

Unlocking credible space-based methane sensing through a year-long single-blind test

Adam Brandt

abrandt@stanford.edu

Stanford <https://orcid.org/0000-0002-2528-1473>

Frances Reuland

Stanford University <https://orcid.org/0000-0001-9463-7931>

Taylor Adams

University of Michigan

Evan Sherwin

Lawrence Berkeley National Laboratory <https://orcid.org/0000-0003-2180-4297>

Sahar El Abbadi

University of Iowa

Eric Kort

Max Plank Institute of Chemistry

Physical Sciences - Article

Keywords:

Posted Date: April 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9110475/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Unlocking credible space-based methane sensing through a year-long single-blind test

Frances E. Reuland¹, Taylor Adams², Evan D. Sherwin³, Sahar H. El Abbadi^{3,4}, Eric A. Kort^{2,5}, Adam R. Brandt¹

¹ Department of Energy Science & Engineering, Stanford University, Stanford, California, 94305, United States

² Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, 48109, United States

³ Energy & Biosciences Institute, University of California-Berkeley, Berkeley, California, 94720, United States

⁴ Department of Civil and Environmental Engineering, University of Iowa, Iowa City, 52242, United States

⁵ Atmospheric Chemistry Department, Max Planck Institute for Chemistry, 55128 Mainz, Germany

Correspondence to: Adam R. Brandt (abrandt@stanford.edu)

Space-based remote sensing now allows detection and monitoring of large, concentrated methane sources worldwide. However, independent evaluation of instruments and algorithms is very limited, and new satellites and AI techniques have not been tested under controlled, single-blind conditions. To address this gap, we conducted a year-long, single-blind testing campaign >15 times larger than prior tests. Testing included 746 events (481 large-scale methane releases and 265 blank controls). We released metered methane quantities in Arizona, USA coincident with overpasses from 12 global satellites (government, commercial, and non-profit). This is the first evaluation of prominent methane instruments EMIT, Tanager-1 (Carbon Mapper), and GHOS. Fifteen analysis teams from nine countries generated >2,250 estimates. Across all sensors and teams, 70.6% correctly identified the presence or absence of emissions. Performance is heterogeneous: “built for purpose” instruments designed to target methane demonstrate excellent performance, with strong detection rates (92-100% true positive rate, 0-11% false positive rate) and quantification accuracy (unbiased parity slopes 0.83-1.00, 44% mean absolute percent error, MAE). Instruments repurposed for methane sensing have more variable performance (detection rates 62-88%, MAE 46%); still, many teams demonstrate significant skill. Broad land scanners performed poorly but have consistent global coverage and vast historical archives, motivating additional work. AI-based methods show some promise, but caution is warranted; 79% of false positives emerge from these methods. Ground-measured cloud cover exceeding 20% caused strong decline in performance across all instrument classes, raising concerns for use in some locations. Our findings show that high-quality methane measurements from space can allow accurate detection and unbiased quantification, initiating a new era in international greenhouse gas mitigation.

Main

40 Methane is a potent short-lived climate pollutant¹. Remote sensing by satellites has emerged as a new tool for detecting and quantifying methane emissions, in particular for intense, localized anthropogenic sources. A global network of methane detecting satellites and analysis teams has emerged in the last decade, and the prominence and importance of these efforts grows. Initial proof-of-concept observations used relatively coarse instruments to detect methane hotspots and ultra-emitting sources (>50 metric tons per hour (t/h))^{2,3}. New satellites with diverse capabilities have been launched, while other existing systems have been repurposed towards methane observation⁴. Today, more than ten different public, public-private, and commercial satellite systems are used to find large plumes (>0.1 -1 t/h) globally⁵⁻¹³. These instruments range from multispectral government satellites with broad Earth-observing missions (e.g. Sentinel-2) to commercial constellations of multiple satellites that offer GHG monitoring as a service¹⁴.

50 In 2024 a step-change in capability occurred with the launch of two new satellites, Carbon Mapper's Tanager-1 and Environmental Defense Fund's MethaneSat. These are the first systems that are both designed specifically to detect methane and are committed to releasing large-scale datasets into the public domain, building on the coarse-resolution monitoring capabilities of the European Space Agency's TROPOMI instrument. Tanager-1 is a high-resolution hyperspectral sensor, which detected more than 2,400 methane plumes globally in its first seven months of operation¹⁵. Carbon Mapper is expected to scale up its constellation with at least three more instruments to achieve daily methane monitoring of high-emitting areas¹⁵. MethaneSat, in operation from March 2024-June 2025, had finer spectral resolution and coarser spatial resolution and was designed for wide-area total emissions estimates¹⁶.

60 The proliferation of observing systems has been accompanied by advances in data processing methods. Recent studies highlight novel and increasingly automated techniques, including automated plume detection and segmentation processes¹⁷. There is also a trend toward AI-driven approaches, typically involving the use of synthetic plumes to train plume detection algorithms¹⁸⁻²⁰.

65 Accordingly, data availability and use have expanded rapidly. Thousands of detected plumes are reported through public data portals, including those managed by the UN International Methane Emissions Observatory (IMEO), Carbon Mapper, and NASA-JPL, and additional subscription-based tools are available for purchase²¹⁻²³. These data

will enable a wide range of regulatory, commercial, and advocacy applications. Satellite observations have been
70 integrated into international leak alert systems, national and sub-national methane policies, and emissions inventory
reconciliation processes^{24–26}. For example, satellite observations will form an important component of the EU
proposed import climate standards for natural gas²⁷.

The global importance of space-based methane detection systems necessitates robust and independent
75 characterization of these systems' performance and capabilities. Single-blind controlled release testing is the
current standard for benchmarking the capabilities of point-source methane sensing systems^{28–32}. Studies
characterizing airplane-based sensing systems have typically gathered hundreds of data points^{28,30}. By contrast,
only a small number of blind satellite detection data points exist in peer-reviewed literature. For example,
combining observations from the three published single-blind validation studies, just nine and five ground-
80 metered point-source releases are available for Sentinel-2 (S2) and GHGSat, respectively^{29,33,34}. Some of the
newest and most prominent systems, including Carbon Mapper, EMIT, and GHOS_t, have not been evaluated in
any blind tests. Furthermore, new AI-based detection algorithms have been developed in recent years but have
not been independently tested using single-blind protocols.

85 Given the relative lack of satellite-scale controlled release testing, many studies rely on theoretical modeling or
multi-instrument comparisons for characterizing detection capabilities and quantifying large emission
sources^{5,10,35–37}. These approaches can approximate instrument performance and are useful for assessing inter-
method variability, but they fall short of offering independent, blind validation.

90 Two relatively small trial satellite validation campaigns were previously conducted by Sherwin et al^{29,33}. The first
campaign lasted 19 days, assessing five analysis teams and five satellites (GHGSat, Landsat, Sentinel-2,
PRISMA, and Worldview-3). That study resulted in 16 successful releases in which satellite tasking was also
performed, and a total of 49 blind estimates were made across all analysis teams. A later campaign expanded in
scope, testing nine satellites over 52 days (previously listed satellites, plus EnMAP, Gaofen 5, Huanjing 2, and
95 Ziyuan 1). That experiment yielded 132 blind estimates. These campaigns highlighted several findings: inter-
platform variability, the influence of atmospheric and surface conditions, and the need for more extensive testing.
Remote sensing experts from across the world in academia, government, and industry have called for larger
experiments to improve statistical confidence in capabilities, a wider range of observing conditions, and inclusion
of a larger number of no-emissions cases to test for false positive and true negative results^{33,38,39}. A challenge with

100 these studies is the large monetary cost and complex logistics associated with performing many large methane releases precisely timed with satellite overpasses.

The present study directly addresses the validation gap by conducting the largest-to-date single-blind controlled methane release experiment for satellite systems. This campaign expands the scope of prior efforts via: (1) a broader
105 set of 12 participating satellites, including the first-ever single-blind tests of Tanager-1 (Carbon Mapper), the GHOS_t constellation, and EMIT, alongside a greatly expanded GHGSat constellation; (2) a larger cohort of 15 analysis teams including variable use of AI tools; (3) more than an order of magnitude more releases (481 releases) and single-blind estimates (2,259); (4) vastly improved weather monitoring for generating an open R&D dataset for future algorithm training; (5) inclusion for the first time of a large number of no-release controls (265 zeros) to
110 test for false positives; (6) a complete year of data, allowing for more variety in weather conditions and sun angles.

Single-blind testing: overview

We generated a novel dataset comprising 746 valid scenes at a fixed location near Casa Grande, Arizona, USA (32.8128, -111.7858). Releases occurred between January 1 and December 31, 2025. A total of 481 controlled
115 methane releases were conducted. Release rates ranged from 9.5 [9.4, 9.6, 95% CI on mean release flow rate] kg CH₄/h to 1,183.7 [1,165.1, 1,202.4] kg CH₄/h. These rates are in the range of point-source emissions observed from real-world infrastructure, such as landfills and oil and gas production and processing sites⁴⁰. In addition to releases, the experiment team created 265 intentional blank controls where no methane was released during a satellite overpass. The experiment was designed as a four-phase, single-blind study.

120 The releases and non-releases coincide with overpasses from 12 satellite instruments or multi-instrument constellations (henceforth simply “satellites”). The satellites included in this study vary in spatial and spectral resolution, swath width, revisit frequency, and methane detection sensitivity (Table 1). Data from the satellites were analyzed independently by 15 teams representing a broad global consortium of academic, government, non-
125 profit, and for-profit institutions (see SI Table S2 for satellite-team pairings). Some instruments had only a single team performing analysis on the data, while other instrument data were analyzed by as many as ten teams. Henceforth, results are shown for 11 satellites, as no results were submitted for the small number of MethaneSat releases that were performed before its mission ended prematurely in June 2025.

