nature communications

Article

https://doi.org/10.1038/s41467-025-66465-6

High-resolution national mapping of natural gas composition substantially updates methane leakage impacts

Received: 13 May 2025

Accepted: 6 November 2025

Published online: 21 November 2025

Check for updates

Methane is emitted from oil and gas operations alongside heavier hydrocarbons and non-hydrocarbon gases, shaping emissions management decision-making, including air quality impacts. Yet, most assessments assume fixed gas composition, overlooking significant spatial and temporal variations. Here, we generate a high-resolution, data-driven map of natural gas composition across the United States, reconstructing methane, heavier hydrocarbons, and non-hydrocarbon species using spatio-temporal interpolation and oil-and-gas production patterns. Our approach is able to reduce composition prediction errors by 39% in terms of Mean Absolute Error (MAE) compared to standard techniques and reveals that methane loss rates have been underestimated by more than 50% in some regions. Beyond methane, we uncover substantial variability in co-emitted gases, exposing blind spots in current emissions inventories and emissions management frameworks. Our work enables more accurate emissions assessments, guides targeted measurement strategies, and informs emissions management decision-making. It also provides a general framework for prediction in environmental applications that integrate sparse measurements with auxiliary variables.

Methane is a potent greenhouse gas, with a global warming potential (GWP) over a 100-year period that is 28–36 times that of carbon dioxide (CO_2)[1]. Consequently, methane emissions from oil and gas operations have been extensively studied due to their large contribution to anthropogenic climate forcing[2–5]. According to the World Bank, emissions from venting, leakage, and flaring in the oil and gas sector account for roughly 25% of global anthropogenic methane emissions[6].

While methane is the primary component of natural gas, upstream gas streams often contain substantial quantities of heavier hydrocarbons, such as ethane, propane, and butane, as well as non-hydrocarbon constituents including CO_2 and N_2. These components are progressively removed or transformed during processing and transport, resulting in substantial variation in gas composition along the supply chain. Emissions from oil and gas systems therefore release not only methane, but also co-emitted species with distinct climate, air quality, and health impacts. For instance, heavier hydrocarbons contribute to radiative forcing and tropospheric ozone formation, while non-methane volatile organic compounds (NMVOCs) are associated with adverse health outcomes.

The produced gas refers here to the gas obtained immediately after separation from liquid. Its composition is a key quantity that was shown to correlate with the composition of emissions and flash emissions factors[7] and is involved in the computation of the production-normalized methane loss rate (ρ), a widely used metric for emissions intensity. This rate is typically defined as the mass of methane emitted (m) divided by the estimated mass of methane

[1]Department of Energy Science & Engineering, Stanford University, Stanford, CA, USA. [2]Systems and Energy Technologies Analysis Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [3]Integrated Ecosystem Sciences Department , Lawrence Berkeley National Laboratory, Berkeley, USA. [4]Integrated Ecosystem Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [5]Insight M, Sunnyvale, CA, USA.
e-mail:

produced, approximated as the product of the mass of metered produced gas production (p) and methane mass fraction in the gas (x), i.e., $\rho = m/(px)$. Because ρ is inversely proportional to x, if x is overestimated by a factor of two, ρ is underestimated by half.

While this composition varies substantially across regions, within basins, and over time, many studies rely on simplified assumptions, such as a fixed national methane fraction (e.g., $x = 90\%$ in ref. 5) or single representative values per basin[4]. Methane contents reported in the literature range from as low as 47% in the Bakken[8] to 97% in the Fayetteville Shale[9], highlighting the limitations of uniform assumptions. Even recent basin-level improvements[9] fail to capture intrabasin variability and temporal dynamics.

Given this variability, assessing the climate and health impacts of emissions requires produced gas composition data that are both spatially localized and densely sampled in space and time. Localized data enable evaluation of air quality impacts at specific sites or regions directly affected by emissions. Dense spatial sampling ensures that high-production areas, which disproportionately influence regional or national emissions estimates, are adequately represented. Temporal resolution is similarly essential for tracking changes due to reservoir depletion, operational shifts, or infrastructure development.

Several datasets, some publicly available, offer partial information on gas composition but remain sparse and incomplete. The United States Geological Survey (USGS) has compiled a dataset of 17,661 gas well samples collected between 1918 and 2014[10]. The U.S. EPA's Greenhouse Gas Reporting Program (GHGRP) aggregates methane content data by county and facility[11], and state-level databases, such as those maintained by the Texas Railroad Commission[12], provide additional information. However, these datasets are fragmented and inconsistent in space and time, limiting their utility for national-scale compositional mapping.

Interpolation techniques, such as kriging, leverage the spatial continuity of the geochemical composition of extracted natural gas, which is influenced by reservoir characteristics and hydrocarbon source materials[13]. For example, Barkley et al.[9] applied kriging to interpolate methane fractions from the USGS database, generating a map at a resolution of 0.25 degrees latitude by 0.25 degrees longitude (~28 km by 28 km). While spatial correlations were effectively utilized, this analysis primarily served as a prior for inferring methane emissions from ethane measurements and did not evaluate the impact of methane fraction on the production-normalized loss rate ρ or account for associated uncertainties. Furthermore, temporal variations and secondary informative variables, such as the gas-to-oil ratio (GOR)—an operational parameter shown to help inform estimates of produced gas composition[7]—were not incorporated.

Here, we introduce a method to estimate spatial and temporal variations in produced gas composition across the US by combining spatio-temporal kriging with a non-linear model taking gas and oil productions as inputs. By applying our method to publicly available datasets, we generate maps of produced gas composition at a 2-by-2 km spatial resolution and 5-year temporal intervals. We make the assumption that the produced gas, immediately after gas-liquid separation, is equivalent to the gas compositions reported in the available datasets (e.g., USGS, GHGRP). With our current data sources, we cannot quantify potential discrepancies between the true produced gas composition and the reported values, such as those arising from changes in reservoir temperature and pressure over time relative to initial sampling. We validate our method using rigorous train/test splits and compare our results to simpler interpolation methods and ablated versions of this work (see Supplementary Section 4.1). Using our framework, we demonstrate that production-normalized loss rates have likely been underestimated in previous studies by 7–54%. Importantly, this study provides a systematic reconciliation of the disparate datasets to which our method is applied. Our results provide a basis for analyzing the environmental impact of higher hydrocarbons emitted alongside methane, as well as the human health implications of other co-emitted components, and motivate new samplings. Finally, as our method includes uncertainty estimates, it can be used to design measurement processes that maximize information gain.

Results

Available data on the produced gas composition

The datasets used as inputs for this work are described in Table 1 and below.

At the national level in the United States, the most comprehensive publicly available produced gas composition databases are: (1) a database released by the USGS in 2021[10], and (2) gas production and gas processing data reported to the GHGRP between 2015 and 2021[11]. GHGRP data have the advantage of being updated yearly under the required reporting from operators.

The USGS database aggregates analyses of 17,661 gas well samples performed by the Bureau of Land Management's Federal Helium Program, the Bureau of Mines, and the USGS from 1918 to 2014[10]. The gas samples were characterized employing chromatographic or mass spectrometric techniques. The dataset records molar percentages of various gases, including multiple hydrocarbons, helium, carbon dioxide, hydrogen, and nitrogen. The gas composition is accompanied by detailed information such as the well's American Petroleum Institute (API) number, geographic coordinates, and collection dates.

Under GHGRP's Subpart W, facilities emitting over 25,000 metric tons of carbon dioxide equivalent annually must report not only their greenhouse gas emissions but also specific operational data. This includes methane and carbon dioxide molar fractions in natural gas at both the production and processing stages. Data from 750 onshore production facilities and 599 processing facilities covering 2015 to 2021 are available at the county and facility levels. We conducted a spot check on three processing facilities by comparing GHGRP-reported data to proprietary operational data, finding close agreement; see Supplementary Section S2.1 for details. For production operations, a "facility" as defined by GHGRP is all wells owned by a given operator in a particular county, grouped into one large aggregate facility. As a result, a single "facility" can contain hundreds of wells. Data linking individual wells to GHGRP facilities is available,[14] making it possible to estimate methane and carbon dioxide fractions at the well level for those wells linked to reporting production facilities.

Table 1 | Description of the datasets and how they are analyzed

Data source	Temporal range	Components	Database size	Continuous output	Weighting variable
United States Geological Survey (USGS)	1918–2014	C1, C2, C3, N-C4, I-C4, N-C5, I-C5, C6+, He, CO_2, H_2, N_2, H_2S, Ar, O_2	17,661 samples	✓	Gas production volume
GHGRP (production facilities)	2015–2021	C1, CO_2	750 facilities	✓	Gas production volume
GHGRP (processing facilities)	2015–2021	C1, CO_2	599 facilities	✗	Plant gas flow

The "Continuous output" column indicates whether spatial interpolation is applied. The components predicted by the non-linear model are underlined. Greenhouse Gas Reporting Program (GHGRP) data reported by processing facilities are not interpolated, as they represent facility-level measurements without geographic correlation.

