

Supporting Information: Methane Emissions from Natural Gas Gathering Pipelines in the Permian Basin

Jevan Yu^{1,2}, Benjamin Hmiel², David R. Lyon², Jack Warren², Daniel H. Cusworth^{3,4}, Riley M. Duren^{3,4}, Yuanlei Chen¹, Erin C. Murphy², and Adam R. Brandt¹

¹Stanford University, Stanford, CA 94305, USA

²Environmental Defense Fund, Austin, TX 78701, USA

³Arizona Institutes for Resilience, University of Arizona, Tucson, AZ 85721, USA

⁴Carbon Mapper, Pasadena, CA 91105, USA

August 24, 2022

Contents

S1 Treatment of "Unknown" Pipeline Emission Sources	S2
S2 Linear Distance of Gathering Pipelines	S3
S3 Minimum Threshold of Overflights	S4
S3.1 Notation	S4
S3.2 Overflight Analysis	S4
S4 Uncertainty Quantification	S5
S4.1 Linear Distance of Gathering Pipelines	S5
S4.2 Aggregate Pipeline Emissions	S6
S4.3 Emission Factor	S7
S5 Monte Carlo Simulation Experiment	S8
S5.1 Experiment	S8
S5.2 Accounting for Emissions Below the Minimum Detection Limit	S8
S6 Figures and Tables	S11
S7 References	S15

S1 Treatment of "Unknown" Pipeline Emission Sources

In the case of some emission sources, we could not definitively determine whether they were from gathering pipelines or transmission pipelines. Although a conservative estimate for a gathering pipeline emission factor would include none of these ambiguous sources (which is the method we follow for our main estimates), it is likely that some of these sources are in fact from gathering pipelines.

To infer whether one of these "unknown" pipeline emission sources is from a gathering or transmission pipeline, we use the activity factors (defined as sources per unit distance of pipeline) for gathering and transmission lines. These activity factors differ from campaign to campaign, and are summarized in Table S1.

If we had no way of discerning whether a particular emission source was from a gathering or transmission line, we might estimate that with probability 0.5 it is gathering and with probability 0.5 it is transmission. However, we can use the activity factors to make a more informed estimate. For a particular campaign, let α_g and α_t be the activity factors for gathering and transmission lines. Then for an "unknown" pipeline source, we say that with probability $\frac{\alpha_g}{\alpha_g + \alpha_t}$ it is gathering and with $\frac{\alpha_t}{\alpha_g + \alpha_t}$ it is transmission.

To build an interval of uncertainty, we run a simulation. For each campaign, we collect all "unknown" pipeline emission sources. With probability $\frac{\alpha_g}{\alpha_g + \alpha_t}$, we (independently) select each "unknown" pipeline emission source as gathering. For those that we select as gathering, we add their emission rates together. We run this process 10,000 times and select the 2.5th and 97.5th percentiles as our interval. For the four campaigns, ordered chronologically, these intervals are 0.1-0.3, 0.1-0.2, 0.1-0.3, and 0.4-0.4 Mg y⁻¹ km⁻¹.

S2 Linear Distance of Gathering Pipelines

Our method for estimating an emission factor for natural gas gathering pipelines involves dividing aggregate gathering pipeline emissions in a study region by the linear distance of gathering pipelines in the study region. To calculate the latter value, we acquired GIS shapefiles of pipeline location and metadata from Enverus Drillinginfo and Enverus Prism in April 2022. Using ArcGIS Pro, we crop the pipeline layer to the particular study region (either “Full” or “ $n_o \geq 3$ ”; recall that “ $n_o \geq 3, n_d > 1$ ” is the same region as “ $n_o \geq 3$,” just with different emissions values). We then calculate the geodesic length of the resulting geometries, export the table, and sum the values of the “operational” (Drillinginfo) or “in service” (Prism) natural gas gathering pipeline (or pipeline fragment) lengths. (Compare this to EPA’s definition of gathering pipelines: “Onshore petroleum and natural gas gathering and boosting means gathering pipelines and other equipment used to collect petroleum and/or natural gas from onshore production gas or oil wells and used to compress, dehydrate, sweeten, or transport the petroleum and/or natural gas to a natural gas processing facility, a natural gas transmission pipeline or to a natural gas distribution pipeline.” Note that the EPA definition includes multi-phase pipelines transporting co-mingled oil and gas production from wellheads to tank batteries, but the Enverus definition does not seem to. In this paper, it is reasonable to exclude these co-mingled product pipelines due to their relative rarity.) The relevant uncertainty quantification is discussed in Section S4.1.