130 We group the instruments into three broad categories used throughout the analysis (Table 1). Group A instruments are purpose-built point-source methane detection instruments with high spectral and spatial resolutions. Group B instruments are repurposed hyperspectral and/or high spatial resolution instruments that were originally launched for broader observing purposes but are used to detect methane. Group C are multi-band global scanning instruments, with much coarser spectral resolution but wider swaths and frequent global revisits. Groupings were made *a priori* based on instrument design and mission intent, not on observed performance. The physics of remote sensing suggests that detection should be most successful for Group A instruments and least successful for Group C. It is therefore appropriate to group instruments in this manner to avoid comparing disparate technologies, as prior expectations for each class differ.

140 Analysis teams included nonprofit Carbon Mapper; research institutions NASA JPL, the International Methane Emissions Observatory (IMEO), and SRON (Space Research Organization Netherlands); academic groups Universitat Politècnica de València (Spain), Nanjing University (China), and IUP-UB (University of Bremen, Germany); and commercial organizations GHGSat (Canada), Google (U.S.A.), Kayrros (France), Orbio Earth (Norway), Arka (U.S.A.), Orbital Sidekick (U.S.A), Geolabe (U.S.A), and Momentick (Israel). Teams were responsible for sourcing imagery and were permitted to analyze any satellite data to which they had access. In many cases, multiple teams processed data from the same satellite using independent algorithms and workflows.

Gas release and data reporting procedures were based on established methods for controlled release validation^{28,33}. Briefly, large volumes of methane were released from a Coriolis-metered release stack supplied by compressed natural gas (CNG) trucks (see “Methods”). Testing was divided into four phases (each three months). During all phases, analysis teams were provided only the stack location and a broad range of possible emission rates (3–1,500 kilograms per hour (kg/h)). Teams were blind to all other ground-truth data, including whether a release was occurring during each satellite overpass, precise emission timing and flow rate, and on-site meteorological conditions. Following each phase, teams had a one-month window to submit their emission estimates. After the submission period closed, ground-truth data for that phase were unblinded as the next phase was underway. Any estimates received after unblinding were considered invalid.

On-site instrumentation captured high-resolution (1 Hz) time series of gas flow rates for every release. Gas composition was obtained daily from a cooperating supply pipeline. A suite of meteorological measurements was also recorded, including multi-height wind speed and direction [m/s, degrees], temperature [C], humidity [% rel.

humidity], soil moisture [m^3 water/ m^3 soil], solar irradiance [W/m^2 GHI], cloud cover [% cloud cover], and 360-degree sky imagery (see Methods).

165 **Table 1. Participating satellites, listed by grouping.** Targeted coverage refers to an instrument that must be tasked
to acquire an observation over a given location, including the controlled test site. Instruments with global coverage
regularly scan most of Earth’s surface. “Sparse” is used to describe where satellite overpass frequency is not
publicly known and could not be inferred from infrequent tasking over the study site. EMIT overpass frequency is
irregular because it is mounted on the ISS and has an unpredictable non-polar orbit, thus limiting the number of
170 possible controlled release observations.

Satellite	Grouping	Coverage	Constella- tion size	Swath [km]	Pixel size [m]	~SWIR spectral resolution [nm]	~Revisit time to site	Data availability
GHGSat ⁴¹	A	Targeted	14	12	25	<1nm	1-2 d	Commercial
Tanager-1 (TG1) ⁴²	A	Targeted	1	19	30	5	7 d	Public/Commercial*
EMIT ⁴³	B	Global	1	75	60	7.4	irregular	Public
EnMAP ^{44,45}	B	Targeted	1	30	30	7.4-10	4-5 d	Public
GHOST ⁴⁶	B	Targeted	4	4	8		2-3 d	Commercial
Gaofen 5 (GF5) ⁴⁷	B	Targeted	2	60	30	5	sparse	Government
PRISMA ⁴⁸	B	Targeted	1	30	30	10-12	7 d	Public
WorldView- 3 (WV3)** ⁴⁹	B	Targeted	1	13	3.7	50	7 d	Commercial
Ziyuan-1 (ZY1) ⁵⁰	B	Targeted	2	60	30	20	sparse	Government
Landsat 8/9 (LS) ⁵¹	C	Global	2	185	30	80/180	8 d	Public
Sentinel-2 (S2) ¹⁴	C	Global	2	290	20	90/180	5 d	Public

* L2 products and above are made publicly available 30 days after acquisition.
** Note on WorldView-3 grouping: while placed in Group B, WorldView-3 is unique. It is the only multispectral instrument
in this group, and its spatial resolution is significantly finer (2-10x) than most others in the group. It is categorized here with
other high-resolution, taskable instruments to avoid a single-instrument category.

Results

Fifteen teams analysed data from between one and seven satellites each. Out of 481 methane releases, 393 (81.7%) resulted in successful image acquisitions, producing the data required to detect a methane emission. The remaining 88 (18.3%) were unsuccessfully imaged. In the 265 blank events (no methane released), successful imagery was achieved in 233 (87.9%) instances. Unsuccessful imaging occurs most frequently when the release site is deprioritized against a competing area of interest by instrument scheduling algorithms (e.g., commercial conflicts in the case of private instruments). Other tasking, communication, or software errors can occur.

Combining results from the 393 release and 233 non-release events with imagery, a total of 2,259 responses were recorded. Analysis teams filtered out 488 (21.6%) of these responses, with 441 (90.4% of the filtered total) of the analyst-filtered releases resulting from excessive cloud cover in the image (see SI Table S5 for filtered counts by team). The team of Stanford and University of Michigan researchers (henceforth, the experiment team) removed an additional six events corresponding to 21 team responses due to improper release timing. Of the 1,750 non-filtered responses, there were 689 true positives (TPs), 547 true negatives (TNs), 443 false negatives (FNs), and 71 false positives (FPs). Thus, nearly three-quarters (70.6%) of the non-filtered responses accurately identified the presence or absence of a release (TP+TN). In seven TP cases, a team reported a successful detection but opted not to estimate source flow rate due to data quality concerns.

True positive rates (TPRs), defined as the proportion of non-zero releases that a team correctly identifies as a release, varied significantly (Figure 1). Group A had TPRs of 92-100%. Group B varied from 62-88%, while Group C TPRs were near 29-37%. This should not be taken as evidence of the general ability of Group C instruments to see very large (>1500 kg/h) emission sources, as our equipment can only release volumes at the expected low end of the detection range for these instruments. Also, the distribution of release rates was instrument-specific to avoid wasting excessive gas on unobservable plumes. Distributions for each instrument were informed by prior controlled release study performance, empirical measurements of real-world plumes, and input from participating teams regarding the capabilities of each platform (see Methods).

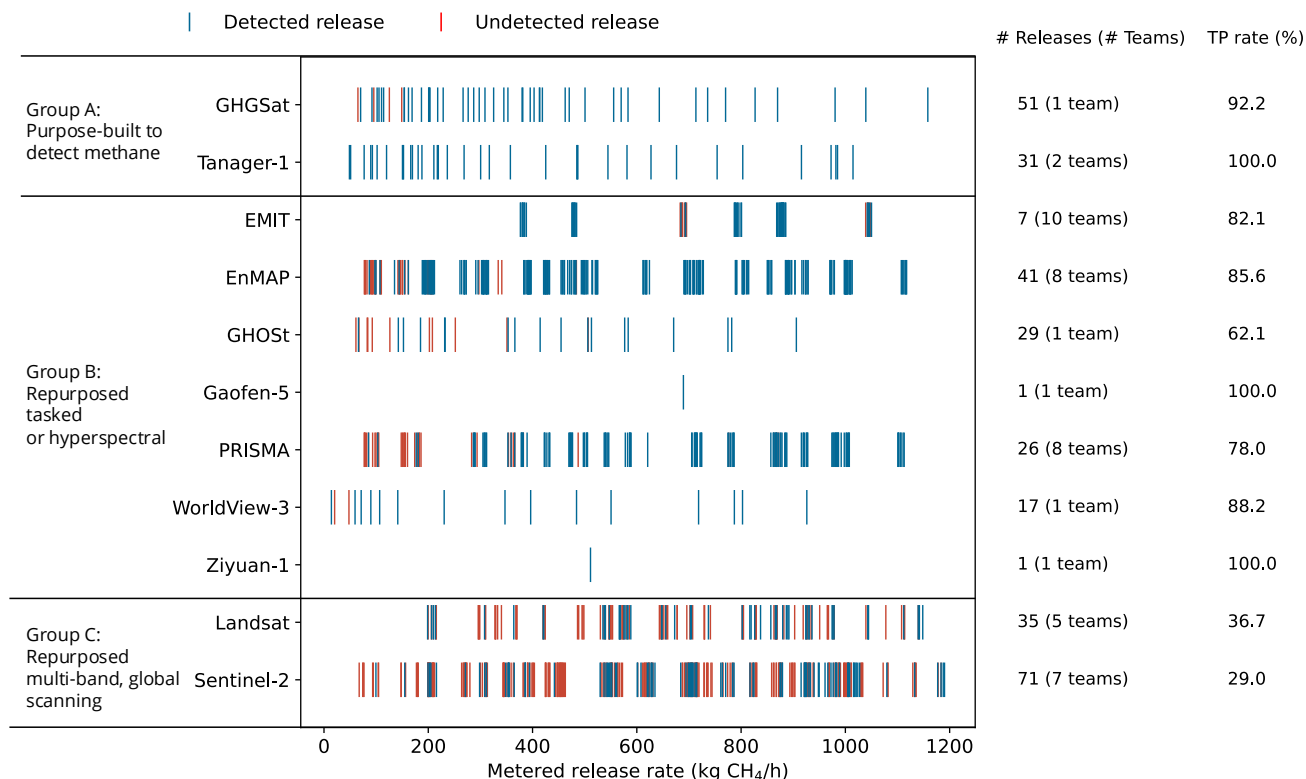


Figure 1. Detection outcomes over the distribution of metered release rates, by satellite and instrument grouping. Each red line (undetected release) or blue line (detected release) represents the performance of a single team. Results from multiple teams analyzing the same satellite are plotted with jitter, which can create the appearance of thicker bars. Release counts, number of analysis teams, and true positive rates (TPRs) listed to right. Instances where a release was performed but the satellite was not tasked or analysis teams filtered the observation due to cloud obstruction are not included. Thus, these results represent rates for successfully acquired, non-cloudy imagery. Rates must therefore be adjusted for cloud coverage in a given location.