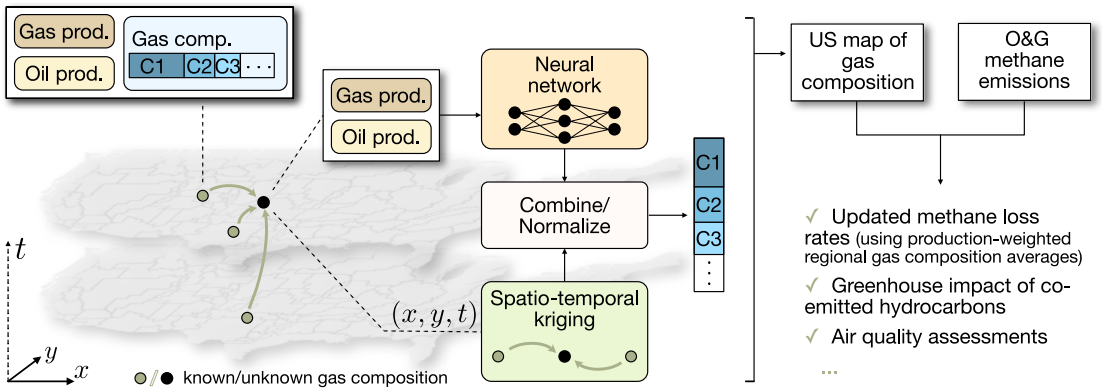


Fig. 1 | Schematic overview of our spatio-temporal interpolation framework for estimating natural gas component fractions. Produced gas composition estimates are derived by integrating production data with available measurements, combining neural network outputs and kriging interpolation. The resulting spatially resolved produced gas composition maps inform methane loss estimates, greenhouse gas impacts, and air quality assessments. O& G refers to oil and gas operations. Figure made using Matplotlib (see Supplementary Section S6 for details and references on all software tools used in this study).

These two datasets have somewhat different characteristics. GHGRP data are more recent and reported more consistently. USGS data, on the other hand, are reported for each well individually, and therefore have somewhat higher spatial resolution, given that GHGRP data are reported for all wells in a county. Therefore, these datasets can be considered as complementary.

Additionally, oil and gas production data were extracted from Enverus DrillingInfo[15], available at the well level, dating back to 1916. These data were aligned with the gas composition data (see Section 4.1).

Spatio-temporal prediction of produced gas composition

Given these local measurements, we developed a method based on kriging to interpolate the composition of natural gas across onshore regions of the United States and reveal intra- and inter-basin variations. In particular, our method can be applied to estimate the produced gas composition at production sites, allowing the refinement of production-normalized methane loss rates and CO_2 equivalent emissions. At production sites, we use gas and oil production as additional variables, as the GOR (in scf bbl^{-1}, based on reported production volumes) correlates with light hydrocarbon fractions[16].

Because this relationship between GOR and light hydrocarbon fractions does not have to be linear, we use a non-linear model with gas and oil production as inputs to improve estimation accuracy (see Supplementary Table S1 and Supplementary Fig. S3).

Produced gas composition estimates obtained from kriging and from our non-linear model are combined using a weighted average, minimizing errors on a held-out dataset. Lastly, we normalize our predictions to ensure that component fractions sum to one. For components with weak correlations with our auxiliary variables (all components except C1 and C2), our method exclusively relies on kriging. Importantly, our method provides uncertainty estimates at the local level and confidence intervals at the aggregate level, facilitating robust comparisons of estimates across data sources. Our method is visually described in Fig. 1, and further methodological details are provided in Section 4.2 and Supplementary Fig. S1.

To assess the added value of our kriging-based method as well as the benefit of incorporating auxiliary variables, we compared our method to a naive baseline (nearest-neighbor) and conducted an ablation study removing the non-linear model, relying solely on kriging. Using four-fold cross-validation (See Supplementary Section S4.1), our method achieves a Mean Absolute Error (MAE) of 5.8 mol% applied to USGS C1 data, representing a 15% reduction relative to ordinary kriging and 39% relative to nearest-neighbor. In data-sparse regions, the improvement obtained using the auxiliary variable reaches ~25% on USGS C1 data, 15% on GHGRP C1 data reported by production facilities, and 4% on USGS C2 data (Supplementary Fig. S10).

National US maps of produced gas composition

The results of our interpolation method for the period 2016–2021 are shown in Fig. 2. When applying our method to USGS and GHGRP data reported by production facilities, we show the estimated means and standard deviations of the methane fraction in regions where production data are available. The basin-level aggregates are then computed as a production-weighted mean, and the standard error (SE) of the new estimate can be derived from the local standard deviations. Basin borders are defined according to the American Association of Petroleum Geologists (AAPG) geologic provinces[17]. For GHGRP processing facilities, the map displays the already-available facility-level methane fractions in the input gas. Final estimates are computed as plant-flow-weighted averages by basin. Because processing plants receive gas from multiple production sites, this processing-site-oriented map presents a different perspective on methane distribution than the maps that show estimates at specific locations.

Due to the sparse and temporally-scattered nature of USGS data, our estimates using USGS data generally exhibit higher standard deviations, except in regions where initial samples are available. Our estimates using GHGRP production facility data exhibit lower standard deviations, likely due to the dataset's higher density and more recent coverage. Additionally, each GHGRP datapoint represents an average across multiple wells within a county for a given operator—potentially aggregating data from hundreds of wells. This averaging inherently reduces variability. Our approach assigns facility-wide average methane and CO_2 fractions to wells (see Section 4.1); we note that incorporating uncertainties in this assignment could increase the standard deviation estimates of the GHGRP-based analysis.

Additional zoomed-in maps of our estimates for all gas components from USGS data for the Anadarko Basin over the same time period are provided in Supplementary Fig. S14, and zoomed-in maps of methane fraction estimates from GHGRP production facility data in 13 major basins are shown in Supplementary Fig. S16.

Basin-level produced gas composition estimates

In order to estimate the molar fraction of a component in the gas produced across an entire basin, we computed a weighted spatial average of the molar fraction estimated by our method across production sites. The weights in this average were given by the gas production volumes at each production site. Based on USGS and GHGRP data reported by production facilities, we applied this method to

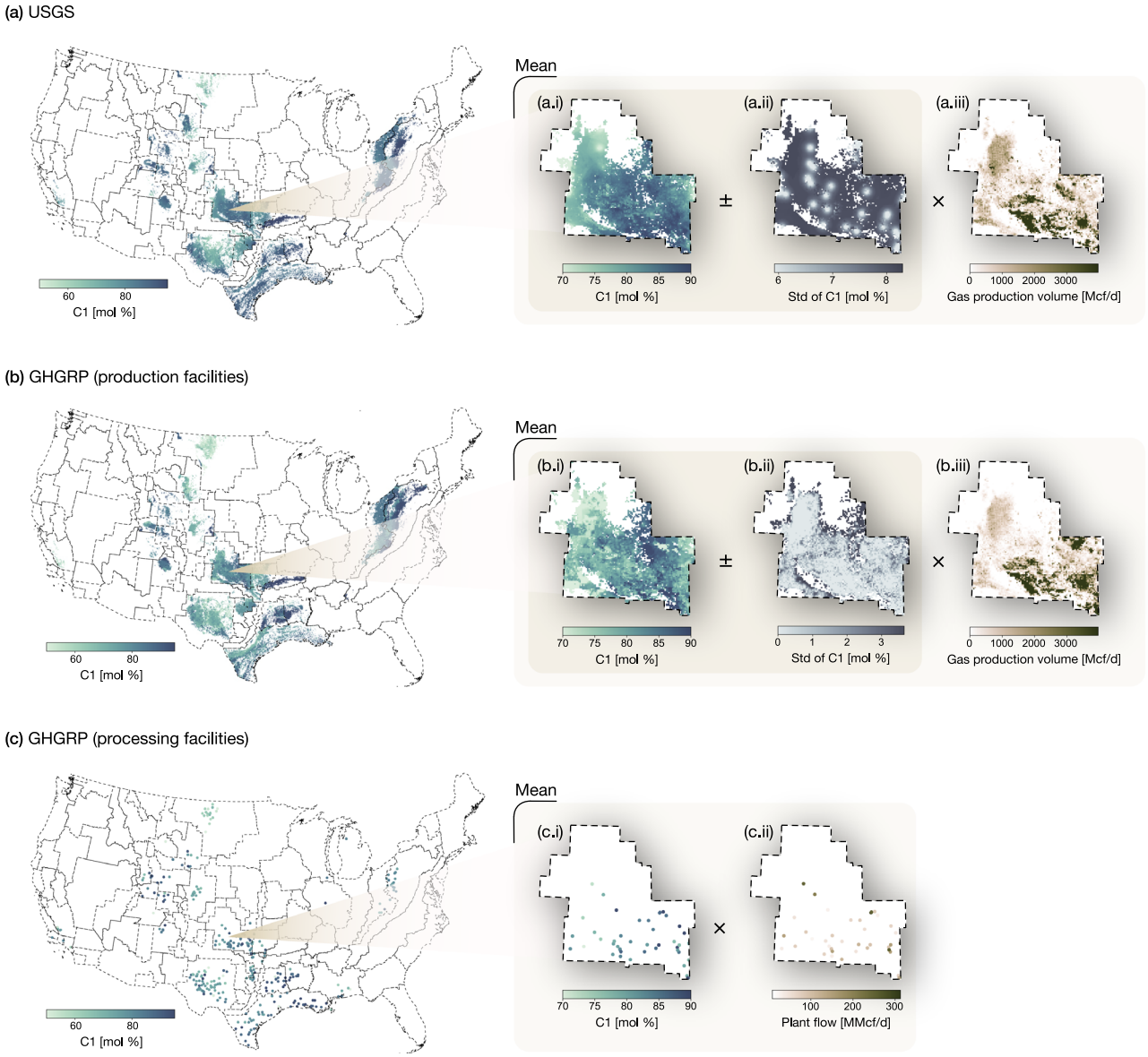


Fig. 2 | Interpolated methane (C1) molar fraction from different input data sources. a United States Geological Survey (USGS), **b** Greenhouse Gas Reporting Program (GHGRP) production facilities, and **c** GHGRP processing facilities. The maps represent estimated average produced gas composition for the 2016–2021 period (see Section 4.3 for details on temporal selection). The left column shows the national-scale spatial distribution of predicted C1 fractions across all basins. The right panels provide a zoomed-in view of the Anadarko Basin, showing C1 estimates (blue-green colormap), associated standard deviations (dark blue/gray colormap), and the weighting variable used in the interpolation (brownish colormap). Basin-level aggregates are computed as weighted averages, visually synthesizing information from these three maps. For GHGRP processing facilities (**c**), raw facility-level data from 2015 to 2021 are shown instead of interpolated values. Figure made using Matplotlib, Seaborn, GeoPandas and Shapely, with shapefiles from ref. 23 (see Supplementary Section S6 for details and references on all software tools used in this study).

determine basin-level averages of C1 and CO_2 fractions in different basins. For GHGRP data reported by processing facilities, we computed a weighted average of all reported fractions, with processing plant flows as weights.

