
Supplementary information

**Evaluating space-based methane detection
through a one-year blind test**

In the format provided by the
authors and unedited

Supplement of

Evaluating space-based methane detection through a one-year blind test

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

S1. Study timeline

10 The year-long study was divided into quarterly phases during calendar year 2025, spanning January 1, 2025 to December 31, 2025 (Table S1). After each quarterly phase, participants could report results during a 30-day reporting period after the final release of the phase. After the end of the reporting period, ground-truth data for that phase were unblinded. Teams that did not submit during the 30-day reporting period forfeited participation in that phase but remained eligible for subsequent phases.

15 Prior the formal start of blind testing in January 2025, a period of cooperative, non-blind releases (“Phase 0”) occurred. These releases were performed between August and December 2024. Phase 0 primarily served as a debugging period for new satellite entrants to test instrument tasking, refine retrieval algorithms, and validate analysis pipelines. Concurrently, the Stanford/University of Michigan experiment team performed equipment upgrades to test and debug remote connection capability and remote control of flow and heating systems. Phase 0

20 releases were conducted in Evanston, Wyoming (Aug–Oct) and Casa Grande, Arizona (Oct–Dec). In Phase 0, satellite teams received advance notice of release times and, in some cases, consulted on target rates. Ground-metered emissions were shared within one week to enable rapid comparison with satellite-based estimates. The Phase 0 dataset has been archived in the Stanford Digital Repository ¹. Phase 0 data should not be compared to the main study results, due to the lack of a blind protocol.

25

Table S1. Experiment schedule, with phase unblinding dates starred

	Aug- Dec '24	Jan '25	Feb '25	Mar '25	Apr '25	May '25	Jun '25	Jul '25	Aug '25	Sep '25	Oct '25	Nov '25	Dec '25	Jan '25
Phase 0 releases (non-blind)														
Phase 1 releases (blind)														
Phase 1 sub-missions due														
Phase 2 releases														
Phase 2 sub-missions due														
Phase 3 releases (blind)														
Phase 3 sub-missions due														
Phase 4 releases (blind)														
Phase 4 sub-missions due														

S2. Participating satellites and analyst teams

35 **Table S2.** Matrix of participating satellites (columns) and analyzing organizations (rows). Dark grey cells indicate no single blind results available for that satellite-team pairing. Light grey cells indicate that satellite/team participated in one phase. White cells indicate that satellite-team pairing participated in two, three or four phases of testing. These markings indicate team-satellite pairings for this study; teams may analyze additional instruments in other contexts.

	TG1	GHGSat	GHOSat	EMIT	EnMAP	PRISMA	WV3	GF5	ZY1	LS	S2	MS
Carbon Mapper	X											
GHGSat		X										
NASA-JPL				X								
UPV				X	X	X						
IMEO				X	X	X				X	X	
Google				X								
NJU				X	X	X		X	X	X	X	
Kayros				X	X	X				X	X	
SRON				X	X	X						
Orbio Earth				X						X	X	
ARKA					X	X					X	
IUB				X	X	X						
Orbital Sidekick			X									
Momentick	X			X	X	X	X			X	X	
Geolabe											X	
MethaneSAT												X

S2.1 Participating satellites

Twelve satellite systems collected at least one measurement during the study period. Extended Data Table 1 summarizes key technical characteristics of each instrument. We have broadly grouped participating satellites into three groups: Group A comprises instruments that were purpose built for methane detection, which includes GHGSat and Tanager-1 (Carbon Mapper); Group B comprises multi-purpose scientific and commercial instruments with hyperspectral and/or tasking capabilities that have been repurposed toward methane monitoring, including EnMAP, Gaofen 5, PRISMA, EMIT, WorldView-3, Ziyuan 1; Group C comprises multi-purpose multispectral global scanning satellites Sentinel-2 and Landsat 8/9. Each observing system is described briefly below.

MethaneSat, a project of the Environmental Defense Fund launched in 2024, was designed to make high fidelity methane concentration measurements at regional scale. Unfortunately, the MethaneSat instrument failed in June 2025. No results were submitted for the 12 MethaneSat releases that were performed during Phase 1 and 2. Reasons for lack of result submission were not made clear to researchers.

The TROPOMI instrument aboard the Sentinel-5P satellite was excluded from this study because its expected detection threshold exceeds the maximum emission rates achievable by the controlled release system used here.

Group A satellites:

GHGSat

The GHGSat network is a constellation of targeted greenhouse gas-monitoring satellites operated by GHGSat Inc., a private Canadian company. Since the launch of GHGSat-D precursor satellite in 2016, the constellation has grown to sixteen satellites. Additional satellites are scheduled to launch in the coming years. The payload is a Wide-Angle Fabry-Perot imaging spectrometer, operating in the Shortwave Infrared (SWIR) region to resolve methane and CO₂ absorption features with 0.3 nm spectral resolution. The satellites typically image specific industrial sites of approximately 12x12 km with a spatial resolution of 25 m. The expanded GHGSat constellation size resulted in a frequent revisit rate to the experiment site of approximately 1-2 days, with some days having multiple overpasses. Data from the GHGSat constellation is available commercially through subscription-based services as well as through select scientific partnerships ².

Tanager-1 (Carbon Mapper)

Tanager-1 (TG1) is the first satellite in the planned Carbon Mapper constellation, launched in early 2024. The mission is a public-private partnership between the non-profit Carbon Mapper Coalition and Planet Labs, which built and operates the satellite. Tanager-1 is a targeted hyperspectral imaging spectrometer with 5 nm spectral sampling, a spatial resolution of 30 m and a swath width of approximately 20 km at target altitude of 406 km. Tanager-1 underwent orbit lowering during various study phases (e.g. 480 km to 430 km in Phase 1), resulting in some changes in spatial resolution over the course of the experiment.

Each Tanager satellite is capable of imaging up to 250,000 km² per day. Tanager-1 had a release site revisit rate of approximately 1 week. Tanager-1 can be operated in multiple modes; for this study, Carbon Mapper used mostly maximum sensitivity mode where the site is imaged 4 times in rapid succession. This system is designed to detect, quantify, and attribute point source emissions of both methane and carbon dioxide. Plume data from the Carbon Mapper constellation is made publicly available through its online data portal with data latency of 30 days. Other data product levels (L1) are available for purchase³. The constellation is expanding, with the next three Tanager satellites currently in assembly ^{4,5}.

90 Group B satellites:

EMIT

The Earth Surface Mineral Dust Source Investigation (EMIT) is a NASA mission developed by the Jet Propulsion Laboratory (JPL). Its primary objective is to map the mineral composition of Earth's arid dust source regions to better understand their role in climate ⁶. The instrument is a hyperspectral imaging spectrometer mounted on the International Space Station (ISS). Due to the ISS's non-sun-synchronous orbit, its overpass schedule over any given location is irregular and less predictable than traditional polar-orbiting satellites. Also, the ISS orbit is less deterministically predictable than other satellites, and observations were not frequent during the study period. EMIT provides hyperspectral data with a spectral sampling of approximately 7.4 nm, a spatial resolution of 60 m, and a swath width of 75 km. Data is available via NASA-JPL's EMIT Open Data Portal⁷.

EnMAP

The Environmental Mapping and Analysis Program (EnMAP) satellite is a German hyperspectral satellite mission managed by the German Aerospace Center (DLR) ⁸. Launched in April 2022, the mission's primary goal is to monitor and characterize the Earth's environment on a global scale. The satellite carries a hyperspectral imaging spectrometer with spectral bands from 420-2,450 nm at 10 nm resolution, spatial resolution of 30 m, and a swath width of 30 km. EnMAP has a maximum revisit frequency of 4 days. Tasking acquisitions are prioritized based on approved user proposals submitted via the EnMAP portal. EnMAP data catalogue is also available via the Data Access Portal⁹.

Gaofen 5

The Gaofen-5 (GF5) satellite series is operated by the China National Space Administration (CNSA). This study included the Gaofen 5-02 instrument, launched in September 2021. The satellite is equipped with an Advanced Hyperspectral Imager that provides data with a spatial resolution of 30 m over a 60 km swath ¹⁰. While prior experiments suggested a potential revisit capability of approximately 7 days, tasking opportunities over the test site for this study were sparse; the experiment team was only informed of a handful of overpasses in advance. Data from GF5 is available upon request from the relevant Chinese government agencies¹¹.

GHOST

120 The GHOST (Global Hyperspectral Observation Satellite) constellation is commercially operated by the U.S.-
based company Orbital Sidekick. The constellation consists of four pushbroom imaging spectrometers designed
to provide high-resolution hyperspectral imagery. The instruments feature a high spatial resolution of 8 m over a
4 km swath ¹². Based on the observation opportunities provided during this study, the constellation demonstrated
an effective revisit capability of approximately 2-3 days. Data from the GHOST constellation is proprietary and
125 available commercially¹³.

PRISMA

PRISMA (PRecursore IperSpettrale della Missione Applicativa) is a hyperspectral satellite mission of the Italian
Space Agency (ASI), launched in March 2019 ¹⁴. Its primary mission objectives are imaging of natural resources
130 and atmospheric characteristics. The instrument includes 173 bands in the NIR/SWIR range of 920-2505 nm at
about 10-12 nm resolution. Its spatial resolution is 30 m over a 30 km swath and has a maximum revisit
frequency of 7 days. PRISMA acquisitions can be tasked upon request, and the mission's data can be requested
from ASI's catalogues¹⁵.

WorldView-3

WorldView-3 (WV3) is a high-resolution commercial Earth observation satellite operated by Vantor, launched in
August 2014. The instrument can be tasked to specific targets with a maximum 1-day revisit capability, although
site revisit in this study occurred approximately every 7 days. WorldView-3 is a multispectral sensor with eight
SWIR bands at 50nm resolution, providing imagery over a 13 km swath at 3.7 m spatial resolution. As a
140 commercial mission, new acquisitions can be tasked upon request; data is available for purchase through Vantor
and its resellers¹⁶.

WorldView-3 is somewhat unique in this study due to its very high spatial resolution, about 1 order of magnitude
finer spatial resolution than the majority of other instruments. Although it is the only multispectral instrument in
145 Group B, it is categorized there with other high-resolution, taskable instruments to avoid a single-instrument
category.

Ziyuan-1

The Ziyuan-1 (ZY-1) series of satellites is operated by the China National Space Administration. This study
150 included the ZY1-02D (launched 2019) and ZY1-02E (launched 2021) satellites. Each satellite is equipped with
an Advanced Hyperspectral Imager providing data with a spatial resolution of 30 m and a swath width of 60 km.
Relevant methane observing wavelength ranges include 90 SWIR bands between 1005–2501 nm and a spectral
resolution of 20 nm ¹⁷. Prior controlled release studies suggested a revisit capability of 1-2 days ¹⁸. However,
observation opportunities over the test site in this study were sparse, and the experiment team was only informed
155 of a small number of overpasses in advance. Data access can be requested with an application through China
Center for Resources Satellite Data and Application¹¹.

Group C satellites:

160 **Sentinel-2**

The Sentinel-2 mission of the European Union's Copernicus Earth observation program is designed for global land monitoring. The constellation comprises Sentinel-2A (launched 2015), Sentinel-2B (launched 2017), and the recently added Sentinel-2C (launched 2024). The satellites carry multispectral instruments and provide broad global coverage with a wide 290 km swath. For methane detection, key spectral information typically comes from
165 two bands in the SWIR region: Band 11 (1610 nm) and Band 12 (2190 nm), both with a spatial resolution of 20 m. Data from the Sentinel-2 mission is openly available through several platforms, including the Copernicus Open Access Hub and cloud-based providers like Google Earth Engine¹⁹.

Landsat 8/9

170 The Landsat program is the world's longest-running civilian Earth observation mission, jointly managed by NASA and the U.S. Geological Survey²⁰. This study utilizes data from the two currently operational satellites, Landsat 8 (launched 2013) and Landsat 9 (launched 2021), which fly in an 8-day offset orbit. Both satellites provide data at a 185 km swath width across several spectral bands. Methane retrievals typically rely on Band 6 (1610 nm) and Band 7 (2200 nm), both delivered at a spatial resolution of 30 m. The Landsat data archive is
175 publicly available through the USGS EarthExplorer data portal and various cloud platforms²⁰.

S2.2 Participating analyst teams

Fifteen analysis teams participated in at least one phase of the single-blind study. Each team analyzed data from between one and seven satellites. We invited all teams that were known by the Stanford/UM researchers to
180 estimate methane emissions from the participating satellites. The GEISAT satellite (Spain) operated by Satlantis, and the analysis team at Satelytics were contacted but declined participation in the study. Dr. Daniel Varon at Massachusetts Institute of Technology, previously of Harvard University, who participated in previous single blind testing (Sherwin et al., 2023), was invited but declined to participate in this round of testing due to limited time availability.

185 The technical contributions of the analysis teams are foundational to the results presented in this paper. The authorship structure for this single-blind study was a deliberate component of the experimental design, established in agreement with all participants prior to the campaign. The primary objective of this structure is to ensure the complete independence and scientific integrity of the final evaluation, avoiding any real or perceived conflict of interest. This model is not unique to our study; it adheres to a well-established precedent set by other
190 major single-blind intercomparison studies, including previous work from our own group as well as studies from other leading research institutions like Colorado State University's METEC facility^{18,21}.

Below, we briefly describe each participating organization, alphabetically. Further details on each team and methods can be found in the Performer Info spreadsheets in the data and code repository.

ARKA

ARKA Group, L.P. is a privately owned aerospace and defense company that acquired Maxar's Radar and Sensor Technology division in 2024. The ARKA team analyzed data from Sentinel-2, EnMAP, and PRISMA in this study, submitting results exclusively for Phase 1 of the blind test. The group was also a participant in the Sherwin et al. (2024) controlled release study.

ARKA's workflow includes a proprietary Methane Mapper algorithm to retrieve methane concentration enhancements from the satellite imagery. Plumes are identified and delineated via manual analyst inspection of the methane enhancement image. Mass flow rate is calculated using the Integrated Methane Enhancement (IME) method. Wind speed and direction data were sourced from the ERDDAP 7 reanalysis product and interpolated to the release stack's location and satellite overpass time.

Carbon Mapper

Carbon Mapper is a non-profit organization based in Pasadena, California, USA. The Carbon Mapper team participated in all four study phases, processing Tanager-1 data with an in-house software package (cmpro).

Methane enhancement retrieval is performed using a matched filter algorithm, with the full methodology described in their public Algorithm Theoretical Basis Document (https://assets.carbonmapper.org/documents/L2B%20Algorithm%20Theoretical%20Basis%20Document_formatted_10-24-25.pdf). Plume segmentation is performed via an automated process involving calculating a dynamic noise background for each scene and then identifying plumes based on a set of rules including pixel clumping, connected components, and proximity to the potential source location. The emission rate is quantified using the IME method. Wind speed data are derived from the High-Resolution Rapid Refresh (HRRR) reanalysis product, and no effective wind speed modulation is applied.

Geolabe

Geolabe LLC is from Los Alamos, New Mexico, USA. Geolabe specializes in developing artificial intelligence (AI) approaches for multispectral and radar satellite imagery. In this study, Geolabe analyzed data from Sentinel-2, participating only in Phase 2. Their methodology uses a deep learning model designed for plume detection in Sentinel-2 imagery (plumenet v63.2). The model was trained on a large dataset of synthetic Gaussian plumes, which were embedded into real Sentinel-2 imagery to create a realistic training set. A full description of this method is available in Rouet-Leduc et al. (2024). For quantification, wind data was sourced from NASA's GEOS reanalysis product and local weather stations near the test site (specifically KAZCASAG250, KAZCASAG158, and KAZCASAG216).

GHGSat

GHGSat Inc. is a Canadian company that operates a commercial methane-monitoring constellation. The team participated in all four phases of the study, analyzing data exclusively from their own satellites. Overpass opportunities were scheduled and coordinated directly with the GHGSat team. The data analyzed in this study were collected using satellites running firmware version 10.29.4 and a specific observation script (N5138CC1.GSB).

GHGSat's source rate retrieval algorithm is described in Jervis et al.²². Two methods were used for plume masking. The primary method uses a floodfill algorithm, initiated from the emission source location. For smaller plumes (e.g., the release on March 29), analysts performed manual pixel selection to define the mask. Wind data for quantification were sourced from the OpenWeatherMap API.

Google

The Google team participated in Phases 2 and 4 of the study, analyzing data from the EMIT instrument. Their approach uses a machine learning (ML) model designed to retrieve methane concentrations directly from satellite imagery. The model is trained on a synthetic plume dataset. Plumes are generated using a wind and plume dispersion model and are then embedded into real EMIT imagery. Model details are expected in a forthcoming publication in 2026 (email correspondence V. Batchu).

The ML model's inputs include Level 1B radiance data and with associated observing metadata. Once the model predicts a methane enhancement map, a plume mask is generated via simple thresholding (in Phase 2). A later version of their model used in Phase 4 also predicted plume instance masks (instead of simple thresholding). The final emission rate is calculated using IME. Wind data is sourced from the ERA5 reanalysis product²³.

International Methane Emissions Observatory (IMEO)

IMEO participated through its Methane Alert and Response System (MARS) initiative. IMEO analyzed data from EMIT, Sentinel-2, Landsat, EnMAP, and PRISMA across all four phases of testing. The IMEO methodology is built around the MARSML (v1.24) module, which integrates methane enhancement retrieval, AI-based plume detection, and IME-based flux rate quantification.

The retrieval approach is sensor-dependent: for hyperspectral data (EMIT, EnMAP, PRISMA), a matched filter based on Roger et al. (2023) is used, while a double-ratio multiband multi-pass method is used for multispectral data²⁴. Following the retrieval and an initial AI-based detection, MARS generates alerts that are passed to Plume Viewer (v1.27), a web-based application designed for analyst review and editing. Any modification prompts MARSML to re-calculate the flux rate. For quantification, 10 m wind speed data are sourced from multiple reanalysis products and adjusted to an effective wind speed following linear relationships described in the literature^{25,26}.

270 IMEO informed the experiment team that they updated coefficients for EMIT effective windspeed in Phase 4. IMEO's quantified estimates from Phase 2 shown are the results at the time of submission and thus reflect the prior method. A revision of Phase 2 estimates using the new coefficients leads to a downward revision and a reduction in fitted slope. See IMEO technical documentation at <https://wedocs.unep.org/items/c77a439f-9689-4164-874d-d49045e6d097>.

275

Institute of Environmental Physics (IUP), University of Bremen

The Institute of Environmental Physics (IUP) at the University of Bremen, Germany, analyzed data from EMIT, EnMAP, and PRISMA, participating in all four phases of the study.

280 Their methodology pairs a matched filter retrieval using High Resolution Fit (HiFi) software with a Cross-Sectional Flux (CSF) method for quantification (CSF v1.6). The CSF algorithm does not require a discrete plume detection and masking step. Instead, it operates by defining a series of cross-sections orthogonal to the wind direction over the target location. The final emission rate (Q) is the average of these individual cross-sectional fluxes, and a 1-sigma uncertainty for Q is also calculated. IUP uses the 10 m wind speed from the ERA5
285 reanalysis product; no correction factor is applied ($U_{\text{eff}} = U_{10}$).

The IUP workflow includes a quality flagging system that flags results based on conditions such as low wind speed (≤ 1 m/s), insufficient data coverage ($< 90\%$ valid pixels), or too few independent cross-sections. For the analysis in this paper, all submitted IUP results were included regardless of their quality flag status; however, the
290 flags are provided for reference in the supplementary data files.

Jet Propulsion Laboratory

The NASA Jet Propulsion Laboratory (JPL) team analyzed data from the EMIT instrument, participating in Phases 2 and 4 of the study. Their methodology involves examination of the Level 2A methane enhancement data
295 in the vicinity of the known release site for visible enhancements. Plume masks are generated using image thresholding. Quantification is performed using IME. Wind data are sourced from the HRRR 10 m product, with values linearly interpolated in space and time to the plume location. The final emission rate uncertainty is calculated by propagating uncertainties from: 1) the concentration-length uncertainty and 2) the wind speed uncertainty, which is defined as the standard deviation of the nine nearest HRRR grid points surrounding the
300 source location.

Kayrros

Kayrros is a French company specializing in satellite-based global environmental monitoring. They analyzed EMIT, EnMAP, PRISMA, Sentinel-2, and Landsat during all four phases of testing.
305

For hyperspectral sensors, their algorithm is based on the matched filter formulation described in Guanter et al.²⁵. For multispectral sensors, the method leverages the correlation between two methane-relevant bands and

background removal to enhance the methane signal. Plume delineation is performed manually. Every confirmed detection is subsequently quantified using IME method ²⁶. Wind estimates were derived from the European
310 Centre for Medium-Range Weather (ECMWF) ERA5 reanalysis product. The 10 m wind speed is converted to an effective wind speed (U_{eff}) using the formula $U_{eff} = 0.34 * U_{10} + 0.44$, described in Guanter et al. (2021).

MethaneSAT

MethaneSAT, a subsidiary of the Environmental Defense Fund (EDF), coordinated and executed targeted
315 overpasses of the release site during Phases 1 and 2. Unfortunately, the MethaneSAT mission ended prematurely in June 2025. The MethaneSAT team did not submit results for single-blind evaluation.

Momentick

Momentick is an Israel-based company specializing in Earth observation for emissions monitoring. Momentick
320 analyzed data from WorldView-3, PRISMA, EnMAP, EMIT, Landsat, and Sentinel-2 during all four phases. They also analyzed Tanager-1 data in Phase 4. Their detection and quantification workflows are based on proprietary deep learning models. Wind data are sourced from the ECMWF ERA5 reanalysis product.

Nanjing University

325 The team from Nanjing University (NJU) analyzed EMIT, PRISMA, EnMAP, Gaofen-5, Ziyuan-1, Landsat, and Sentinel-2 in Phases 1-4. NJU uses a matched filter for methane concentration retrieval. Plume masks are generated semi-automatically using a watershed segmentation technique applied to a denoised methane enhancement map.

330 To derive the wind speed for quantification, an effective wind speed (U_{eff}) is calculated from the 10 m wind speed (U_{10}) sourced from the GEOS-FP reanalysis product. The calculation uses the formula: $U_{eff} = 0.34 * U_{10} + 0.44$. The final emission rate is calculated using IME.

Orbio Earth

335 Orbio Earth is a Germany-based company founded in 2021 focused on monitoring global methane emissions from the oil and gas industry using satellite imagery. The team participated in three out of four phases of testing, analyzing data from EMIT, Landsat, and Sentinel-2.

Their methodology is centered on an open-source codebase that applies computer vision (CV) models to each
340 sensor. For multispectral imagery, the CV model generates 1) a binary classification map indicating the probability of each pixel belonging to a methane plume, and 2) a regression layer that predicts the fractional reduction in the ratio of the two SWIR bands. This fractional reduction is then converted into a physical concentration retrieval (in mol/m^2). For EMIT, the CV model is trained to predict both the binary classification map and the final concentration retrieval map. For all sensors, once the retrieval is complete, plume segmentation
345 is performed using a watershed algorithm applied to the binary probability mask. Wind data are sourced from the

ERA5 reanalysis product, and the final emission rate is quantified using IME method. More details can be found at: <https://orbio-earth.github.io/Project-Eucalyptus/>.

Orbital Sidekick

350 Orbital Sidekick (OSK) is a U.S.-based company that operates the GHOST constellation and provides commercial monitoring services for sustainability and security objectives. In this study, OSK analyzed data from its own GHOST satellite constellation in Phases 1-4. Their data processing workflow is proprietary but includes steps based on matched filtering for methane retrieval and the IME method for quantification. Wind data is sourced from the High-Resolution Rapid Refresh (HRRR) weather model.

355

SRON

SRON, Space Research Organization Netherlands, analyzed EnMAP, PRISMA, and EMIT in all four rounds of the study. Their data processing is performed using the HyperGas v0.2 software package, which is based on the methodology detailed in their preprint (doi.org/10.5194/egusphere-2025-6127), with additional improvements
360 noted in their submission files.

The workflow uses a matched filter retrieval which incorporates water masks. Following retrieval and denoising of the methane enhancement map, plume detection is performed manually. The team also made a point to note instances where a faint enhancement was visible at the source location but would not have been confidently
365 reported in an operational "in the wild" scenario. Once a plume is confirmed, a mask is generated using a watershedding technique. The source location is defined as the most upwind high-enhancement pixel within the mask.

Quantification is performed using the IME method. Wind data is sourced from the ERA-5 10 m reanalysis
370 product. Sensor-specific calibrations are used to determine the effective wind speed (U_{eff}). For EMIT, the U_{eff} is calculated as $0.35 \times U_{10} + 0.43$. For EnMAP, the formula used is $U_{\text{eff}} = 0.32 \times U_{10} + 0.45$, and for PRISMA, it is $U_{\text{eff}} = 0.31 \times U_{10} + 0.43$.

Universitat Politècnica de València (UPV)

375 The team from the Universitat Politècnica de València (UPV), Spain participated in all four study phases, analyzing data from PRISMA and EnMAP. UPV's methodology for detection is a combo-matched filter²⁷. Enhancements are visually inspected, and plumes are manually delineated. For quantification, a different retrieval algorithm is used, based on two lognormal matched filters that leverage different unit methane absorption spectra²⁸. The final emission rate is calculated using the IME method, which combines the quantification retrieval, the
380 manual delineation, and wind data. Wind information is sourced from the GEOS-FP reanalysis 10 m wind speed (U_{10}) product. For EnMAP and PRISMA, the formula is $U_{\text{eff}} = 0.34 * U_{10} + 0.44$ ^{25,27}; the formula for EMIT is: $U_{\text{eff}} = 0.31 * U_{10} + 0.40$ ²⁹.

385 **S2.3 Sources of methodological variance**

It is well-documented that different analysis teams can produce divergent quantification results from identical satellite data¹⁸. This variance stems from diverse methodological choices, including retrieval algorithms, data filtering thresholds, plume masking, and wind data selection (Table S3). Recognizing this challenge, the remote sensing community has launched major, multi-team intercomparison initiatives, such as the ESA Project

390 MEDUSA and the Data Foundation’s Technical Working Groups, which aim to systematically decompose these sources of uncertainty^{30,31}.

It should be noted that a full decomposition of methods was not a primary goal of this paper, partly because our study design respected the proprietary nature of some commercial algorithms to ensure broad participation. Our

395 study contributes to external intercomparison efforts by providing a well-characterized emissions dataset, which is often lacking. The data presented here provide a resource for future studies aimed at decomposing sources of analytical variance.

Table S3. Participating teams’ wind data source

Team	Wind reanalysis product(s)
Arka	ERDDAP* (NOAA)
Carbon Mapper	HRRR (NOAA)
Geolabe	GEOS-FP (NASA)
GHGSat	OpenWeatherMap**
Google	ERA5 (ECMWF)
IMEO	ERA5 (ECMWF), GEOS-FP (NASA)
IUP	ERA5 (ECMWF)
JPL	HRRR (NOAA)
Kayros	ERA5 (ECMWF)
Momentick	ERA5 (ECMWF)
NJU	GEOS-FP (NASA)
Orbio Earth	ERA5 (ECMWF)
Orbital Sidekick	HRRR (NOAA)
SRON	ERA5 (ECMWF)
UPV	GEOS-FP (NASA)

400 *ERDDAP is a data server hosted by NOAA. The reanalysis product wasn’t specified by ARKA.

**OpenWeatherMap is a commercial weather data service and aggregator. It draws on multiple sources including weather stations, ERA5 reanalysis product, and its own proprietary modeling.³²

405 **S3. Supplementary results**

S3.1 Submission dates

During Phase 1, 12 teams submitted results: ARKA, Carbon Mapper, GHGSat, IMEO, IUP-UB, Kayrros, Momentick, NJU, Orbio Earth, Orbital Sidekick, SRON, and UPV. In Phase 2, 14 teams submitted results: the same 12 teams as Phase 1 minus ARKA plus Google, Geolabe, and JPL. In Phase 3, 11 teams submitted results: the same 12 teams as Phase 1 minus ARKA. And in Phase 4, 12 teams submitted results: the same 12 teams as Phase 1 minus ARKA plus Google and JPL. Teams were permitted to submit results at any time during the 30-day window following the final release of the phase (i.e., throughout April 2025 for Phase 1, throughout July 2025 for Phase 2, throughout October 2025 for Phase 3, and throughout January 2026 for Phase 4). The date of each team’s initial submission is provided in the table below.

The experiment team reviewed submissions for completeness and followed up with teams in cases where materials were missing or needed clarification. This prompted some teams to submit revised versions, which are also noted in the table. In some cases, minor revisions were submitted after the 30-day window as the experiment team was processing and checking submissions. Slightly late responses were accepted if the ground truth results had not yet been unblinded. In Phase 1, Geolabe attempted to submit results after some data had been unblinded, and their data were therefore not accepted. Similarly, NJU submitted three additional data points for PRISMA on May 11, after unblinding had occurred; these were also excluded from analysis.

Table S4. Results submission dates for each team by phase, listed alphabetically

Analysis team	Phase 1 (Date of first submission, Date of last accepted edits)	Phase 2 (Date of first submission, Date of last accepted edits)	Phase 3 (Date of first submission, Date of last accepted edits)	Phase 4 (Date of first submission, Date of last accepted edits)
ARKA (ARK)	4/9/2025, 4/28/2025	n/a	n/a	n/a
Carbon Mapper (CM)	4/29/2025	7/25/2025	10/30/2025	1/27/2026
Geolabe (GEO)	n/a	8/1/2025, 8/4/2025	n/a	n/a
Google (GOO)	n/a	7/30/2025, 7/31/2025	n/a	2/1/2026 2/3/2026
GHGSat (GHG)	4/30/2025	7/30/2025, 8/5/2025	10/31/2025	1/30/2026
IMEO	4/18/2025, 4/24/2025	7/17/2025, 7/18/2025	10/6/2025, 10/9/2025	1/22/2026, 1/29/2026, 2/4/2026
IUP-UB	4/16/2025, 4/23/2025	7/9/2025, 7/14/2025	10/9/2025, n/a	1/12/2026, 1/15/2026
JPL	n/a	7/25/2025, n/a	n/a	1/29/2026
Kayrros (KAY)	4/30/2025, n/a	7/30/2025, n/a	10/31/2025, 11/3/2025	1/30/2026
MethaneSAT (MS)	n/a	n/a	n/a	n/a
Momentick (MO)	4/29/2025	7/31/2025, 8/1/2025	10/30/2025, 11/2/2025	1/28/2026
NJU	5/1/2025, 5/2/2025	8/1/2025, n/a	10/31/2025, 11/3/2025	1/29/2026
Orbio Earth (OE)	4/30/2025, n/a	7/25/2025, 7/31/2025	10/28/2025, 10/31/2025	n/a
Orbital Sidekick (OSK)	4/30/2025, n/a	7/29/2025, n/a	10/31/2025	1/31/2026

SRON	4/30/2025, n/a	7/29/2025, n/a	10/31/2025	1/29/2026
UPV	4/20/2025, 4/24/2025	7/17/2025, n/a	10/9/2025, n/a	1/27/2026

S3.2 Exclusion criteria

The Stanford/University of Michigan experiment team (henceforth, the experiment team,) excluded release events or intentional non-release events from consideration in results if:

1. A release was severely mis-timed. For example, if it began much later than the intended start time. Releases were designed to last 8-12 minutes on average to provide sufficient time for plume buildup. Occasionally, delays were caused by issues with remotely accessing the heater or flow control equipment.

2. The experiment team learned about overpasses during the results submission period and a release had been occurring within or near the overpass window for another instrument or continuous monitor.

3. The metered flow standard deviation, accounting for all sources of variability and uncertainty, exceeded 5% of the mean flow. This occurred for one Landsat release on September 8, 2025. These points were still considered for detection but removed from quantification.

Note on unforeseen overpasses: In several instances, the experiment team was not aware of a satellite overpass prior to becoming aware of it during the reporting period. This happened with Sentinel-2A during Phase 1 of the experiment when Sentinel-2C was phased in and made the Sentinel-2A overpass schedule unpredictable. It also occurred occasionally with taskable commercial instruments like GHOS_t where a new observation opportunity was not communicated in advance. In these cases, we adopted a “data centric” approach—if an analyst team processed the satellite data while remaining fully blind to the ground-truth conditions, their results were included in the final analysis, provided the event did not violate the three exclusion criteria listed above.