Detection outcomes can be analyzed by team with more detail (Figure 2). The most frequently analyzed instruments, EMIT (10 teams), EnMAP (8), and PRISMA (8), displayed relatively consistent detection outcomes. The seven teams who participated in all four phases of blind testing correctly detected between 28 and 35 of the non-cloudy EnMAP release events and achieved comparable F1 scores of 0.89 to 0.96 (see SI Table S6 for performance metrics). Given the heterogeneity of analysis approaches used, the very similar TP rate suggests some insensitivity to methodological approach and that detection limits are governed by the underlying spectral data. In contrast, outcomes for Sentinel-2 and Landsat varied substantially among teams. Out of 71 non-zero releases for Sentinel-2, two teams made 0 or 1 TP detections while two other teams made >35 TP detections (see SI Figure S2 for each team's outcomes by release rate); Sentinel-2 F1 scores range 0.0-0.73 (SI Table S6).

Eight different instruments displayed at least one false positive (Fig.2). This result differs from previous controlled release studies in which no FPs occurred^{29,33}. Of the total 71 FPs, 56 (78.9%) occurred with Group C instruments. With these instruments, teams that reported more TPs generally also reported more FPs. For instance, analyzing Group C satellite data, Orbio Earth had the most TPs (58) and FPs (33) (precision = 0.64), while Kayrros made one-third as many TPs (19) and no FPs (precision = 1). Thus, aggressive algorithm tuning likely allows detection of smaller emissions sources (higher TPs) but can cause artifacts to be detected as plumes (higher FPs). Depending on the context, these FPs can be more or less impactful.

225

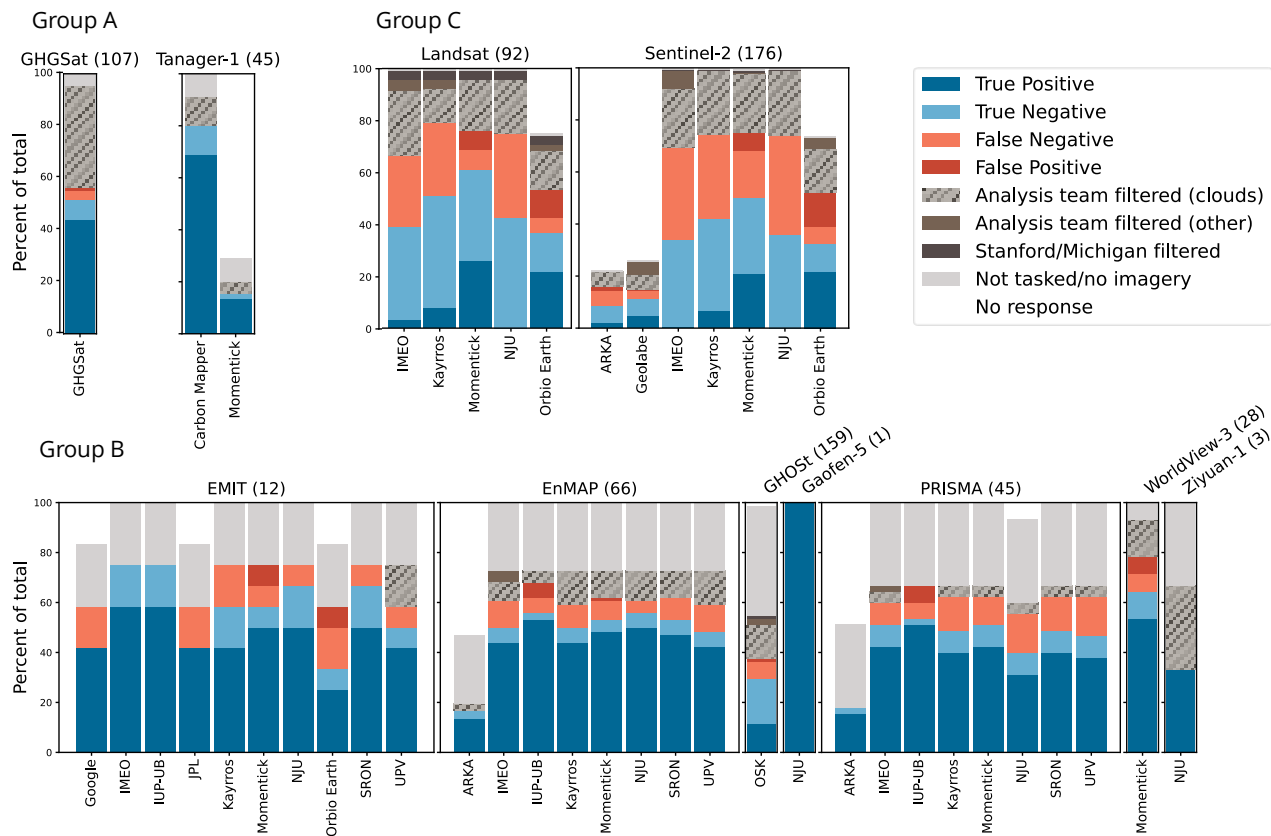


Figure 2. Detection results by satellite and analysis team, grouped by instrument type. Total number of observations listed in parentheses above chart. Between one and ten teams analyzed data from each satellite. For most satellite-team pairings, the number of accurate detection responses exceeds the number of incorrect classifications. In several cases for tasking instruments, the satellite was not tasked and collected no data. Bar heights will vary for teams that did not participate in all four study phases.

230

Figure 3 shows example methane plume images from releases of approximately 800 kg CH₄/h. All plumes are visualized on the same concentration enhancement scale. The white dot near the center of each image is the actual methane release location. The three previously untested instruments (EMIT, GHOS_t, and Tanager-1) are included, as is GHGSat, so teams who participated in all study phases are represented (see SI for images from all team-satellite pairings). As observed in prior work, teams analyzing the same spectral data can produce enhancements, and subsequently plume masks, with very different shapes³³. For example, while nine of the ten teams analyzing EMIT for the 3 June case show plumes traveling in the same downwind direction, two stand out as having significantly different smoothing and brightness. Moreover, visually similar enhancements can yield varying source rate estimates; eight visually similar enhancements (excepting Google and Orbio Earth) result in rate estimates ranging 348 kg CH₄/h [0, 1,196] to 1,406 kg CH₄/h [10, 2,803]. There is no apparent trend between errors in assumed windspeed and enhancement concentration in this example. Future work should examine in detail the drivers of high variance cases.

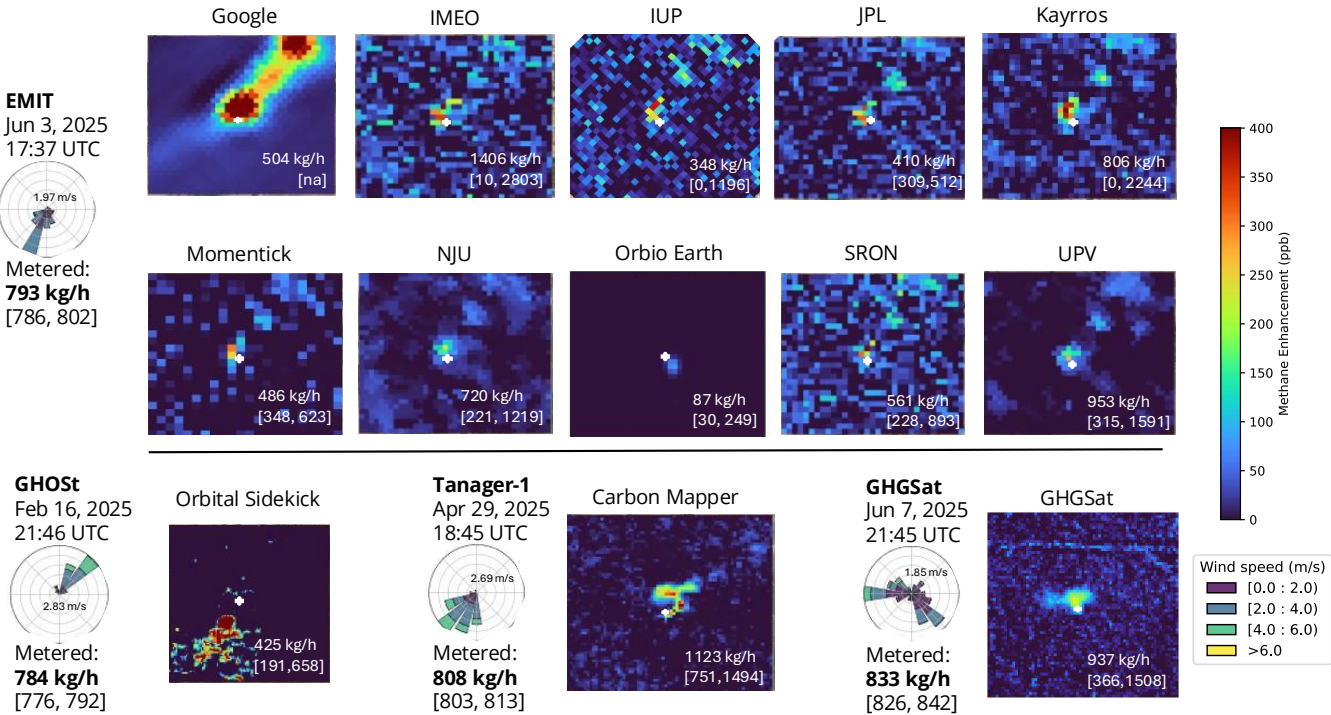


Figure 3. Methane enhancements from three newly tested satellites, EMIT, GHOS_t and Tanager-1. Releases for each satellite occurred on different days but are all approximately 800 kg/h. The 95% confidence interval around the mean metered flow rate is shown. Also shown are mean estimated emission rate and 95% confidence interval from each team-satellite pairing and a 10 min wind rose of measured 10 m wind speed.

