multiple chemical species and integrating observations from diverse spatial and temporal resolutions, e.g., [2, 3, 4, 5]. Despite these improvements, there remains uncertainty as well as potential data gaps, particularly at finer geographic scales or in lower resource regions, where estimates could be improved through the inclusion of more specialized data sets and relevant prior knowledge. An open challenge is to rigorously integrate this diverse information while maintaining defensible uncertainty quantification.

Bayesian methods offer a principled framework to help meet these challenges. By enabling the integration of heterogeneous emissions and metadata across multiple scales, Bayesian reasoning can estimate and correct for data set-specific biases, quantify uncertainties and degeneracy in posterior estimates, and infer missing data through probabilistic imputation. This could not only improve the completeness and accuracy of inventories, but also inform where additional sensing resources might be most beneficial. Our proposal is to explore the application of these methods to improve both modern emissions inventories and climate mitigation planning.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

2 Data and methods

We conducted an initial case study grounded in data from the Climate TRACE (Climate TRACE) [5] emissions inventory, a data set with worldwide scope including over 600 million emissions sources spanning multiple sectors. Many Climate TRACE sector teams leverage machine learning (ML) to estimate activity based on signatures within remotely sensed data; however, not all facilities generate sufficiently strong signals. Thus, facility-level data may be non-uniformly noisy, or potentially missing altogether. In the event that a facility is missing, selection bias is likely since observability and emissions are interrelated. Such selection effects manifest in other scientific domains where inference of the underlying population, not just the directly observable population, is of value e.g., [6, 7, 8].

Climate TRACE models bottom-up emissions via the coupled equations in eq. (1)

$$A_i(t) = C_i(t) \cdot CF_i(t), \qquad E_{i,g}(t) = A_i(t) \cdot EF_{i,g,}(t), \tag{1}$$

where t denotes time (typically discrete; months or years), i facility, g gas species, A facility activity level, C capacity of the facility, CF capacity factor (proportion of capacity utilized), EF an emissions factor which converts activity into flux, and E emissions. For our study, we focus solely on activity; however, models that incorporate all the emissions terms are of ultimate interest.

Here we consider a single sector (iron and steel manufacturing), a single country (Brazil), and propose a model for analyzing monthly emissions data. The steel sector manifests interesting challenges – there are competitive economic incentives for withholding facility-level production levels and, in many cases, production information may only be available at the country level [9]. Our study uses data from January of 2022, consisting of 22 known facilities for Brazil which we partition by the facility-type used in manufacturing crude steel. The two relevant types in Brazil currently tracked by Climate TRACE are blast-furnace-to-basic-oxygenation-furnace (BF-BOF) facilities and electric-arc furnace (EAF) facilities. The former are a primary pathway for steel production and rely on iron ore, whereas the latter rely on scrap steel and are less energy intensive [10]. Figure 1 shows a prototype Bayesian hierarchical model where the values for μ_{BF}, σ_{BF}, μ_{EAF}, and σ_{EAF} are determined by conditioning on Climate TRACE data. We make an initial assumption that the missing facilities will be selected at random from the BF-BOF population and introduce shared hyperparameters to model this population. Our initial approach assumes the modeler is willing to hypothesize the number of missing facilities n a priori; the idea is that one can sweep over hypotheses related to the number of missing facilities and compare the posterior estimates to conclude which scenarios are most plausible. Other approaches might be considered in a full study.

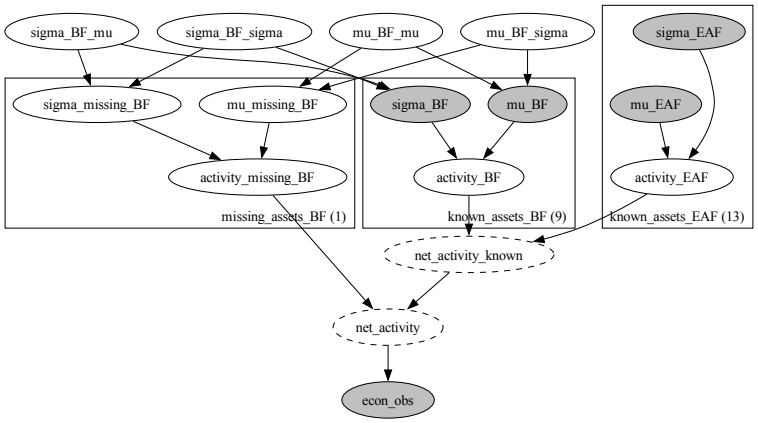

Figure 1: Plate diagram for scenario with $n = 1$ unobserved facilities. Shaded nodes indicate observations, dashed nodes indicate deterministic variables, and rectangles indicate repeated variables that are conditionally independent. The "net" activity nodes indicate aggregation of data from facility- to country-level. Full definition of the conditional distributions is omitted due to space constraints.

To further inform parameters related to missing facilities, we incorporate an exogenous country-level estimate of steel manufacturing activity taken from [11] (denoted `econ_obs` in fig. 1). This observed value is approximately 17% higher than the sum of facility-level activities currently represented in Climate TRACE suggesting one or more possible missing facilities[1]. We implemented this model using the NumPyro probabilistic programming framework [12, 13].