S3 Minimum Threshold of Overflights

S3.1 Notation

As a convention, we define an emission source as a collection of plumes located within a 150 m radius, the approximate length of a well pad. The persistence-adjusted emission rate from each source i is calculated as $q_i = \frac{1}{n_o} \sum_{j=1}^{n_d} q_j$, where n_o is the number of overflights of the source (defined as the number of days the source was flown over), n_d is the number of days non-zero emissions were detected from the source (so necessarily $n_o \geq n_d$), and q_j is the emission rate associated with the j th overflight of the source (for $j \in \{1, \dots, n_o\}$). When there are two or more observations collected on the same day, q_j is the mean of all the detections on that day. The quantity n_d/n_o is defined as the persistence of the source.

S3.2 Overflight Analysis

To estimate aggregate emissions from gathering pipelines, we use all measured sources. Doing so produces an unbiased estimate of aggregate measurable emissions: if q_i is the actual emission rate from asset i and \hat{q}_i is our estimate of its emission rate, we know $\mathbb{E}[\hat{q}_i] = q_i$, and hence $\mathbb{E}[\sum_i \hat{q}_i] = \sum_i \mathbb{E}[\hat{q}_i] = \sum_i q_i$, as desired. Despite this, we wish to further explore the dataset to maintain that the emissions observations are consistent in regions of high and low coverage.

For our analyses of individual sources and their properties, we choose to not use all sources. For values with low n_o , we find a positive persistence bias. Thus, we select a threshold of $n_o \geq 3$, which allows for reasonable convergence in persistence while also preserving sufficient data. See Figure S2.

Cusworth et al. 2021 also used the threshold of $n_o \geq 3$, justifying it as the minimum number of overflights to be able to confidently capture the persistence of a source [1]. That is, a 25% persistent source would be detected more than half the time, or $1 - (1 - 0.25)^3 > 0.5$.

S4 Uncertainty Quantification

S4.1 Linear Distance of Gathering Pipelines

In calculating the linear distance of gathering pipelines in the study region, there are two potential sources of uncertainty that we address.

First, there is potential uncertainty in the quality and representativeness of the infrastructure data. Indeed, this is the case: Drillinginfo and Prism are separate commercial databases of oilfield metadata provided by a single commercial aggregator, Enverus; however, further analysis reveals that the two datasets (at least in the Permian Basin) are not homogenous with respect to pipeline locations and distances. For uncertainty bounds on pipeline distance, we adopt the following method. For a particular study region, let d_{DI} and d_P be the linear distance of natural gas gathering pipelines calculated from Drillinginfo and Prism data. Then our estimate with uncertainty bounds (we take this as 1 standard deviation in each direction) is given by

$$\hat{d} = \frac{d_{DI} + d_P}{2} \pm \frac{|d_{DI} - d_P|}{2}.$$

That is, the lower bound is $\min\{d_{DI}, d_P\}$ and the upper bound is $\max\{d_{DI}, d_P\}$.

Second, there is a potential temporal change in gathering pipeline distance. The data suggest that this change is negligible: Drillinginfo data acquired in April 2022 contain 120,012 km of operational natural gas gathering pipelines in the Permian Basin, and Drillinginfo data acquired in July 2021 contain 119,512 km of the same category; Prism data shows an even higher level of stability between February 2021 and April 2022. The temporal difference over 9 months represents a <1% difference in distance, which is much smaller than the difference in distance observed between the Drillinginfo and Prism products. As such, we disregard temporal uncertainty in our estimates of pipeline distance. We use data acquired in April 2022 for temporal consistency across platforms.