Analyst teams were allowed to exclude measurements for any reason. The submission sheet provided a space to specify reason for exclusion (see reporting template below). The most common reason was cloud cover.

In a handful of instances, a team submitted multiple responses corresponding to a single release event, typically when multiple images were captured in rapid succession during a single satellite pass. For detection analysis, these submissions were consolidated into a single outcome. In all such cases, the duplicate detection outcomes were identical (e.g. both were True Positives). If multiple quantified estimates were provided for a release, these are treated as separate points on the parity charts.

S3.3 Release and non-release event outcomes

3.3.1 Detection summary

Table S5. Detection results by satellite, team, and phase. Satellites GH: GHGSat, GO: GHOST, TG: Tanager, EN: EnMAP, ET: EMIT, PR: PRISMA, GF: Gaofen5, WV: WorldView, ZY: Ziyuan, LS: Landsat, S2: Sentinel-2. *NT* denotes not tasked/no imagery, *AT* denotes analysis team (acronyms above), and *ET* denotes experiment team. Any differences between total counts and those reported in Figure 2 of the main text are missing responses.

Sat	AT	True positive (TP)				False negative (FN)				True negative (TN)				False positive (FP)				NT	Cloudy				ET filtered	AT filtered				Total
		P1	P2	P3	P4	P1	P2	P3	P4	P1	P2	P3	P4	P1	P2	P3	P4		P1	P2	P3	P4		P1	P2	P3	P4	
TG	CM MO	6	10	9	6	0	0	0	0	1	0	2	2	0	0	0	0	4	1	1	2	1	0	0	0	0	0	45
		-	-	-	6	-	-	-	0	-	-	-	1	-	-	-	0	4	0	0	0	2	0	0	0	0	0	13
GH	GHG	11	15	9	12	2	1	0	1	1	2	1	4	1	0	0	0	5	9	9	13	11	0	0	0	0	0	107
EN	ARK	9	-	-	-	0	-	-	-	2	-	-	-	0	-	-	-	18	2	-	-	-	0	0	-	-	-	31
	IMEO	9	6	7	7	2	0	1	4	2	1	1	0	0	0	0	0	18	0	1	3	1	0	0	1	1	1	66
	IUP	9	7	9	10	0	1	2	1	0	0	2	0	2	1	0	1	18	2	0	0	1	0	0	0	0	0	66
	KAY	9	7	6	7	0	0	2	4	2	1	1	0	0	0	0	0	18	2	1	4	2	0	0	0	0	0	66
	MO	9	7	7	9	1	0	2	2	2	0	1	0	0	1	0	0	18	1	1	3	2	0	0	0	0	0	66
	NJU	9	7	9	8	0	0	0	3	2	1	1	0	0	0	0	0	18	2	1	3	2	0	0	0	0	0	66
	SRON	9	7	8	7	0	0	2	4	2	1	1	0	0	0	0	0	18	2	1	2	2	0	0	0	0	0	66
	UPV	9	7	5	7	0	0	3	4	2	1	1	0	0	0	0	0	18	2	1	4	2	0	0	0	0	0	66
ET	GOO	-	3	-	2	-	2	-	0	-	0	-	0	-	0	-	0	3	-	0	-	0	0	-	0	-	0	10
	IMEO	-	5	0	2	-	0	0	0	-	0	2	0	-	0	-	0	3	-	0	0	0	0	-	0	-	0	12
	IUP	-	5	0	2	-	0	0	0	-	0	2	0	-	0	0	0	3	-	0	0	0	0	-	0	0	0	12
	JPL	-	3	-	2	-	2	-	0	-	0	-	0	-	0	-	0	3	-	0	-	0	0	-	0	-	0	10
	KAY	0	3	0	2	0	2	0	0	0	0	2	0	0	0	0	0	3	-	0	0	0	0	-	0	0	0	12
	MO	0	4	0	2	0	1	0	0	0	0	1	0	0	0	1	0	3	-	0	0	0	0	-	0	0	0	12
	NJU	-	4	0	2	-	1	0	0	-	0	2	0	-	0	0	0	3	-	0	0	0	0	-	0	0	0	12
	OE	0	3	0	-	0	2	0	-	0	0	1	-	0	0	1	-	3	-	0	0	-	0	-	0	0	-	10
	SRON	0	4	0	2	0	1	0	0	0	0	2	0	0	0	0	0	3	-	0	0	0	0	-	0	0	0	12
	UPV	-	3	0	2	-	1	0	0	-	0	1	0	-	0	0	0	3	-	1	1	0	0	-	0	0	0	12
PR	ARK	7	-	-	-	0	-	-	-	1	-	-	-	0	-	-	-	15	0	-	-	-	0	0	-	-	-	23
	IMEO	7	4	5	3	0	1	1	2	1	1	1	1	0	0	0	0	15	0	1	1	0	0	0	1	0	0	45
	IUP	6	6	7	4	1	1	0	1	0	0	0	1	1	1	1	0	15	0	0	0	0	0	0	0	0	0	45
	KAY	6	6	4	2	1	0	2	3	1	1	1	1	0	0	0	0	15	0	1	1	0	0	0	0	0	0	45
	MO	6	6	5	2	1	0	1	3	1	1	1	1	0	0	0	0	15	0	1	1	0	0	0	0	0	0	45
	NJU	3	5	4	2	1	1	2	3	1	1	1	1	0	0	0	0	15	0	1	1	0	0	0	0	0	0	41
	SRON	6	6	4	2	1	0	2	3	1	1	1	1	0	0	0	0	15	0	1	1	0	0	0	0	0	0	45
	UPV	6	5	4	2	1	1	2	3	1	1	1	1	0	0	0	0	15	0	1	1	0	0	0	0	0	0	45
GF	NJU	1	-	-	-	0	-	-	-	0	-	-	-	0	-	-	-	0	0	-	-	-	0	0	-	-	-	1
GO	OSK	5	6	7	0	3	5	1	2	7	15	5	2	1	0	1	0	70	1	8	7	5	2	0	0	2	2	157
WV	MO	3	4	5	3	1	0	0	1	2	0	1	0	0	1	0	1	2	1	0	2	1	0	0	0	0	0	28
ZY	NJU	1	-	-	-	0	-	-	-	0	-	-	-	0	-	-	-	1	1	-	-	-	0	0	-	-	-	3
LS	IMEO	1	0	0	2	6	8	8	3	7	11	7	8	0	0	0	0	1	8	2	7	6	3	0	0	0	4	92
	KAY	2	3	1	1	7	5	8	6	9	12	9	10	0	0	0	0	1	4	1	4	3	3	0	0	0	3	92
	MO	4	6	8	6	3	2	1	1	5	10	9	8	1	2	1	3	1	9	1	3	5	3	0	0	0	0	92
	NJU	0	0	0	0	6	8	11	5	9	12	11	7	0	0	0	0	1	7	1	0	11	3	0	0	0	0	92
	OE	7	6	7	-	1	2	2	-	4	6	4	-	3	4	3	-	1	7	2	5	-	3	0	1	1	-	69
S2	ARK	4	-	-	-	10	-	-	-	11	-	-	-	3	-	-	-	1	10	-	-	-	0	0	-	-	-	39
	GEO	-	8	-	-	-	5	-	-	-	12	-	-	-	1	-	-	1	-	10	-	-	0	-	9	-	-	46
	IMEO	0	0	0	1	13	19	20	10	14	17	11	17	0	0	0	0	1	11	9	13	7	1	0	0	2	10	176
	KAY	4	5	1	2	10	14	20	13	13	17	13	19	0	0	0	0	1	12	9	12	11	1	0	0	0	0	176
	MO	8	12	11	6	6	7	10	9	11	17	10	13	3	1	3	5	1	9	8	12	11	1	0	0	0	1	175
	NJU	0	0	0	0	13	18	23	13	13	17	17	16	0	0	0	0	1	12	10	6	16	1	0	0	0	0	176
	OE	7	15	16	-	5	3	4	-	8	5	6	-	5	11	7	-	1	9	8	12	-	0	4	3	1	-	130

3.3.2 Detection outcome by release size

475 Figure S1 is alternative version of Figure 1 that appears in the main text that adopts a stricter definition of non-detection. Here, non-detections include missed releases for any reason, including instances of failed tasking and analyst-filtered datapoints due to clouds. As expected, true positive rates are consistently lower.

480 Figure S2 breaks out detection performance by team and satellite, binned by release rate. Some teams participated in different study phases and filtered different releases, leading to different counts for the same satellite. Figure S3 aggregates these results by satellite.

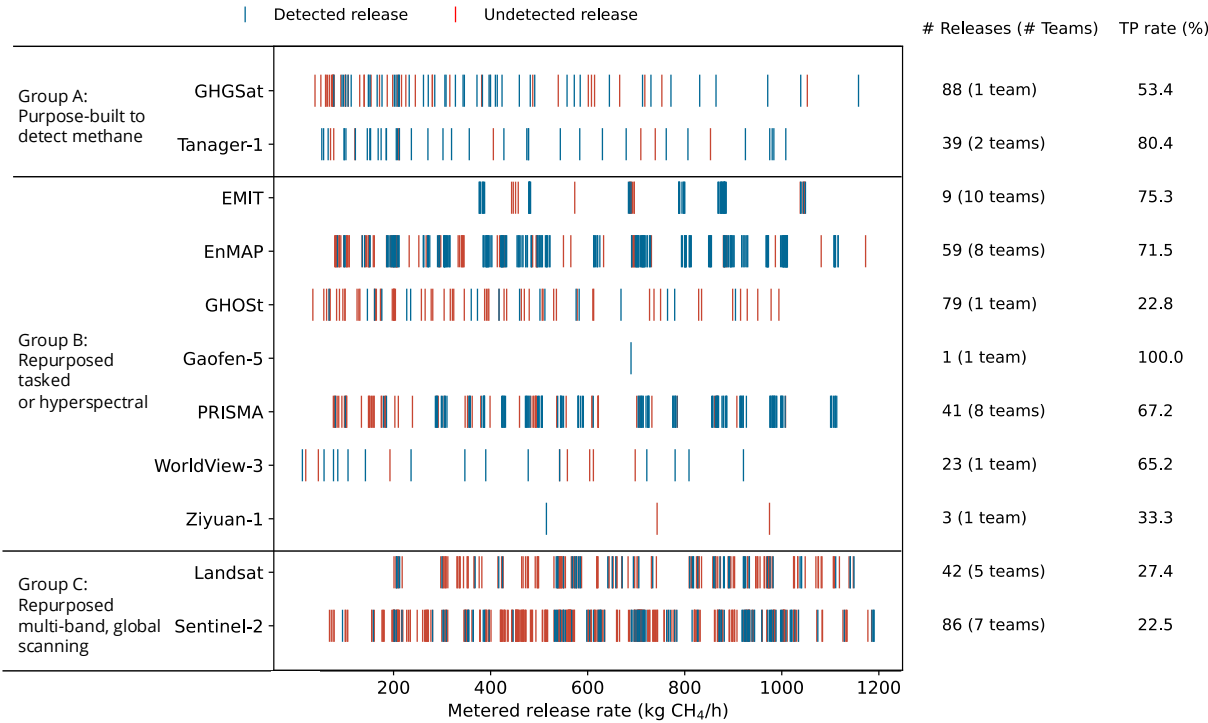
We did not derive formal Probability of Detection (POD) curves from this data, as a robust analysis is complicated by several factors:

- 485 1. Methodological heterogeneity: The significant variation in algorithms and performance among teams makes aggregating their results into a single composite curve inappropriate.
2. Insufficient sample size: The number of data points for any individual team is generally too small to support robust statistical curve fitting.
3. Release rates tested: Testing for Group C instruments was concentrated in their marginal detection range below 1500 kg/h, so we cannot characterize performance at higher rates.

490 However, this dataset is a valuable input for external studies focused on performance modeling. For instance, our data contributed to a recent preprint by Ramier et al. (2026) that generates a probabilistic, conditions-dependent detection model for GHGSat, demonstrating the utility of our data for advancing community-wide performance assessments³³.

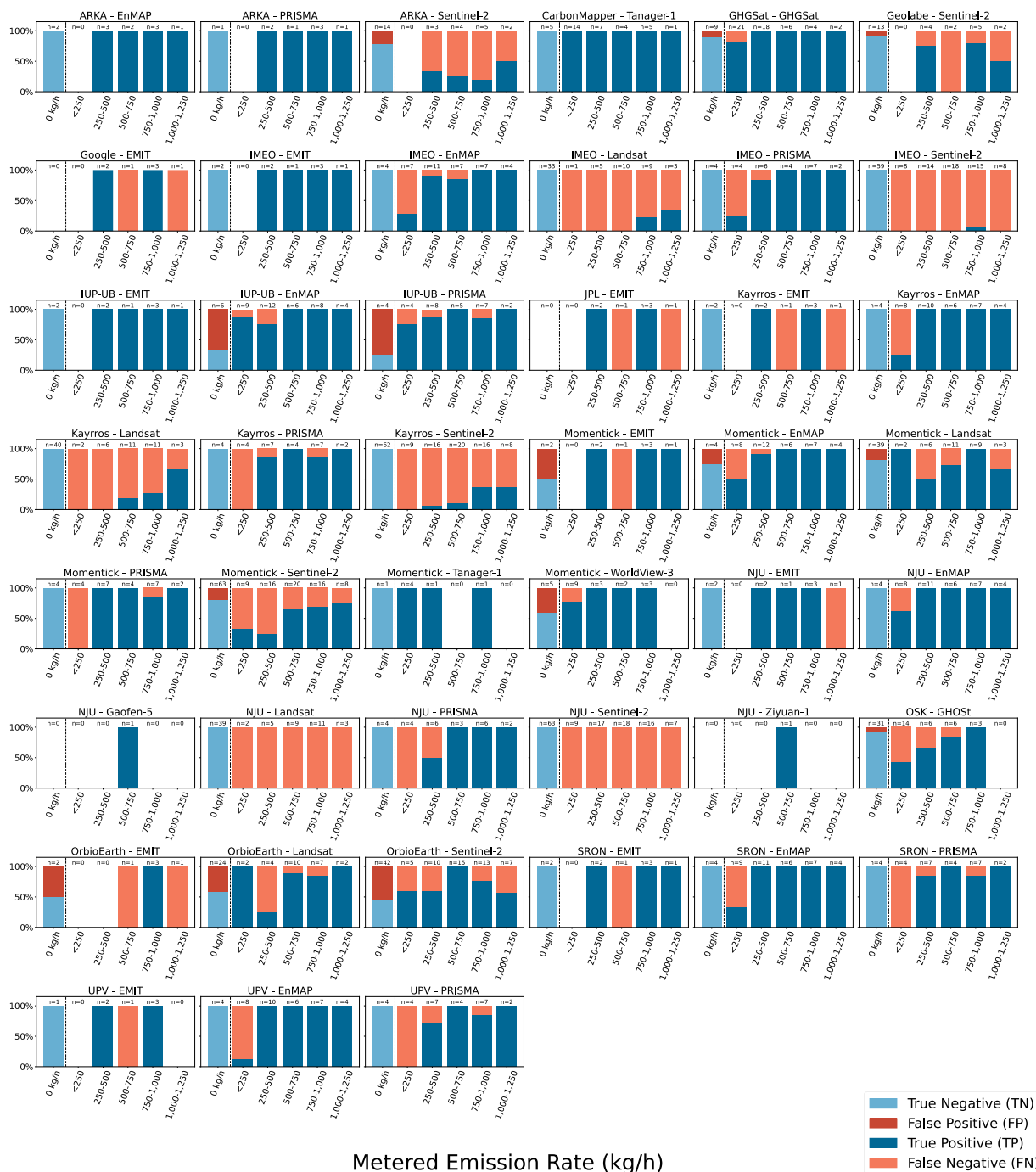
495 Figure S4 provides detail on results for blank control events by team and satellite, binned by team-estimated release rate. Zero estimates correspond to True Negatives, while any non-zero estimates are False Positives (FPs). Figure S5 aggregates these results by satellite. Most FP estimates with Landsat were less than 500 kg/h, while FPs estimates with Sentinel-2 were slightly larger.

500 In a handful of events, a team submitted two detection outcomes corresponding to a single release or non-release event. This occurred when multiple images were captured in rapid succession during a single satellite pass. For the detection analysis (e.g., calculating True Positive Rates), these submissions were consolidated to ensure a one-to-one correspondence between a unique observing event and its outcome. In all such cases, the two submitted outcomes were identical (e.g., both were True Positives, or both were filtered for clouds), making the consolidation straightforward. Each satellite-team-event combination is therefore represented by only a single data point in the detection statistics. For quantification analysis, if a team submitted multiple quantified estimates, each estimate was treated as an individual data point. These are represented as separate points on the parity charts and no averaging was performed.



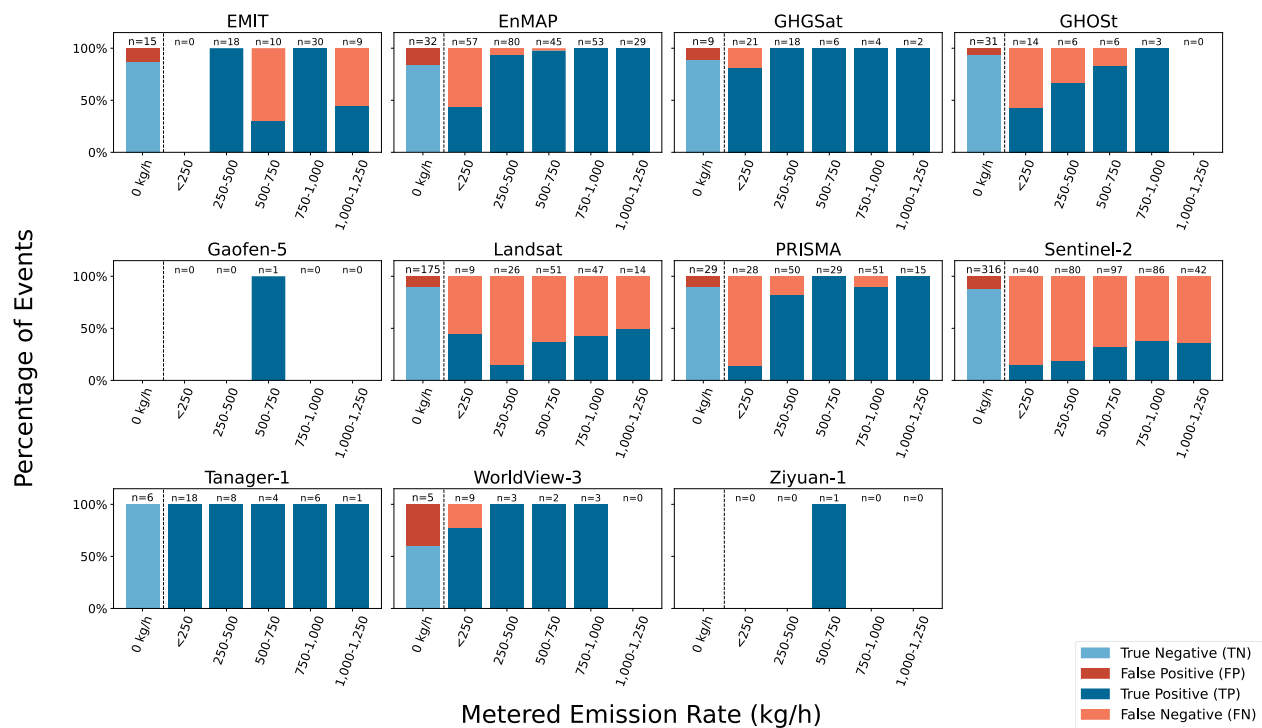
510

Figure S1. Detection outcomes over the distribution of metered release rates, by satellite and instrument grouping, including releases that were undetected for any reason. This chart differs from Figure 1 in the main text because it includes releases wherein the team did not detect for any reason (e.g., clouds).



Metered Emission Rate (kg/h)

Figure S2. Detection performance binned by release rate. Bars left of the dashed line correspond to zero release events (blanks), while bars to the right correspond to non-zero releases. Some teams participated in different study phases and filtered different releases, leading to different counts for the same satellite.



525 **Figure S3.** Detection performance aggregated by satellite, binned by release rate. Bars left of the dashed line correspond to zero release events (blanks), while bars to the right correspond to non-zero releases. Counts represent number of releases and analyzing teams (e.g. 2 releases x 9 teams analyzing = 18).



Figure S4 . Detection outcomes for blank control events by satellite and team, binned by team-estimated rate (zero estimates are True Negatives, non-zero estimates are False Positives).

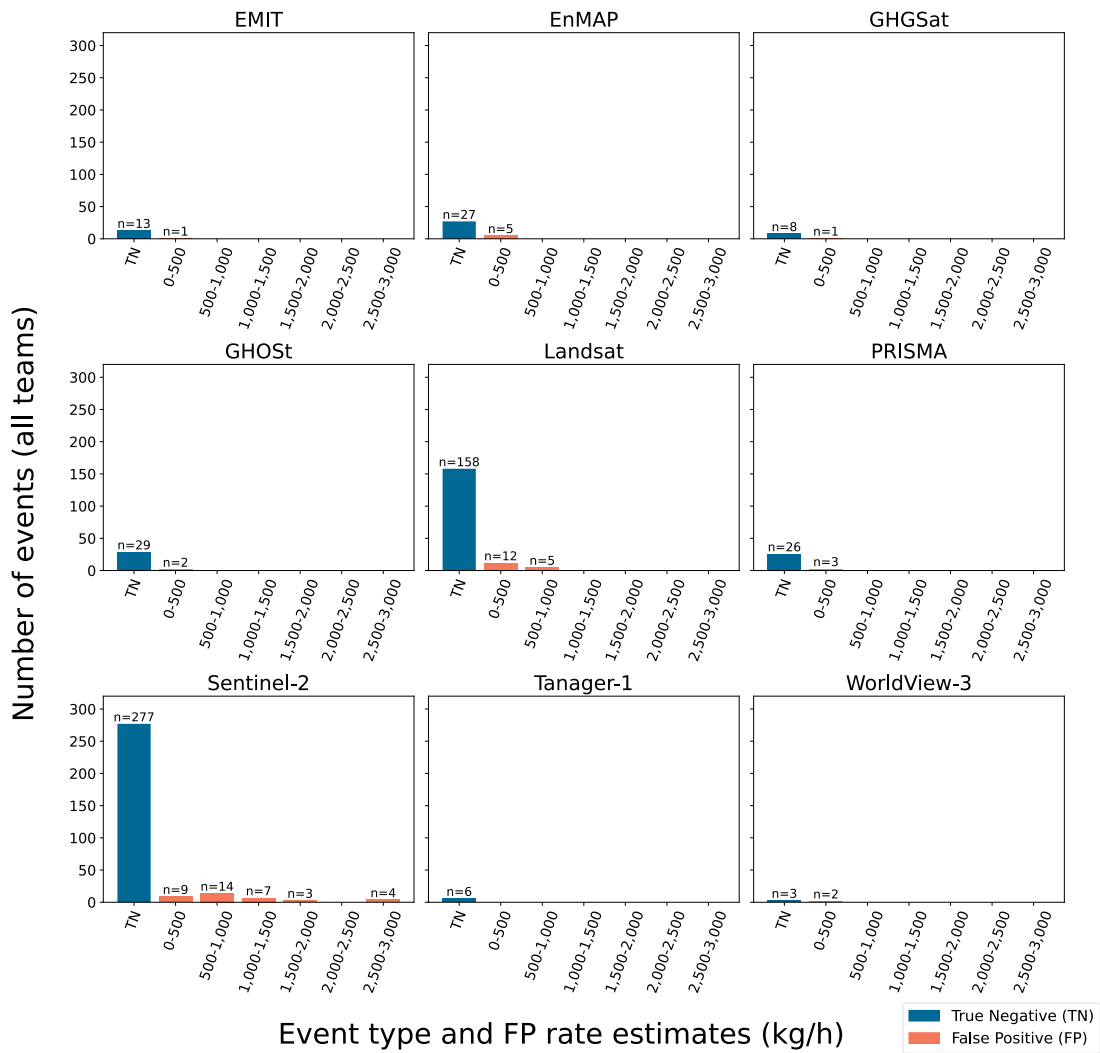


Figure S5. Detection outcomes for blank control events by satellite, binned by team-estimated rate (zero estimates are True Negatives, non-zero estimates are False Positives).

540 **S3.4 Classification performance by team and satellite**

The following metrics were calculated for each team-satellite pairing to evaluate their classification performance (Table S6). The calculations are based on the four possible outcomes for an unfiltered release or non-release event: TN, TP, FN, FP.

- 545
- Observed accuracy: the overall proportion of correct classifications out of all classifications made.
 - $(TP + TN) / (TP + FP + TN + FN)$
 - Baseline accuracy (prevalence): the accuracy a naive model would achieve by always guessing the most prevalent class. In this experiment, the most prevalent class is most often “True” since more events were releases than non-releases. It represents the benchmark to beat, and methods that improve upon baseline accuracy may be evidencing detection skill.

550

 - p_value: the probability of a team achieving their Observed Accuracy or better, under the null hypothesis where no detection skill exists. That null world is created through permutation analysis: their labels were assigned randomly with a 10,000-fold permutation test. By permuting their reported results, this accounts for the team's specific guess rate (i.e., how often they guessed "True" vs. "False"). A small P-value suggests that the team's performance is statistically significant and unlikely to be the result of random chance. The lowest possible P-value is 0.00001 due to the use of 10,000 simulations.

555

 - Sensitivity (recall): measures the ability of the team to find all the actual "True" cases.
 - $TP / (TP + FN)$
 - Specificity: measures the ability of the team to find all the actual “False” cases.
 - $TN / (TN + FP)$

560

 - Precision: measures the accuracy of the positive predictions
 - $TP / (TP + FP)$
 - F1 score: the harmonic mean of Precision and Sensitivity
 - Balanced accuracy: the average of Sensitivity and Specificity. This metric provides a more representative measure of overall performance than Observed Accuracy when the actual number of "True" and "False" cases is unequal (as it is in our case, where we performed about 2/3 “True” releases and 1/3 “False” non-release controls).

565

We present significance P-value thresholds of 0.05 and 0.01. Some teams with high accuracy are non-significant statistically. This tends to occur where baseline accuracy is high and overall sample sizes are small. No correction for multiple comparison is performed, so significant results near thresholds could represent random chance. For cases with very low P-values, any correction for multiple comparisons would have trivial impacts.

570

575 **Table S6. Metrics of classification performance by satellite and team.** p_value * <0.05, ** <0.01

Satellite	Team	Observed accuracy	Baseline accuracy	p_value	Sensitivity (recall)	Specificity	Precision	F1	Balanced accuracy
Tanager	CM	1	0.861	9.999E-05**	1	1	1	1	1
	MO	1	0.857	0.143	1	1	1	1	1
GHGSat	GHG	0.916	0.85	9.999E-05**	0.923	0.889	0.97917	0.945	0.905
EMIT	GOO	0.714	1	1	0.714	n/a	1	0.833	n/a

	IMEO	1	0.778	0.029 *	1	1	1	1	1
	IUP	1	0.778	0.029 *	1	1	1	1	1
	JPL	0.714	1	1	0.714	n/a	1	0.833	n/a
	KAY	0.778	0.778	0.167	0.714	1	1	0.833	0.857
	MO	0.778	0.778	0.413	0.857	0.5	0.85714	0.857	0.679
	NJU	0.889	0.778	0.084	0.857	1	1	0.923	0.929
	OE	0.571	0.714	0.711	0.6	0.5	0.75	0.667	0.55
	SRON	0.889	0.778	0.084	0.857	1	1	0.923	0.929
	UPV	0.889	0.778	0.086	0.857	1	1	0.923	0.929
EnMAP	ARKA	1	0.818	0.016 *	1	1	1	1	1
	IMEO	0.825	0.9	0.004 **	0.806	1	1	0.892	0.903
	IUP	0.822	0.867	0.174	0.897	0.333	0.897	0.897	0.615
	KAY	0.846	0.897	0.003 **	0.829	1	1	0.906	0.914
	MO	0.854	0.902	0.018 *	0.865	0.75	0.970	0.914	0.807
	NJU	0.925	0.9	0.0004 **	0.917	1	1	0.957	0.958
	SRON	0.854	0.902	0.002 **	0.838	1	1	0.912	0.919
	UPV	0.821	0.897	0.005 **	0.8	1	1	0.889	0.9
GHOSSt	OSK			9.999E-05 **					
		0.783	0.517		0.621	0.935	0.9	0.735	0.778
Gaofen5	NJU	1	1	1	1	n/a	1	1	n/a
PRISMA	ARKA	1	0.875	0.125	1	1	1	1	1
	IMEO	0.852	0.852	0.005 **	0.826	1	1	0.905	0.913
	IUP	0.8	0.867	0.453	0.885	0.25	0.885	0.885	0.567
	KAY	0.786	0.857	0.010 *	0.75	1	1	0.857	0.875
	MO	0.821	0.857	0.005 **	0.792	1	1	0.884	0.896
	NJU	0.72	0.84	0.025 *	0.667	1	1	0.8	0.833
	SRON	0.786	0.857	0.010 *	0.75	1	1	0.857	0.875
	UPV	0.75	0.857	0.016 *	0.708	1	1	0.829	0.854
WV3	MO	0.818	0.773	0.055	0.882	0.6	0.882	0.882	0.741
ZY1	NJU	1	1	1	1	n/a	1	1	n/a
Landsat	IMEO	0.590	0.541	0.090	0.107	1	1	0.194	0.554
	KAY	0.644	0.548	0.002 **	0.212	1	1	0.35	0.606
	MO	0.8	0.557	9.999E-05 **	0.774	0.821	0.774	0.774	0.797
	NJU	0.565	0.565	1	0	1			0.5
	OE	0.694	0.510	0.005 **	0.8	0.583	0.667	0.727	0.692
Sentinel-2	ARKA	0.536	0.5	0.498	0.286	0.786	0.571	0.381	0.536
	GEO	0.769	0.5	0.005 **	0.615	0.923	0.889	0.727	0.769
	IMEO	0.492	0.516	0.516	0.016	1	1	0.031	0.508
	KAY	0.565	0.527	0.001 **	0.174	1	1	0.296	0.587
	MO	0.667	0.523	9.999E-05 **	0.536	0.80952	0.755	0.627	0.673
	NJU	0.485	0.515	1	0	1	n/a	n/a	0.5
	OE	0.620	0.543	0.026 *	0.76	0.452	0.623	0.685	0.606

580 **S3.5 Statistical analysis of false positive rates**

To evaluate if AI-based methods were significantly more likely to produce false positives (FPs), we run a bootstrap analysis. In our dataset, AI-based methods produced 58 FPs across 302 blank scenes (False Positive Rate [FPR] = 19.2%), while non-AI methods produced 13 FPs across 317 blank scenes (FPR = 4.1%). We
585 perform 100,000 bootstrap iterations. In each iteration, the FPRs for the AI and non-AI groups are simulated by resampling their respective true negative and false positive outcomes with replacement. The difference between the simulated AI and non-AI FPRs is recorded.

This procedure yields a 95% confidence interval for the difference in FPRs (AI FPR - Non-AI FPR) of [0.1018,
590 0.2004]. The probability of observing a difference less than or equal to zero under this model was negligible ($p < 0.00001$). Since the confidence interval does not contain zero, this provides strong statistical evidence that AI-based methods have a higher FPR.

We repeat the analysis with subsets 1) Group A and B satellites only and 2) Group C satellites only. The higher
595 FPRs for AI-based methods remains statistically significant for the Group C subset. There was no statistically significant difference in the FPRs between AI and non-AI methods for Group A and B satellites.