Detection efficacy and meteorological conditions

Cloud cover constrains satellite methane monitoring, even in arid locations like our study site (see SI for meteorological distributions). Almost one-fifth (19.5%) of total responses were filtered by analysis teams due to cloud cover. Cloud filtering varies between analysis teams processing the same underlying data (Figure 2). For example, Kayrros filtered 12 Landsat observations due to clouds while IMEO filtered almost twice as many (23 observations).

How cloudy is too cloudy for satellites to make detections? Figure 4 investigates detection performance across instrument classes binned by cloud percentage. Cloud cover percentage is calculated from a Wemantics cloud mask product applied to sky images captured every ~30s taken by an upward facing 360-degree fish-eye sky camera (see Methods). This camera includes RGB imaging with an integrated pyranometer (spectral range 385-2015 nm) to detect and segment clouds across the entire sky. These data represent cloud cover seen from the ground and can differ from satellite-derived cloud masks. Group A and Group B instruments maintained high detection accuracy in clear or low-cloud (<20% cloud cover) conditions. Detection performance deteriorated significantly under partly cloudy or cloudy conditions; most detection outcomes in the 20-40% cloud cover range (or greater) resulted in either no determination (filtered for clouds) or an incorrect classification (FN or FP). Group C detection performance was also significantly challenged by cloudiness and degraded somewhat even at low cloud levels (10-20%).

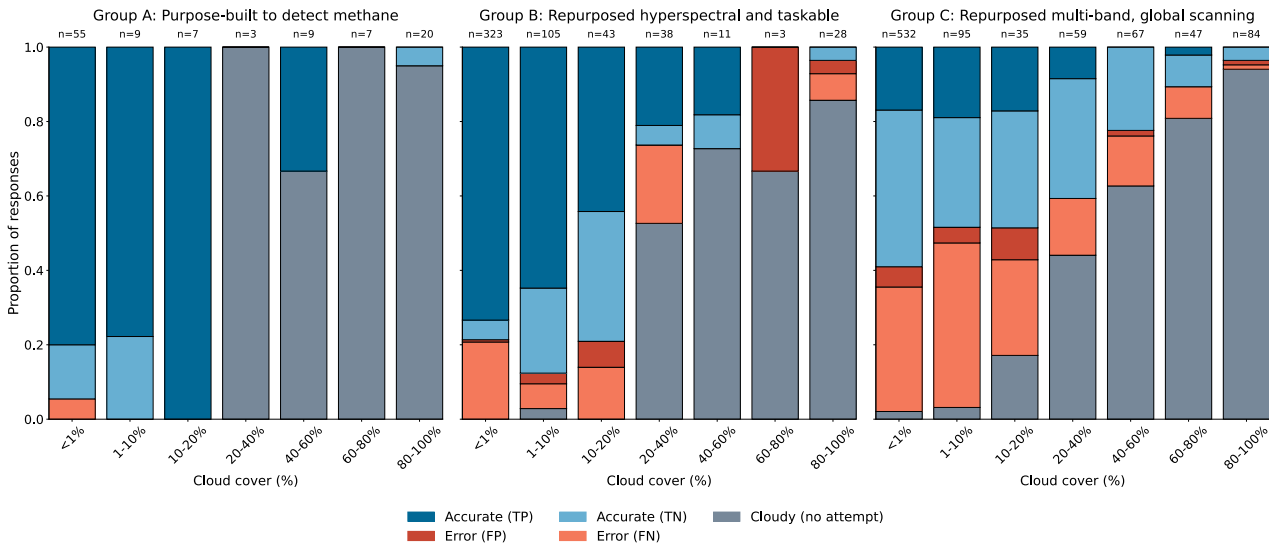


Figure 4. Detection accuracy across satellite typologies, binned by cloud cover percentage. Sample per bin listed on chart. Purpose-built methane detection instruments (Group A), and to a similar extent, other hyperspectral and

taskable instruments (Group B) maintain strong detection accuracy under <20% cloud cover. Detection
275 performance deteriorates significantly across all instrument types in the 20-40% cloud cover range.

Wind speed can also affect plume detection. Group A and C satellites demonstrated modest changes in detection
efficacy as a function of wind speed (see SI Figure S4). Group B satellites showed pronounced sensitivity to wind
speed—detection accuracy declined for winds > 3 m/s. This suggests that stronger winds can quickly dilute
280 plumes and push methane enhancements below detectable limits. Further examination of windspeed and
detection capabilities is warranted for future study.

Variable quantification performance across instrument classes

In addition to detection, teams reported estimates of the methane release rate [kg CH₄/h]. Figure 5 shows all
285 reported quantity estimates by satellite, including FP and FN observations on the y-axis and x-axis respectively.
Individual estimate errors ranged from -98% to 678% of the metered value. The median absolute error was 37%,
while the mean absolute percent error (MAE) was 49%. By group: median absolute errors were 34.8%, 35.6%,
43.0% for Groups A, B and C, respectively, while group-wise MAEs were 43.5% [95% CI: 34.9, 52.2], 45.8%
[42.0, 49.7], and 60.7% [46.3, 75.1]. Nearly two-thirds (64%) of all estimates fell within +/- 50% of the metered
290 value. Regarding uncertainty ranges, 72.8% of all reported 95% CIs intersected the 1:1 parity line, suggesting
moderately overconfident (narrow) uncertainty ranges.

Ordinary Least Squares (OLS) fits between actual (x) and estimated (y) mass emissions rate were performed against
all satellite-team pairs with five or more TP detections. To characterize each team's quantification performance,
295 we fit a 2-parameter (2P) regression (free slope and intercept) that includes both TP and FP estimates. This approach
differs from prior experiments that forced a zero intercept^{29,33} and is based on simulations showing that a 2P model
is more robust when FPs are non-negligible (see SI S3.9.1). Full fitting statistics are provided in the SI (Table S6).

Group A instruments showed strong overall performance, with average slopes of 0.83-1.00 and R^2 of 0.70-0.71.
300 Estimates based on both Group A instruments exhibited modest overconfidence, with 67-83% of the error bars
intersecting the parity line. Group B instruments also showed strong quantification performance with more
variation. Average slopes ranged from 0.61-1.19 and R^2 from 0.34-0.70. For satellites processed by multiple teams,
variable quantification performance reflects diversity in analytical approaches. For example, with EnMAP, eight
teams produced slopes ranging from 0.39 to 0.95 ($R^2 = 0.28-0.45$). For Group C instruments, performance was

305 mixed and generally below that of Group A and B. Some cases show flattened slopes, low correlation, and large
positive intercepts. While this suggests weak Group C performance at the tested rates, these outcomes are not
entirely unexpected. When testing in an instrument’s partial detection range, quantification becomes convolved
with Y/N detection—signals that are diminished by noise will be undetected, and signals boosted by noise will be
over-quantified—this effect will tend to flatten a linear fit. Landsat, relative to other satellites in the study, was also
310 tested over a more limited dynamic range (~500-1200), which contributes to a point cloud effect.

The lowest successfully quantified releases provide a first bound on minimum detection capabilities of several new
instruments, under idealized conditions. Carbon Mapper and Momentick detected the lowest Tanager-1 release of
49.6 kg CH₄/h [95% CI: 49.4, 49.7]; Momentick quantified the release to within 27% error. The mean 10 m wind
315 speed during Tanager-1 releases was 2.55 m/s [min .50, max 8.58]. Duren et al. (2025) predicts minimum detection
limits of about 64-126 kg CH₄/h for 3 m/s winds, 25% albedo and 45-degree zenith¹⁵. GHOSSt successfully detected
one out of seven <100 kg CH₄/h releases, the detection being a release of 67.0 kg CH₄/h [95% CI: 66.2, 67.8] under
0.9 m/s winds, quantified with 1.2% rate error. To the best of the authors’ knowledge, this is the first such public
estimate of this instrument’s detection limit. The lowest EMIT release of 382.4 [95% CI: 380.1, 384.8] kg CH₄/h
320 under 0.2 m/s winds was detected by all analysis teams. Ayasse et al. (2024) estimated a 90% probability of
detection (POD) for EMIT at approximately 1,500 kg CH₄/h for 6 m/s winds³⁶. See the SI for all quantification
error summary statistics by satellite and team.

Discussion

A performance hierarchy clearly emerges among our broadly grouped satellite classes. Group A systems demonstrated the strongest detection performance, characterized by high true positive detection rates and few to no false positives for releases spanning <100 kg CH₄/h to over 1,000 kg CH₄/h. Their F1 scores range 0.95 to 1, indicating ability to accurately identify both positive and negative cases (see performance metrics in SI Table S6). Quantification was also largely unbiased. For example, Carbon Mapper successfully detected every non-filtered release (n=31) with Tanager-1 in the year-long experiment. They also reported zero false positives across five no-release controls. Carbon Mapper's performance cannot be attributed to random chance. Under a permutation analysis where Carbon Mapper's Y/N labels were shuffled 10,000 times and randomly assigned to each release/non-release event (simulating a null hypothesis of no detection skill), the likelihood of obtaining their observed accuracy (100%) by chance is very small ($p < 0.0001$) (see SI S3.4). Among their successful detections were eleven releases less than 200 kg CH₄/h (between 49.6 [49.4, 49.7] kg CH₄/h and 190.9 [189.1, 192.7] kg CH₄/h) with 2.2 m/s mean windspeeds. They quantified ten of these eleven emissions with mean absolute percent error (MAE) of 61% (one release not quantified). Momentick, who analyzed Tanager-1 only in the fourth experiment phase, also detected every release. Similarly, GHGSat successfully detected 92% of all releases, including eleven out of fifteen emissions smaller than the lowest quantified source in previous tests [$\sim$ 197 kg CH₄/h]. MAE across those eleven detections was 51.7% under 2.4 m/s mean 10 m windspeeds. GHGSat reported one false positive, estimated at 76.9 kg CH₄/h [8.3, 145.4]. Their observed accuracy (92%) is also highly statistically significant ($p < 0.0001$) (Table S6).