For the sake of argument, though, assume that the actual temporal change were more significant

than the data indicated. Note that the emission factor may be (point) estimated as

$$EF = \frac{\sum_i \hat{q}_i}{\frac{d_{DI} + d_P}{2}},$$

where i indexes the pipeline emission sources in the survey region. A reasonable assumption is that the linear distance of gathering pipelines in the Permian is a generally increasing function of time due to increasing development and gas production. Thus, if $\frac{d_{DI} + d_P}{2}$ were an overestimate, then the resulting emission factor would be an underestimate. This means that if data from April 2022 were not representative of the conditions during those campaigns, using such data would most likely render conservative the emission factor estimates for the four campaigns.

S4.2 Aggregate Pipeline Emissions

For each emission source i , there is an associated σ_i value that represents one standard deviation of uncertainty (assuming a Gaussian distribution) for the estimated emission rate. This value factors in uncertainty in both integrated methane enhancement (IME) and wind speed [1]. To be conservative, we do not assume independence of errors, given the potential for correlated errors for nearby plumes. Hence, we do not add the errors in quadrature, which would be the canonical method for quantifying uncertainty of a sum of independent Gaussian random variables. Rather, we sum the individual standard deviations and use that figure as a wide measure of uncertainty for the aggregate estimate. Mathematically, this is represented as

$$\sigma_{tot} = \sum_i \sigma_i,$$

so our estimate of aggregate emissions, with uncertainty bounds, is given by

$$\hat{q}_{tot} = \sum_i \hat{q}_i \pm \sigma_{tot}.$$

S4.3 Emission Factor

The emission factor calculation essentially consists of two pieces of input data: (1) linear distance of gathering pipelines and (2) aggregate emissions from gathering pipelines. Thus, to quantify uncertainty in our emission factor estimates, we preserve the uncertainties from each of these components by joining the methods in Sections S4.1 and S4.2.

We use simulation to calculate a 95% confidence interval for each emission factor. We assume that both the linear distance of gathering pipelines and the aggregate emissions from gathering pipelines are Gaussian random variables. Using the means and standard deviations derived in Sections S4.1 and S4.2, we simulate 1,000,000 draws of each value (aggregate pipeline emissions and pipeline distance), for each survey subset. We take the median of the resulting emission factors as the point estimate, and the 2.5th and 97.5th percentiles as the 95% confidence interval.

S5 Monte Carlo Simulation Experiment

S5.1 Experiment

In this experiment, we take all gathering pipeline emission sources with $n_o \geq 3$ from the Fall 2019 campaign (which had the highest number of such sources) and place them randomly on a “line” of length 27,518 km, the estimated linear distance of gathering pipelines covered in the Fall 2019 survey ($n_o \geq 3$ region). We then consider a hypothetical ground survey of length 100 km. To do this, we randomly select a 100 km interval along the 27,518 km line. The 100 km interval may wrap around the 27,518 km line, so every source is equally likely to fall within the 100 km interval. This 100 km interval is the simulated ground survey. (It should be noted that, in reality, this 100 km survey could not be randomly placed due to restrictions on access to certain gathering pipelines.) We then sum the persistence-adjusted emissions of the pipeline sources that fall in that 100 km interval and divide by the survey distance (100 km) to calculate an emission factor. We conduct this process 10,000 times total for the 100 km distance, and report the 25th percentile, 50th percentile (median), and 75th percentile. Then, we repeat this process for lengths of multiple 100 between 200 km and 2,000 km inclusive. We compare these estimates to the estimate derived from the entire survey region. The results are displayed in Figure S3.

S5.2 Accounting for Emissions Below the Minimum Detection Limit

Our goal in this study is to produce a conservative estimate of a gathering pipeline methane emission factor in the Permian. To that end, for our main estimates, we disregard emission sources in the partial detection range and below the minimum detection limit.