Table S7. False positive rate and reported use of AI method, by analyzing team

Analysis team (alphabetical)	Organization type	False Positive Rate (# blanks analyzed)	Reports AI-based detection method
ARKA	Commercial	0.18 (n = 17)	No
Carbon Mapper	Non-Profit	0.0 (n = 5)	No
GHGSat	Commercial	0.11 (n = 9)	No
Geolabe	Commercial	0.08 (n = 13)	Yes
Google	Commercial	0.0 (n = 0)	Yes
IMEO	Intergovernmental/NGO	0.0 (n = 102)	Yes
IUP-UB	Academic/Research	0.58* (n = 12)	No
JPL	Government/Public	0.0 (n = 0)	No
Kayros	Commercial	0.0 (n = 112)	No
Momentick	Commercial	0.19 (n = 118)	Yes
NJU	Academic/Research	0.0 (n = 112)	No
Orbio Earth	Commercial	0.50 (n = 68)	Yes
Orbital Sidekick	Commercial	0.06 (n = 31)	No
SRON	Research/Public	0.0 (n = 10)	No
UPV	Academic/Research	0.0 (n = 9)	No

*Note: Interpreting FP detections by IUP-UB with the EnMAP and PRISMA instruments requires consideration
600 of their analytical method. Their cross-sectional flux method does not include a plume detection step and instead directly estimates an emission rate (Q). As agreed upon with the analyst team, any final $Q > 0$ was classified as a

detection for this study even if the 1-sigma uncertainty interval included zero. If a more conservative threshold were applied that considers estimates close to zero as non-detections, these FPs would be eliminated and the number of reported TPs reduced.

605

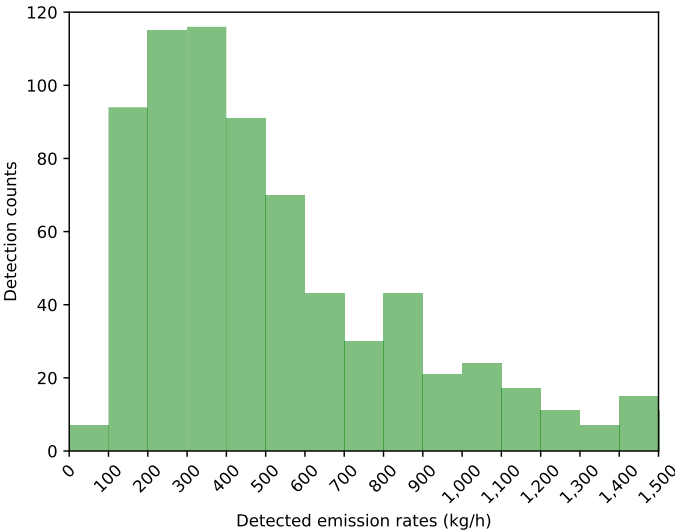
S3.6 Real-world plume observations with Tanager-1

610

To contextualize the performance observed in the experiment, we analyzed a set of real-world Tanager-1 plume detections. We downloaded all publicly available (as of February 18, 2026) Tanager-1 plumes from the Permian Basin, a desert region with numerous emission sources and an environment relatively similar to the test site.

615

Figure S6 shows the distribution of the 762 quantified emission rates (plumes >1500 kg/h not shown). The count of plumes in the 200-300 kg/h bin is comparable to the 300-400 kg/h bin, and both are substantially larger than the count in the 100-20 kg/h bin. Given expected power-law behavior of emission source populations in the Permian down to at least 100 kg/h, the drop-off in detections below 200-400 kg/h suggests that Tanager-1 is operating within its partial detection range.



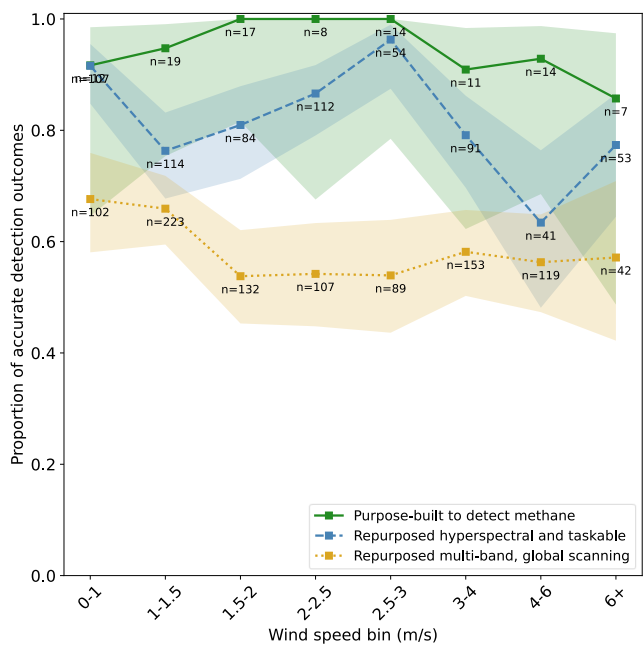
620

Figure S6. Tanager-1 plume detections in the Permian basin

S3.7 Detection and meteorology effects

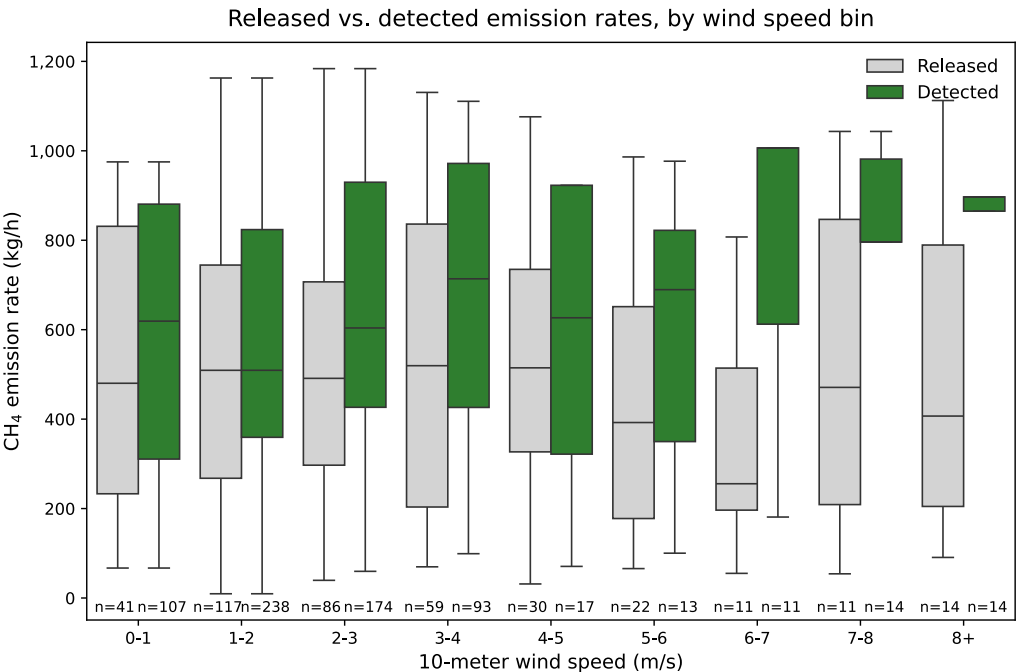
S3.7.1 Wind

- 625 Figure S7 illustrates the proportion of accurate detection outcomes (the sum of True Positives and True Negatives divided by total attempts) for each satellite typology, binned by 10 m wind speed. Uncertainty bands are generated using the Wilson score interval (<https://www.mwsug.org/proceedings/2008/pharma/MWSUG-2008-P08.pdf>).
- 630 The purpose-built Group A satellites demonstrated a high proportion of accurate detections (>0.9) across a broad range of wind speeds (0–6 m/s). A slight reduction in accuracy was observed in the highest wind bin (>6 m/s); however, high wind speed samples were limited.
- Of the three groups, the repurposed hyperspectral instruments (Group B) showed the most sensitivity to wind speed, although the pattern is somewhat erratic. Observed accuracy generally declined for winds exceeding 3 m/s, suggesting that in some cases stronger winds more quickly dilute plumes and push methane enhancements below detectable limits. A deeper examination of wind effects and statistical significance is reserved for future investigation.
- 635
- 640 The multispectral scanners (Group C) exhibited a consistent but more modest proportion of accurate detections (~0.6) that was largely independent of wind speed across the tested range.



645 **Figure S7.** Detection accuracy across satellite typologies, binned by wind speed. Samples per bin listed on chart

Figure S8 presents the distribution of all attempted release rates alongside the distribution of successfully detected rates, both binned by 10 m wind speed. The distribution of release rates attempted during the experiment (grey bars), including the mean and interquartile range, remains relatively consistent across all wind speed bins. In contrast, the distribution of successfully detected release rates (green bars) shows dependence on wind speed. As expected, the mean detected emission rate increases with increasing wind speed, indicating teams were less able to detect smaller releases during windier conditions. At the highest wind speed bin, only the largest of releases are able to be detected. Although not broken out here, all three satellite groupings (A, B, C) are represented in the highest windspeed bin. A deeper examination of how windspeed affects detection capabilities is warranted for future study.



660 **Figure S8.** Distribution of release rates and detected release rates, binned by wind speed. Number of releases and number of analyzed detections are listed.

Figure S9 presents a comparison of the 10 m wind speeds used by analysis against the ground-truth measurements recorded on-site, averaged over a five-minute period prior to each overpass. The scatter plot reveals only weak to moderate agreement.

665 As detailed in the individual team descriptions and in Table S3, teams sourced wind information from several different reanalysis products, including ERA5, HRRR, GEOS-FP, and OpenWeatherMap. Each team may apply has its own spatial and temporal interpolation techniques to estimate the wind conditions at the location and time of the overpass. Also, each team may convert the resulting wind speed to an “effective” wind speed that is used in

the IME or IME-like method. While not explored in depth here, how these meteorological discrepancies affect final quantification results merits future study.

The use of the HRRR model, which is limited to the Continental U.S., also raises an important question about the generalizability of these results to other regions. Comparisons between HRRR and global models like ERA5 suggest that while region-specific biases exist, quantification results can be broadly comparable. For instance, while our specific study location in Arizona was not covered, Millstein et al. found good agreement between the models over arid West Texas during midday (relevant to satellite overpass times)³⁴. Ramier et al. used both HRRR and ERA5 for GHGSat emission estimates in North America also found broadly comparable performance³³.

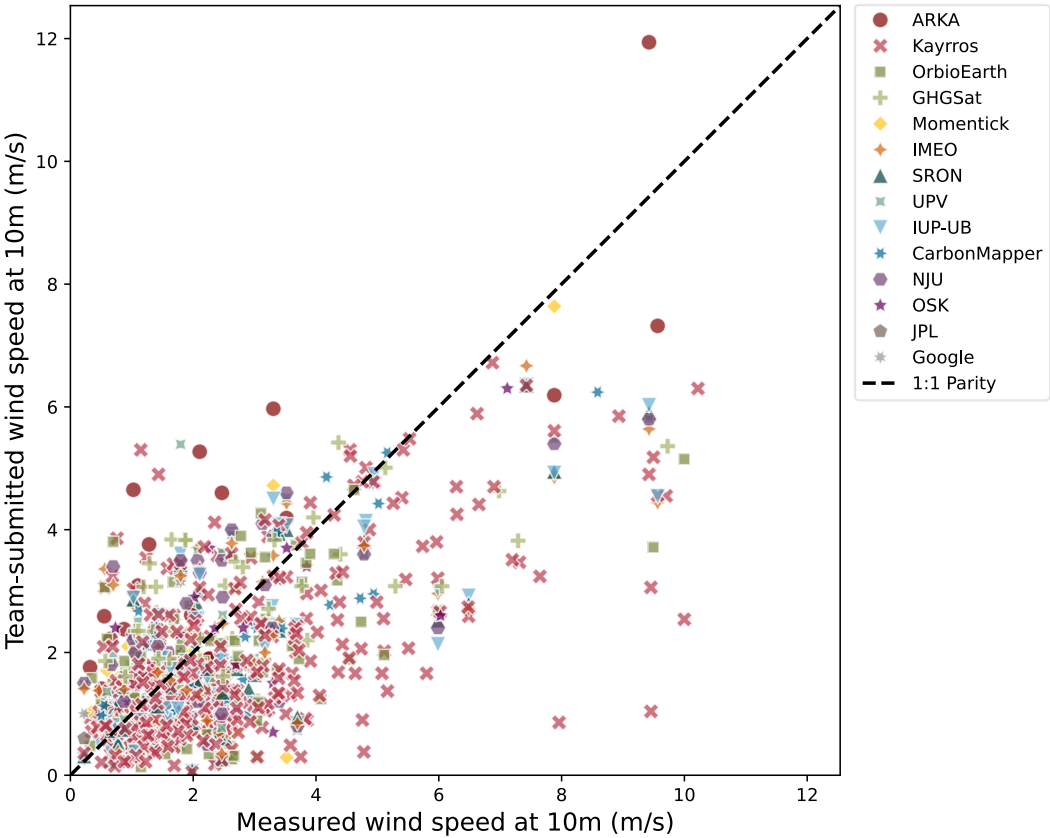


Figure S9. Scatter plot of 5-minute averaged windspeeds at 10 m measured onsite, compared to team-submitted windspeeds at 10 m. Analysing teams source wind information from different reanalysis products. Geolabe did not report 10 m windspeeds.

Note on the blinding of wind data:

A methodological choice in this study was to keep *in-situ* meteorological observations blinded from the analysis teams. While this introduces variability in the wind data used by each team, this decision serves a primary experimental goal: to evaluate the performance of each team's workflow. In real-world monitoring, ground-based anemometers are almost never available at the precise location and time of a satellite overpass. Therefore, a team's choice of a wind reanalysis product and their method for interpolation are integral components of their overall performance and were thus considered part of the system being tested.

This approach is informed by findings from prior two-stage controlled release studies. Sherwin et al. (2024) conducted an experiment where teams first submitted estimates using their own wind data and then re-submitted after being provided with *in-situ* measurements. The result was only a modest improvement in mean aggregate error (from 29% to 27%), indicating that wind data selection was not the sole dominant driver of error¹⁸. Given this evidence, and the substantial logistical burden of a two-stage analysis, we opted for a single-stage blind design.

S3.7.2 Sun angle

Figure S10 shows proportion of accurate detection outcomes (TP fraction and TN fraction showed separately) for each satellite typology, binned by 10-degree solar zenith angle. In this convention, an angle of 0-degrees is sun directly overhead.

All groups showed modest decline in TP detections as sun angle increased, although error bars are fairly wide. This could suggest that lower photon availability associated with the sun being closer to the horizon can reduce signal, making methane detection more challenging. Sun angle showed no effect on TN fraction.

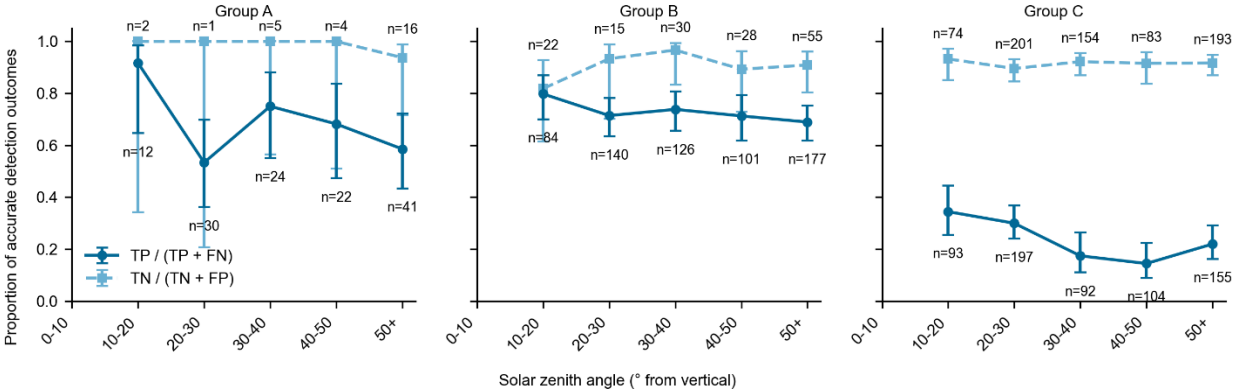


Figure S10. Proportion of accurate detection outcomes by solar zenith angle bin and instrument grouping.

S3.7.3 Clouds

Figure S11 breaks out the results from Figure 4 in the main text, further broken down by cloud cover and emission rate size. This provides further context for a few datapoints that deviate from the main pattern observed, where Group A and Group B instruments maintained high detection accuracy in clear or low-cloud (<20% cloud cover) conditions. The new figure reveals that the False Negatives for Group A under clear skies were all low-magnitude releases (e.g., <250 kg/h). The "perfect" record seen in the 40-60% cloud bin, while encouraging, is based on a smaller sample size (n=9). Similarly, the False Negatives observed for Group B under low cloud skies were also predominantly associated with low-rate releases.

Our study provides empirical data on methane detection under cloudy conditions, a topic with limited published literature. The findings should be interpreted within the context of the experimental design, in which participants had approximate knowledge of the source location. While other work has demonstrated detections through cloud gaps, the influence of prior location knowledge is often unspecified³⁵. Disentangling the interdependent effects of cloud fraction, cloud geometry, and prior knowledge remains an area for future investigation.

The cloud challenge is particularly acute in major oil and gas producing nations in the tropics, such as Nigeria and Venezuela. These countries were selected as representative examples in the main text because they are among the highest methane emitters from oil and gas in their respective regions and present a significant monitoring challenge due to persistent cloud cover, which has been shown to cause data sparseness for satellite sensors^{36,37}. For context, Nigeria is the largest oil producer in Africa, while Venezuela is one of the largest producers in South America.

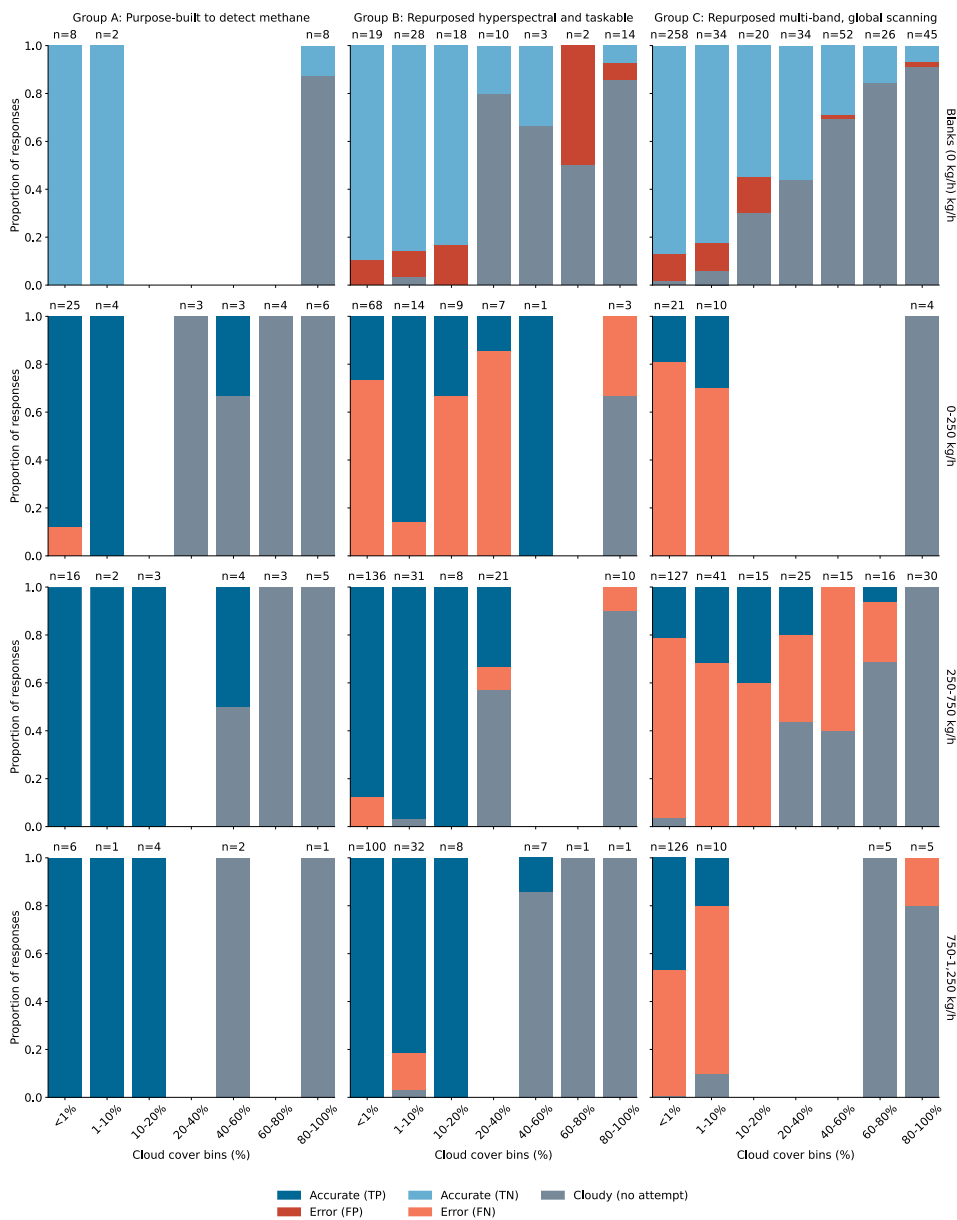


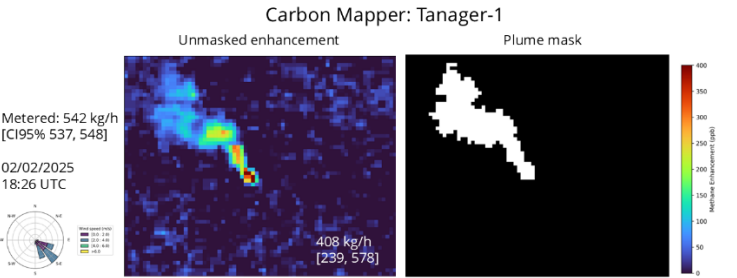
Figure S11. Detection accuracy across satellite typologies, binned by cloud cover percentage and emission rate. Sample per bin listed on chart.

S3.8 Enhancements and plume masks

Figures S12 through S56 show an enhancement and plume mask (if applicable) for each satellite-team pairing. Each figure shows a wind rose of the 10-min average of 10 m wind speed prior to the satellite overpass. Image selection was random, with priority given to choosing events from days that featured multiple satellite overpasses.

750

Figure S12.



755 Figure S13.

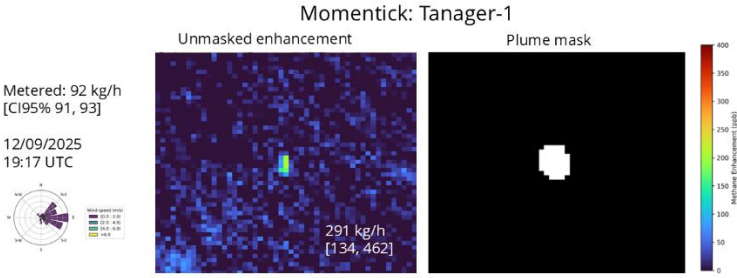
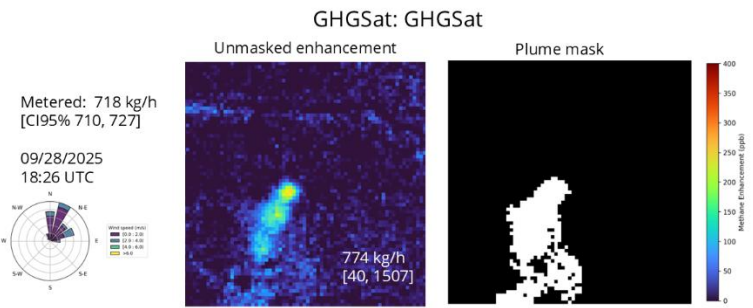


Figure S14.



760

Figure S15.

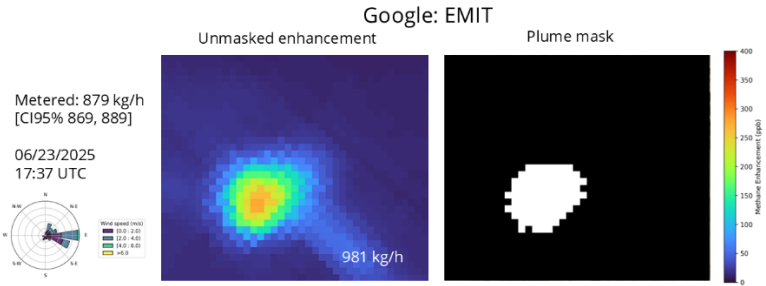


Figure S16.

765

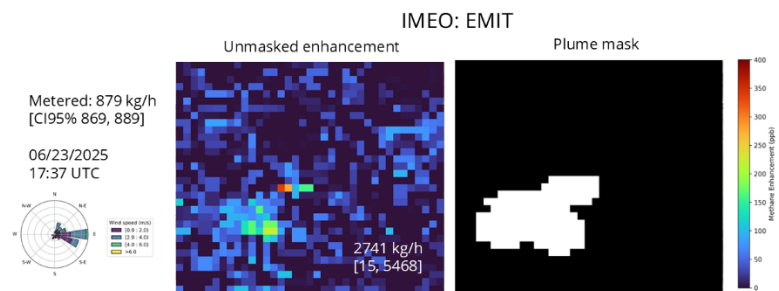


Figure S17.

770

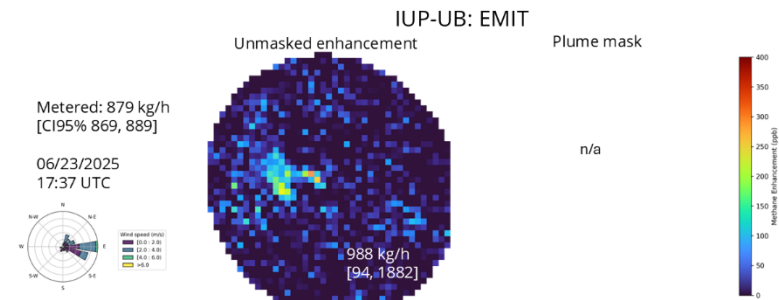


Figure S18.

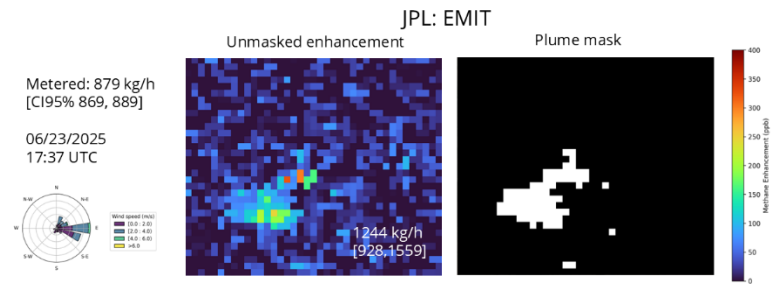
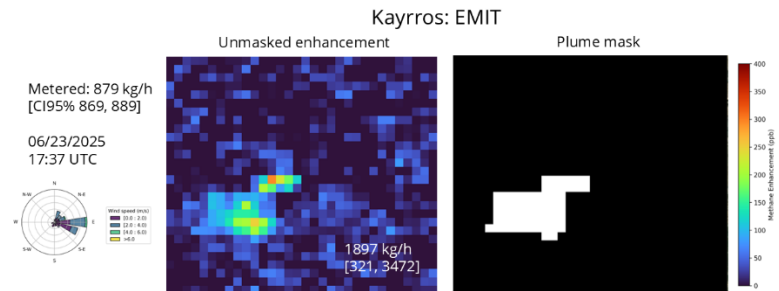
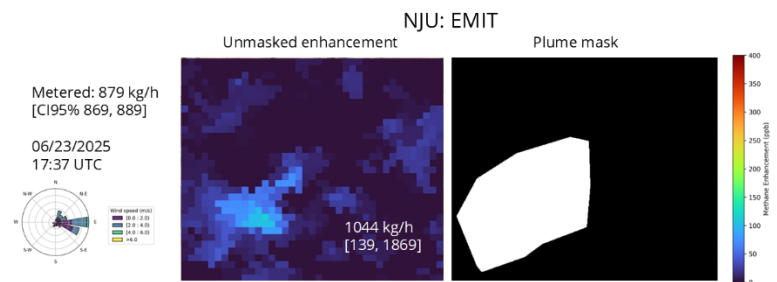


Figure S19.

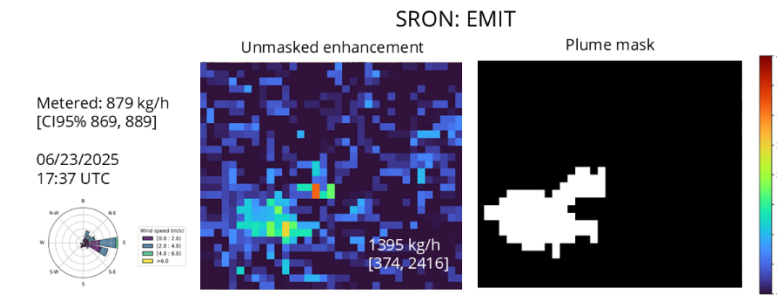
775



780 **Figure S21.**



785



Note: Change in color scale on this image to show very small enhancement.

Figure S23.

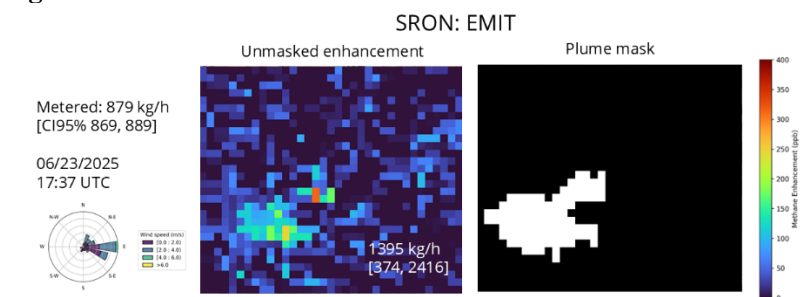
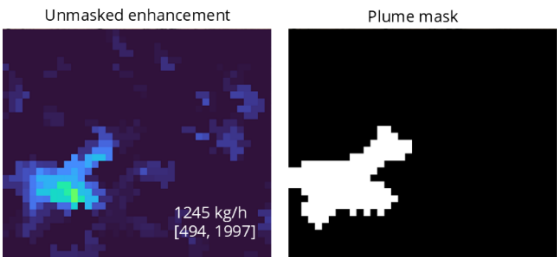
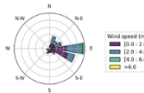


Figure S24.

UPV: EMIT

Metered: 879 kg/h
[CI95% 869, 889]

06/23/2025
17:37 UTC



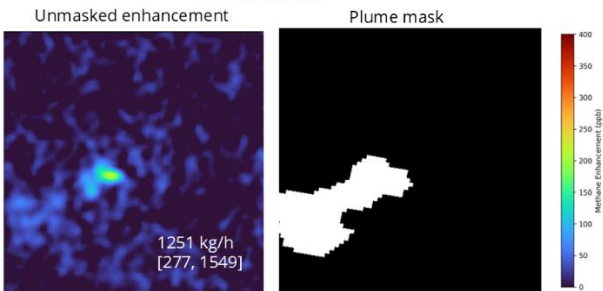
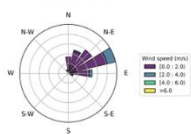
790

Figure S25.

Arka: EnMAP

Metered: 499 kg/h
[CI95% 488, 511]

03/10/2025
18:48 UTC



795 Figure S26.

IMEO: EnMAP

Metered: 809 kg/h
[CI95% 806, 814]

04/29/2025
18:44 UTC

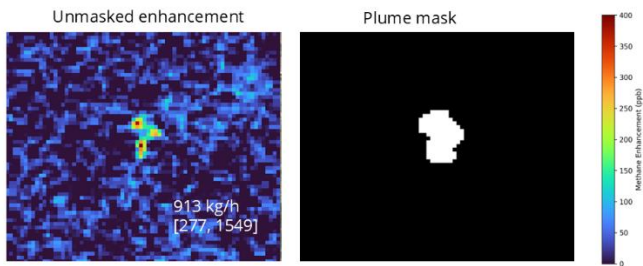
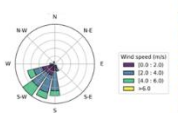
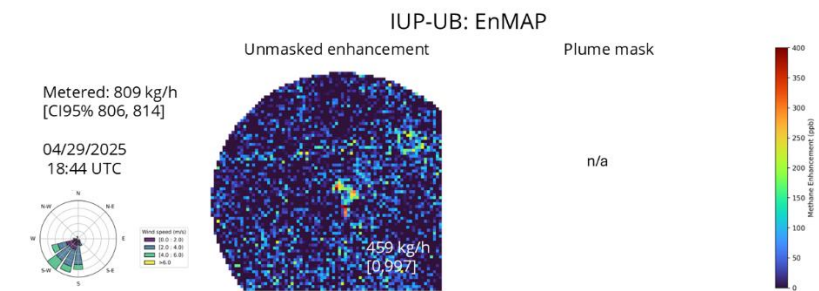


Figure S27.