Significant skill was also demonstrated by several teams analyzing Group B instruments. Using EnMAP, EMIT, and PRISMA, eight teams (IMEO, IUP, ARKA, SRON, NJU, UPV, Momentick, and Kayros) demonstrated statistically significant balanced accuracies above 90%, on par with Group A (Table S8). JPL (EMIT), Google (EMIT), Kayros (EnMAP, PRISMA), and Orbital Side Kick (GHOSat) achieved near-parity slopes. While Group C performance was challenged overall, Kayros (Sentinel-2) and Momentick (Sentinel-2) had among the lowest average error for True Positive quantifications on a percentage basis (Table S8). These results give confidence that a growing number of organizations have credible analytic competencies to support global methane mitigation.

Spectral resolution in the shortwave infrared (SWIR) is one factor that appears to differentiate study performance. The top-performing Group A instruments combine <30 m pixels with high spectral resolution (<5 nm). Group B instruments tend to have coarser spectral resolution. For example, PRISMA has the same spatial resolution as

Tanager-1 (30 m), but half the spectral resolution (10 nm). It is significantly less effective. Even more apparent, Landsat has the same spatial resolution but a very coarse spectral resolution (80/180 nm), resulting in poor performance. But there are tradeoffs between spatial and spectral resolution, and performance is not driven by a single factor. While not a hyperspectral instrument, WorldView-3's very high spatial resolution (3.7 m) enabled Momentick to detect the smallest release in this study of 9.47 kg CH₄/h [9.38, 9.57]. Momentick detected four of six <100 kg CH₄/h WV3 releases and quantified them with 51.5% MAE. WV3 was tested against the smallest rates in the study. If detection performance is compared in an overlapping ~50-200 kg CH₄/h range, WV3's true positive rate (85%) is comparable with Group A instruments (73-100%), as is IUP's analysis of EnMAP (83%) and PRISMA (75%). This illustrates again that very effective performance is possible with Group B instruments as well.

These results suggest strong evidence for reliable Y/N detection of sources under 200 kg CH₄/h by Group A and some Group B satellites in idealized conditions. But test performance is unlikely to be totally representative of performance in the field. A recent evaluation comparing GHGSat to airborne instruments in the Permian Basin found that its 50% POD threshold is near 250 kg CH₄/h. And analysis of Tanager-1 quantified plumes in the same basin detected using the satellite's lowest sensitivity mode shows a drop-off in detections near 200-400 kg CH₄/h (see SI S3.6). Factors affecting this 'test-to-field gap' include favorable test conditions, such as a constrained emissions location, high solar insolation and albedo, low cloud cover, as well as the possibility that analysis teams apply different operational workflows when not in a test environment (e.g. additional human quality control steps)³⁸. Another relevant factor for some satellites is the use of different imaging modes that can be used to trade off higher detection limits for higher spatial coverage (or vice versa). The Tanager-1 Permian observations were made in the lowest sensitivity mode to prioritize higher spatial coverage, while observations of the test site were made in higher sensitivity modes to prioritize detection accuracy. Had higher-sensitivity modes been used in the Permian, the drop-off in detections would likely be closer to 100-200 kg CH₄/h¹⁵. More research is needed on the translation of controlled test results to real-world performance.

Group C multispectral global imagers performance was challenged by the emission sizes in this study. Quantification error for Group C instruments was substantially higher (64% MAE) for smaller releases (<1,000 kg CH₄/h) compared to larger ones (39% MAE). This is not unexpected: simulations have predicted a roughly 1,000 kg CH₄/h detection limit for Sentinel-2 over homogeneous surfaces⁵².

395 There is some evidence that emerging AI-driven analytical methods may be improving performance. Researchers at Geolabe previously reported a deep learning method with stated Sentinel-2 detection capabilities down to 200 kg CH₄/h²⁰. They participated in one phase of this study, detecting three of four releases <500 kg CH₄/h with 67% MAE. With Sentinel-2, Orbio Earth detected and quantified an emission rate of 100.1 [98.8, 101.3] kg CH₄/h to within 15% MAE, and Momentick detected an emission of 304.8 [302.8, 306.8] kg CH₄/h to within 22% MAE.

400 The performance of AI-utilizing teams is unlikely to be by chance; their observed accuracies of 0.62-0.77 were statistically significant compared to a null hypothesis of random guessing ($p < 0.05$). But these new methods deserve careful analysis and judicious application. Most false positives (79%) in this study originated from teams applying AI-based methods to Sentinel-2 and Landsat data, leading to low precision values (see SI Table S6 and section S3.5 confirming statistical significance of higher FPs). Ideal methods will minimize false positives to maintain trust in

405 results and avoid wasted repair efforts. However, very strict detection thresholds that avoid all false positives may result in many missed plumes (large FN rate); the optimal balance between these two error types depends on the application. Our results suggest that further development of AI methods is needed to improve reliability and robustness. Given S2 and LS's global coverage, high revisit rates (5-8 d), and deep historical archives, improving the reliability of methane detection from these instruments remains a high-value goal for the monitoring

410 community. The dataset generated in this study may allow for training and refining these algorithms.

Clouds present an ongoing challenge for satellite methane monitoring. Our results confirm that while there is some inter-team variability in cloud filtering thresholds, all instrument classes are challenged to make detections by partly-to-fully cloudy conditions. This poses problems in some key methane-emitting regions, including tropical

415 oil producers Nigeria and Venezuela, as well as cold cloudy climates like Canada and Russia⁵³. Instruments with smaller spatial footprints may be able to observe between gaps in clouds⁵⁴. In this study, Carbon Mapper made two TP detections when ground-measured cloud cover exceeded 45%.

Importantly, most applications of satellite monitoring also require consideration of revisit frequency and total

420 spatial coverage. Many methane sources are highly intermittent, meaning their detection is fundamentally a function of how often a satellite observes them⁵⁵. Furthermore, emitting infrastructure is often geographically dispersed. Constellation systems tend to have superior spatial coverage over single, taskable assets with comparable spatial resolution.

425 Also, any satellite's ability to see a high fraction of total emissions depends on how its detection threshold interacts with a region's specific emissions profile. A sensor (or constellation of sensors) with a 100 kg CH₄/h detection limit could capture a large fraction of emissions (>70%) in the high-emitting Permian basin if coverage was sufficiently high. It would detect a much smaller fraction of cumulative emissions in a basin characterized by more numerous, smaller sources⁵⁶.

430

We develop instrument- and team-specific quantification fits, enabled by this study's large dataset. Ideally, these models would be useful for characterizing each method's performance and for adjusting estimates of real-world emission sources. However, further study is needed to determine applicability of test-based calibration factors to globally measured plumes. Future validation campaigns should be prioritized to include: 1) locations with varying
435 surface terrains and albedos; 2) locations with very different weather conditions and sun availability; 3) offshore releases to test performance over water; and 4) "fully blind" tests where the source location is undisclosed. Ultimately, a controlled test that is completely unknown to participating teams would be the best evaluation of the true efficacy of operational workflows.

440 This study provides the most comprehensive evidence to date that space-based methane monitoring is effective. We have shown that purpose-built sensors can robustly detect relatively small emissions, while numerous teams can also wield repurposed instruments to detect and quantify methane with significant skill. By establishing these findings, we hope this moves the conversation beyond *if* these systems work, towards discussions of *how*, given algorithmic trade-offs, and *where* they can be most effectively applied to drive global methane mitigation.

445

450

Methods

The release stack was installed in Casa Grande, Arizona with coordinates (32.8128, -111.7858). Arizona's high sun hours and arid, low-cloud conditions were advantageous for year-round operation. All gas release and meteorological equipment were powered by solar photovoltaic (PV). Participants were informed of the stack location. Participants were not informed whether a release would occur during a given overpass. Participants were only provided broad rate bounds (3–1,500 kg/h) and were blind to metered release rates. No data were given on meteorological conditions during each release. All data became available after unblinding, following the conclusion of each phase. Gas release instrumentation and data reporting procedures were based on methods used in prior controlled release experiments^{28,33}.

Controlled release site and equipment

Four hundred eighty-one (481) controlled methane releases were conducted at the Casa Grande site: 130 during Phase 1 (Jan–Mar 2025), 122 during Phase 2 (Apr–Jun 2025), 128 during Phase 3 (Jul–Sept 2025), and 101 during Phase 4 (Oct–Dec 2025). Compressed natural gas (CNG) was purchased from local filling stations, stored onsite in two trailers, and filled to approximately 170 bar(g) (2,500 psig). The two trailers had water volumes of 9.7 and 15.8 m³ (342 and 557 ft³), respectively.

Gas was routed from the trailers to a Rawhide Leasing pressure regulation unit (RT-30) via high pressure hosing. Gas flows through two parallel regulation lines. Two pressure cuts bring trailer pressure to 65.5 bar(g) (950 psig) and then to 13.8 bar(g) (200 psig) using sequential inline pressure regulators. Eight Catco catalytic heaters were installed at the pressure cut points to prevent freeze-out due to the Joule-Thomson expansion cooling effect. All valves and tubing were stainless steel, rated to approximately -200°C. The most temperature-sensitive component was the largest Coriolis meter (CMFS150M) on the metering trailer, rated to approximately -50°C. The system was equipped with multiple pressure relief valves for overpressure protection. Pressure is digitally monitored upstream of the first regulator, as well as between the regulators and at the inlet of the metering system.