However, the methane plume distributions suggest that the aerial instrument misses a sizable range of emission sizes. Our hypothesis is that each bin containing lower emission rates should have a higher frequency. Chen et al. 2022 used ground measurements and a bootstrapping approach to estimate emissions in the partial detection range and below the minimum detection limit [2]. However, here we are constrained by a limited number of existing ground-based measurements of

gathering pipelines. Furthermore, the range of measurements is limited, as the largest recorded gathering pipeline leak in a ground-based survey was 4.0 kg hr^{-1} , significantly lower than the minimum detection limit of the aerial instrument. We propose a method of simulating emissions below the minimum detection limit using the exponential distribution, although we recognize the limitations of such an approach. Our steps are summarized below.

1. Collect all emission sources q_i for which $n_o \geq 3$.
2. Select an interval $[a, b]$ of the data which could feasibly be modeled by the exponential distribution. (We test multiple ranges here and find that the results are not unreasonably sensitive, as long as the tail of the distribution is not included in the interval.) Let $Q = \{q_i | q_i \in [a, b]\}$. Calculate the mean of Q , and call this \bar{Q} .
3. Choose a value $\lambda > 0$. Draw $X_1, \dots, X_{100,000} \sim \text{Expo}(\lambda)$. Let $S = \{X_i | X_i \in [a, b]\}$. Calculate the mean of S , and call this \bar{S} . Repeat this process for different values of λ until $\bar{S} \approx \bar{Q}$; linear regression is helpful here. Record this value of λ as λ_0 .
4. Count the number of q_i values for which $q_i > a$, and call this m . To determine the quantity of small emission sources we should draw, we can solve for n :

$$\frac{n}{n+m} = \int_{x=0}^a \lambda_0 e^{-\lambda_0 x} dx.$$

5. Now, we draw $Y_1, Y_2, \dots \sim \text{Expo}(\lambda_0)$, and take the first n values of Y_i that are less than a (call this set Y). These are our simulated emission sources below the minimum detection limit.
6. We can take one of two approaches. The first, which makes generous assumptions about the presence of below-detection-limit emissions, adds all elements of Y to the observed emission sources. The second, which makes more conservative assumptions, joins all observed emission sources above a with the elements of Y .

The result of this process is displayed in Figure S3, with the bars indicating the contribution of simulated emissions below the detection limit. We use the first approach from step 6 to illustrate

that, even if the more generous assumptions are made, the contribution of these small emission sources is relatively modest compared to the contribution of larger sources.