To quantify uncertainty in these basin-level estimates, for methods relying on spatio-temporal interpolation, we simulated multiple plausible distributions of produced gas composition that account for spatio-temporal dependencies. This ensures that confidence intervals for basin-level values account for both local uncertainties and their aggregation. For GHGRP data reported by processing facilities, where individual uncertainties are not available, we used a resampling method. See Section 4.4 for more details on how uncertainties are computed. These confidence intervals allow for consistent comparisons between basins and data sources.

Results of these basin-level estimates and their uncertainties for the period 2016–2021 are presented in Fig. 3. Additionally, as illustrative examples, temporal trends in methane fraction estimates for the Permian and Uinta basins, derived from USGS data, are presented in Supplementary Fig. S15.

The methane fractions reported by GHGRP processing facilities are, in many cases, higher than those from production facilities (e.g., in the Denver, East Texas, Gulf Coast, Powder River, San Joaquin, Uinta, and Williston basins), suggesting the influence of processing steps or heavier hydrocarbon condensation, either of which would alter the gas composition before it is delivered. Notably, almost all estimates presented here have mean values below the 90% assumed in refs. 4,5, except for those from GHGRP processing facilities in the Uinta Basin. The methane fractions in the San Joaquin Basin are significantly lower,

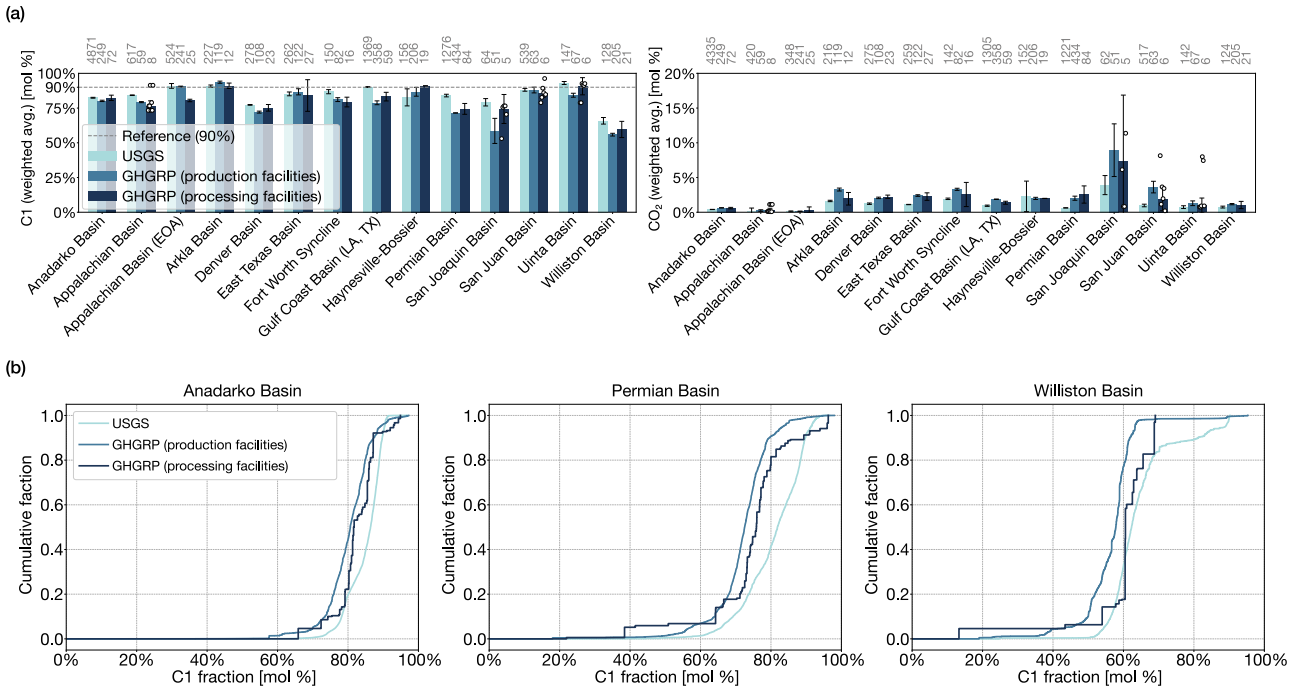


Fig. 3 | Basin-level aggregated methane (C1) and carbon dioxide (CO_2) fractions for the 2016–2021 period. a Weighted average CO_2 (left) and C1 (right) molar fractions across multiple basins, comparing data from the United States Geological Survey (USGS), Greenhouse Gas Reporting Program (GHGRP) production facilities, and GHGRP processing facilities. Error bars represent 95% confidence intervals (2 standard errors, computed as described in Section 4.4). The number of input data points (all years considered, for each basin) used to derive the reported average is displayed in gray above the bars. The dashed line in the C1 plot represents the default 90 mol% assumption for C1 made in refs. 4,5. The label "Appalachian Basin (EOA)" refers to the AAPG Appalachian Basin (Eastern Overthrust Area). Note that the *y* axis scales differ between the CO_2 and C1 plots; this should be considered when visually comparing values across the two panels. **b** Weighted cumulative distribution functions (CDFs) of C1 molar fractions for three selected basins: Anadarko, Permian, and Williston.

which is consistent with previously reported high methane loss rates in this area. Estimates for the Williston Basin exceed the values reported in ref. 8, although they fall within the uncertainty bounds. Importantly, the overlap of 95% confidence intervals between methane and CO_2 fractions in USGS and GHGRP data reported by production facilities indicates statistical consistency. This means that these datasets can be meaningfully integrated to obtain a complete representation of produced gas composition. This is particularly valuable because GHGRP data are updated annually, while USGS data capture the full range of gas components.

Integrated estimates of full produced gas composition

To obtain a unified basin-level produced gas composition by combining available datasets, we relied on our basin-level estimates for each component and each basin based on USGS data, except for methane and carbon dioxide, for which we relied on our estimates from GHGRP data reported by production facilities, which have more complete data. We then summed and normalized the remaining component fractions so that all components add to 100%, keeping the methane and carbon dioxide values unchanged. Results are summarized in Table 2. We plan to release a national unified gridded spatial produced gas composition map constructed using the same combination logic (see Section 4.3).

The results reveal substantial regional variation in methane fractions, ranging from <60% in the Williston and San Joaquin basins to more than 90% in the Arkla basin. Beyond methane, minor gas components further differentiate regional compositions. Carbon dioxide (CO_2) fractions are particularly elevated in the San Joaquin Basin (8.94%) and San Juan Basin (4.24%), while they remain below 1% in the Appalachian and Anadarko basins. Nitrogen content also varies widely, with the Williston Basin reaching 10.89% and Anadarko at 7.94%, compared to under 1.5% in Arkla and East Texas. H_2S levels range from negligible in basins such as the Appalachian and San Joaquin basins, to 6.19% in the Williston and 5.99% in East Texas. These compositional differences reflect distinct geochemical signatures across basins.

We also computed the average produced gas composition across the United States. On average, methane (C1) constitutes 80.42% (±7.36%) of the gas mixture. Other major components include ethane (C2, 8.58% ± 3.28%), propane (C3, 3.64% ± 2.16%), nitrogen (N_2, 3.25% ± 2.94%), and carbon dioxide (CO_2, 1.99% ± 1.42%). All other components, including hydrogen, hydrogen sulfide, and heavier hydrocarbons (C4+), occur at concentrations below 1%.

While these values allow us to report a U.S. average produced gas composition, it is important to note that individual basins can differ substantially from this national average. For example, methane content ranges from over 93% in the Arkla Basin to under 60% in the Williston and San Joaquin basins.

Updates on recent methane loss rate estimates

To evaluate the impact of our updated produced gas composition estimates, we examined how basin-specific methane fractions alter inferred production-normalized methane loss rates compared to the conventional assumption of a fixed methane fraction. Specifically, we updated the methane loss rates reported by Sherwin et al.[5] by replacing their fixed 90% methane molar fraction with our basin-specific methane compositions. The methane loss rate ρ is defined as $\rho = \frac{m}{p \times x}$ where m is the emitted methane mass, p is the total mass of metered produced gas, and x is the methane mass fraction in the produced gas. We use the same emissions and production datasets as in ref. 5; the only difference is that we substitute their fixed methane fraction with our basin-specific methane mass fractions. These mass fractions are computed by converting our estimated molar compositions of produced gas to a mass basis using molecular weights (see Section 4.5). Because our composition estimates represent the produced gas

Table 2 | Full produced gas composition across selected basins for the period 2016–2021