800

Figure S28.

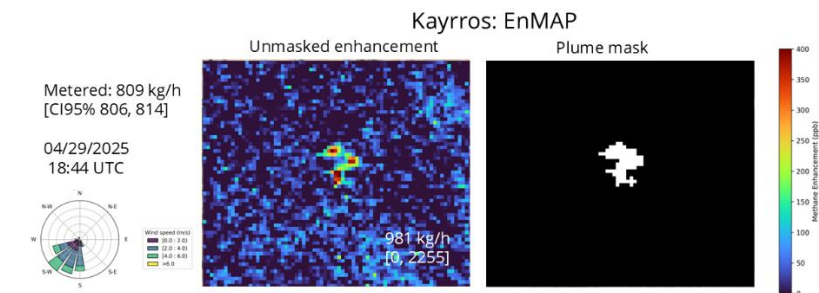
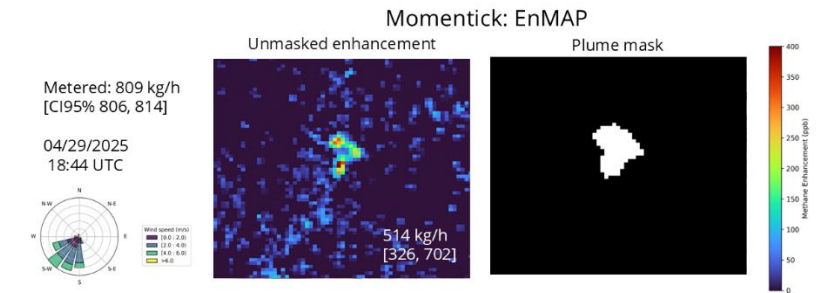
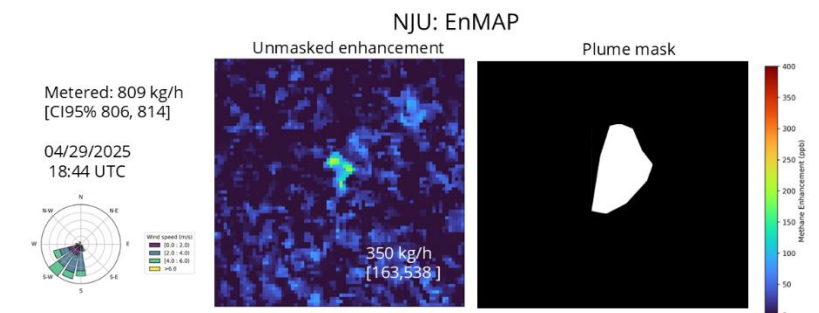


Figure S29.



805

Figure S30.



Note: No CRS was provided for the enhancement; Mask location is approximated to the enhancement.

810 **Figure S31.**

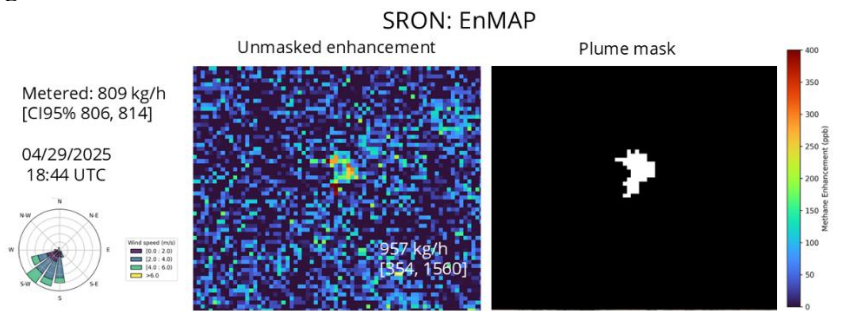


Figure S32.

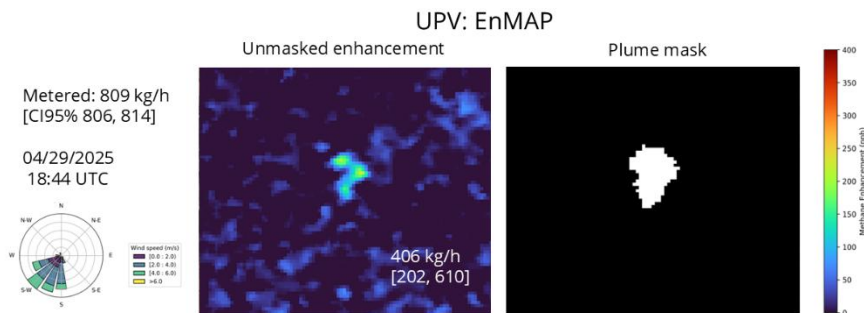
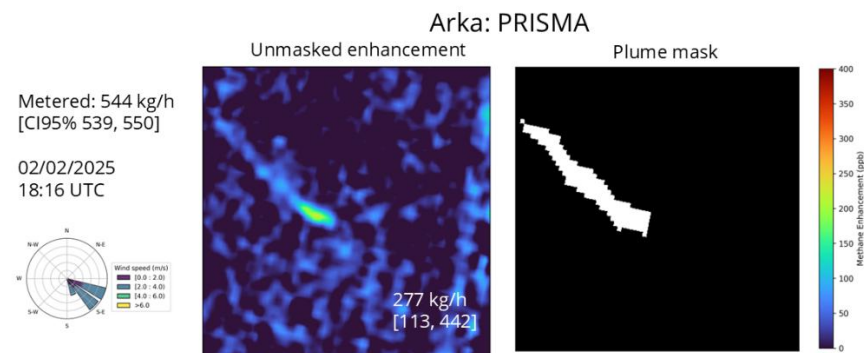
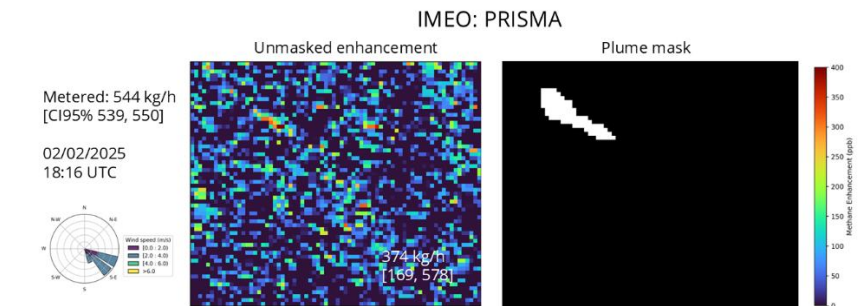


Figure S33.



815

Figure S34.



820



Figure S41.

855

Figure S42.

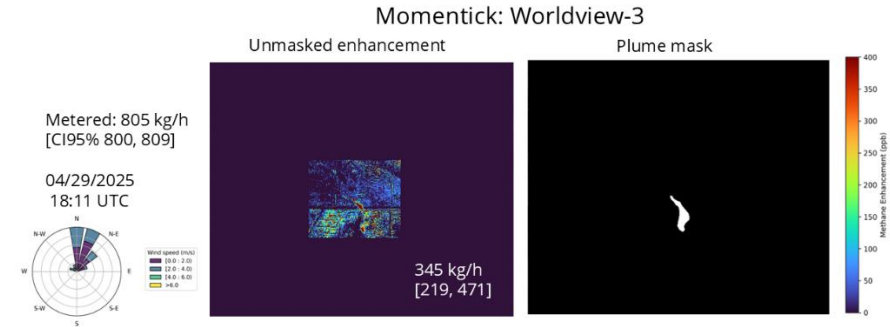
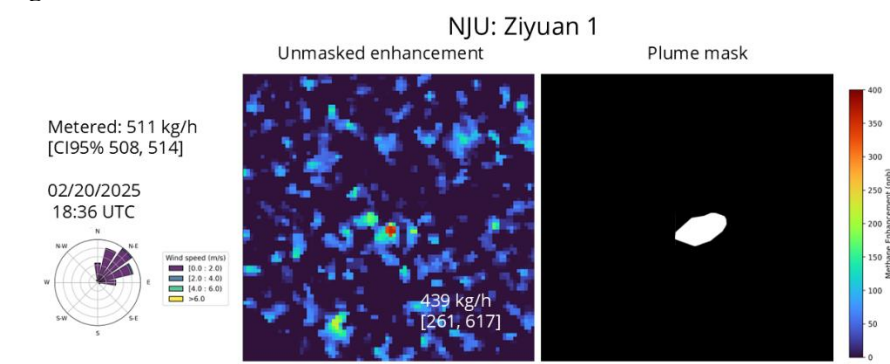


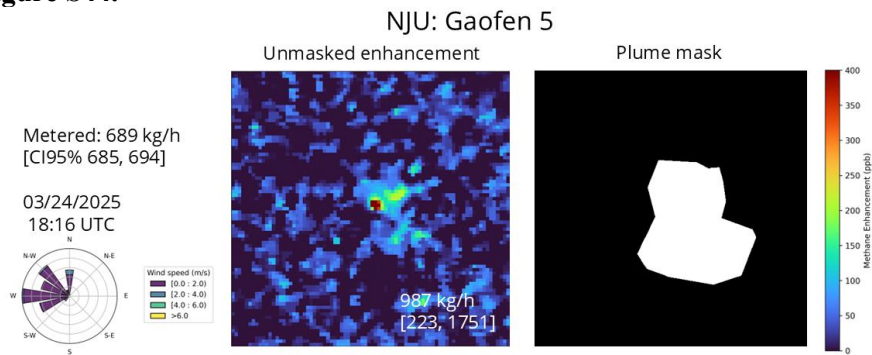
Figure S43.



860 Note: No CRS was provided for the enhancement; Mask location is approximated to the enhancement.

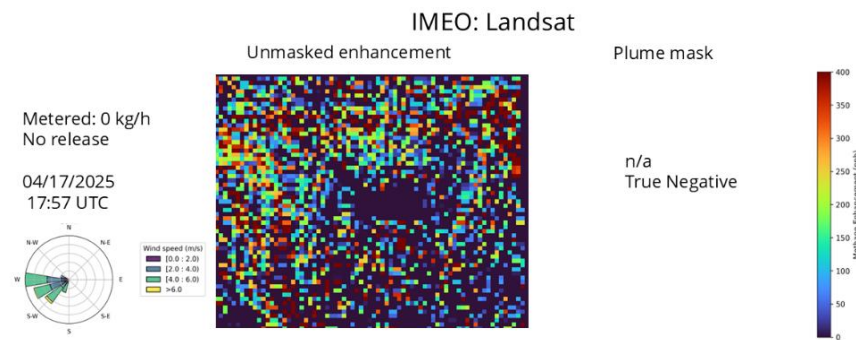
865

870 **Figure S44.**



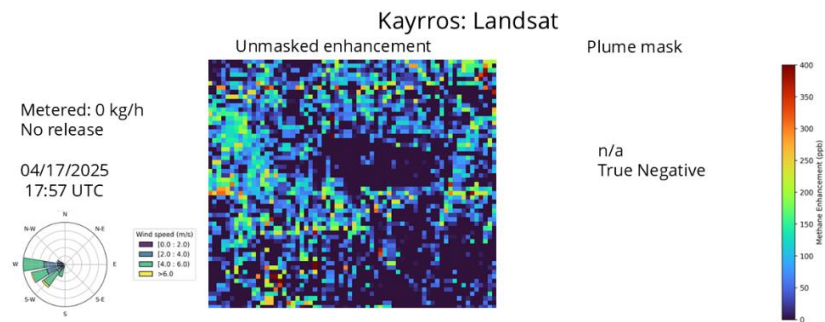
Note: No CRS was provided for the enhancement; mask location is approximated to the enhancement.

Figure S45.



875

Figure S46.



880

885

Figure S47.

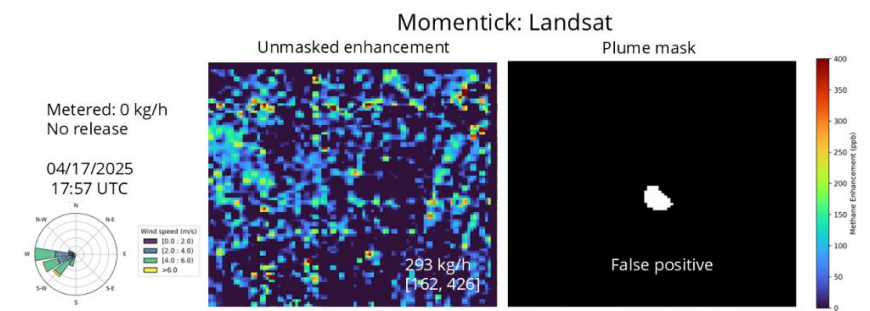
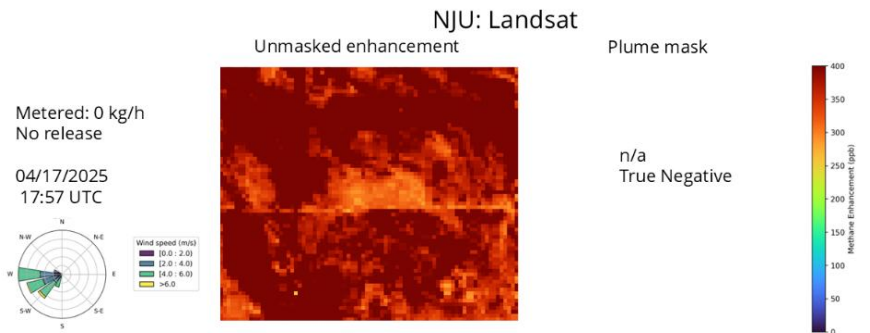


Figure S48.



890

Figure S49.

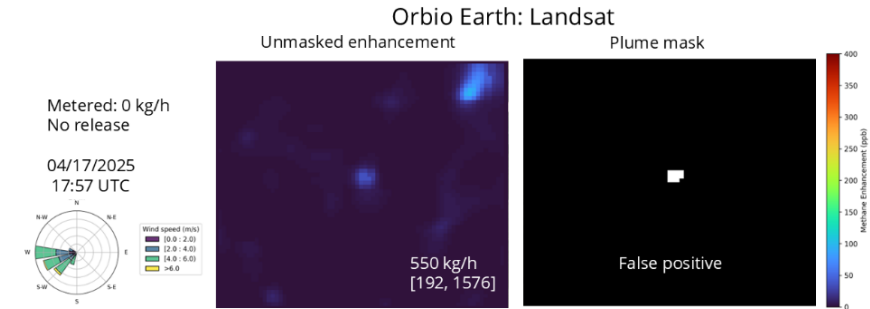
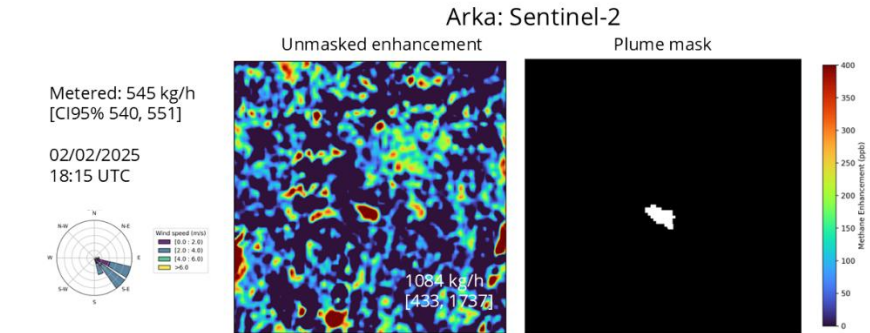


Figure S50.



895

Figure S51.

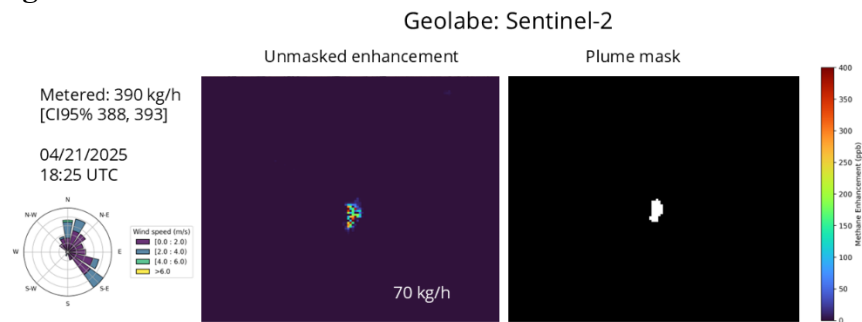
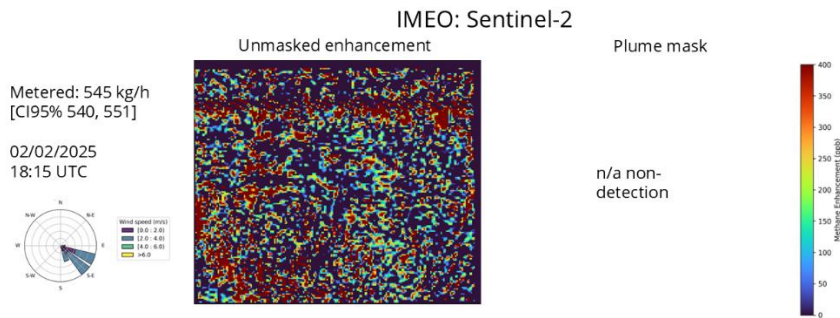
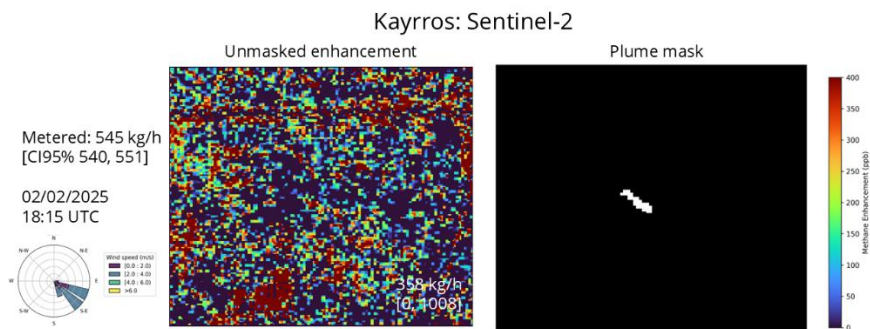


Figure S52.



900

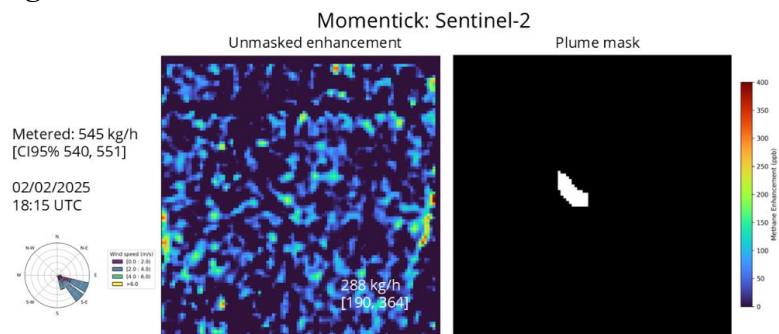
Figure S53.



905

910

Figure S54.



915

Figure S55.

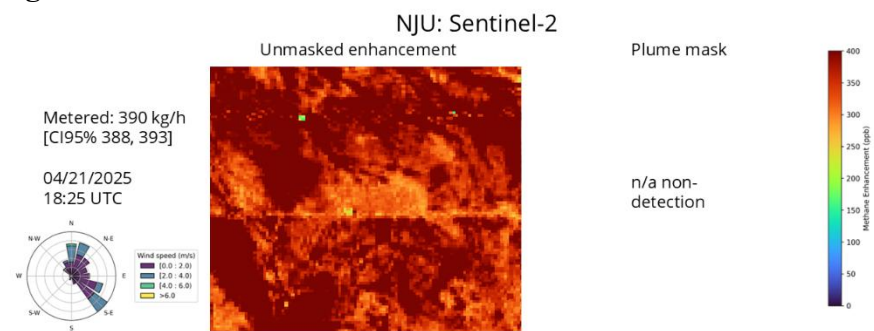
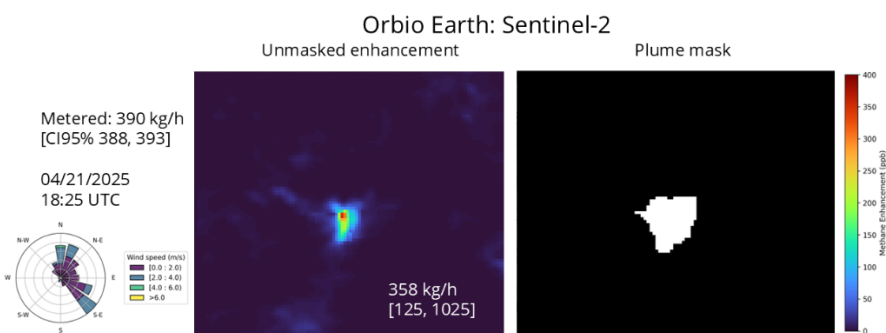


Figure S56.



920

S3.9 Quantification results

S3.9.1 Methodology for quantification fits and simulation study

925

To characterize the quantification performance of each satellite-team pairing, we required a minimum of five true positive (TP) detections to fit an Ordinary Least Squares (OLS) regression. Most pairings met this threshold since, in most cases, dozens of release opportunities occurred over the year-long study. Exceptions included the NJU team analyzing GF5 and ZY1, which had fewer than five overpasses each. Some teams had fewer than five
930 detections with Sentinel-2 and Landsat.

For each satellite-team pairing we applied two regression models:

935

1. 2-Parameter ("2P") Model: A regression with a free slope and intercept, fit only to TP estimates.
2. 2-Parameter ("2P") Model: A regression with a free slope and intercept, fit to both TP and false positive (FP) estimates.

940

This approach differs from prior work that exclusively used a forced-zero-intercept model^{18,38}. The decision to use both models was informed by a simulation study, detailed below. Alternative regressions to OLS, including York and standardized major axis, are available and useful for accounting for error in x values, when relevant. However, in this case, x-direction errors with a mean 95% confidence interval of +/- 1.1% (max +/- 8.0%) are small compared to y-direction errors. Sherwin et al. (2023) also shows these alternatives can introduce undesirable upward bias in the slopes.

945

Simulation Design for Quantification Fit

The primary goal of the simulation was to answer the question: Which regression model, when fit to controlled release data, provides the most useful correction factor for measurements of real-world sources of unknown rates? That is, if we use our regression results to later correct real-world measurements for instrument bias and noise, which fit is likely to produce corrected results that are most aligned with the actual volumes?

950

To determine the optimal fitting strategy, we designed a simulation to mimic the process of calibration and application. Simulation steps were:

955

1. Generate a synthetic calibration dataset
 - A synthetic emission rates dataset was generated to mimic 100 test releases, with rates drawn from a uniform random distribution between 0 and 1250 kg/h, in line with the operational range of our release equipment

960

2. Simulate sensor performance for hypothetical sensors

- A hypothetical sensor's performance was simulated by systematically varying key parameters; parameter values were altered by sweeping over pre-defined value ranges. This artificial performance simulation allowed us to determine which fitting method would give the best representation of the real-world observations across a range of sensor performance settings.
- Sensor performance parameters were:
 - i. Probability of detection (POD) curves characterized by variable sigmoid width scales (sweep range: 10 to 150 kg/h) and midpoint (50%) parameters (sweep range 0 to 500 kg/h) were applied to the synthetic emission rates
 - ii. Quantification Bias: Both additive bias (sweep range: -100 to 100 kg/h) and multiplicative scaling bias (multiplier range: 0.6 to 1.4) were applied to the measured emission rates of detected plumes
 - iii. Noise: Random noise was added to bias multiplier (sweep range: 0.15 to 0.3) and the quantified estimates (gaussian noise sigma sweep range: 10 to 300 kg/h)
 - iv. False Positives: FPs were introduced at varying rates (0% to 15%), with estimated emission rates either simulated randomly or drawn from a half-normal distribution based on the quantification noise level

3. Calibrate models

- For each simulated dataset generated, five regression models were fit to each sensor's observed emission rates:
 - A. 1-Parameter fit (forced zero intercept) using all points (TPs and false negatives)
 - B. 1-Parameter fit using only TPs
 - C. 2-Parameter fit (free intercept) using all points
 - D. 2-Parameter fit using only TPs
 - E. 2-Parameter fit using TPs and FPs
- Note: in scenarios with no FPs, Fit D and E are the same.
- Figure S57 shows an example of the Fits C, D and E for one simulation. Note that the inclusion of FPs in the fit results in larger y-intercepts and flatter slopes

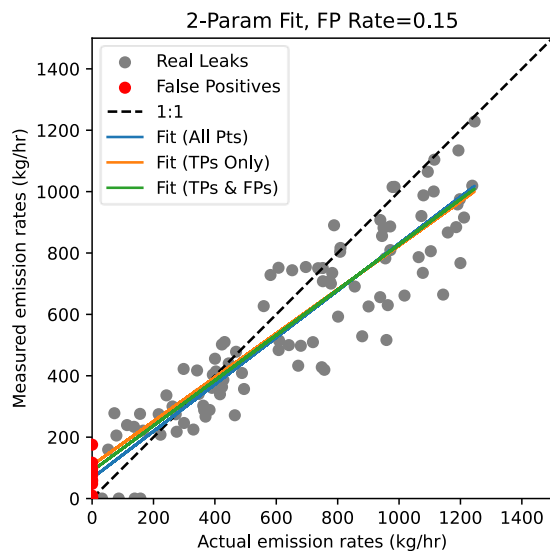


Figure S57. Example 2-parameter simulated fits

4. Generate a synthetic validation dataset

- A separate "real-world" validation dataset of 100 emission rates was generated, with rates drawn from a log-normal distribution to represent typical “in-the-wild” emission profiles
- The same sensor performance characteristics (e.g. POD, bias, noise) from Step 2 were applied in each case.

5. Evaluate model performance

- Each of the five calibration models A-E derived in Step 3 was used to "correct" the quantified estimates in the synthetic validation dataset. For example, if a 1P fit was derived to have a slope of 1.5, measurements would be corrected by dividing by 1.5 to estimate true emission rate.
- The performance of each model was then evaluated by calculating the ratio of the total *corrected* mass of detected emissions to the total *true* mass of those same detected emissions. Because we know the true mass of the synthetic validation dataset, we can exactly determine the impact of each correction. A ratio of 1.0 indicates a perfect correction.

Figures S58-S62 chart the distribution of the ratio of corrected emissions to true emissions for each calibration method when FPs were present. A mean ratio of 1.0 indicates a perfect correction. In scenarios with non-negligible FPs, the **2P model fit to both TPs and FPs (Fit E)** consistently provided the most accurate correction, with a mean ratio closest to 1.0. Calibration models that did not incorporate the FPs in their fit were shown to overestimate total emissions when applied to the validation set. Meanwhile, in scenarios with no FPs, the 1P model fit only to TPs (Fit B) provides the best correction, while Fit E also performs well. Because we cannot determine the real-world FP rate, and because the 2-parameter fit to all data also works well in cases without false positives, we use **the 2P fit to both TPs and FPs** as the default result.

While it might be useful in some cases, we do not recommend using the 2P fit to only the TP data (Fit D). In some sense, this could be considered the “best” correction to the true results, in that it only includes TP data in making the correction. However, in a real-world case where corrections are to be applied to observations in real images, one does not know if any given measurement is a TP or FP measurement. Therefore, a fit only to TP data might underestimate the upward bias caused by the existence of FP measurements. The recommended fit for each satellite-team pairing in this study is denoted in Table S8.

Note on the exclusion of True Negatives

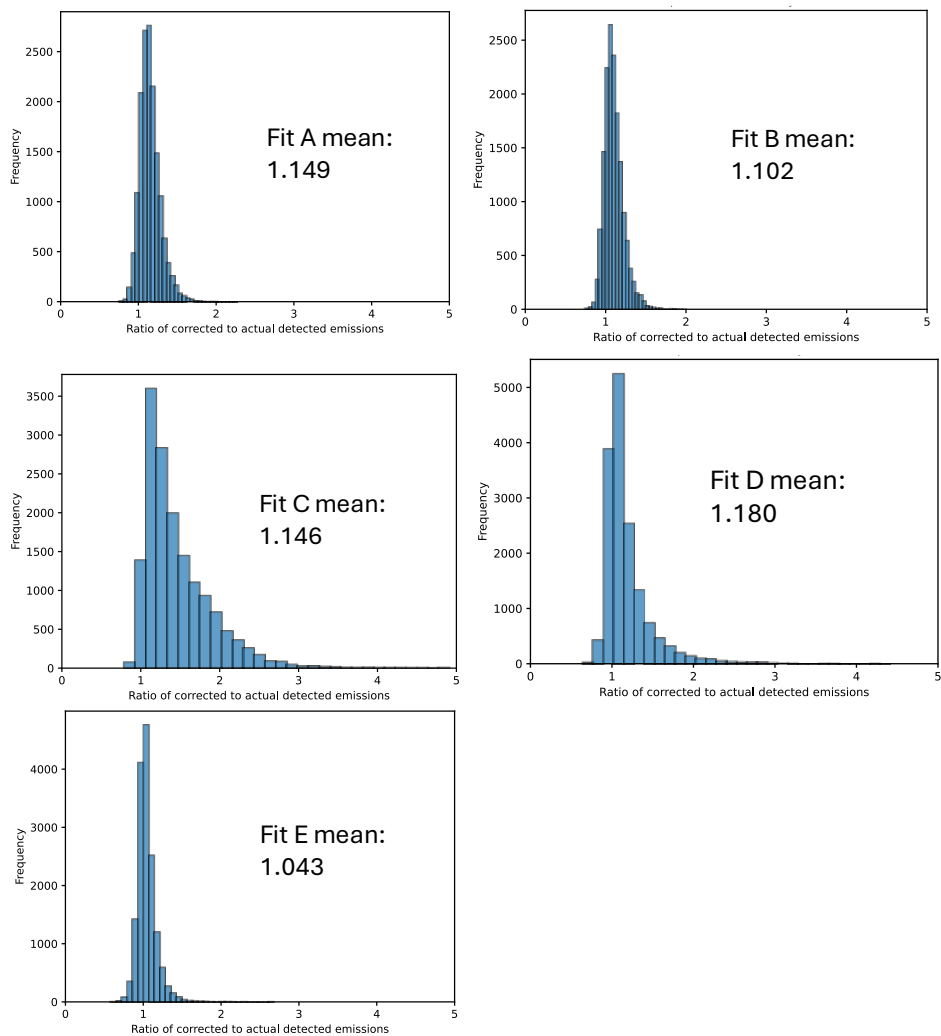
True Negatives (TNs) were intentionally excluded from the regression models based on the practicalities of operational monitoring. In any real-world monitoring scenario, the number of TNs is effectively unbounded; it depends on the total area and time surveyed and will always vastly outnumber the quantity of true detections. Including a potentially immense number of (0,0) data points would overwhelmingly anchor any regression to the origin. This could obscure the true performance characteristics of the system, such as a non-zero bias, when a plume is actually present. Furthermore, in a non-experimental scenario, a "non-detection" cannot be definitively confirmed as a TN without ground-truth knowledge, making the inclusion of such points in a generalizable calibration methodology more difficult.

Limitations of this approach

While the derivation of calibration factors is a primary goal, we caution that factors developed under these specific test conditions may not be directly applicable to measurements made under varying global scenarios. The favorable conditions of this experiment (e.g., constrained source location, generally high sunlight, high and spatially uniform albedo) will differ significantly from some of those encountered in operational plumes monitoring in diverse locations (e.g., cloudy regions, complex terrain). Further study is required to determine the global applicability of these findings.

This approach is expected to behave well for instruments tested well beyond the partial detection range. However, there are statistical challenges when testing in an instrument’s partial detection range (i.e., Sentinel-2 and Landsat). In this regime, 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.

Figure S58-S62. Histograms of ratio of corrected emissions to true emissions, using Fits A-E, when false positives are present.