S6 Figures and Tables

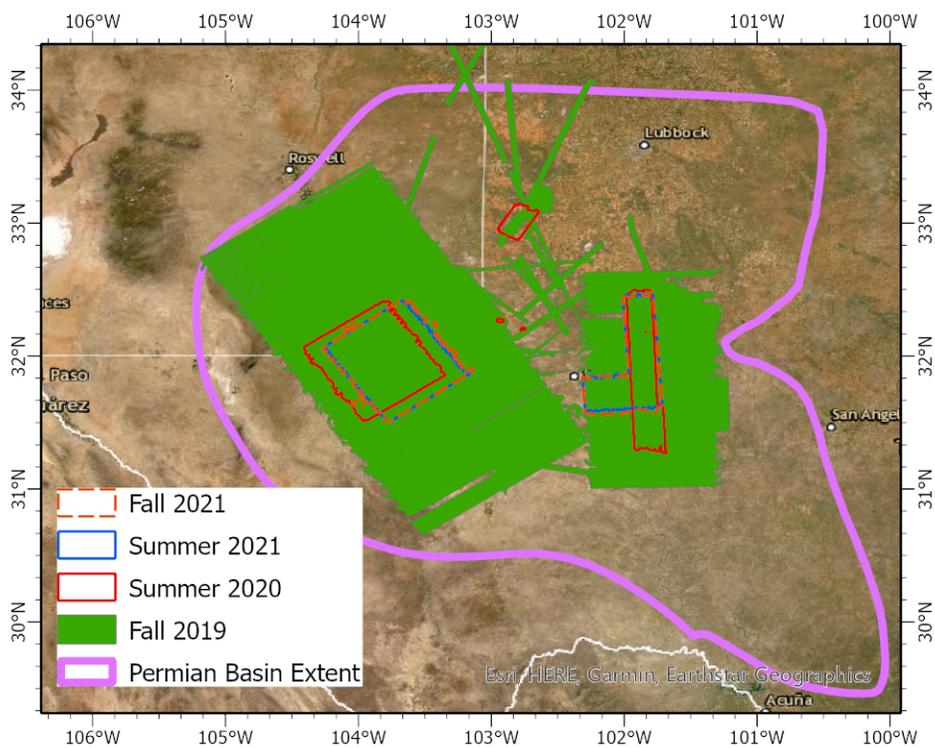


Figure S1: **Coverage of the four aerial campaigns in the Permian Basin.** Two roughly box-shaped figures are visible; these covered the Delaware (west) and Midland (east) basins. The entire Permian Basin extent is shown in pink [3]. Note that the Fall 2019 campaign coverage is larger than — and generally a superset of — the coverage of the other campaigns. To see the effect, we try limiting the gathering pipeline sources from the Fall 2019 campaign to those falling within the coverage of later campaigns. We estimate aggregate gathering pipeline emissions (and corresponding emission factors) of 17,000 (9.1), 20,000 (9.6), and 20,000 kg hr^{-1} (9.3 $\text{Mg y}^{-1} \text{km}^{-1}$) in Fall 2019 for the regions encompassed by the Summer 2020, Summer 2021, and Fall 2021 campaigns respectively, compared to 90,000 kg hr^{-1} (10.0 $\text{Mg y}^{-1} \text{km}^{-1}$) for the full Fall 2019 region.

Campaign	GPD (km)	TPD (km)	GPES (count)	TPES (count)	UPES (count)	GPAF (sources/km)	TPAF (sources/km)
Fall 2019	79000	9900	331	11	13	0.0042	0.0011
Summer 2020	17000	1600	56	1	3	0.0034	0.0006
Summer 2021	18000	1800	80	1	8	0.0044	0.0006
Fall 2021	19000	1800	50	0	5	0.0026	0.0000

Table S1: **Estimated activity factors for gathering and transmission lines in each aerial campaign.** GPD=Gathering Pipeline Distance, TPD=Transmission Pipeline Distance, GPES=Gathering Pipeline Emission Sources, TPES=Transmission Pipeline Emission Sources, UPES=Unknown Pipeline Emission Sources, GPAF=Gathering Pipeline Activity Factor, TPAF=Transmission Pipeline Activity Factor.

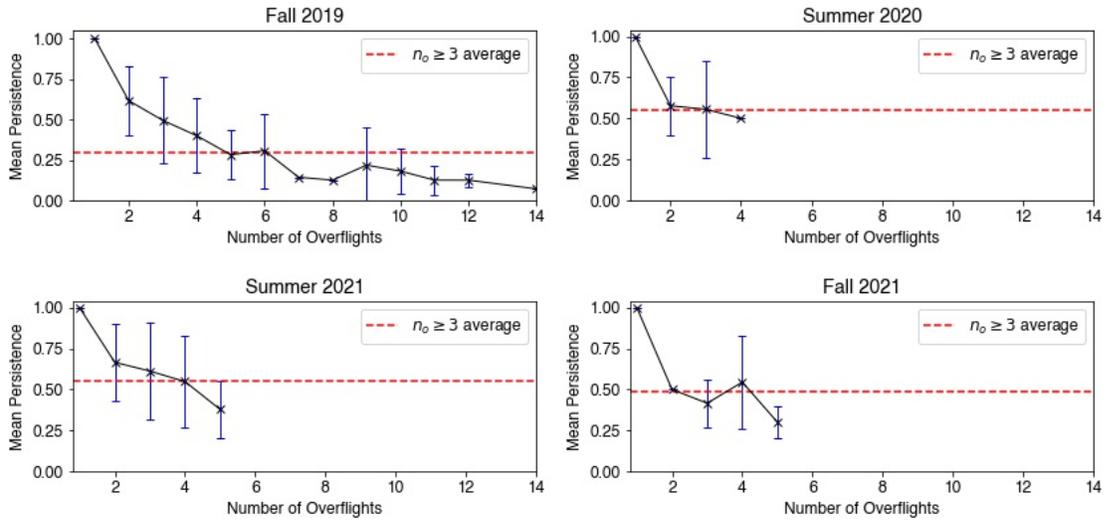


Figure S2: Mean persistence of gathering pipeline emission sources versus number of overflights, for the four aerial campaigns. The uncertainty bars represent one standard deviation in each direction.