	Ana- darko	Appalachian	Arkla	Denver	East Texas	Fort Worth	Gulf Coast	Perm- ian	San Joaquin	San Juan	Uinta	Willi- ston	US
C1	80.19 ± 0.25	89.50 ± 0.12	93.69 ± 0.42	72.10 ± 0.36	86.71 ± 1.13	81.10 ± 0.71	78.82 ± 0.71	71.50 ± 0.05	58.57 ± 4.61	86.50 ± 1.18	84.26 ± 0.73	56.25 ± 0.23	**80.42 ± 7.36**
C2	6.31 ± 0.07	5.32 ± 0.27	3.42 ± 0.45	11.38 ± 0.48	3.67 ± 0.35	9.15 ± 0.38	11.21 ± 0.20	12.52 ± 0.33	12.68 ± 1.32	7.53 ± 0.29	7.68 ± 1.73	14.05 ± 1.75	**8.58 ± 3.28**
C3	3.08 ± 0.04	1.21 ± 0.40	0.72 ± 0.08	8.27 ± 0.26	0.98 ± 0.15	4.01 ± 0.11	3.86 ± 0.08	6.02 ± 0.17	12.15 ± 0.58	2.94 ± 0.12	2.66 ± 0.42	6.27 ± 2.04	**3.64 ± 2.16**
N-C4	0.89 ± 0.02	0.43 ± 0.09	0.30 ± 0.07	2.41 ± 0.28	0.31 ± 0.04	1.31 ± 0.04	1.21 ± 0.12	1.78 ± 0.12	4.32 ± 0.26	0.73 ± 0.04	0.72 ± 0.24	2.00 ± 0.24	**1.11 ± 0.68**
I-C4	0.39 ± 0.01	0.23 ± 0.03	0.32 ± 0.07	1.16 ± 0.06	0.26 ± 0.10	0.65 ± 0.03	1.14 ± 0.05	0.80 ± 0.03	2.67 ± 0.22	0.51 ± 0.06	0.54 ± 0.09	1.49 ± 0.17	**0.70 ± 0.43**
N-C5	0.23 ± 0.00	0.15 ± 0.03	0.22 ± 0.04	0.69 ± 0.06	0.13 ± 0.03	0.33 ± 0.02	0.50 ± 0.02	0.49 ± 0.06	1.46 ± 0.15	0.23 ± 0.01	0.31 ± 0.08	0.84 ± 0.37	**0.36 ± 0.21**
I-C5	0.21 ± 0.01	0.15 ± 0.02	0.16 ± 0.02	0.70 ± 0.05	0.10 ± 0.02	0.33 ± 0.04	0.51 ± 0.06	0.46 ± 0.03	1.79 ± 0.13	0.25 ± 0.01	0.37 ± 0.10	0.71 ± 0.27	**0.36 ± 0.24**
C6+	0.27 ± 0.01	0.14 ± 0.02	0.14 ± 0.03	0.91 ± 0.04	0.22 ± 0.04	0.26 ± 0.02	0.49 ± 0.03	0.47 ± 0.01	1.36 ± 0.18	0.28 ± 0.02	1.23 ± 0.16	0.47 ± 0.14	**0.36 ± 0.20**
N_2	7.94 ± 0.22	2.30 ± 0.31	0.57 ± 0.23	2.12 ± 0.07	1.29 ± 1.18	2.44 ± 0.15	1.33 ± 0.26	4.60 ± 0.44	4.38 ± 1.27	0.47 ± 0.33	1.81 ± 1.27	10.89 ± 2.34	**3.25 ± 2.94**
CO_2	0.64 ± 0.01	0.14 ± 0.02	3.25 ± 0.16	2.07 ± 0.05	2.41 ± 0.06	3.31 ± 0.08	1.87 ± 0.02	2.02 ± 0.15	8.94 ± 1.94	4.24 ± 0.45	1.30 ± 0.17	0.90 ± 0.02	**1.99 ± 1.42**
H_2S	0.04 ± 0.03	0.00 ± 0.00	0.03 ± 0.01	0.02 ± 0.01	5.99 ± 0.71	0.03 ± 0.02	0.10 ± 0.04	0.70 ± 0.05	0.00 ± –	0.24 ± 0.05	0.00 ± –	6.19 ± 3.95	**0.68 ± 1.45**
O_2	0.11 ± 0.04	0.35 ± 0.07	0.16 ± 0.02	0.13 ± 0.01	0.17 ± 0.03	0.11 ± 0.04	0.32 ± 0.18	0.25 ± 0.09	0.26 ± 0.82	0.24 ± 0.35	0.16 ± 0.18	0.20 ± 0.51	**0.23 ± 0.14**
H_2	0.02 ± 0.00	0.12 ± 0.04	0.19 ± 0.02	0.04 ± 0.02	0.09 ± 0.01	0.11 ± 0.05	0.27 ± 0.01	0.26 ± 0.10	0.10 ± 0.04	0.02 ± 0.04	0.09 ± 0.04	0.28 ± 0.04	**0.15 ± 0.05**
HE	0.24 ± 0.00	0.00 ± –	0.02 ± 0.00	0.05 ± 0.00	0.01 ± 0.00	0.09 ± 0.01	0.02 ± 0.00	0.08 ± 0.01	0.13 ± 0.02	0.03 ± 0.04	0.03 ± 0.01	0.09 ± 0.03	**0.08 ± –**
Ar	0.06 ± 0.02	0.09 ± 0.02	0.07 ± 0.01	0.02 ± 0.01	0.06 ± 0.01	0.07 ± 0.02	0.10 ± 0.04	0.08 ± 0.00	0.11 ± 0.02	0.03 ± 0.05	0.16 ± 0.08	0.27 ± 0.05	**0.08 ± 0.03**

The reported basins account for approximately 90% of total US gas production. Values represent molar fractions (%) ± standard errors, where standard errors are computed for each basin-level estimate, as described in Section 4.4. For the U.S. average (in bold font), the reported uncertainty is the standard deviation across basin-level estimates. The Appalachian value corresponds to the average across the two basins: Appalachian Basin and Appalachian Basin Eastern Overthrust Area. The last column presents the US average.

immediately after separation, they provide an accurate basis for estimating methane content in metered production.

As shown in Fig. 4, incorporating basin-specific compositions can substantially alter inferred methane loss rates, particularly in regions where actual methane fractions differ significantly from 90%. In the Permian, the revised methane loss rate is 26% [23%, 28%] higher than the previous estimate, while in the San Joaquin Basin, the adjustment leads to a 54% [40%, 67%] increase. While most basins show increased loss rates due to methane fractions below 90%, the adjustment can also reduce loss estimates in cases where actual methane content exceeds 90%, as in the Arkla Basin.

Discussion

In this paper, we fill a critical gap in assessing the environmental, human health, and economic impacts of emissions along the natural gas supply chain, all of which depend on the composition of the produced gas, including methane and other co-emitted species. We introduce an interpolation method that combines spatio-temporal kriging with a non-linear model incorporating gas and oil production data to provide detailed estimates of produced gas composition across the continental United States. The result is a granular, data-driven, national time series of onshore natural gas composition, including methane, heavier hydrocarbons, and non-hydrocarbon species.

We show that our interpolation and production-weighted averaging, leveraging three different data sources (USGS, GHGRP production, GHGRP processing), lead to consistent estimates when applied to the three data sources. In particular, this highlights the value of gas composition reporting to the GHGRP for both production and gas processing.

Combining our updated per-basin produced gas composition estimates with methane emissions measurement data reveals that production-normalized loss rates ρ are likely underestimated in most basins, with a relative increase ranging from 7% to 54%.

Our findings, therefore, emphasize the need for incorporating high-quality gas composition data when computing methane loss estimates from oil and gas operations.

Importantly, our method estimates the composition of produced natural gas, not that of emissions along the oil and gas supply chain. Because produced gas is altered during processing, emissions from the gas phase can have compositions that vary by processing stage, and because additional emissions arise from processes such as flash evaporation from the liquid phase, emission composition can differ substantially from that of produced gas (see 4.5). To bridge the gap between the composition of produced gas and that of the emissions, these processes and associated phase changes need to be modeled. This constitutes an interesting and rich avenue for future work.

In this study, we found substantial variability in the non-methane composition of produced gas, both across and within basins. Although not quantified in this paper, these variations are likely to impact the composition of emitted natural gas, with important implications for climate and air quality. NMVOCs are recognized pollutants by agencies such as the US Environmental Protection Agency (EPA) due to their adverse health effects, including respiratory irritation and neurological damage[11]. NMVOCs also act as precursors to ozone, extend methane's atmospheric lifetime, and ultimately oxidize to CO_2. The magnitude of these effects on climate is an active area of research, with multiple studies employing atmospheric modeling to quantify their impact[18–20], and the IPCC recommends accounting for NMVOC oxidation to CO_2 in total carbon emissions from fugitive sources[21]. Furthermore, current greenhouse gas reporting frameworks, such as the Greenhouse Gas Inventory (GHGI) and the GHGRP, primarily account only for methane and, in some cases, CO_2, in CO_2-equivalent emissions. Our estimates of produced gas composition, therefore, offer a foundation for extending