S3.9.2 Quantification results by team

Applying the conclusions of the above simulation, the recommended fit for each satellite-team pairing is Fit E (Table S8); these are the fits plotted in Figure 5 of the main text, alongside results interpretation. Here, we offer additional assessment of Fit D (TP only) results.

For well-behaved systems with negligible FPs, there is little difference between Fit D and Fit E. Group A instruments showed strong overall performance, with average Fit D slopes of 0.84-1.01 and R^2 of 0.69-0.71. Group B instruments also showed strong quantification performance with more variation. Average slopes ranged

from 0.59-1.19 and R^2 from 0.33-0.65. For satellites processed by multiple teams, variable quantification performance reflects diversity in analytical approaches. For example, with PRISMA, eight teams produced slopes ranging from 0.55 to 1.33 ($R^2 = 0.17$ -0.55).

For Group C instruments, performance was mixed and generally below that of Group A and B. As with Fit E, some cases show flattened slopes, low correlation, and large positive intercepts. In some cases, TP-only improved slopes. For example, average S2 slope improved from 0.32 to 0.69. While this suggests weak Group C performance at the tested rates, these outcomes are not entirely unexpected. As discussed in the main text, 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.

Table S8. OLS fits, corresponding to Figure 5 in main text. Fit E is a 2-parameter model with a free slope and intercept fit to both True Positive and False Positive estimates. Fit D is a 2-parameter model fit only to True Positive estimates.

Instrument	Team	OLS Fit E (recommended) (slope [CI], R^2 , intercept [CI])	OLS Fit D (slope, R^2 , intercept)
GHGSat	GHGSat	0.83 [0.66, 1.00], 0.69, 38.5 [-41.1, 118.0]	0.84, 0.69, 35.7
Tanager-1	Carbon Mapper	0.94 [0.65, 1.24], 0.64, 84.9 [-56.0, 225.8]	0.95, 0.64, 84.9
	Momentick	1.05 [0.26, 1.86], 0.77, 125.1 [-243.0, 493.2]	1.06, 0.77, 125.1
EMIT	Google	0.99 [-0.45, 2.42] 0.61, -50.3 [-1077.7, 977.2]	0.99, 0.61, -50.3
	IMEO	2.99 [-0.33, 6.31], 0.52, -268.4 [-2813.6, 2276.7]	2.99, 0.52, -268.4
	IUP-UB	1.57 [0.22, 2.91], 0.64, -687.3[-1718.4, 343.7]	1.57, 0.64, -687.3
	JPL	1.00 [-1.17, 3.17], 0.42, -51.8 [-1600.1, 1496.5]	1.00, 0.42, -51.8
	Kayrros	1.60 [-1.10, 4.30], 0.54, 6.0 [-1920.0, 1931.9]	1.60, 0.54, 6.0
	Momentick	0.26 [-0.43, 0.94], 0.21, 324.4 [-211.3, 860.1]	0.26, 0.21, 324.4

	Orbio Earth	Not enough quantified points	
	NJU	0.57 [-1.25, 2.40], 0.16, 492.7 [-806.7, 1792.2]	0.57, 0.16, 492.7
	SRON	0.53 [-0.96, 2.02], 0.20, 425.4 [-736.3, 1587.2]	0.53, 0.20, 425.4
	UPV	1.22 [-1.24, 3.69], 0.45, -115.9 [-1873.7, 1641.8]	1.22, 0.45, -115.9
EnMAP	ARKA	0.81 [-0.30, 1.91], 0.30, 417.4 [-384.3, 1219.1]	0.81, 0.30, 417.4
	IMEO	0.65 [0.24, 1.06], 0.28, 316.3 [30.4, 602.1]	0.65, 0.28, 316.3
	IUP-UB	0.71 [0.37, 1.04], 0.33, 50.8 [-154.9, 256.6]	0.72, 0.28, 42.3
	Kayrros	0.95 [0.53, 1.36], 0.45, 342.2 [54.2, 630.3]	0.95, 0.45, 342.2
	Momentick	0.76 [0.45, 1.07], 0.44, 94.0 [-110.3, 298.2]	0.77, 0.42, 84.5
	NJU	0.39 [0.19, 0.60], 0.33, 255.7 [121.9, 389.5]	0.39, 0.33, 255.7
	SRON	0.73 [0.30, 1.16], 0.30, 232.2 [-61.0, 525.4]	0.73, 0.30, 232.2
	UPV	0.57 [0.20, 0.95], 0.28, 168.7 [-93.9, 431.3]	0.57, 0.28, 168.7
GHOSt	Orbital Sidekick	0.85 [0.49, 1.20], 0.58, 18.1 [-149.6, 185.8]	0.99, 0.62, -64.2
PRISMA	ARKA	1.19 [-1.84, 4.22], 0.17, 289.6 [-2029.3, 2608.5]	1.19, 0.17, 289.6
	IMEO	0.77 [0.21, 1.33], 0.33, 232.3 [-183.2, 647.7]	0.77, 0.33, 232.3
	IUP-UB		0.71, 0.38, -20.8

	Kayrros	0.65 [0.32, 0.97], 0.42, 30.5 [-174.7, 235.8]	1.07, 0.42, 234.2
	Momentick	1.07 [0.40, 1.74], 0.42, 234.2 [-258.9, 727.4]	0.55, 0.32, 44.4
	NJU	0.55 [0.14, 0.96], 0.32, 44.4 [-250.5, 339.2]	1.28, 0.39, -259.7
	SRON	1.28 [0.28, 2.29], 0.39, -259.7 [-1065.5, 546.1]	0.70, 0.27, 255.7
	UPV	0.70 [0.10, 1.3], 0.27, 255.7 [-189.8, 701.1]	1.33, 0.55, -207.0
		1.33 [0.66, 1.99], 0.55, -207.0 [-707.8, 293.9]	
WorldView-3	Momentick	0.60 [0.40, 0.81] , 0.70, 75.9 [-19.0, 170.8]	0.59, 0.65, 90.4
Landsat	IMEO	Not enough quantified points	
	Kayrros	-0.75 [-3.16, 1.67], 0.11, 1470.7 [-649.7, 3591.1]	-0.75, 0.11, 1470.7
	Momentick	0.45 [0.16, 0.74], 0.27, 327.5 [138.4, 516.6]	0.50, 0.14, 288.6
	NJU	Not enough quantified points	
	Orbio Earth	0.44 [0.09, 0.79], 0.19, 410.6 [193.2, 628.1]	0.57, 0.10, 303.6
Sentinel-2	ARKA	0.12 [-1.93, 2.17], .00, 1313.1 [133.3, 2492.9]	Not enough quantified points
	Geolabe	0.68 [-0.42, 1.79], 0.23, 186.4 [-649.3, 1022.0]	0.99, 0.30, -77.8
	IMEO	Not enough quantified points	
	Kayrros	0.71 [-0.33, 1.74], 0.19, 196.4 [-700.1, 1092.9]	0.71, 0.19, 196.4

	Momentick	0.32 [0.14, 0.50], 0.23, 422.3 [305.9, 538.6]	0.51, 0.27, 266.5
	NJU	Not enough quantified points	
	Orbio Earth	-0.21 [-0.69, 0.28], 0.01, 1151.7 [875.1, 1428.4]	0.55, 0.06, 566.0

1090

1095 Figure S63 presents the quantification results gridded by analysis team, offering a different assessment from Figure 5 in the main text, which is gridded by satellite. For consistency across panels, Fit E is displayed for every satellite-team pairing. This allows for visual assessment of whether performance characteristics are consistent for a given team's methodology when applied to different satellite data streams. For example, SRON's results show consistent performance for EnMAP and PRISMA. IUP-UB's analyses of EnMAP and PRISMA are also consistent, exhibiting a modest low bias. Both ARKA and Kayrros show a tendency toward a high bias.

1100

In main text Figure 5, WorldView-3 shows higher error for larger release rates than smaller ones. This can be contextualized here. This analysis was performed by the Momentick team, and a similar tendency to underestimate at higher rates can be observed in their analysis of PRISMA data, suggesting this behavior may be characteristic of the analytical method. Additionally, for high-resolution sensors like WorldView-3, it is plausible that the large, bright footprint of the ground infrastructure interfered with the retrieval at high concentrations.

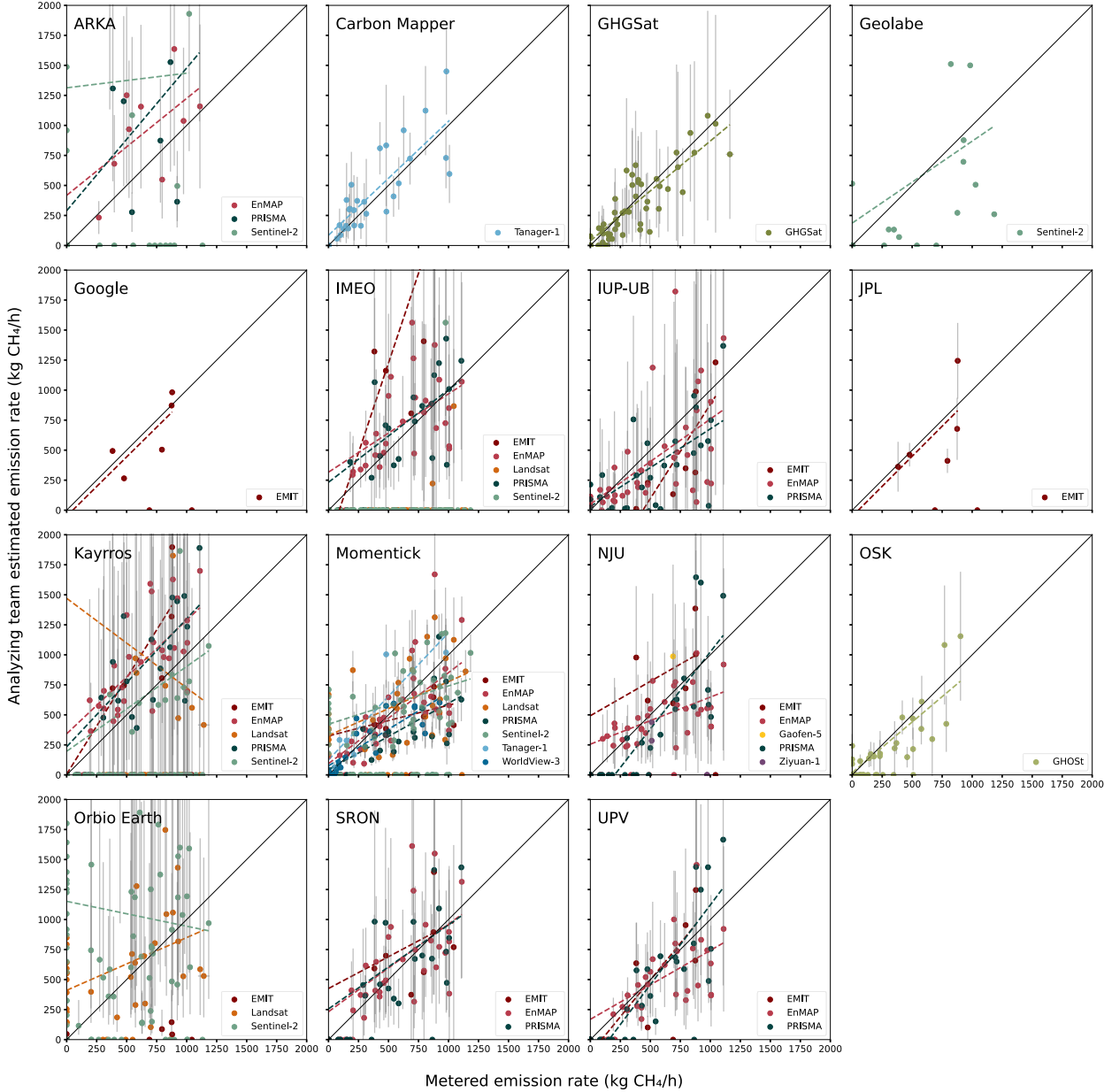
1105

1110

1115

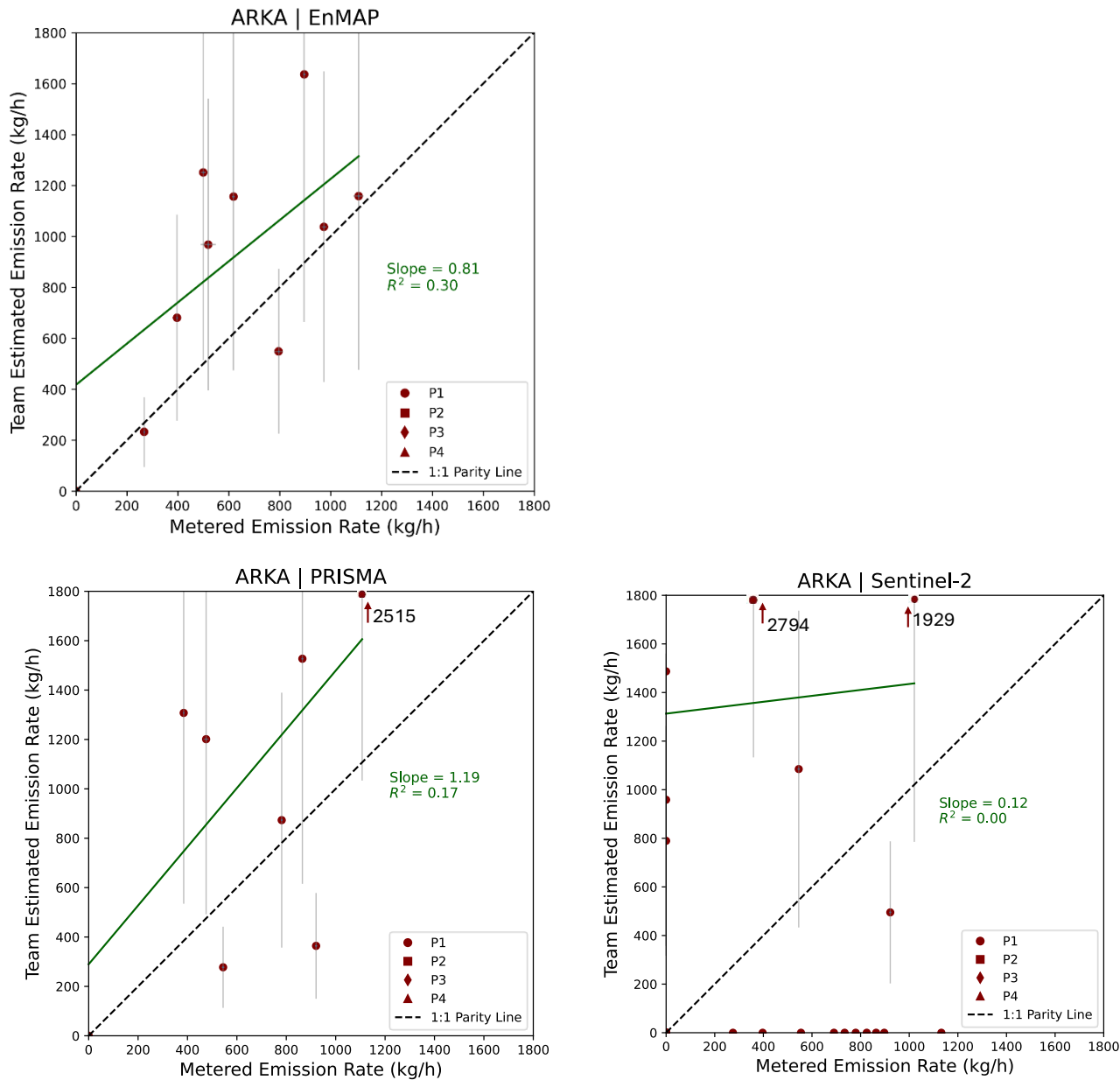
1120

Figure S63. Parity plots by team comparing reported emission rates to metered values. The solid black lines denote exact 1:1 agreement. Each point is color-coded by satellite. All plots display ordinary least squares (OLS) regression fits with a free intercept (2P fit described above). For most teams, 95% CIs were calculated as 1.96 times the submitted 1-sigma uncertainty. Momentick submitted only minimum and maximum estimates, so their error bars reflect this range rather than a 95% confidence interval. Orbio Earth submitted a 2-sigma uncertainty statistic that was not always symmetric about the mean; we used the reported values as provided.



1125

S3.9.3 Satellite- and team-specific parity charts



1130 **Figure S64.** Quantification performance for ARKA for each analyzed satellite, with 95% X and Y confidence intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R² shown for ordinary least squares fit to quantified points (green line).

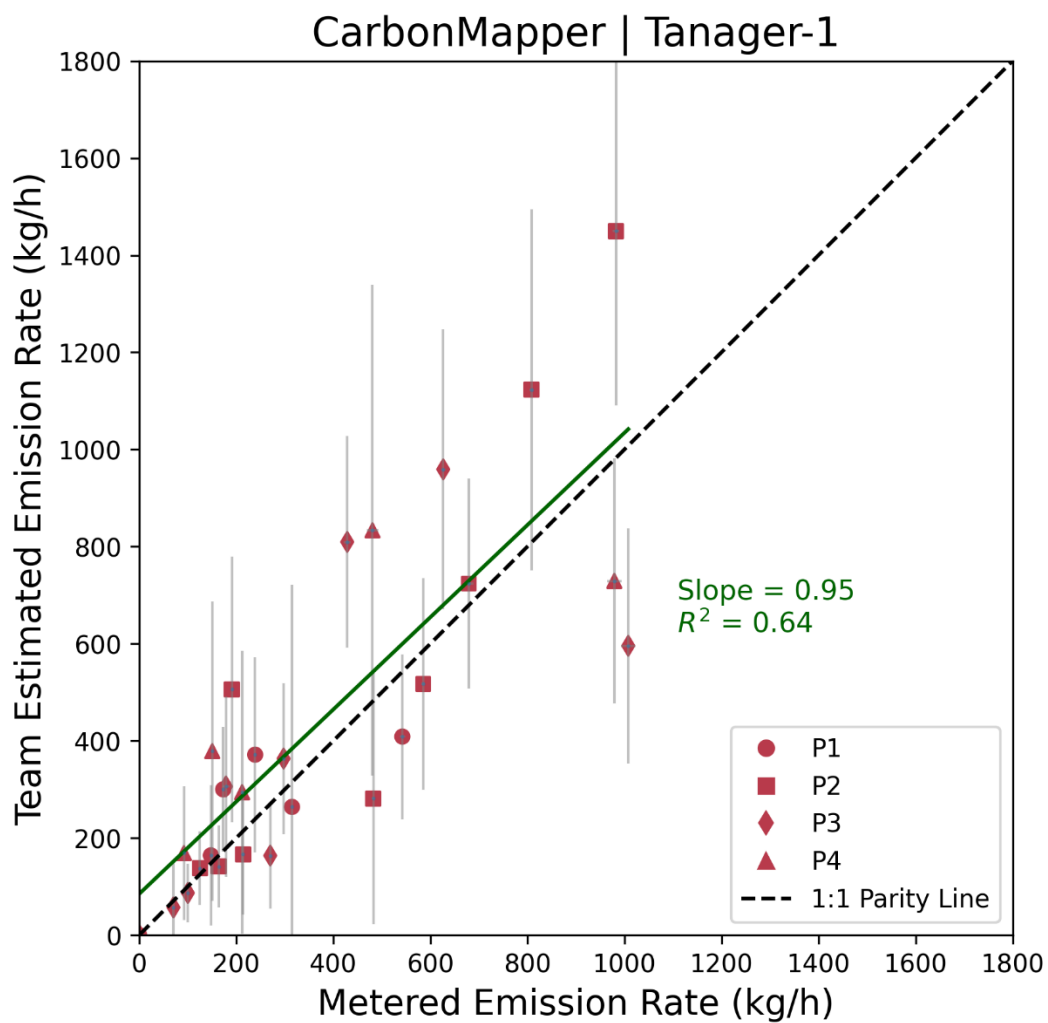


Figure S65. Quantification performance for Carbon Mapper, with 95% X and Y confidence intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

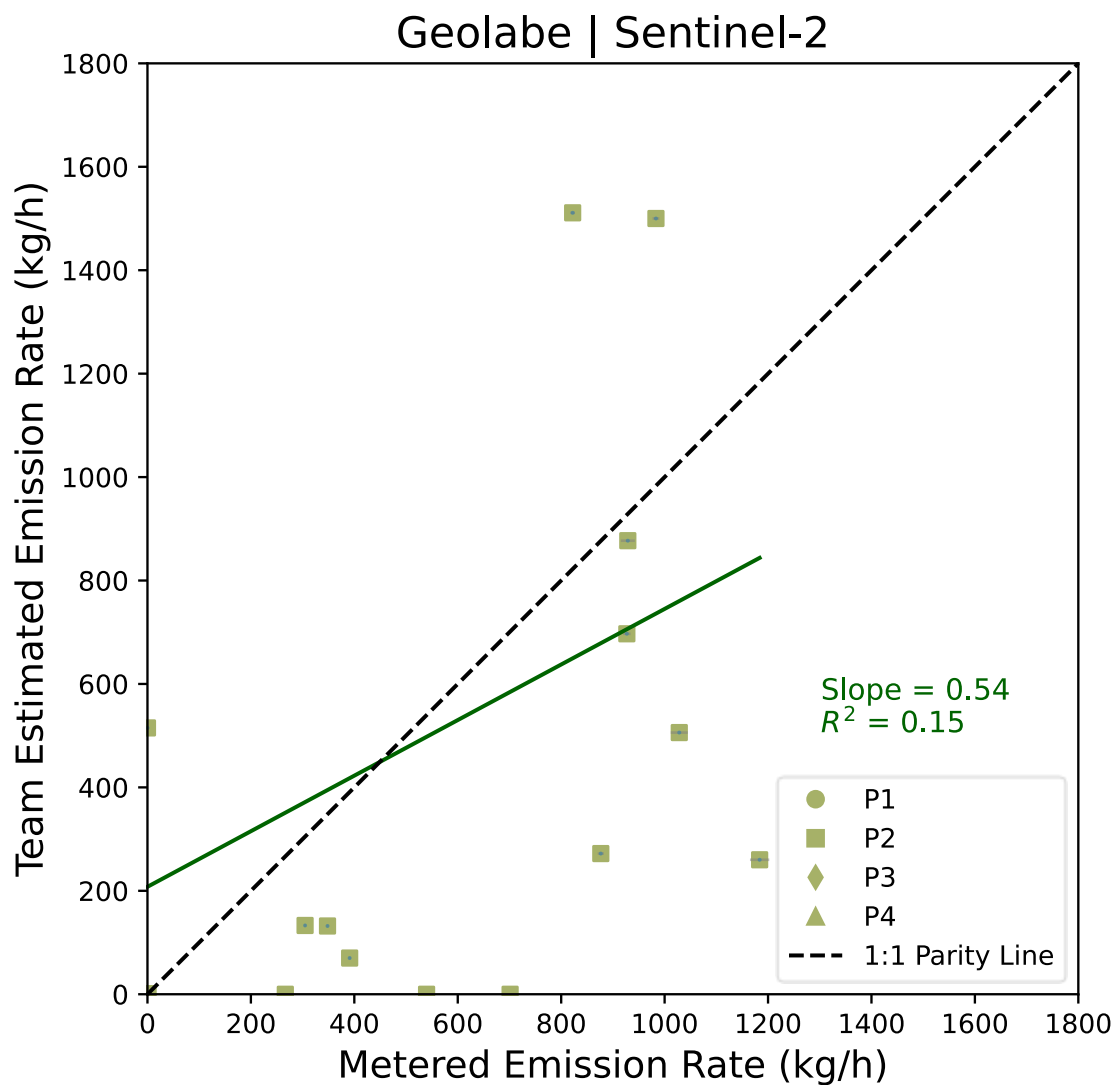


Figure S66. Quantification performance for Geolabe, with 95% X confidence intervals. Geolabe did not submit emission rate uncertainty intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

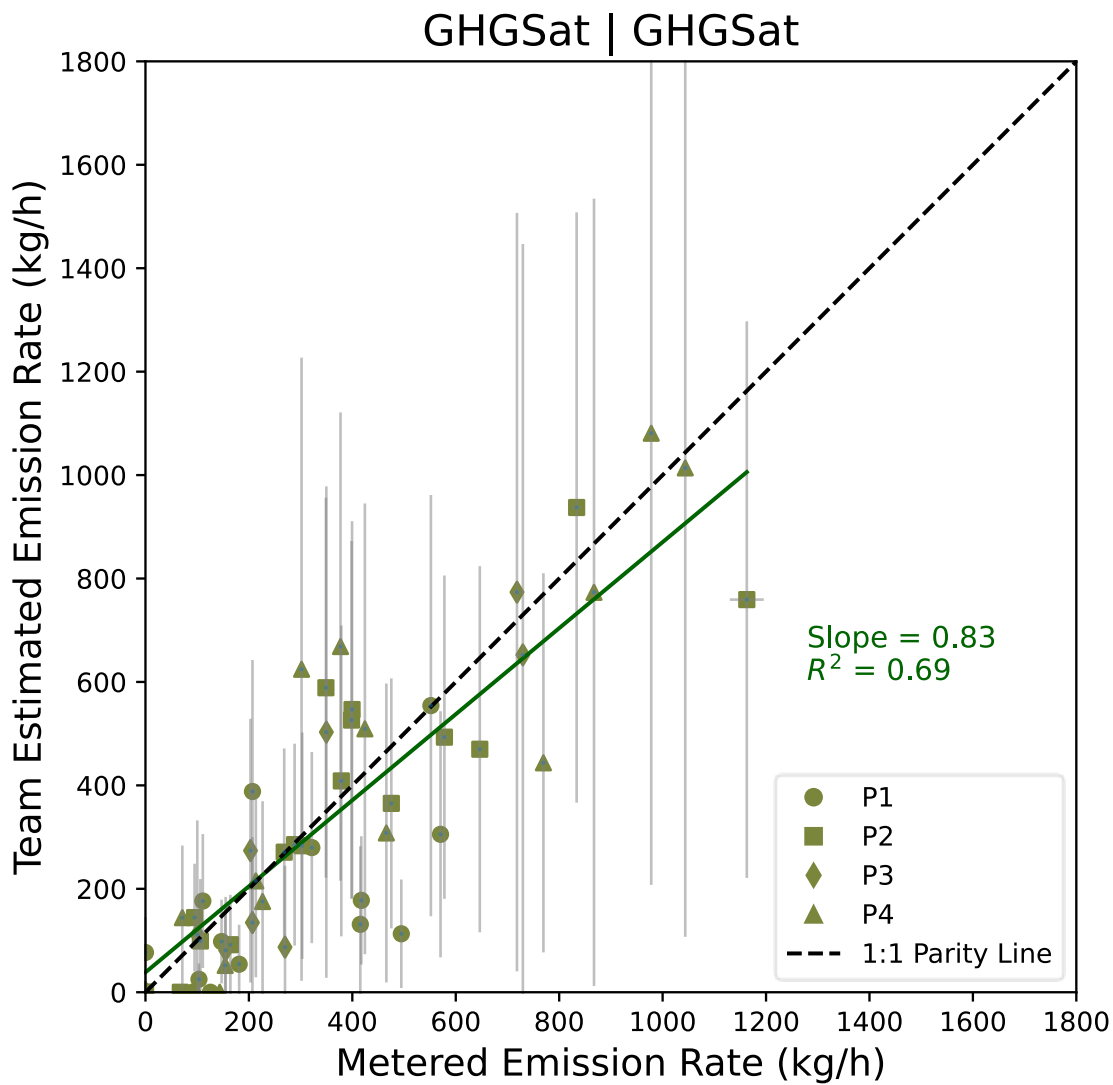


Figure S67. Quantification performance for GHGSat, with 95% X and Y confidence intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

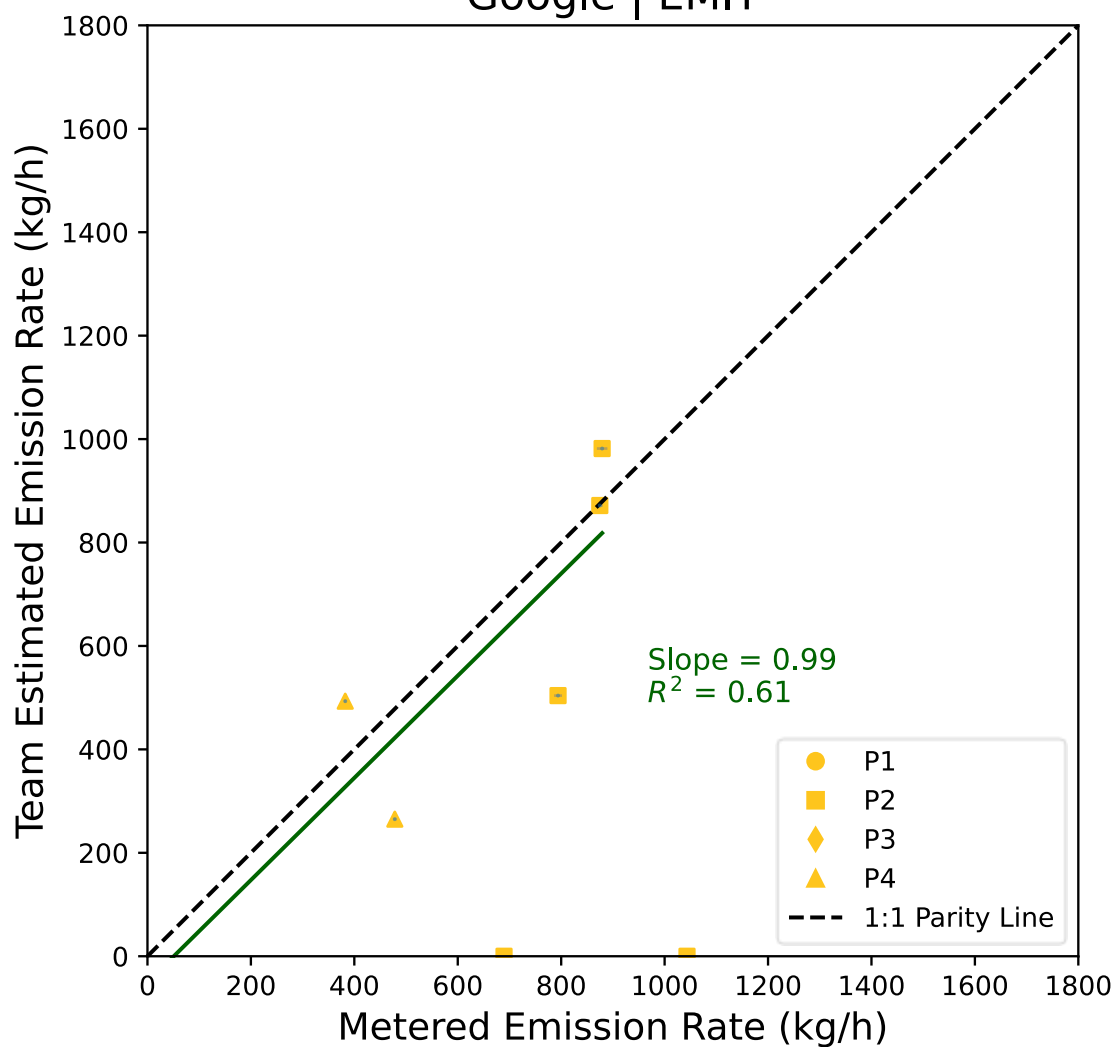


Figure S68. Quantification performance for Google, with 95% X confidence intervals. Google did not submit emission rate uncertainty intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

1165

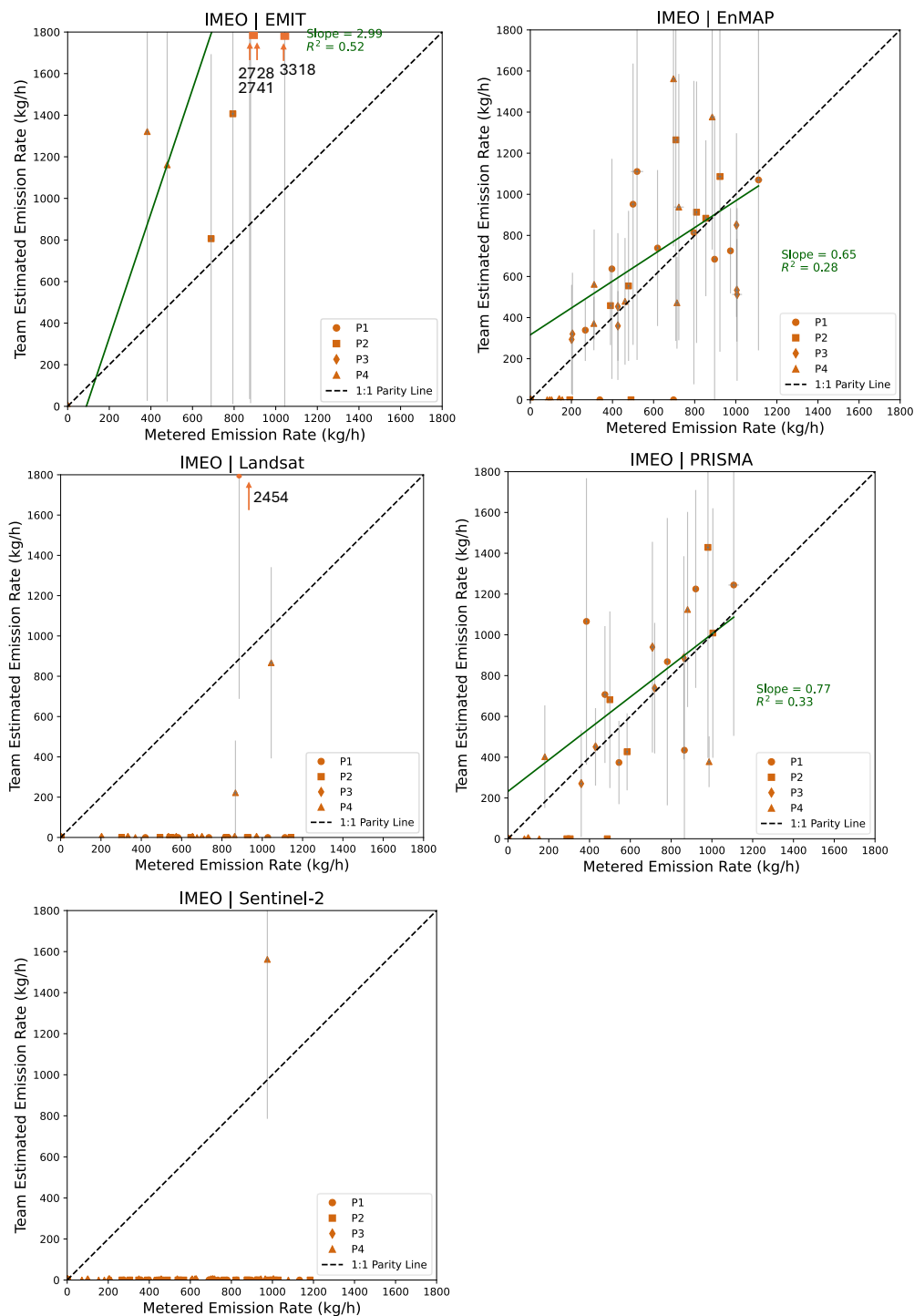


Figure S69. Quantification performance for IMEO, with 95% X and Y confidence intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line). IMEO's quantification methodology for EMIT changed between Phases 2 and 4 (see S2.2).