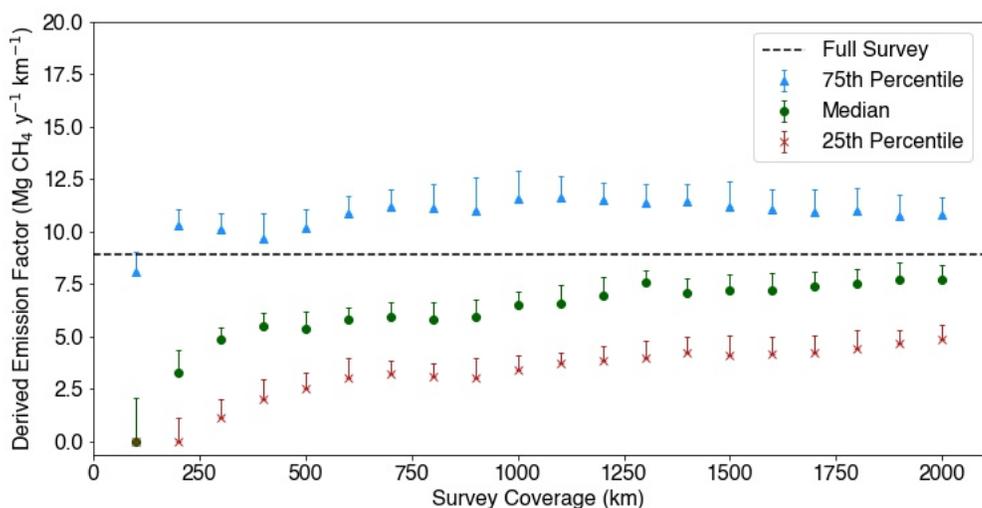


Figure S3: **Monte Carlo simulation experiment of emission factors derived from hypothetical ground surveys.** Under the liberal assumption that ground surveys have the potential to identify any of the sources found in aerial surveys, a ground campaign covering 500 km or less will systematically underestimate (in median) a basin-wide emission factor. In the relatively unlikely scenario that a ground survey identifies one or more high-emitting sources, the aggregate estimate may be excessively high, as in the 75th percentile path. The bars represent the contribution of simulated emission sources below the minimum detection limit (see Section S5.2).

S7 References

- (1) Cusworth, D. H.; Duren, R. M.; Thorpe, A. K.; Olson-Duvall, W.; Heckler, J.; Chapman, J. W.; Eastwood, M. L.; Helmlinger, M. C.; Green, R. O.; Asner, G. P.; Dennison, P. E.; Miller, C. E. Intermittency of Large Methane Emitters in the Permian Basin. *Environ. Sci. Technol. Lett.* **2021**, *8* (7), 567–573. <https://doi.org/10.1021/acs.estlett.1c00173>.
- (2) Chen, Y.; Sherwin, E. D.; Berman, E. S. F.; Jones, B. B.; Gordon, M. P.; Wetherley, E. B.; Kort, E. A.; Brandt, A. R. Quantifying Regional Methane Emissions in the New Mexico Permian Basin with a Comprehensive Aerial Survey. *Environ. Sci. Technol.* **2022**, *56* (7), 4317–4323. <https://doi.org/10.1021/acs.est.1c06458>.
- (3) U.S. Energy Information Administration. *Permian Basin: boundary, structure and tectonic features (shapefile)*. <https://www.eia.gov/maps/maps.htm> (accessed 2022-02-24).