3 Preliminary results

Inference for this model produced $\hat{r}$ (Gelman-Rubin statistic) values generally close to 1, indicating reasonable convergence [14]. Figure 2 presents a subset of the posterior estimates. The left panel shows that the model can successfully refine loose priors on the hyperparameters associated with the population of BF assets; however, there is still substantial uncertainty that might be improved with more refined modeling. In the right panel, we observe that, as the number of presumed missing assets increases, the estimated net activity approaches the observed value, with the missing assets explaining an increasing share of the activity gap. While we would not want to draw specific conclusions from this preliminary model, it does suggest viability of the overall approach.

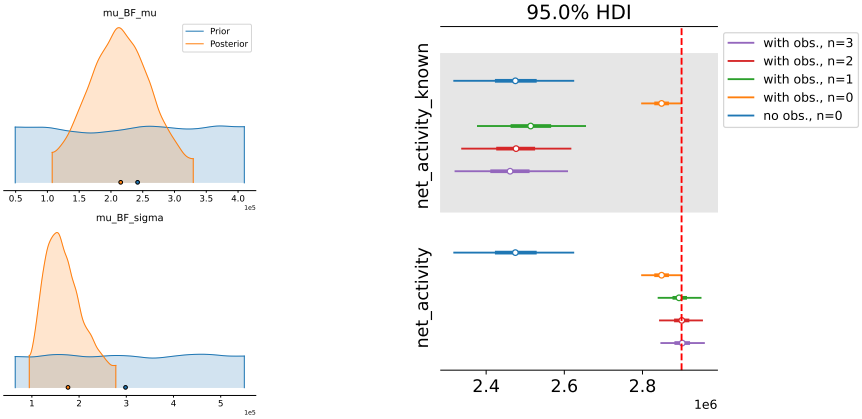

Figure 2: Posterior estimates for select variables from fig. 1. The left panel shows estimates for variables associated with BF population location. There is still fairly high uncertainty, likely due to the wide range of activity values for BF assets. The right subplot shows the 95% highest density interval (HDI) for two country-level aggregate activity variables, as a function of the number n of presumed missing BF assets. The red vertical dashed line denotes the economic activity observation.

4 Conclusions and proposed future work

Our study suggests that Bayesian methods are applicable to representative climate inventory challenges. A compete work would broaden the model's scope (e.g., spatial, temporal), relax some of the strong initial assumptions, conduct rigorous model comparison (e.g., [15]), and compare other approaches for dealing with missing assets (e.g., [16]). We are also enthusiastic about identifying additional auxiliary data sets that might be incorporated to improve these estimates.

Acknowledgments and Disclosure of Funding

The authors would like to thank the Climate TRACE coalition for its organizational support, and Climate TRACE's funders for their financial support. For a full list of funders, please visit the Climate TRACE website: https://climatetrace.org/team.

[1]There are other possible causes for this discrepancy, including noise and the fact that the economic data is steel-specific and omits iron. For our initial exploration, we presume the unique explanation is missing facilities.

References

[1] IPCC. Special report on the impacts of global warming of 1.5 C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change. *Sustainable Development, and Efforts to Eradicate Poverty*, 32, 2018.

[2] M Crippa, D Guizzardi, F Pagani, M Banja, M Muntean, E Schaaf, F Monforti-Ferrario, WE Becker, et al. GHG emissions of all world countries. Publications Office of the European Union, Luxembourg (2023), 2024.

[3] R. M. Hoesly, S. J. Smith, L. Feng, Z. Klimont, G. Janssens-Maenhout, T. Pitkanen, J. J. Seibert, L. Vu, R. J. Andres, R. M. Bolt, T. C. Bond, L. Dawidowski, N. Kholod, J.-I. Kurokawa, M. Li, L. Liu, Z. Lu, M. C. P. Moura, P. R. O'Rourke, and Q. Zhang. Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the community emissions data system (CEDS). *Geoscientific Model Development*, 11(1):369–408, 2018.

[4] T. Oda and S. Maksyutov. A very high-resolution (1 km×1 km) global fossil fuel CO_2 emission inventory derived using a point source database and satellite observations of nighttime lights. *Atmospheric Chemistry and Physics*, 11(2):543–556, 2011.

[5] Climate TRACE database. https://climatetrace.org. Accessed: 2025-08-14.

[6] Daniel Foreman-Mackey, David W Hogg, and Timothy D Morton. Exoplanet population inference and the abundance of earth analogs from noisy, incomplete catalogs. *The Astrophysical Journal*, 795(1):64, 2014.

[7] Thomas J Loredo. Accounting for source uncertainties in analyses of astronomical survey data. In *AIP Conference Proceedings*, volume 735, pages 195–206. American Institute of Physics, 2004.

[8] Ilya Mandel, Will M Farr, and Jonathan R Gair. Extracting distribution parameters from multiple uncertain observations with selection biases. *Monthly Notices of the Royal Astronomical Society*, 486(1):1086–1093, 2019.

[9] Verity Crane and George Ebri. Manufacturing and industrial processes sector: Iron & steel manufacturing emissions. https://github.com/climatetracecoalition/methodology-documents/tree/main/2025, 2025.

[10] IEA. Iron and steel technology roadmap. https://www.iea.org/reports/iron-and-steel-technology-roadmap, 2020.

[11] World Steel Association. World steel in figures. https://worldsteel.org/data/world-steel-in-figures/world-steel-in-figures-2024/, 2024.

[12] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554*, 2019.

[13] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20:28:1–28:6, 2019.

[14] Osvaldo Martin. *Bayesian analysis with python*, volume 48. Packt Publishing Birmingham, UK, 2016.

[15] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27(5):1413–1432, 2017.

[16] Craig K Enders. *Applied missing data analysis*. Guilford Publications, 2022.