Downstream of the second pressure cut, gas entered a stainless-steel manifold fitted with a digital pressure gauge, then flowed to a metered release trailer. There, it passed through one of three Coriolis flow meters and was released from a 7.3 m vertical stack (Phase 1) or a 3.0 m stack (Phase 2, Phase 3, Phase 4). The large stack approximates an unlit oilfield flare or gas processing unit. The small stack approximates a vent or leak from a smaller piece of equipment such as a compressor. We did not perform ground-level releases, such as those that might occur from a

wellhead, for safety reasons. The three flow paths were designed for release rates of approximately 1–30 kg/h, 30–300 kg/h, and 300–2,000 kg/h, each with an appropriately sized Emerson Micromotion Coriolis meter (see SI of El Abbadi et al., 2024). At the highest flow rates, significant pressure drop caused flow instability; the maximum
485 stable release rate achieved was approximately 1,300 kg/h. Each Coriolis meter was connected to a Micro Motion 5700 transmitter, providing real-time data to the metering trailer’s programmable logic controller. An automated control system, operated through an internet-connected laptop, adjusted valve positions to maintain a user-inputted flow set point.

490 All external wiring used Class I, Division 1 hazardous location-approved components to ensure safety and compliance with relevant standards. Solar PV panels with battery storage supplied electricity for catalytic heaters, remote communication via cellular networks, and flow control for the metering systems.

The Emerson Coriolis flow meters were calibrated by the manufacturer according to the schedules described in El
495 Abbadi et al. (2024). These meters are known for their durability, with minimal drift expected under typical field conditions. To avoid risks from removing and shipping the meters for laboratory recalibration, the research team developed a field-based verification method (see Supplementary methods). An Apollo GX-102000LS Balance was used to validate meter performance. Inert gas from 50 lb cylinders was flowed at 5-20 kg/h through the small (CMFS015H, 0.166 inch) and at 50-75 kg/h thorough the medium (CMF050M, 0.5 inch) meters, while CNG was
500 disconnected. Each test run lasted a minimum of five minutes, and average flow rate values were calculated over the interval (see SI Table S9). The large meter, with a minimum rate of 300 kg/h, could not be tested with this method. Instead, a mass-balance method was applied (see SI S4.1), comparing volumes of purchased gas to the sum of metered flow and other emission sources, including small leaks reported throughout the experiment. Mass balance accounted for ~98.5% of input mass over the year-long experiment. Given the consistent results for the
505 smaller meters, the larger meter’s identical model generation and calibration history, and the overall mass balance accounting, we can have confidence in metering accuracy.

Meteorological equipment

Meteorological instruments captured basic meteorological data, multi-height wind profiles, sky conditions, and soil
510 moisture. These measurements provide a canonical dataset for supporting and improving satellite-based modelling and are permanently archived⁵⁷. Instrument deployment timelines and any outages are detailed in Table S10.

A Metsens500 sensor mounted on a steel tripod measured wind speed and direction at 2 m using an integrated ultrasonic sensor (<https://www.campbellsci.com/metsens500>). At 5 and 10 m, 3D wind velocity and direction were
515 measured using Campbell Scientific CSAT3B sonic anemometers (<https://www.campbellsci.com/csat3b>). These anemometers were mounted on an aluminum hoistable tower (Aluma Tower) oriented toward prevailing winds (270° from N) per manufacturer guidance (see Supplementary methods). These were located ~60 m WNW of the release stack (see SI Figure S75). The Metsens500 also recorded air temperature, relative humidity, and barometric pressure. All sensors were connected to a CR1000X datalogger (Campbell Scientific), which recorded data at 0.2
520 Hz.

A ZX 300 Doppler wind lidar (serial no. 2340) manufactured by Zephir Ltd (<https://www.zxlidars.com/>) was deployed ~50 m WNW of the release stack to measure vertical wind profiles. The ZX 300 performs continuous conical scans to detect aerosol backscatter and retrieve wind speed and direction at 11 user-selected heights between
525 10 and 300 m above sensor height. Since the sensor is 1 m above ground, the lowest effective measurement height is 11 m. We sampled at 5-m increments between 11 and 100 m, including the required 39-m calibration height. A full scan across all selected heights occurs approximately every 15 seconds. The unit's self-cleaning system was supplied by a multi-gallon water reservoir. LIDAR systems like the ZX 300 are known to exhibit a small number of samples where the wind direction is flipped 180 degrees. To ensure consistency with the co-located 10-m
530 measurement, a standard correction was applied to the lidar-derived wind direction: when the 11-m direction differed from the 10-m sonic anemometer by more than 90 degrees, the lidar-derived wind direction at all heights was flipped by 180 degrees. This correction, consistent with methods used in peer-reviewed studies affected fewer than 10% of all measurements⁵⁸.

535 Sky conditions were monitored using a Wemantics dual-camera sky imager mounted ~2.5 m above ground level on a tripod (<https://www.wemantics.com/>). The camera faced directly upward with an unobstructed hemispheric field of view. The system features a visible imaging sensor and integrates a thermopile pyranometer for measuring global horizontal irradiance (GHI) across a spectral range of 385–2,105 nm. The camera captures 12.3-megapixel high-dynamic-range (HDR) images every 30 seconds and computes a cloud coverage factor and binary cloud mask for
540 each frame (see SI Figure S80). Timestamped images were accessed via the Wemantics API.

Soil moisture was measured using two METER Group TEROS 11 sensors, each installed approximately 60 meters from the release point (NE and SW directions) (<https://metergroup.com/products/teros-11/>). The sensors were

buried horizontally at a depth of two cm and connected to the datalogger. Volumetric water content (VWC) was
545 recorded at 30-second intervals and reported at 5-second intervals (.2 Hz) to align with the reported wind data.

Data were processed using Campbell Scientific datalogging and recording software (Loggernet version 4.7) and
further processed using custom Python data handling scripts. Daily readings from each instrument were combined
into a single date file in Coordinated Universal Time (UTC). Reported wind direction is referenced to true north
550 (geographic north). Processing scripts take raw wind direction output from each sensor and correct for both the
instrument's physical orientation and the site-specific magnetic declination.

All measured release and meteorological data are archived in a permanent public repository hosted by Stanford
University at an unchanging URL. See: <https://doi.org/10.25740/qh001qt3946>.
555

Safety

All methane release activities followed strict safety protocols and were conducted only by trained personnel.
Fabrication, transport, and operation of natural gas equipment were handled by Rawhide Leasing, a professional
gas services contractor. Experiment team members did not operate gas-handling equipment and only approached
560 the release infrastructure when valves were closed and no gas was flowing.

The release system was fully automated and remotely operated. Before each release, the team confirmed site
vacancy via real-time security cameras. They also monitored equipment performance and verified that releases
occurred only in intended windows using a remote dashboard with methane detections made by an infrared Kuva
565 camera. Gas trailers were refilled by licensed staff at Armadillo Trucking.

Releases occurred from a stack 3-7 m above ground, where methane, being lighter than air, dispersed rapidly in
ambient winds. The high exit velocity further minimized risk of accumulation near the ground.

570 Gas composition

To convert total gas flow rates to CH₄ flow rates, we applied location-specific gas composition data. Composition
data were obtained from the two nearest upstream measurement stations serving the CNG fill point. The gas
supplier that serves the CNG station provided composition data files with mole fractions for 15 species. Stanford
researchers logged fill dates and usage. For each trailer fill, we averaged the reported CH₄ content over the five

575 days leading up to and including the fill date, following El Abbadi et al. 2024. This procedure enabled each release event to be matched with the corresponding trailer fill and composition (see SI Table S11). Across all fills, mean CH₄ mole percent was 94.9%.

Flow rate uncertainty

580 Uncertainty in ground-metered methane flow rates arises from three sources: (1) short-term variability in the actual gas flow rate as the automated control system continuously adjusts to reach the user-defined target, (2) Coriolis meter uncertainty, and (3) gas composition uncertainty. These sources of error are propagated into the reported ground-metered methane flow rate time series using a quadrature approach.

585 The Coriolis meters used in this campaign have a stated mass flow accuracy of $\pm 0.25\%$ within meter performance ranges (cross validated using Emerson’s online meter sizing tool)⁵⁹. The uncertainty (1-sigma) in the CH₄ mole fraction ranged from 0.149% to 1.09%, with most values under 1%. Following the methodology in Sherwin et al. (2023), we assess the flow rate and its stability over a 5-minute window prior to each satellite overpass by computing the mean, standard deviation, and standard error of the mean. The 95% confidence interval for the mean
590 flow rate is then calculated using the standard error of the mean and the uncertainty in composition. Separately, to assess flow stability, we calculate a 95% prediction interval to quantify the second-by-second variability. If this prediction interval exceeds $\pm 10\%$ of the mean flow value, the release is flagged with an internal quality indicator. These flags are included in the CSV data shared with analysis teams. The 5-minute stability check is intended as an internal consistency filter rather than an absolute pass/fail criterion. Data points are retained in the analysis
595 unless there is evidence of instrumentation error or release instability, such as loss of network connectivity, control system failure to reach the target setpoint, or other operational anomalies.