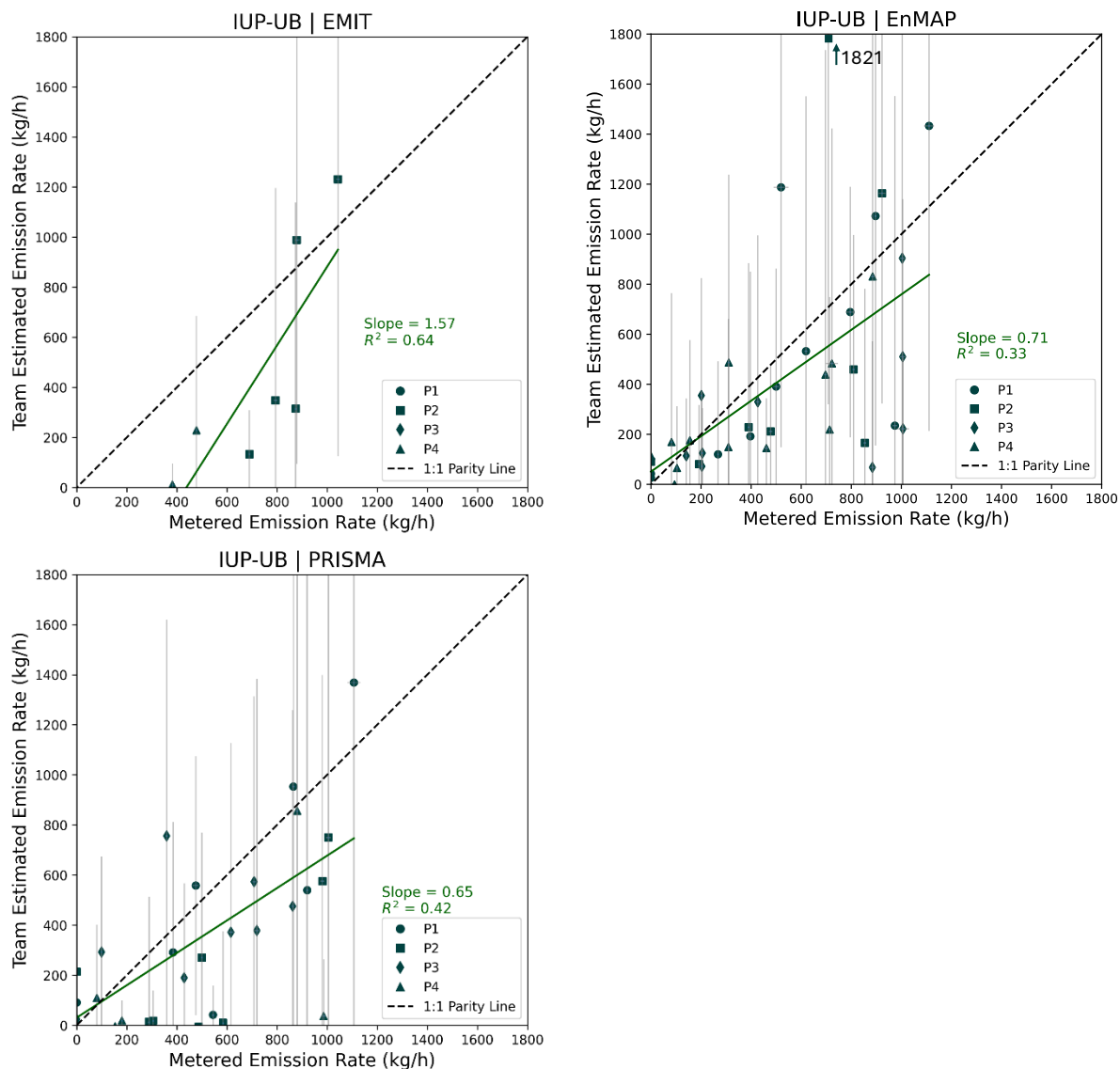
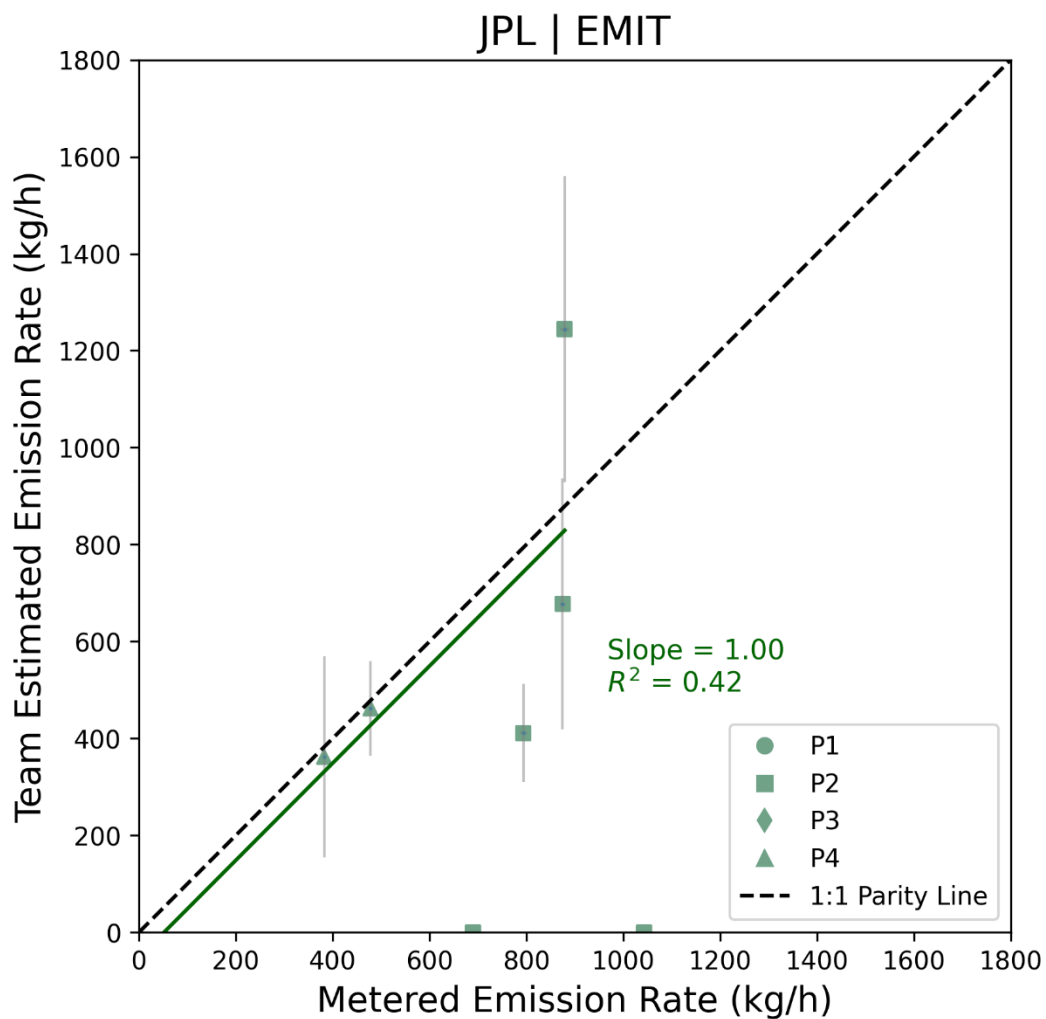


Figure S70. Quantification performance for IUP-UB, with 95% X and Y confidence intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

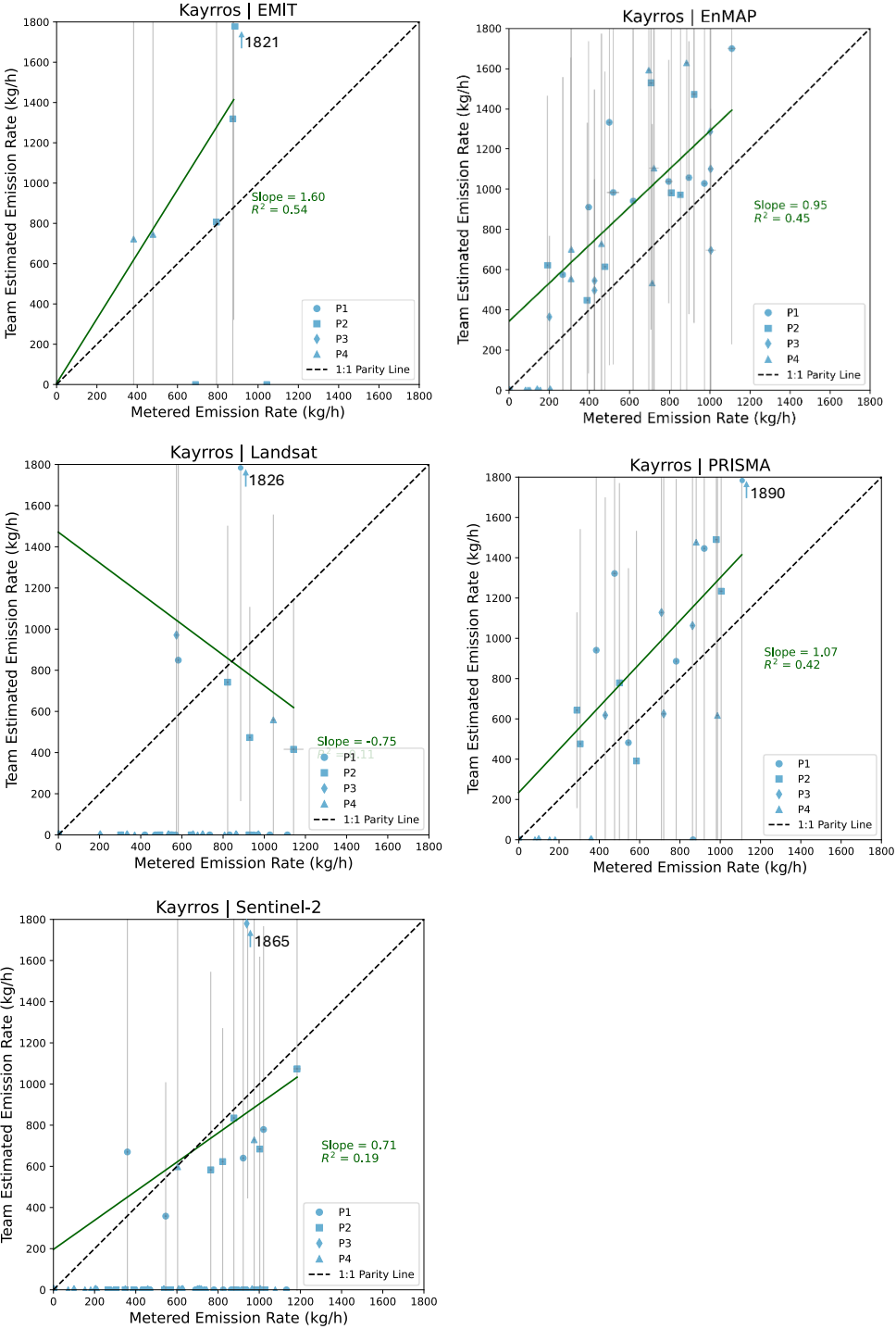
Note: Interpreting FP detections by IUP-UB with the EnMAP and PRISMA instruments requires consideration of their analytical method. Their cross-sectional flux method does not include a plume detection step and instead directly estimates an emission rate (Q). As agreed upon with the analyst team, any final $Q > 0$ was classified as a detection for this study even if the 1-sigma uncertainty interval included zero. If a more conservative threshold were applied that considers estimates close to zero as non-detections, these FPs would be eliminated and the number of reported TPs reduced.



1185

Figure S71. Quantification performance for JPL, with 95% X and Y confidence intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

1190



1195

Figure S72. Quantification performance for Kayrros, with 95% X and Y confidence intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

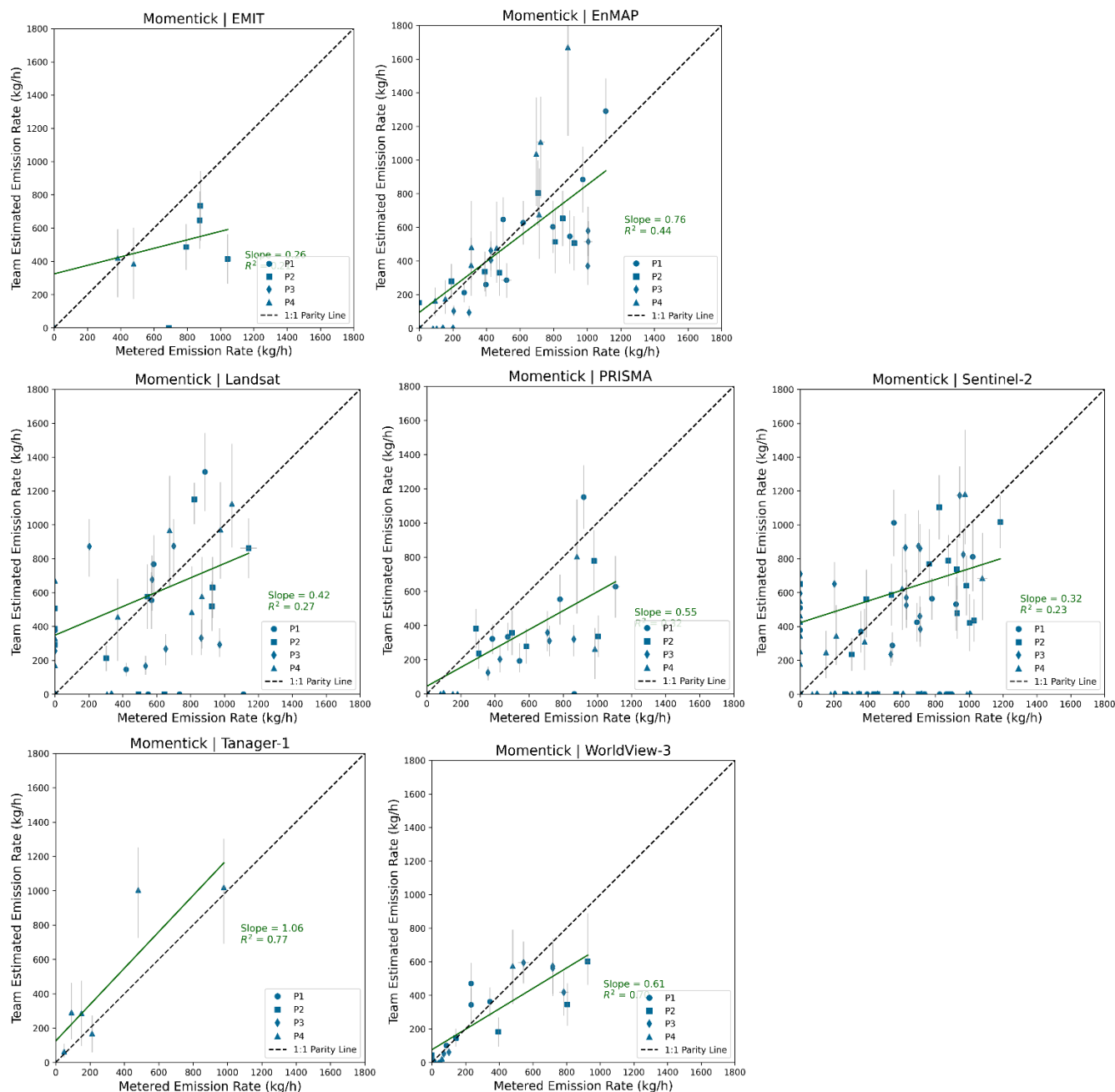


Figure S73. Quantification performance for Momentick, with 95% X confidence intervals. Momentick submitted minimum and maximum estimates, so their error bars reflect this range rather than a 95% confidence interval. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

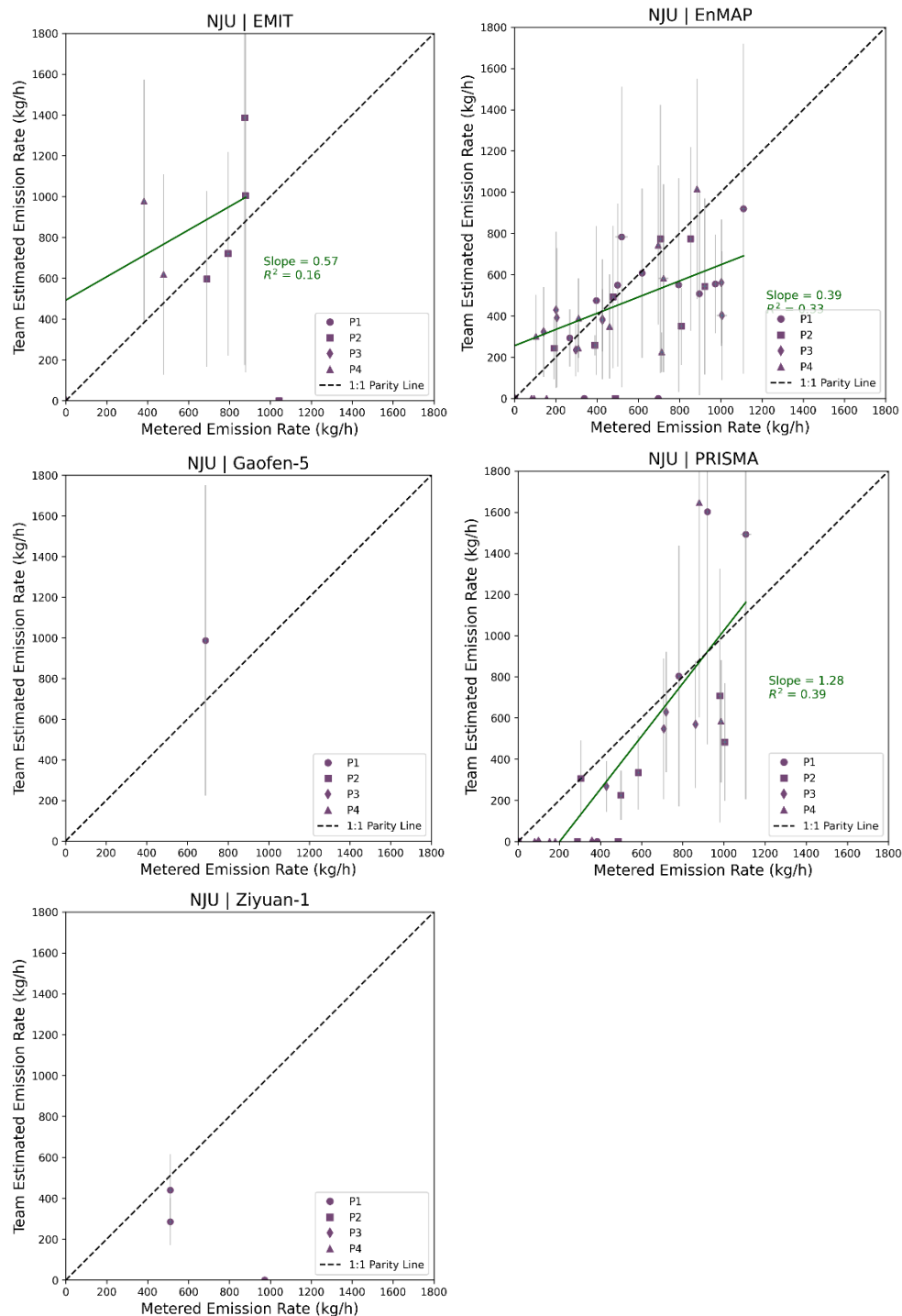
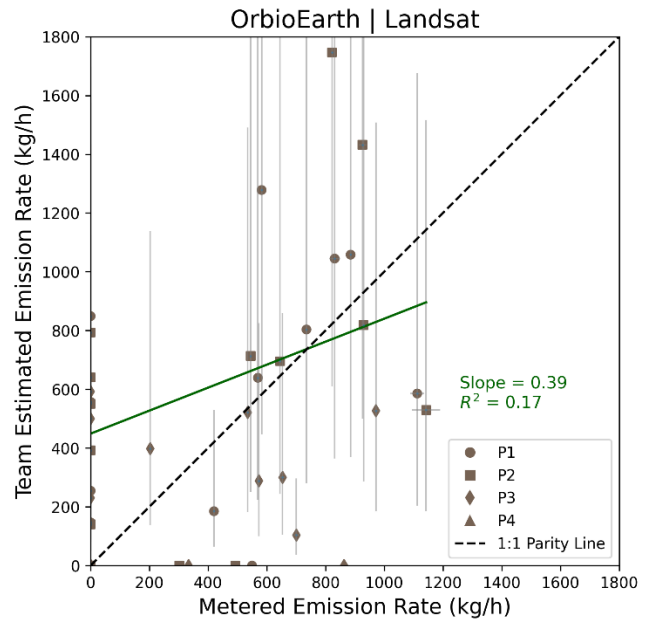
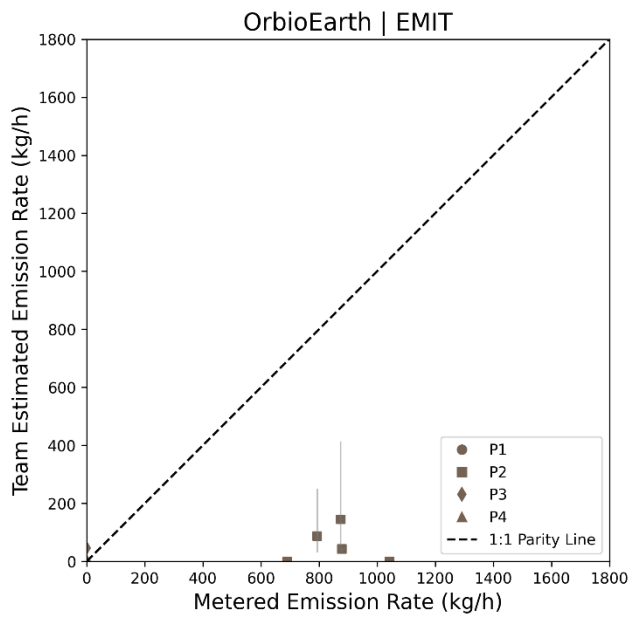


Figure S74. Quantification performance for NJU, with 95% X and Y confidence intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

1210



1215

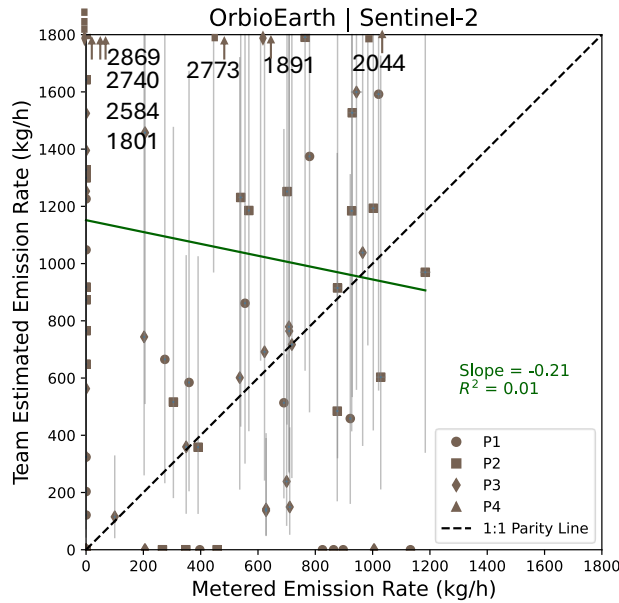


Figure S75. Quantification performance for Orbio Earth, with 95% X confidence intervals. Orbio Earth submitted a 2-sigma uncertainty statistic that was not always symmetric about the mean; rather than adjust these intervals, we used the reported values as provided. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

OSK | GHOST

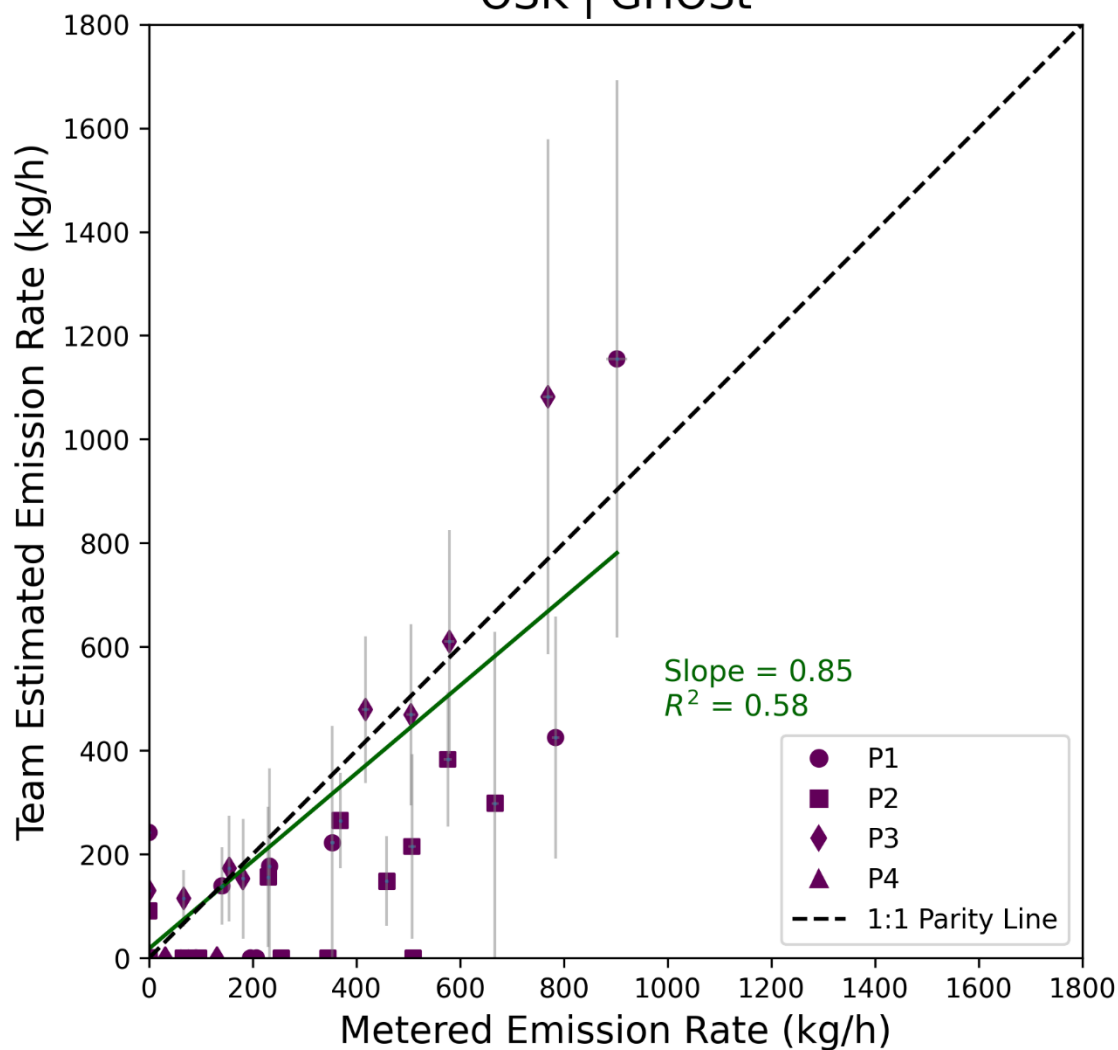
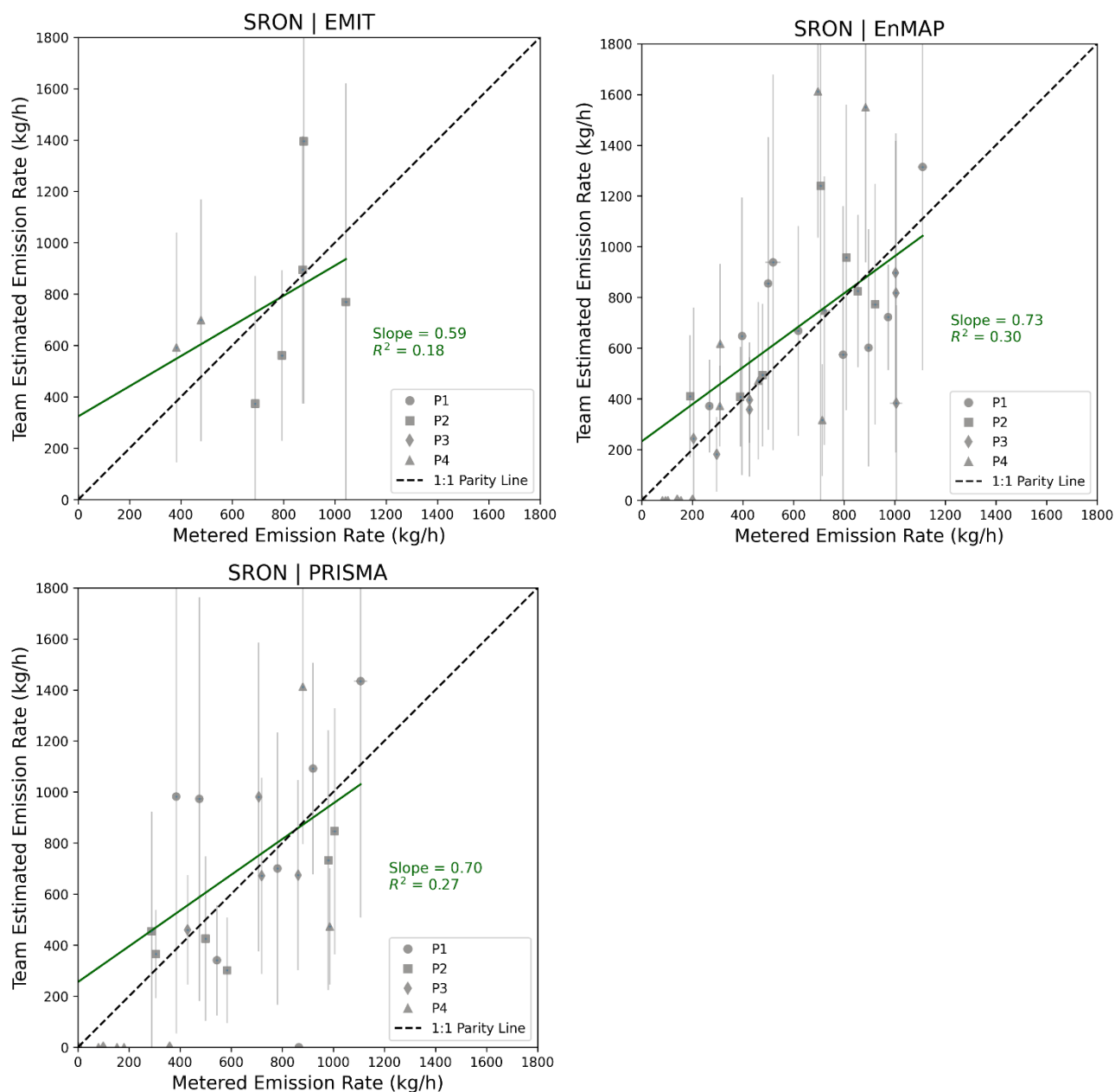


Figure S76. Quantification performance for Orbital Sidekick, with 95% X and Y confidence intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).