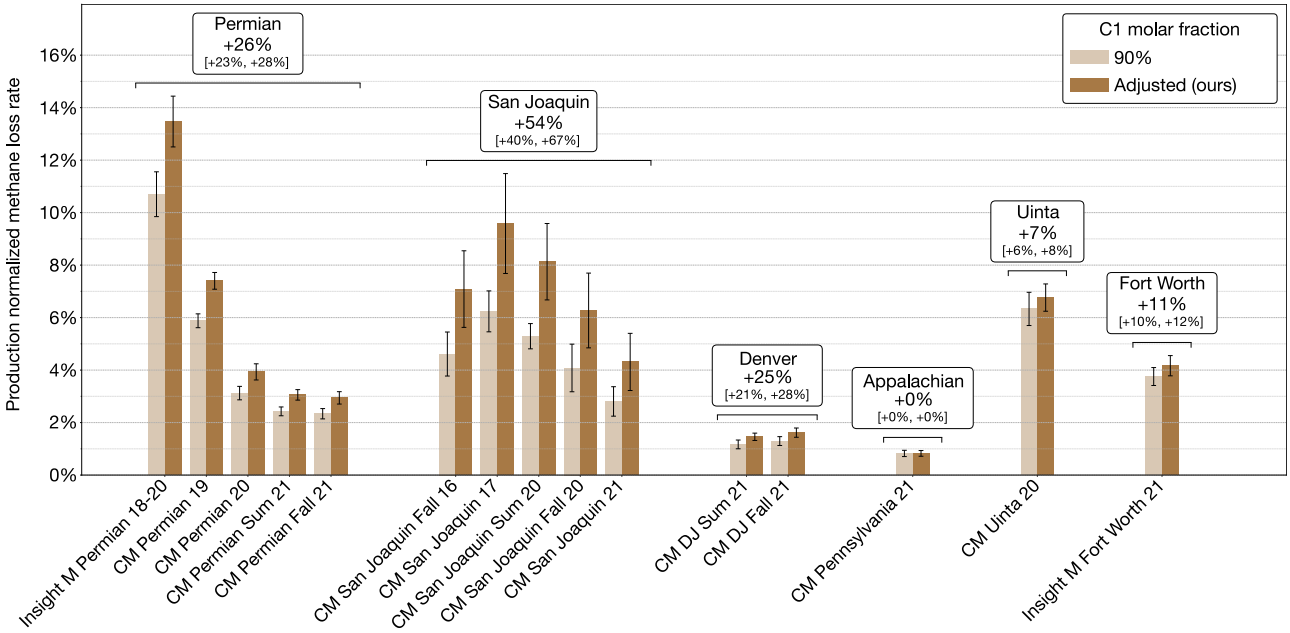


Fig. 4 | Updates of production-normalized methane loss. Methane loss (%) by region and time period, comparing estimates based on a fixed 90% C1 assumption from ref. 5 with our composition-adjusted values, which use our average produced gas composition estimates for the period 2016–2021 by basins. The 2018–2020 Insight M campaign in the Permian was conducted only in New Mexico; for consistency, we apply our Permian-wide average estimate. CM denotes Carbon Mapper. DJ denotes the Denver-Julesburg basin, for which we apply our Denver basin estimate.

emissions assessments to include non-methane species, accounting for their variable abundance and associated impacts.

Because of the heterogeneous nature of the publicly available data on gas composition, our work made different simplifying assumptions depending on the data source. Although USGS wellstream data may theoretically be the most relevant for geospatial interpolation analysis, in practice, not all sampled wells necessarily contribute to marketed natural gas. To address this, we assessed the impact of filtering based on non-hydrocarbon fractions (see Section S7) and found a ≈2 mol% relative difference between the unfiltered data and the most restrictive filtering approach. For GHGRP data reported by production facilities, differences between produced and wellstream gas composition due to separation are not accounted for, potentially limiting the validity of kriging, which assumes geological continuity.

Regarding our spatio-temporal interpolation method, while we focused here on the surface variations of gas composition, integrating depth as an auxiliary variable or an additional dimension would be an interesting extension of this work. Additionally, other well characteristics could be incorporated into our non-linear model to further refine the estimates. Accounting for measurement uncertainties could also improve robustness and is inherently compatible with the kriging framework.

Finally, since our method is partly based on kriging, it provides local estimates of uncertainty (with a higher uncertainty further from measurements). This property allows for targeted improvements in data collection: optimizing future measurement campaigns by prioritizing sampling in regions with both high uncertainty and high gas production could maximize information gain under physical and financial constraints. Developing such applications, drawing from Bayesian optimization frameworks, is a promising direction for future research. New measurement campaigns would provide additional and more recent data, helping to improve our estimates.

By capturing variability in produced gas composition, our approach improves methane loss estimates and reveals the broader climate and health relevance of non-methane constituents. It provides a foundation for more accurate inventories and targeted measurements.

Methods

Data pre-processing

Shapefiles for states and counties in the US were acquired from the US Census bureau[22], and geologic basin boundaries used by EPA's GHGRP were acquired from the US EPA[23]. The coordinate reference system (CRS) was transformed to "EPSG:26914" (Projected coordinate system for North America).

We used Enverus DrillingInfo[15] to retrieve well-level monthly oil and gas production data in the US (1916–2023). In most states, we used well-level production data directly. However, there were states where production was reported only at the lease level (Kansas and Michigan). In these cases, we used the representative well by lease. A detailed description of the different reporting levels and resolution of discrepancies is provided in Supplementary Section S1.1. The Production values were then summed by year, and only the American Petroleum Institute (API14) well number, latitude, longitude, gas production, oil production, and production year data columns were retained. Data were converted into geometry points using latitude and longitude values, and the CRS was transformed to "EPSG:26914" (Projected coordinate system for North America). The well production was then spatially joined with the basins to assign the basin names to each well. In Enverus data, oil volumes are reported in barrels (bbl), with one bbl equal to 42 U.S. gallons or approximately 0.159 cubic meters. Natural gas volumes are reported in standard cubic feet (scf), measured at 60°F and 14.73 psi as defined by the American Gas Association (AGA)[24].

The US Geological Survey (USGS) data were sourced from a CSV file from ref. 10 and processed to remove entries with missing latitude, longitude, or sampling date values. Invalid dates (e.g., the month or day being zero) were manually corrected (removing them when not clear, or e.g., replacing 0 by 1). Geometry points in "EPSG:26914" were created and spatially joined with county shapefiles. Null API14 values were removed, and the filtered data was merged with well production data based on API14 and year. The GORs (scf bbl^{-1}) were calculated as the ratio of reported gas production (in standard cubic feet) to oil production (in barrels).

Data from GHGRP in Subpart W were retrieved from the Query Builder[14] and filtered to include only onshore petroleum and natural gas

production facilities. Facilities where the methane mole fraction was zero or lower than the carbon dioxide mole fraction were removed as likely enhanced oil recovery wells or CO_2 production operations. Well-to-facility data (files named EF_W_ONSHORE_WELLS) were sourced from the GHGRP Query Builder for the years 2016 to 2023. The column WELL_ID_NUMBER contained a variety of formats that needed standardization. A custom function was implemented to convert these values to standardized API14 numbers using county and state numeric codes and the API14 number definition[25]. The county number and state abbreviations were extracted from the SUB_BASIN field, and state abbreviations were mapped to their respective numeric codes. A well-to-facility was then created, associating each well API14 number to the corresponding facility ID. Enverus well production data from 2015 to 2022 were merged with GHGRP data based on facility ID using our well-to-facility dictionary, year, and sub-basin identifier, and further spatially joined with basin data. The GORs (scf bbl^{-1}) were calculated as the ratio of reported gas production (in standard cubic feet) to oil production (in barrels).

Subpart W GHGRP data for processing facilities was retrieved from the Query Builder[14] and filtered to include only onshore natural gas processing facilities. Latitude and longitude coordinates of the plants were obtained from the Query Builder[26] and merged with Subpart W data based on facility ID. Facilities where the methane mole fraction was zero or less than the carbon dioxide mole fraction were removed. Processing capacity and plant flow data were taken from the US Energy Information Administration (EIA) Natural Gas Processing Plants dataset[27], which was saved as a structured table containing plant capacity (MMcfd) and flow rates. These data were matched to GHGRP facilities using a nearest-neighbor approach based on spatial proximity. Facilities related to CO_2 recovery were excluded. The clean dataset was then spatially joined with counties and basins.

Prediction of gas composition

In this study, gas composition is predicted on a spatio-temporal grid for each basin and each database, including USGS data and data reported by production facilities to the GHGRP. The USGS gas composition data contain 15 gas component fractions (including 'C6+'), while the GHGRP (production) data include only methane and carbon dioxide fractions. Gas and oil production data are used as secondary variables to enhance predictions derived from spatio-temporal interpolation of known gas composition samples, effectively incorporating the GOR in the modeling process.

For each basin and each database, two sets of predictions are obtained: one from spatio-temporal kriging using known gas component fractions as inputs, and another from a non-linear model using gas and oil production volumes as inputs. These two sets are then combined. A visual description of the method is available in Supplementary Fig. S1.

The initial dataset is divided into training, validation, and testing subsets (See 4.6). First, spatio-temporal kriging is applied to the training set, while a non-linear model is trained on the same subset using gas and oil production volumes as inputs. The outputs of the two methods are then combined linearly, with inverse-variance weighting. The uncertainty of the prediction coming from the non-linear model is estimated by minimizing the MAE on a validation set (Section 4.2). Second, the training and validation sets are merged, and the two models (kriging and non-linear model) are re-optimized on the combined data. Finally, the method is evaluated on the held-out testing set.

We first recall the theoretical basis of ordinary kriging.

Ordinary kriging is a linear interpolation technique that provides unbiased estimates of minimal variance at unsampled points. In this study, we use spatio-temporal ordinary kriging to estimate gas component fractions. At the core of the kriging method is a tool called the "variogram", which essentially characterizes how correlated a pair of observations is, depending on their relative position. In dimension n, the variogram $\gamma(\mathbf{h})$ is defined as:

$$\gamma(\mathbf{h}) = \frac{1}{2}\mathrm{Var}\left[Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})\right], \tag{1}$$

where $\mathbf{h} \in \mathbb{R}^n$ is the spatial lag vector, $Z(\mathbf{x})$ is the value of the (stochastic) process Z at location $\mathbf{x} \in \mathbb{R}^n$. The variogram is estimated from observed data on a discrete set of lag vectors and interpolated using a mathematical model. Common models include spherical, exponential, and Gaussian models. In this work, we use an exponential model to ensure positive definiteness. The obtained variogram is noted $\hat{\gamma}(\mathbf{h})$.

Following the classical assumptions of ordinary kriging, we consider the stochastic process Z to be Gaussian, stationary (i.e., its statistical properties such as mean and variance are constant over the spatio-temporal domain) and isotropic (i.e., the variogram only depends on h, the norm of $\mathbf{h}$).