Experimental design

The experiment was conducted in sequential phases, beginning with an initial cooperative, non-blind stage between
600 August 1 and December 31, 2024. On January 1, 2025, the experiment transitioned to one year of single-blind testing, conducted in quarterly phases. In blind phases, participants were aware of the fixed release coordinates but were not informed whether or not a release would occur during a given satellite overpass (see protocol exception noted below). Participating teams were instructed not to cross-communicate or exchange any information related to the active phase of controlled release testing. Teams were permitted to analyze and submit results for any of the
605 12 satellite observing platforms for which they had access to data. All submissions followed a standardized

reporting template provided by the research team (see SI). Teams were permitted to submit results during a 30-day reporting period after the final release of the phase. Teams that did not submit results during the 30-day reporting period forfeited participation in that phase but remained eligible for subsequent phases. After the reporting period, ground-truth data for that phase were unblinded.

610

An exception to the standard blind protocol occurred on two dates, June 3 and June 27, 2025. On these days, the NASA JPL team conducted AVIRIS-NG airborne campaigns that were intentionally scheduled to be coincident with expected EMIT overpasses of the site. Coincident flights are scientifically valuable for instrument intercomparison. They are also rare—fewer than ten EMIT observing opportunities occurred over the test site during the entire year-long study period. Given the significant logistical and financial resources required to mobilize an aircraft, the airborne flight team needed to anticipate that a release would be scheduled to occur during their flight window. We note that all teams that analyzed the EMIT data for these two dates successfully detected the releases that occurred; therefore, we don’t believe it provided a significant advantage in detection to JPL over other teams.

620

Target release rates

The experiment team allocated approximately half of release events to each satellite’s expected partial detection range. The other half of observations were intended to span the full operational release range. This strategy aimed to balance two goals: (1) probing sensitivity thresholds and (2) assessing quantification accuracy over a wide range of emission magnitudes. Intentional non-releases (“zeros,” “controls,” or “blanks”) were included to test for the presence of false positives but were limited in number to maximize non-zero release observations.

For each overpass, the experiment team randomly selected a target release rate from a satellite-specific distribution. The distributions were developed using prior controlled release studies, empirical plume measurements, and input from participating teams familiar with the capabilities of each platform. For satellites without prior controlled testing (EMIT, GHOST, and Tanager-1), distributions relied primarily on empirical plume data and expert judgment. For example, EMIT detections in the IMEO data portal (over 2,700 global plumes, accessed early 2025) cluster around 1,600–1,700 kg/h range, with some as low as 200 kg/h²³. Given evidence of a power-law distribution of global methane point sources, the scarcity of smaller plumes likely reflects instrument sensitivity rather than absence of small sources⁶⁰. Accordingly, EMIT-targeted releases were set between 200–1,500 kg/h. This procedure was repeated for all satellites. For S2 and LS8/9, the partial detection range spans nearly the entire achievable

635

release range, so all releases effectively occur near the limit of the satellite’s detection capabilities. More sensitive instruments, such as Tanager-1 and GHGSat, received releases spanning both marginally and confidently detectable levels.

640

Operator data collection and metered data unblinding

The experiment team maintained a log of satellite overpasses using email communications, online user portals, and predictable orbital schedules. Each release was assigned a unique release ID. Raw gas flow meter files were cataloged by date and hour (UTC). Metered emissions (1 Hz) and meteorological data (0.2 Hz) were unblinded
645 after the reported period completed and posted to a permanent repository⁵⁷.

Participating teams submitted results using a standardized reporting template provided by the research team., based on that of Sherwin et al. (2024) and the Advancing Development of Emissions Detection (ADED) protocol⁶¹. Reported materials included: a qualitative sheet describing the measurement system under test and analysis
650 methodology, including methane retrieval steps and plume delineation approach; a quantitative sheet detailing satellite overpass times, viewing geometry, detection outcomes, emission rate estimates with uncertainties, and wind conditions with associated uncertainty; unmasked retrieval enhancement images for all measurement attempts; detected methane plume masks, when applicable (see SI Tables S8 and S9).

655 In Phase 1, 12 teams submitted results: Arka, Carbon Mapper, GHGSat, IMEO, IUP-UB, Kayrros, Momentick, NJU, Orbio Earth, Orbital Sidekick, SRON, and UPV. The same teams except Arka, plus JPL, Geolabe, and Google submitted in Phase 2, totalling 14 submissions. In Phase 3, 11 teams submitted (all Phase 1 teams except Arka). In Phase 4, 12 teams submitted (all Phase 1 teams except Arka and Orbio Earth, plus JPL and Google). MethaneSAT did not submit results in any phase. The date of each team’s initial submission is provided in SI Table S4. The
660 experiment team reviewed submissions and followed up if materials were missing or needed clarification. This prompted some teams to submit revised versions, the dates of which are also noted in Table S4. Late submissions were only accepted if they occurred before data unblinding.

Acknowledgements

665 This project relied on the contributions of many satellite operators and participating organizations. We thank Steve Wofsy, Maryann Sargent, Jasna Pittman, and Jonathan Franklin of MethaneSat; Dan Cusworth, Riley Duren, Alana Ayasse, Joseph Heckler, and Ashton Jackson of Carbon Mapper; Andrew Thorpe, Robert Green, Michael Eastwood, Jay Fahlen, Philip Brodrick, and Dan Jensen of NASA Jet Propulsion Lab; Gonzalo Mateo Garcia, Manuel Montesino San Martin, and Itziar Irakulis Loitxate of IMEO; Bijoy Shetty, Vishal Batchu, Alex Wilson, 670 Michelangelo Conserva, Christopher Van Arsdale, Anna Micihalak, and Mansi Kansal of Google; Carlo de Franchis and Julian Akani Guery of Kayrros; Zani Mouton, Maxime Rischard, and Robert Huppertz of Orbio Earth; Marianne Girard, Dylan Jervis, Jason McKeever, Ariane Deslieres, and Noe Schultz of GHGSat; Javier Roger Juan, Adriana Valverde Iglesias, and Luis Guanter Palomar of UPV; Fei Li and Yongguang Zhang of NJU; Andreas Hayden and Jenny Wan of Arka; Claudia Hulbert and Bertrand Rouet-Leduc of Geolabe; Alon Gilat and Adam 675 Eshel of Momentick; Xin Zhang, Joannes D. Maasakkers, and Ilse Aben of SRON ; Kevin Brodie and Alvin Spivey of Orbital Sidekick; and Stefan Noel, Michael Buchwitz, Jonas Hachmeister, Michael Hilker, and Maximilian Reuter of the University of Bremen. We also thank the German Aerospace Center and the Italian Space Agency for assistance with tasking the EnMAP and PRISMA satellites. We thank Kevin Galvin, Josh Romo, Raffy Xie, and Erica Whiting for assisting with field visits, site setup, and meteorology analyses. The teams at Volta, Rawhide 680 Leasing (Merritt Norton, Kim Loflin, Jason Elliot, Vince Christensen, Daniel, Paco, and Robert), Wematics (Max and Paul), and Armadillo Trucking (Kim and Adis) for their logistical and engineering support throughout the experiment.

Financial support

685 This research has been supported by (alphabetically) the Global Methane Hub, the Grantham Foundation, Saudi Aramco, and the Stanford Natural Gas Initiative (industrial affiliates program). FER's research time was supported by the Stanford Knight-Hennessy Scholars program.

Author contributions

690 Conceptualization—FER, TA, EDS, EAK, ARB. Methods—FER, TA, EDS, SHE, ARB. Resources—ARB. Data curation—FER, TA. Formal analysis—FER, TA, ARB. Project administration—FER, TA, ARB. EAK and ARB supervised the project. Writing (original draft)—FER, ARB. Writing (review and editing)—FER, TA, EAK, EDS, SHE, ARB. Funding acquisition—EAK, ARB.

695 References

1. Intergovernmental Panel On Climate Change (IPCC). *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. (Cambridge University Press, 2023). doi:10.1017/9781009157896.
2. Kort, E. A. *et al.* Four corners: The largest US methane anomaly viewed from space. *Geophys. Res. Lett.* **41**, 6898–6903 (2014).
3. Thompson, D. R. *et al.* Space-based remote imaging spectroscopy of the Aliso Canyon CH₄ superemitter. *Geophys. Res. Lett.* **43**, 6571–6578 (2016).
4. Jacob, D. J. *et al.* Quantifying methane emissions from the global scale down to point sources using satellite observations of atmospheric methane. *Atmospheric Chem. Phys.* **22**, 9617–9646 (2022).
5. Cusworth, D. H. *et al.* Multisatellite Imaging of a Gas Well Blowout Enables Quantification of Total Methane Emissions. *Geophys. Res. Lett.* **48**, e2020GL090864 (2021).
6. Fei, L. *et al.* Mapping methane super-emitters in China and United States with GF-5 02 hyperspectral imaging spectrometer. *Natl. Remote Sens. Bull.* **28**, 1986–2001 (2024).
7. Guanter, L. *et al.* Multisatellite Data Depicts a Record-Breaking Methane Leak from a Well Blowout. *Environ. Sci. Technol. Lett.* <https://doi.org/10.1021/acs.estlett.4c00399> (2024) doi:10.1021/acs.estlett.4c00399.
8. Irakulis-Loitxate, I. *et al.* Satellite-based survey of extreme methane emissions in the Permian basin. *Sci. Adv.* **7**, eabf4507 (2021).
9. Lauvaux, T. *et al.* Global assessment of oil and gas methane ultra-emitters. *Science* <https://doi.org/10.1126/science.abj4351> (2022) doi:10.1126/science.abj4351.
10. Pandey, S. *et al.* Satellite observations reveal extreme methane leakage from a natural gas well blowout. *Proc. Natl. Acad. Sci.* **116**, 26376–26381 (2019).
11. Sánchez-García, E., Gorroño, J., Irakulis-Loitxate, I., Varon, D. J. & Guanter, L. Mapping methane plumes at very high spatial resolution with the WorldView-3 satellite. *Atmospheric Meas. Tech.* **15**, 1657–1674 (2022).
12. Varon, D. J. *et al.* Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes. *Atmospheric Meas. Tech.* **11**, 5673–5686 (2018).