1225 **Figure S77.** Quantification performance for SRON, with 95% X and Y confidence intervals. The black dashed
 1230 line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green
 line). Note: in one instance on Apr 6, 2025, SRON quantified an EMIT plume although they did not categorize it
 as a confident detection. In the main text, this is considered a False Negative point and not included in
 quantification fits. However, their emission rate estimate is included in the fit here. See Notes column in
 submitted data sheets.

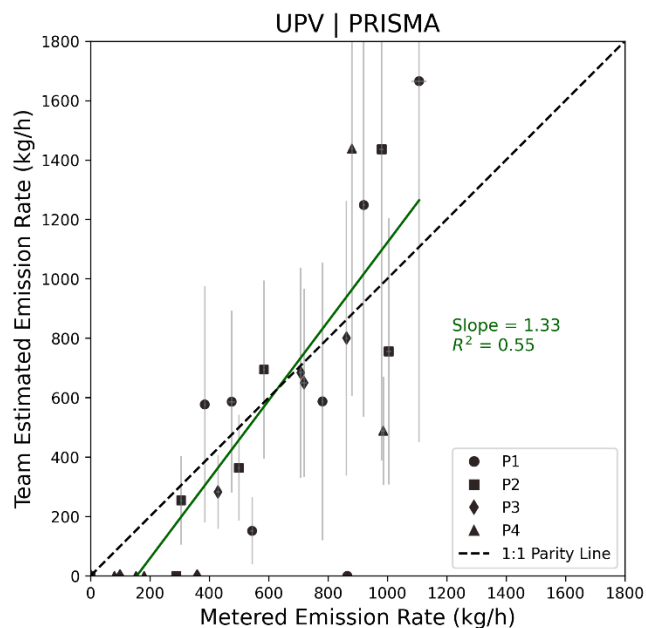
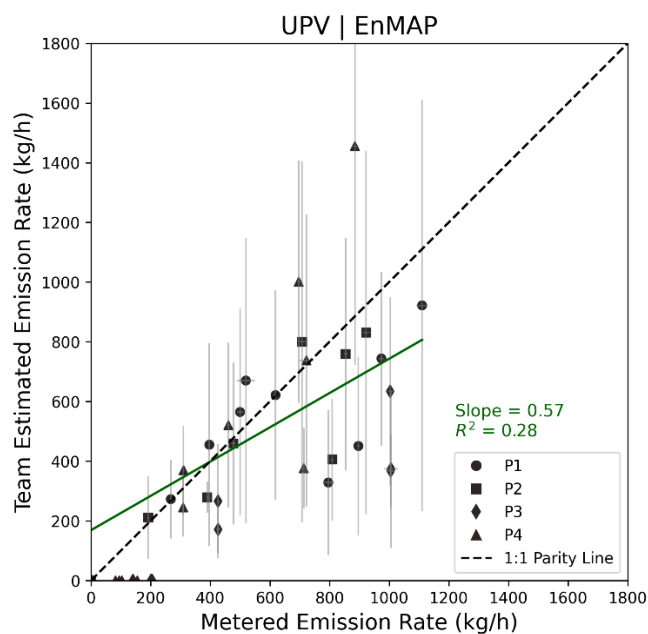
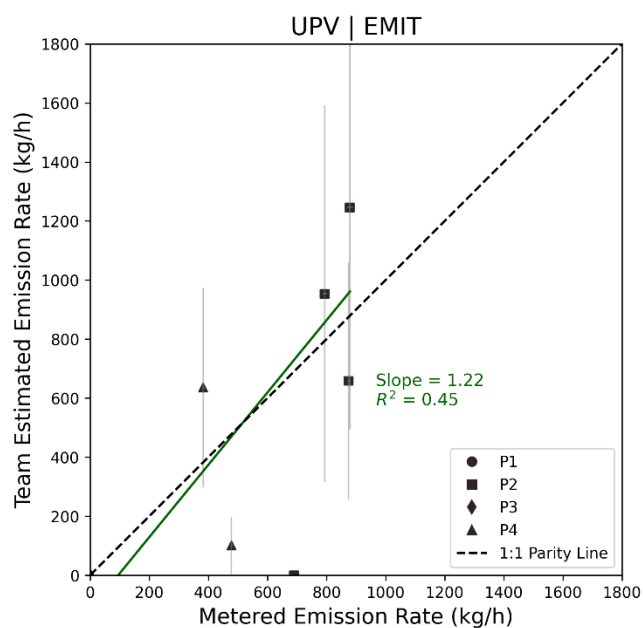


Figure S78. Quantification performance for UPV, with 95% X and Y confidence intervals. The black dashed line denotes 1:1 agreement. Fitted slope and R^2 shown for ordinary least squares fit to quantified points (green line).

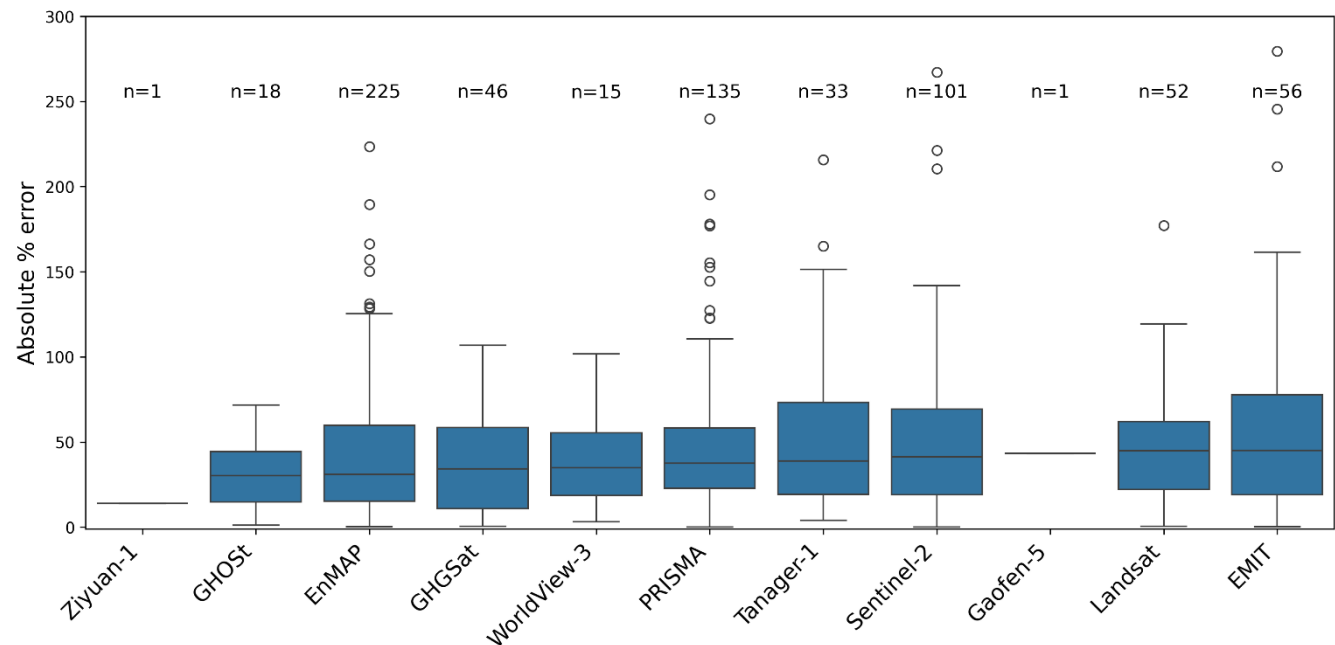


Figure S79. Absolute value of percent error for quantified estimates by satellite, ordered by median.

S3.9.4 Ensemble performance analysis

1250 To investigate whether an ensemble estimate from multiple teams outperforms individual estimates—a "wisdom of crowds" effect—we performed an aggregation analysis for all metered release events observed by public satellites (EMIT, EnMAP, PRISMA, Landsat, and Sentinel-2).

1255 For each release event, we evaluated the effect of creating an ensemble of k teams, where k ranged from 1 to the maximum number of participating teams (N). For each k , all possible combinations of teams were generated, and the mean of their emission estimates was calculated. False Negatives (non-detections) were included as zero-value estimates.

1260 We then calculated the signed percent error (to assess bias) and the absolute percent error (to assess precision) for the mean of each combination against the metered ground truth. The final error metrics plotted in Figures S80 and S81 represent the average of these errors across all combinations of size k and all release events for a given satellite.

1265 The signed percent error (Figure S80) remains largely constant across k , with fluctuations attributable to the unbalanced participation of teams across different release events. This reinforces the conclusion that ensemble averaging primarily reduces random error, not systematic bias. For example, EMIT and EnMAP ensembles consistently show a slight high bias, while PRISMA is nearly unbiased in the aggregate. The apparent low bias for Sentinel-2 and Landsat is influenced by the high number of zero-value estimates from non-detections.

1270 Mean Absolute Percent Error (MAPE) (Figure S81) demonstrates a "wisdom of crowds" effect. For EMIT, EnMAP, and PRISMA, MAPE declines as k increases from 1 to ~ 4 , after which the curve flattens, indicating diminishing returns. This shows that aggregating even a few independent analyses improves the precision and reliability of the quantification. Sentinel-2 and Landsat show a similar trend for $k=1-3$, although this is limited by the smaller number of teams providing estimates.

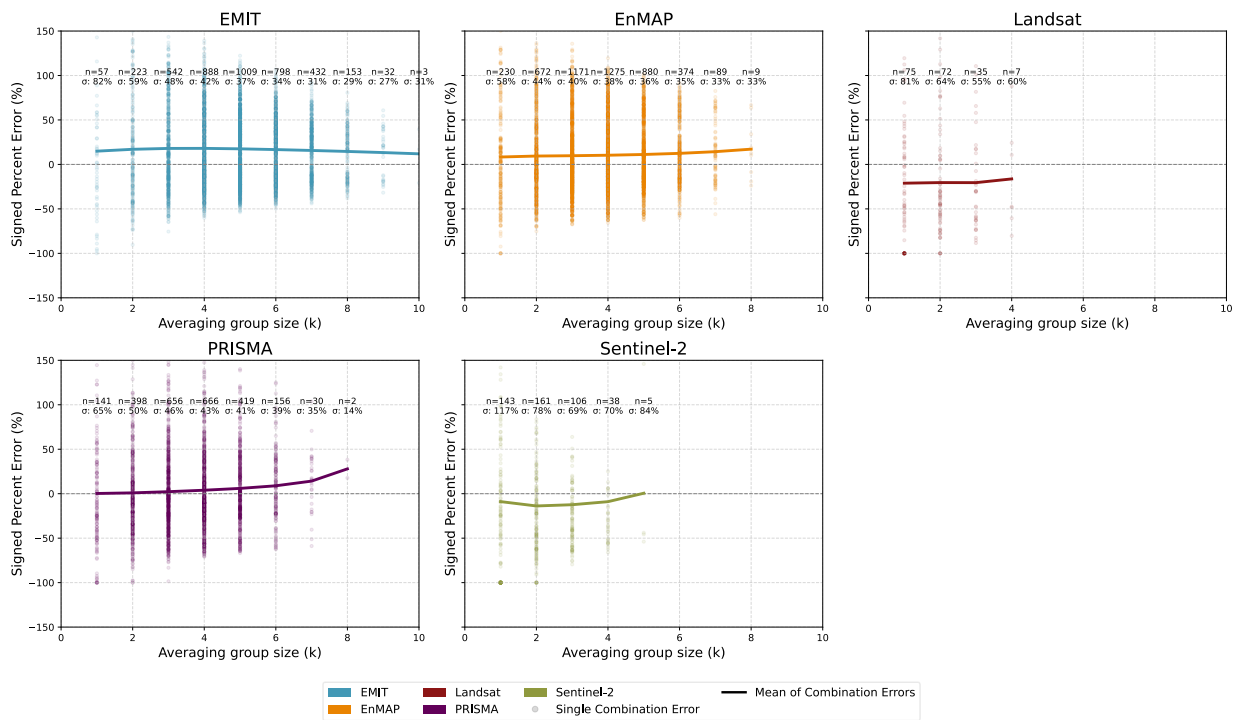


Figure S80. Signed percent error across averaging group sizes k . For each release event, we evaluated the effect of creating an ensemble of k teams, where k ranged from 1 to the maximum number of participating teams (N). For each k , all possible combinations of teams were generated, and the mean of their emission estimates was calculated. The signed percent error remains largely constant across k , with fluctuations attributable to the unbalanced participation of teams across different release events. This reinforces the conclusion that ensemble averaging primarily reduces random error, not systematic bias.

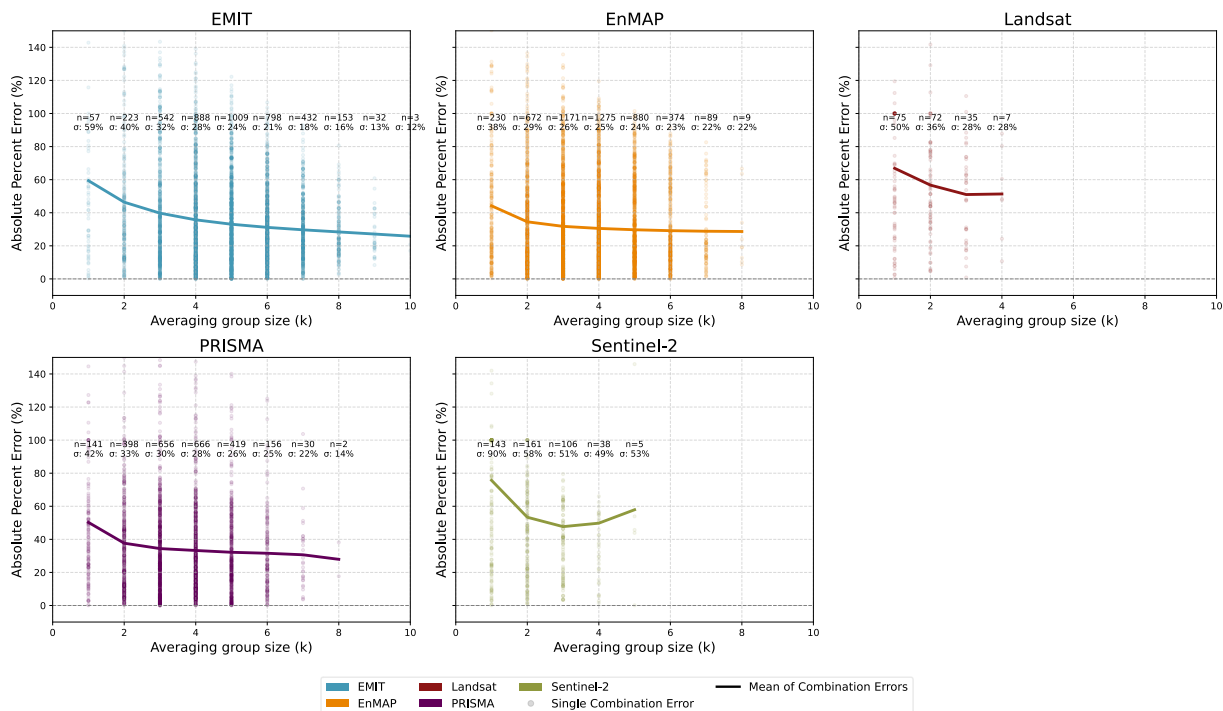


Figure S81. Mean absolute percent error across averaging group sizes k . For each release event, we evaluated the effect of creating an ensemble of k teams, where k ranged from 1 to the maximum number of participating teams (N). For each k , all possible combinations of teams were generated, and the mean of their emission estimates was calculated. For EMIT, EnMAP, and PRISMA, MAPE declines as k increases, demonstrating a "wisdom of crowds" effect.

S3.9.5 Performance metrics ranking

We present satellite-team rankings of five performance metrics in Table S9 with the following columns:

- **Balanced Accuracy rank.** Ranked from highest (best) to lowest. Balanced Accuracy the average of Sensitivity and Specificity (see Table S6). Balanced Accuracy rankings are not provided for pairings where either the Sensitivity or Specificity metric could not be calculated.
- **p _value rank.** Ranked from the lowest p -value (highest statistical significance of skill) to the highest. p _value is the probability of a team achieving their Observed Accuracy (see Table S6) or better, under the null hypothesis where no detection skill exists.
- **Slope deviation rank.** Ranked from smallest deviation (best) to largest. We calculated the absolute difference between the fitted slope and an ideal value of 1.0. Rankings are only provided for pairings with sufficient data (>5 datapoints, see Table S8).
- **Mean absolute percent error rank.** Ranked from lowest percent error (best) to highest. Mean absolute percent error is calculated as the average distance of quantified estimates from the metered rate, relative to metered rate size, on quantified True Positive detections only.
- **Mean absolute error rank.** Ranked from lowest (best) to highest. Mean absolute error is calculated as the average distance of quantified estimates from the metered rate on True Positive detections only.

Multiple metrics related to both detection and quantification were intentionally chosen to avoid over-reliance on any single performance indicator. Each metric emphasizes different things. For example, p_value is aided by larger sample sizes. Rankings by absolute error may favor systems that observed smaller releases and rankings by percent error may favor those that observed larger ones.

The rankings for each metric are independent. They are based on numerical value only and do not imply that the difference between any two ranks is statistically significant.

Table entries are colored by satellite category (Group A, B, C). This facilitates visual inspection of overall trends, allowing for a quick assessment of where teams may have over performed in a given instrument grouping.

Table S9. Performance metric rankings by satellite and team. Groupings are colored (Group A purple, Group B yellow, Group C green).

Balanced accuracy rank	p_value rank	Slope rank	Absolute % error rank	Absolute error rank
(tie)	(tie)			(1) GHGSat - GHGSat
(1) Tanager-1 – Carbon Mapper	(1) Tanager-1 - CarbonMapper	(1) EMIT - JPL	(1) EMIT - JPL	(2) WorldView-3 - Momentick
(1) Tanager-1 - Momentick	(1) GHGSat - GHGSat	(2) EMIT - Google	(2) EMIT - Google	(3) EMIT - Google
(1) EnMAP - ARKA	(1) Landsat - Momentick	(3) Tanager-1 - CarbonMapper	(3) EMIT - Momentick	(4) GHOS - OSK
(1) PRISMA - ARKA	(1) GHOS - OSK	(4) EnMAP - Kayrros	(4) EnMAP - UPV	(5) Tanager-1 - Momentick
(1) EMIT - IMEO	(5) Sentinel-2 - Momentick	(5) Tanager-1 - Momentick	(5) GHOS - OSK	(6) Tanager-1 - CarbonMapper
(1) EMIT - IUP-UB	(6) Sentinel-2 - Kayrros	(6) PRISMA - Kayrros	(6) Sentinel-2 - Kayrros	(7) EMIT - JPL
(7) EnMAP - NJU	(7) EnMAP - NJU	(7) GHOS - OSK	(7) PRISMA - UPV	(8) EnMAP - NJU
(tie)	(8) Landsat - Kayrros	(8) GHGSat - GHGSat	(8) EnMAP - Momentick	(9) EnMAP - Momentick
(8) EMIT - SRON				(10) EnMAP - UPV
(8) EMIT - NJU	(9) EnMAP - SRON	(9) PRISMA - ARKA	(9) Sentinel-2 - Momentick	(11) PRISMA - IMEO
(10) EnMAP - SRON	(10) EnMAP - Kayrros	(10) EnMAP - ARKA	(10) GHGSat - GHGSat	(12) EnMAP - IMEO
(11) EMIT - UPV	(11) PRISMA - IMEO	(11) EMIT - UPV	(11) PRISMA - NJU	(13) EnMAP - SRON
(12) EnMAP - Kayrros	(12) EnMAP - UPV	(12) PRISMA - IMEO	(12) EMIT - SRON	(14) EMIT - Momentick
(13) PRISMA - IMEO	(13) EnMAP - IMEO	(13) EnMAP - Momentick	(13) EnMAP - IMEO	(15) PRISMA - UPV
(14) GHGSat - GHGSat	(14) PRISMA - Momentick	(14) EnMAP - SRON	(14) EnMAP - SRON	(16) PRISMA - SRON
(15) EnMAP - IMEO	(15) Sentinel-2 - Geolabe	(15) PRISMA - NJU	(15) PRISMA - IMEO	(17) Sentinel-2 - Kayrros
(16) EnMAP - UPV	(16) Landsat - OrbioEarth	(16) EnMAP - IUP-UB	(16) EnMAP - NJU	(18) EMIT - SRON
(17) PRISMA - Momentick	(17) PRISMA - SRON	(17) Sentinel-2 - Kayrros	(17) WorldView-3 - Momentick	(19) EMIT - NJU
(tie)	(18) PRISMA - Kayrros	(18) PRISMA - SRON	(18) PRISMA - SRON	(20) EnMAP - IUP-UB
(18) PRISMA - SRON				(21) PRISMA - IUP-UB
(18) PRISMA - Kayrros	(19) PRISMA - UPV	(19) Sentinel-2 - Geolabe	(19) PRISMA - Momentick	(22) EMIT - UPV
(20) EMIT - Kayrros	(20) EnMAP - ARKA	(20) PRISMA - UPV	(20) EMIT - UPV	(23) Sentinel-2 - Momentick
(21) PRISMA - UPV	(21) EnMAP - Momentick	(21) EnMAP - IMEO	(21) Tanager-1 –CarbonMapper	(24) Landsat - Momentick
(22) PRISMA - NJU	(22) PRISMA - NJU	(22) PRISMA - IUP-UB	(22) EMIT - NJU	(25) PRISMA - Momentick
(23) EnMAP - Momentick	(23) EMIT - IMEO	(23) WorldView-3 - Momentick	(23) Landsat - OrbioEarth	(26) PRISMA - NJU
(24) Landsat - Momentick	(24) EMIT - IUP-UB	(24) EnMAP - UPV	(24) Landsat - Momentick	(27) EnMAP - Kayrros
(25) GHOS - OSK	(25) Sentinel-2 - OrbioEarth	(25) EMIT - NJU	(25) EnMAP - IUP-UB	(28) EnMAP - ARKA
(26) Sentinel-2 - Geolabe	(26) WorldView-3 - Momentick	(26) PRISMA - Momentick	(26) EMIT - IUP-UB	(29) EMIT - IUP-UB
(27) WorldView-3 - Momentick	(27) EMIT - SRON	(27) EMIT - SRON	(27) Landsat - Kayrros	(30) PRISMA - Kayrros
(28) Landsat - OrbioEarth	(28) Landsat - IMEO	(28) Landsat - Momentick	(28) Sentinel-2 - Geolabe	(31) Landsat - OrbioEarth
(29) EMIT - Momentick	(29) EMIT - NJU	(29) Landsat - OrbioEarth	(29) PRISMA - IUP-UB	(32) EMIT - Kayrros
(30) Sentinel-2 - Momentick	(30) PRISMA - ARKA	(30) EMIT - IUP-UB	(30) PRISMA - Kayrros	(33) Sentinel-2 - Geolabe
(31) EnMAP - IUP-UB	(31) Tanager-1 - Momentick	(31) EMIT - Kayrros	(31) EnMAP - ARKA	(34) Landsat - Kayrros
(32) Sentinel-2 - OrbioEarth	(32) EMIT - Kayrros	(32) EnMAP - NJU	(32) EMIT - Kayrros	(35) PRISMA - ARKA
(33) Landsat - Kayrros	(33) EnMAP - IUP-UB	(33) Sentinel-2 - Momentick	(33) EnMAP - Kayrros	(36) Sentinel-2 - OrbioEarth
(34) Sentinel-2 - Kayrros	(34) EMIT - UPV	(34) EMIT - Momentick	(34) Tanager-1 - Momentick	(37) Sentinel-2 - ARKA
(35) PRISMA - IUP-UB	(35) EMIT - Momentick	(35) Sentinel-2 - ARKA	(35) Sentinel-2 - OrbioEarth	(38) EMIT - IMEO
(36) Landsat - IMEO	(36) PRISMA - IUP-UB	(36) Sentinel-2 - OrbioEarth	(36) PRISMA - ARKA	
(37) EMIT - OrbioEarth	(37) Sentinel-2 - ARKA	(37) Landsat - Kayrros	(37) EMIT - IMEO	
(38) Sentinel-2 - ARKA	(38) Sentinel-2 - IMEO	(38) EMIT - IMEO		
(39) Sentinel-2 - IMEO	(39) EMIT - OrbioEarth			
(tie)	(tie)			
(40) Landsat - NJU	(40) Landsat - NJU			
(40) Sentinel-2 - NJU	(40) Sentinel-2 - NJU			
	(40) EMIT - Google			
	(40) EMIT - JPL			
	(40) Gaofen-5 - NJU			
	(40) Ziyuan-1 - NJU			

S4. Supplementary methods

Figure S82. Annotated Google Earth imagery of the Casa Grande release site.
Maps data: Google Earth, Landsat/Copernicus



1325

Figure S83. Photograph of the release stacks, pressure regulation unit, and CNG trailer



1330 **S4.1 Coriolis meter calibration**

The Emerson Micro Motion Coriolis flow meters used in this study were calibrated by the manufacturer according to the schedules described in El Abbadi et al. (2024)³⁹. These meters are known for their durability, with minimal drift expected under typical field conditions. To avoid potential risks from uninstalling, shipping, and reinstalling the meters for laboratory recalibration, the research team developed a field-based verification method.

An Apollo High Capacity Balance (model GX-102000LS) was used to validate meter performance. The scale has a 102 kilogram capacity, repeatability of 1 gram, and accuracy of ± 10 grams, and includes a built-in function for calculating flow rates. Inert gas (N_2) from 50-pound cylinders flowed through the small (CMFS015H, 0.166 inch) and medium (CMF050M, 0.5 inch) Coriolis meters, while CNG was disconnected. Flow was tested at 5, 10, 15, and 20 kg/h on the small meter, and at 50 and 75 kg/h on the medium meter. Each test run lasted a minimum of five minutes, and average flow rate values were calculated over that time window. A wind shield was used to minimize cylinder rocking and associated weight fluctuations on the scale. Table S10 summarizes the five-minute averaged flow rates as recorded by both the scale and the Coriolis meters, along with the intended flow setpoints. Figure S85 shows time series for the 50 kg/h test case. The two readings show good agreement: five-minute average flow measured by the scale was 54.63 kg/h, while the Coriolis meter reported 54.85 kg/h (0.4% difference).

The high-flow meter, used for all releases greater than 300 kg/h, could not be calibrated using the same methods as the small and medium meters. Instead, we used a mass-balance analysis over the year-long experiment. This method reconciles the total mass of purchased gas against the sum of all known metered and unmetered outputs from the system.

Twenty CNG trailer refills were conducted, corresponding to a total purchased gas mass of 50,053 kg of natural gas, accounting for the gas composition specific to each purchase event. We use gas composition as reported by the pipeline operator and averaged over 5 days before the refill, as in the main text. Fill volumes are reported as SCF, and we use the GPSA-recommended conversion value of 1.1953 moles per SCF to convert to moles then to mass.

1360 Mass balance terms:

1. 20 fill events, excluding a fill on afternoon of January 1st 2025. This fill was not included as this was before the first release of the study.
2. Difference between initial and final mass loading of the trailers. An initial pressure before the first release of the study was noted to be 2280 psig in experimental notes (F. Reuland). This resulted in initial gas storage of 3693 kg. Final gas mass in storage on December 31st was 803 kg, derived from 603 psig reading.

3. The sum of all metered releases conducted for satellite, aircraft, and continuous monitoring tests, after January 1st 2025. Time series of 1 Hz mass flow rates summed over the year account for 91% of filled trailer mass.
- 1370 4. Heater system consumption is estimated using the time to flow plus 15 minutes to pre-heat the heaters before flow begins. A total of 8 CATCO catalytic heaters are installed on the RT-20 regulation skids. These are 1500 W heaters with consumption rates of 11.8 scf/h.
- 1375 5. Trailer fitting leaks: two leaks on trailer fittings occurred in June and July 2025. Mass emissions rates were estimated based on the time required to notice the issue and procure repair equipment (approx. 20 days), and we assume a leak rate of 0.1 kg/h for each event; these were small events that were not noticeable from far away by refill technician, but were noticeable upon approaching the rear of the trailer to disconnect the pressure hose.
- 1380 6. Pressure release valve (PRV) leak: a persistent small leak from a pressure release valve on the RT-20 pressure regulation trailer was observed by multiple ground sensors (Konica on June 11; QLM continuously). Based on these observations, we assume an ongoing leak rate of approximately 0.25 kg/h throughout the experiment.
- 1385 7. Stack vent: a small seep from the stack vent was first noted by a Kuva camera on August 1 and was estimated to have a rate of approximately 1 kg/h.
8. Heater system leaks: small fugitive leaks from the system's CATCO heaters and associated piping were noticed by the KonicaMinolta team, and occasionally by odorant by experiment team onsite. KonicaMinolta quantification estimate was a combined rate of approximately 0.25 kg/h.

Summing these terms accounts for ~98.5-100% of the total purchased gas mass, depending on various duration assumptions for the small leaks noted above. This mass balance closure, combined with the demonstrated accuracy of the identically calibrated small and medium meters, provides strong confidence in the accuracy of the high-flow meter. It is reasonable to assume that any residual error in this meter is small compared to the much larger uncertainties in satellite-based source rate quantification.

Figure S84. Field site Coriolis meter calibration with gas cylinder on scale.