Given these assumptions, the kriging prediction $\hat{Z}(\mathbf{x})$ is defined as the "best" linear unbiased estimate of the true value at a location $\mathbf{x}$, meaning the linear and unbiased estimate with minimal variance.

We note the set of observed locations $\{\mathbf{x}_i\}_{i=1,\ldots,n}$. Let $\boldsymbol{\lambda} = \{\lambda_i\}_{i=1,\ldots,n}$ be the kriging weights, i.e.,

$$\hat{Z}(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{x}_i). \tag{2}$$

We can show that minimizing the variance of $\hat{Z}(\mathbf{x})$ is equivalent to solving the following linear system of equations:

$$\begin{gathered} \forall i \in \{1, \ldots, n\}, \quad \gamma(\mathbf{x}_i, \mathbf{x}) = \mu + \sum_{j=1}^{n} \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j), \\ \sum_{i=1}^{n} \lambda_i = 1. \end{gathered} \tag{3}$$

which can be solved efficiently with a linear solver. $\mu \in \mathbb{R}$ is a Lagrange multiplier coming from the unbiasedness constraint.

The kriging variance, which provides a measure of the uncertainty associated with the prediction, is then given by

$$\sigma^2(\mathbf{x}) = \gamma(\mathbf{x}, \mathbf{x}) - \sum_{i=1}^{n} \lambda_i \gamma(\mathbf{x}, \mathbf{x}_i) - \mu, \tag{4}$$

where $\sigma_K^2(\mathbf{x})$ is the kriging variance at location $\mathbf{x}$. This variance quantifies the expected squared difference between the predicted and actual values, taking into account the spatial correlation structure described by the variogram.

We next describe our implementation of ordinary kriging.

In this work, we rely on the source code of the GStatSim package from ref. 28. We modify the source code to extend its functionality from two to three dimensions (two spatial and one temporal).

We construct a 3D spatio-temporal grid for each production basin, with a 2 km × 2 km spatial resolution and 5-year temporal intervals. The grid extent is defined by the spatial and temporal distribution of oil and gas production data. Gas composition samples from USGS and GHGRP are then aggregated onto this grid, and interpolation is then performed at the grid-cell level.

To ensure that the quantities we interpolate follow Gaussian distributions, we apply a quantile transformer (with 500 quantiles) to the gas component fractions within each basin prior to interpolation. This transformation is monotonic while mapping the original data to a nearly standard normal distribution. The effect of this transformation is illustrated in Supplementary Fig. S6.

For each data source (USGS and GHGRP production), an empirical variogram is computed separately for each gas component and each

basin using the `scikit-gstat`[29] library (as implemented in ref. 28). The variogram model parameters (azimuth, nugget, range, and sill) are extracted for each case. If the computation fails, default parameters from stored average variograms are used (see Supplementary Section S3.3).

We now define the spatio-temporal scaling required to perform kriging.

To perform spatio-temporal kriging, we need to define a norm in space-time. We use the Euclidean norm and scale space with an "anisotropy factor" α. Given spatial coordinates (x_1, y_1) and (x_2, y_2), and temporal coordinates t_1 and t_2, the spatio-temporal Euclidean distance scaled with α is defined as:

$$\sqrt{\left(\frac{\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}}{\alpha}\right)^2+(t_2-t_1)^2} \tag{5}$$

To determine the optimal anisotropy factor α^*, we follow the method described in ref. 30 (implemented in `R`). We first estimate the temporal variogram γ_t by considering all pairs of measures made at the same location at different times. We linearly approximate this variogram around $\Delta t = 0$. The coefficients of the linear approximation, β_0 and β_1, are obtained by solving a least-squares problem:

$$\beta_0, \beta_1 = \arg\min_{b_0, b_1} \sum_{i,j} \left((b_0 + b_1 \cdot |t_i - t_j|) - \gamma_t(|t_i - t_j|)\right)^2, \tag{6}$$

where the sum is taken over the pair of observations made at the same spatial location. The optimal anisotropy factor α^* is then defined as:

$$\alpha^* = \arg\min_{\alpha} \sum_{i,j} \left(\left(\beta_0 + \beta_1 \cdot \frac{\| \mathbf{x}_i - \mathbf{x}_j \|}{\alpha}\right) - \gamma_s(\| \mathbf{x}_i - \mathbf{x}_j \|)\right)^2, \tag{7}$$

where the sum is taken over the pair of observations made at the same time, and γ_s is the spatial variogram associated with these measures.

The data used to compute the optimal value of the α^* are $C1$ molar fraction from the USGS data, which are considered the most representative as they are more numerous and are directly derived from well samples.

A value of $\alpha^* = 6.3$ m/day is obtained.

The optimal anisotropy factor is determined by minimizing the root mean square error (RMSE) in the spatial variogram fit, as shown in Supplementary Fig. S5. This value results in approximately isotropic gas composition variations.

The impact of different α values is shown in Supplementary Fig. S5, where methane fraction differences primarily follow a temporal trend when $\alpha = 1.0$, indicating weak spatial dependence, while at $\alpha = 40.0$, the differences are dominated by spatial variation, with minimal temporal correlation. We also show in Supplementary Fig. S4 the difference between the spatial variogram and its linear approximation using α, for different α values.

The non-linear model used to complement the kriging estimates is described below.

Our non-linear model uses gas and oil production volumes to provide a second estimate of the component fractions. We implement a multilayer perceptron (MLP) in PyTorch[31]. The input features to the MLP include logarithmically-transformed gas and oil production values, as well as the GOR. The model architecture consists of three hidden layers, each containing 16 neurons, with ReLU activation functions. We use a sigmoid activation function at the output layer.

The model is optimized using the Adam[32] optimizer with a batch size of 128, a learning rate of 10^{-4} for methane (C1) and 10^{-5} for ethane (C2). The number of training epochs depends on the dataset, with 2000 epochs for USGS data and 100 epochs for GHGRP data. The loss function minimized during training is the mean squared error (MSE). Additional subsurface or operational variables, such as formation depth, could also be incorporated when available. These may help capture geological influences on gas composition and further refine local estimates.

We explain below how the two predictions are combined using inverse-variance weighting.

The two predictions ($\hat{Z}$ from ordinary kriging, $\hat{Z}_{\text{nn}}$ from the neural network) are combined by inverse-variance weighting. The combined prediction $\hat{Z}_{\text{combined}}$ is therefore given by

$$\hat{Z}_{\text{combined}}(\mathbf{x}) = \frac{\hat{Z}(\mathbf{x})/\sigma^2(\mathbf{x}) + \hat{Z}_{\text{nn}}(\mathbf{x})/\sigma^2_{\text{nn}}}{1/\sigma^2(\mathbf{x}) + 1/\sigma^2_{\text{nn}}} \tag{8}$$

The variance for the kriging estimate, $\sigma^2(\mathbf{x})$, is provided by Equation (4). The variance of the neural network estimate, σ^2_{nn}, is estimated by minimizing the MAE on the validation set. The resulting value of σ_{nn} provides information on the uncertainty associated with the neural network predictions, with higher values indicating lower confidence in neural network estimates relative to kriging predictions.

Unified gas composition dataset

To support estimation of recent methane leak impacts, we construct a unified gas composition dataset representing the 2016–2021 period. We collapse our 3D spatio-temporal prediction grid (two spatial dimensions and one temporal dimension) into a 2D spatial grid by selecting, for each location and gas component, the prediction from the 5-year grid cell whose final year matches the most recent available data—2014 for the USGS dataset, and 2021 for data reported by production facility to the GHGRP. Methane and carbon dioxide fractions, along with their associated SEs, are taken from GHGRP results and used to replace the corresponding values from USGS data. The remaining component fractions are summed and normalized so that all components add to 100%, keeping the C1 and CO_2 values unchanged. Uncertainties are propagated using SE propagation formulas.

The interpolated composition map corresponding to this time period underlies the methane loss estimates and basin-wide averages presented in Section 2. We will release this recommended dataset.

Basin aggregates and uncertainties

For each basin and data source, an aggregate value of each component fraction is computed by calculating a weighted average within the basin. The weighting variable for USGS and GHGRP production is the gas production volumes. For GHGRP processing, it is the plant flow obtained from ref. 27.

For USGS and GHGRP production data, uncertainties are estimated using Monte Carlo sampling.

The aggregate prediction across a basin b can be expressed as

$$\hat{Z}_b = \sum_i w_i \hat{Z}(\mathbf{x}_i, t_i) \tag{9}$$

where the sum is taken over all known production sites in the basin b. Since the predictions $\hat{Z}(\mathbf{x}_i, t_i)$ are derived from kriging, they are spatially and temporally correlated. Treating their variances as independent would underestimate the uncertainty by neglecting the covariance terms.

To accurately capture the spatio-temporal correlations in our uncertainty estimation, we use Monte Carlo sampling, i.e., we simulate multiple realizations from the Gaussian process. By construction, these simulations incorporate the covariance structure. The variance of the aggregate on the basin b is then given by:

$$\text{Var}(\hat{Z}_b) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{Z}_b^{(i)} - \bar{Z}_b\right)^2, \quad \text{where } \bar{Z}_b = \frac{1}{n}\sum_{i=1}^{n}\hat{Z}_b^{(i)}. \tag{10}$$

and the sum is taken over the n simulations. This approach provides an unbiased estimate of the SE for the aggregate:

$$\mathrm{SE}(\hat{Z}_b) = \frac{\sqrt{\mathrm{Var}(\hat{Z}_b)}}{\sqrt{n}} \tag{11}$$

For GHGRP processing data, since we do not make use of the kriging method, standard deviations for individual reported component fractions are not available. The uncertainties of the aggregated component fractions are estimated using the bootstrap method. This method is a resampling technique that approximates the distribution of an estimator by repeatedly sampling the data with replacement.