13. Varon, D. J. *et al.* Satellite Discovery of Anomalously Large Methane Point Sources From Oil/Gas Production. *Geophys. Res. Lett.* **46**, 13507–13516 (2019).
14. ESA. S2 Mission. <https://sentiwiki.copernicus.eu/web/s2-mission> (2021).
15. Duren, R. *et al.* The Carbon Mapper emissions monitoring system. *Atmospheric Meas. Tech.* **18**, 6933–6958 (2025).
- 725 16. MethaneSAT. The Satellite | MethaneSAT. <https://www.methanesat.org/satellite> (2025).
17. Bruno, J. H., Jervis, D., Varon, D. J. & Jacob, D. J. U-Plume: automated algorithm for plume detection and source quantification by satellite point-source imagers. *Atmospheric Meas. Tech.* **17**, 2625–2636 (2024).
18. Jongaramrungruang, S., Thorpe, A. K., Matheou, G. & Frankenberg, C. MethaNet – An AI-driven approach to quantifying methane point-source emission from high-resolution 2-D plume imagery. *Remote Sens. Environ.* **269**, 112809 (2022).
- 730 19. Plewa, T., Butz, A., Frankenberg, C., Thorpe, A. K. & Marshall, J. Improvements of AI-driven emission estimation for point sources applied to high resolution 2-D methane-plume imagery. *Remote Sens. Environ.* **331**, 115002 (2025).
20. Rouet-Leduc, B. & Hulbert, C. Automatic detection of methane emissions in multispectral satellite imagery using a vision transformer. *Nat. Commun.* **15**, 3801 (2024).
21. Carbon Mapper. Dashboard | Carbon Mapper. <https://data.carbonmapper.org/#1/25.3/50.5> (2025).
- 735 22. NASA JPL. earth.jpl.nasa.gov/data/data-portal/coverage-and-forecasts. <https://earth.jpl.nasa.gov/data/data-portal/coverage-and-forecasts> (2025).
23. UNEP. Eye on Methane Map. <https://methanedata.unep.org/map> (2025).
24. CARB. California Satellite Methane Project | California Air Resources Board. <https://ww2.arb.ca.gov/our-work/programs/california-satellite-methane-project/about> (2025).
- 740 25. UNEP. Methane Alert and Response System (MARS) | UNEP - UN Environment Programme. <https://www.unep.org/topics/energy/methane/international-methane-emissions-observatory/methane-alert-and-response-system> (2022).
26. US EPA, O. Methane Super Emitter Program. <https://www.epa.gov/compliance/methane-super-emitter-program> (2023).

27. Nordin, S., Hoffmann, E., de Kersauson, Q. & De Catelle, W. Are you ready for the new EU rules on methane emissions
745 in the energy sector? | White & Case LLP. <https://www.whitecase.com/insight-alert/are-you-ready-new-eu-rules-methane-emissions-energy-sector> (2025).
28. El Abbadi, S. H. *et al.* Technological Maturity of Aircraft-Based Methane Sensing for Greenhouse Gas Mitigation. *Environ. Sci. Technol.* **58**, 9591–9600 (2024).
29. Sherwin, E. D. *et al.* Single-blind validation of space-based point-source detection and quantification of onshore methane
750 emissions. *Sci. Rep.* **13**, 3836 (2023).
30. Bell, C. *et al.* Single-blind determination of methane detection limits and quantification accuracy using aircraft-based LiDAR. *Elem. Sci. Anthr.* **10**, (2022).
31. Bell, C., Ilonze, C., Duggan, A. & Zimmerle, D. Performance of Continuous Emission Monitoring Solutions under a Single-Blind Controlled Testing Protocol. *Environ. Sci. Technol.* **57**, 5794–5805 (2023).
- 755 32. Ravikumar, A. P. *et al.* Single-blind inter-comparison of methane detection technologies – results from the Stanford/EDF Mobile Monitoring Challenge. *Elem. Sci. Anthr.* **7**, 37 (2019).
33. Sherwin, E. D. *et al.* Single-blind test of nine methane-sensing satellite systems from three continents. *Atmospheric Meas. Tech.* **17**, 765–782 (2024).
34. Darynova, Z., Blanco, B., Juery, C., Donnat, L. & Duclaux, O. Data assimilation method for quantifying controlled
760 methane releases using a drone and ground-sensors. *Atmospheric Environ. X* **17**, 100210 (2023).
35. Worden, J. R., Nickeson, J. & Policelli, F. CSDA GHGSat Evaluation PI Report v2. *Commercial Satellite Data Acquisition Program GHGSat Principal Investigator Evaluation Summary* <https://earthdata.nasa.gov/s3fs-public/2025-03/CSDA%20GHGSat%20Evaluation%20PI%20Report%20v2.pdf> (2024).
36. Ayasse, A. K. *et al.* Probability of Detection and Multi-Sensor Persistence of Methane Emissions from Coincident
765 Airborne and Satellite Observations. *Environ. Sci. Technol.* **58**, 21536–21544 (2024).
37. Zhang, X. *et al.* Global Identification of Solid Waste Methane Super Emitters Using Hyperspectral Satellites. *Environ. Sci. Technol.* **59**, 18134–18145 (2025).

38. Worden, J., Green, P., Eldering, A. & Sherwin, E. *Common Practices for Quantifying Methane Emissions from Plumes Detected by Remote Sensing*. NIST IR 8575 <https://nvlpubs.nist.gov/nistpubs/ir/2025/NIST.IR.8575.pdf> (2025) doi:10.6028/NIST.IR.8575.
39. Zhang, X. *et al.* HyperGas 1.0: A Python package for analyzing hyperspectral data for greenhouse gases from retrieval to emission rate quantification. Preprint at <https://doi.org/10.5194/egusphere-2025-6127> (2026).
40. Duren, R. *et al.* California's methane super-emitters. *Nature* **575**, 180–184 (2019).
41. Jervis, D. *et al.* The GHGSat-D imaging spectrometer. *Atmospheric Meas. Tech.* **14**, 2127–2140 (2021).
- 775 42. Duren, R. *et al.* The Carbon Mapper emissions monitoring system. *EGUsphere* 1–41 (2025) doi:10.5194/egusphere-2025-2275.
43. Green, R. O. *et al.* The Earth Surface Mineral Dust Source Investigation: An Earth Science Imaging Spectroscopy Mission. in *2020 IEEE Aerospace Conference* 1–15 (IEEE, Big Sky, MT, USA, 2020). doi:10.1109/AERO47225.2020.9172731.
- 780 44. Earth Observation Center of DLR. Mission - EnMAP. <https://www.enmap.org/mission/> (2023).
45. Roger, J. *et al.* High-Resolution Methane Mapping With the EnMAP Satellite Imaging Spectroscopy Mission. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–12 (2024).
46. eoPortal. GHOS (Global Hyperspectral Observation Satellite) - eoPortal. <https://www.eoportal.org/satellite-missions/ghost#eop-quick-facts-section> (2024).
- 785 47. Liu, Y.-N. *et al.* The Advanced Hyperspectral Imager: Aboard China's GaoFen-5 Satellite. *IEEE Geosci. Remote Sens. Mag.* **7**, 23–32 (2019).
48. OHB Italia. Satellites & Missions: PRISMA. *PRISMA* <https://www.ohb-italia.it/satellites-missions/> (2022).
49. ESA. WorldView-3 - Earth Online. <https://earth.esa.int/eogateway/missions/worldview-3> (2022).
50. Chen, L. *et al.* Mapping Alteration Minerals Using ZY-1 02D Hyperspectral Remote Sensing Data in Coalbed Methane Enrichment Areas. *Remote Sens.* **15**, 3590 (2023).
- 790 51. USGS. Landsat 8 | U.S. Geological Survey. <https://www.usgs.gov/landsat-missions/landsat-8> (2023).

52. Gorroño, J., Varon, D. J., Irakulis-Loitxate, I. & Guanter, L. Understanding the potential of Sentinel-2 for monitoring methane point emissions. *Atmospheric Meas. Tech.* **16**, 89–107 (2023).
53. Frankenberg, C. *et al.* Data Drought in the Humid Tropics: How to Overcome the Cloud Barrier in Greenhouse Gas Remote Sensing. *Geophys. Res. Lett.* **51**, e2024GL108791 (2024).
54. Carbon Mapper. Data Dispatch - Seeing around the clouds. <https://carbonmapper.org/articles/seeing-around-the-clouds> (2025).
55. Cusworth, D. H. *et al.* Intermittency of Large Methane Emitters in the Permian Basin. *Environ. Sci. Technol. Lett.* **8**, 567–573 (2021).
56. Sherwin, E. D. *et al.* US oil and gas system emissions from nearly one million aerial site measurements. *Nature* **627**, 328–334 (2024).
57. Reuland, F., Adams, T., Galvin, K., Kort, E. A. & Brandt, A. R. Large-Scale Controlled Methane Releases for Satellite-Based Detection and Emission Quantification of Point-Sources. Preprint at <https://doi.org/10.25740/QH001QT3946> (2025).
58. Knoop, S., Bosveld, F. C., De Haij, M. J. & Apituley, A. A 2-year intercomparison of continuous-wave focusing wind lidar and tall mast wind measurements at Cabauw. *Atmospheric Meas. Tech.* **14**, 2219–2235 (2021).
59. Emerson. Products Sizing Tool. <https://www.emerson.com/catalog/CommonProductAdvisorDisplayView?langId=-1&storeId=20151&divCode=MMI&language=US> (2025).
60. Ehret, T. *et al.* Global Tracking and Quantification of Oil and Gas Methane Emissions from Recurrent *Sentinel-2* Imagery. *Environ. Sci. Technol.* **56**, 10517–10529 (2022).
61. Bell, C. & Zimmerle, D. METEC controlled test protocol: survey emission detection and quantification. Preprint at <https://doi.org/10.25675/10217/235363> (2022).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplemental03122026.docx](#)