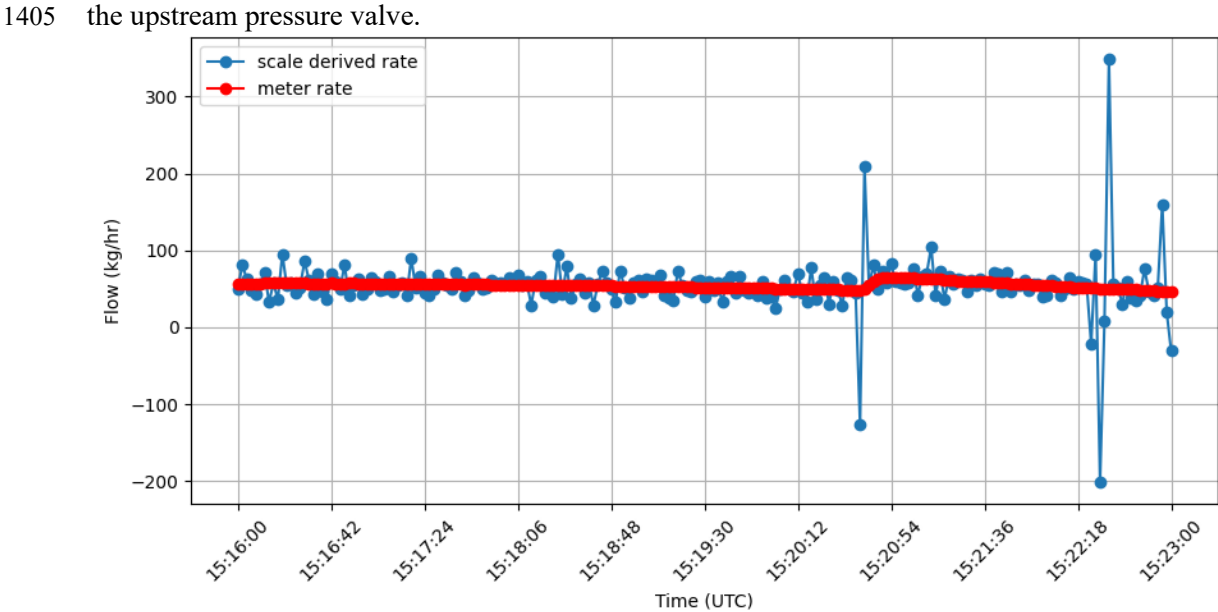


1400 **Table S10. Coriolis meter calibration results of scale-derived flow rates compared to meter-reported flow rates**

Test set point (kg/h)	Scale-measured 5-min flow rate average	Meter-reported 5-min flow rate average
5	5.508	5.097
10	10.056	9.886
15	15.773	15.21
20	19.20	19.23
50	54.63	54.848
75	71.789	72.363

Figure S85. Coriolis meter calibration time series gas flow for the 50 kg/h test case

Note: A small increase in meter rate (red) around the 15:20:50 UTC mark corresponds to a manual adjustment of the upstream pressure valve.



S4.2 Meteorological equipment

Power for the 2-, 5-, and 10-meter sensors, CR1000X Datalogger, sky camera, and soil probes was provided by a PV system with four 100 Watt (W) monocrystalline SunShare panels. The panels were connected in parallel via a 4-to-1 pigtail configuration and fed into a Tracer3210AN maximum power point tracking (MPPT) solar charge controller, rated for a maximum current of 30 Amps (A). Energy storage was supplied by two Li Time 100 Amp-hour (Ah), 12.8 Volt (V) lithium iron phosphate (LiFePO₄) batteries wired in parallel, providing reliable overnight and low-sunlight power for continuous operation. The ZX 300 LIDAR had a dedicated PV system with two 410 W Jinko panels and MidNite Solar MNPV3 combiner. Power was regulated through Morningstar Tristar solar charge controllers and supplemented by a Meanwell PB-600-24 three-stage switching-mode battery charger. The battery system consists of ten 99 Ah gel batteries (2-s 5-p) housed in an insulated enclosure. Power was delivered to the LIDAR via 6 AWG armored cabling.

Table S11. Meteorological equipment install and reinstall dates and recorded data outages

Equipment	Install date	Outages
Sky camera	Install Feb 18, 2025 (Phase 1)	February 21 and 22, 2025 before 18:20:00 UTC March 11, 2025 April 8, 10-12, 2025 June 20, 2025 July 4-11, 17, 31, 2025 August 21-28, 31 2025 September 2-16, 25-30, 2025 October 1, 6, 8, 30 2025
2m wind sensor	Before start of Phase 1	Oct 8, 2024 – Oct 16, 2024 (Phase 0) April 10-12, 2025
5m ultrasonic anemometer	Apr 9, 2025 (Phase 2)	April 10-12, 2025
10m ultrasonic anemometer	Before start of Phase 1	Oct 8, 2024 – Oct 16, 2024 (Phase 0) April 8, 10-12, 2025
Soil moisture sensor 1	Before start of Phase 1	Oct 8, 2024 – Oct 16, 2024 (Phase 0) April 8, 11-12
Soil moisture sensor 2	Before start of Phase 1	Oct 8, 2024 – Oct 16, 2024 (Phase 0) April 8, 11-12
ZX Lidar wind data	Jan 13, 2025 (Phase 1)	Periodic gaps overnight to mid-morning due to power loss on: January 29, 30, 31, February 1, 2, 4, 9, 11, 12, 13, 15, 16, 17, April 8, 10

		May 7, May 8, May 10, May 17, May 25, May 31, 2025, June 27, July 9
--	--	---

Figure S86A-C. A. Wematics sky camera and Metsens500 wind sensor mounted on tripod. B. Campbell Scientific 3-D ultrasonic anemometer mounted at 10 m height. C. ZX Lidar Unit

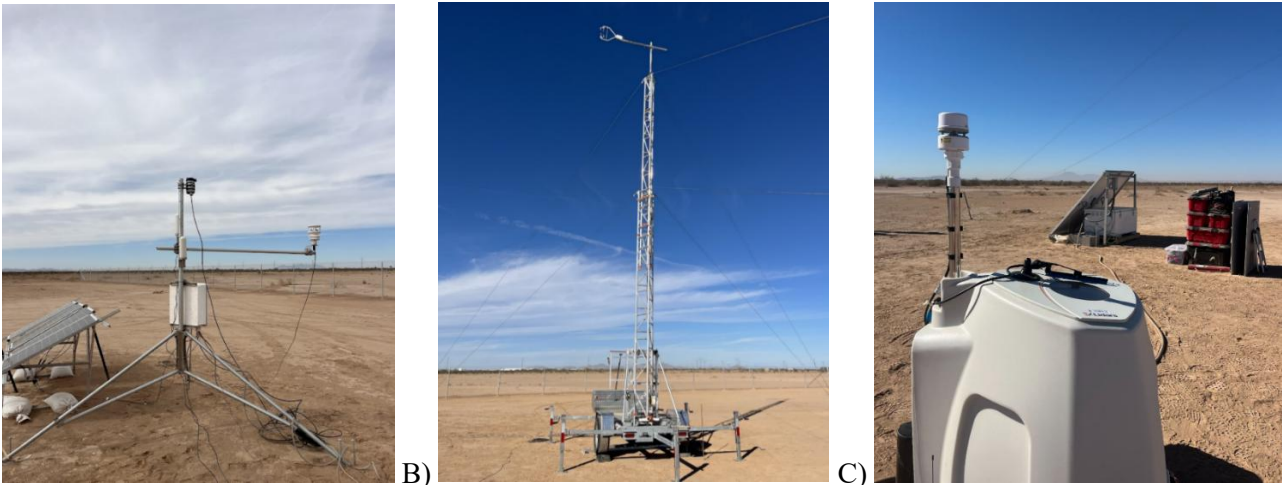
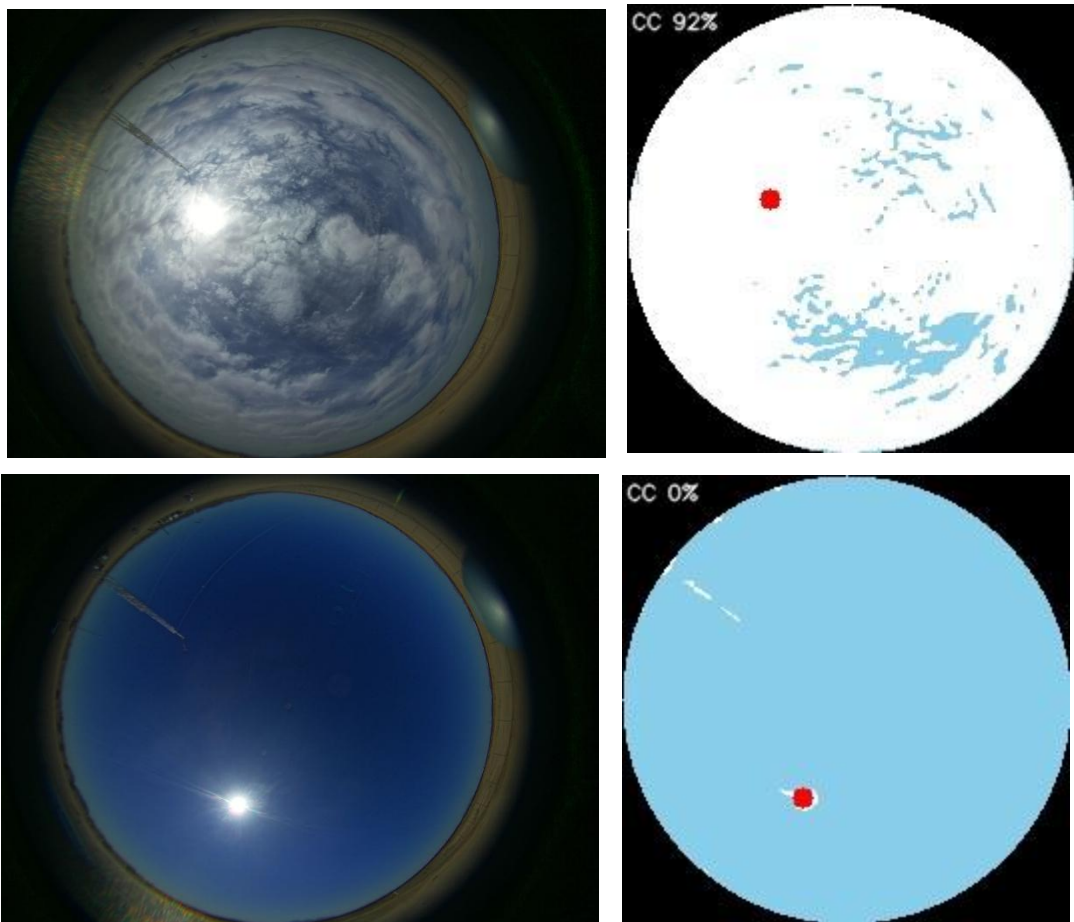


Figure S87. Sky camera HDR images and cloud masks on two days in March 2025



1430

1435

1440

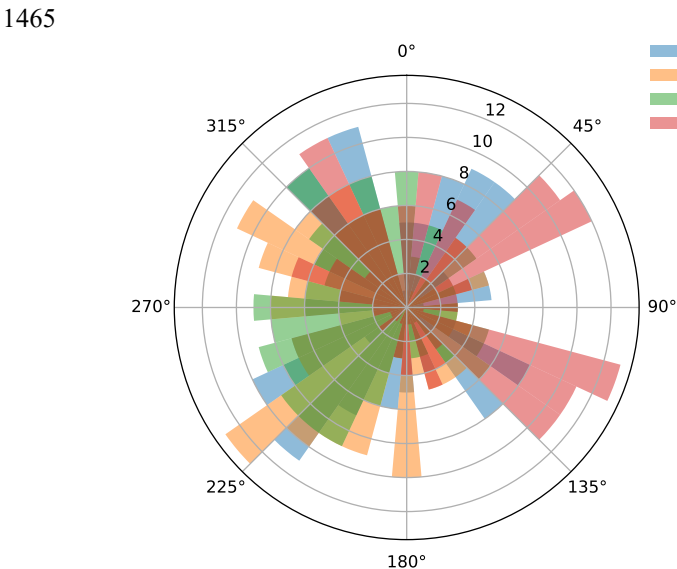
S4.3 Site meteorology during each study phase

1445 This section details the meteorological conditions at the release site during the study period. The histograms presented below (Figures S88-S93) display key variables, time-averaged over the 5-minute window preceding each satellite overpass. These satellite observations primarily occurred between 11:00 and 15:00 local time due to the nature of the polar-orbiting satellite orbits common among these observing systems.

1450 The study spanned one calendar year and distinct seasonal profiles are visible. Across Phase 1, 2, and 3, the predominant wind direction was from the southwest. In Phase 4 the wind direction shifted to be more variable with more winds from the southeast. Phase 1 was characterized by the coolest temperatures, with midday averages around 20°C, and had a higher frequency of low wind speeds (< 2 m/s) compared to the other phases. Phase 2 represented the driest period of the study and, correspondingly, experienced the highest solar irradiances. Phase 3 was the hottest and most humid, with temperatures regularly surpassing 35°C, which aligns with the region's late summer monsoon season.

1460 Overall cloud cover remained extremely low year-round. This is one reason the site is considered to have idealized conditions. Note that a small number of solar irradiance measurements below 200 W/m² are present in the dataset. These values are abnormally low for midday, clear-sky conditions and likely correspond to sensor interference (e.g., from dust storms or temporary dust accumulation) or transient sensor failure rather than true atmospheric conditions. For future analyses using this public dataset, we recommend filtering out these specific data points.

Figure S88. Polar histogram of 10 m wind direction during overpasses, by phase



1470

Figure S89. Histogram of 10 m wind speed during overpasses, by phase

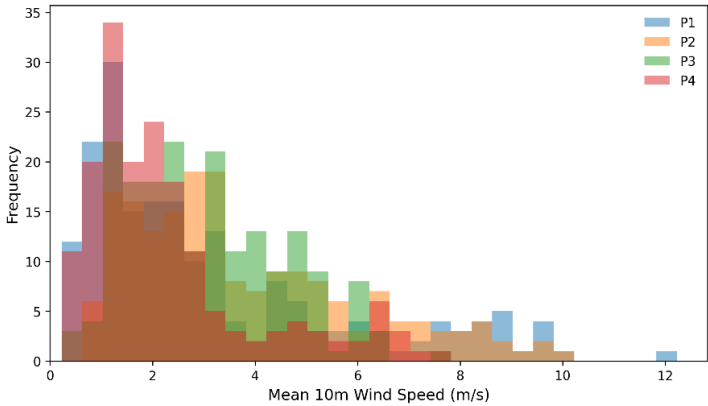
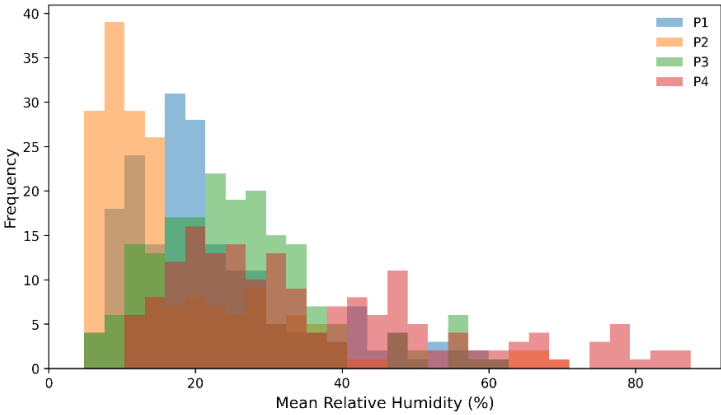
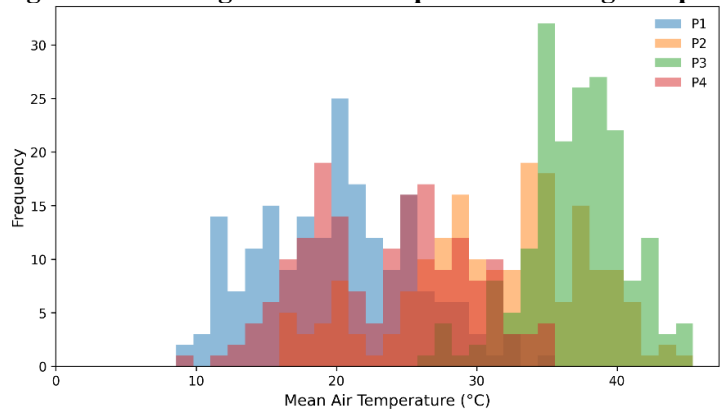


Figure S90. Histogram of relative humidity by phase



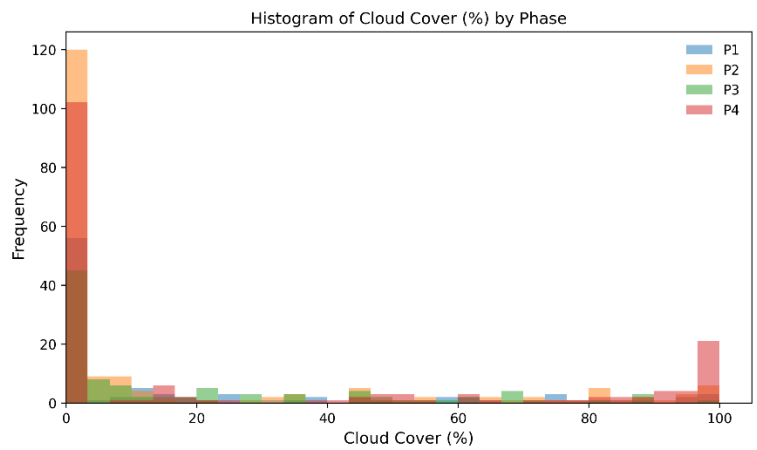
1475

Figure S91. Histogram of air temperature during overpasses, by phase

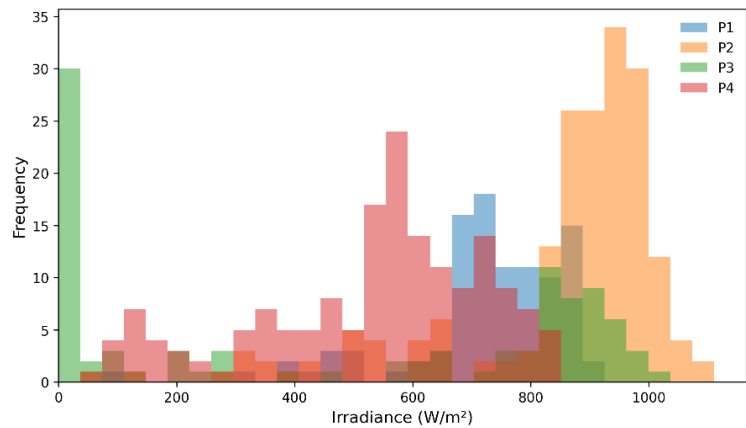


1480

Figure S92. Histogram of cloud cover % during overpasses, by phase



1485 **Figure S93. Histogram of irradiances (W/m2) during overpasses, by phase**



1490

S4.4 Gas composition

Gas composition data were obtained from the two nearest gas composition measurement stations located upstream of the CNG fill point in Casa Grande, AZ. We obtained datafiles for gas composition provided by the gas pipeline supplier that serves the CNG station. Trailer refills were performed by trained and licensed staff from Armadillo Trucking. Stanford researchers maintained detailed logs of trailer fill dates and usage. These records allowed us to associate each gas release event with the correct trailer fill and corresponding gas composition. Batch numbers start at 12 because 11 refills took place prior the start of blind testing during the five-month testing period occurring in Evanston, WY and Casa Grande, AZ.

Table S12. Fill dates and methane composition for each gas batch at Casa Grande

Batch number	Fill date	CH ₄ content (avg %, std dev %)
12	1/1/2025	93.4, 0.149
13	1/27/2025	92.7, 0.448
14	2/16/2025	93.1, 0.189
15	3/6/2025	94.2, 1.09
16	3/20/2025	95.0, 0.291
17	4/18/2025	96.5, 0.215
18	5/11/2025	95.5, 0.694
19	5/23/2025	95.6, 0.443
20	6/9/2025	95.1, 0.224
21	6/25/2025	96.1, 0.267
22	7/3/2025	95.4, .0278
23	7/22/2025	94.5, 0.408
24	8/8/2025	94.9, 0.609

25	8/25/2025	95.0, .0791
26	9/16/2025	95.9, 0.575
27	10/10/2025	95.5, 0.221
28	10/30/2025	95.0, 0.142
29	11/23/2025	95.8, 0.226
30	12/16/2025	96.5, 0.314

S4.5 Advancing Development of Emissions Detection Protocol

1505 Some of this section has been reproduced from Sherwin et al. 2024 with permission.

The single-blind controlled methane release testing in this study followed the Advancing Development of Emissions Detection (ADED) protocol, developed at the Methane Emissions Technology Evaluation Center (METEC) in Colorado ⁴⁰

1510 We followed protocols from Section 10, “Aerial Survey Emission Detection And Quantification,” which was designed to apply to remote sensing technologies in general, including satellites.

1515 Documentation of the system under test is included both within this paper as well as in the “Performer Info” tab of the data spreadsheets submitted by each team, which are available in the data and code repository. Submitted emission estimates used a standard template spreadsheet modified based on the Aerial Survey Emission Detection and Quantification version of the ADED protocol. See Table S13 and S14 for the quantitative and qualitative reporting fields.

1520 Following Section 10.4.1 of the ADED protocol, the test location was at least 1 km away from all potential confounding methane sources, e.g. the local landfill, and from all nearby water features.

Table S13. Single-blind quantitative reporting template sheet columns and descriptions

Column	Description	Unit	Type
InstrumentName	Name of satellite instrument or platform		Drop down
DateAnalysis	Date the analysis form is completed, format MM/DD/YYYY	MM/DD/YYYY	User entered

DateObs	Date of satellite observation attempt, format MM/DD/YYYY	MM/DD/YYYY	User entered
ObsTimestamp	Time of overpass, UTC	UTC	User entered
StartTime	If applicable, start time for overpass/acquisition, UTC	UTC	User entered
EndTime	If applicable, end time for overpass/acquisition, UTC	UTC	User entered
ImagingMode	If applicable, specify imaging mode for this observation, e.g. pushbroom, specific target mode, glint, etc.		User entered
OffNadirAngle	Off-nadir sensor viewing angle, in degrees	degrees	User entered
ViewingAzimuth	Viewing azimuth angle, in degrees clockwise from 0 as North	degrees	User entered
SolarZenith	Solar Zenith Angle at time of satellite observation, in degrees	degrees	User entered
SolarAzimuth	Solar Azimuth Angle at time of satellite observation, in degrees clockwise from 0 N	degrees	User entered
DetectionAttempt	Indicate whether plume detection attempt is FEASIBLE or INFEASIBLE. Note, DetectionAttempt may be feasible while Detection is FALSE. If infeasible, please select reason or use Notes Column to indicate.		Drop down
Detection	Indicate whether an emissions plume is detected (TRUE) or not (FALSE). Leave blank if DetectionAttempt (col L) was indicated to be infeasible. Detection may be TRUE even if quantification attempt is not made.		Binary (True/False)
QuantAttempt	Indicate whether an emissions rate quantification attempt is made. If emissions rate estimate is not made, please select reason or use Notes Column to indicate.		Drop down
Gas	Gas species measured		Drop down
PlumeOriginLat	Latitude coordinate of estimated origin for controlled release plume	decimal degrees	User entered
PlumeOriginLon	Longitude coordinate of estimated origin for controlled release plume	decimal degrees	User entered
PlumeLength	Plume length included in quantification, in meters	meters	User entered
IME	Integrated mass enhancement, in kg	kilograms	User entered
EmissionRate	Enter the mass emission rate in kg/hr. Enter 0 to signify no emissions are detected in a successful observation. Enter N/A	kilograms/hour	User entered

	if emission rate estimate is not attempted. Use drop down or notes to indicate reason.		
EmissionRateUpper	Enter the upper bound estimate of mass emission rate in kg/hr	kilograms/hour	User entered
EmissionRateLower	Enter the lower bound estimate of mass emission rate in kg/hr	kilograms/hour	User entered
EmissionRateUncertaintyType	The form of uncertainty presented (e.g. min/max, 95% confidence interval, standard deviation) for EmissionRate.		User entered
U10WindSpeed	If applicable, report here the 10-meter altitude wind speed estimate used in quantifying emission rate, in meters per second.	meters/sec	User entered
UeffWindSpeed	If applicable, report the Ueff windspeed parameter used in quantification, in meters per second.	meters/sec	User entered
WindSpeedUpper	If applicable, enter the upper uncertainty estimate of Ueff wind speed in m/s.	meters/sec	User entered
WindSpeedLower	Enter the lower uncertainty estimate of Ueff wind speed in m/s.	meters/sec	User entered
WindSpeedUncertaintyType	The form of uncertainty presented for WindSpeed (e.g. min/max, 95% confidence interval, standard deviation).		User entered
WindDirection	In degrees measured clockwise relative to 0 as geographic North. Here, the angle corresponds to the direction FROM where the wind comes. For example, 180 degree wind comes from the South.	degrees	User entered
Notes	Leave any other notes in this cell		User entered

1525

Table S14. Single-blind methods description reporting sheet

Question

(1) Company/Organization name

(2) Please provide a detailed description of system configuration and primary components including the sensor(s) and satellite platform(s). If any auxiliary equipment (such as ground meteorological stations) is used please specify.

(3) Please record the software revision installed on the components in (2), including performer-specific software components, revisions, or customizations, where relevant.

(4) Please record the revision number of any data processing software used in analysis. For example: retrieval software version or source rate estimation tools. If multiple software versions are used throughout this study, please specify what software versions apply to which dates.

(5) If your equipment is capable of multiple imaging modes, please describe what they are and which imaging mode(s) is used in this study.

(6) Please provide a detailed description of the methodology used during emission detection and quantification surveys.

*If multiple distinct methodologies are used, please use the Notes column in the Overpass Summary tab to specify which method applies on which dates.

(7) Please describe your method for obtaining methane concentration retrieval.

(8) Please describe your method of plume delineation/segmentation (e.g., plume mask generation).

(9) Please describe how plume length is determined for quantification, if applicable.

(10) If wind speed is used in computing total emission rate, please describe how the wind estimate is obtained, including the precise instrument or wind reanalysis product used and if/how Ueff is generated.

(11) Please record the number of personnel participating in the surveys and their roles. Names of individual personnel are not required.

(12) Did any major changes or discontinuities occur during the study that are relevant to the main analysis (i.e. change in satellite orbit altitude, major change to data processing methodology)? If so, please describe what the change was and when it occurred.

Supplement references

- 1530 1. 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. Stanford Digital Repository <https://doi.org/10.25740/QH001QT3946> (2025).
2. GHGSat. Emission Sensing Technology - GHGSat. <https://www.ghgsat.com/en/technology/> (2025).
3. Planet. Hyperspectral Imaging Satellites | Planet. <https://www.planet.com/products/hyperspectral/>.
- 1535 4. Carbon Mapper. Carbon Mapper Coalition's First Methane-Sensing Satellite to Launch August 16th. <https://carbonmapper.org/articles/carbon-mapper-coalitions-first-methane-sensing-satellite-to-launch> (2024).
5. Duren, R. *et al.* The Carbon Mapper emissions monitoring system. *Atmos. Meas. Tech.* **18**, 6933–6958 (2025).

6. 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).
1540 doi:10.1109/AERO47225.2020.9172731.
7. NASA JPL. VISIONS: The EMIT Open Data Portal. <https://earth.jpl.nasa.gov/data/data-portal/coverage-and-forecasts> (2025).
8. Earth Observation Center of DLR. Mission - EnMAP. <https://www.enmap.org/mission/> (2023).
9. DLR, EnMAP & Federal Ministry of Research, Technology, and Space. EnMAP Instrument Planning.
1545 <https://planning.enmap.org/>.
10. Liu, Y.-N. *et al.* The Advanced Hyperspectral Imager: Aboard China’s GaoFen-5 Satellite. *IEEE Geosci. Remote Sens. Mag.* **7**, 23–32 (2019).
11. CRESDA. WMO OSCAR | Space Agency: CRESDA. <https://space.oscar.wmo.int/spaceagencies/view/cresda>.
12. eoPortal. GHOSt (Global Hyperspectral Observation Satellite) - eoPortal. [https://www.eoportal.org/satellite-](https://www.eoportal.org/satellite-missions/ghost#eop-quick-facts-section)
1550 [missions/ghost#eop-quick-facts-section](https://www.eoportal.org/satellite-missions/ghost#eop-quick-facts-section) (2024).
13. Orbital Sidekick. Orbital Sidekick - Spectral Intelligence. <https://www.orbitalsidekick.com/>.
14. OHB Italia. Satellites & Missions: PRISMA. *PRISMA* <https://www.ohb-italia.it/satellites-missions/> (2022).
15. Agenzia Spaziale Italiana. PRISMA. <https://www.asi.it/en/earth-science/prisma/>.
16. Vantor. WorldView Tasking Suite – On-Demand Satellite Imagery of Earth and Space | Vantor.
1555 <https://vantor.com/product/worldview/>.
17. 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).
18. Sherwin, E. D. *et al.* Single-blind test of nine methane-sensing satellite systems from three continents. *Atmos. Meas. Tech.* **17**, 765–782 (2024).
19. Copernicus. Sentinel-2 | Copernicus Data Space Ecosystem. [https://dataspace.copernicus.eu/data-collections/copernicus-](https://dataspace.copernicus.eu/data-collections/copernicus-sentinel-missions/sentinel-2)
1560 [sentinel-missions/sentinel-2](https://dataspace.copernicus.eu/data-collections/copernicus-sentinel-missions/sentinel-2).
20. USGS. Landsat 8 | U.S. Geological Survey. <https://www.usgs.gov/landsat-missions/landsat-8> (2023).

21. 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).
- 1565 22. Jervis, D. *et al.* The GHGSat-D imaging spectrometer. *Atmos. Meas. Tech.* **14**, 2127–2140 (2021).
23. Setchell, H. ECMWF Reanalysis v5. (2020).
24. Irakulis-Loitxate, I., Gorroño, J., Zavala-Araiza, D. & Guanter, L. Satellites Detect a Methane Ultra-emission Event from an Offshore Platform in the Gulf of Mexico. *Environ. Sci. Technol. Lett.* **9**, 520–525 (2022).
25. Guanter, L. *et al.* Mapping methane point emissions with the PRISMA spaceborne imaging spectrometer. *Remote Sens. Environ.* **265**, 112671 (2021).
- 1570 26. Varon, D. J. *et al.* Quantifying methane point sources from fine-scale satellite observations of atmospheric methane plumes. *Atmos. Meas. Tech.* **11**, 5673–5686 (2018).
27. Roger, J., Guanter, L., Gorroño, J. & Irakulis-Loitxate, I. Exploiting the entire near-infrared spectral range to improve the detection of methane plumes with high-resolution imaging spectrometers. *Atmos. Meas. Tech.* **17**, 1333–1346 (2024).
- 1575 28. Pei, Z. *et al.* Improving quantification of methane point source emissions from imaging spectroscopy. *Remote Sens. Environ.* **295**, 113652 (2023).
29. 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.
30. Akani-Guéry:, J., Roger, J. & Buchwitz, M. ESA Project MEDSUA: Algorithm Intercomparison Report (AIR). (2026).
- 1580 31. Eldering, A. Methane Plume Detection and Quantification Intercomparison Analysis Based on EMIT Measurements. <https://datafoundation.org/news/reports/812/812-Methane-Plume-Detection-and-Quantification-Intercomparison-Analysis-Based-on-EMIT-Measurements->.
32. OpenWeather. OpenWeather APIs for Weather, Energy, and Environment. https://openweathermap.org/payload/api/media/file/clouds_bg.jpg.
- 1585 33. Ramier, A. *et al.* Performance validation of the GHGSat Methane Constellation via controlled releases. Preprint at <https://doi.org/10.31223/X5N78R> (2026).

34. Millstein, D., Jeong, S., Ancell, A. & Wiser, R. A database of hourly wind speed and modeled generation for US wind plants based on three meteorological models. *Sci. Data* **10**, 883 (2023).
35. Carbon Mapper. Data Dispatch - Seeing around the clouds. <https://carbonmapper.org/articles/seeing-around-the-clouds> (2025).
36. Qu, Z. *et al.* Global distribution of methane emissions: a comparative inverse analysis of observations from the TROPOMI and GOSAT satellite instruments. *Atmos. Chem. Phys.* **21**, 14159–14175 (2021).
37. International Energy Agency. Methane Tracker – Data Tools - IEA. <https://www.iea.org/data-and-statistics/data-tools/methane-tracker>.
38. Sherwin, E. D. *et al.* Single-blind validation of space-based point-source detection and quantification of onshore methane emissions. *Sci. Rep.* **13**, 3836 (2023).
39. El Abbadi, S. H. *et al.* Technological Maturity of Aircraft-Based Methane Sensing for Greenhouse Gas Mitigation. *Environ. Sci. Technol.* **58**, 9591–9600 (2024).
40. Bell, C. & Zimmerle, D. METEC controlled test protocol: survey emission detection and quantification. (2022) doi:10.25675/10217/235363.