The process for computing uncertainties using bootstrap involves generating n bootstrap samples from the original dataset. We choose $n = 1000$ to ensure a stable estimate of the SE. Each bootstrap sample is created by randomly sampling with replacement $\hat{Z}_j$ from $\{\hat{Z}_i\}$, where each bootstrap sample has the same size as the original dataset. This ensures that each resample reflects the full variability of the original data. For each bootstrap sample, the weighted mean $\hat{Z}_b^{(i)}$ of the component fraction for the basin is computed using the weights w:

$$\hat{Z}_b^{(i)} = \sum_j w_j \hat{Z}_j. \tag{12}$$

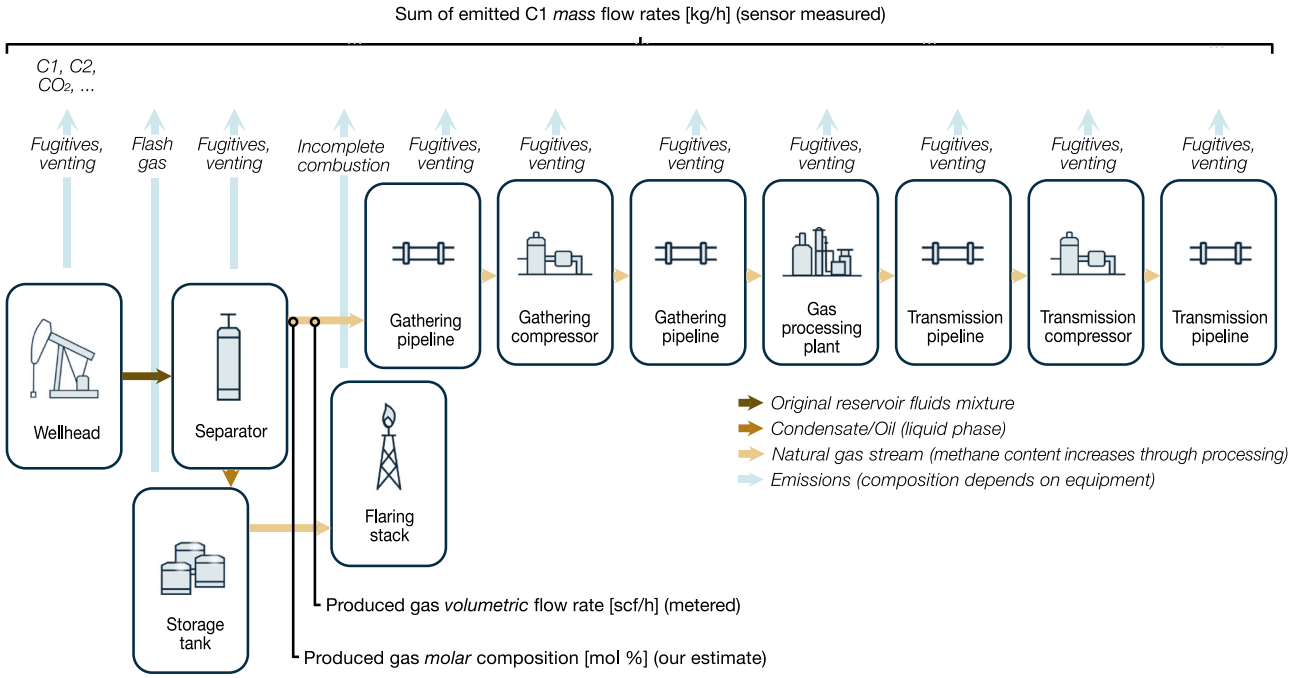
The SE is then given by:

$$\mathrm{SE}(\hat{Z}_b) = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\hat{Z}_b^{(i)} - \bar{Z}_b\right)^2}, \quad \text{where } \bar{Z}_b = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_b^{(i)}. \tag{13}$$

Basin-level production-normalized methane loss rate

We compute a dimensionless metric, the production-normalized methane loss rate ρ, to quantify how much of the methane extracted from the ground is ultimately lost to the atmosphere through fugitive emissions, venting, or incomplete flare combustion. Because direct measurements of extracted methane mass are not available, the standard approach—used in prior studies such as Sherwin et al.[5]—relies on two inputs: the volume of gas metered at the production site and independent measurements of methane emissions. These studies typically lack information on the composition of the produced gas, which is required to convert volumetric gas production into methane mass. This is the gap our analysis addresses. Fig. 5 shows where and how our composition estimates are applied. We emphasize that our estimates of produced gas composition are not suitable for estimating the full climate or air quality impacts of emissions: such assessments would require modeling how gas composition evolves along the supply chain

(a) Schematic of gas-phase emission sources along the oil and gas supply chain



(b) Computation of production-normalized methane loss rate

$$\textit{Defined as}\ \frac{\text{Emitted methane}}{\text{Produced methane}} = \textit{Computed as}\ \frac{\text{Sum of emitted C1 } \textit{mass} \text{ flow rates [kg/h] (sensor}}{\text{Produced gas } \textit{mass} \text{ C1 fraction [kg \%] (our estimate)} \times \text{Produced gas } \textit{mass} \text{ flow rate [kg/h]}}$$

Fig. 5 | Context for interpreting our estimates of natural gas composition and their application to methane loss rate calculation. a Schematic of gas-phase emission sources along the oil and gas supply chain, including venting, fugitives, flash gas, and flaring. Combustion-related CO_2 emissions from fully combusted fuel are not included. We show the location in the supply chain (the separator outlet) where our method estimates the produced gas composition. This estimate differs from the composition of emitted gases (e.g., flash gas, fugitives) and from that of processed gas transmitted to end users. **b** Definition and computation of the production-normalized methane loss rate, based on the ratio of emitted C1 to produced C1. The input quantities shown in (**a**) are expressed in different physical bases: the produced gas flow rate is volumetric (scf/h), the gas composition is estimated as a molar fraction, and the emitted methane rate is in mass units (kg/h). For the loss rate calculation shown in (**b**), all quantities are converted to a consistent mass basis.

and how thermodynamic processes (e.g., flashing from the liquid phase) alter the composition of emission streams. These additional assumptions are outside the scope of our methane loss rate calculation.

For each basin studied in ref. 5, we compute the production-normalized methane loss rate as the ratio of aggregate methane emissions to the methane contained in produced gas: $\hat{\rho} = \frac{\hat{\mathcal{E}}_{C1}}{\mathcal{P}_{C1}}$, where $\hat{\mathcal{E}}_{C1}$ is the estimated aggregate methane emissions (kg h^{-1}) and $\mathcal{P}_{C1}$ is the methane production rate (kg h^{-1}). Writing the methane share of total produced gas on a molar basis gives $\frac{\mathcal{P}_{C1}}{\mathcal{P}_{\text{total}}} = \frac{M_{C1}\hat{Z}_{C1}}{\sum_i M_i \hat{Z}_i}$, where M_i is the molar mass of component i (kg mol^{-1}) and $\hat{Z}_i$ its estimated molar fraction. Substituting yields

$$\hat{\rho} = \frac{\hat{\mathcal{E}}_{C1}}{\mathcal{P}_{\text{total}} \cdot \frac{M_{C1}\hat{Z}_{C1}}{\sum_i M_i \hat{Z}_i}}. \tag{14}$$

Note that $\frac{M_{C1}\cdot\hat{Z}_{C1}}{\sum_i (M_i\cdot\hat{Z}_i)}$ is the mass fraction of $C1$, not its molar fraction.

The uncertainty in $\hat{\rho}$, denoted as SE_ρ, is calculated by propagating the uncertainties in $\hat{\mathcal{E}}_{C1}$, $\hat{Z}_{C1}$, and $\hat{Z}_i$:

$$\text{SE}_\rho = \hat{\rho} \cdot \sqrt{\left(\frac{\text{SE}_{\hat{\mathcal{E}}_{C1}}}{\hat{\mathcal{E}}_{C1}}\right)^2 + \left(\frac{\text{SE}_{\hat{Z}_{C1}}}{\hat{Z}_{C1}}\right)^2 + \sum_i \left(\frac{\text{SE}_{\hat{Z}_i}}{\hat{Z}_i}\right)^2}. \tag{15}$$

Evaluation method

For each data source (USGS and GHGRP production), a disjoint pair of validation and testing sets is constructed per basin. To define these sets, we apply k-means clustering among the measurement locations, with $k = 4$. Each centroid is then used to define a cluster of points. For each centroid $\mathbf{x}_k$ ($k = 1, \ldots, 4$), we consider the set of points located, at most, at a distance r of $\mathbf{x}_k$, where

$$r = \kappa \times \min(x_{\text{range}}, y_{\text{range}}). \tag{16}$$

x_{range} and y_{range} represent the spatial extent of the dataset, and κ is a parameter set to 0.15. To maximize separation, we first choose a validation cluster and then select the farthest available cluster for testing.

This approach follows the standard machine learning practice of block cross-validation, which reduces data leakage and provides a more realistic evaluation of model generalization to unseen spatial regions. Using cylindrical rather than fully spherical (3D ball) held-out regions ensures that distances to the nearest training points retain a meaningful physical interpretation. In particular, the spatial distance to the closest training point can be directly used to assess extrapolation performance, without introducing artificial distortions from temporal scaling effects.

To evaluate our method, we calculate two performance metrics: MAE and RMSE on the testing set. These metrics are computed for both our method and baseline approaches, including ordinary kriging only and the nearest-neighbor algorithm. When comparing our method to the ordinary kriging baseline, we evalute the improvement brought by the processing of gas and oil data by the non-linear model. We define the relative improvement by $\frac{X-Y}{Y}$, where X is the error obtained with the non-linear model and Y the error obtained without. We also evaluate the uncertainty of our evaluation by reporting SEs.

Data availability

The processed datasets generated in this study have been deposited in Zenodo under accession code[33]. The interactive online tool—Interactive U.S. map of produced gas composition—is publicly available at ref. 34. Raw gas composition datasets used in this study were obtained from publicly available sources (USGS and EPA GHGRP) as described in the Methods. The raw Enverus oil and gas production data used in this study are available under restricted access due to commercial license restrictions. Access can be obtained through a paid subscription from Enverus, and equivalent well-level production data are also available from state regulatory agencies.

Code availability

The code used to process raw datasets, perform geostatistical modeling, and generate figures in this study has been deposited in Zenodo under accession code[35].

References

1. Intergovernmental Panel on Climate Change (IPCC). Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (eds. Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S. et al.) (Cambridge University Press, Cambridge, UK and New York, NY, USA). https://doi.org/10.1017/9781009157896 (2021).
2. International Energy Agency. *Global Methane Tracker 2024*. International Energy Agency, Paris, France https://www.iea.org/reports/global-methane-tracker-2024 (2024).
3. Brandt, A. R. et al. Methane leaks from North American natural gas systems. *Science* **343**, 733–735 (2014).
4. Alvarez, R. A. et al. Assessment of methane emissions from the U.S. oil and gas supply chain. *Science* **361**, 186–188 (2018).
5. Sherwin, E. D. et al. U.S. oil and gas system emissions from nearly one million aerial site measurements. *Nature* **627**, 328–334 (2024).
6. World Bank. *Methane Explained*. World Bank, Washington, DC, USA https://www.worldbank.org/en/programs/gasflaringreduction/methane-explained (2024).
7. Cardoso-Saldaña, F., Pierce, K., Chen, Q., Kimura, Y. & Allen, D. T. A searchable database for prediction of emission compositions from upstream oil and gas sources. *Environ. Sci. Technol.* **55**, 3210–3218 (2021).
8. Brandt, A. et al. Energy intensity and greenhouse gas emissions from crude oil production in the Bakken Formation: input data and analysis methods. Argonne National Laboratory, U.S. Department of Energy, Argonne, IL, USA https://greet.anl.gov/publication-bakken-oil (2015).
9. Barkley, Z. R. et al. Analysis of oil and gas ethane and methane emissions in the south-central and eastern United States using four seasons of continuous aircraft ethane measurements. *J. Geophys. Res. Atmos.* **126**, e2020JD034194 (2021).
10. Brennan, S.T., Rivera, J.D., Creitz, R.L., Varela, B.A. & Park, A.J. Natural gas compositional analyses dataset of gases from United States wells. U.S. Geological Survey data release, https://doi.org/10.5066/P9TR93E3 (2021).
11. U.S. Environmental Protection Agency (EPA). *Greenhouse Gas Reporting Program*. U.S. Environmental Protection Agency, Washington, DC, USA https://www.epa.gov/ghgreporting (2023).
12. Railroad Commission of Texas. *Oil and Gas Well Records*. Railroad Commission of Texas, Austin, TX, USA https://rrcsearch3.neubus.com/esd3-rrc/index.php?_module_=esd&_action_=keysearch&profile=17 (2025).
13. Zhang, T., Sun, X., Milliken, K. L., Ruppel, S. C. & Enriquez, D. Empirical relationship between gas composition and thermal maturity in Eagle Ford Shale, South Texas. *AAPG Bull.* **101**, 1277–1307 (2017).
14. U.S. Environmental Protection Agency (EPA). *Greenhouse Gas Reporting Program (GHGRP) Query Builder: Subpart W - Petroleum and Natural Gas Systems (Reporting Years 2015-2023)*. U.S. Environmental Protection Agency, Washington, DC, USA. https://enviro.epa.gov/query-builder/ghg/SUBPART. (2025).

15. Enverus. *Enverus homepage*. Enverus, Austin, TX, USA https://www.enverus.com (2023).
16. Doe, J. & Smith, A. Gas-oil ratio correlations with hydrocarbon composition in natural gas reservoirs. *J. Mar. Sci. Eng.* **11**, 1891 (2023).
17. U.S. Geological Survey. *AAPG Geologic Provinces*. U.S. Geological Survey, Reston, VA, USA https://ngmdb.usgs.gov/Geolex/stratres/provinces (2025).
18. Collins, W. J., Derwent, R. G., Johnson, C. E. & Stevenson, D. S. The oxidation of organic compounds in the troposphere and their global warming potentials. *Climatic Change* **52**, 453–479 (2002).
19. Fry, M. M., Shindell, D. T. & Wagner, L. E. The influence of ozone precursor emissions from four world regions on tropospheric composition and radiative climate forcing. *J. Geophys. Res. Atmos.* **119**, 1390–1412 (2014).
20. Hodnebrog, Ø., Dalsøren, S. B. & Myhre, G. Lifetimes, direct and indirect radiative forcing, and global warming potentials of ethane, propane and butane. *Atmos. Sci. Lett.* **19**, e804 (2018).
21. Radunsky, K. & Gillenwater, M. Chapter 7: Precursors and indirect emissions. In *2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Volume 1: General Guidance and Reporting* (Intergovernmental Panel on Climate Change, Geneva, Switzerland). https://www.ipcc-nggip.iges.or.jp/public/2019rf/pdf/1_Volume1/19R_V1_Ch07_Precursors_Indirect.pdf (2019).
22. U.S. Census Bureau. *Cartographic Boundary Files*. U.S. Census Bureau, Washington, DC, USA https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html (2025).
23. U.S. Environmental Protection Agency (EPA). *Geologic Basin Boundaries (GHGRP GIS Layer)*. U.S. Environmental Protection Agency, Washington, DC, USA https://catalog.data.gov/dataset/geologic-basin-boundaries-basins_ghgrp-gis-layer9 (2025).
24. American Gas Association, *BSR-B109.5 ANSI-PR Draft: Standard Conditions for Gas Volume* Available at: https://www.aga.org/wp-content/uploads/2023/08/BSR-B109.5-ANSI-PR-Draft-rev.pdf (2023).
25. West Virginia Department of Environmental Protection (WVDEP). *API Well Number Explanation*. West Virginia Department of Environmental Protection, Charleston, WV, USA https://dep.wv.gov/oil-and-gas/GI/API_Explanation/Pages/default.aspx (2024).
26. U.S. Environmental Protection Agency (EPA). *GHGRP Facility Information Query Builder*. U.S. Environmental Protection Agency, Washington, DC, USA https://enviro.epa.gov/query-builder/ghg/FACILITY. (2025).
27. U.S. Energy Information Administration (EIA). *Natural Gas Processing Plants*. U.S. Energy Information Administration, Washington, DC, USA https://atlas.eia.gov/datasets/eia::natural-gas-processing-plants-2/explore (2025).
28. MacKie, E. et al. GStatSim v1.0: a Python package for geostatistical interpolation and conditional simulation. *Geosci. Model Dev.* **16**, 3765–3783 (2023).
29. Mälicke, M. SciKit-GStat 1.0: a SciPy-flavored geostatistical variogram estimation toolbox written in Python. *Geosci. Model Dev.* **15**, 2505–2532 (2022).
30. Pebesma, E. & Graeler, B. *gstat: Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation*. R package version 2.1-2 https://github.com/r-spatial/gstat/ (2024).
31. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (Curran Associates, Inc., 2019).
32. Kingma, D.P. & Ba, J. Adam: a method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* https://arxiv.org/abs/1412.6980 (2015).
33. Burdeau, P. M. et al. *Interactive U.S. map of produced gas composition (processed datasets)*. Zenodo v1.0.0 https://doi.org/10.5281/zenodo.17246944 (2025).
34. Burdeau, P.M., Sherwin, E.D., Biraud, S.C., Berman, E.S.F. & Brandt, A.R. Interactive U.S. map of produced gas composition. Available at: https://pburdeau.github.io/us_map_gas_composition/ (2025).
35. Burdeau, P. M., Sherwin, E. D., Biraud, S. C., Berman, E. S. F. & Brandt, A. R. *Code for: High-resolution national mapping of natural gas composition substantially updates methane leakage impacts*. Zenodo v1.0.2. https://doi.org/10.5281/zenodo.17254285 (2025).

Acknowledgements

This work was partially supported by the by the Stanford Natural Gas Initiative, an industry consortium that supports independent research at Stanford University. This work was partially supported by the US Department of Energy Office of Fossil Energy and Carbon Management under Award Number DE-FE0032310. This work was partially supported by the California Energy Commission (SUMMATION project, agreement number PIR-17-015). It does not necessarily represent the views of the Energy Commission, its employees, or the State of California. The Energy Commission, the State of California, its employees, contractors, and subcontractors make no warranty, express or implied, and assume no legal liability for the information in this report; nor does any party represent that the uses of this information will not infringe upon privately owned rights. This paper has not been approved or disapproved by the California Energy Commission, nor has the California Energy Commission passed upon the accuracy or adequacy of the information in this paper. This manuscript has been authored by authors at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges, that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes. Additional support for this work was provided by the Archie Initiative of Aramco.

Author contributions

P.M.B., E.D.S., and A.R.B. conceptualized the study. P.M.B., E.D.S., and A.R.B. conducted the formal analysis and developed the methodology. Supervision was provided by A.R.B. and E.D.S. Funding was acquired by E.S.F.B., S.C.B., A.R.B., and E.D.S. Data curation was handled by P.M.B. P.M.B. wrote the manuscript with feedback from all the authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-66465-6.

Correspondence and requests for materials should be addressed to Philippine M. Burdeau.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.